

What do you understand by Natural Language Processing?

It is used to create automated software that helps understand human spoken languages to extract useful information from the data it gets in the form of text or audio. Techniques in NLP are used to interpret data in the form of natural languages.

List any two real-life applications of Natural Language Processing.

Two real-life applications of Natural Language Processing are as follows:

1. **Google Translate:** Google Translate is one of the famous applications of Natural Language Processing. It helps convert written or spoken sentences into any language.

Chatbots: To provide a better customer support service, companies have started using chatbots for 24/7 service. **AI Chatbots** help resolve the basic queries of customers.

Define the terminology in NLP.

Tokenization

The processes of splitting sentence into its constitute words.

Usually of three types

Unigram,Bigram and Trigram

PoS Tagging

Parts of speech tagging,process of tagging words within sentence into their respective PoS and labelling.

Text normalization

There are words in sentence, they are spelled differently but meaning is same.

example Gurgaon, Gurugram

Stemming

If words ending with ed remove ed if words ending with ing remove ing

Lemmatization

additional check is done by looking through the dictionary to extract base form of a word. **The method of mapping all the various forms of a word to its base word (also called “lemma”) is known as Lemmatization.**

Named Entity Recognition

Typically there are words in sentence which are not in dictionary,we need to treat them separately called

Word sense disambiguation

A words meaning depends on its association with other words in the sentence, disambiguation is process of mapping words to correct sense

Sentence boundary detection

Method of detecting where the one sentence ends and where the another sentence begins

Types of tokenization

MWE tokenizer: multi word expression tokenizer

Certain group of multiple words are treated as one entity during tokenization.

Regular expression tokenizer:

Sentences are split based on occurrence of pattern

White space tokenizer: split a string whenever there is space.

Bag of words

Process of converting unstructured data into structured data.

What are stop words?

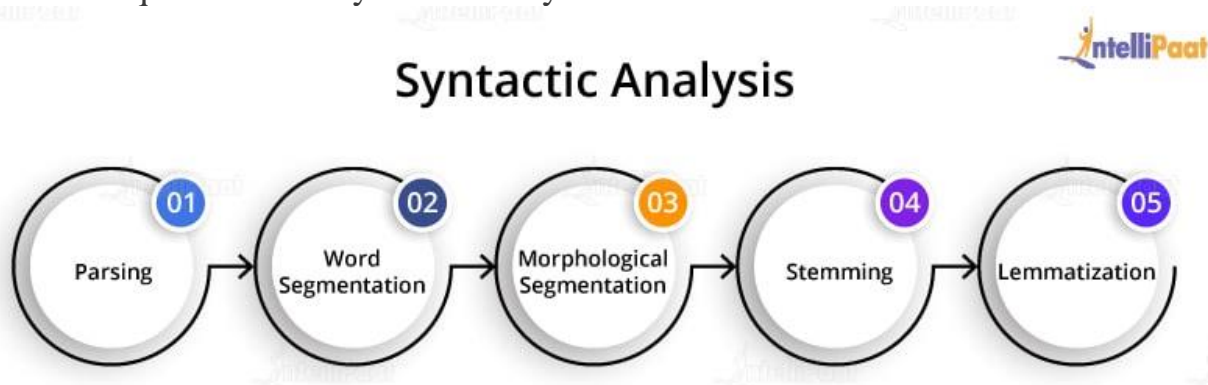
Stop words are said to be useless data for a search engine. Words such as articles, prepositions, etc. are considered as stop words. There are stop words such as was, were, is, am, the, a, an, how, why, and many more.

What is NLTK?

NLTK is a Python library, which stands for Natural Language Toolkit. We use NLTK to process data in human spoken languages. NLTK allows us to apply techniques such as parsing, tokenization, lemmatization, stemming, and more to understand natural languages.

What is Syntactic Analysis?

Syntactic analysis is a technique of analyzing sentences to extract meaning from it. **Using syntactic analysis, a machine can analyze and understand the order of words arranged in a sentence.** NLP employs grammar rules of a language that helps in the syntactic analysis of the combination and order of words in documents. The techniques used for syntactic analysis are as follows:



1. **Parsing:** It helps in deciding the structure of a sentence or text in a document. It helps analyze the words in the text based on the grammar of the language.

2. **Word segmentation:** The segmentation of words segregates the text into small significant units.
3. **Morphological segmentation:** The purpose of morphological segmentation is to break words into their base form.
4. **Stemming:** It is the process of removing the suffix from a word to obtain its root word.
5. **Lemmatization:** It helps combine words using suffixes, without altering the meaning of the word.

What is Semantic Analysis?

Semantic analysis helps make a machine understand the meaning of a text. It uses various algorithms for the interpretation of words in sentences. It also helps understand the structure of a sentence.

1. **Named entity recognition:** This is the process of information retrieval that helps identify entities such as the name of a person, organization, place, time, emotion, etc.
2. **Word sense disambiguation:** It helps identify the sense of a word used in different sentences.
3. **Natural language generation:** It is a process used by the software to convert the structured data into human spoken languages. By using NLG, organizations can automate content for custom reports.

List the components of Natural Language Processing.

- **Entity extraction:** Entity extraction refers to the retrieval of information such as place, person, organization, etc. by the segmentation of a sentence. It helps in the recognition of an entity in a text.
- **Syntactic analysis:** Syntactic analysis helps draw the specific meaning of a text.
- **Pragmatic analysis:** To find useful information from a text, we implement pragmatic analysis techniques.
- **Morphological and lexical analysis:** It helps in explaining the structure of words by analyzing them through parsing.

What is Latent Semantic Indexing (LSI)?

Latent semantic indexing is a mathematical technique used to improve the accuracy of the information retrieval process. The design of LSI algorithms allows machines to detect the hidden (latent) correlation between semantics (words).

What are Regular Expressions?

A regular expression is used to match and tag words. It consists of a series of characters for matching strings.

What is Regular Grammar?

Regular grammar is used to represent a regular language.

A regular grammar comprises rules in the form of $A \rightarrow a$, $A \rightarrow aB$, and many more. The rules help detect and analyze strings by automated computation.

What is Parsing in the context of NLP?

Parsing in NLP refers to the understanding of a sentence and its grammatical structure by a machine. Parsing allows the machine to understand the meaning of a word in a sentence and the grouping of words, phrases, nouns, subjects, and objects in a sentence. Parsing helps analyze the text or the document to extract useful insights from it. To understand parsing, refer to the below diagram:

What is TF-IDF?

TFIDF or Term Frequency-Inverse Document Frequency indicates the importance of a word in a set. It helps in information retrieval with numerical statistics. For a specific document, TF-IDF shows a frequency that helps identify the keywords in a document. The major use of TF-IDF in NLP is the extraction of useful information from crucial documents by statistical data. It is ideally used to classify and summarize the text in documents and filter out stop words.

TF helps calculate the ratio of the frequency of a term in a document and the total number of terms. Whereas, **IDF** denotes the importance of the term in a document.

The formula for calculating TF-IDF:

TF(W) = (Frequency of W in a document)/(The total number of terms in the document)

IDF(W) = \log_e (The total number of documents/The number of documents having the term W)

Explain Dependency Parsing in NLP.

Dependency parsing helps assign a syntactic structure to a sentence. Therefore, it is also called syntactic parsing. Dependency parsing is one of the critical tasks in NLP. It allows the analysis of a sentence using parsing algorithms. Also, by **using the parse tree in dependency parsing, we can check the grammar and analyze the semantic structure of a sentence.**

Explain the concept of Feature Engineering.

After a variety of pre-processing procedures and their applications, we need a way to input the pre-processed text into an NLP algorithm later when we employ ML methods to complete our modelling step. The set of strategies that will

achieve this goal is referred to as feature engineering. Feature extraction is another name for it. The purpose of feature engineering is to convert the text's qualities into a numeric vector that NLP algorithms can understand. This stage is called "text representation"

What is an ensemble method in NLP?

An ensemble approach is a methodology that derives an output or makes predictions by combining numerous independent similar or distinct models/weak learners

What do you mean by a Bag of Words (BOW)?

The **Bag of Words** model is a popular one that uses word frequency or occurrences to train a classifier. This methodology generates a matrix of occurrences for documents or phrases, regardless of their grammatical structure or word order.

A bag-of-words is a text representation that describes the frequency with which words appear in a document. It entails two steps:

- A list of terms that are well-known.
- A metric for determining the existence of well-known terms.

What is Latent Semantic Indexing (LSI) in NLP?

Latent Semantic Indexing (LSI), also known as Latent Semantic Analysis, is a mathematical method for improving the accuracy of information retrieval. It aids in the discovery of hidden(latent) relationships between words (semantics) by generating a set of various concepts associated with the terms of a phrase in order to increase information

What is the difference between NLP and NLU?

Natural Language Processing (NLP)	Natural Language Understanding (NLU)
NLP is a system that manages end-to-end conversations between computers and people at the same time.	NLU aids in the solving of Artificial Intelligence's complex problems.
Humans and machines are both	NLU allows machines to interpret

Natural Language Processing (NLP)	Natural Language Understanding (NLU)
involved in NLP.	unstructured inputs by transforming them into structured text.
NLP focuses on interpreting language in its most literal sense, such as what was said.	NLU, on the other hand, concentrates on extracting context and meaning, or what was meant.
NLP can parse text-based on grammar, structure, typography, and point of view.	It'll be NLU that helps the machine deduce the meaning behind the language content.

What are some metrics on which NLP models are evaluated?

The following are some metrics on which NLP models are evaluated:

- **Accuracy:** When the output variable is categorical or discrete, accuracy is used. It is the percentage of correct predictions made by the model compared to the total number of predictions made.
- **Precision:** Indicates how precise or exact the model's predictions are, i.e., how many positive (the class we care about) examples can the model correctly identify given all of them?
- **Recall:** Precision and recall are complementary. It measures how effectively the model can recall the positive class, i.e., how many of the positive predictions it generates are correct.
- **F1 score:** This metric combines precision and recall into a single metric that also represents the trade-off between accuracy and recall, i.e., completeness and exactness.

$$(2 \text{ Precision Recall}) / (\text{Precision} + \text{Recall})$$
is the formula for F1.
- **AUC:** As the prediction threshold is changed, the AUC captures the number of correct positive predictions versus the number of incorrect positive predictions

10. What are the best NLP Tools?

Some of the best NLP tools from open sources are:

- SpaCy
- TextBlob

- Textacy
- Natural language Toolkit ([NLTK](#))
- Retext
- NLP.js
- Stanford NLP
- CogcompNLP

13. Which of the following techniques can be used for keyword normalization in NLP, the process of converting a keyword into its base form?

- Lemmatization
- Soundex
- Cosine Similarity
- N-grams

Answer: a)

Lemmatization helps to get to the base form of a word, e.g. are playing -> play, eating -> eat, etc..

Other options are meant for different purposes.

Which of the following techniques can be used to compute the distance between two word vectors in NLP?

- Lemmatization
- Euclidean distance
- Cosine Similarity
- N-grams

Answer: b) and c)

Distance between two word vectors can be computed using Cosine similarity and Euclidean Distance. **Cosine Similarity** establishes a cosine angle between the vector of two words. A cosine angle close to each other between two word vectors indicates the words are similar and vice a versa.

E.g. cosine angle between two words “Football” and “Cricket” will be closer to 1 as compared to angle between the words “Football” and “New Delhi”

What are the possible features of a text corpus in NLP?

- Count of the word in a document
- Vector notation of the word

- c. Part of Speech Tag
- d. Basic Dependency Grammar
- e. All of the above

Answer: e)

All of the above can be used as features of the text corpus.

You created a document term matrix on the input data of 20K documents for a Machine learning model. Which of the following can be used to reduce the dimensions of data?

- 1. Keyword Normalization
 - 2. Latent Semantic Indexing
 - 3. Latent Dirichlet Allocation
- a. only 1
 - b. 2, 3
 - c. 1, 3
 - d. 1, 2, 3

Answer: d)

Which of the text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection, and object detection in NLP.

- a. Part of speech tagging
- b. Skip Gram and N-Gram extraction
- c. Continuous Bag of Words
- d. Dependency Parsing and Constituency Parsing

Answer: d)

Dissimilarity between words expressed using cosine similarity will have values significantly higher than 0.5

- a. True
- b. False

Ans: a)

Which one of the following are keyword Normalization techniques in NLP

- a. Stemming
- b. Part of Speech
- c. Named entity recognition
- d. Lemmatization

Answer: a) and d)

Part of Speech (POS) and Named Entity Recognition(NER) are not keyword Normalization techniques. Named Entity help you extract Organization, Time, Date, City, etc..type of entities from the given sentence, whereas Part of Speech helps you extract Noun, Verb, Pronoun, adjective, etc..from the given sentence tokens.

Which of the below are NLP use cases?

- a. Detecting objects from an image
- b. Facial Recognition
- c. Speech Biometric
- d. Text Summarization

Ans: d)

a) And b) are Computer Vision use cases, and c) is Speech use case.
Only d) Text Summarization is an NLP use case.

In a corpus of N documents, one randomly chosen document contains a total of T terms and the term “hello” appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term “hello” appears in approximately one-third of the total documents?

- a. $KT * \log(3)$
- b. $T * \log(3) / K$
- c. $K * \log(3) / T$
- d. $\log(3) / KT$

Answer: (c)

formula for TF is K/T

formula for IDF is $\log(\text{total docs} / \text{no of docs containing "data"})$

$= \log(1 / (1/3))$

$= \log(3)$

Hence correct choice is $K\log(3)/T$

In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

- a. Term Frequency (TF)
- b. Inverse Document Frequency (IDF)
- c. Word2Vec
- d. Latent Dirichlet Allocation (LDA)

Ans: b)

In NLP, The process of removing words like “and”, “is”, “a”, “an”, “the” from a sentence is called as

- a. Stemming
- b. Lemmatization
- c. Stop word
- d. All of the above

Ans: c)

In Lemmatization, all the stop words such as a, an, the, etc.. are removed. One can also define custom stop words for removal.

In NLP, The process of converting a sentence or paragraph into tokens is referred to as Stemming

- a. True
- b. False

Ans: b)

The statement describes the process of tokenization and not stemming, hence it is False.

In NLP, Tokens are converted into numbers before giving to any Neural Network

- a. True
- b. False

Ans: a)

In NLP, all words are converted into a number before feeding to a Neural Network.

identify the odd one out

- a. nltk
- b. scikit learn
- c. SpaCy
- d. BERT

Ans: d)

All the ones mentioned are NLP libraries except BERT, which is a word embedding

TF-IDF helps you to establish?

- a. most frequently occurring word in the document
- b. most important word in the document

Ans: b)

TF-IDF helps to establish how important a particular word is in the context of the document corpus. TF-IDF takes into account the number of times the word appears in the document and offset by the number of documents that appear in the corpus.

Which one of the following is not a pre-processing technique in NLP

- a. Stemming and Lemmatization
- b. converting to lowercase
- c. removing punctuations
- d. removal of stop words
- e. Sentiment analysis

Ans: e)

Sentiment Analysis is not a pre-processing technique. It is done after pre-processing and is an NLP use case. All other listed ones are used as part of statement pre-processing.

In text mining, converting text into tokens and then converting them into an integer or floating-point vectors can be done using

- a. CountVectorizer
- b. TF-IDF
- c. Bag of Words
- d. NERs

Ans: a)

CountVectorizer helps do the above, while others are not applicable.
text = ["Rahul is an avid writer, he enjoys studying understanding and presenting. He loves to play"]
vectorizer = CountVectorizer()
vectorizer.fit(text)
vector = vectorizer.transform(text)
print(vector.toarray())

n NLP, Words represented as vectors are called as Neural Word Embeddings

- a. True
- b. False

Ans: a)

Word2Vec, GloVe based models build word embedding vectors that are multidimensional.

In NLP, Context modeling is supported with which one of the following word embeddings

1. a. Word2Vec
2. b) GloVe
3. c) BERT
4. d) All of the above

Ans: c)

Only BERT (Bidirectional Encoder Representations from Transformer) supports context modelling where the previous and next sentence context is taken into consideration. In Word2Vec, GloVe only word embeddings are considered and previous and next sentence context is not considered.

What is a *Probability Distribution*?

A probability distribution is a statistical function that describes all the possible values and probabilities for a random variable within a given range. This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution will be determined by a number of factors. The mean (average), standard deviation, skewness, and kurtosis of the distribution are among these factors. There are two main types of probability distribution:

- **Discrete probability distributions:** used for random variables with discrete outcomes, for example, the number of heads in five consecutive coin tosses, the number of rainy days in a given week, the number of goals scored by a player, and so on.
- **Continuous probability distributions:** used for random variables with continuous outcomes, for example, the height of male students, median house prices in San Francisco, claim amounts experienced by an insurance company, and so on.

Q4:

What is the difference between a *Combination* and a *Permutation*?

Junior

Answer

- A **Combination** is the choice of r elements from a set of n elements *without replacement* and where *order does not matter*. Is most used to group data. For example, picking three team members from a group, picking two colors from a color brochure, etc. It is mathematically defined as:

$$C_r^n = \frac{n!}{(n-r)!r!} \quad C_{rn} = (n-r)r!n!$$

- A **Permutation** is the choice of r elements from a set of n elements *without replacement* and where the *order matters*. Is used to list data, for example picking first, second and third place winners, picking two favorite colors -in order- from a color brochure, etc. It is mathematically defined as:

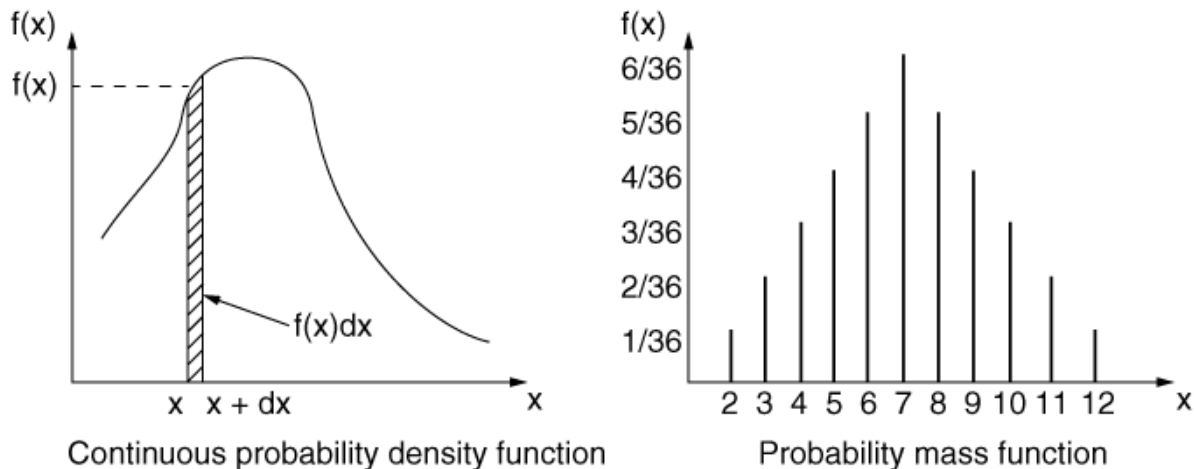
$$P_{\{n,r\}} = \frac{n!}{(n-r)!} \quad P_{n,r} = (n-r)!n!$$

What's the difference between *Probability Mass Functions* and *Density Probability Functions*?

Junior

Answer

- **Probability mass functions** are used to describe *discrete probability* distributions and allow us to determine the probability of an observation being *exactly equal* to a target value.
- **Density functions** are used to describe *continuous probability* distributions and allows us to determine the probability of an observation being within a *range around our target value* by computing the *area under the curve* for our interval.



How do you transform a *Skewed Distribution* into a *Normal Distribution*?

Mid

Probability 34

Answer

To transform a **Skewed Distribution** into a **Normal Distribution** we apply some **linearized function** on it. Some common functions that achieve this goal are:

- **Logarithmic function:** We can use it to make *extremely* skewed distributions less skewed, especially for *right-skewed distributions*. The only condition is that this function is defined only for **strictly positive numbers**.

$$f(x) = \ln(x)$$
- **Square root transformation:** this one has an average effect on distribution shape: it's weaker than *logarithmic transformation*, and it's also used for reducing *right-skewed distributions*, but is defined only for **positive numbers**.

$$f(x) = \sqrt{x}$$
- **Reciprocal transformation:** this one reverses the order among values of the same sign, so *large values* become *smaller*, but the *negative reciprocal* preserves the order among values of the same sign. The only condition is that this function is not defined for **zero values**.

$$f(x) = \frac{1}{x}$$

- **Exponential or Power transformation:** has a reasonable effect on distribution shape; generally, we apply power transformation (power of two usually) to reduce *left skewness*. We could also try any exponent to see which one provides better results. $f(x) = x^n$
- **Box-Cox Transformation:** in this transformation, we're searching and evaluating all the other transformations and *choosing the best one*. It's defined as:

What's the difference between *Cumulative Distribution Functions* and *Probability Density Functions*?

Mid

Probability 34

Answer

- The **Cumulative Distribution Function (CDF)** can be defined for any kind of random variable, i.e. *discrete* or *continuous*, and it tells us the probability that the random variable **X** takes a value less than or equal to a particular value **x**:

$$F(x) = \Pr[X \leq x]$$

The CDF is used to determine the probability that an observation will be greater than a certain value, or between two values.

- A **Probability Density Function (PDF)** can be defined only for *continuous random variables* and it tells us the probability of the random variable **X** falling within a range of values (**a, b**) by computing the integral of this variable's PDF over that range:

$$F(x) = \Pr[a \leq X \leq b] = \int_a^b f_X(x) dx$$

When is an event *Independent of Itself*?

Mid

Answer

An event can only be *independent of itself* when either there is *no chance of it happening* or when it is *certain to happen*. To demonstrate this, let's remember that two events **A** and **B** are independent if:

$$P(A \cap B) = P(A)P(B) \quad P(A \cap B) = P(A)P(B)$$

Facebook - Easy] There is a fair coin (one side heads, one side tails) and an unfair coin (both sides tails). You pick one at random, flip it 5 times, and observe that it comes up as tails all five times. What is the chance that you are flipping the unfair coin?

We can use Bayes Theorem here. Let U denote the case where we are flipping the unfair coin and F denote the case where we are flipping a fair coin. Since the coin is chosen randomly, we know that $P(U) = P(F) = 0.5$. Let $5T$ denote the event where we flip 5 heads in a row. Then we are interested in solving for $P(U|5T)$, i.e., the probability that we are flipping the unfair coin, given that we saw 5 tails in a row.

We know $P(5T|U) = 1$ since by definition the unfair coin will always result in tails. Additionally, we know that $P(5T|F) = 1/2^5 = 1/32$ by definition of a fair coin. By Bayes Theorem we have:

$$P(U|5T) = \frac{P(5T|U) * P(U)}{P(5T|U) * P(U) + P(5T|F) * P(F)} = \frac{0.5}{0.5 + 0.5 * 1/32} = 0.97$$

Therefore the probability we picked the unfair coin is about 97%.

2[Two Sigma - Easy] Say you are running a multiple linear regression and believe there are several predictors that are correlated. How will the results of the regression be affected if they are indeed correlated? How would you deal with this problem?

Problem #2 Solution:

There will be two main problems. The first is that the coefficient estimates and signs will vary dramatically, depending on what particular variables you include in the model. In particular, certain coefficients may even have confidence intervals that include 0 (meaning it is difficult to tell whether an increase in that X value is associated with an increase or decrease in Y). The

second is that the resulting p-values will be misleading - an important variable might have a high p-value and deemed insignificant even though it is actually important.

You can deal with this problem by either removing or combining the correlated predictors. In removing the predictors, it is best to understand the causes of the correlation (i.e. did you include extraneous predictors or such as both X and $2X$). For combining predictors, it is possible to include interaction terms (the product of the two). Lastly, you should also 1) center data, and 2) try to obtain a larger sample size (which will lead to narrower confidence intervals).

1. [**Uber - Easy**] Describe p-values in layman's terms.
2. [**Facebook - Easy**] How would you build and test a metric to compare two user's ranked lists of movie/tv show preferences?
3. [**Microsoft - Easy**] Explain the statistical background behind power.
4. [**Twitter - Easy**] Describe A/B testing. What are some common pitfalls?
5. [**Google - Medium**] How would you derive a confidence interval from a series of coin tosses?
6. [**Stripe - Medium**] Say you model the lifetime for a set of customers using an exponential distribution with parameter λ , and you have the lifetime history (in months) of n customers. What is your best guess for λ ?
7. [**Lyft - Medium**] Derive the mean and variance of the uniform distribution $U(a, b)$.
8. [**Google - Medium**] Say we have $X \sim \text{Uniform}(0, 1)$ and $Y \sim \text{Uniform}(0, 1)$. What is the expected value of the minimum of X and Y ?
9. [**Spotify - Medium**] You sample from a uniform distribution $[0, d]$ n times. What is your best estimate of d ?
10. [**Quora - Medium**] You are drawing from a normally distributed random variable $X \sim N(0, 1)$ once a day. What is the approximate expected number of days until you get a value of more than 2?
11. [**Facebook - Medium**] Derive the expectation for a geometric distributed random variable.
12. [**Google - Medium**] A coin was flipped 1000 times, and 550 times it showed up heads. Do you think the coin is biased? Why or why not?

- 13.[**Robinhood - Medium**] Say you have n integers $1 \dots n$ and take a random permutation. For any integers i, j let a swap be defined as when the integer i is in the j th position, and vice versa. What is the expected value of the total number of swaps?
- 14.[**Uber - Hard**] What is the difference between MLE and MAP? Describe it mathematically.
- 15.[**Google - Hard**] Say you have two subsets of a dataset for which you know their means and standard deviations. How do you calculate the blended mean and standard deviation of the total dataset? Can you extend it to K subsets?
- 16.[**Lyft - Hard**] How do you randomly sample a point uniformly from a circle with radius 1?
- 17.[**Two Sigma - Hard**] Say you continually sample from some i.i.d. uniformly distributed $(0, 1)$ random variables until the sum of the variables exceeds 1. How many times do you expect to sample?
- 18.[**Uber - Hard**] Given a random Bernoulli trial generator, how do you return a value sampled from a normal distribution
- 19.**Problem #5 Solution:**
- 20.
- 21.

#solution of 5:

By definition, a chord is a line segment whereby the two endpoints lie on the circle. Therefore, two arbitrary chords can always be represented by any four points chosen on the circle. If you choose to represent the first chord by two of the four points then you have:

$$\binom{4}{2} = 6$$

choices of choosing the two points to represent chord 1 (and hence the other two will represent chord 2). However, note that in this counting, we are duplicating the count of each chord twice since a chord with endpoints p_1 and p_2 is the same as a chord with endpoints p_2 and p_1 . Therefore the proper number of valid chords is:

$$\frac{1}{2} \binom{4}{2} = 3$$

Among these three configurations, only exactly one of the chords will intersect, hence the desired probability is:

$$p = \frac{1}{3}$$

Problem #13 Solution:

Let X be the number of coin flips needed until two heads. Then we want to solve for $E[X]$. Let H denote a flip that resulted in heads, and T denote a flip that resulted in tails. Note that $E[X]$ can be written in terms of $E[X|H]$ and $E[X|T]$, i.e. the expected number of flips needed, conditioned on a flip being either heads or tails respectively.

Conditioning on the first flip, we have:

$$E[X] = \frac{1}{2}(1 + E[X|H]) + \frac{1}{2}(1 + E[X|T])$$

Note that $E[X|T] = E[X]$ since if a tail is flipped, we need to start over in getting two heads in a row.

To solve for $E[X|H]$, we can condition it further on the next outcome: either heads (HH) or tails (HT).

Therefore, we have:

$$E[X|H] = \frac{1}{2}(1 + E[X|HH]) + \frac{1}{2}(1 + E[X|HT])$$

Note that if the result is HH, then $E[X|HH] = 0$ since the outcome was achieved, and that $E[X|HT] = E[X]$ since a tail was flipped, we need to start over again, so:

$$E[X|H] = \frac{1}{2}(1 + 0) + \frac{1}{2}(1 + E[X]) = 1 + \frac{1}{2}E[X]$$

Plugging this into the original equation yields $E[X] = 6$ coin flips

Problem #15 Solution:

Consider the first n coins that A flips, versus the n coins that B flips.

There are three possible scenarios:

1. A has more heads than B
2. A and B have an equal amount of heads
3. A has less heads than B

Notice that in scenario 1, A will always win (irrespective of coin $n+1$), and in scenario 3, A will always lose (irrespective of coin $n+1$). By symmetry, these two scenarios have an equal probability of occurring.

Denote the probability of either scenario as x , and the probability of scenario 2 as y .

We know that $2x + y = 1$ since these 3 scenarios are the only possible outcomes. Now let's consider coin $n+1$. If the flip results in heads, with probability 0.5, then A will have won after scenario 2 (which happens with probability y). Therefore, A's total chances of winning the game are increased by $0.5y$.

Thus, the probability that A will win the game is:

$$x + \frac{1}{2}y = x + \frac{1}{2}(1 - 2x) = \frac{1}{2}$$

Problem #18 Solution:

Let B be the event that all n rolls have a value less than or equal to r . Then we have:

$$P(B_r) = \frac{r^n}{6^n}$$

since all n rolls must have a value less than or equal to r . Let A be the event that the largest number is r . We have:

$$B_r = B_{r-1} \cup A_r$$

and since the two events on the right hand side are disjoint, we have:

$$P(B_r) = P(B_{r-1}) + P(A_r)$$

Therefore, the probability of A is given by:

$$P(A_r) = P(B_r) - P(B_{r-1}) = \frac{r^n}{6^n} - \frac{(r-1)^n}{6^n}$$

Problem #9 Solution:

For $X \sim U(a, b)$ we have the following:

$$f_X(x) = \frac{1}{b-a}$$

AutoSave NLP, TM, Probability Distribution, CIT, Cldock - C...

Search (Alt+Q)

Dr. Radhakrishna Naik

File Home Insert Draw Design Layout References Mailings Review View Help

Top 30 NLP Interview Questions x 40 Probability & Statistics Data S

nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street

Comments Share

Undo Clipbo

Reuse Files

Use Files

Nick Singh

Previously at data startup SafeGraph, and Software Engineer on Facebook's Growth Team. Author of [Ace the Data Science Interview](#).

[Join the 46,000 readers who subscribe to my tech career newsletter!](#)

I send an email just once a month with guides on Tech Careers, Data Science, & Startups, as well as a few links to interesting articles & books on careers and technology.

BLOG

ACE THE DATA SCIENCE INTERVIEW

MASSAPPLY - COLD EMAIL FOR TECH JOBS

14 BOOKS THAT CHANGED MY LIFE

ABOUT ME

in tw ig em

you should also 1) center data, and 2) try to obtain a larger sample size (which will lead to narrower confidence intervals).

Problem #9 Solution:

For $X \sim U(a, b)$ we have the following:

$$f_X(x) = \frac{1}{b-a}$$

Therefore we can calculate the mean as:

$$E[X] = \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{a+b}{2}$$

Similarly for variance we want:

$$\text{Var}(X) = E[X^2] - E[X]^2$$

And we have:

$$E[X^2] = \int_a^b x^2 f_X(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{x^3}{3(b-a)} \Big|_a^b = \frac{a^2+ab+b^2}{3}$$

Therefore:

$$\text{Var}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Problem #12 Solution:

Page 23 of 23 4918 words Text Predictions: On Accessibility: Unavailable 5:00 AM 6/8/2022

AutoSave NLP, TM, Probability Distribution, CIT, Cldock - C...

Search (Alt+Q)

Dr. Radhakrishna Naik

File Home Insert Draw Design Layout References Mailings Review View Help

Top 30 NLP Interview Questions x 40 Probability & Statistics Data S

nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street

Comments Share

Undo Clipbo

Reuse Files

Use Files

Nick Singh

Previously at data startup SafeGraph, and Software Engineer on Facebook's Growth Team. Author of [Ace the Data Science Interview](#).

[Join the 46,000 readers who subscribe to my tech career newsletter!](#)

I send an email just once a month with guides on Tech Careers, Data Science, & Startups, as well as a few links to interesting articles & books on careers and technology.

BLOG

ACE THE DATA SCIENCE INTERVIEW

MASSAPPLY - COLD EMAIL FOR TECH JOBS

14 BOOKS THAT CHANGED MY LIFE

ABOUT ME

in tw ig em

Therefore:

$$\text{Var}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Problem #12 Solution:

Since X is normally distributed, we can look at the cumulative distribution function (CDF) of the normal distribution:

$$\Phi(x) = P(X \leq x)$$

To check the probability X is at least 2, we can check (knowing that X is distributed as standard normal):

$$\Phi(2) = P(X \leq 2) = P(X \leq \mu + 2\sigma) = 0.977$$

Therefore $P(X > 2) = 1 - 0.977 = 0.023$ for any given day. Since the draws are independent each day, then the expected time until drawing an $X > 2$ follows a geometric distribution, with $p = 0.023$. Let T be a random variable denoting the number of days, then we have:

$$E[T] = \frac{1}{p} = \frac{1}{.024} \approx 43 \text{ days}$$

Problem #14 Solution:

Because the sample size of flips is large (1000), we can apply the Central Limit Theorem. Since each individual flip is a Bernoulli random variable, we can assume it has a probability of showing up heads as p . Then we want to test whether p is 0.5 (i.e. whether it is fair). The Central Limit Theorem allows us to approximate the total number of

Page 23 of 23 4918 words Text Predictions: On Accessibility: Unavailable 5:00 AM 6/8/2022

AutoSave: On NLP, TM, Probability Distribution, C11, C12.docx - C:\...

Search (Alt+Q)

Dr. Radhakrishna Naik

File Home Insert Draw Design Layout References Mailings Review View Help

Top 30 NLP Interview Questions x 40 Probability & Statistics Data x

nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street

Nick Singh

Previously at data startup SafeGraph, and Software Engineer on Facebook's Growth Team. Author of [Ace the Data Science Interview](#).

[Join the 46,000 readers who subscribe to my tech career newsletter!](#)

I send an email just once a month with guides on Tech Careers, Data Science, & Startups, as well as a few links to interesting articles & books on careers and technology.

BLOG
ACE THE DATA SCIENCE INTERVIEW
MASSAPPLY - COLD EMAIL FOR TECH JOBS
14 BOOKS THAT CHANGED MY LIFE
ABOUT ME

in tw ig ex

expected time until drawing an $X > z$ follows a geometric distribution, with $p = 0.025$. Let T be a random variable denoting the number of days, then we have:

$$E[T] = \frac{1}{p} = \frac{1}{.025} \approx 43 \text{ days}$$

Problem #14 Solution:

Because the sample size of flips is large (1000), we can apply the Central Limit Theorem. Since each individual flip is a Bernoulli random variable, we can assume it has a probability of showing up heads as p . Then we want to test whether p is 0.5 (i.e. whether it is fair). The Central Limit Theorem allows us to approximate the total number of heads seen as being normally distributed.

More specifically, the number of heads seen should follow a Binomial distribution since it a sum of Bernoulli random variables. If the coin is not biased ($p = 0.5$), then we have the following on the expected number of heads:

$$\mu = np = 1000 * 0.5 = 500$$

and the variance is given by:

$$\sigma^2 = np(1-p) = 1000 * 0.5 * 0.5 = 250, \sigma = \sqrt{250} \approx 16$$

Since this mean and standard deviation specify the normal distribution, we can calculate the corresponding z-score for 550 heads:

$$z = \frac{550 - 500}{16} > 3$$

This means that, if the coin were fair, the event of seeing 550 heads should occur with a < 1% chance under normality assumptions. Therefore, the coin is likely biased.

Page 23 of 23 4918 words Text Predictions: On Accessibility: Unavailable

5:01 AM 6/8/2022

Top 30 NLP Interview Questions x 40 Probability & Statistics Data x

nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street

Nick Singh

Previously at data startup SafeGraph, and Software Engineer on Facebook's Growth Team. Author of [Ace the Data Science Interview](#).

[Join the 46,000 readers who subscribe to my tech career newsletter!](#)

I send an email just once a month with guides on Tech Careers, Data Science, & Startups, as well as a few links to interesting articles & books on careers and technology.

BLOG
ACE THE DATA SCIENCE INTERVIEW
MASSAPPLY - COLD EMAIL FOR TECH JOBS
14 BOOKS THAT CHANGED MY LIFE
ABOUT ME

in tw ig ex

Problem #20 Solution:

Assume we have n Bernoulli trials each with a success probability of p :

$$x_1, x_2, \dots, x_n, x_i \sim \text{Ber}(p)$$

Assuming iid trials, we can compute the sample mean for p from a large number of trials:

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

We know the expectation of this sample mean is:

$$E[\hat{\mu}] = \frac{np}{n} = p$$

Additionally, we can compute the variance of this sample mean:

$$\text{Var}(\hat{\mu}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Assume we sample a large n . Due to the Central Limit Theorem, our sample mean will be normally distributed:

$$\hat{\mu} \sim N(p, \frac{p(1-p)}{n})$$

Therefore we can take a z-score of our sampled mean as:

$$z(\hat{\mu}) = \frac{\hat{\mu} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

This z-score will then be a simulated value from a standard normal distribution.

Stay-in-the-loop: How To Get More Data Science Interview Prep Resources

5:02 AM 6/8/2022

Top 30 NLP Interview Questions x 40 Probability & Statistics Data x

nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street

Nick Singh

Previously at data startup SafeGraph, and Software Engineer on Facebook's Growth Team. Author of [Ace the Data Science Interview](#).

Join the 46,000 readers who subscribe to my tech career newsletter!

I send an email just once a month with guides on Tech Careers, Data Science, & Startups, as well as a few links to interesting articles & books on careers and technology.

BLOG

ACE THE DATA SCIENCE INTERVIEW

MASSAPPLY - COLD EMAIL FOR TECH JOBS

14 BOOKS THAT CHANGED MY LIFE

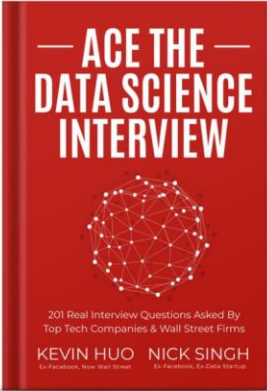
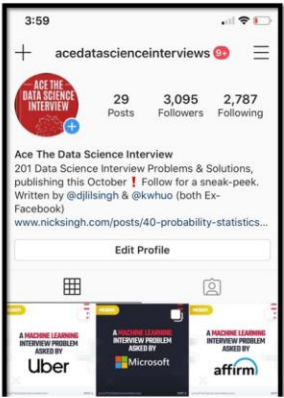
ABOUT ME

in

This z-score will then be a simulated value from a standard normal distribution.

Stay-in-the-loop: How To Get More Data Science Interview Prep Resources

Make sure to [buy the full 301-page book on Amazon](#) and follow along the [Acing The Data Science Interview Instagram](#) & [Nick's tech careers email newsletter](#) to get more like this.

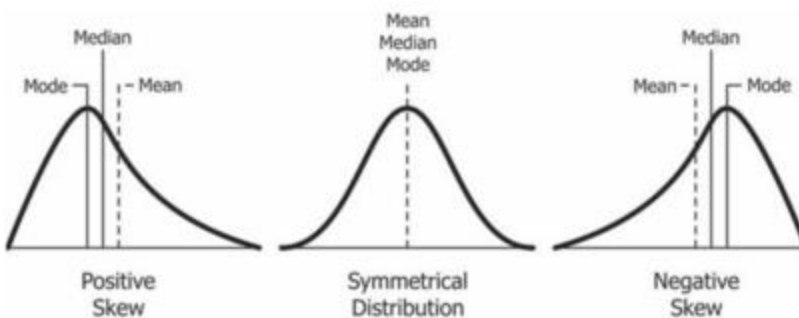
For general Data Science career advice, make sure you've read the [Breaking Into Data Science Guide](#) and the [Guide To Creating Kick-Ass Machine Learning & Data Science Portfolio Projects](#). And feel free to connect with Nick personally on [Instagram](#), [LinkedIn](#), and [Twitter](#). You can also watch video Q&A we did

1. Which of the following relation is correct for a negative skewed distribution?

- (a) Mean=Mode=Median
- (b) Mean>Median>Mode
- (c) Mode>Median>Mean
- (d) Mean>Mode=Median

Solution:(c)

Explanation:



2. In the symmetric covariance matrix:

- (a) Diagonal elements must be positive and other elements are always zero.
- (b) Diagonal elements can never be negative and other elements are always positive.
- (c) Diagonal elements can never be negative and other elements can be negative or positive.
- (d) Diagonal elements can be negative and positive and other elements are always negative.

Solution: (c)

Explanation: In a covariance matrix, the diagonal entries represent covariance of the variable with itself which is equal to the variance of that variable and is calculated as the square of standard deviation. Since variance is always positive, therefore diagonal entries are always positive.

Presence of Outliers in a dataset not affects:

- (a) Standard deviation
- (b) Range
- (c) Mean
- (d) Inter-quartile Range(IQR)

Solution: (d)

Explanation: The IQR is essentially the range of the middle 50% of the data. Since it uses the middle 50%, therefore it is not affected by the outliers.

If X and Y are independent random variables, then which of the following is TRUE?

- (a) $E(XY)=E(X)E(Y)$ [E represents Expectation value]

(b) $\text{Cov}(X, Y) = 0$
variables]

[Cov represents covariance between

(c) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

[Var represents variance]

(d) All of the above

Solution: (d)

Explanation: If X and Y are independent then $\text{Cov}(X, Y) = 0$ and $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$ ($\because 2\text{Cov}(X, Y) = 0$)

5. For a normal distribution Z, which option is TRUE?

(a) Coefficient of skewness ($E(Z^3)$) = 0

(b) $E(Z) = 0$; $E(Z^2) = \text{Var}(Z) = 1$

(c) Kurtosis ($E(Z^4)$) = 3

(d) Its density is symmetric about the mean.

Solution: (d)

Explanation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

6. Let X and Y be normal random variables with their respective means 3 and 4 and variances 9 and 16, then 2X-Y will have normal distribution with parameters:

(a) Mean=2 and Variance=52

(b) Mean=0 and Variance=1

(c) Mean=2 and Variance=1

(d) None of the above

Solution: (d)

Hint: $\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab\text{Cov}(X, Y)$

Suppose X and Y take values {0,1} and are independent with $P(X=1)=1/2$ and $P(Y=1)=1/3$. What is the probability that $P(X+Y=1)$?

(a) 5/18

(b) 1/2

(c) 5/6

(d) 1/6

Solution:(b)

Explanation: $P(X + Y = 1) = P(X=0).P(Y=1) + P(X=1).P(Y=0) = (1/2)(1/3) + (1/2)(2/3) = 1/2$.

8. Let X and Y are random variables with $E(X)=\mu/2$ and $E(Y)=\mu$, then which one is TRUE?

(a) $g=X+Y$ is an unbiased estimator of μ

(b) $g = X+Y$ is a biased estimator of μ with bias equals to μ

(c) $h=X+(Y/2)$ is an unbiased estimator of μ

(d) $h= X+(Y/2)$ is a biased estimator of μ with bias equals to $\mu/2$

Solution: (c)

**Explanation: $E(g) = E(X+Y) = E(X) + E(Y) = 3\mu/2$; $\text{Bias}(g) = E(g) - \mu = \mu/2$
 $E(h) = E(X+(Y/2)) = E(X) + 1/2E(Y) = \mu$, $\text{Bias}(h) = E(h) - \mu = 0$**

9. Suppose that X takes values between 0 and 1 and has probability density function(PDF) $2x$, then the value of Variance of X^2 is :

- (a) $1/12$
- (b) $1/18$
- (c) $1/6$
- (d) $5/18$

Solution:(a)

Hint: Use $\text{Var}(X^2) = E(X^4) - (E(X^2))^2$

10. For random variables X and Y, we have $\text{Var}(X)=1$, $\text{Var}(Y)=4$, and $\text{Var}(2X-3Y)=34$, then the correlation between X and Y is:

- (a) $1/2$
- (b) $1/4$
- (c) $1/3$
- (d) None of the above

Solution:(b)

Explanation: $\text{Var}(2X-3Y) = 34$

$$= 4\text{Var}(X) + 9\text{Var}(Y) - 12\text{Cov}(X, Y)$$

$$= 4(1) + 9(4) - 12\text{Cov}(X, Y) = 34$$

$$\therefore \text{Cov}(X, Y) = 1/2$$

11. A fair die is rolled repeatedly until a number larger than 4 is observed. If K is the total number of times that the die is rolled, then $P(K=4)$ is equal to:

- (a) $16/81$
- (b) $8/81$

(c) $8/27$

(d) $16/27$

Solution: (b)

Explanation: $P(K=4) = (P(\text{\#less than 4 or equal}))^3 \cdot P(\{4\}) = (2/3)^3 \cdot (1/3) = 8/81$.

12. Let X and Y be independent uniform $(0, 1)$ random variables. Define $A=X+Y$ and $B=X-Y$. Then,

(a) A and B are independent random variables

(b) A and B are uncorrelated random variables

(c) A and B are both uniforms $(0,1)$ random variables.

(d) None of these

Solution: (b)

Explanation: $\text{Cov}(X+Y, X-Y) = \text{Cov}(X, X) - \text{Cov}(X, Y) + \text{Cov}(Y, X) - \text{Cov}(Y, Y) \Rightarrow \text{Var}(X) - \text{Var}(Y) = 0$

13. If g is a point estimator of X , then Mean Square error(MSE) for g is:

(a) $\text{Variance}(g) + \text{Bias}(g)$

(b) $\text{Variance}(g) + \text{Bias}(g^2)$

(c) $\text{Variance}(g) + (\text{Bias}(g))^2$

(d) $\text{Variance}(g^2) + \text{Bias}(g)$

Solution: (c)

Explanation: $\text{MSE}(g) = E[(g-X)^2] = \text{Var}(g-X) + (E[g-X])^2 = \text{Var}(g) + (\text{Bias}(g))^2$

14. Let X and Y be two random variables and let a, b, c, d be real numbers, then which one of the following is FALSE?

- (a) $\text{Cov}(X+b, Y+d) = \text{Cov}(X, Y)$
- (b) $\text{Cov}(aX, cY) = ac \cdot \text{Cov}(X, Y)$
- (c) $\text{Cov}(aX+b, cY+d) = ac \cdot \text{Cov}(X, Y)$
- (d) $\text{Corr}(aX+b, cY+d) = ac \cdot \text{Corr}(X, Y)$ for $a, c > 0$

Solution: (d)

Explanation: $\text{Corr}(aX+b, cY+d) = \text{Corr}(X, Y)$

15. Let X and Y be jointly(bivariate) normal with $\text{Var}(X) = \text{Var}(Y)$, then:

- (a) $X+Y$ and $X-Y$ are jointly normal
- (b) $X+Y$ and $X-Y$ are uncorrelated
- (c) $X+Y$ and $X-Y$ are independent
- (d) All of the above

Solution: (d)

Explanation: If X and Y be the bivariate normal distribution, then any linear combination of X and Y is also normally distributed.

16. Let $X_1, X_2, X_3, \dots, X_n$ be a random sample from a distribution with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ Now, consider two estimators:

$$g_1 = X_1 \qquad g_2 = \bar{X} = (X_1 + X_2 + X_3 + \dots + X_n)/n$$

Which of these estimator has high mean squared error(MSE)?

- (a) g_1
- (b) g_2
- (c) Same for both g_1 and g_2
- (d) None of the above

Solution: (a)

Explanation: $MSE(g_1) = E[(g_1 - \mu)^2] = E[(X_1 - E(X_1))^2] = \text{Var}(X_1) = \sigma^2$

$MSE(g_2) = E[(g_2 - \mu)^2] = E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X} - \mu) + (E[\bar{X} - \mu])^2 = \text{Var}(\bar{X}) = \sigma^2/n$

17. A random sample of $n=6$ taken from the population has the elements 6, 10, 13, 14, 18, 20. Then, which option is False?

- (a) Point estimate for population mean is 13.5
- (b) Point estimate for population standard deviation is 4.68
- (c) Point estimate for population standard deviation is 3.5
- (d) Point estimate for standard error of mean is 1.91

Solution: (c)

Explanation: Population mean(\bar{X}) = $(\sum X_i/n) = 13.5$

Population standard deviation(S) = $\sqrt{(\sum X_i^2/n) - (\sum X_i/n)^2} = 4.68$

Standard error of mean = $S/\sqrt{n} = 4.68/\sqrt{6} = 1.91$

True or False: If the Pearson's correlation between 2 variables is zero, then they are necessarily independent.

Solution: False

Explanation: Correlation is a measure of linear dependence between the variables.

True or False: Let g be an unbiased estimator of X and U be a random variable with zero means, then $h=g+U$ is also unbiased for X .

Solution: True.

Explanation: $E(h) = E(g) + E(U) = 0+0 = 0$ ($\because E(g)=0$ due to unbiased estimator)

True or False: Let X and Y be two independent standard normal random variables and $T=XY^2+X+1$ and $P=X-3$, then $\text{Cov}(T, P)=1$

Solution: False.

Hint: Use properties mentioned in Question-14.

21. True or False: Let X has a normal distribution with parameters μ and σ^2 , then X^2 follows a chi-square distribution with parameter 1.

Solution: False.

Explanation: For the given statement to be True, X should be Standard normal distribution ($\mu=0, \sigma^2=1$)

22. True or False: If the characteristic function of a random variable exists, then its expectation and variance will also exist.

Solution: False.

Hint: Moment Generating Function(MGF)

23. True or False: Let X has uniform distribution $U(a, b)$ such that $E(X)=2$ and $\text{Var}(X)=3/4$, then $P(X<1)=1/6$.

Solution: True.

Explanation: $E(X) = (a+b)/2 = a+b=4$; $\text{Var}(X) = (b-a)^2/12 = (b-a)=3 \Rightarrow X \sim U(0.5, 3.5)$

24. True or False: The correlation coefficient between $X+Y$ and $X-Y$, where X and Y are independent random variables with variances 36 and 16 respectively is $6/13$.

Solution: False.

Explanation: $\text{Corr}(X+Y, X-Y) = \text{Cov}(X+Y, X-Y) / \text{Std}(X+Y) \cdot \text{Std}(X-Y)$ [Std= Standard Deviation]

25. True or False: In interval estimation, As the confidence level increases the margin of error decreases.

Solution: False.

Explanation: The Confidence Interval is defined as $\bar{X} \pm Z(s/\sqrt{n})$