Descriptive Analytics (Descriptive Statistics)

Ms Swapnil Shrivastava

C-DAC Bangalore

Content

- Pareto Analysis
- Populations and Samples
- Measures of Location
  - Mean, median, mode, mid range
- Measures of Dispersion
  - Variance, standard deviation, inter quartile range
- Measures of Shapes
  - Skewness, kurtosis
- Measures of Association
  - Covariance, correlation
- Summary
- References

Pareto Analysis

- An Italian economist, Vilfredo Pareto, observed in 1906 that a large proportion of the wealth in Italy was owned by a small proportion of the people.
- Similarly, businesses often find situations where a large proportion of sales come from a small proportion of customers.
- A Pareto analysis involves sorting data and calculating cumulative proportions.
- The Excel file Bicycle Inventory lists the inventory of bicycle models In a sporting good store.

Applying Pareto Analysis to Bicycle Inventory

	A	B	C	D	E	F	G	H	I
1	**Bicycle Inventory**								
2									
3	**Product Category**	**Product Name**	**Purchase Cost**	**Selling Price**	**Supplier**	**Quantity on Hand**	**Inventory Value**	**Percentage**	**Cumulative %**
4	Road	Runroad 5000	$450.95	$599.99	Run-Up Bikes	5	$ 2,254.75	11.2%	11.2%
5	Road	Runroad 1000	$250.95	$350.99	Run-Up Bikes	8	$ 2,007.60	10.0%	21.1%
6	Road	Elegant 210	$281.52	$394.13	Bicyclist's Choice	7	$ 1,970.64	9.8%	30.9%
7	Road	Runroad 4000	$390.95	$495.99	Run-Up Bikes	5	$ 1,954.75	9.7%	40.6%
8	Mtn.	Eagle 3	$350.52	$490.73	Bike-One	5	$ 1,752.60	8.7%	49.3%
9	Road	Classic 109	$207.49	$290.49	Bicyclist's Choice	7	$ 1,452.43	7.2%	56.5%
10	Hybrid	Eagle 7	$150.89	$211.46	Bike-One	9	$ 1,358.01	6.7%	63.3%
11	Hybrid	Tea for Two	$429.02	$609.00	Simpson's Bike Supply	3	$ 1,287.06	6.4%	69.7%
12	Mtn.	Bluff Breaker	$375.00	$495.00	The Bike Path	3	$ 1,125.00	5.6%	75.2%
13	Mtn.	Eagle 2	$401.11	$561.54	Bike-One	2	$ 802.22	4.0%	79.2%
14	Leisure	Breeze LE	$109.95	$149.95	The Bike Path	5	$ 549.75	2.7%	81.9%
15	Children	Runkidder 100	$50.95	$75.99	Run-Up Bikes	10	$ 509.50	2.5%	84.5%
16	Mtn.	Jetty Breaker	$455.95	$649.95	The Bike Path	1	$ 455.95	2.3%	86.7%
17	Leisure	Runcool 3000	$85.95	$135.99	Run-Up Bikes	5	$ 429.75	2.1%	88.9%
18	Children	Coolest 100	$69.99	$97.98	Bicyclist's Choice	6	$ 419.94	2.1%	91.0%
19	Mtn.	Eagle 1	$410.01	$574.01	Bike-One	1	$ 410.01	2.0%	93.0%
20	Children	Green Rider	$95.47	$133.66	Simpson's Bike Supply	4	$ 381.88	1.9%	94.9%
21	Leisure	Breeze	$89.95	$130.95	The Bike Path	4	$ 359.80	1.8%	96.7%
22	Leisure	Blue Moon	$75.29	$105.41	Simpson's Bike Supply	4	$ 301.16	1.5%	98.2%
23	Leisure	Supreme 350	$50.00	$70.00	Bicyclist's Choice	3	$ 150.00	0.7%	98.9%
24	Children	Red Rider	$15.00	$25.50	Simpson's Bike Supply	8	$ 120.00	0.6%	99.5%
25	Leisure	Starlight	$100.47	$140.66	Simpson's Bike Supply	1	$ 100.47	0.5%	100.0%
26	Hybrid	Runblend 2000	$180.95	$255.99	Run-Up Bikes	0	$ -	0.0%	100.0%
27	Road	Twist & Shout	$490.50	$635.70	Simpson's Bike Supply	0	$ -	0.0%	100.0%
28						Total	$ 20,153.27		

75% of the bicycle inventory value comes from 40% (9/24) of items.

Populations and Samples

- Population consists of all items of interest for a particular decision or investigation.
  - All subscribers to Netflix
- Sample is a subset of a population.
  - All subscribers who rented a comedy from Netflix

Sampling is to obtain sufficient information to draw a valid inference about a population.

e.g. Market researchers use sampling to gauge consumer perceptions on new or existing goods and services.

Quality control analysts sample production output to verify quality levels and identify opportunities for improvement.

Measures of Location

- The measures that are used for describing the data using a single value.
- Measures of central tendency or location are frequently used to summarize and comprehend the data.
- Arithmetic Mean, Median, Mode and Midrange

Mean

- Mean is the arithmetical average value of the data and is one of the most frequently used measure of central tendency.
- It is the sum of observations divided by the number of observations.
- The most common is average
  - Measure students accomplishment in college: grade point average
  - Measure the performance of sports team: batting average
  - Measure performance in business: average delivery time
- In Python, using mean function of numpy module
- Mean is meaningful for both interval and ratio data. It can be affected by outliers.

Median

- Median is the value that divides the date into two equal parts, that is, proportion of observations below median and above median will be 50%.
- Ito find the median, arrange the data from least to greatest. For an
  - Odd number of observations, the median is the middle of the sorted numbers.
  - Even number of observations, the median is the mean of two middle numbers.
- In Python, using median function of numpy module
- It is meaningful for ratio, interval and ordinal data.

Mode

- Mode is the most frequently occurring value in the dataset.
- Most useful for data sets that contain a relatively small number of unique values.
- In Python, using mode function of scipy stats module.
- Mean and median are meaningless for qualitative data.

Midrange

- It is the average of the greatest and least value in the data set.
- Two ways:
  - Sort the data and take average of first and last element
  - Use amin and amax function of numpy to find minimum and maximum values respectively. Then take average of them using mean function of numpy package.
- Usually a rougher estimate than the mean

Quoting Computer Repair Time

Dataset includes 250 repair times for customers.

What repair time would be reasonable to quote to a new customer?

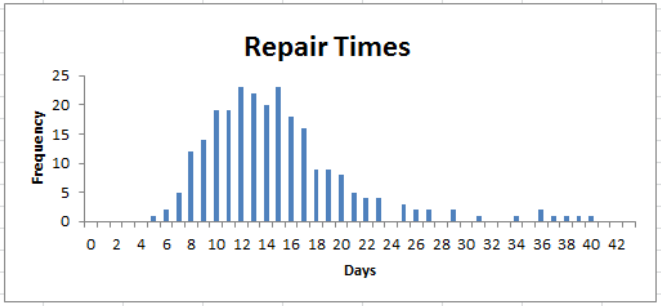
Mean, median and mode are all very close and show that the typical repair time is about 2 weeks.

What if repair time of 2 weeks is quoted?

	A	B
1	Computer Repair Times	
2		
3	Sample	Repair Time (Days)
4	1	18
5	2	15
6	3	17
7	4	9
250	247	31
251	248	6
252	249	17
253	250	13
254		
255	Mean	14.9
256	Median	14
257	Mode	15

Quoting Computer Repair Time

	A	B	C	D
1	Computer Repair Times			
2				
3			Relative	Cumulative
4	Days	Frequency	Frequency	Percentage
5	0	0	0.000	0.0%
6	1	0	0.000	0.0%
7	2	0	0.000	0.0%
8	3	0	0.000	0.0%
9	4	0	0.000	0.0%
10	5	1	0.004	0.4%
11	6	2	0.008	1.2%
12	7	5	0.020	3.2%
13	8	12	0.048	8.0%
14	9	14	0.056	13.6%
15	10	19	0.076	21.2%
16	11	19	0.076	28.8%
17	12	23	0.092	38.0%
18	13	22	0.088	46.8%
19	14	20	0.080	54.8%
20	15	23	0.092	64.0%
21	16	18	0.072	71.2%
22	17	16	0.064	77.6%
23	18	9	0.036	81.2%
24	19	9	0.036	84.8%
25	20	8	0.032	88.0%
26	21	5	0.020	90.0%
27	22	4	0.016	91.6%



Longest repair time took almost 6 weeks. Can the company give guaranteed repair time of 6 weeks?
90% of time repairs are completed within 3 weeks. Rare occasion it take longer.

Decision: The customers could expect their computers back within 2 to 3 weeks and inform them that it might take longer if a special part was needed.

Measures of Dispersion

- Dispersion refers to the degree of variation in the data, that is, the numerical spread (or compactness) of the data.
- Several statistical measures characterize dispersion:
  - Range
  - Variance
  - Interquartile Range
  - Standard deviation

Range and Interquartile Distance

- Range: difference between the maximum value and the minimum value in the dataset.
  - amax(data range) – amin(data range)
  - Range for cost per order data in Purchase Orders database.
- Interquartile Range (IQR): difference between the first and third quartiles, Q3-Q1
  - Middle 50% of data
  - for cost per order data in Purchase Orders database.
  - Use iqr function if scipy stats module.

Variance

- Its computation depends on all the data. The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations.
  - In Python var function of numpy package could be used.
- The dimension of the variance is the square of the dimension of the observations. Hence difficult to use in practical applications.

Standard Deviation

- It is the square root of the variance. Its unit of measure are the same as the units of data.
  - In Python std function of numpy package could be used.
- It is a popular measure of risk particularly in financial analysis.
- A larger standard deviation implies that while a greater potential of higher return exists, there is also greater risk of realizing a lower return.
  - Find mean and standard deviation for Intel and GE stocks in Closing Stock Prices Excel file.

Chebyshev's Theorem

- It states that for any set of data, the proportion of values that lie within k standard deviations (k>1) of the mean is at least $1 – 1/k^2$.
  - For k=2, at least ¾ or 75% of data lie within two standard deviations of the mean
  - For k=3, at least 8/9 or 89% of data lie within three standard deviations of the mean
- It provide a basic understanding of the variation in a set of data using only the computed mean and standard deviation.
- For cost per order data in the Purchase Orders database find two standard deviation across mean.

Empirical Rules

- Approximately 68% of the observations will fall within one standard deviation of the mean or between x-s and x+s
- Approximately 95% of the observations will fall within two standard deviation of the mean or between x-2s and x+2s
- Approximately 99.7% of the observations will fall within three standard deviation of the mean or between x-3s and x+3s
- Two or three standard deviations around the mean are commonly used to describe the variability of most practical sets of data. E.g. Good transportation Company

Process Capability Index

To measure how well a manufacturing process can achieve the specifications, we usually take a sample of output, measure the dimension, compute the total variation using the third empirical rule, and then compare the result to the specifications by dividing the specification range by the total variation.

$$Cp = \frac{\text{upper specification} - \text{lower specification}}{\text{total variation}}$$

Manufacturers use this index to evaluate the quality of their products and determine when they need to make improvements in their process.

Manufacturing Measurement Excel file

Z-score

- It provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement.
- The distance from mean is expressed in units of standard deviation
  - Z-score = -1.5 means that observation is 1.5 standard deviation to the left of the mean.
- Use zscore function of scipy stats module.
- Calculate z-scores for cost per order data

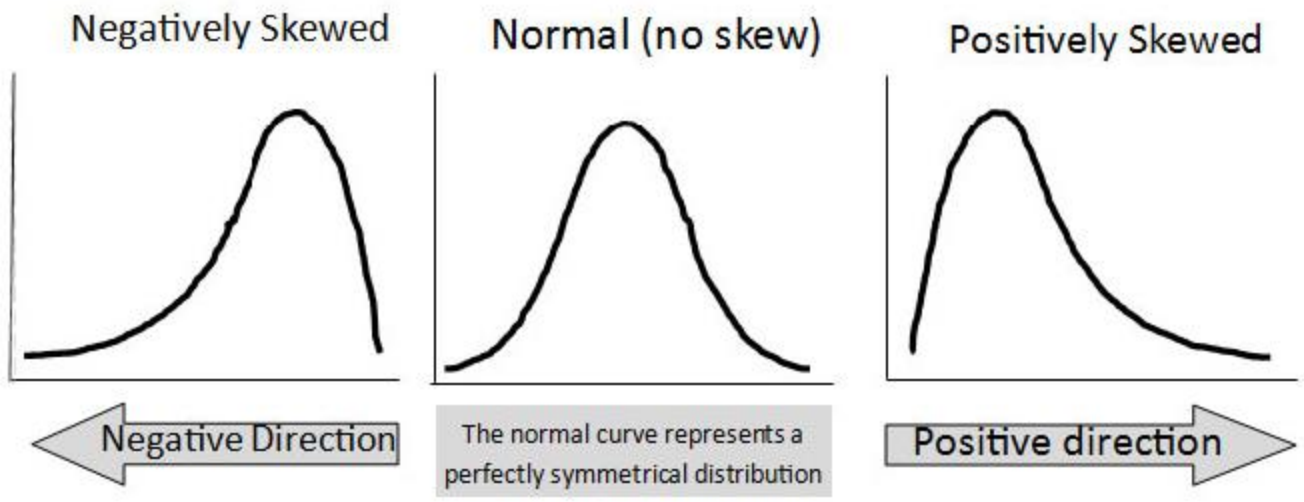
Coefficient of Variation

- It provides a relative measure of the dispersion in data relative to the mean and is defined as

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

- Multiplied by 100 to express it as percent.
- For comparing the variability of two or more datasets.
- It provides a relative measure of risk to return. The smaller the coefficient of variation, the smaller the relative risk is for the return provided.
- Apply to Closing Stock Prices worksheet.

Measures of Shape

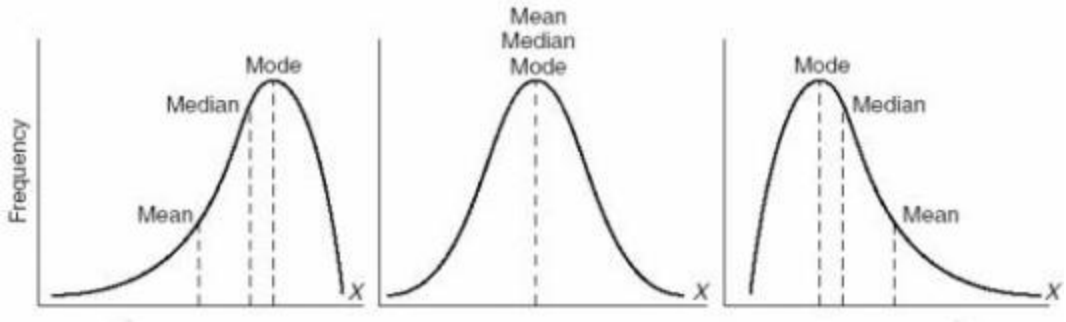


- Histograms of sample data can take on a variety of different shapes.
  - Symmetrical: no skew
  - Asymmetrical: skewed i.e. more of mass concentrated on one side
- Skewness describes lack of symmetry of data

Measuring Skewness

- The coefficient of skewness (CS) measures the degree of asymmetry of observations around the mean.
- Using skew function of scipy stats module
  - +ve CS means positively skewed
  - -ve CS means negatively skewed
  - CS near 0 means less degree of skewness
- Also
  - CS>1 or CS<-1 : higher degree of skewness
  - 1>CS>0.5 or -0.5>CS>-1 : moderate skewness
  - 0.5>CS>-0.5 : relative symmetry
- Find CS for cost per order in Purchase Order Excel file

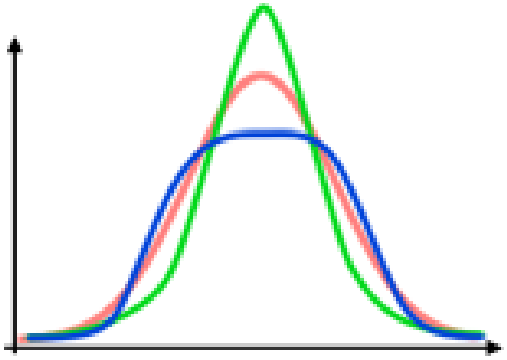
Comparing Measure of Location with Skewness



- Perfectly symmetrical: mean, median and mode would all be the same
- Negatively skewed: mean<median<mode
- Positively slewed: mode<median<mean

Kurtosis

- It refers to the peakedness(i.e. high, narrow) or flatness (i.e. short, flat-topped) of a histogram.
- The coefficient of kurtosis (CK) measures the degree of kurtosis of a population.
  - CK < 3 are more flat with a wide degree of dispersion
  - CK>3 are more peaked with less dispersion
- Computes using Kurtosis function of scipy stats module.



Skewness and Kurtosis

- Skewness and kurtosis can help provide more information to evaluate risk than just using the standard deviation.
- Both a negatively and positively skewed distribution may have the same standard deviation. If objective is to achieve high return, the negatively skewed distribution will have higher probabilities of larger returns.
- For higher kurtosis, histogram has more area in the tails rather than in the middle. This can indicate a greater potential for extreme and possibly catastrophic outcomes.

Describe Function

- It provides a summary of numerical statistical measures that describe location, dispersion and shape for sample data.
- scipy.stats.describe(array, axis=0) computes the descriptive statistics of the passed array elements along the specified axis of the array.
- Apply this function to cost per order and A/P terms data column in Purchase Order Excel file

Grouped Data

- The sample data are summarized in a frequency distribution.
- Extracting information from government databases such as the Census Bureau of Labor Statistics
- The mean and variance are calculated differently. Calculations of Mean and Variance Using a Frequency Distribution in Computer Repair Time Excel file.

Proportion

- Proportion are key descriptive statistics for categorical data.
  - Defects or errors in quality control applications
- The proportions are numbers between 0 and 1.
- In Purchase Orders database, find the number of orders placed with Spacetime Technologies

Measures of Association

- Two variables have a strong statistical relationship with one another if they appear to move together.
  - Ice cream sales likely have a strong relationship with daily temperature.
- Scatter plot is used to visualize relationships between two variables.
- Covariance and Correlations

Covariance

- It is a measure of the linear association between two variables, X and Y.
- It is the average of the product of the deviations of each pair of observations from their respective means.
- The sign of the covariance tells whether there is a direct relationship or an inverse relationship.
- The numpy cov() method returns the covariance matrix.
  - Using Colleges and Universities Excel file

Correlation

- It is a measure of the linear relationship between two variables, X and Y, which does not depend on the unit of measurements.
- It is measured by correlation coefficient.
- The pearsonr method of stats package computes the pairwise correlation of columns.
- Correlation of
  - 0 indicates that the two variables have no linear relationship to each other.
  - +ve
  - -ve

Outliers

- It is an unusual observation as compared with the rest.
- How to find them?
  - Check the data for possible errors
  - Histograms
  - Z-scores
  - Inter quartile range
- Using Home Market Value Excel file

Summary

- Descriptive statistics provides a summary of numerical statistical measures that describe
  - Location
  - Dispersion
  - Shape
  - Association

  for sample data

References

- James Evans, Business Analytics: Methods, Models and Decisions, Second Edition, Pearson Publication, 2017.
- Manaranjan Pradhan and U Dinesh Kumar, Machine Learning using Python, Wiley Publication, 2019.
- U Dinesh Kumar, Business Analytics- The Science of Data-Driven Decision Making, Wiley Publication, 2017.
- Python pandas: https://pandas.pydata.org/
- Matplotlib: Visualization with Python https://matplotlib.org/