VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF ARUNACHAL PRADESH USING

PYTHON AND R

DEEPTHI ANNA ALEX

V01101949

Date of Submission: 16/06/2024

# CONTENTS

|  |  |
| --- | --- |
| Content: | Page no: |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANC | 3-4 |
| RESULTS AND  INTERPRETATIONS | 4-9 |
| CODES | 10-14 |

Analyzing Consumption in the State of Arunachal Pradesh Using R

# INTRODUCTION

This study focuses on the state of Arunachal Pradesh using NSSO data to identify the top and bottom three districts in terms of consumption. To obtain the necessary data for analysis, we clean and alter the dataset during the process. We have compiled a dataset of consumption-related data, including information on the rural and urban sectors and district-level variances, to make this research easier. R, a potent statistical programming language well-known for its adaptability in managing and analysing big datasets, has received the dataset.   
  
Finding missing values, dealing with outliers, standardizing district and sector names, district- and regional-level summaries of consumption data, and determining the significance of mean differences are some of our goals. The study's conclusions can help legislators and

# OBJECTIVES

1. Check if there are any missing values in the data, identify them and if there are replace them with the mean of the variable.
2. Check for outliers and describe the outcome of your test and make suitable amendments.
3. Rename the districts as well as the sector, viz. rural and urban.
4. Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption.
5. Test whether the differences in the means are significant or not.

# BUSINESS SIGNIFICANCE

The focus of this study on Andhra Pradesh's consumption patterns from NSSO data holds significant implications for businesses and policymakers. By identifying the top and bottom three consuming districts, the study provides valuable insights for market entry, resource allocation, supply chain optimization, and targeted interventions. Through data cleaning, outlier detection, and significance testing, the findings facilitate informed decision-making, fostering equitable development and promoting Andhra Pradesh's economic growth.

# RESULTS AND INTERPRETATION

**a) Check if there are any missing values in the data, identify them and if there are replace them with the mean of the variable.**

*#Identifying the missing values.*

Code and Result:

> # Check for missing values in the subset > cat("Missing Values in Subset:\n")

Missing Values in Subset: > print(colSums(is.na(WBnew)))

state\_1 District Region Sector 0 0 0 0

State\_Region Meals\_At\_Home ricepds\_v Wheatpds\_q 0 111 0 0 chicken\_q pulsep\_q wheatos\_q No\_of\_Meals\_per\_day 0 0 0 5

Interpretation: From the selected variables, after sorting the data for the state of Arunachal Pradesh, it is

seen that only the column ‘Meals\_At\_Home has missing variables. Since missing values in the dataset can be problematic as they lead to incomplete or biased analyses, hindering the accuracy of results and potentially skewing interpretations and decision-making processes.Therefore we replace the missing values with the mean of the variable using following code.

*#Imputing the values, i.e. replacing the missing values with mean.*

Code and Result:

Impute missing values with mean for specific columns

+ impute\_with\_mean <- function(column) {

+ if (any(is.na(column))) {

+ column[is.na(column)] <- mean(column, na.rm = TRUE)

+ }

+ return(column)

+ }

+ arpnew$Meals\_At\_Home <- impute\_with\_mean(arpnew$Meals\_At\_Home)

Interpretation: The above code has successfully replaced the missing values with the mean value of the variable. As can be seen from the result above, there are no missing values in the selected data.

**b) Check for outliers and describe the outcome of your test and make suitable amendments.**

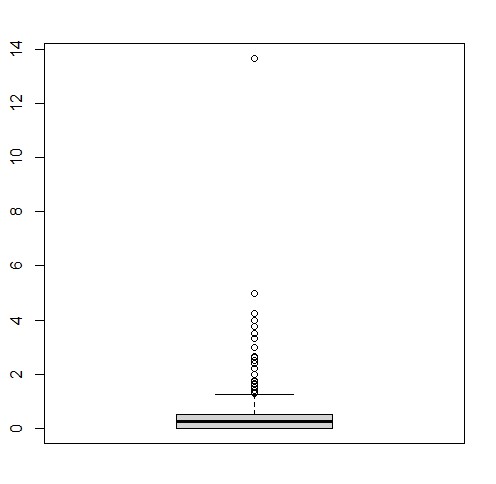
Boxplots can be used to find outliers in the dataset. Boxplots visually reveal outliers in a dataset by displaying individual points located beyond the whiskers of the boxplot.

*#Checking for outliers*

Plotting the boxplot to visualize outliers.

Code and Result:

> boxplot(arpnew$ricepds\_v)



Interpretation: From the boxplot above, which is a visual representation of the variable ‘ricepds\_v’ shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. The outliers can be removed using the following code.

*#Setting quartiles and removing outliers*

Code and results:

Setting quartile ranges to remove outliers

> # Calculate quartiles and IQR

> Q1 <- quantile(arpnew$ricepds\_v, 0.25)

> Q3 <- quantile(arpnew$ricepds\_v, 0.75)

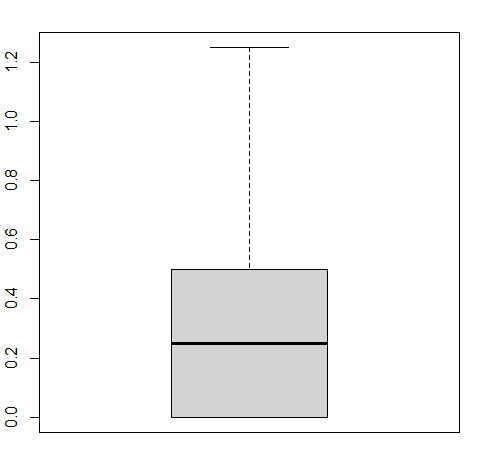
> IQR <- Q3 - Q1

> # Define outlier thresholds

> lower\_threshold <- Q1 - (1.5 \* IQR)

> upper\_threshold <- Q3 + (1.5 \* IQR)

> boxplot(arpnew$ricepds\_v)



Interpretation: Interpreting quartile ranges allows for outlier detection and removal. By calculating the interquartile range (IQR) as the difference between the upper and lower quartiles, data points beyond 1.5 times the IQR from either quartile are identified as outliers and can be excluded or treated to ensure the robustness of the analysis.

In the similar way the outliers in all other variables can be removed

**c) Rename the districts as well as the sector, viz. rural and urban**.

Each district of a state in the NSSO of data is assigned an individual number. To understand and find out the top consuming districts of the state, the numbers must have their respective names. Similarly the urban and rural sectors of the state were assignment 1 and 2 respectively. This is done by running the following code.

Interpretation: The result as show above has successfully assigned the district names to the given number. Also the sectors 1 and 2 have been replaced as urban and rural sectors respectively.

**d) Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption**.

By summarizing the critical variables as total consumption we can estimate the top 3 and bottom 3 consuming districts.

Code and Result:

> arpnew$total\_consumption=

arpnew$ricepds\_v+apnew$Wheatpds\_q+apnew$chicken\_q+apnew$pulsep\_q+apnew$wheatos\_q > apnew%>%

+ group\_by(District)%>%

+ summarise(total=sum(total\_consumption))%>%

+ arrange(-total,District)

Similarly the bottom three districts can be found by sorting the total consumption.

Interpretation:.

**e) Test whether the differences in the means are significant or not.**

The first step to this is to have a Hypotheses Statement.

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.

> rural=arpnew%>%

+ select(Sector,total\_consumption)%>%

+ filter(Sector=="RURAL")

> fix(rural)

> urban=apnew%>%

+ select(Sector,total\_consumption)%>%

+ filter(Sector=="URBAN")

> fix(urban)

> cons\_rural=rural$total\_consumption

> cons\_urban=urban$total\_consumption

> z.test(cons\_rural,

+ cons\_urban,

+ alternative="two.sided",

+ mu=0,

+ sigma.x = 2.56,sigma.y=2.34,

+ conf.level = 0.95)

Result:

Two-sample z-Test

data: cons\_rural and cons\_urban z = 29.202, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval:

1.614254 1.846533 sample estimates: mean of x mean of y

Interpretation:

# CODES

setwd('C:\\Users\\SERVICE POINT\\Desktop\\SCMA\\A1A') getwd() library(dplyr) library(readr) library(readxl) library(tidyr) install.packages("ggplot2") library(ggplot2)

#READING THE FILE INTO R

data=read.csv("4. NSSO68 data set.csv")

#FILTERING FOR ARP df=data%>%

filter(state\_1=="ARP") names(df) head(df) dim(df)

#FINDING MISSING VALUES is.na(df) any(is.na(df)) sum(is.na(df))

sort(colSums(is.na(df)),decreasing=T)

# SUBSETIING arpnew = df%>%

select(state\_1,District,Region,Sector,State\_Region,Meals\_At\_Home,ricepds\_v,Wheatpds\_q,chicken

\_q,pulsep\_q,wheatos\_q,No\_of\_Meals\_per\_day) fix(apnew)

any(is.na(apnew)) sum(is.na(apnew)) head(apnew) sort(colSums(is.na(apnew)),decreasing=T)

#IMPUTING THE VALUES i.e REPLACING MISSING VALUES WITH MEAN arpnew=arpnew%>%

mutate(across(all\_of(c("Meals\_At\_Home")), ~ifelse(is.na(.), mean(., na.rm = TRUE), .))) any(is.na(arpnew)) fix(arpnew)

# FINDING OUTLIERS AND MAKING AMENDMENTS

boxplot(arpnew$ricepds\_v) boxplot(arpnew$Wheatpds\_q) boxplot(arpnew$chicken\_q) boxplot(arpnew$pulsep\_q) boxplot(arpnew$No\_of\_Meals\_per\_day)

# Calculate quartiles and IQR

Q1 <- quantile(arpnew$ricepds\_v, 0.25)

Q3 <- quantile(arpnew$ricepds\_v, 0.75)

IQR <- Q3 - Q1

# Define outlier thresholds lower\_threshold <- Q1 - (1.5 \* IQR) upper\_threshold <- Q3 + (1.5 \* IQR)

apnew = subset(arpnew,apnew$ricepds\_v>=lower\_threshold & arpnew$ricepds\_v<=upper\_threshold) fix(arpnew) boxplot(arpnew$ricepds\_v)

Q1 <- quantile(arpnew$chicken\_q, 0.25)

Q3 <- quantile(arpnew$chicken\_q, 0.75)

IQR <- Q3 - Q1

# Define outlier thresholds lower\_threshold <- Q1 - (1.5 \* IQR) upper\_threshold <- Q3 + (1.5 \* IQR)

apnew = subset(arrpnew,apnew$chicken\_q>=lower\_threshold & apnew$chicken\_q<=upper\_threshold) fix(arpnew) boxplot(arpnew$chicken\_q)

#Renaming the districts as well as the sector, viz. rural and urban.

arpnew$Sector <- ifelse(arpnew$Sector == 2, "URBAN", ifelse(arpnew$Sector == 1, "RURAL",apnew$Sector)) fix(arpnew)

# Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption.

# 1. Districts

arpnew$total\_consumption= arpnew$ricepds\_v+arpnew$Wheatpds\_q+arpnew$chicken\_q+arpnew$pulsep\_q+arpnew$wheatos\_q arpnew%>%

group\_by(District)%>% summarise(total=sum(total\_consumption))%>% arrange(total,District)

# 2. Region arpnew%>% group\_by(Region)%>% summarise(total=sum(total\_consumption))%>% arrange(-total,Region)

# Region 3,1 and 5 are the top 3 consuming regions.

#e) Test whether the differences in the means are significant or not.

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.

rural=arpnew%>% select(Sector,total\_consumption)%>% filter(Sector=="RURAL") fix(rural)

urban=arpnew%>%

select(Sector,total\_consumption)%>% filter(Sector=="URBAN") fix(urban)

cons\_rural=rural$total\_consumption cons\_urban=urban$total\_consumption

length(cons\_rural) length(cons\_urban)

install.packages("BSDA") library(BSDA)

z.test(cons\_rural, cons\_urban, alternative="two.sided", mu=0, sigma.x = 2.56,sigma.y=2.34, conf.level = 0.95)

# P value is <0.05, Therefore we reject the null hypothesis.

#There is difference between mean consumptions of urban and rural.