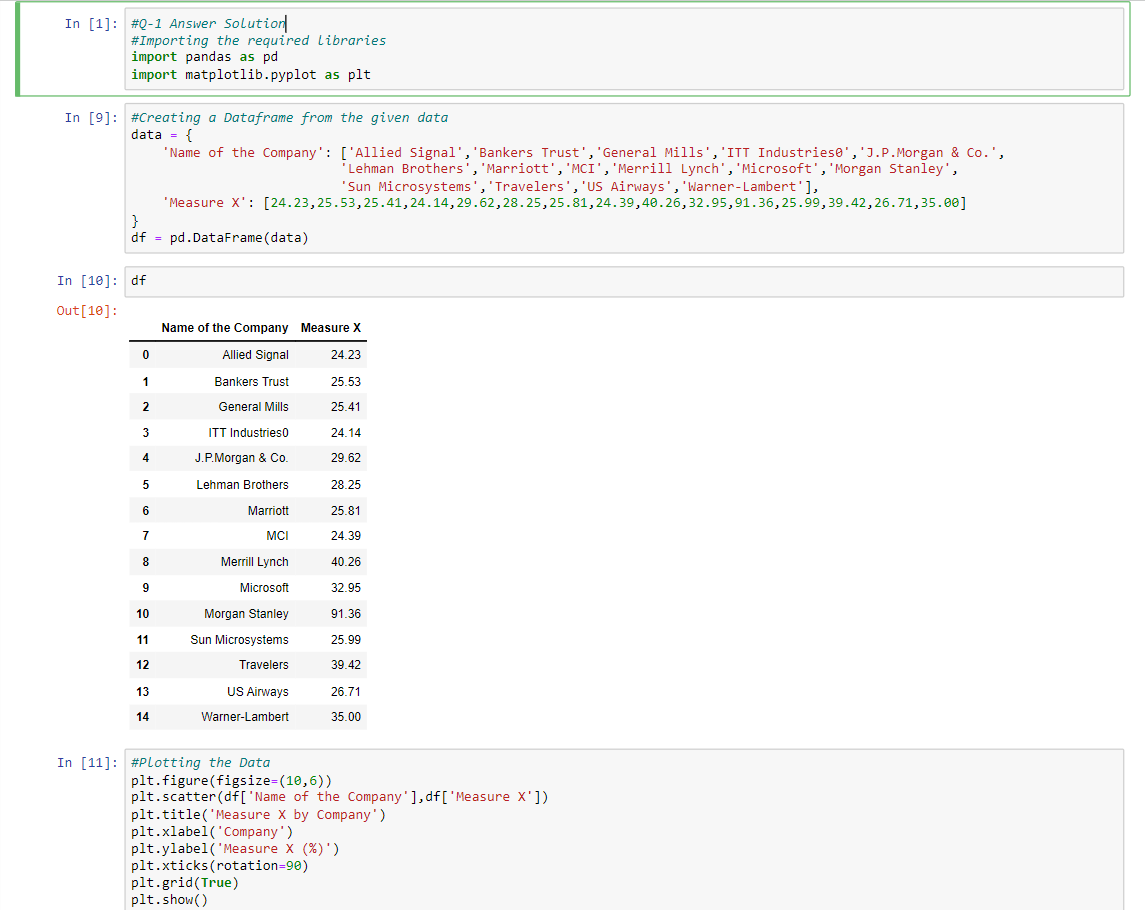
Basic-Stats Assignment-02

**Topics: Descriptive Statistics and Probability**

Q-1) Look at the data given below. Plot the data, find the outliers, and find out .

Ans-1) 

A screenshot of a graph

Description automatically generated

In the given two screenshots, I have used the libraries required for finding out the mean, standard deviation, and Variance.

Further, I created the data frame and put all the data as given in the question and then created the plotting graph using the matplotlib library.

Q-2) 

Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.
2. What can we say about the skewness of this dataset?
3. If it was found that the data point with the value 25 is 2.5, how would the new boxplot be affected?

Ans-2) i) Approximately, (First Quantile Range) Q1 = 5

(Third Quantile Range) Q3 = 12,

Median (Second Quantile Range) = 7 (Inter-Quartile Range)

IQR = Q3 -Q1 = 12 -5 = 7

Second Quartile Range is the Median value.

ii) The Skewness of this dataset is that the Right-Skewed median is towards the left side it is not normal distribution.

iii) In this case, there would be no

Outliers on the given dataset because of the outlier the data had positive skewness it will reduce, and the data will be normally distributed.

Q-3 

Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?
2. Comment on the skewness of the dataset.
3. Suppose that the above histogram and the boxplot in question 2 are plotted for the same dataset. Explain
4. how these graphs complement each other in providing information about any dataset.

Ans-3 i) The mode of the dataset lies in between 5 to 10 and approximately between 4 to 8.

ii) The skewness of the dataset is rightly skewed

Where, Mean>Median>Mode.

iii) Both the graphs are right-skewed, and both have outliers the median can be easily visualized in box plot whereas in histogram mode is more visible.

Q-4 AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that “could happen.” Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Ans-4 **IF** 1 in 200 long-distance telephone calls are getting misdirected.

probability of call misdirecting   = 1/200

Probability of call not Misdirecting = 1-1/200 = 199/200

**The** probability for at least one in five attempted telephone calls reaches the wrong number.

Number of Calls = 5

n = 5

p = 1/200

q = 199/200

P(x) = at least one in five attempted telephone calls reaches the wrong number

P(x) = ⁿCₓ pˣ qⁿ⁻ˣ

P(x) = (nCx) (p^x) (q^n-x) # **nCr = n! / r!** **\* (n - r)!**

P(1) = (5C1) (1/200)^1 (199/200)^5-1

P(1) = 0.0245037

Q-5 Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution.

|  |  |
| --- | --- |
| x | P(x) |
| -2,000 | 0.1 |
| -1,000 | 0.1 |
| 0 | 0.2 |
| 1000 | 0.2 |
| 2000 | 0.3 |
| 3000 | 0.1 |

1. What is the most likely monetary outcome of the business venture?
2. Is the venture likely to be successful? Explain
3. What is the long-term average earning of business ventures of this kind? Explain
4. What is the good measure of the risk involved in a venture of this kind? Compute this measure.

Ans-5 i) E(X) =Sum X.\*P(X) | E(X^2) =X^2\*P(X)

-200             | 400000

-100                 | 100000

0             | 0

200       | 200000

600         | 1200000

300         | 900000

Total: 800         | 2800000

1. The most likely monetary outcome of the business venture is 2000$

As for 2000$ the probability is 0.3 which is maximum as compared to others.

1. Ans: Yes, the probability that the venture will make more than 0 or a profit.

p(x>0) +p(x>1000) +p(x>2000) +p(x=3000) = 0.2+0.2+0.3+0.1 = 0.8

this states that there are a good 80% chances for this venture to be making a profit.

1. Ans: The long-term average is Expected value = Sum (X \* P(X)) = 800$ which means on an average the returns will be + 800$
2. Ans: The good measure of the risk involved in a venture of this kind depends on the Variability in the distribution. Higher Variance means more chances of risk.

Var (X) = E(X^2) –(E(X)) ^2

= 2800000 – 800^2

= 2160000

Topics: Normal Distribution, Functions of Random Variables

Q-1 The time required for servicing transmissions is normally distributed with *μ* = 45 minutes and *σ* = 8 minutes. The service manager plans to have work begin on the transmission of a customer’s car 10 minutes after the car is dropped off and the customer is told that the car will be ready within 1 hour from drop-off. What is the probability that the service manager cannot meet his commitment?

1. 0.3875
2. 0.2676
3. 0.5
4. 0.6987

Ans-1 We have a normal distribution with *μ* = 45 and *σ* = 8.0.

Let X be the amount of time it takes to complete the repair on a customer's car.

P (X > 50) = 1 - P (X ≤ 50).

Z = (X -*μ* )/ *σ* = (X - 45)/8.0

Therefore,

P (X ≤ 50) = P (Z ≤ (50 - 45)/8.0) = P (Z ≤ 0.625) =73.4%

Probability that the service manager will not meet his demand will be = 100-73.4 = 26.6% or 0.2676

Option B is the Correct one for this question.

Q-2 The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean *μ* = 38 and Standard deviation *σ* =6. For each statement below, please specify True/False. If false, briefly explain why.

1. More employees at the processing center are older than 44 than between 38 and 44.
2. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Ans-2 We have a normal distribution with = 38 and = 6. Let X be the number of employees. So according to question

a) Probability of employees greater than age of 44= P (X>44)

P (X > 44) = 1 – P (X ≤ 44).

Z = (X - *μ* )/*σ* = (X - 38)/6

P (X ≤ 44) = P (Z ≤ (44 - 38)/6) = P (Z ≤ 1) =84.1345%

Probability that the employee will be greater than age of 44 = 100-84.1345=15.86%

So, the probability of number of employees between 38-44 years of age = P (X<44)-0.5=84.1345-0.5= 34.1345%

Therefore, the statement that “More employees at the processing center are older than 44 than between 38 and 44” is TRUE.

b) Probability of employees less than age of 30 = P (X<30).

Z = (X - *μ* )/*σ* = (30 - 38)/6

P (X ≤ 30) = P (Z ≤ (30 - 38)/6) = P (Z ≤ -1.333) =9.12%

So, the number of employees with probability 0.912 of them being under age 30 = 0.0912\*400=36.48(or 36 employees).

Therefore, the statement B of the question is also TRUE.

Q-3 If *X1* ~ *N*(μ, σ2) and *X*2 ~ *N*(μ, σ2) are *iid* normal random variables, then what is the difference between 2 *X*1 and *X*1 + *X*2? Discuss both their distributions and parameters.

Ans-3

As we know that if X ∼ N (µ1, σ1^2), and Y ∼ N (µ2, σ2^2) are two independent random variables then X + Y ∼ N (µ1 + µ2, σ1^2 + σ2^2), and X − Y ∼ N (µ1 − µ2, σ1^2 + σ2^2).

Similarly, if Z = aX + bY, where X and Y are as defined above, i.e. Z is linear combination of X and Y, then Z ∼ N (aµ1 + bµ2, a^2σ1^2 + b^2σ2^2).

Therefore, in the question

2X1~ N (2 u,4 σ^2) and

X1+X2 ~ N (µ + µ, σ^2 + σ^2) ~ N (2 u, 2σ^2)

2X1-(X1+X2) = N (4µ,6 σ^2)

Q-4 Let X ~ N (100, 202). Find two values, *a* and *b*, symmetric about the mean, such that the probability of the random variable taking a value between them is 0.99.

1. 90.5, 105.9
2. 80.2, 119.8
3. 22, 78
4. 48.5, 151.5
5. 90.1, 109.9

Ans-4

The Probability of getting value between a and b should be 0.99.

So, the Probability of going wrong, or the Probability outside the a and b area is 0.01 (i.e. 1-0.99).

The Probability towards left from a = -0.005 (i.e. 0.01/2).

The Probability towards right from b = +0.005 (i.e. 0.01/2).

So, since we have the probabilities of a and b, we need to calculate X, the random variable at a and b which has got these probabilities.

By finding the Standard Normal Variable Z (Z Value), we can calculate the X values.

Z= (X- μ) / σ

For Probability 0.005 the Z Value is -2.57 (from Z Table).

Z \* σ + μ = X

Z (-0.005)\*20+100 = -(-2.57)\*20+100 = 151.4

Z (+0.005) \*20+100 = (-2.57) \*20+100 = 48.6

So, option D is correct.

Q-5 Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions Profit1 ~ N (5, 32) and Profit2 ~ N (7, 42) respectively. Both the profits are in $ Million. Answer the following questions about the total profit of the company in Rupees. Assume that $1 = Rs. 45

1. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.
2. Specify the 5th percentile of profit (in Rupees) for the company.
3. Which of the two divisions has a larger probability of making a loss each year?

Ans-5 A) Total profit = Profit1 + Profit2

Mean of total profit = Mean1 + Mean2 = 5 + 7 = 12 million dollars

Variance of total profit = Variance1 + Variance2 = 32 + 42 = 74

Standard deviation of total profit = √74

To convert this into rupees, we'll multiply by the conversion rate:

Mean of total profit (in rupees) = 12 \* 45 = 540 million rupees

Standard deviation of total profit (in rupees) = √74 \* 45

Finding out the z-score corresponding to the 97.5th percentile (since it’s a two tailed test for 95% probability) and will use it to calculate the rupee range around the mean as follows:

Mean of total profit (in rupees) = 12 \* 45 = 540 million rupees

Standard deviation of total profit (in rupees) = √74 \* 45

Now, we'll find the z-score corresponding to the 97.5th percentile (since it's a two-tailed test for 95% probability):

Z

97.5th percentile = invNorm (0.975)

Z

97.5th percentile =1.96

Using the z-score formula, we'll calculate the rupee range around the mean:

Rupee range=Mean of total profit (in rupees) ±Z

97.5th percentile×Standard deviation of total profit (in rupees)

Rupee range=540±1.96× (74×45)

Therefore, Rupee Range for the annual profit of the company = 540±1.96× (74×45).

B) To find the 5th percentile of profit in rupees for the company, we'll use the formula:

5th percentile of profit (in rupees) =Mean of total profit (in rupees) −Z5th percentile×Standard deviation of total profit (in rupees)5th percentile of profit (in rupees) =Mean of total profit (in rupees) −Z 5th percentile×Standard deviation of total profit (in rupees)

To find the 5th percentile of profit in rupees for the company, we'll use the formula:

Z 5th percentile=invNorm (0.05)

Z 5th percentile=−1.64

5th percentile of profit (in rupees) =540−(−1.64) × (8.6×45)

5th percentile of profit (in rupees) =540+1.64×387

5th percentile of profit (in rupees) =540+634.68

5th percentile of profit (in rupees) =1174.68

Therefore, the 5th percentile of profit for the company is approximately 1174.68 million rupees.

C) To determine which division has a larger probability of making a loss, we'll compare the probabilities of profit being negative for each division using their respective cumulative distribution functions (CDFs).

Calculating the probabilities of making a loss for each division using their mean and standard deviation:

For Profit1: P(Profit1<0) = P (Z< 0−5/underroot32) = P (Z < -1.77)

For Profit2: P(Profit2<0) = P (Z< 0−7/underroot42) = P(Z<−2.42)

Using the standard normal distribution table or a calculator to find the probabilities associated with these z-scores:

For Profit1: P(Profit1<0) = P(Z<−1.77)

For Profit2: P(Profit2<0) = P(Z<−2.42)

Using the standard normal distribution table, we find:

For Profit1: P(Profit1<0) = P(Z<−1.77) ≈ 0.0384

For Profit2: P(Profit2<0) = P(Z<−2.42) ≈ 0.0073

Therefore, profit1 has a probability of approximately 0.0384 of making a loss each year, while profit2 has a probability of approximately 0.0073 of making a loss.

So, profit2 has a larger probability of making a loss each year.

Topics: Confidence Intervals

Q-1 For each of the following statements, indicate whether it is True/False. If false, explain why.

1. The sample size of the survey should at least be a fixed percentage of the population size to produce representative results.
2. The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.
3. Larger surveys convey a more accurate impression of the population than smaller surveys.

Ans-1 I) The given statement is false as the sample size of a survey should not necessarily be a fixed percentage of the population size to produce representative results.

The appropriate sample size depends on various factors such as the desired level of confidence, the margin of error, the heterogeneity of the population and the research objectives.

Calculating sample size often involves statistical methods that consider these factors to ensure the survey results are representative of the population.

II) The given statement is false. The reason is that the sampling frame is a list of every item that comprises the population, not just those that appeared in the survey sample.

It is the list from which the sample is drawn. The sampling frame should ideally include all elements of the population to ensure that the sample is representative.

III) The given statement is false, while larger may provide more precise estimates of population parameters, they do not necessarily convey a more accurate impression of the population.

The accuracy of a survey depends on various factors including the sampling method, the representativeness of the sample, and the quality of data collection.

A smaller well-designed survey with a representative sample can provide accurate insights into the population, while a larger survey with biases or methodological issues may produce inaccurate results.

2. PC Magazine asked all its readers to participate in a survey of their satisfaction with different brands of electronics. In the 2004 survey, which was included in an issue of the magazine that year, more than 9000 readers rated the products on a scale from 1 to 10. The magazine reported that the average rating assigned by 225 readers to a Kodak compact digital camera was 7.5. For this product, identify the following:

A. The population

B. The parameter of interest

C. The sampling frame

D. The sample size

E. The sampling design

F. Any potential sources of bias or other problems with the survey or sample

Ans-2 A) The population:

The population in this scenario would be all readers of PC Magazine who participated in the survey of their satisfaction with different brands of electronics. This includes all readers who received the survey invitation, regardless of whether they responded or rated the Kodak compact digital camera.

B) The parameter of interest:

The parameter of interest in this scenario is the average satisfaction rating of all readers of PC Magazine for the Kodak compact digital camera. It is estimated to be 7.5 based on the average rating reported by the 225 readers who rated the camera.

C) The sampling frame:

The sampling frame is the list of all readers of PC Magazine who received the survey invitation. It includes all individuals who had the opportunity to participate in the survey.

D) The sample size:

The sample size in this scenario is 225 readers who rated the Kodak compact digital camera. These readers are a subset of the total population of readers who received the survey invitation.

E) The sampling design:

The sampling design used in this scenario appears to be convenience sampling, where all readers of PC Magazine were invited to participate in the survey voluntarily. However, without more information about the survey methodology, it's difficult to determine the exact sampling design.

F) Potential sources of bias or other problems with the survey or sample:

Self-selection bias: The survey relies on voluntary participation, which may lead to self-selection bias if individuals who chose to respond have different characteristics or opinions compared to those who did not respond.

Response bias: The reported average rating may be influenced by response bias if individuals who rated the Kodak compact digital camera had different opinions compared to those who did not participate in the survey.

Sampling bias: If the survey invitation was not distributed randomly or if certain groups of readers were more likely to respond to the survey, it could introduce sampling bias.

Lack of representativeness: The sample of 225 readers may not be representative of all PC Magazine readers, particularly if certain demographic groups are overrepresented or underrepresented in the sample.

Limited sample size: The sample size of 225 readers may be considered relatively small compared to the total population of readers, which could affect the reliability and generalizability of the survey results.

Scale limitations: The use of a 1 to 10 scale for rating products may not capture the full range of reader opinions and may result in less nuanced or accurate ratings.

Q-3 For each of the following statements, indicate whether it is True/False. If false, explain why.

I. If the 95% confidence interval for the average purchase of customers at a department store is $50 to $110, then $100 is a plausible value for the population mean at this level of confidence.

II. If the 95% confidence interval for the number of moviegoers who purchase concessions is 30% to 45%, this means that fewer than half of all moviegoers purchase concessions.

III. The 95% Confidence-Interval for μ only applies if the sample data are nearly normally distributed.

Ans-3 I) The given statement is true as a 95% confidence interval for the average purchase of customers at a department store of $50 to $110 means that if we were to repeat the sampling process many times and construct confidence intervals for each sample, we would expect approximately 95% of those intervals to contain the true population mean.

Since $100 falls within the confidence interval, it is indeed a plausible value for the population mean at this level of confidence.

II) The statement is false. A 95% confidence interval for the number of moviegoers who purchase concessions of 30% to 45% means that we are 95% confident that the true proportion of moviegoers who purchase concessions falls within this range.

It does not necessarily mean that fewer than half of all moviegoers purchase concessions, as the true proportion could be anywhere within the confidence interval.

III) The given statement is false as the 95% confidence interval for the population mean (μ) applies regardless of the distribution of the sample data, if the sample size is sufficiently large (typically n ≥ 30) due to the Central Limit Theorem.

Whereas, if the sample size is small and the population distribution is significantly non-normal, other methods such as the bootstrap method may be more appropriate for constructing confidence intervals.

Q-4 What are the chances that Xbar > μ?

A. ¼

B. ½

C. ¾

D. 1

Ans-4 The chance that the sample mean (Xbar) is greater than the population mean (μ) depends on the distribution of the sample means and the population mean.

If the sample means are normally distributed, and the population mean is equal to the mean of the sampling distribution of the sample means (Xbar), then the probability that Xbar>μ is 0.5 (or 50%).

Therefore, the correct answer is:

B. ½

Q-5 In January 2005, a company that monitors Internet traffic (WebSideStory) reported that its sampling revealed that the Mozilla Firefox browser launched in 2004 had grabbed a 4.6% share of the market.

I. If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?

II. WebSideStory claims that its sample includes all the daily Internet users. If that’s the case, then can Microsoft conclude that Mozilla has a less than 5% share of the market?

Ans-5 I) As (p\_value = 0.2058) > (α = 0.05); so, we Accept Null Hypothesis i.e. Mozilla market share > 5%.

Thus, Microsoft cannot conclude that Mozilla has a less than 5% share of the market.

II) Given that WebSideStory claims that its sample includes all the daily Internet users.

This means that the 4.6% is the population percentage.

Comparing it with Microsoft's claim that Mozilla has a less than 5% share of the whole market is True.

Hence, we can conclude that Mozilla has a less than 5% share of the market.

Q-6 A book publisher monitors the size of shipments of its textbooks to university bookstores. For a sample of texts used at various schools, the 95% confidence interval for the size of the shipment was 250 ± 45 books. Which, if any, of the following interpretations of this interval are correct?

1. All shipments are between 205 and 295 books.
2. 95% of shipments are between 205 and 295 books.
3. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.
4. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.
5. We can be 95% confident that the range 160 to 340 holds the population mean.

Ans-6 Option is True and correct as it describes the purpose of a confidence interval.

The 95% confidence interval indicates that if we were to take many samples and compute confidence intervals for each sample, approximately 95% of those intervals would contain the true population mean.

Q-7 Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ =s?

1. The z-interval is shorter.
2. The t-interval is shorter.
3. Both are equal.
4. We cannot say.

Ans-7 When the sample size is large (typically n≥30), the t-distribution approaches the standard normal distribution.

Therefore, for large sample sizes, the critical values for the t-distribution are like those for the standard normal distribution, and the lengths of the confidence intervals are comparable.

Whereas, when the sample size is small (typically

n<30), the t-distribution has heavier tails than the standard normal distribution.

As a result, the critical values for the t-distribution are larger than those for the standard normal distribution, leading to wider confidence intervals.

Therefore, the correct answer is:

B The t-interval is shorter.

Q-8 How many randomly selected employers (minimum number) must we contact to guarantee a margin of error of no more than 4% (at 95% confidence)?

A. 600

B. 400

C. 550

D. 1000

Ans-8 To determine the minimum sample size needed to guarantee a margin of error of no more than 4% at a 95% confidence level, we can use the formula for sample size calculation:

n=(Z×σ/E) 2

Where:

n = sample size

Z = Z-score corresponding to the desired confidence level (for 95% confidence level, Z is approximately 1.96)

σ = standard deviation (unknown, but can be estimated from a pilot study or assumed to be 0.5 for a conservative estimate in binary scenarios)

E = margin of error (in decimal form, so 4% would be 0.04)

Plugging in the values:

n = (1.96x0.5/0.04)2

n = (0.98/0.04)2

n = (24.5/0.04)2

n = (612.5)2

n ≈375156.25

To guarantee a margin of error of no more than 4% at a 95% confidence level, will be rounding up to the nearest whole number.

So, the minimum number of randomly selected employers we must contact is 613.

But we have been provided the closest option as 600 so looking at the option, we can see that the closest option is:

1. 600

While this doesn’t guarantee a margin of error of exactly 4%, it’s the closest option provided.

Q-9 Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?

1. 1000
2. 757
3. 848
4. 543

Ans-9 To determine the minimum sample size needed to guarantee a margin of error of no more than 4% at a 95% confidence level, we can use the formula for sample size calculation:

n=(Z×σ/E) 2

However, this time, we use the Z-score corresponding to a 98% confidence level, which is approximately 2.33.

Plugging in the values:

n = (2.33x0.5/0.04)2

n = (1.165/0.04)2

n = (29.125/0.04)2

n = (728.125)2

n ≈530109.7656

To guarantee a margin of error of no more than 4% at a 98% confidence level, we need to round up to the nearest whole number. So, the minimum sample size needed is 729.

Among the given options, the closest one to 729 is:

1. 757

Topics: Sampling Distributions and Central Limit Theorem

Q-1 Examine the following normal Quantile plots carefully. Which of these plots indicates that the data …

1. Are nearly normal?
2. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.)
3. Are skewed (i.e. not symmetric) ?
4. Have outliers on both sides of the center?



Ans-1 I) The option C indicates that the data are nearly Normal.

II) The option B and D have a bimodal distribution.

III) The option A, B and D are skewed (i.e. not symmetric)

IV) The option A and B have outliers on both sides of the center.

Q-2 For each of the following statements, indicate whether it is True/False. If false, explain why.

The manager of a warehouse monitors the volume of shipments made by the delivery team. The automated tracking system tracks every package as it moves through the facility. A sample of 25 packages is selected and weighed every day. Based on current contracts with customers, the weights should have *μ* = 22 lbs. and *σ* = 5 lbs.

1. Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that weights of individual packages are normally distributed.
2. The standard error of the daily average SE () = 1.

Ans-2 i) The statement is false as a sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

In our case the samples contain 25 packages, and the larger number of samples contain of each such 25 packages taken into different samples (25+25+25+25…and so on).

The mean for one these samples is 22lbs and standard deviation of 5lbs which means each individual package is having a weight varying between + or – 5lbs with respect to mean(22lbs).

Hence it is invalid to take a weight of individual packages and confirm that it follows normal distribution before using a normal model for the sampling distribution.

The Sample Central Limit Theorem states that the sampling distribution of the samples mean approaches normal distribution as the sample size is large enough.

ii) The statement is true as

E (Standard Error) = sample standard deviation / Square root of (number of sample) SE = 5 / (25) ^1/2SE=1

Q-3 Auditors at a small community bank randomly sample 100 withdrawal transactions made during the week at an ATM machine located near the bank’s main branch. Over the past 2 years, the average withdrawal amount has been $50 with a standard deviation of $40. Since audit investigations are typically expensive, the auditors decide to not initiate further investigations if the mean transaction amount of the sample is between $45 and $55. What is the probability that in any given week, there will be an investigation?

A. 1.25%

B. 2.5%

C. 10.55%

D. 21.1%

E. 50%

Ans-3 D t=(x-mean)/sigma/sqrt(n); t-test because standard deviation is not given for the long term = (45-50) or (55-50)/40/sqrt (100) =+/- 5/40/sqrt (100) =+/- 1.25.

The probability of z between those values is 0.7857, so probability of an investigation is 1-0.7887, or 0.214.

Q-4 The auditors from the above example would like to maintain the probability of investigation to 5%. Which of the following represents the minimum number transactions that they should sample if they do not want to change the thresholds of 45 and 55? Assume that the sample statistics remain unchanged.

1. 144
2. 150
3. 196
4. 250
5. Not enough information

Ans-4 The option D is correct as

For 5%, t-value is +/-1.96.

t\_value = (x\_bar – mew)/(sample\_standard\_deviation/sqrt(n))

So, 1.96=(5)/(sqrt(n)/40) sqrt(n)= (40\*tvalue)/ (5) n=248

Q-5 An educational startup that helps MBA aspirants write their essays is targeting individuals who have taken GMAT in 2012 and have expressed interest in applying to FT top 20 b-schools. There are 40000 such individuals with an average GMAT score of 720 and a standard deviation of 120. The scores are distributed between 650 and 790 with a very long and thin tail towards the higher end resulting in substantial skewness. Which of the following is likely to be true for randomly chosen samples of aspirants?

1. The standard deviation of the scores within any sample will be 120.
2. The standard deviation of the mean of across several samples will be 120.
3. The mean score in any sample will be 720.
4. The average of the mean across several samples will be 720.
5. The standard deviation of the mean across several samples will be 0.60

Ans-5 The statement E is true as

Standard Error = sigma / (n)^0.5 = standard deviation / (sample size)^0.5 = 120 / (40000)^0.5 = 0.60.