KEY STATISTICS

TOPICS

- Summary Statistics
- Inferential Statistics
- Exploratory Analysis
- Distribution Modeling
- Numerical Computation using python

SUMMARY STATISTICS

- Summary statistics summarize and provide information about the sample data.
- It describes and provides values in the data set.
- Summary statistics falls into three categories
  - Measures of location.
  - Measures of spread.
  - Graphs/charts.

Stephanie Glen. "Summary Statistics: Definition and Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/summary-statistics/

MEASURES OF LOCATION

- Measures of location tell you where your data is centered at, or where a trend lies.
- Examples:
  - Mean (arithmetic average value)
  - Geometric mean (used for interest rates and other types of growth)
  - Trimmed mean (mean with outliers excluded)
  - Median (middle of data set)

MEASURES OF SPREAD

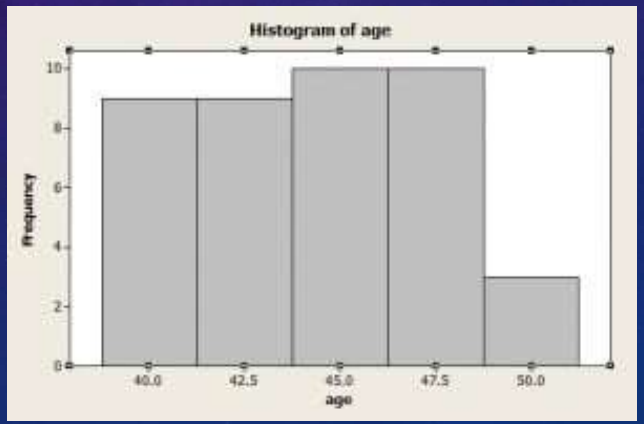
- Describes how spread out or varied the data set.
- Examples:
  - Range (how spread out the data is)
  - Interquartile range (where the "middle fifty" percent of data is)
  - Quartiles (boundaries for the lowest, middle, and upper quarters of data)
  - Skewed (does the data have mainly low, or mainly high values?)
  - Kurtosis (a measure of how much data is in the tails)

GRAPHS AND CHARTS

- There are literally dozens of ways to display summary data using graphs or charts.
- Examples:
  - Histogram
  - Frequency Distribution Table
  - Box Plot
  - Bar Chart
  - Scatter Plot
  - Pie Chart

HISTOGRAM

- Histograms are like bar charts; they are a way to display counts of data.
- A bar graph charts actual counts against categories; The height of the bar indicates the number of items in that category.
- A histogram displays the same categorical variables in “bins”.



FREQUENCY DISTRIBUTION TABLE

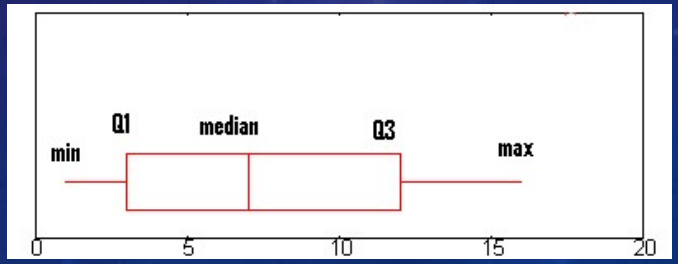
- Frequency tells you how often something happened.
- The frequency of an observation tells you the number of times the observation occurs in the data.

Number of Pets (x)	Tally	Frequency (f)
0	𝍸	4
1	𝍸 I	6
2	𝍸	5
3	III	3
4	II	2

Method	Number
Abstinence	14
Condoms	47
Injectables	1
Norplant	1
Pill	35
None	307
Total	405

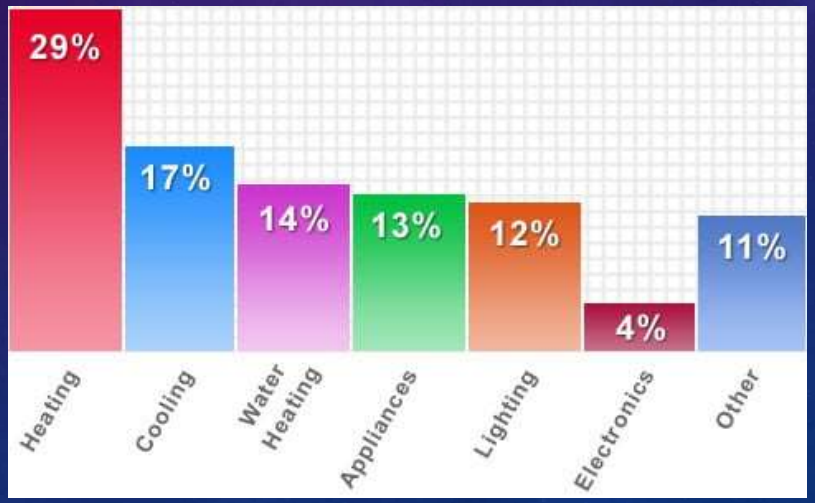
BOX PLOT

- A boxplot is a way to show a five number summary in a chart.
- The minimum (the smallest number in the data set). The minimum is shown at the far left of the chart, at the end of the left "whisker."
- First quartile, Q1, is the far left of the box (or the far right of the left whisker).
- The median is shown as a line in the center of the box.
- Third quartile, Q3, shown at the far right of the box (at the far left of the right whisker).
- The maximum (the largest number in the data set), shown at the far right of the box.



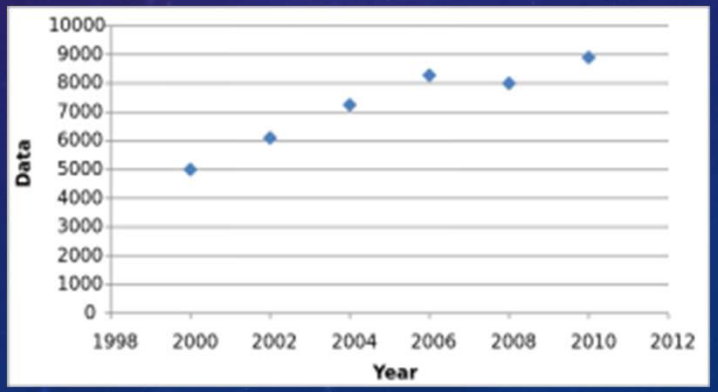
BAR CHART

- A bar graph is useful for looking at a set of data and making comparisons.
- For example, it's easier to see which items are taking the largest chunk of your budget by glancing at the above chart rather than looking at a string of numbers. They can also show trends over time or reveal patterns in periodic sequences.



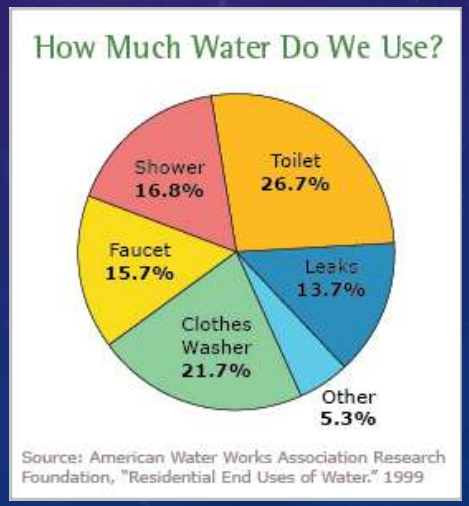
SCATTER PLOT

- A scatter plot uses dots to represent individual pieces of data.
- In statistics, these plots are useful to see if two variables are related to each other. For example, a scatter chart can suggest a linear relationship (i.e., a straight line).
- The relationship between variables is called correlation. Correlation is just another word for "relationship." For example, how much you weigh is related (correlated) to how much you eat.
- There are two type of correlation: positive correlation and negative correlation.



PIE CHART

- A Pie Chart is a type of graph that displays data in a circular graph.
- The pieces of the graph are proportional to the fraction of the whole in each category. In other words, each slice of the pie is relative to the size of that category in the group.
- The entire "pie" represents 100 percent of a whole, while the pie "slices" represent portions of the whole.



INFERENTIAL STATISTICS

- Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data.
- With inferential statistics, you take data from samples and generalize about a population.
- For example, you might stand in a mall and ask a sample of 100 people if they like shopping at Sears. You could make a bar chart of yes or no answers (that would be descriptive statistics) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (all shoppers in all malls) like shopping at Sears.

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis (EDA) is an approach to analyzing data. It's where the researcher takes a bird's eye view of the data and tries to make some sense of it.
- It's often the first step in data analysis, implemented before any formal statistical techniques are applied.
- Although specific statistical techniques can be used, like creating histograms or box plots, EDA is not a set of techniques or procedures; the Engineering Statistics Handbook calls EDA a "philosophy."
- EDA is considered by some to be more of an art form than a science.
- EDA involves the analyst trying to get a "feel" for the data set, often using their own judgment to determine what the most important elements in the data set are.

PURPOSE OF EDA

- Check for missing data and other mistakes.
- Gain maximum insight into the data set and its underlying structure.
- Check assumptions associated with any model fitting or hypothesis test.
- Create a list of outliers or other anomalies.
- Find parameter estimates and their associated confidence intervals or margins of error.
- Identify the most influential variables.

DISTRIBUTION MODELING

- Data distribution is a function that determines the values of a variable and quantifies relative frequency, it transforms raw data into graphical methods to give valuable information. It becomes substantial to understand the kind of distribution that a population has that assists in applying proper statistical techniques/methods.
- On the other hand, when statisticians or data experts analyze datasets, the very first step is to conduct exploratory data analysis (EDA) for learning about characteristics of a specific feature in datasets that help in understanding any pattern present in the data distributions.
- Through this way, they can tailor machine learning models suitable for case studies as ML models are designed under some data distribution assumptions.

TYPES OF DATA/STATISTICAL DISTRIBUTION MODEL

- Bernoulli's Distribution
- Binomial Distribution
- Normal (Gaussian) Distribution
- Poisson Distribution
- Exponential Distribution
- Multinomial Distribution
- Beta Distribution
- Beta-Binomial Distribution
- T-Distributions
- Uniform Distribution

BERNOULLI'S DISTRIBUTION

- Bernoulli's distribution has possibly two outcomes (success or failure) and a single trial.
- For example, tossing a coin, the success probability of an outcome to be heads is p, then the probability of having tail as outcome is (1-p). Bernoulli's distribution is the special case of binomial distribution with a single trial.

$$f(x) = p^x \ (1-p)^{(1-x)} \quad \text{where } x \in (0,1)$$

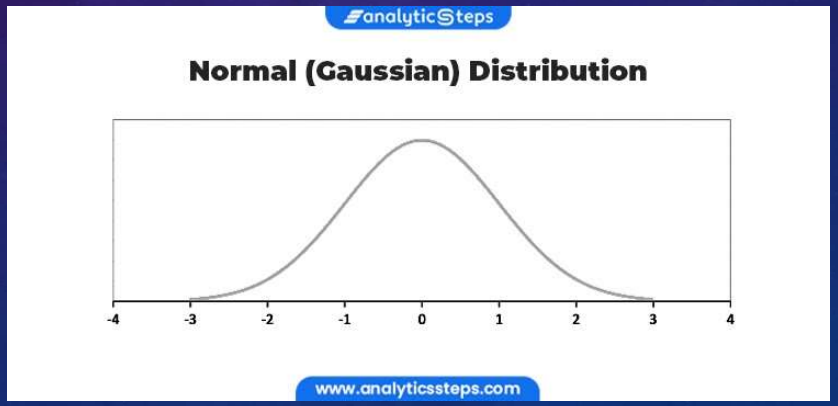
- The number of trials, to be performed, need to be predefined for a single experiment.
- Each trial has only two possible outcomes-success or failure.
- The probability of success of each event/experiment must be the same.
- Each event must be independent of each other.

BINOMIAL DISTRIBUTION

- The binomial distribution is applied in binary outcomes events where the probability of success is equal to the probability of failure in all the successive trials. Its example includes tossing a biased/unbiased coin for a repeated number of times.
- As input, the distribution considers two parameters, and is thus called as bi-parametric distribution. The two parameters are;
  - The number of times an event occurs, n, and
  - Assigned probability, p, to one of the two classes

NORMAL (GAUSSIAN) DISTRIBUTION

- Being a continuous distribution, the normal distribution is most used in data science.
- A very common process of our day-to-day life belongs to this distribution - income distribution, average employees report, average weight of a population, etc.
- The distribution is said to be normal if mean (μ) = 0 and standard deviation (σ) = 1



POISSON DISTRIBUTION

- Being a part of discrete probability distribution, Poisson distribution outlines the probability for a given number of events that take place in a fixed time period or space, or particularized intervals such as distance, area, volume.
- For example, conducting risk analysis by the insurance/banking industry, anticipating the number of car accidents in a particular time interval and in a specific area.
- Poisson distribution considers following assumptions;
  - The success probability for a short span is equal to success probability for a long period of time.
  - The success probability in a duration equals to zero as the duration becomes smaller.
  - A successful event can't impact the result of another successful event.

EXPONENTIAL DISTRIBUTION

- Like the Poisson distribution, exponential distribution has the time element; it gives the probability of a time duration before an event takes place.
- Exponential distribution is used for survival analysis, for example, life of an air conditioner, expected life of a machine, and length of time between metro arrivals.

NUMERICAL COMPUTATION USING PYTHON

ACTIVITY

THE END