Descriptive Analytics

Prof. Abhinava Tripathi

Introduction

- As a fund manager, several prospective clients are requesting to compare the performance of different funds
- They have several questions such as: Are all the values relatively similar?
- And does any variable have outlier values that are either extremely small or extremely large?
- While doing a complete search of the retirement funds data could lead to answers to the preceding questions, you wonder if there are better ways than extensive searching to uncover those answers

Introduction

- Descriptive analytics is a commonly used form of data analysis whereby historical data is collected, organized, and then presented in a way that is easily understood
- In Descriptive analysis, we describe our data with the help of various representative methods like charts, graphs, tables, excel files, etc
- The descriptive statistic can be categorized into three parts:
  - Measures of central tendency
  - Measures of variation
  - Measures of shape

Measures of central tendency

Measures of central tendency

- A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset
- In statistics, the three most common measures of central tendency are the mean, median, mode, and quartiles
  - Mean: It is the sum of observations divided by the total number of observations
  - Median: It is the middle value of the data set. It splits the data into two halves
  - Mode: It is the value that has the highest frequency in the given data set
  - Quartiles: Quartiles are measures of central tendency that divide a group of data into four subgroups or parts (Q1, Q2, Q3, Q4)

Measures of central tendency: Mean

- The arithmetic mean (in everyday usage, the mean) is the most common measure of central tendency
- To calculate a mean, sum the values in a set of data and then divide that sum by the number of values in the set
  - $\bar{X} = \frac{sum\ of\ n\ values}{n}$ or $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ or $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- Consider the following data on typical time-to-get-ready for the office in the morning

Day:	1	2	3	4	5	6	7	8	9	10
Time (minutes)	39	29	43	52	39	44	40	31	44	35

Measures of central tendency: Mean

- Consider the following data on typical times to get ready for the office in the morning
- $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{39+29+43+52+39+44+40+31+44+35}{10} = \frac{396}{10} = 39.6$
- On Day 3, a set of unusual circumstances delayed the person getting ready by an extra hour, so that the time for that day was 103 minutes
- $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{39+29+103+52+39+44+40+31+44+35}{10} = \frac{456}{10} = 45.6$

Measures of central tendency: Median

- It is the middle value of the data set as It splits the data into two halves
- Extreme values do not affect the median, making the median a good alternative to the mean
- $Median = \frac{n+1}{2} th\ ranked\ value$
- Calculate the median by following one of two rules
  - Rule 1: If the data set contains an odd number of values, the median is the measurement associated with the middle-ranked value
  - Rule 2: If the data set contains an even number of values, the median is the measurement associated with the average of the two middle-ranked values

Measures of central tendency: Median

- We will again use the example of 10 time-to-get-ready values, first we will rank them from low to high

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- The result of dividing n + 1 by 2 for this sample of 10 is (10 + 1)>2 = 5.5
- As per rule two: Median= (39+40)/2=39.5
- Substituting 103 minutes on Day 3 (As earlier) does not affect the value of median, which would remain 39.5
- This example illustrates that the median is not affected by extreme values

Measures of central tendency: Mode

- The mode is the value that appears most frequently
- Like the median and unlike the mean, extreme values do not affect the mode

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- There are two modes, 39 minutes and 44 minutes, because each of these values occurs twice

Measures of central tendency: Mode

- The mode is the value that appears most frequently
- Like the median and unlike the mean, extreme values do not affect the mode

Observed Data	**1**	**3**	**0**	**3**	**26**	**2**	**7**	**4**	**0**	**2**	**3**	**3**	**6**	**3**
Ranked values	**0**	**0**	**1**	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**4**	**6**	**7**	**26**

- Because 3 occurs five times, more times than any other value, the mode is 3

Measures of central tendency: Quartiles

- Quartiles are measures of central tendency that divide a group of data into four subgroups or parts
- The three quartiles (Q1, Q2, Q3, Q4) split a set of data into four equal parts.
- First quartile, Q1, Q1 = (n + 1)/4th ranked value
- Third quartile, Q3, Q3 =3(n + 1)/4th ranked value
- The second quartile (Q2), the median, divides the set such that 50% of the values are smaller than or equal to the median, and 50% are larger than or equal to the median

Measures of central tendency: Quartiles

- Rules for Calculating the Quartiles from a Set of Ranked Values
  - Rule 1: If the ranked value is a whole number, the quartile is equal to the measurement that corresponds to that ranked value
  - Rule 2: If the ranked value is a fractional half (2.5, 4.5, etc.), the quartile is equal to the measurement that corresponds to the average of the measurements corresponding to the two ranked values involved
  - Rule 3: If the ranked value is neither a whole number nor a fractional half, round the result to the nearest integer and select the measurement corresponding to that ranked value

Measures of central tendency: Quartiles

- Consider our example of time-to-get-ready values

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- Q1: (n + 1)/4 = (10 + 1)/4 = 2.75, thus Q1= 35
- Q3: 3(n + 1)>4 = 3(10 + 1)>4 = 8.25 , thus Q3= 44
- Q2 is same as median= 39.5 (corresponding to 5.5)
- Percentiles: Related to quartiles are percentiles that split a variable into 100 equal parts

Measures of central tendency: The Interquartile Range

- The interquartile range (also called the midspread) measures the difference in the center of a distribution between the third and first quartiles
- Interquartile range (IQR) = $Q_3 - Q_1$

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- IQR= 44-35= 9

Measures of variation

Measures of variability

- Measures of variability describe the spread or the dispersion of a data set
- Measures of variability are
  - **Range:** The Range describes the difference between the largest and smallest data point in our data set
  - **Variance:** The variance is the average of the squared deviations about the arithmetic mean for a set of numbers
  - **Standard Deviation (SD):** Standard deviation measures the dispersion of a dataset relative to its mean. It is defined as the square root of the variance
  - **Mean Absolute deviation:** The mean absolute deviation (MAD) is the average of the absolute values of the deviations around the mean for a set of numbers.

Measures of variability: Range

- A simple measure of variation, the range is the difference between the largest and smallest value and is the simplest descriptive measure of variation for a numerical variable
- $Range = X_{largest} - X_{smallest}$

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- As per the formula, the range is 52-29=23 minutes
- The range measures the total spread in the set of data
- However, the range does not take into account how the values are distributed between the smallest and largest values

Measures of variability: Variance or Standard Deviation

- Two commonly used measures of variation that account for how all the values are distributed are the variance and the standard deviation
- Two commonly used measures of variation that account for how all the values are distributed are the variance and the standard deviation
- The calculation of variance squares the difference between each value and the mean and then sums those squared differences
- For sample variance these sum of squares are divided by sample size-1
- For population variance these sum of squares are divided by population size (N)

Measures of variability: Variance or Standard Deviation

- For a sample containing n values $X_1, X_2 \ldots. X_n$, the sample variance (S^2) is defined as
- Sample variance $S^2 = \frac{[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]}{n-1}$
- For a Population containing N values $X_1, X_2 \ldots. X_n$, the Population variance (σ^2) is defined as
- Population variance $\sigma^2 = \frac{[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2]}{N}$
- Observe that the difference between dividing by n and by n - 1 becomes smaller as the sample size increases and converges to large population size N

Measures of variability: Variance or Standard Deviation

- This can be put in a more compact manner as shown here.
- $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ or in standard deviation form
- $S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$
- For population SD: $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}}]$
- Observe that the difference between dividing by n and by n - 1 becomes smaller as the sample size increases and converges to large population size N

Measures of variability: Variance or Standard Deviation

- Consider the example of 10 observations from time-to-get-ready

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean=40		Sum =412.40
		Sum Divide by (n-1)=45.82

- $S^2 = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} = \frac{[(39-39.6)^2 + (29-39.6)^2 + \cdots + (35-39.6)^2]}{10-1} = \frac{412.4}{9} = 45.82$
- $S = 6.77$

Measures of variability: Variance or Standard Deviation

- Consider the example of 10 observations from time-to-get-ready

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean=40		Sum =412.40
		Sum Divide by (n-1)=45.82

- $\sigma^2 = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\text{n}}} = \frac{[(39-39.6)^2 + (29-39.6)^2 + \cdots + (35-39.6)^2]}{10} = \frac{412.4}{10} = 41.24$
- $\sigma = 6.42$

Measures of variability: MAD

- The steps to calculate the mean absolute deviation are shown provided here
  - Step 1: Calculate the mean
  - Step 2: Calculate how far away each data point is from the mean using positive distances. These are called absolute deviations
  - Step 3: Add those deviations together
  - Step 4: Divide the sum by the number of data points
- $MAD = \frac{[\sum_{i=1}^{n} |(x_i - \bar{x})|]}{n}$

Measures of variability: MAD

- Consider the example of 10 time-to-get-ready values and MAD computation for the data

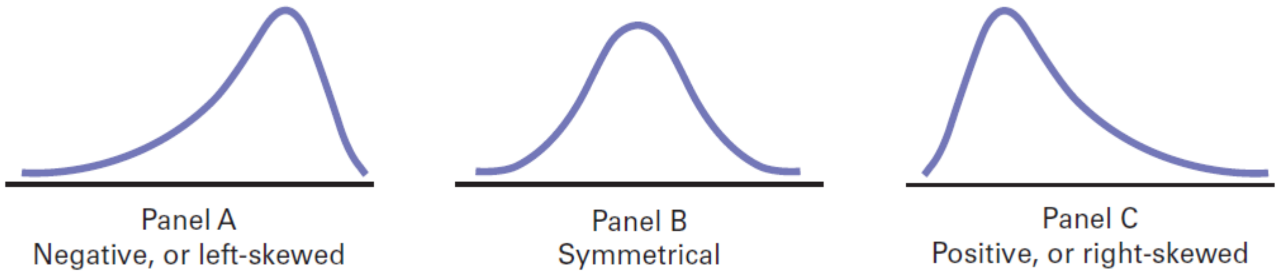
Time (X)	S2: absolute(Xi - $\bar{X}$)
39	0.60
29	10.60
43	3.40
52	12.40
39	0.60
44	4.40
40	0.40
31	8.60
44	4.40
35	4.60
S1: Mean=40	S3: Sum=50.00
	S4: Sum/10=5

Measures of shape

- A measure of shape is the tool that can be used to describe the shape of a distribution of data
    - **Skewness:** Skewness refers to a distortion or asymmetry that deviates from the symmetrical nature of data around its mean
    - **Kurtosis:** Kurtosis measures the peakedness of the curve of the distribution

Measures of shape: Skewness

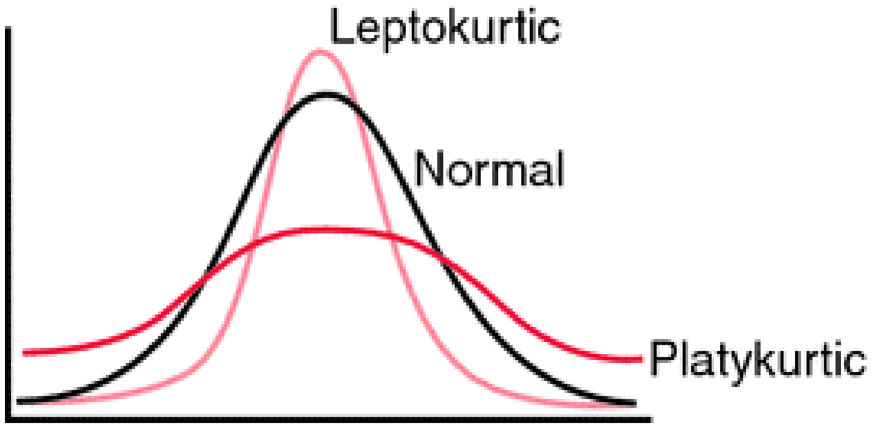
- The distribution of data in which the right half is a mirror image of the left half is said to be symmetrical



- Panel A: Mean < median: negative, or left-skewed distribution
- Panel B: Mean = median: symmetrical distribution (zero skewness)
- Panel C: Mean > median: positive, or right-skewed distribution

Measures of shape: Kurtosis

- Kurtosis measures the peakedness of the curve of the distribution
- That is, how sharply the curve rises approaching the center of the distribution



- **Leptokurtic:** A distribution that has a sharper-rising center peak than the peak of a normal distribution has positive kurtosis
- **Platykurtic:** A distribution that has a slower-rising (flatter) center peak than the peak of a normal distribution has negative kurtosis

Thanks!