**OM 305**

**Spring 2023**

**Quiz 1**

**Statistics Class Notes**

**as of 1/29/23**

**1. Statistical inference tools:** tools that allow a decision maker to reach a conclusion about a population of data based on a subset of data from population.

**2. Experiment:** a process that produces a single outcome whose result cannot be predicted with certainty.

**3. Sample:** a subset of the population.

**4. Population Census:** an enumeration of the entire set of measurements taken from the whole population. Each parameter should be reached with a survey.

**5. Non-statistical sampling techniques:** those methods of selecting samples using convenience, judgement, or other non-chance processes.

**6. Quantitative data:** measurements whose values are inherently numerical.

**7. Qualitative data:** data whose measurement scale is inherently categorical.

**8. Parameter:** a measure computed from the entire population. As long as the population does not change, the value of the parameter will not change.

**9. Statistic:** a measure computed from a sample that has been selected from a population. The value of the statistic will depend on which sample is selected.

**10. Dependent variable:** a variable whose values are thought to be a function of, or dependent on, the values of another variable called the independent variable.

**11. Independent variable:** a variable whose values are thought to impact the values of the dependent variable. The independent variable, or explanatory variable, is often within direct control of the decision maker.

**12. Right-skewed data:** a data distribution when the mean of the data is larger than the median.

**13. Left-skewed data:** a data distribution where the mean for the data is smaller than the median.

**14. Population Variance:** is the average of the squared distances of the data values from the mean. It a measure of variability that utilizes all data. The variance is based on the deviation about the mean.

**15. Standard deviation:** is the positive square root of the variance. It is a distance from the mean on the x axis.

**16. Standardization of data values (standard normal distribution**): a normal distribution that has a mean = 0 and a standard deviation = 1,

Standardized data values are referred to as z-scores. It is a distance from the mean on the x-axis.

**17. Coefficient of variation:** the ratio of the standard deviation to the mean expressed as a percentage. The coefficient of variation is used to measure the relative variation in data.

**18. Decision tree:** a diagram that illustrates the correct ordering of actions and events in a decision-analysis problem. Each act or event is represented by a branch on the decision tree.

**19.** **Axioms:** givens and assumptions that are necessary in analytical work. In [logic](https://www.britannica.com/topic/logic), an indemonstrable first principle, rule, or maxim, that has found general acceptance or is thought worthy of common acceptance whether by virtue of a claim to [intrinsic](https://www.merriam-webster.com/dictionary/intrinsic) merit or on the basis of an appeal to self-evidence. An example would be: “Nothing can both be and not be at the same time and in the same respect”, or the “80-20 rule.”

**20. Consumer Research process (Burke video):** is never fully reliable or unbiased or can yield perfect information. Prospective consumers are irrational in their behavior.

**21. Forensic research:** a retrospective, descriptive research based on historical data. Typically, it is applied in GAPS analysis (explaining differences between stated goals and the actual results).

**22. Predictive analysis** – a decision making process about the future when the future is uncertain.

**23. Empirical Rule:** The empirical rule states that for a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), nearly all of the data will fall within three [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.datasciencecentral.com/mean/).

The empirical rule can be broken down into three parts:

* 68.28% of data falls within the first standard deviation from the mean.
* 95.44% falls within two standard deviations from the mean.
* 99.72% falls within three standard deviations from the mean.

**24. Mean:** a numerical measure of the center of a set of the quantitative measures computed by dividing the sum of the values by the number of values in the data.

**25. Weighted mean:** the mean value of data values that have been weighted according to their relative importance.

**26. Median**: is the center value that divides a data array into two halves.

**27. Mode:** is the value on a data set that occurs most frequently.

**28.** **The concept of “self”:** it is a product of all external influences on an individual and all internal characteristics of that individual.

**29. Bias:** is a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief. In science and engineering, a bias is a systematic error. In statistics, it is the tendency of a predictive model to overestimate or underestimate the value of a continuous outcome. It is undesirable and unavoidable.

**30. Business Statistics:** This is a form of mathematical analyses that is used to convert data into meaningful information. It uses quantified models, representations and synopses for a given set of experimental data or real-life studies. The discipline of Business Statistics studies methodologies to gather, review, analyze and draw conclusions from data. It focuses on issues relevant to business.

**31. Factorials:** i.e. :0! =1

**32. Utility:** In [economics](https://en.wikipedia.org/wiki/Economics) and [finance](https://en.wikipedia.org/wiki/Finance), risk aversion (one’s utility) is the behavior of [humans](https://en.wikipedia.org/wiki/Human) (especially [consumers](https://en.wikipedia.org/wiki/Consumer) and [investors](https://en.wikipedia.org/wiki/Investor)), who, when exposed to [uncertainty](https://en.wikipedia.org/wiki/Uncertainty), attempt to lower that uncertainty.

It is the hesitation of a person to agree to a situation with an unknown payoff rather than another situation with a more predictable payoff, but possibly a lower [expected payoff](https://en.wikipedia.org/wiki/Expected_value).

For example, a risk-averse investor might choose to put their money into a [bank](https://en.wikipedia.org/wiki/Bank) account with a low but guaranteed interest rate, rather than into a [stock](https://en.wikipedia.org/wiki/Stock) that may have high expected returns, but also involves a chance of losing value. It is a measure of the total worth or relative desirability of a particular outcome.

**33. Normal probability distribution:** is a bell-shaped distribution with the following properties. It is:

a) unimodal: that is, the normal distribution peaks at a single value.

b) symmetrical: this means that the two areas under the curve between the mean and any two points equidistant on either side of the mean are identical.

c) is asymptotic to the x-axis (the normal curve approaches the x-axis as to moves toward the + or- infinity but never cross them.

also,

d) one side of the distribution is the mirror image of the other.

e) the mean, medial, and mode are equal.

f) the amount of variation in the random variable determines the width of the normal distribution.

g) the tails never cross the x-axis.

**34. Probability**: the chance that a particular event will occur. The probability of an event will have a value in the range from 0 to1.

**35. Random variable:** a variable that assigns a numerical value to each outcome of a random experiment or trial.

**36. Discrete random variable:** a random variable that can only assume a countable number of values.

**37. Continuous random variable:** random variables that can assume any value in an interval.

**38. Outliers:** observations with unusually large or unusually small values. These extreme values are called outliers. Numbers larger than 2 sigma should be tested for validity.

**39. Frequency distribution:** is a tabular summary of data showing the number (frequency) of data values in each of several categories.

**40. Variation:** captures differences in values of a variables of interest.

**41. Cumulative relative frequency distribution:** is a running total of the proportions of observations with a given array of numbers.

**42. Uncertainty:** no one can predict the future, yet business decisions about the future have to be made. Good decisions will create value and bad decisions will destroy it.

**43. Big data:** any set of data that is so large or too complex to be handled by standard data processing techniques and typical desktop software.

**44. Data mining:** the use of analytical techniques for better understanding patterns and relationships that exist in large data sets.

**45. Data query:** a request for information with certain characteristics from a database.

**46 Data scientist:** analysts trained in both computer science and statistics who know how to effectively process and analyze massive amounts of data.

**47. Data security:** protecting stored data from destructive forces or unauthorized use.

**48. Descriptive analytics:** analytical tools that describe what has happened.

**49. Optimization model:** a mathematical model that gives the best decision, subject to decision’s constraints.

**50. Decision-making:** involves the following steps: 1. Identify and define the problem,2. Determine the criteria that will be used to evaluate alternative solutions, 3. Identify the set of alternative solutions,4. Evaluate the alternatives,4. Choose the best alternative.

**51. For Vs of big data:** volume, velocity, variety, veracity

**52. Correlation coefficient:** a standardized measure of linear association between two variables that takes on values between -1 and +1. Values near -1 indicate a strong negative linear relationship, values meat +1 indicate a strong positive linear relationship, and the values near zero indicate the lack of linear relationship.

**53. Cross-sectional data:** data collected at the same or approximately the same point in time at different locations.

**54. Binomial probability distribution:** a probability distribution for a discrete random variable showing the probability of “x” successes in “n” trials. There are always two possible outcomes.

**55. Uniform probability distribution:** a continuous probability distribution for which the probability that the random variable will assume a value in any interval is the same for each interval of equal length.

**56. Poisson probability distribution**: a discrete distribution that is constructed from the probability of occurrence of rare events over an interval; focuses only on the number of discrete occurrences over some interval or continuum.

**57. Exponential probability distribution:** a continuous probability distribution that is useful in computing probabilities for the time it takes to complete a task or the time between two points in time.

**58.Variable:** a characteristic or quantity of interest that can take of different values.

**59. z-score:** denotes the number of standard deviations from the mean.

**60. Observation:** a set of values corresponding to a set of variables.

**61. Population:**  the set of parameters of interest in a particular study.

**62. Operations decision:** a decision concerned with how the organization is run from day-to-day.

**63. Prescriptive analytics:** techniques that analyze input data and yield a best course of action.

**64. Strategic decision:** a decision that involves higher-level issues and that is concerned with the overall direction of an organization, defining the overall goals and aspirations for the organization’s future.

**65. Tactical decisions:** a decision concerned with how the organization should achieve the goals and objectives set by its strategy.

**66. Simulation:** the use of probabilities to construct a computer model to study the impact of uncertainty on the decision at hand.

**67. Simulation optimization:** the use of probabilities to model uncertainty. When combined with optimization techniques, it enables researchers to identify the best solution to highly complex and highly uncertain business challenges.

**68. Covariance:** a measure of linear association between two variables. Positive values indicate positive relationship. Negative values indicate the lack of linear relationship.

**69. Time-series data:** data points that are collected over a period of time (i.e.: hours, minutes, years, etc.)

**70. Bayes Method:**  an extension of the conditional law of probabilities discovered by Thomas Bayes that can be used to revise probabilities.

**71. Correlation:** a measure of the degree of relatedness of two or more variables.

**72. Making Decisions with Uncertainty:** these decisions are made without perfect information. Hence, they need to consider the range of possible outcomes and the probability of the most likely outcome.

**73. Permutations:** is a subgroup of objects selected from a larger group of objects, where the order of selection or position in the subgroup does matter.

**74. Combination:** is a subgroup of objects selected from a larger group of objects, where the order of selection or position in the subgroup does not matter.

**75. Venn diagram:** a visual representation of a probability experiment using circles to represent events.