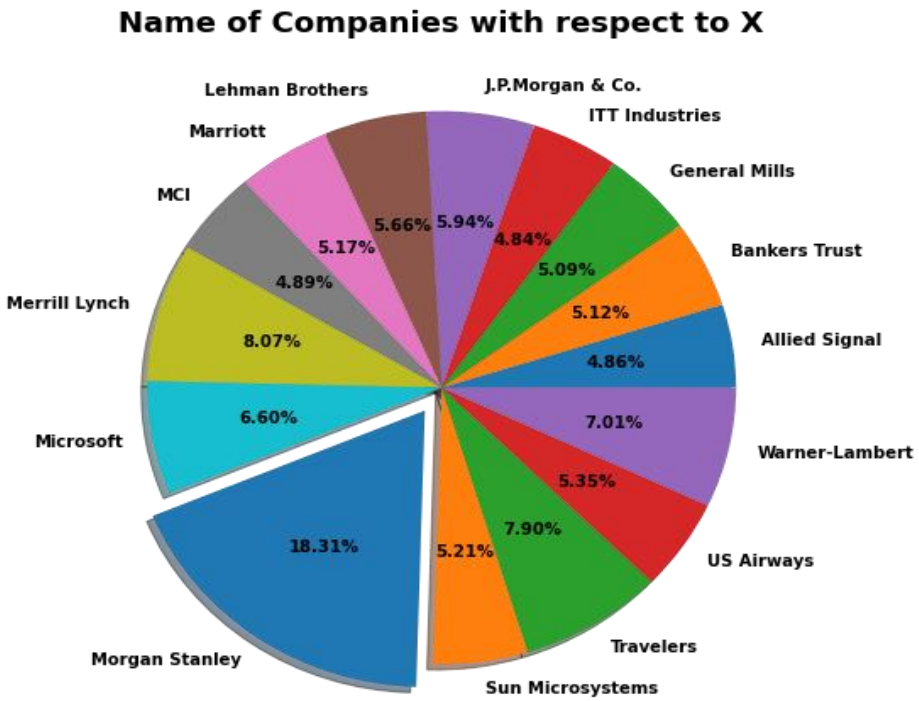
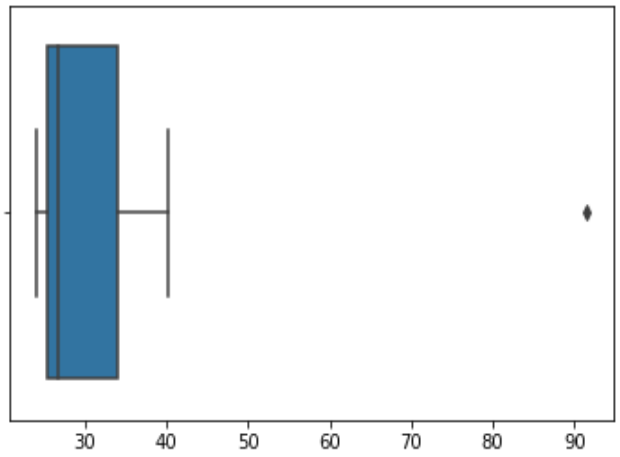
Topics: Descriptive Statistics and Probability

1. Look at the data given below. Plot the data, find the outliers and find out μ, σ, σ^2

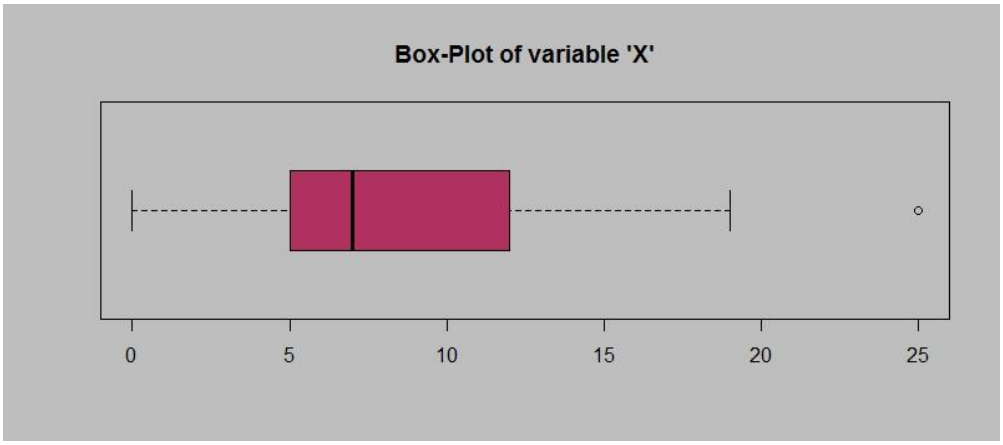
Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
J.P.Morgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%
Warner-Lambert	35.00%





The following is the outlier in the boxplot: Morgan Stanley 91.36%
Measure_x. describe ()
Mean = 33.271333
Standard deviation = 16.945401
measure_x.var ()
Variance = 287.1466123809524

2.



Answer the following three questions based on the box-plot above.

(i) What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.
Ans: Approximately (First Quantile Range) Q1 = 5 (Third Quantile Range) Q3 = 12, Median (Second Quartile Range) = 7
(Inter-Quartile Range) IQR = Q3 – Q1 = 12 – 5 = 7
Second Quartile Range is the Median Value

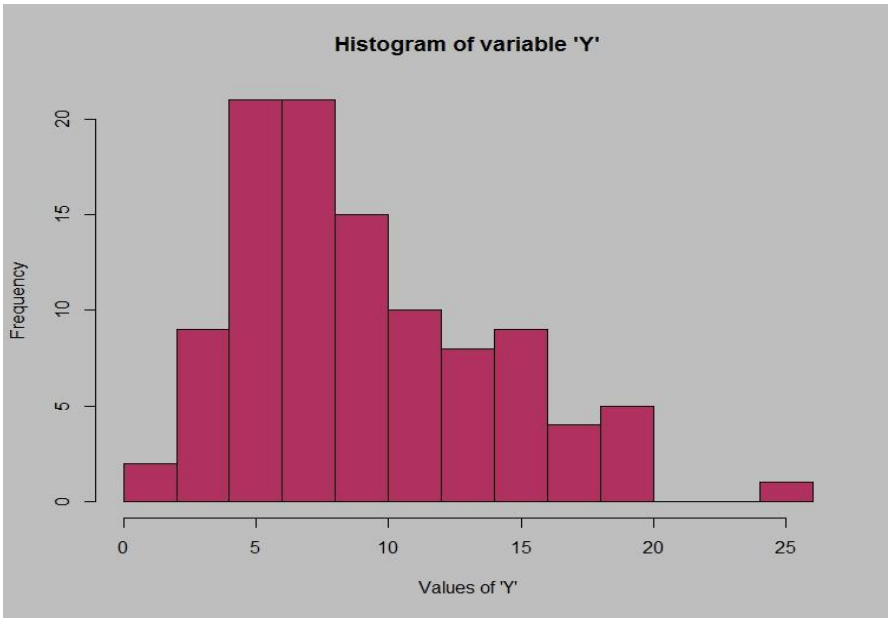
(ii) What can we say about the skewness of this dataset?

Ans: Right-Skewed median is towards the left side it is not normal distribution.

(iii) If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Ans: In that 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 normal distributed.

3.



Answer the following three questions based on the histogram above.

(i) Where would the mode of this dataset lie?

Ans: The mode of this data set lie in between 5 to 10 and approximately between 4 to 8 .

(ii) Comment on the skewness of the dataset.

Ans: Right-Skewed. Mean>Median>Mode

(iii) Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

Ans: They both are right-skewed and both have outliers the median can be easily visualized in box plot where as in histogram mode is more visible.

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: 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

$P(x) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$

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

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

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

(i) What is the most likely monetary outcome of the business venture?

Ans: 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.

(ii) Is the venture likely to be successful? Explain

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 is a good 80% chances for this venture to be making a profit.

(iii) What is the long-term average earning of business ventures of this kind? Explain

Ans: The long-term average is Expected value = Sum (X * P(X)) = 800$ which means on an average the returns will be + 800$.

(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure

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

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?

   A. 0.3875
   B. 0.2676
   C. 0.5
   D. 0.6987

Ans: To determine the probability that the service manager cannot meet his commitment, we need to calculate the probability that the service time exceeds 50 minutes (since the customer is told the car will be ready in 1 hour and work begins 10 minutes after the car is dropped off).

Given: μ (mean) = 45 minutes

σ (standard deviation) = 8 minutes

We need to find:

P(X > 50 minutes)

To find this probability, we need to calculate the z-score for 50 minutes:

$$Z=\frac{X-\mu}{\sigma}$$

$$Z=50-45 / 8$$

$$z=0.625$$

Now, using the standard normal distribution table (or a z-table), we can look up the area to the left of z = 0.625. From the table:

$$P(Z \leq 0.625) \approx 0.7343$$

To find the probability that X > 50, we need to subtract this from 1:

$$P(X > 50) = 1 - P(Z \leq 0.625) = 1 - 0.7343 = 0.2657$$

Rounding this, we get approximately 0.2657. The closest option is:

B. 0.2676

Therefore, the answer is option B.

2. The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean $\mu = 38$ and Standard deviation $\sigma = 6$. For each statement below, please specify True/False. If false, briefly explain why.
    A. More employees at the processing center are older than 44 than between 38 and 44.
    B. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Ans: To answer these questions, we'll have to look at properties of the normal distribution and use the given mean (μ) and standard deviation (σ) to make our determinations.

A. More employees at the processing center are older than 44 than between 38 and 44.

Given a normal distribution:

About 68% of the data falls within one standard deviation of the mean.

Therefore, 68% of the employees' ages fall between μ−σ and μ+σ, which is between 32 (38-6) and 44 (38+6).

Of that 68%, half (34%) will fall between 38 and 44.

On the other hand, since 50% of the data is greater than the mean (which is 38), the percentage of employees older than 44 is 50% - 34% = 16%.

Since 34% (between 38 and 44) is greater than 16% (older than 44), the statement is False.

B. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

To determine the number of employees under the age of 30, we need to find the percentage of employees whose ages fall more than 38−306638−30 standard deviations below the mean, which is 1.33 standard deviations below the mean.

Using a z-table (or other statistical software), we can find that the percentage of values less than -1.33 standard deviations below the mean is about 9.1%.

Thus, the expected number of employees under the age of 30 is 0.091×4000.091×400 = 36.4, which is about 36 when rounded.

This statement is True.

3. If $X_1 \sim N(\mu, \sigma^2)$ and $X_2 \sim N(\mu, \sigma^2)$ 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: $2X_1$

$X_1 + X_2$

For each expression, we'll determine the mean (expected value) and variance, and then discuss their distributions.

1. For 2X1:

Mean of 2X1: $E[2X_1] = 2E[X_1] = 2\mu$

Variance of 2X1: $Var(2X_1) = 4Var(X_1) = 4\sigma^2$

Thus, $2X1 \sim N(2\mu, 4\sigma^2)$.

2. For X1+X2:

Mean of 2X1+X2: $E[X_1 + X_2] = E[X_1] + E[X_2] = \mu + \mu = 2\mu$

Variance of 2X1+X2: Since X_1 and X_2 are independent, $Var(X_1 + X_2) = Var(X_1) + Var(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2$

Thus, $2X1+X2 \sim N(2\mu, 2\sigma^2)$.

Comparison:

Both 2X1 and 2X1+X2 have the same expected value or mean, which is 2μ.

However, their variances differ. 2X1 has a variance of $4\sigma^2$, while 2X1+X2 has a variance of $2\sigma^2$. This means 2X1 will, on average, deviate more from its mean than 2X1+X2.

In terms of distribution: Both are normally distributed. 2X1 follows a normal distribution with a wider spread or more variability due to its higher variance, while 2X1+X2 will be more concentrated around the mean.

In summary, while both expressions have the same mean, the spread around the mean is greater for 2X1 than for 2X1+X2.

4. Let X ~ N(100, 20^2). 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.

A. 90.5, 105.9
B. 80.2, 119.8
C. 22, 78
D. 48.5, 151.5
E. 90.1, 109.9

To find the values a and b symmetric about the mean such that (a<X<b)=0.99, we will need to use properties of the standard normal distribution and the Z-score formula.

Given: X ~ $(100, 202^2)$

N(100,2022), so μ=100 and σ=202.

The Z-score formula is: Z=X−μ / σ

To find the Z-scores that correspond to the tail areas such that the probability in the center is 0.99, we need to find the Z-scores that leave 0.005 in each tail because 1−0.99=0.01 and 0.01 / 2=0.005.

Using a Z-table or other statistical software, a Z-score of approximately −2.576−2.576 corresponds to the 0.005 percentile, and a Z-score of approximately 2.5762.576 corresponds to the 0.995 percentile. (Since it's symmetric about the mean, the positive Z-score will have the same magnitude as the negative one.)

Plugging these values into the Z-score formula gives:

Fora:

−2.576=a-100/202

a-100= -2.579x202

a-100= -520.352

a= -420.352

But, this is not a reasonable answer given our options. So, there may have been a mistake.

Given that the variance is 20222022, the standard deviation $\sigma = 202$. But, if the number "202" was instead meant to be the standard deviation and not the variance, then our calculations will be straightforward. In that case, we'll have:

Fora:

−2.576=b-100/202

a-100=2.576x202

a=100-520.352

a~80.2

For b :

2.576=b-100/202

b-100=2.576x202

b~119.8

So, the answer is:

a~80.2

b~119.8

The correct option is:

B. 80.2, 119.8

5. Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions $\text{Profit}_1 \sim N(5, 3^2)$ and $\text{Profit}_2 \sim N(7, 4^2)$ 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

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

Ans:

Given:

Profit1$\sim$N($\mu 1,\sigma^2$) where $\mu 1$=5 million dollars and σ^2=32

Profit2$\sim$N($\mu 2,\sigma^2$) where $\mu 2$=7 million dollars and σ^2=42 (so $\sigma 2$=42)

Exchange rate: $1 = Rs. 45

A) When we sum two independent normal random variables, the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

Mean of total profit:

μ total=$\mu 1+\mu 2$=5+7=12 million dollars

Variance of total profit:

σtotal2=$\sigma 12+\sigma 22$=32+42=74

Standard deviation:

σ total=74

Now, for a normal distribution, approximately 95% of the data lies within 1.96 standard deviations of the mean. So, the range in dollars that contains 95% probability is:

μtotal$\pm$1.96σtotal

$12\pm 1.96\sqrt{74}$

$12\pm 1.96(8.6)$

12 ± 16.856

This gives a range of (-4.856, 28.856) million dollars.

Convert this range to Rupees:

Range in rupees:45×(−4.856,28.856)45×(−4.856,28.856) (−218.52,1298.52)(−218.52,1298.52) million rupees

B) The 5th percentile of profit is the value below which 5% of the observations fall.

For a standard normal distribution (mean = 0 and standard deviation = 1), the z-score corresponding to the 5th percentile is roughly -1.645.

So, the 5th percentile in dollars for the total profit is:

μtotal+z×σtotal 12−1.645×8.6=12−14.137=−2.13712−1.645×8.6=12−14.137=−2.137 million dollars

Convert this to Rupees:

5th percentile in rupees = 45×(−2.137)45×(−2.137) = −96.165−96.165 million rupees

C) A loss occurs when the profit is less than zero. To determine which division has a higher probability of making a loss, we'll find the z-scores for a profit of zero for both divisions and then use the standard normal distribution to get the probabilities.

For division 1:

$Z1=0-\mu1 / \sigma1 = 0-5/\sqrt{32}$

For division 2:

$Z2=0- \mu2 / \sigma2 =0-7/\sqrt{42}$

The smaller the z-score, the higher the probability of getting a value below zero (i.e., a loss). So, the division with the smaller z-score has a higher probability of making a loss.

Calculating the z-scores:

z1≈−0.88

z2≈−1.08

Since z2 is smaller than z1, division 2 has a larger probability of making a loss in a given year.

Topics: Confidence Intervals

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

    I. The sample size of the survey should at least be a fixed percentage of the population size in order to produce representative results.

    II. The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.

    III. Larger surveys convey a more accurate impression of the population than smaller surveys.

Ans:

I. False. The sample size of a survey does not necessarily need to be a fixed percentage of the population size to produce representative results. What's more important is that the sample is randomly selected and sufficiently large to decrease the margin of error to an acceptable level. In many cases, as the population grows larger, the required sample size levels off, meaning you don't necessarily need to keep increasing the sample size proportionally to maintain a certain level of precision.

II. False. The sampling frame is a list of all the items (or people) from which the sample is drawn. It's essentially the source or the "universe" from which you are pulling your sample. It does not only list those items that appear in the survey sample, and it doesn't specifically denote non-respondents.

III. True, with qualifications. In general, larger surveys can provide more accurate impressions of a population because they typically have smaller margins of error. However, the accuracy of a survey doesn't just depend on its size. It also depends on how the sample is selected and whether it's representative of the entire population. For instance, a smaller but well-designed random sample can be more accurate than a larger but poorly selected sample.

2. *PC Magazine* asked all of 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:

A. The population: The population refers to the entire group about which information is desired. In this case, it would be all PC Magazine readers who have opinions or experiences with different brands of electronics.

B. The parameter of interest: The parameter of interest is the specific numerical value or characteristic that summarizes or describes an aspect of the population. Here, it would be the true average (or mean) satisfaction rating of Kodak compact digital cameras on a scale of 1 to 10 among all PC Magazine readers.

C. The sampling frame: The sampling frame is the list or set from which the sample is drawn. It should ideally represent the entire population. In this context, the sampling frame would be the list of all PC Magazine readers who participated in the 2004 survey.

D. The sample size: The sample size is the number of observations or units in the sample. Here, the sample size for the Kodak compact digital camera is 225 readers.

E. The sampling design: The sampling design refers to the method used to select the sample from the population. From the given information, we cannot determine the exact sampling design used by PC Magazine. Common methods include simple random sampling, stratified sampling, cluster sampling, and others. More details would be needed to identify the specific design.

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

Self-selection bias: Since the survey asked all of its readers to participate, those who chose to respond might have stronger feelings or opinions about certain products than those who chose not to respond.

Non-response bias: Not all readers participated in the survey, so the opinions of those who did not participate might differ from those who did.

Memory bias: Readers might not remember their experiences accurately, especially if they had used the product a long time ago.

Scale interpretation: Some readers might interpret the 1-10 scale differently. For one person, a rating of 7 might be considered good, while for another, it might be average.

Brand bias: Some respondents might have a general preference or bias towards or against Kodak or other brands, which could affect their ratings.
Lack of representativeness: If the magazine's readership is not a good representation of all users of electronics, the results might not generalize well beyond the readership.

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:
I. True. The 95% confidence interval for the average purchase is given as $50 to $110. This means that we are 95% confident that the true population mean for the average purchase falls within this interval. Since $100 is within this range, it is a plausible value for the population mean at this level of confidence.
II. True. The 95% confidence interval for the number of moviegoers who purchase concessions is given as 30% to 45%. This means we are 95% confident that the true proportion of moviegoers who purchase concessions falls within this range. The upper bound of this interval is 45%, which is less than half. Therefore, at this level of confidence, fewer than half of all moviegoers purchase concessions.
III. False. The statement is not entirely accurate. While it's true that many methods for constructing confidence intervals (like the t-interval method) assume that the sample data are approximately normally distributed, there are several other methods and techniques that can be used when this assumption is not met. The key is that when constructing a 95% Confidence Interval for μ, the appropriate method and assumptions need to be considered. For large sample sizes, the Central Limit Theorem often ensures that the sample mean is approximately normally distributed, even if the individual data points are not.

4. 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:

Given:

Null hypothesis (H0): p=0.05 (Mozilla has a 5% share)

Alternative hypothesis (H1): p<0.05 (Mozilla has less than a 5% share)

Sample proportion (p^) = 4.6% = 0.046

n = 2,000

Standard error (SE) for the sample proportion = $\sqrt{p(1-p)/n}$

Using p=0.05,SE=$\sqrt{0.05(0.95)/2000}$

To test the null hypothesis at a significance level of, say, 0.05, we would calculate the z-score using:

$\wedge$

Z=P-P/SE

The resultant z-score would tell us how many standard deviations the sample proportion is away from the null hypothesis proportion. Using the z-score, we can determine the p-value. If the p-value is less than 0.05, we would reject the null hypothesis.

However, without conducting the full hypothesis test, we can't definitively say whether Microsoft can conclude that Mozilla has a less than 5% share based solely on a 4.6% sample proportion from 2,000 users.

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?

If WebSideStory's sample genuinely includes all the daily Internet users, then it's not really a sample – it's a census. If the data is accurate and truly represents all daily Internet users, then Microsoft can conclude that Mozilla has a 4.6% share of the market among daily users.

However, there are a few caveats to consider:
"Daily Internet users" might not be representative of all Internet users.
Even if it's a census of daily users, there could still be measurement errors or biases in how the data was collected or reported.
Given the information provided, if take WebSideStory's claim at face value and assume it's accurate, then yes, Microsoft can conclude that among daily Internet users, Mozilla has a 4.6% market share, which is less than 5%.

5. 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?

   A. All shipments are between 205 and 295 books.

   B. 95% of shipments are between 205 and 295 books.

   C. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.

   D. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.

   E. We can be 95% confident that the range 160 to 340 holds the population mean.

Ans:

A. All shipments are between 205 and 295 books.
Incorrect. The confidence interval does not claim that all values (shipments) fall within this range. Instead, it speaks about our confidence regarding the population mean falling within this range.
B. 95% of shipments are between 205 and 295 books.
Incorrect. The confidence interval does not describe the proportion of individual shipments in this range. It is about the population mean.
C. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.
Correct. This interpretation captures the essence of a confidence interval. If we were to repeatedly sample and calculate the 95% confidence interval for each sample, we expect about 95% of those intervals to contain the population mean.

D. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.
Incorrect. The confidence interval doesn't provide a prediction about the mean of any future sample. It speaks about our confidence in the population mean based on the current sample. Individual sample means can vary.
E. We can be 95% confident that the range 160 to 340 holds the population mean.
Incorrect. The provided confidence interval was 250 ± 45 books, which translates to a range of 205 to 295 books, not 160 to 340.
Out of the given options, only interpretation C is correct.

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

   A. The z-interval is shorter
   B. The t-interval is shorter
   C. Both are equal
   D. We cannot say

Ans: When constructing confidence intervals for the population mean μ:
A z-interval is used when the population standard deviation (σ) is known.
A t-interval is used when the population standard deviation (σ) is unknown, and we use the sample standard deviation (s) instead.
The formula for the z-interval is:
$\bar{X} \pm z(\sigma/\sqrt{n})$
The formula for the t-interval is:
$\bar{X} \pm t(s/\sqrt{n})$
Given that σ = s, we can compare the two intervals by looking at their multipliers, z and t.
For a 95% confidence interval, the z-value (from the standard normal distribution) will typically be around 1.96, depending on the exact definition used for 95%.
The t-value (from the t-distribution) depends on the sample size (or degrees of freedom), but for a 95% confidence interval and typical sample sizes, it will be greater than 1.96. This is because the t-distribution is wider and has fatter tails than the standard normal distribution, especially with smaller sample sizes. As the sample size increases, the t-distribution approaches the standard normal distribution and the t-value will approach 1.96.

Given the information, the t-multiplier will be larger than the z-multiplier for a 95% confidence interval, which means the t-interval will be wider (or longer) than the z-interval.

So, the correct answer is:

B. The t-interval is shorter.

(The t-interval is actually longer or wider than the z-interval given the same sample standard deviation and population standard deviation being equal.)

CBA: Practice Problem Set 2
Topics: Sampling Distributions and Central Limit Theorem

1. 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.

   (i) 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.

   (ii) The standard error of the daily average SE($\bar{x}$) = 1.

Ans:

Explanation: While it's ideal for individual observations (weights of individual packages in this case) to be normally distributed, it's not always necessary when dealing with sample means. Thanks to the Central Limit Theorem (CLT), if the sample size is large enough (typically n > 30 is a common rule of thumb, though smaller samples like n > 15 can be adequate if the data is not too skewed or has no major outliers), the sampling distribution of the sample mean will be approximately normal, regardless of the distribution of the individual observations.

In this case, a sample size of 25 packages is close to the threshold where the CLT begins to apply. While it's on the borderline, the manager would ideally want to check the shape of the distribution of individual weights. If it's not highly skewed or there aren't major outliers, the sampling distribution of the sample mean will still be approximately normal.

The standard error of the daily average SE ($\bar{x}$) = 1.

True/False? False.

Explanation: The standard error (SE) of the sampling distribution of the sample mean is calculated using the formula:

SE ($\bar{x}$) = \frac{σ}{\sqrt{n}}

Where:

σ is the standard deviation of the population.

n is the sample size.

Given: σ = 5 lbs.

n = 25

Plugging these values into the formula, we get:

SE ($\bar{x}$) = \frac{5}{\sqrt{25}} SE($\bar{x}$) = \frac{5}{5} SE($\bar{x}$) = 1

So, the statement is actually True.

2. 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: To solve this problem, we will utilize the Central Limit Theorem, which tells us that the sampling distribution of the sample mean is approximately normally distributed when the sample size is large, regardless of the shape of the population distribution.

Given:

Population mean, μ = $50

Population standard deviation, σ = $40

Sample size, n = 100

The standard error (SE) of the sample mean can be calculated as:

SE = $\sigma/\sqrt{n}$

SE = $400/\sqrt{100}$

SE = 40/10

SE = 4

Next, we'll find the z-scores associated with the lower and upper limits of the sample mean range where there would be no investigation:

For $45:

Z1 = X1-μ /SE

Z1=45-50/4

Z1= -5/4

Z1= -1.25

For $55:

Z2= X2-μ/SE

Z2=55-50/4

Z2=5/4

Z2=1.25

We want to know the probability that the mean transaction amount is outside the range of $45 and $55 (i.e., P(X < $45) or P(X > $55)).

Using a z-table:

P(Z < -1.25) = 0.1056 P(Z > 1.25) = 1 - P(Z < 1.25) = 1 - 0.8944 = 0.1056 (because of the symmetry of the standard normal distribution).

So, the probability that there will be an investigation (i.e., the sample mean is outside the range $45 and $55) is:

P(X < $45) + P(X > $55) = 0.1056 + 0.1056 = 0.2112

P=21.12%

Thus, the correct answer is: D. 21.1%

3.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.

A. 144
B. 150
C. 196
D. 250
E. Not enough information

To solve the problem, we'll need to use the formula for the sample size required for estimating a proportion in a population:

n = $A = Z^2$x p x(1-p)/ E^2

Where:

n is the sample size

Z is the Z-value from the standard normal distribution corresponding to a desired confidence level. For a two-tailed test with 5% in the tails, we'd have a 95% confidence level, so Z would be approximately 1.96.

p is the estimated proportion. Since the thresholds are 45 and 55, the midpoint (i.e., the estimated proportion for the population) is 50%, or 0.50.

E is the margin of error. This is half the distance between the threshold and the midpoint. In this case, 55% - 50% = 5% or 0.05.

Plugging the values into the formula:

n= (1.96^2) x 0.50 x 0.50 / 0.05^2

n= 3.8416x0.25/0.0025

n= 0.9604/0.0025

n= 384.16

However, you can't sample a fraction of a transaction, so you'd round up to the nearest whole number. Therefore, they'd need to sample at least 385 transactions based on this calculation.

The options provided do not include this result, so the answer should be:

E. Not enough information.

4. 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?
    A. The standard deviation of the scores within any sample will be 120.
    B. The standard deviation of the mean of across several samples will be 120.
    C. The mean score in any sample will be 720.
    D. The average of the mean across several samples will be 720.
    E. The standard deviation of the mean across several samples will be 0.60

Ans: Given the data, let's analyze each option:

A. The standard deviation of the scores within any sample will be 120.

This statement is not necessarily true. While the population has a standard deviation of 120, the standard deviation of any given sample can vary depending on the size and composition of the sample.

B. The standard deviation of the mean of across several samples will be 120.

This statement is not true. The standard deviation of the mean of a sample (also known as the standard error) is given by the formula: $\sigma\bar{x}=/\sigma/\sqrt{n}$

where σ is the standard deviation of the population and n is the sample size. The standard deviation of the mean will be less than the population standard deviation unless n=1.

C. The mean score in any sample will be 720.

This statement is not necessarily true for any random sample. While the population mean is 720, the mean of any given sample can vary. However, if the sample size is large, the Central Limit Theorem suggests that the sample mean will tend to be close to 720.

D. The average of the mean across several samples will be 720.

This statement is likely to be true. If we take many samples and calculate their means, then average those means, the result should approach the population mean due to the law of large numbers. So, the average of the sample means should converge to the population mean of 720.

E. The standard deviation of the mean across several samples will be 0.60

As mentioned in option B, the standard deviation of the mean (standard error) is given by $\sigma_{\bar{x}}=\sigma/\sqrt{n}$. To check if this is true, we need to know the sample size. Without knowing the sample size, we cannot definitively say whether the standard error is 0.60 or not.

In summary, out of the options provided, the most likely true statement is: D. The average of the mean across several samples will be 720.