Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Answer is a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

Answer is a)Central Limit Theorem

Explanation : The Central Limit Theorem is one of the most important theorems in all of statistics. It states that the distribution of averages of id variables (properly normalized) becomes that of a standard normal as the sample size increases.

3. Which of the following is incorrect with respect to use of Poisson distribution?
a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

Answer is b) Modeling bounded count data

Explanation
Poision Distribution is the discrete probability distribution of the number of events occuring in a given time period,given the average number of times the event occurs over that time period.Poisson distribution is used for modeling unbounded count data.

4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Answer is a)All of the mentioned.

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the
Answer is c) Poision

6.Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Answer is b) False
Usually replacing the standard error by its estimated value does change the CLT.

7. 1. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Answer is b)Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Answer is a) 0

9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Answer is c)

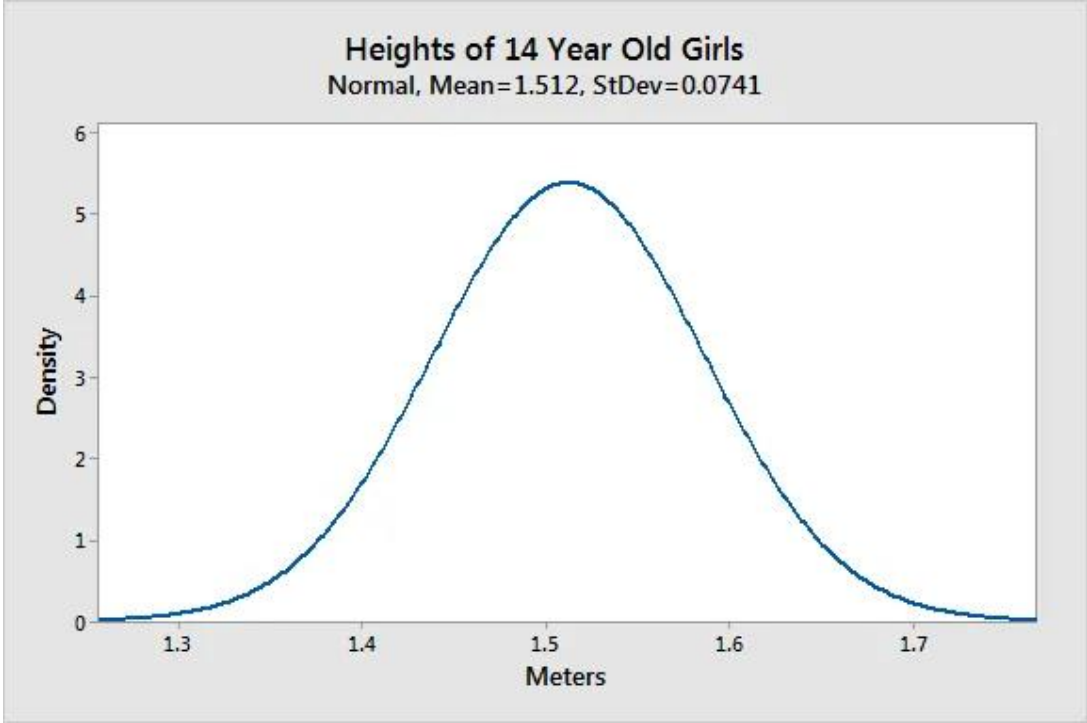
Explanation:

Outliers cannot conform to the regression relationship
In statistics and applications of statistics, normalization can have a range of meanings

Q10: What do you understand by the term Normal Distribution?

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.



As you can see, the distribution of heights follows the typical bell curve pattern for all normal distributions. Most girls are close to the average (1.512 meters). Small differences between an individual's height and the mean occur more frequently than substantial deviations from the mean. The standard deviation is 0.0741m, which indicates the typical distance that individual girls tend to fall from mean height.

The distribution is symmetric. The number of girls shorter than average equals the number of girls taller than average. In both tails of the distribution, extremely short girls occur as infrequently as extremely tall girls.

Common Properties for All Forms of the Normal Distribution

Despite the different shapes, all forms of the normal distribution have the following characteristic properties.

- They're all symmetric bell curves. The Gaussian distribution cannot model skewed distributions.
- The mean, median, and mode are all equal.
- Half of the population is less than the mean and half is greater than the mean.
- The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean. More on this below!

Q11 : How do you handle missing data ? What imputation techniques do you recommend?

Dealing with missing values in Python

When no data value is stored for feature for particular observation, we say this feature has missing values.

Usually this missing value in Data Science could be represented as "?",'N/A',0 or just blank cell.

In the example below the normalized losses has a missing value which is represented as Nan.

	Symboling	Normalized-Losses	Make	Fuel-Type	Aspiration	No of doors	Body Style	Drive Wheels	Engine Location
0	3	Nan	Alfa-romero	Gas	Std	2	Convertible	Rwd	Front

How to deal with missing data

There are many ways to deal with missing values. Of course, each situation is different and should be judged differently. However below are the typical options which we can consider.

1) Check with data source collection.
   Check with the person or group and go find what the actual data should be.
2) Drop the missing values.
   Just remove the data where the missing value is found. We can either do the following:
   a) Drop the variable.
   b) Drop the data entry. If you don't have lots of observations, then it best to drop the entry from the missing tables.
3) Replacing the missing values.
   Replacing data is better since no data is wasted, however it is less accurate since we are replacing the missing data with a guess of what the data should be. Below is the standard technique of replacing the data:
   a) Replacing it with an average (of similar data points)
   As for an example, suppose we have missing entry for the entry Normalized columns and the column average for entries if data is 4500. For this we don't have accurate guess as what missing

values for normalized losses columns should have been. We can approximate the values which is using the average values of the columns 4500.

We should keep in mind that values cannot be average in case of categorical variables. For example, for a variable Fuel type, there can't be average of fuel type. So in this case we should try to use the below.

b) Replace it by frequency. It means replace it with what is most common. For this we can use mode function.
c) Replace it based on other functions. Finally, sometimes we may find other way to replace missing data. This is usually because there is something additional about missing data. For example, we may know missing values tend to be old cars and the normalized losses for the old cars are significantly higher than the average vehicle

4) Leave it as missing data.
5) Use encoder and imputer technique to transform the data. If we have data in terms of object data type. In such cases, the model doesn't understand such data type. So we have to come with an idea to convert object data type into numeric (integer or float). For instance, in case we have continuous data type, where we have salary based on experience, and at one place experience data is missing. In such cases, it is best to use Encoder and Imputer techniques to resolve our problem.

Q12 : What is A/B Testing?

A/B testing is generally conducted for websites, where business want to see how users will react to new website design. They create multiple websites design and test which design attracts most users. Whichever website attract most users, then that website will be made final to the customers.

Here ,A/B is just the name given for this kind of testing.

Q13 : Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Its so simple. And yet, so dangerous. Perhaps that's a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.

But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

Let's first see the many reasons not to use mean imputation (and to be fair, its advantages).

First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

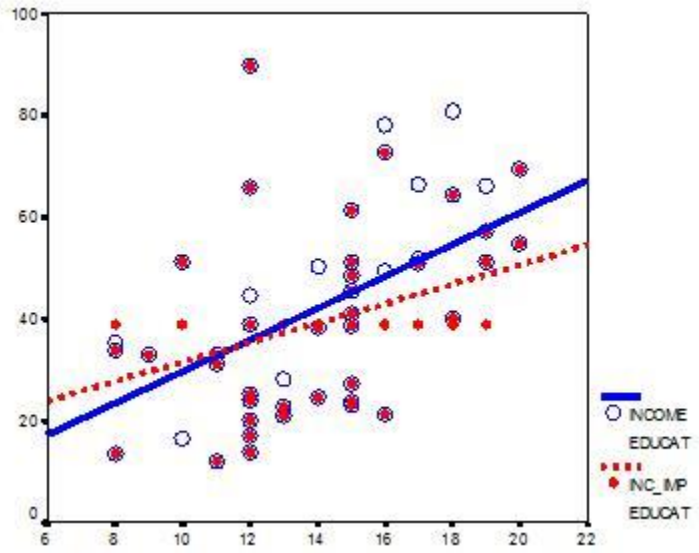
Problem #1: Mean imputation does not preserve the relationships among variables.

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.This is the original logic involved in mean imputation. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

It *will* still bias your standard error, but I will get to that in another post.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The following graph illustrates this well:



This graph illustrates hypothetical data between X=years of education and Y=annual income in thousands with n=50. The blue circles are the original data, and the solid blue line indicates the best fit regression line for the full data set. The correlation between X and Y is r = .53.

I then randomly deleted 12 observations of income (Y) and substituted the mean. The red dots are the mean-imputed data.

Blue circles with red dots inside them represent non-missing data. Empty Blue circles represent the missing data. If you look across the graph at Y = 39, you will see a row of red dots without blue circles. These represent the imputed values.

The dotted red line is the new best fit regression line with the imputed data. As you can see, it is less steep than the original line. Adding in those red dots pulled it down.

The new correlation is r = .39. That's a lot smaller that .53.

The real relationship is quite underestimated.

Of course, in a real data set, you wouldn't notice so easily the bias you're introducing. This is one of those situations where in trying to solve the lowered sample size, you create a bigger problem.

One note: if X were missing instead of Y, mean substitution would artificially *inflate* the correlation.

In other words, you'll think there is a stronger relationship than there really is. That's not good either. It's not reproducible and you don't want to be overstating real results.

This solution that is so good at preserving unbiased estimates for the mean isn't so good for unbiased estimates of relationships.

Problem #2: Mean Imputation Leads to An Underestimate of Standard Errors

A second reason is applied to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

Ultimately, because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.

That's not good.

Q14 : What is linear regression in Statistics?

This is explained in Linear Regression Python file

Q15 : What are the various branches of statistics?

Types of statistics and parameters involved in statistics

We'll learn types of statistics and different parameters involved in the statistics , we will understand how each type differs from one another and how we can interpret the parameters .

Two types are:

- Descriptive
- Inferential

1.Descriptive statistics : if data can be described without any statistical tools then it is called descriptive statistics . ex, marks in class , height of student.

2.Inferential statistics: if data is too big then then we use inferential statistics ,

We take a few samples from different data and we find the average. This is called inferential statistics. The average is then applicable to all the data from where we have selected our samples.

population and sample

We can understand this topic by taking an example of election. Ever wondered how the media came up with the exit polls?

Assume an Indian state. Karnataka, now the media will take a few samples from all the cities in Karnataka.

Collecting the data from a few populations in order to form an exit poll is one of the examples of inferential statistics and with this example population and samples are also included.

After taking samples from the few populations, we will now start making predictions.

Another example can be of the electronics, in order to check the quality of the television, not all televisions have to go through the quality control process. A few items are selected randomly and through the test, if they pass the test, all the television sets are labeled as approved.

2. Analytics Methodology and How Industry Use Statistics

1.Weather forecasting

2.Giving Insurance

3.Stock Market

4.Drug Effectiveness before releasing to the market

5.Diseased survival probability

6.Election winning and exit poll prediction

7.Loan approval and fraud detection

8.Netflix/Amazon recommendation

9.New Campaign Effectiveness

Parameters and Statistics Mean,Median and Mode

Mean= Average

Median= Centre Data

Centre data if the sample is in an odd number.

If the sample is even then we add both the middle value and divide by 2.

Mode= Also called frequency, the most number of occurrences in a sample is termed as mode.

Standard Deviation

Number of data deviated from the given mean is called standard deviation.

Variance

It is the Square of standard deviation.

The command for finding mean is

```
Print ( st.mean(x))
```

The command for finding median is

```
Print ( st.median(x))
```

The command for finding mode is

```
Print ( st.mode(x))
```