41 questions on Statistics for data scientists & analysts

analyticsvidhya.com/blog/2017/05/41-questions-on-statisitics-data-scientists-analysts

Dishashree Gupta Dishashree is passionate about statistics and is a machine learning enthusiast. She has an experience of 1.5 years of Market Research using R, advanced Excel, Azure ML.

May 4, 2017

Introduction

Statistics forms the back bone of data science or any analysis for that matter. Sound knowledge of statistics can help an analyst to make sound business decisions.

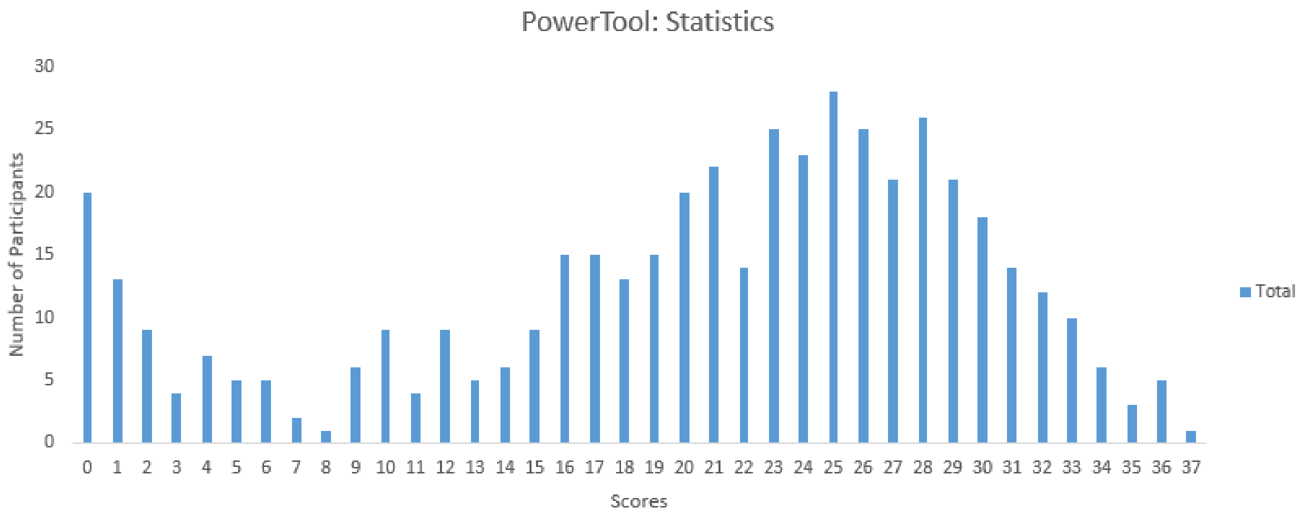
On one hand, descriptive statistics helps us to understand the data and its properties by use of central tendency and variability. On the other hand, inferential statistics helps us to infer properties of the population from a given sample of data. Knowledge of both descriptive and inferential statistics is essential for an aspiring data scientist or analyst.

To help you improve your knowledge in statistics we conducted this practice test. The test covered both descriptive and inferential statistics in brief. I am providing the answers with explanation in case you got stuck on particular questions.

In case you missed the test, try solving the questions before reading the solutions.

Overall Scores

Below are the distribution scores, they will help you evaluate your performance.



You can access the final scores here. More than 450 people took this test and the highest score obtained was 37. Here are a few statistics about the distribution.

Mean Score: 20.40

Median Score: 23

Mode Score: 25

Questions & Solution

1) Which of these measures are used to analyze the central tendency of data?

A) Mean and Normal Distribution

B) Mean, Median and Mode

C) Mode, Alpha & Range

D) Standard Deviation, Range and Mean

E) Median, Range and Normal Distribution

Solution: (B)

The mean, median, mode are the three statistical measures which help us to analyze the central tendency of data. We use these measures to find the central value of the data to summarize the entire data set.

2) Five numbers are given: (5, 10, 15, 5, 15). Now, what would be the sum of deviations of individual data points from their mean?

A) 10

B)25

C) 50

D) 0

E) None of the above

Solution: (D)

The sum of deviations of the individual will always be 0.

3) A test is administered annually. The test has a mean score of 150 and a standard deviation of 20. If Ravi’s z-score is 1.50, what was his score on the test?

A) 180
B) 130
C) 30

D) 150
E) None of the above

Solution: (A)

X= μ+Zσ where μ is the mean, σ is the standard deviation and X is the score we're calculating. Therefore X = 150+20*1.5 = 180

4) Which of the following measures of central tendency will always change if a single value in the data changes?

A) Mean

B) Median

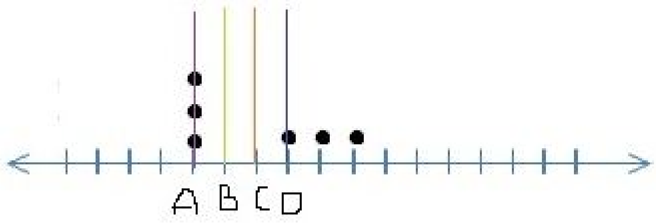
C) Mode

D) All of these

Solution: (A)

The mean of the dataset would always change if we change any value of the data set. Since we are summing up all the values together to get it, every value of the data set contributes to its value. Median and mode may or may not change with altering a single value in the dataset.

5) Below, we have represented six data points on a scale where vertical lines on scale represent unit.



Which of the following line represents the mean of the given data points, where the scale is divided into same units?

A) A
B) B
C) C
D) D

Solution: (C)

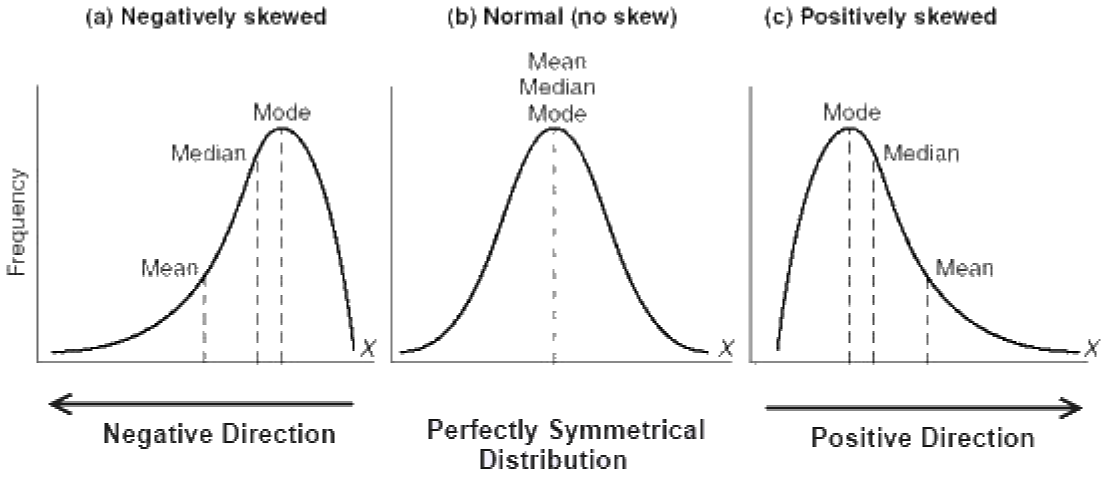
It’s a little tricky to visualize this one by just looking at the data points. We can simply substitute values to understand the mean. Let A be 1, B be 2, C be 3 and so on. The data values as shown will become {1,1,1,4,5,6} which will have mean to be 18/6 = 3 i.e. C.

6) If a positively skewed distribution has a median of 50, which of the following statement is true?

A) Mean is greater than 50
B) Mean is less than 50
C) Mode is less than 50
D) Mode is greater than 50
E) Both A and C
F) Both B and D

Solution: (E)

Below are the distributions for Negatively, Positively and no skewed curves.



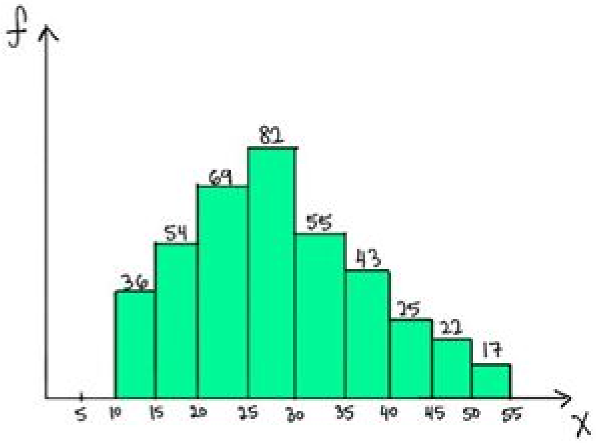
As we can see for a positively skewed curve, Mode<Median<Mean. So if median is 50, mean would be more than 50 and mode will be less than 50.

7) Which of the following is a possible value for the median of the below distribution?

A) 32
B) 26
C) 17
D) 40

Solution: (B)

To answer this one we need to go to the basic definition of a median. Median is the value which has roughly half the values before it and half the values after. The number of values less than 25 are (36+54+69 = 159) and the number of values greater than 30 are (55+43+25+22+17= 162). So the median should lie somewhere between 25 and 30. Hence 26 is a possible value of the median.



8) Which of the following statements are true about Bessels Correction while calculating a sample standard deviation?

1. **Bessels correction is always done when we perform any operation on a sample data.**
2. **Bessels correction is used when we are trying to estimate population standard deviation from the sample.**
3. **Bessels corrected standard deviation is less biased.**

A) Only 2

B) Only 3

C) Both 2 and 3

D) Both 1 and 3

Solution: (C)

Contrary to the popular belief Bessel's correction should not be always done. It's basically done when we're trying to estimate the population standard deviation using the sample standard deviation. The bias is definitely reduced as the standard deviation will now(after correction) be depicting the dispersion of the population more than that of the sample.

9) If the variance of a dataset is correctly computed with the formula using (n – 1) in the denominator, which of the following option is true?

A) Dataset is a sample
B) Dataset is a population
C) Dataset could be either a sample or a population
D) Dataset is from a census
E) None of the above

Solution: (A)

If the variance has n-1 in the formula, it means that the set is a sample. We try to estimate the population variance by dividing the sum of squared difference with the mean with n-1.

When we have the actual population data we can directly divide the sum of squared differences with n instead of n-1.

10) [True or False] Standard deviation can be negative.

A) TRUE

B) FALSE

Solution: (B)

Below is the formula for standard deviation

Since the differences are squared, added and then rooted, negative standard deviations are not possible.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

11) Standard deviation is robust to outliers?

A) True

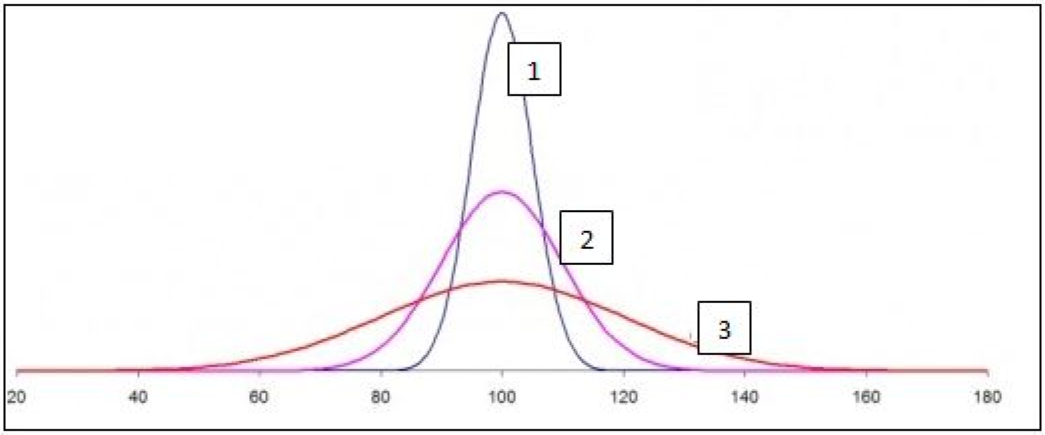
B) False

Solution: (B)

If you look at the formula for standard deviation above, a very high or a very low value would increase standard deviation as it would be very different from the mean. Hence outliers will effect standard deviation.

12) For the below normal distribution, which of the following option holds true ?

σ1, σ2 and σ3 represent the standard deviations for curves 1, 2 and 3 respectively.



A) σ1> σ2> σ3

B) σ1< σ2< σ3

C) σ1= σ2= σ3

D) None

Solution: (B)

From the definition of normal distribution, we know that the area under the curve is 1 for all the 3 shapes. The curve 3 is more spread and hence more dispersed (most of values being within 40-160). Therefore it will have the highest standard deviation. Similarly, Curve 1 has a very low range and all the values are in a small range of 80-120. Hence, curve 1 has the least standard deviation.

13) What would be the critical values of Z for 98% confidence interval for a two-tailed test ?

A) +/- 2.33
B) +/- 1.96
C) +/- 1.64
D) +/- 2.55

Solution: (A)

We need to look at the z table for answering this. For a 2 tailed test, and a 98% confidence interval, we should check the area before the z value as 0.99 since 1% will be on the left side of the mean and 1% on the right side. Hence we should check for the z value for area>0.99. The value will be +/- 2.33

14) [True or False] The standard normal curve is symmetric about 0 and the total area under it is 1.

A)TRUE

B) FALSE

Solution: (A)

By the definition of the normal curve, the area under it is 1 and is symmetric about zero. The mean, median and mode are all equal and 0. The area to the left of mean is equal to the area on the right of mean. Hence it is symmetric.

Context for Questions 15-17

Studies show that listening to music while studying can improve your memory. To demonstrate this, a researcher obtains a sample of 36 college students and gives them a standard memory test while they listen to some background music. Under normal circumstances (without music), the mean score obtained was 25 and standard deviation is 6. The mean score for the sample after the experiment (i.e With music) is 28.

15) What is the null hypothesis in this case?

A) Listening to music while studying will not impact memory.
B) Listening to music while studying may worsen memory.
C) Listening to music while studying may improve memory.
D) Listening to music while studying will not improve memory but can make it worse.

Solution: (D)

The null hypothesis is generally assumed statement, that there is no relationship in the measured phenomena. Here the null hypothesis would be that there is no relationship between listening to music and improvement in memory.

16) What would be the Type I error?

A) Concluding that listening to music while studying improves memory, and it's right.
B) Concluding that listening to music while studying improves memory when it actually doesn't.
C) Concluding that listening to music while studying does not improve memory but it does.

Solution: (B)

Type 1 error means that we reject the null hypothesis when its actually true. Here the null hypothesis is that music does not improve memory. Type 1 error would be that we reject it and say that music does improve memory when it actually doesn't.

17) After performing the Z-test, what can we conclude ____ ?

A) Listening to music does not improve memory.

B)Listening to music significantly improves memory at p

C) The information is insufficient for any conclusion.

D) None of the above

Solution: (B)

Let's perform the Z test on the given case. We know that the null hypothesis is that listening to music does not improve memory.

Alternate hypothesis is that listening to music does improve memory.

In this case the standard error i.e. $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$

The Z score for a sample mean of 28 from this population is

Z critical value for α = 0.05 (one tailed) would be 1.65 as seen from the z table.

$$Z = \frac{sample\ mean - population\ mean}{\text{standard error}} = \frac{28-25}{1} = 3$$

Therefore since the Z value observed is greater than the Z critical value, we can reject the null hypothesis and say that listening to music does improve the memory with 95% confidence.

18) A researcher concludes from his analysis that a placebo cures AIDS. What type of error is he making?

A) Type 1 error

B) Type 2 error

C) None of these. The researcher is not making an error.

D) Cannot be determined

Solution: (D)

By definition, type 1 error is rejecting the null hypothesis when its actually true and type 2 error is accepting the null hypothesis when its actually false. In this case to define the error, we need to first define the null and alternate hypothesis.

19) What happens to the confidence interval when we introduce some outliers to the data?

A) Confidence interval is robust to outliers

B) Confidence interval will increase with the introduction of outliers.

C) Confidence interval will decrease with the introduction of outliers.

D) We cannot determine the confidence interval in this case.

Solution: (B)

We know that confidence interval depends on the standard deviation of the data. If we introduce outliers into the data, the standard deviation increases, and hence the confidence interval also increases.

Context for questions 20- 22

A medical doctor wants to reduce blood sugar level of all his patients by altering their diet. He finds that the mean sugar level of all patients is 180 with a standard deviation of 18. Nine of his patients start dieting and the mean of the sample is observed to 175. Now, he is considering to recommend all his patients to go on a diet.

Note: He calculates 99% confidence interval.

20) What is the standard error of the mean?

A) 9
B) 6
C) 7.5
D) 18

Solution: (B)

The standard error of the mean is the standard deviation by the square root of the number of values. i.e.

Standard error = $18/\sqrt{9}$ = 6

21) What is the probability of getting a mean of 175 or less after all the patients start dieting?

A) 20%
B) 25%
C) 15%
D) 12%

Solution: (A)

This actually wants us to calculate the probability of population mean being 175 after the intervention. We can calculate the Z value for the given mean.

$$Z=\frac{\text{sample mean}-\text{population}}{\text{standard error}}=\frac{175-180}{6}$$

$$Z=-\frac{5}{6}=-0.833$$

If we look at the z table, the corresponding value for z = -0.833 ~ 0.2033.

Therefore there is around 20% probability that if everyone starts dieting, the population mean would be 175.

22) Which of the following statement is correct?

A) The doctor has a valid evidence that dieting reduces blood sugar level.

B) The doctor does not have enough evidence that dieting reduces blood sugar level.

C) If the doctor makes all future patients diet in a similar way, the mean blood pressure will fall below 160.

Solution: (B)

We need to check if we have sufficient evidence to reject the null. The null hypothesis is that dieting has no effect on blood sugar. This is a two tailed test. The z critical value for a 2 tailed test would be ±2.58.

The z value as we have calculated is -0.833.

Since Z value < Z critical value, we do not have enough evidence that dieting reduces blood sugar.

Question Context 23-25

A researcher is trying to examine the effects of two different teaching methods. He divides 20 students into two groups of 10 each. For group 1, the teaching method is using fun examples. Where as for group 2 the teaching method is using software to help students learn. After a 20 minutes lecture of both groups, a test is conducted for all the students.

We want to calculate if there is a significant difference in the scores of both the groups.

It is given that:

- Alpha=0.05, two tailed.
- Mean test score for group 1 = 10
- Mean test score for group 2 = 7
- Standard error = 0.94

23) What is the value of t-statistic?

A) 3.191
B) 3.395
C) Cannot be determined.
D) None of the above

Solution: (A)

The t statistic of the given group is nothing but the difference between the group means by the standard error.

=(10-7)/0.94 = 3.191

24) Is there a significant difference in the scores of the two groups?

A) Yes
B) No

Solution: (A)

The null hypothesis in this case would be that there is no difference between the groups, while the alternate hypothesis would be that the groups are significantly different.

The t critical value for a 2 tailed test at α = 0.05 is ±2.101. The t statistic obtained is 3.191. Since the t statistic is more than the critical value of t, we can reject the null hypothesis and say that the two groups are significantly different with 95% confidence.

25) What percentage of variability in scores is explained by the method of teaching?

A) 36.13
B) 45.21
C) 40.33
D) 32.97

Solution: (A)

The % variability in scores is given by the R^2 value. The formula for R^2 given by

$$R^2 = \frac{\text{t square}}{\text{t square} + \text{degree of freedom}}$$

The degrees of freedom in this case would be 10+10 -2 since there are two groups with size 10 each. The degree of freedom is 18.

$$R^2 = \frac{3.191*3.191}{(3.191*3.191)+18} = 36.13$$

26) [True or False] F statistic cannot be negative.

A) TRUE

B) FALSE

Solution: (A)

F statistic is the value we receive when we run an ANOVA test on different groups to understand the differences between them. The F statistic is given by the ratio of between group variability to within group variability

Below is the formula for f Statistic.

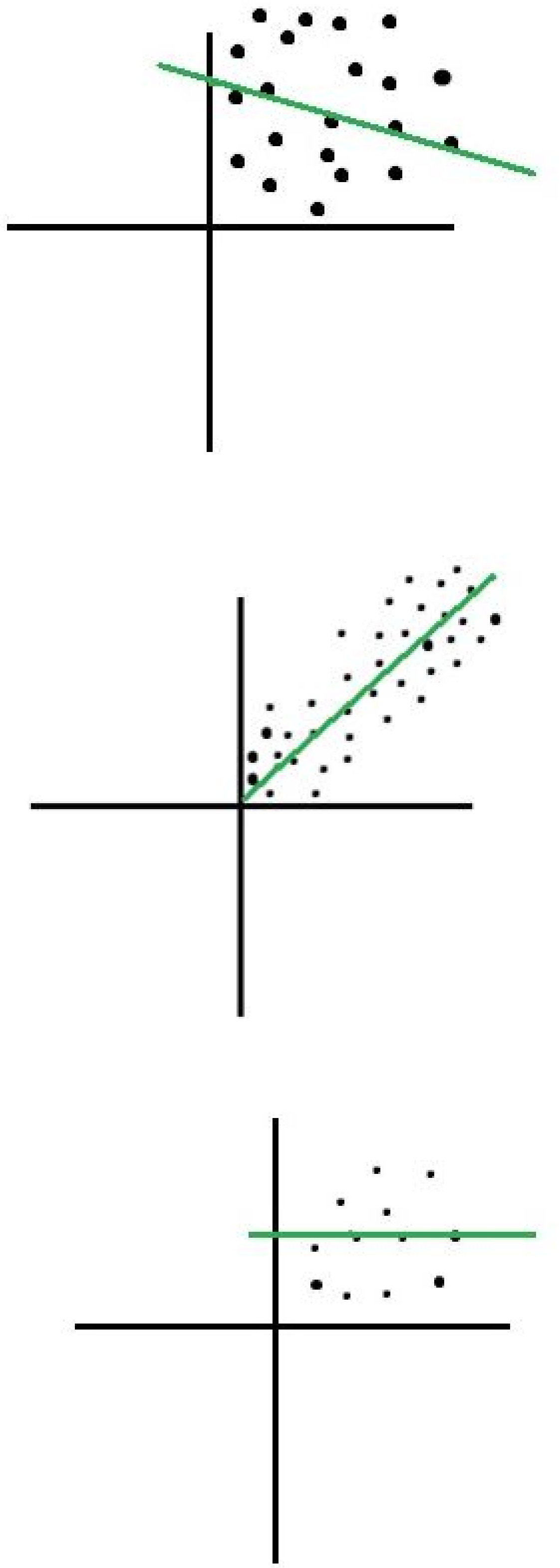
$$\frac{\text{Sum of squared error for between group/degree of freedom of between group}}{\text{Sum of squared error for within group/degree of freedom of within group}}$$

Since both the numerator and denominator possess square terms, F statistic cannot be negative.

27) Which of the graph below has very strong positive correlation?
A)
B)
C)
D)

Solution: (B)

A strong positive correlation would occur when the following condition is met. If x increases, y should also increase, if x decreases, y should also decrease. The slope of the line would be positive in this case and the data points will show a clear linear relationship. Option B shows a strong positive relationship.



28) Correlation between two variables (Var1 and Var2) is 0.65. Now, after adding numeric 2 to all the values of Var1, the correlation co-efficient will_______ ?

A) Increase
B) Decrease
C) None of the above

Solution: (C)

If a constant value is added or subtracted to either variable, the correlation coefficient would be unchanged. It is easy to understand if we look at the formula for calculating the correlation.

If we add a constant value to all the values of x, the x_i and will change by the same number, and the differences will remain the same. Hence, there is no change in the correlation coefficient.

29) It is observed that there is a very high correlation between math test scores and amount of physical exercise done by a student on the test day. What can you infer from this?

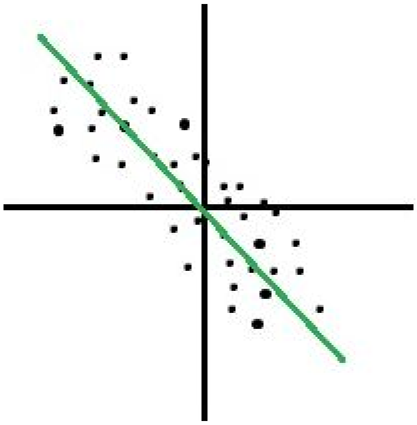
1. **High correlation implies that after exercise the test scores are high.**
2. **Correlation does not imply causation.**
3. **Correlation measures the strength of linear relationship between amount of exercise and test scores.**

A) Only 1
B) 1 and 3
C) 2 and 3
D) All the statements are true

Solution: (C)

Though sometimes causation might be intuitive from a high correlation but actually correlation does not imply any causal inference. It just tells us the strength of the relationship between the two variables. If both the variables move together, there is a high correlation among them.

30) If the correlation coefficient (r) between scores in a math test and amount of physical exercise by a student is 0.86, what percentage of variability in math test is explained by the amount of exercise?



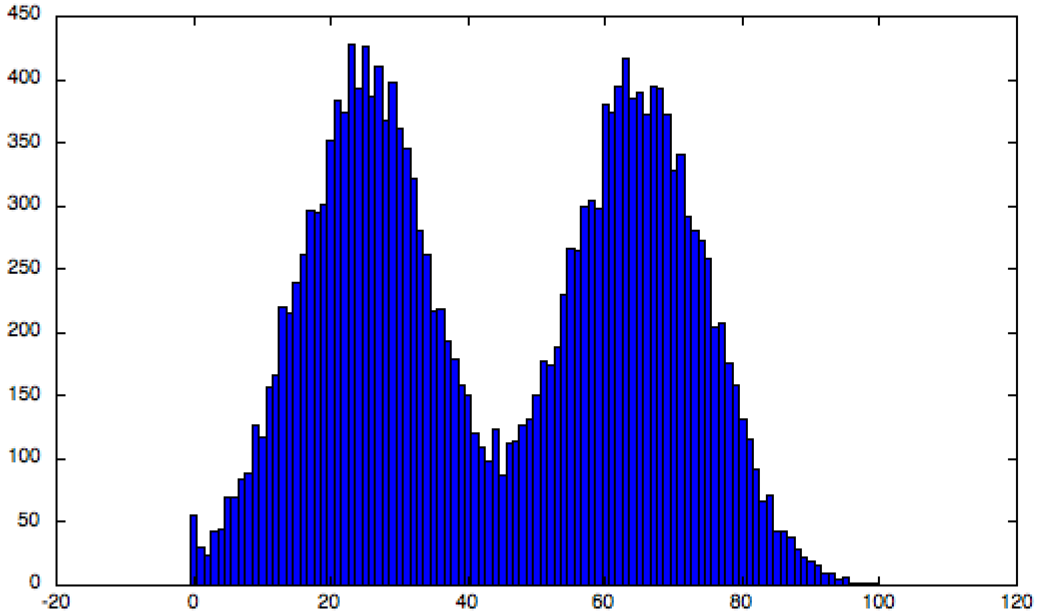
A) 86%
B) 74%
C) 14%
D) 26%

Solution: (B)

The % variability is given by r^2, the square of the correlation coefficient. This value represents the fraction of the variation in one variable that may be explained by the other variable. Therefore % variability explained would be 0.86^2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

31) Which of the following is true about below given histogram?



A) Above histogram is unimodal

B) Above histogram is bimodal

C) Given above is not a histogram

D) None of the above

Solution: (B)

The above histogram is bimodal. As we can see there are two values for which we can see peaks in the histograms indicating high frequencies for those values. Therefore the histogram is bimodal.

32) Consider a regression line y=ax+b, where a is the slope and b is the intercept. If we know the value of the slope then by using which option can we always find the value of the intercept?

A) Put the value (0,0) in the regression line True

B) Put any value from the points used to fit the regression line and compute the value of b False

C) Put the mean values of x & y in the equation along with the value a to get b False

D) None of the above can be used False

Solution: (C)

In case of ordinary least squares regression, the line would always pass through the mean values of x and y. If we know one point on the line and the value of slope, we can easily find the intercept.

33) What happens when we introduce more variables to a linear regression model?

A) The r squared value may increase or remain constant, the adjusted r squared may increase or decrease.

B) The r squared may increase or decrease while the adjusted r squared always increases.

C) Both r square and adjusted r square always increase on the introduction of new variables in the model.

D) Both might increase or decrease depending on the variables introduced.

Solution: (A)

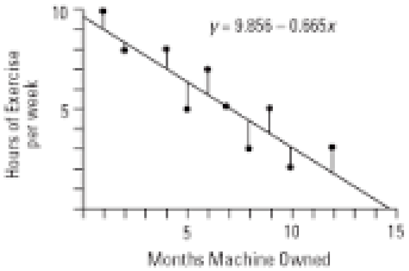
The R square always increases or at least remains constant because in case of ordinary least squares the sum of square error never increases by adding more variables to the model. Hence the R squared does not decrease. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

34) In a scatter diagram, the vertical distance of a point above or below regression line is known as ____ ?

A) Residual
B) Prediction Error
C) Prediction
D) Both A and B
E) None of the above

Solution: (D)

The lines as we see in the above plot are the vertical distance of points from the regression line. These are known as the residuals or the prediction error.



35) In univariate linear least squares regression, relationship between correlation coefficient and coefficient of determination is ______ ?

A) Both are unrelated False

B) The coefficient of determination is the coefficient of correlation squared True

C) The coefficient of determination is the square root of the coefficient of correlation False

D) Both are same F

Solution: (B)

The coefficient of determination is the R squared value and it tells us the amount of variability of the dependent variable explained by the independent variable. This is nothing but correlation coefficient squared. In case of multivariate regression the r squared value represents the ratio of the sum of explained variance to the sum of total variance.

36) What is the relationship between significance level and confidence level?

A) Significance level = Confidence level
B) Significance level = 1- Confidence level
C) Significance level = 1/Confidence level
D) Significance level = sqrt (1 – Confidence level)

Solution: (B)

Significance level is 1-confidence interval. If the significance level is 0.05, the corresponding confidence interval is 95% or 0.95. The significance level is the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained when the null hypothesis is true. The confidence interval is the range of likely values for a population parameter, such as the population mean. For example, if you compute a 95% confidence interval for the average price of an ice cream, then you can be 95% confident that the interval contains the true average cost of all ice creams.

The significance level and confidence level are the complementary portions in the normal distribution.

37) [True or False] Suppose you have been given a variable V, along with its mean and median. Based on these values, you can find whether the variable "V" is left skewed or right skewed for the condition

```
mean(V) > median(V)
```

A) True
B) False

Solution: (B)

Since, its no where mentioned about the type distribution of the variable V, we cannot say whether it is left skewed or right skewed for sure.

38) The line described by the linear regression equation (OLS) attempts to ____ ?

A) Pass through as many points as possible.

B) Pass through as few points as possible

C) Minimize the number of points it touches

D) Minimize the squared distance from the points

Solution: (D)

The regression line attempts to minimize the squared distance between the points and the regression line. By definition the ordinary least squares regression tries to have the minimum sum of squared errors. This means that the sum of squared residuals should be minimized. This may or may not be achieved by passing through the maximum points in the data. The most common case of not passing through all points and reducing the error is when the data has a lot of outliers or is not very strongly linear.

39) We have a linear regression equation (Y = 5X +40) for the below table.

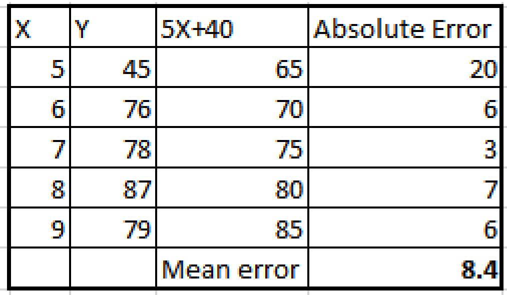
X	Y
5	45
6	76
7	78
8	87
9	79

Which of the following is a MAE (Mean Absolute Error) for this linear model?

A) 8.4
B) 10.29
C) 42.5
D) None of the above

Solution: (A)

To calculate the mean absolute error for this case, we should first calculate the values of y with the given equation and then calculate the absolute error with respect to the actual values of y. Then the average value of this absolute error would be the mean absolute error. The below table summarises these values.



X	Y	5X+40	Absolute Error
5	45	65	20
6	76	70	6
7	78	75	3
8	87	80	7
9	79	85	6
		Mean error	**8.4**

40) A regression analysis between weight (y) and height (x) resulted in the following least squares line: y = 120 + 5x. This implies that if the height is increased by 1 inch, the weight is expected to

A) increase by 1 pound
B) increase by 5 pound
C) increase by 125 pound
D) None of the above

Solution: (B)

Looking at the equation given y=120+5x. If the height is increased by 1 unit, the weight will increase by 5 pounds. Since 120 will be the same in both cases and will go off in the difference.

41) [True or False] Pearson captures how linearly dependent two variables are whereas Spearman captures the monotonic behaviour of the relation between the variables.

A)TRUE

B) FALSE

Solution: (A)

The statement is true. Pearson correlation evaluated the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

The spearman evaluates a monotonic relationship. A monotonic relationship is one where the variables change together but not necessarily at a constant rate.

End Notes

I hope you had fun solving the questions and they did make you scratch your head sometime. Please share your thoughts on the above topics and also your feedback.

We shall be happy to incorporate your ideas in further articles and tests. Also, one question might have multiple approaches and the solution above might show just one. I have tried to be descriptive with the solutions but feel free to investigate further in case of doubts using the comments below.

Learn, compete, hack and get hired!

You can also read this article on Analytics Vidhya's Android APP