EP219: Data Analysis and Interpretation

Assignment Report 2

By Team: *Significantly Different*

August/20/2018 - August/27/2018

Contents

Problem Statement

Our aim is to:

1. Extract the excel sheet to pandas data frame
2. Finding the fraction of ODF villages for each district in Uttar Pradesh
3. Making it's histogram. X-axis of histogram should have fraction of villages and y-axis should have number of Districts.
4. Finding Mean and Sample variance of given data and showimg it on the plot.

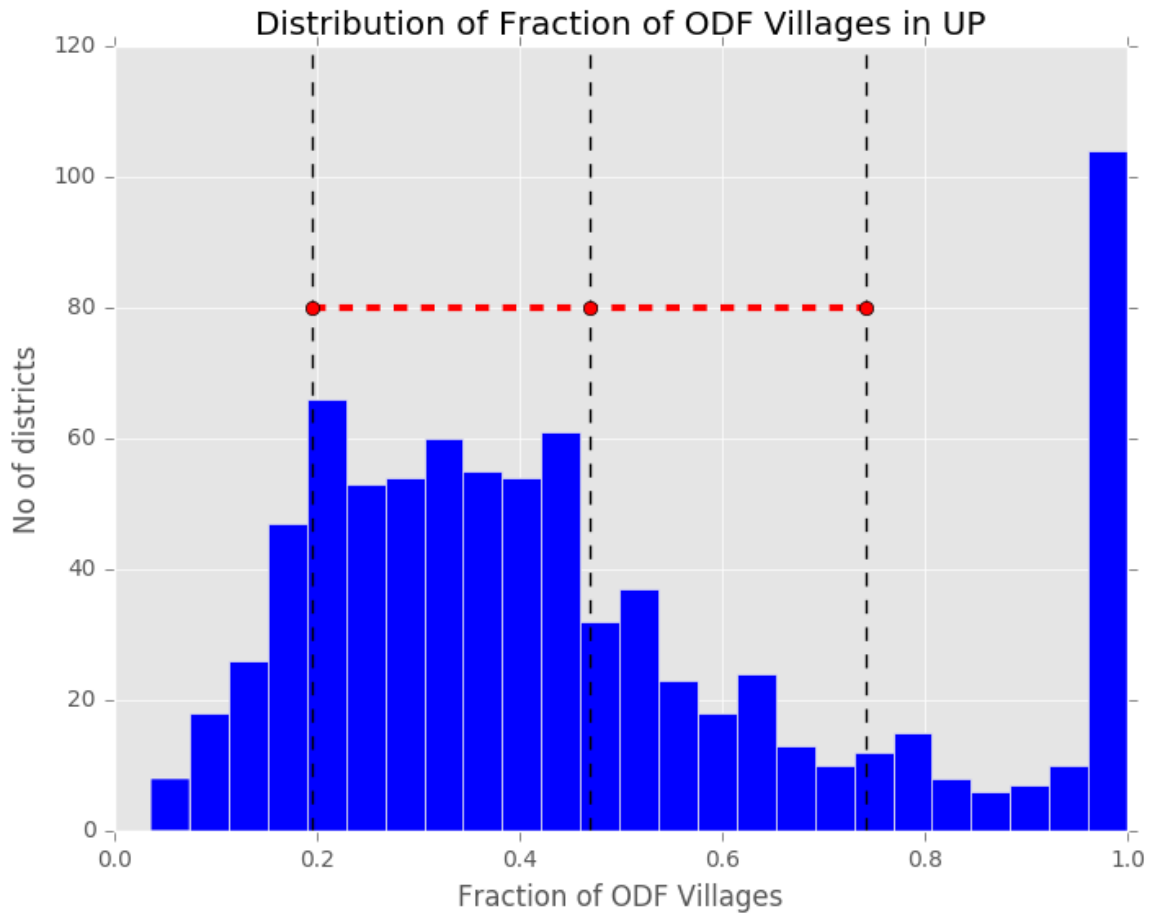
Code

```
#Importing libraries pandas for data handling and matplotlib for plotting histogram
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import style
style.use('ggplot')

#Function to calulate sample mean
def sample_mean(iterable):
    total=0
    for element in iterable:
        total+=element
    return(total/len(iterable))

#Function to calulate sample variance
def sample_variance(iterable):
    s_mean = sample_mean(iterable)
    total = 0
    for element in iterable:
        total += (element-s_mean)**2
    return(total/(len(iterable)-1))

#Function to calulate sample standard deviation
def sample_std_dev(iterable):
    return((sample_variance(iterable))**(0.5))

#Reading full data from excel sheet to a pandas data frame
data = pd.read_excel('swachhbharat.xlsx')

#Creating another data frame and storing relevant data into it
cols=['TotalVillage','TotalODFVillage']
UP_data=data.loc[data['StateName'] == 'Uttar Pradesh',cols]

#Adding a column to the data frame containing the fraction of ODF Villages
num=UP_data['TotalODFVillage']
den=UP_data['TotalVillage']
UP_data['Frac of ODF Villages'] = num / den

#Creating a histogram plot of the fraction of ODF Villages
UP_data['Frac of ODF Villages'].plot.hist(bins = 25,color = 'b')

#Labelling the axes and title of the plot
plt.xlabel('Fraction of ODF Villages')
plt.ylabel('No of districts')
plt.title('Distribution of Fraction of ODF Villages in UP')

#Adding a vertical line to the plot denoting the sample mean
samp_mean=sample_mean(UP_data['Frac of ODF Villages'])
samp_std_dev=sample_std_dev(UP_data['Frac of ODF Villages'])
plt.axvline(samp_mean, c = 'k', ls = 'dashed', lw = 1)
plt.axvline(samp_mean+samp_std_dev, c = 'k', ls = 'dashed', lw = 1)
plt.axvline(samp_mean-samp_std_dev, c = 'k', ls = 'dashed', lw = 1)
```

```
#Adding a horizontal line segment to the plot denoting the sample standard deviation
x_list=[samp_mean-samp_std_dev,samp_mean,samp_mean+samp_std_dev]
plt.plot(x_list,[80,80,80],c='r',lw=3, ls='--', marker='o')

#Showing the plot
plt.show()
```

Histogram

Histogram of number of Fraction of ODF Villages for each district of Uttar Pradesh



Here, Middle Vertical line represent Mean of the Data and distance between two extreme vertical line represent standard deviation.

Conclusion

Apart from when the fraction is equal to one, the distribution is approximately normal,but there is a sharp peak at one. This implies that the government's policy is to select a district and work to make its fraction one.

Team Contribution

a) **Vashishtha Kochar** - Team Leader 25%

b) **Nihal Barde** - Web Developer 25%

c) **Adeem Jassani** - Report Writer 25%

d) **Ram** - Programmer 25%