### **Terms and Concepts**

1. **Bessel's Correction**
   * Bessel's correction is the use of n−1 instead of n in the formula for the sample variance and sample standard deviation, where n is the number of observations in a sample.
   * This method corrects the bias in the estimation of the population variance.
   * It also partially corrects the bias in the estimation of the population standard deviation.
2. **Bias**
   * Bias is a systematic error that can occur in a statistical analysis, leading to inaccurate or misleading results.
   * It can be caused by various factors, such as sampling bias, measurement bias, or human bias.
3. **Degrees of Freedom**
   * Degrees of freedom are the number of independent values that a statistical analysis can estimate.
   * You can also think of it as the number of values that are free to vary as you estimate parameters.
   * Degrees of freedom is a combination of how much data you have and how many parameters you need to estimate.
4. **Standard Deviation**
   * The standard deviation is a measure of the spread or dispersion of data points around the mean.
   * It is calculated as the square root of the variance.
   * A higher standard deviation indicates that the data is more spread out, while a lower standard deviation indicates that the data is more clustered around the mean.
5. **Boosting**
   * In machine learning, boosting is an ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones.
   * Boosting is based on the question posed by Kearns and Valiant (1988, 1989): "Can a set of weak learners create a single strong learner?".
   * A weak learner is defined to be a classifier that is only slightly correlated with the true classification (it can label examples better than random guessing).
6. **Bootstrapping**
   * Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples.
   * This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics.
   * These resamples are known as sampling distributions of estimates.
7. **Central Limit Theorem**
   * The central limit theorem (CLT) is a statistical concept that states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
   * Sample means will be normally distributed, even if the population isn't normally distributed.
   * The central limit theorem is important in hypothesis testing and confidence intervals.
8. **Law of Large Numbers**
   * In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times.
   * According to the law, the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to the expected value as more trials are performed.
   * Guarantees stable long-term results for the averages of some random events.
9. **Distribution of Sample Means**
   * Imagine taking many random samples from a population and calculating the mean of each sample. The distribution of these sample means is called the "distribution of sample means."
   * Crucially, even if the original population isn't normally distributed, the distribution of sample means tends towards a normal distribution as the sample size increases. This is the foundation of the Central Limit Theorem.
   * This distribution has its own mean (close to the population mean) and standard deviation (called the standard error).
10. **Confidence Intervals**
    * A confidence interval is a range of values around a sample statistic (like the mean) that is likely to contain the true population parameter with a certain level of confidence (e.g., 95%).
    * It essentially gives you a plausible range for where the true population value lies, based on your sample data.
    * Confidence intervals are affected by sample size and variability. Larger samples and lower variability lead to narrower confidence intervals, indicating more precise estimates.
11. **Uncertainty**
    * In the statistical sense, uncertainty refers to the lack of complete knowledge about a population or phenomenon. We use samples to estimate population characteristics, but there's always some level of uncertainty because we haven't measured the entire population.
    * Uncertainty can be quantified using concepts like confidence intervals, standard errors, and margins of error.
12. **Margin of Error**
    * The margin of error is the range of values above and below the sample statistic in a confidence interval.
    * It essentially tells you how much your sample estimate might differ from the true population value due to random sampling error.
    * A larger margin of error indicates greater uncertainty in the estimate.
13. **Standard Error**
    * The standard error is the standard deviation of the distribution of sample means.
    * It measures how much the sample means vary from the true population mean.
    * A smaller standard error indicates that the sample means are clustered more closely around the population mean, suggesting a more precise estimate.
14. **Relationships and Comparisons**

* The distribution of sample means is the foundation for understanding confidence intervals and standard errors.
* Confidence intervals are constructed using the margin of error, which is calculated using the standard error.
* All these concepts help quantify and manage uncertainty in statistical analysis.
* By understanding these concepts and their relationships, you can better interpret statistical results, estimate population parameters, and make informed decisions based on data analysis.

1. **Natural Language Processing (NLP)**
   * Natural language processing (NLP) is a field of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language.
   * It involves developing algorithms and models that can analyze text and speech, extract meaning, and perform tasks such as translation, summarization, and question answering.
   * NLP is used in various applications, including chatbots, machine translation, sentiment analysis, and text classification.
2. **One-Hot Encoding**
   * In machine learning, one-hot encoding is a frequently used method to deal with categorical data.
   * Because many machine learning models need their input variables to be numeric, categorical variables need to be transformed in the pre-processing part.
   * One-hot encoding creates new binary columns for each unique value in a categorical variable, indicating the presence or absence of that value for each observation.

Natural Language Processing (NLP) is a branch of artificial intelligence that deals with enabling computers to understand, interpret, and generate human language.

1. **Word Tokens**
   * Tokens are the total number of words in a corpus regardless of whether they are repeated.
   * Word tokenization splits text into words.
   * Tokens are the basic building blocks of natural language understanding.
2. **CountVectorizer**
   * CountVectorizer is a tool used in NLP to convert a collection of text documents into numerical feature vectors.
   * It works by counting the occurrences of each word in each document and creating a matrix representation of the data.
   * This matrix can then be used as input for machine learning algorithms.
3. **Bag of Words**
   * The bag-of-words model is a simplifying representation used in NLP and information retrieval (IR).
   * In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.
   * Keeping multiplicity means that the number of times a word appears in the text is recorded.
4. **Stemming**
   * Stemming is the process of reducing a word to its base or root form, called the stem.
   * This is done by removing suffixes and prefixes from the word.
   * Stemming is used in NLP to normalize words and reduce the number of unique words in a corpus.
5. **Lemmatization**
   * Lemmatization is the process of grouping together the inflected forms of a word so they can be analyzed as a single item.
   * This is done by using a dictionary or a set of rules to identify the base form of a word, called the lemma.
   * Lemmatization is used in NLP to improve the accuracy of text analysis.
6. **TF-IDF (Term Frequency Inverse Document Frequency)**
   * TF-IDF is a numerical statistic that reflects how important a word is to a document in a collection or corpus.
   * It is used to score words in the context of the document as well as in the context of the corpus.
   * The higher the TF-IDF score, the more useful the word is in the analysis of the document.
7. **Stop Words**
   * Stop words are commonly used words (such as "the", "a", "is", "to") that are often removed from text during NLP tasks.
   * This is because stop words typically do not carry much meaning and can clutter the analysis.
   * Removing stop words can improve the efficiency and accuracy of NLP tasks.
8. **spaCy**
   * spaCy is an open-source library for advanced NLP tasks.
   * It provides tools for various NLP tasks, such as tokenization, part-of-speech tagging, named entity recognition, and dependency parsing.
   * It is designed to be fast, efficient, and easy to use.
9. **Language Models**
   * A language model is a probability distribution over sequences of words.
   * It assigns a probability to the whole sequence.
   * Language models are used in various NLP tasks, such as machine translation, speech recognition, and text generation.
10. **N-grams**
    * N-grams are contiguous sequences of n items from a given sample of text or speech.
    * They are used to model the probability of a sequence of words, which is useful in various NLP tasks.
    * Examples of n-grams include unigrams (single words), bigrams (two words), and trigrams (three words).
11. **Markov Chains**
    * A Markov chain is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states.
    * It is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
    * Markov chains are used in NLP to model the probability of a sequence of words, which is useful in various NLP tasks.
12. **Word Embeddings**
    * In natural language processing (NLP), a word embedding is a learned representation for text where words that have the same meaning have a similar representation.
    * Word embeddings are used to capture the semantic relationships between words.
    * They are used in various NLP tasks, such as machine translation, sentiment analysis, and text classification.
13. **Similarity Measures**
    * Similarity measures are used to quantify the similarity between two pieces of text.
    * There are various similarity measures, such as cosine similarity, Jaccard distance, and Euclidean distance.
    * Similarity measures are used in NLP tasks such as information retrieval, document clustering, and text classification.
14. **Part of Speech (POS)**
    * Part-of-speech (POS) tagging is the process of assigning a part-of-speech tag to each word in a text.
    * POS tags indicate the grammatical role of a word in a sentence, such as noun, verb, adjective, or adverb.
    * POS tagging is used in various NLP tasks, such as parsing, machine translation, and information retrieval.
15. **Named Entity Recognition (NER)**
    * Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
    * NER is used in various NLP applications, such as information retrieval, question answering, and machine translation.
16. **Sentence Segmentation**
    * Sentence segmentation, also known as sentence boundary disambiguation, is the problem in natural language processing of deciding where sentences begin and end.
    * Sentence segmentation is used to divide a text into individual sentences, which is a necessary step for many NLP tasks.
17. **Topic Modeling**
    * Topic modeling is a statistical method for discovering abstract "topics" that occur in a collection of documents.
    * It is used to uncover hidden themes or topics in a corpus of text data.
    * Topic modeling is used in various NLP applications, such as document summarization, information retrieval, and text classification.
18. **Latent Dirichlet Allocation (LDA)**
    * Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
19. **Non-Negative Matrix Factorization**
    * Non-Negative Matrix Factorization is a statistical method that helps us to reduce the dimension of the input corpora or corpora. Internally, it uses the factor analysis method to give comparatively less weightage to the words that are having less coherence.
    * <https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/>
    * Performs dimensionality reduction and clustering
    * Used with TF-IDF
20. **Similarity Measures**
    * <https://flavien-vidal.medium.com/similarity-distances-for-natural-language-processing-16f63cd5ba55>
    * Longest Common Substring
      1. India and Indiana would return 5
    * Levenshtein Edit Distance
      1. Finds the minimum number of single-character edits such as replacement, deletion, and insertion, needed to convert 1 text into another
      2. India and Indiana would return 2
    * Hamming Distance
      1. Finds the number replacements needed to change one text into another of equal size
      2. Indians and Indiana returns 2
    * Jaccard Distance
      1. Finds how disimilar two words are by distance and the lower the distance, the more similar
    * Euclidean Distance
      1. Finds the length between two points
    * l2 norm
    * Dot Product
      1. Considers orientation, the direction, that Euclidean lacks
      2. Uses magnitude with orientation
    * Cosine Similarity
      1. In data analysis, cosine similarity is a measure of similarity between two non-zero vectors defined in an inner product space. Cosine similarity is the cosine of the angle between the vectors; that is, it is the dot product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle. The cosine similarity always belongs to the interval -1, 1. For example, two proportional vectors have a cosine similarity of 1, two orthogonal vectors have a similarity of 0, and two opposite vectors have a similarity of -1. For example, in information retrieval and text mining, each word is assigned a different coordinate and a document is represented by the vector of the numbers of occurrences of each word in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be, in terms of their subject matter, and independently of the length of the documents. The technique is also used to measure cohesion within clusters in the field of data mining.
      2. https://en.wikipedia.org/wiki/Cosine\_similarity