Applied Statistics and Queuing Theory

CSE 3107 by MAH

RUET
Edited by: rmShoeb_CSE_16

Introduction

Statistics and its importance

Statistics is a branch of mathematics, working with data collection, organization, analysis, interpretation and presentation. In applying statistics to a scientific, industrial or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.

A **population** includes all of the elements from a set of data. A **sample** consists one or more observations drawn from the population. The main difference between a population and sample has to do with how observations are assigned to the data set. Depending on the sampling method, a sample can have fewer observations than the population, the same number of observations or more observations. More than one sample can be derived from the same population. A measurable characteristic of a population, such as a mean or standard deviation, is called a **parameter**; but a measurable characteristic of a sample is called a **statistic**.

The numerical descriptors of a sample are called statistics. These statistics may be categorized as location, spread, shape indicators, percentiles and interval estimates.

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics. Descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data. There are a number of items that belong in this portion of statistics, such as:

- The average or measure of the center of a data set, consisting of the mean, median, mode or midrange.
- The spread of a data set, which can be measured with the range or standard deviation.
- Overall descriptions of data such as the five number summary.
- Measurements such as skewness and kurtosis.
- The exploration of relationships and correlation between paired data.
- The presentation of statistical results in graphical form.

These measures are important and useful because they allow scientists to see patterns among data and thus to make sense of that data. Descriptive statistics can only be used to describe the population or data set under study. The results cannot be generalized to any other group or population.

There are two kinds of descriptive statistics that scientists use:

- **Measures of central tendency** capture general trends within the data and are calculated and expressed as the mean, median and mode. A mean tells scientists the mathematical average of all of a data set, such as the average age at first marriage; the median represents the middle of the data distribution, like the age that sits in the middle of the range of ages at which people first marry; and, the mode might be the most common age at which people first marry.
- **Measures of spread** describe how the data are distributed and relate to each other, including:
  - The range, the entire range of values present in a data set.
  - The frequency distribution, which defines how many times a particular value occurs within a data set.
  - Quartiles, subgroups formed within a data set when all values are divided into four equal parts across the range.
  - Mean absolute deviation, the average of how much each value deviates from the mean.
  - Variance, which illustrates how much of a spread exists in the data.
  - Standard deviation, which illustrates the spread of data relative to the mean

Measures of spread are often visually represented in tables, pie and bar charts, and histograms to aid in the understanding of the trends within the data.

Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population. Scientists use inferential statistics to examine the relationships between variables within a sample and then make generalizations or predictions about how those variables will relate to a larger population.

Although descriptive statistics is helpful in learning things such as the spread and center of the data, nothing in descriptive statistics can be used to make any generalizations. In descriptive statistics, measurements such as the mean and standard deviation are stated as exact numbers.

Even though inferential statistics uses some similar calculations — such as the mean and standard deviation — the focus is different for inferential statistics. Inferential statistics start with a sample and then generalizes to a population. This information about a population is not stated as a number. Instead, scientists express these parameters as a range of potential numbers, along with a degree of confidence.

Why statistics are important in our life? Statistics are the sets of mathematical equations that we used to analyze the things. It keeps us informed about, what is happening in the world around us. Statistics are important because today we live in the information world and much of this information's are determined mathematically by Statistics Help**.** It means to be informed correct data and statics concepts are necessary. To be more specific about the importance of statics in our life, here are 9 amazing reasons that we have heard on several occasions:

1. Everybody watches weather forecasting. Have you ever think how do you get that information? There are some computers models build on statistical concepts. These computer models compare prior weather with the current weather and predict future weather.
2. Statistics is mostly used by the researchers. They use their statistical skills to collect the relevant data. Otherwise, it results in a loss of money, time and data.
3. What do you understand by insurance? Everybody has some kind of insurance, whether it is medical, home or any other insurance. Based on an individual application, some businesses use statistical models to calculate the risk of giving insurance.
4. In financial market also statistic plays a great role. Statistics are the key of how traders and businessmen invest and make money.
5. Statistics play a big role in the medical field. Before any drugs prescribed, scientist must show a statistically valid rate of effectiveness. Statistics are behind all the study of medical.
6. Statistical concepts are used in quality testing. Companies make many products on a daily basis and every company should make sure that they sold the best quality items. But companies cannot test all the products, so they use statistics sample.
7. Doctors predict disease based on statistics concepts. Suppose a survey shows that 75%-80% people have cancer and not able to find the reason. When the statistics become involved, then you can have a better idea of how the cancer may affect your body or is smoking is the major reason for it.
8. News reporter makes a prediction of winner for elections based on political campaigns. Here statistics play a strong part in who will be your governments.
9. Statistics data allow us to collect the information around the world. The internet is a devise which help us to collect the information. The fundamental behind the internet is based on statistics and mathematics concepts.

Variable and constants

A **variable** is a characteristic, often but not always quantitatively measured, containing two or more values or categories that can vary from person to person, object to object or from phenomenon to phenomenon. Variables can be classified as Qualitative (or Categorical**)** and Quantitative (or Numerical).

- **Qualitative variables** take on values that are names or labels. Numerical measurements are not possible. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of qualitative or categorical variables.
  - Nominal variable: categorical variable without an intrinsic order such as where a person lives in the U.S. Used for labeling variables.
  - Ordinal variable: categorical variable with some intrinsic order such as education, agreement. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort etc.
- **Quantitative variables** are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.
  - **Interval scales**
    Interval scales are numeric scales in which we know both the order and the exact differences between the values. The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.
    Interval scales are nice because the realm of statistical analysis on these data sets opens up. For example, central tendency can be measured by mode, median, or mean; standard deviation can also be calculated.
    Here's the problem with interval scales: they don't have a "true zero." For example, there is no such thing as "no temperature," at least not with Celsius. In the case of interval scales, zero doesn't mean the absence of value, but is actually another number used on the scale, like 0 degrees Celsius. The zero isn't meaningful, it doesn't mean a true absence of something. Negative numbers also have meaning. Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide.
  - **Ratio scales**
    Everything above about interval data applies to ratio scales, plus ratio scales have a clear definition of zero. Good examples of ratio variables include height, weight, and duration.
    Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.
- **Continuous variable:** A continuous variable is a numeric variable. Observations can take any value between a certain set of real numbers. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.
- **Discrete variable:** A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of which measured as whole units (i.e. 1, 2, 3 cars).

Univariate data: When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

Bivariate data: When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

Multivariate data: When we conduct a study that examines the relationship among more than two variables, we are working with multivariate data. Suppose we conducted a study to see if there were a relationship

among the height, weight, and age of high school students. Since we are working with three variables (height, weight, age), we would be working with multivariate data.

The term **constant** refers to a property whereby the members of a group or category remain fixed and do not vary one from another.

Statistical data

Data are the facts and figures collected, summarized, analyzed and interpreted. The data collected in a particular study are referred to as the **data set.** The **elements** are the entities on which data are collected. A **variable** is a characteristic about each individual element. The set of measurements collected for a particular element is called an observation. The total number of data values in a data set is the number of elements multiplied by the number of variables

Data collection and presentation

Measures of position: Measures of position are used to describe the relative location of an observation. Quartiles and percentiles are two of the most popular measures of position. A measure of relative position tells where data values fall within the ordered set.

- **Percentiles:** The pth percentile is a value such that p percent of the observations fall below or at that value.

  $$l = n\frac{p}{100} \qquad where, l\ is\ the\ location\ of\ the\ data\ value$$

  When using this formula to find the location of the percentile's value in the data set, you must make sure to follow these two rules:
  1. If the formula results in a decimal value for l, the location is the next largest integer.
  2. If the formula results in a whole number, the percentile's value is the average of the value in that location and the one in the next largest location.

  When calculating the percentile, always round up to the next integer.
- **Quartiles**
  Values of the variable that divide the ranked data into quarters. The 25th percentile is called the *first (lower) quartile*, the 50th percentile is the *median*, and the 75th percentile is called the *third (upper) quartile*.
- **Interquartile range**: The interquartile range is the distance between the third and first quartile, giving spread of middle 50% of the data.
- Advantages and disadvantages of the techniques
- **Z-score**
  Center থেকে কত standard deviation দ্বারা data অবস্থিত।
  The z-score for an observation is the number of standard deviations that it falls from the mean. For sample data, the z-score is calculated as

  $$z = \frac{observation - mean}{standard\ deviation}$$

  An observation from a bell-shaped distribution is a potential outlier if its z-score < -3 or > +3.
  Typically, the calculated value of z is rounded to the nearest hundredth. The z-score measures the number of standard deviations above/below, or away from, the mean. z-scores typically range from -3.00 to +3.00. z-scores may be used to make comparisons of raw score.
  Why z-scores? Transforming raw scores to z-scores facilitates making comparisons, especially when using different scales. A z-score provides information about the relative position of a score in relation to other scores in a sample or population. A raw score provides no information regarding the relative standing of the score relative to other scores. A z-score tells one how many standard deviations the score is from the mean. It also provides the approximate percentile rank of the score relative to other scores.

Measures of relative position

- **Midquartile:** The numerical value midway between the first and third quartile.

$$Midquartile = \frac{Q_1 + Q_3}{2}$$

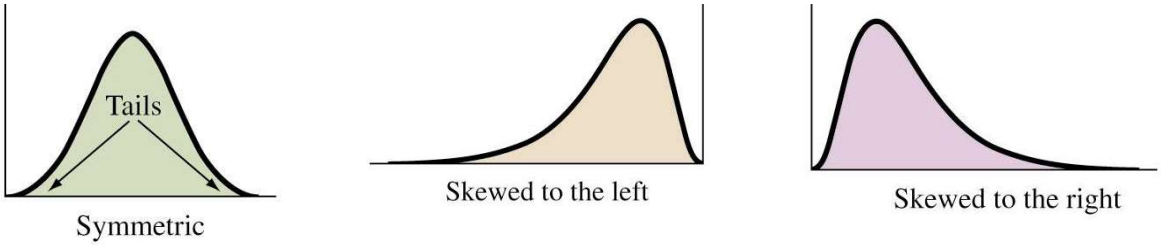
- **Midrange**

5 number summary: The 5-number summary indicates how much the data is spread out in each quarter.

1. Minimum value
2. first quartile
3. median
4. third quartile
5. maximum value.

Construction of frequency distribution

- Symmetrical frequency distribution
- Skewed frequency distribution
- Grouped and ungrouped distribution
- Cumulative frequency distribution: The cumulative frequency for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values.
  The cumulative relative frequency for any given class is the sum of the relative frequency for that class and the relative frequencies of all classes of smaller values.



Graphical presentation

- **Bar graph**
- **Histograms:** A bar graph representing a frequency distribution of a quantitative variable.
  The histogram is a traditional way of displaying the shape of a group of data. It is constructed from a frequency distribution, where choices on the number of bins and bin width have been made. These choices can drastically affect the shape of the histogram. The ideal shape to look for in the case of normality is a bell-shaped distribution.
- Difference between bar graphs and histograms
  With bar charts, each column represents a group defined by a categorical variable. With histograms, each column represents a group defined by a quantitative variable. With bar charts, however, the X axis does not have a low end or a high end; because the labels on the X axis are categorical - not quantitative. As a result, it is not appropriate to comment on the skewness of a bar chart.
- **Pie charts**
  Pareto charts
  Stem and Leaf display
  The stem-leaf plot is a type of histogram which retains much of the identity of the original data. It is useful for finding data-entry errors as well as for studying the distribution of a variable. The stem is the first digit of the actual number. The leaf is the second digit of the actual number.
  Example: 15 17 52 25 20 98 47 44 18 19 45 21 30 33 35 90 92 46 41 37

	STEM & LEAF
1	5 7 8 9
2	0 1 5
3	0 3 5 7
4	1 4 5 6 7
5	2

- **Box and Whisker Display**
  The five numerical values are located on a scale, either vertical or horizontal. The box is used to depict the middle half of the data that lies between the two quartiles. The whiskers are line segments used to depict the other half of the data. One line segment represents the quarter of the data that is smaller in value than the first quartile. The second line segment represents the quarter of the data that is larger in value that the third quartile.
- **Box plot**
  Boxplots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions. It is a graphical representation of a five number summary.
  Steps for creating a box plot:
  1. Begin with a horizontal (or vertical) number line.
  2. Draw a small line segment above (or next to) the number line to represent each of the numbers in the five-number summary.
  3. Connect the line segment that represents the first quartile to the line segment representing the third quartile, forming a box with the median's line segment in the middle.
  4. Connect the "box" to the line segments representing the minimum and maximum values to form the "whiskers".

Empirical rule

If the data distribution is bell-shaped, then the interval:

- $\mu \pm \sigma$ contains about 68% of the values in the population or the sample.
- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample.
- $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample.

Chabyshev's theorem (probably not important, for now)

Misleading data displays

- The frequency scale should start at zero to present a complete picture. Graphs that do not start at zero are used to save space.
- Graphs that start at zero emphasize the size of the numbers involved.
- Graphs that are chopped off emphasize variation.

The role of statistics in Computer Science

Measures of Central Tendency

One of the first impressions that we like to get from a variable is its general location. You might think of this as the center of the variable on the number line. The average (mean) is a common measure of location. When investigating the center of a variable, the main descriptors are the mean, median, mode and the trimmed mean. Other averages, such as the geometric and harmonic mean, have specialized uses.

If the data come from the normal distribution, the mean, median, mode and the trimmed mean are all equal. If the mean and median are very different, most likely there are outliers in the data or the distribution is skewed. If this is the case, the median is probably a better measure of location. The mean is very sensitive to extreme values and can be seriously contaminated by just one observation.

Arithmetic mean

Sample mean, $\bar{X} = \frac{\sum X}{n}$

Population mean, $\mu = \frac{\sum X}{N}$

Properties:

- Every set of interval-level and ratio-level data has a mean.
- All the values are included in computing the mean.
- A set of data has a unique mean.
- The mean is affected by unusually large or small data values.
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.
- It is the most common measure of central tendency.

Weighed mean

The weighted mean of a set of numbers X_1, X_2, ..., X_n, with corresponding weights w_1, w_2, ...,w_n, is computed from the following formula:

$$\overline{X_w} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n}$$

Geometric mean

The geometric mean (GM) is an alternative type of mean that is used for business, economic, and biological applications. Only nonnegative values are used in the computation. If one of the values is zero, the geometric mean is defined to be zero. One example of when the GM is appropriate is when a variable is the product of many small effects combined by multiplication instead of addition. The geometric mean (GM) of a set of n numbers is defined as the n^{th} root of the product of the n numbers. The geometric mean is used to average percent, indexes, and relatives.

$$GM = \sqrt[n]{(X_1)(X_2)\dots(X_n)}$$

Harmonic mean

The harmonic mean is used to average rates. For example, suppose we want the average speed of a bus that travels a fixed distance every day at speeds $s_1, s_2\ and\ s_3$. The average speed, found by dividing the total distance by the total time, is equal to the harmonic mean of the three speeds. The harmonic mean is appropriate when the distance is constant from trial to trial and the time required was variable. However, if the times were constant and the distances were variable, the arithmetic mean would have been appropriate. Only nonzero values may be used in its calculation.

$$HM = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Relation between arithmetic, geometric and harmonic mean

Median

The median is the midpoint of the values after they have been ordered from the smallest to the largest. There are as many values above the median as below it in the data array. For an even set of values, the median will be the arithmetic average of the two middle numbers.

Properties:

- There is a unique median for each data set.
- It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.

- It can be computed for ratio-level, interval-level, and ordinal-level data.
- It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.
- Robust measure of central tendency
- Not affected by extreme values

In an ordered array, the median is the "middle" number

- If n or N is odd, the median is the middle number
- If n or N is even, the median is the average of the two middle numbers

Mode

The mode is the value of the observation that appears most frequently.

Properties:

Measures of Dispersion

After establishing the center of a variable's values, the next question is how closely the data fall about this center. The pattern of the values around the center is called the *spread*, *dispersion* or *variability*. There are numerous measures of variability: range, variance, standard deviation, interquartile range and so on. All of these measures of dispersion are affected by outliers to some degree, but some do much better than others.

The *standard deviation* is one of the most popular measures of dispersion. Unfortunately, it is greatly influenced by outlying observations and by the overall shape of the distribution. Because of this, various substitutes for it have been developed. It will be up to you to decide which is best in a given situation.

Range

The difference between the largest and smallest values for a variable. If the data for a given variable is normally distributed, a quick estimate of the standard deviation can be made by dividing the range by six.

Deviation from the mean $(\boldsymbol{x - \bar{x}})$ is the difference between the value of x and the mean $\bar{x}$.

Mean absolute deviation $\left(\frac{1}{n}\sum|\boldsymbol{x - \bar{x}}|\right)$ is the mean of the absolute values of the deviation from the mean.

Sample Variance

The sample variance, s^2, is the mean of the squared deviations, calculated using $n - 1$ as the divisor:

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 \qquad \text{where, } n \text{ is the sample size}$$

The numerator for the sample variance is called the sum of squares for x, denoted $SS(x)$:

$$s^2 = \frac{SS(x)}{n-1} \qquad where, SS(x) = \sum(x - \bar{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

Standard deviation

The sample standard deviation, *s*, is a popular measure of dispersion. It measures the average distance between a single observation and its mean. The use of *n-1* in the denominator instead of the more natural *n* is often of concern. It turns out that if *n* (instead of *n-1*) were used, a biased estimate of the population standard deviation would result. The use of *n-1* corrects for this bias.

Unfortunately, *s* is inordinately influenced by outliers. For this reason, you must always check for outliers in your data before you use this statistic. Also, *s* is a biased estimator of the population standard deviation.

$$\text{Sample Standard Deviation, } s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

$$\text{Population Standard Deviation, } \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Mean and Standard Deviation of Frequency Distribution

If the data is given in the form of a frequency distribution, we need to make a few changes to the formulas for the mean, variance, and standard deviation. Complete the extension table in order to find these summary statistics. In order to calculate the mean, variance, and standard deviation for data:

- In an *ungrouped* frequency distribution, use the frequency of occurrence, *f*, of each observation
- In a *grouped* frequency distribution, we use the frequency of occurrence associated with each class midpoint:

$$\bar{x} = \frac{\sum xf}{\sum f} \qquad s^2 = \frac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}$$

Outliers

Outliers in a univariate data set are defined as observations that appear to be inconsistent with the rest of the data. An outlier is an observation that sticks out at either end of the data set.

The visualization of univariate outliers can be done in three ways: with the stem-and-leaf plot, with the box plot, and with the normal probability plot. In each of these informal methods, the outlier is far removed from the rest of the data. A word of caution: the box plot and the normal probability plot evaluate the potentiality of an outlier assuming the data are normally distributed. If the variable is not normally distributed, these plots may indicate many outliers. You must be careful about checking what distributional assumptions are behind the outliers you may be looking for.

Outliers can completely distort descriptive statistics. For instance, if one suspects outliers, a comparison of the mean, median, mode, and trimmed mean should be made. If the outliers are only to one side of the mean, the median is a better measure of location. On the other hand, if the outliers are equally divergent on each side of the center, the mean and median will be close together, but the standard deviation will be inflated. The interquartile range is the only measure of variation not greatly affected by outliers. Outliers may also contaminate measures of skewness and kurtosis as well as confidence limits.

Procedure to check for outliers:

- *Step 1:* Arrange the data in order from smallest to largest.
- *Step 2:* Determine the first quartile Q_1 and the third quartile Q_3.
- *Step 3:* Find the interquartile range (IQR). IQR = $Q_3 - Q_1$.
- *Step 4:* Compute ($Q_1 - 1.5\times$IQR) and ($Q_3 + 1.5\times$IQR).
- *Step 5:* Let x be the data value that is being checked to determine whether it is an outlier.

  a) If the value of x is smaller than ($Q_1 - 1.5\times$IQR), then x is classified as an outlier.
  b) If the value of x is larger than ($Q_3 + 1.5\times$IQR), then x is classified as an x is an outlier.

Outlier based on z-score:

- When we consider the empirical rule, an observation with a z-score < -2.00 or z-score > 2.00 might be characterized as a mild outlier.
- Any observation with a z-score < - 3.00 or z-score > 3.00 might be characterized as an extreme outlier.

Moments

The shape of the distribution describes the pattern of the values along the number line. Are there a few unique values that occur over and over, or is there a continuum? Is the pattern symmetric or asymmetric? Are the data bell shaped? Do they seem to have a single center or are there several areas of clumping? These are all aspects of the shape of the distribution of the data.

Two of the most popular measures of shape are skewness and kurtosis. *Skewness* measures the direction and lack of *symmetry*. The more skewed a distribution is, the greater the need for using robust estimators, such as the median and the interquartile range. *Kurtosis* measures the heaviness of the tails. A kurtosis value less than three indicates lighter tails than a normal distribution. Kurtosis values greater than three indicate heavier tails than a normal distribution.

The measures of shape require more data to be accurate. For example, a reasonable estimate of the mean may require only ten observations in a random sample. The standard deviation will require at least thirty. A reasonably detailed estimate of the shape (especially if the tails are important) will require several hundred observations.

Skewness

This statistic measures the direction and degree of asymmetry. A value of zero indicates a symmetrical distribution. A positive value indicates skewness to the right while a negative value indicates skewness to the left. Values between -3 and +3 indicate are typical values of samples from a normal distribution.

Let $x_1, x_2, \dots, x_n$ be n obsrvations. Then, Skewness $= \dfrac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$

Kurtosis

This statistic measures the heaviness of the tails of a distribution. The usual reference point in kurtosis is the normal distribution. If this kurtosis statistic equals three and the skewness is zero, the distribution is normal. Unimodal distributions that have kurtosis greater than three have heavier or thicker tails than the normal. These same distributions also tend to have higher peaks in the center of the distribution. Unimodal distributions whose tails are lighter than the normal distribution tend to have a kurtosis that is less than three. In this case, the peak of the distribution tends to be broader than the normal. Be forewarned that this statistic is an unreliable estimator of kurtosis for small sample sizes.

Let $x_1, x_2, \dots, x_n$ be n obsrvations. Then, Kurtosis $= \dfrac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$

Correlation Theory

Type 1:

- Positive correlation
- Negative correlation

Type 2:

- Simple
- Multiple
- Partial
- Total

Type 3:

- Linear
- Non-linear

Scattergrams(advantages and disadvantages)

Covariance

- Converting covariance to correlation
- Covariance calculation
- Difference between covariance and correlation. Which one is better?

Interpretation of correlation coefficient

Advantage and limitations of Pearson's

Time consuming method

Linear correlation and its measures and significance

Rank correlation

Regression Analysis

Linear and non-linear regression

Least-Square method of Curve fittings

Probability

Probability and Statistical Inference_-_Hogg and Tannis

Page(1-64)

Elementary Concepts

Laws of probability

Additive law

Multiplicative law

Conditional probability and Bayes theorem

Random variables, x

Given a random experiment with an outcome space S, a function X that assigns one and only one real number $X(s) = x$ to each element s in S is called a random variable.

PMF/PDF: probability mass function/probability density function, f(x)

Mean of discrete random variables

Variance standard deviation

$$\sigma^2 = E[(X-\mu)^2]$$
$$= E[X^2 - 2X\mu + \mu^2]$$
$$= E(X^2) - 2\mu E(X) + \mu^2$$
$$= E(X^2) - 2\mu^2 + \mu^2 \quad [where, E(X) = \mu]$$
$$= E(X^2) - \mu^2$$

Standard deviation $= \sqrt{\sigma^2}$

Variance random variable

Continuous random variable

Measure of central tendency for random variables

Probability distribution function

Uniform distribution: if the probability of every outcome is same.

Uniform probability

Hyper geometric distribution

Descriptive probability

Mathematical expectation, μ

Mathematical expectation will give middle value of any random variable.

Necessity of mathematical expectation

Variance from mathematical expectation

Probability Distributions

Probability and Statistical Inference_-_Hogg and Tannis

Page(65-94, 105-113)

Discrete distribution and continuous distribution

Binomial distribution

Binomial histogram

Bernoulli trials

Bernoulli distribution

Moment generating functions

Properties

Advantages

Application

Central moment

Rules for expectation

MGF of a RVX

Properties

Poisson distribution

Properties

Probability mass function of Poisson's

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0,1,2,\ldots \qquad \lambda \geq 0$$

Moment generating function

$$E(e^{tx}) = \sum_{x \in S} e^{tx} f(x)$$

$$M(t) = E(tx) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x e^{tx}}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$$= e^{\lambda(e^t - 1)}$$

CDF: Cumulative Distribution Function

Normal distribution

Most used continuous distribution

Queuing Theory

What is queuing theory?

Why is it important?

Stochastic processes

Discrete time Markov Chain and Continuous time Markov Chain

Birth-death process in Queuing

Queuing models

M/M/1

M/M/C

M/G/1

M/D/1

G/M/1

Solution of network of Queue-closed queuing models and approximate models

Application of Queuing models in Computer Science

Cumulative probability, F(x)

Exponential distribution function

Derivation of poisson distribution

Mean, variance ber kora shikhte hobe.

Base theorem

[Resources:

1. https://medium.com/@john_marsh7/10-awesome-reasons-why-statistics-are-important-96b87e283640
2. https://stattrek.com/tutorials/ap-statistics-tutorial.aspx

3. https://www.probabilitycourse.com/
4. Taylor, Courtney. "The Difference Between Descriptive and Inferential Statistics." ThoughtCo. https://www.thoughtco.com/differences-in-descriptive-and-inferential-statistics-3126224 (accessed October 13, 2019).
5. https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
6.