DATA-DRIVEN DECISION MAKING
LECTURE-2: DESCRIPTIVE ANALYTICS

AVIJIT MALLIK
ASSISTANT PROFESSOR,
IBA, UNIVERSITY OF DHAKA

DESCRIPTIVE ANALYTICS

looks at past performance and use historical data to gain insight by employing explanatory statistical methods and reporting tools (graphs, dashboards)

DESCRIPTIVE ANALYTICS

- Two complimentary components:
  - Analysis
    - Descriptive Statistics
  - Reporting
    - Graphs/Dashboards/Tables/Infographics

DESCRIPTIVE STATISTICS

- Involve arranging, summarising and presenting a set of data in such way that useful information is produced
- Help managers and decision makers gain a better understanding of the business and economic environment and thus enables them to make better informed decisions (For example: Pareto model in identifying profitable customers)

EXAMPLES

- Detailed Sales Reports
- Financial Statement Analysis
- Demand Trends
- Aggregated Survey Results
- Progress to Goals

POPULATION VS. SAMPLE

- Population
  - Set of all relevant values for a decision
- Sample
  - A subset of the population of interest
- Parameter
  - Characteristic of the *population* (such as the mean, proportion, standard deviation)
- Statistic
  - Characteristic of the *sample* that is used to estimate a parameter

MEASURES OF LOCATION

- Mean
  - Sum of observations divided by the number of observations
- Median
  - Middle of the data
  - 50% of the observations fall on each side of the median
- Mode
  - Most frequently occurring observation
- Midrange
  - Average of the smallest and largest values within a dataset

MEAN (ARITHMETIC MEAN)

A simple average computed by summing the observations and dividing by the number of observations

- Population Mean $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, N is the number of data in the population
- Sample Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, n is the number of data in the sample

MEAN (ARITHMETIC MEAN)

- Example: the following data correspond to the monthly phone bill of eight customers, what is the sample monthly average phone bill?
  - Phone bill x: 42.19, 48.54, 104.88, 44.32, 53.90, 31.77, 45.77, 33.40
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{42.19+48.54+104.88+44.32+53.90+31.77+45.77+33.40}{8}$
  - $\bar{x} = \frac{404.77}{8} = 50.60$

ARITHMETIC MEAN: PROS & CONS

- PROS:
  - It is the most commonly used measure of location
  - Easily interpreted
- CONS:
  - Not appropriate for nominal/ordinal data
  - It is often NOT the most helpful, can be distorted by extreme values

MEDIAN

- The median is the value in the middle when the data are placed in order (most commonly ascending).
- The position of the median is given by (n+1)/2, where n is the number of observations
- For an odd number of observations the median will correspond to a real datum in the dataset
- For an even number of observations it will be a value that is not present in the dataset (one of its disadvantages)

MEDIAN

- Example 1: 2020, 2075, 2125, 2040, 1980, 2010, 2050, 2165, 2070, 3930, 2040

  Sorted: 1980 2010 2020 2040 2040 2050 2070 2075 2125 2165 3930

  Position = (n+1)/2 = (11+1)/2 = 6, hence **median = 2050**

- Example 2: 2020, 2075, 2125, 2040, 1980, 2010, 2050, 2165, 2070, 3930, 2060, 2040

  Sorted: 1980 2010 2020 2040 2040 2050 2060 2070 2075 2125 2165 3930

  Position: (n+1)/2 = (12+1)/2 = 6.5, hence **median = (2050+2060)/2 = 2055**

MODE

- Defined as the observation(s) that occurs with the greatest frequency

- Example: previous example (average starting salaries)

- Sorted values: 1980 2010 2020 2040 2040 2050 2060 2070 2075 2125 2165 3930

- Mode is £2040

MODE

- PROS
  - It is the only average measure to describe categorical data
- CONS
  - It is not always representative (in the example it is significantly lower than mean)
  - It may not be unique (i.e. multimodal)
  - It is unstable due to its sensitiveness to the number of observations

MEASURES OF LOCATION

- All measures of central of location should be considered when the objective is to describe a single set of data
- However, not all of them are appropriate for all different type of data
    - Nominal data: Mode
    - Ordinal data: Mode, Median
    - Interval/Ratio data: Mean, Median, Mode

EXAMPLE

Sample 1	Sample 2
5	0
5	7
6	1
4	1
5	0
6	5
4	5
4	5
6	11
5	15

If we calculate the most common measures of central tendency, what do we observe?

MEAN		
MEDIAN		
MODE		

MEASURES OF DISPERSION

- Comparing similar datasets using only measures of location is inappropriate
- We also need to have information on the range of our data and the level of dispersion around the mean
- Consider two nations with the same median household income. This does not mean that both nations are having similar household incomes if for example, one nation has extremes or wealth and poverty and the other has little variation amongst households
- The simplest useful numerical description of a dataset consists of both a measure of location and a measure of dispersion (spread)

MEASURES OF DISPERSION

- **Range:** The difference between the largest and smallest observations
- **Variance:** Measures spread within the data set; squared value of the standard deviation
- **Standard Deviation:** Measures the spread within the data set
- **Coefficient of Variance:** Ratio of the standard deviation of the sample divided by the mean of the sample

RANGE

- The difference between the largest and smallest observations
- Example: Recall the dataset of the starting monthly salary of business graduates.
- Range = £3930 – £1980 = £1950
- Pros: It gives an idea of the broad spread of data and is useful in spotting extreme values and errors
- Cons: It is not really representative since it take into consideration the most extreme values in the dataset which can often be quite extreme

VARIANCE

- The variance (s^2) is a measure of variability of the data that make use of all the observations. It is essentially the average of the squares of the deviations of the observations from the mean

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, where\ n\ the\ sample\ size$

- The above formula gives the sample variance, which is also an unbiased estimate of the population variance

VARIANCE

- Example: Recall the example with the average monthly starting salaries of 12 business graduates - 2020, 2075, 2125, 2040, 1980, 2010, 2050, 2165, 2070, 3930, 2060, 2040
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 294564.20$
- PROS:
  - It uses all observations
  - It is an unbiased estimate of the population variance
- CONS:
  - Difficult to interpret since the measurement units are in squares
  - Usually a large number compared to the mean

STANDARD DEVIATION

- It is the squared root of the variance
- $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{294564.20} = £542.74$
- PROS
  - It has the same measurement unit as our data
  - Therefore, it can be directly compared with the sample mean
- CONS
  - Like the mean it is not resistant. Strong skewness or a few outliers can greatly increase the standard deviation

COEFFICIENT OF VARIANCE

- Variance as well as standard deviation are difficult to consider when comparing different variables
  - they are affected by the range of the values considered and their measurement units
- Example:
  - Does a variable with standard deviation of 10 have greater spread of values than variable with a standard deviation of 100?
  - What can we say about the variability of two variables that are measuring the same thing but have different measurement units (e.g. the average salary of skilled UK worker to that of a skilled EU worker)?
- A measure that gives an indication of how large is the standard deviation in relation to the mean is the coefficient of variance (CV).
- $CV = \frac{standard\ deviation}{mean}$ (i.e. frequently given in %: CV*100%)

COEFFICIENT OF VARIANCE

- The average salary of an EU skilled worker is €27,560 with a standard deviation of €436, while for the same worker in the UK it is £23,860 with a standard deviation of £656. Are the salaries equally spread?

- EU worker, $CV = \frac{standard\ deviation}{mean} = \frac{436}{27560} = 0.0158 = 1.58\%$

- UK worker, $CV = \frac{standard\ deviation}{mean} = \frac{656}{23860} = 0.0275 = 2.75\%$

EXAMPLE

Sample 1	Sample 2
5	0
5	7
6	1
4	1
5	0
6	5
4	5
4	5
6	11
5	15

Central tendency statistics told us that the samples were the same

MEAN	5.0	5.0
MEDIAN	5	5
MODE	5	5

Dispersion shows us that the samples are different

RANGE	2	15
VARIANCE	0.67	25
ST DEV	0.82	5
CV	0.16	1

EMPIRICAL RULES

- Empirical rules:
  - Approximately 68% of the observations will fall within one standard deviation of the mean, or within $\bar{x}$ ± s
  - Approximately 95% of the observations will fall within two standard deviations of the mean, or within $\bar{x}$ ± 2s
  - Approximately 99.7% of the observations will fall within three standard deviations of the mean, or within $\bar{x}$ ± 3s

PRACTICE TIME

- Please open '*Day_2-Practice2'* excel file.
- From the given data set, please provide a descriptive summary of 'values of the items' using the 'Descriptive statistics' tool from 'Data Analysis' tab.
- Please find the mean, median, mode, maximum and minimum values.
- Go to Insert>Pivot Table. By dropping 'Department' in row, 'Value' in value, create a basic pivot table.
- Create another column showing the average values, values as percentage of Grand Total.
- Finally drop 'Branch' in Column. Can you construct a pie chart using 'Branch' as the slicers?

Thank You!