**KANAKALA TARUN KUMAR**

[**Kanakalatarunkumar0212@gmail.com**](mailto:Kanakalatarunkumar0212@gmail.com)

**Assignment -2**

**Topics: Descriptive Statistics and Probability**

1. **Look at the data given below. Plot the data, find the outliers and find out**

|  |  |
| --- | --- |
| **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%** |

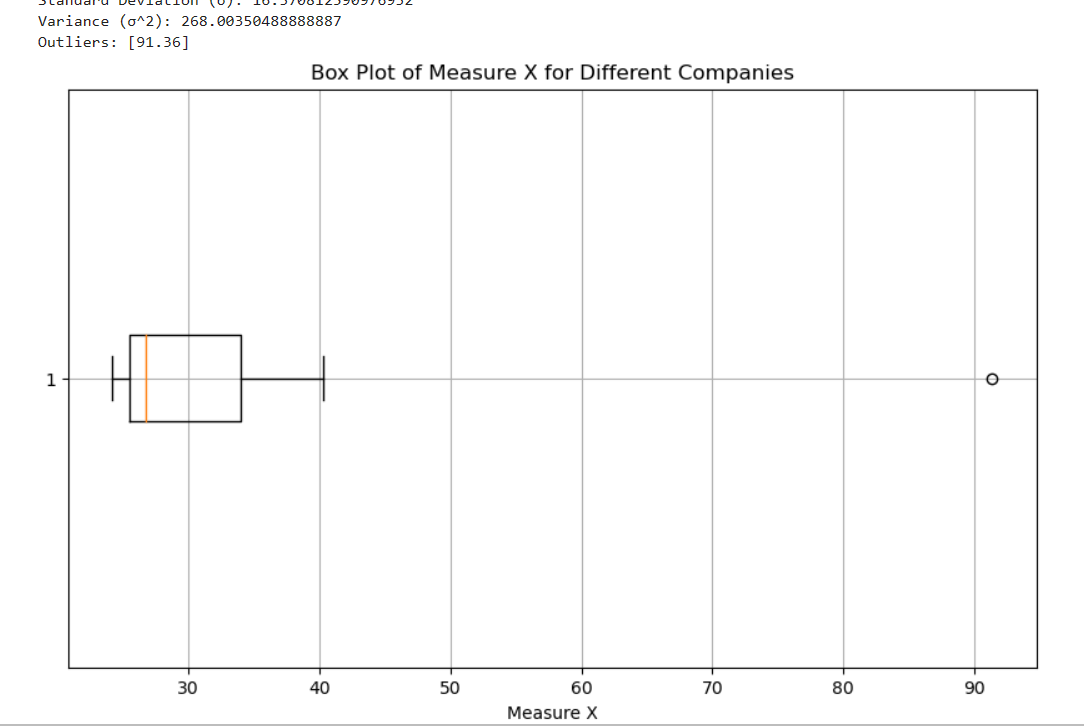
**Ans:**

Mean (μ): 33.27133333333333

Standard Deviation (σ): 16.370812590976932

Variance (σ^2): 268.00350488888887

Outliers: [91.36]





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

**(i) What is the interquartile range of this dataset, and what does it imply?**

Ans: The interquartile range (IQR) of the given dataset is approximately 5.67. This value represents the spread of the middle 50% of the data and does not consider outliers or extreme values. Therefore, it provides a robust measure of variability in the dataset.

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

Ans: The dataset exhibits positive skewness because the median is less than the mean. This indicates that the distribution is right-skewed, with extreme values and outliers predominantly on the right side of the graph.

**(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: If the data point with the value 25 is corrected to 2.5, it would not significantly affect the box plot because 2.5 would still be within the lower fence of the boxplot. The shift in the mean would be relatively small, and the overall shape and characteristics of the box plot would remain largely unchanged.



**Answer the following three questions based on the histogram above.**

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

**Ans:** The mode of the dataset is approximately between 4 and 8, where the frequency is greater than 20.

**(ii) Comment on the skewness of the dataset.**

**Ans:** The dataset exhibits positive skewness, indicating a right-skewed, right-tailed distribution. There are more positive values towards the right of the distribution.

**(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:** Both the histogram and box plot complement each other in providing a comprehensive understanding of the dataset. The histogram provides a visual representation of the data's distribution, helping to grasp its probability distribution and the frequency of values within different ranges. On the other hand, the box plot summarizes key statistical measures such as the median, quartiles, and potential outliers. Together, these graphs offer insights into the central tendency, spread, and presence of outliers in the dataset, making it easier to assess its overall characteristics and compare it with other datasets.

1. **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.)**

The probability of the event E is

P(E)= 1/200

Probability of having at least one successful call will be

1-P(X)= 1-1/200= 199/200= 0.967

As every event is independent of other event the probability will be

1- (0.967) ^5

0.02475 = 2% chance.

1. **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?**

**Ans:** 2000$ have most change of occurrence.

1. **Is the venture likely to be successful? Explain**

**Ans:** Yes, it is, because 60% chance to give results and 20% change of failure. And 20% chance to no loss and no profit

1. **What is the long-term average earning of business ventures of this kind? Explain**

**Ans:**

(-2000\*0.1) +(-1000\*0.1) +(0\*0.2) +(1000\*0.2) +(2000\*0.3) +(3000\*0.1) =800

1. **What is the good measure of the risk involved in a venture of this kind? Compute this measure**

**Ans:** We can calculate the risk by using the variance and standard deviation.

Variance = 3500000

Standard deviation = 1870.83

This venture at high risk because the standard deviation is $1870 with the average return of $800

**SET -2**

**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?**
2. 0.3875
3. **0.2676**
4. 0.5
5. 0.6987

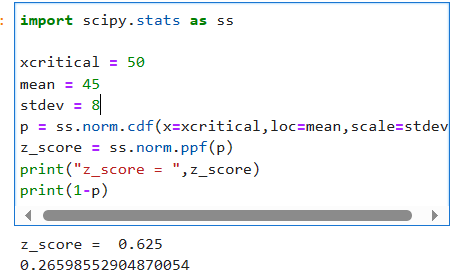
Let the prob of not meeting commitment be P(E).

We have to calculate the z-score first for the given scenario

Given : μ = 45 , σ = 8 , time = 60 – 10 = 50 Minutes

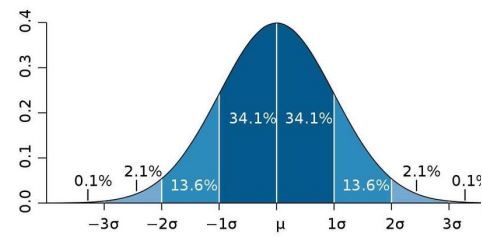
Z-Score at 50 => (time – mean time)/std dev => (50-45)/8 = 0.625

Corresponding probability from Z-table = 0.7324 P(E) = 1 – 0.7324 = 0.2676 (Answer = Option B)



1. **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.**
2. **More employees at the processing center are older than 44 than between 38 and 44.From Normal Distribution Graph.**

The range between ages 38 and 44 is within one standard deviation from the mean. This means that it contains about 34.1% of 400 approx. = 136 people. Going beyond age 44 will result in about 16% approx. = 64 people which is less than the former. Therefore, the answer is False.



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

Z=(X-µ)/ *σ*

P(X≤30) =p (Z≤ (30-38)/6) =p(Z≤-1.33) = 0.0918

Expected count=0.0918\*400= 36.72

1. **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:** 1. 2\*X1:

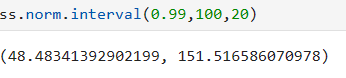
- Distribution: Normal (μ' = 2μ, σ'^2 = 4σ^2).

2. X1 + X2:

- Distribution: Normal (μ'' = 2μ, σ''^2 = 2σ^2).

Both are normal distributions with the same mean (2μ), but the variance differs: 2\*X1 has a larger variance (4σ^2) compared to X1 + X2 (2σ^2).

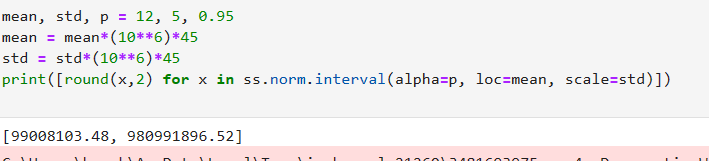
1. **Let X ~ N(100, 202). Find two values, *a* and *b*, symmetric about the mean, such that the probability of the random variable taking a value between them is 0.99.**
2. 90.5, 105.9
3. 80.2, 119.8
4. 22, 78
5. **48.5, 151.5**
6. 90.1, 109.9



Answer D

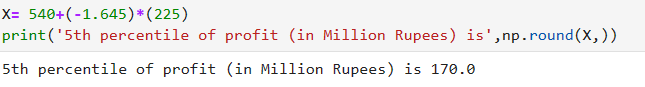
1. **Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions Profit1 ~ N(5, 32) and Profit2 ~ N(7, 42) respectively. Both the profits are in $ Million. Answer the following questions about the total profit of the company in Rupees. Assume that $1 = Rs. 45**
2. **Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.**

Range is Rs (99008103.48, 980991896.52) in Millions



1. **Specify the 5th percentile of profit (in Rupees) for the company**

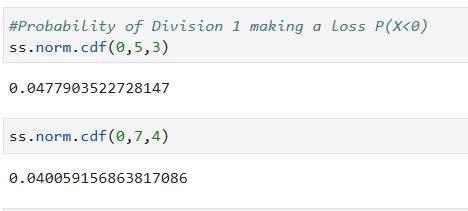
5th percentile of profit (in Million Rupees) is 170.0



1. **Which of the two divisions has a larger probability of making a loss in a given year?**

0.0477903522728147

0.040059156863817086



**SET -3**

**Topics: Confidence Intervals**

1. **For each of the following statements, indicate whether it is True/False. If false, explain why.**
2. **The sample size of the survey should at least be a fixed percentage of the population size in order to produce representative results.**

Ans: False.

The result depend on the size(n) of the sample. The sample size should have at least 30 observation.

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

Ans: True

The sampling frame is a list or database that includes all the items or elements that are eligible to be selected in a survey. This includes not only those who responded to the survey questions but also those who were contacted or eligible to be contacted but did not respond. A complete sampling frame is essential for estimating response rates, calculating biases, and ensuring the representativeness of the sample.

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

True: Generally, larger surveys tend to provide a more accurate impression of the population than smaller surveys, assuming that the larger survey maintains a random or representative sampling method. A larger sample size reduces the margin of error and increases the precision of estimates, making it more likely to capture the true characteristics of the population. However, the cost and practical considerations associated with larger surveys may limit their feasibility in some cases, so the sample size should be chosen carefully to balance accuracy with resource constraints.

1. ***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:**
2. **The population**

Ans: All the readers of the PC magazine

1. **The parameter of interest**

Ans: The population mean that rated the digital camera

1. **The sampling frame**

Ans: 9000

1. **The sample size**

Ans: 225

1. **The sampling design**

Ans: Sampling Design = n/N

Where n – number of units to be samples

N – number of units in total population

Sampling Design = 225/9000 = 0.025 (Answer)

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

Ans: Selection of the readers, Selection of the issue which will contain the survey

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

Ans: True, Confidence interval identifies the collection of values for the population parameter that are consistent with the observed sample

1. **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**.

Ans: False,

We have evidence in that direction but we cannot confirm 100% based on this data we have to consider the values out this range

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

Ans: False

We should have a moderately large sample (usually at least large than 30 for many cases the central limit theorem implies that the sampling distribution is normal regardless of the data itself

1. **What are the chances that ?**
2. ¼
3. **½**
4. ¾
5. 1

The is pure assumption . there is a 50 % chance that the sample mean(X) is greater than the population mean.

1. **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.**
2. **If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?**

ANS: NO

1. **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: YES

1. **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?**
2. **All shipments are between 205 and 295 books.**

Ans: False

1. **95% of shipments are between 205 and 295 books.**

Ans: TRUE

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

Ans: TRUE

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

Ans: TRUE

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

Ans: False

1. **Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ =s?**
2. **The z-interval is shorter**
3. The t-interval is shorter
4. Both are equal
5. We cannot say

Ans : 95% z-interval for μ is shorter than a 95% t-interval for μ when the sample standard deviation (s) is equal to the population standard deviation (σ). This is because the t-distribution has heavier tails compared to the standard normal (z) distribution.

In situations where σ = s, the t-distribution effectively becomes a standard normal distribution, but the critical values for the t-distribution are slightly larger (i.e., the tails are slightly heavier). As a result, you need a slightly larger margin of error (and therefore a wider interval) with a t-interval compared to a z-interval to achieve the same level of confidence.

Answer is (A)

Questions 8 and 9 are based on the following: To prepare a report on the economy, analysts need to estimate the percentage of businesses that plan to hire additional employees in the next 60 days.

1. **How many randomly selected employers (minimum number) must we contact in order to guarantee a margin of error of no more than 4% (at 95% confidence)?**
2. **600**
3. 400
4. 550
5. 1000

We are required to find the value of n in order to create an estimate where we are 95% confident

with a margin or error 4%.

Margin of error is given by

-----------------------(1)

Let n represent the size of the sample.

Let p^ be the sample proportion and q^ = 1-p^ Since value of sample proportion p^ has not been given and then we can take p^ =0.5 (which implies q^ = 1 -0.5 = 0.5 and

p^ x (1−p^) = 0.25), because this will result in the largest possible .sample size and this will largest possible sample size will be appropriate for all sample proportions.

z-score corresponding to 95% = 1.96

Margin of Error, M.E = 0.04

Calculating n using eq(1)

n = 0.25/((M.E/z)^2) = 600.25 ~ approx. 600 (Ans0.25/((M.E/z)^2) )

Answer: A

1. **Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?**
2. 1000
3. 757
4. **848**
5. 543

z-score corresponding to 98% = 2.33 , M.E = 0.04 n = 0.25/((M.E/z)^2) = 848.265 ~ approx. 848

**CBA: Practice Problem Set 2**

**Topics: Sampling Distributions and Central Limit Theorem**

1. **Examine the following normal Quantile plots carefully. Which of these plots indicates that the data …**
2. **Are nearly normal?**

Ans: C

1. **Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.)**

Ans: D

1. **Are skewed (i.e. not symmetric) ?**

Ans: A

1. **Have outliers on both sides of the center?**

Ans: B



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.**

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.**

Ans:True:

The normal model requires either a normal population or a large sample size (n > 30).

1. **The standard error of the daily average SE() = 1.**

Ans :False

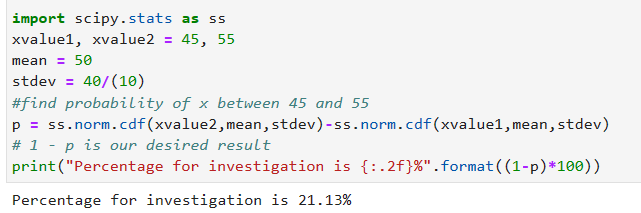
Standard error (SE) is σ/√n, not always 1. Here, SE = 5/√25 = 1 lb, but it varies with σ and n.

1. **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?**
2. 1.25%
3. 2.5%
4. 10.55%
5. **21.1%**
6. 50%

It is given that n = 100, Mean = 50, std dev (pop) = 40, std dev(sample) = 40/sqrt(100)

We need to calculate z scores and find probability of x in range(45,55)

P(45<x<55)



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

ANS:250 We are required to find the n value.

Mean and Population std dev remains the same.

P(E) = 0.05 => 0.05/2 => 1-0.025 = 0.975 [since it is two sided distribution]

Corresponding z-score = 1.644 (value – Mean)/std\_sample = 1.644

Std\_sample = (55-50)/1.64 = 3.041

Std\_sample = Std\_population/root(n)

n = (std\_population/std\_sample)^2

n = 15.68^2 = 246.05~ approx. 250

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

The standard deviation of scores within any sample (option A) is not necessarily 120. The standard deviation of scores within a sample depends on the size of the sample and the distribution of scores within that specific sample. In this case, the sample standard deviation could be smaller or larger than 120 for any randomly chosen sample.

The standard deviation of the mean across several samples (option B) is a different concept. It refers to the standard error of the mean (SEM) and is given by the formula:

SEM = σ / √n

Where:

σ is the population standard deviation (given as 120).

n is the sample size.

So, the standard deviation of the mean (SEM) across several samples will depend on the sample size. It will be smaller as the sample size (n) increases.

The mean score in any sample (option C) is not necessarily 720. It can vary from sample to sample depending on the random selection of individuals.

The average of the mean across several samples (option D) is likely to be close to 720 if the samples are random and unbiased. This is because the law of large numbers suggests that the average of sample means from multiple random samples will converge to the population mean.

The standard deviation of the mean across several samples (option E) is not 0.60. It depends on the sample size, as mentioned earlier. The SEM is calculated as σ / √n, so it will depend on the square root of the sample size, not a fixed value like 0.60.

So, the most likely true statement among the given options is:

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