Module 2: bs of big data & stats 101

Thursday, June 15, 2023 22:41

- Why is data so important?
  - Brief history of statistics:
    - A systematic collection of data on the population and the economy was begu
      Italian city-states of Venice and Florence during the Renaissance
    - Term statistics is derived from the word state = used to refer to a collection o
      interest to the state
    - 1662 English tradesman John Graunt published a book "natural and political
      observations made upon the Bills of morality'
      - London bills of morality used to survey households in parishes and disc
        that on average there were approx. 3 deaths for every 88 people.
      - 13200 deaths /year - estimate London population = 13200 x88/3 = 387
        387,200
- How to mislead through poor sampling
  - Sample = data collected
  - Sample is collected from a population
  - Data analysis = gathering, modeling and transforming data, highlight useful informa
    conclusions, supporting decision making
- How to mislead through interpreting
  - Want to lie, graphical charts
  - Invented x-axis
- Python and stats 101
  - Defining data analytics:
- data: facts and figures collected, summarized, analyzed, : -r
- quantitative: age (18)
- qualitative: age (young)
- Continuous - data is infinitely divisible into whatever units
  - Age = 0- 100
- Ordinal or rank:
  - In order but not necessarily equal (abcd)
- Categorical or discrete:
  - Data consists of indivisible categories.
- cross-sectional data

un in the

of facts of

covered

200

ation,

- cross sectional data
- time-series data from previous year
- Types of studies and sampling errors
    - Descriptive analytics:
        - Methods of organizing and summarizing and presenting data in an informativ
            - frequency table
            - Histogram
            - Mean
            - Variance
    - Inferential analytics
        - The methods used to determine something about a population on the basis o sample (ml/ai for big data)
        - Population: the entire set of individuals or objects of interest or the measure obtained from all individuals or objects of interest
        - sample: a portion, a part, of the population of interest
- Types of studies
    - Experimental study
        - One variable is manipulated
        - Second variable is observed and measured to determine effect of manipulate variable
        - Measurements are compared to see if there are differences between conditi
    - Correlation study
        - Determining if there is a relationship between two variables and to describe relationship
        - Observes two variables as they exist naturally
    - quasi-experimental
        - Compares groups based on avariable that differentiate the groups (male/fem
    - Sampling error
        - Discrepancy between a sample statistic and its population parameter
- mean, median, mode
    - Center measurement is a summary measure of the overall level of a dataset
    - Geometric mean
    - Mean - arithmetic average
- Median
    - Middle number = odd
    - Even add too middle numbers and divide by 2
    - Middle value in an ordered sequence of numbers
    - Sort data first
- Mean or median?

ve way

of a

ements

ed

ons

the

nale)

- Mean is best for symmetric distributions
- Median is less sensitive to outliers than the mean and thus better measure than th for highly skewed distributions (family income, housing prices)
- 88.8 guns per 100 people
    - Civilian five arms 270,000, ooo
    - Total us population 304,000,000
    - 270, 000,000 / 304,000,000 X 100 = 0 88.8

- Mode
    - Most frequently occurring number (score, measurement, value, cost)
    - Frequency distribution, it's the highest point
    - Value observed most frequently
    - If no observation is repeated the mode is undefinedfor that sequence
    - Average number of tickets purchased per person for a GT football game, for examp almost always going to be accurately reflected by the mode
- Frequency Distribution
    - Number of times a data item occurs
    - Cumulative frequency distribution - running total of frequencies
        - Tells you the total number of data items at different stages in the data set
- Variability (dispersion) - measures amount -of scatter in a dataset
    - Gives us an indication of how well the average characterizes the data Asa whole
    - Average characterizes a set of observations
    - A: 30,50,70
    - b: 40,50,60
    - Mean of both two data setsis 50
    - But the distance of the observations from the mean in data set a is larger than in T B
    - Data set bis a better representation of Tre dataset than is the casefor set A
    - Commonly used methods for calculating variability: range, variance, standard devia interquartile range, coefficient of variation
    - range= difference between largest and smallest observations
        - 10, 5, 2,100
        - 100 - 2 = 98
    - Variance - average of the squares of the deviations of the observations from their m
        - Variance of 5,7,3? Mean (5+7+3) / 3 =5
        - Variance (5-5) 2 + (3 - 5)2 + (7- 5) 2 /3 -I =4
    - Quartile = data can be divided into four regions that cover the total range of observ values
        - Q 1=25%

e mean

ble, is

e data set

ation

mean

ved

- Q2 = 25-50%
- upper bound of Q2 = median
- Q3 = 25% - 75%
- Max observation = Q4