Day 6 of #100daysofmathandstats: Exploration of Data Distribution

By Harsh Kathiriya

Outline

- Importance
- Boxplot
- Frequency table
- Histogram
- Density plot

Importance of exploring the data distribution

- The goal of data exploration is to learn about characteristics and potential problems of a data set without the need to formulate assumptions about the data beforehand.
- In statistics, data exploration is often referred to as “exploratory data analysis” and contrasts traditional hypothesis testing.

Boxplot (Box and whiskers plot)

- Box plots are among the most used types of graphs in the business, statistics and data analysis.
- It is especially useful when you want to see if a distribution is skewed and whether there are potential unusual data values (outliers) in a given dataset.
- These plots are also widely used for comparing two data sets.
- A box plot is a great way of summarizing data set measured on an interval scale

Some real examples of box and whisker plot

- Comparison of heart rate for sitting, walking, and running.
- Distribution of bills in week
- Month-by-month expenses over a year

Frequency Table

- A tally of the count of numeric data values that fall into a set of intervals (bins).
- They help you understand which data values are common and which are rare.
- These tables organize your data and are an effective way to present the results to others.
- Frequency tables are also known as frequency distributions because they allow you to understand the distribution of values in your dataset.

Some real examples of box and whisker plot

- Count number of customers visited on daily basis
- Number of wickets a bowler has taken against a particular team to analyze the performance.

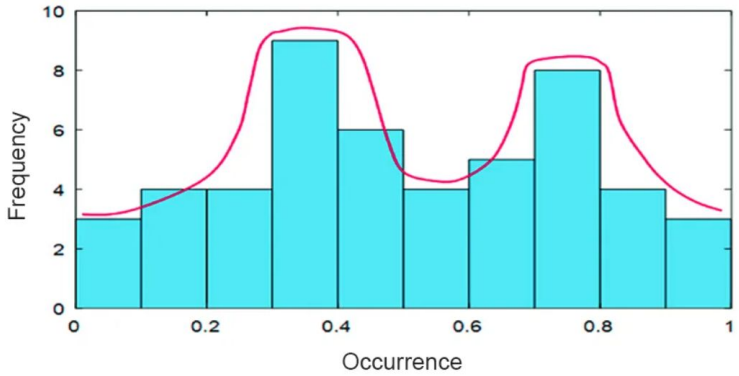
Day	Number of customers	Frequency
Monday	𝍸 𝍸 𝍸 |||	18
Tuesday	𝍸 𝍸 |||	13
Wednesday	𝍸 𝍸 𝍸 𝍸	20
Thursday	𝍸 𝍸 ||||	14
Friday	𝍸 𝍸 𝍸 𝍸 |	21
Saturday	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
Sunday	𝍸 𝍸 𝍸 𝍸 𝍸 |	26

Histogram

- A histogram is a way to visualize a frequency table, with bins on the x-axis and the data count on the y-axis.
- In statistical theory, location and variability are referred to as the first and second moments of a distribution.
- The third and fourth moments are called skewness and kurtosis. Skewness refers to whether the data is skewed to larger or smaller values, and kurtosis indicates the propensity of the data to have extreme values.
- Generally, these are discovered through visual displays of histograms.

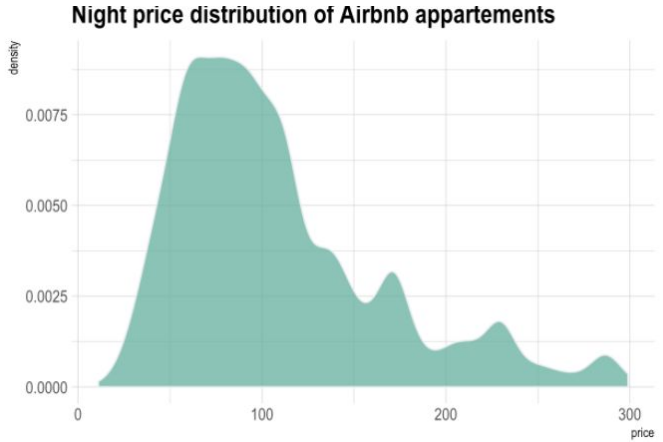
Some real examples of histograms

- In any general office, employees tend to drink less tea or coffee, but as a later hour approaches, their tiredness increases and they tend to drink more tea and coffee. Such data can be represented by the Skewed left histogram as shown in the graph below.



Density Plot

- A **smoothed version of the histogram**, often based on a kernel density estimate.
- Used for the same concepts as of histogram, the main difference is the visual representation as seen here.



Thank you

Github Link: https://github.com/harsh9898/100daysofstatandmath

Don’t forget to post your queries or feedbacks on the post.

Share or like for the benefit of others.