Descriptive Statistics in Pandas Series

Farivar Zarvandeh — Zarvandeh@gmail.com

July 27, 2023

1 Introduction

Descriptive statistics provide summary information about the data in a Series, giving us insights into its central tendency, variability, and distribution. Pandas offers a variety of methods to compute descriptive statistics on a Series. We will cover mean, median, standard deviation, and quantiles.

2 Mean

The mean represents the average of all values in the Series.

Method: `Series.mean()`

3 Median

The median is the middle value in the sorted Series. If the Series has an odd number of elements, it is the middle value. If the Series has an even number of elements, it is the average of the two middle values.

Method: `Series.median()`

4 Standard Deviation

The standard deviation measures the dispersion of values around the mean. It indicates how much the data deviates from the average.

Method: `Series.std()`

5 Quantiles

Quantiles are points that divide the data distribution into equal parts. Common quantiles include quartiles (dividing the data into four parts) and percentiles (dividing the data into hundred parts).

Method: `Series.quantile(q)`, where `q` is the desired quantile (0 to 1).

6 Handling Missing Values During Statistical Computations

Handling missing values is crucial when computing descriptive statistics. Pandas provides methods to handle missing data during statistical computations:

1. **Series.dropna():** Removes missing values from the Series before performing computations.

2. **Series.fillna(value):** Fills missing values with a specified value before computations.

7 Examples

```python
# Import Pandas library
import pandas as pd

# Create a sample Series
data = pd.Series([10, 20, 30, 40, 50, None, 60, 70])

# Calculate mean
mean_value = data.mean()
print("Mean:", mean_value)

# Calculate median
median_value = data.median()
print("Median:", median_value)

# Calculate standard deviation
std_value = data.std()
print("Standard Deviation:", std_value)

# Calculate 25th percentile (first quartile)
q25 = data.quantile(0.25)
print("25th Percentile (First Quartile):", q25)

# Handling missing values
data_dropna = data.dropna()
print("Series after removing missing values:")
print(data_dropna)

data_fillna = data.fillna(0)
print("Series after filling missing values with 0:")
print(data_fillna)
```

```python
import pandas as pd

# Sample Series with missing values
data = pd.Series([10, 20, 30, None, 50, 60, 70])

# Descriptive Statistics
mean_value = data.mean()
median_value = data.median()
std_deviation = data.std()
quantile_25 = data.quantile(0.25)  # First quartile (25th percentile)
quantile_75 = data.quantile(0.75)  # Third quartile (75th percentile)

print("Descriptive Statistics:")
print("Mean:", mean_value)
print("Median:", median_value)
print("Standard Deviation:", std_deviation)
print("25th Percentile (First Quartile):", quantile_25)
print("75th Percentile (Third Quartile):", quantile_75)

# Handling Missing Values
data_filled = data.fillna(0)  # Filling missing values with 0 before computations
mean_filled = data_filled.mean()
median_filled = data_filled.median()
std_deviation_filled = data_filled.std()

print("\nDescriptive Statistics with Missing Values Filled:")
print("Mean (Filled):", mean_filled)
print("Median (Filled):", median_filled)
print("Standard Deviation (Filled):", std_deviation_filled)
```