**SET 1**

**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% |
|  |  |

* The problem is being solved in python
* Answer is as follows:

1. Outlier : Morgan Stanley - 91.36%
2. Mean : 33.2713
3. Median : 26.71
4. Standard Deviation : 16.3708
5. Variance : 268.0035



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

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

Ans: it is being assumed that Lowest value is 0, highest value 19, Q1 is 5, Q2 is 7, Q3 is 12.

Hence, Inter-quartile range(IQR) is:

IQR = Q3 – Q1 = 12 – 5 = 7

1. What can we say about the skewness of this dataset?

Ans: The data is skewed to the right. More data points are present towards the lower end of the range.

1. 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: The boxplot could shift towards lower end of the data range. Q1 and Median could shift at lower part of the range.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Ans: Mode would be between 4 to 8 value of Y.

1. Comment on the skewness of the dataset.

Ans: The data is positively skewed. It means there are more datapoints towards the lower end of the distribution.It is not symmetrical.

1. 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: The box plot shows the total range of the data which the histogram cannot show. The box plot shows a particular boxed area where 50% or more of the data point is concentrated, but cannot show frequency for each data point. Here histogram is helpful as it can show frequencies for each datapoint. Also, Median can be located and shown in a boxplot, which is not easy in histogram.

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.)

Ans: let P(A) = probability of misdirection = 1/200

P(B) = probability of call reaching at correct number = 199/200

P(C) = Probability of one in five calls being correct (As all calls are independent)

= 199/200 \* 199/200 \* 199/200 \* 199/200 \* 199/200

= 0.9925

P(D) = Probability of one at-least one call getting misdirected

= 1 – 0.9925 = 0.0075

Hence, is the probability that at least one in five attempted telephone calls reaches the wrong number is 0.0075

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: The most likely outcome is of the business earning $2000 with highest probability of 0.3

1. Is the venture likely to be successful? Explain

Ans: Yes, because any business venture can continue to operate as long as its not getting into losses. The probability of this business venture getting into loss is 0.2 out of 1

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

Ans: Expected value of long-term average earning =

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

= $800

Long-term average earning of business ventures = $800

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

Ans: Mean = Long-term average earning of business ventures = $800

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Profit/loss(X) | P(x) | Deviation from mean (X – mean) | Squaring of deviation(x-mean)2 | (x – mean)2 \* P(x) |
| -2000 | 0.1 | -2800 | 7840000 | 784000 |
| -1000 | 0.1 | -2200 | 4840000 | 484000 |
| 0 | 0.2 | -800 | 640000 | 128000 |
| 1000 | 0.2 | 200 | 40000 | 8000 |
| 2000 | 0.3 | 1200 | 1440000 | 432000 |
| 3000 | 0.1 | 2200 | 4840000 | 484000 |

Sum of squared deviations:

=784000 + 484000 + 128000 + 8000 + 432000 + 484000 = 8928000

Standard Deviation: = 2955

Hence, $2955 deviation is the risk involved in a venture of this kind

**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

Ans: 0.2676

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.

Ans: False, more employees are between 38 to 44. The standard deviation is 6, it means most of the values are in the range of 32 to 44. Hence number of employees being above 44 is very less.

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

Ans: False, Most likely false. 36 mployees out of 400 is 9% of total employees. But most of the employees are already between 32 to 44. So to get 36 employees ofr training program may be impossible. New recruits are necessary to have training program under age 30.

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: Meaning of 2X1:

µ, in 2X1 = µ + µ = 2µ

σ2 in 2X1 = σ2 + σ2 = 2σ2

on squaring, it will be 4σ2

Hence 2X1 = N(2µ + 4σ2 )

Meaning of (X1 + X2 ) :

µ, in X1 +X2 = µ + µ = 2µ

σ2 in X1 +X2 = σ2 + σ2 = 2σ2

Hence X1 +X2 = N(2µ + 2σ2 ).

Hence, The difference between 2X1 is and (X1 + X2) is 2σ2

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

According to z-score normal distribution, 99% value lie in standard of 2.33 deviation, means +2.33 and -2.33. Hence we can say that values of a and b are:

a = 100 - 2.33\* 20 = 53

b= 100 + 2.33 \* 20 = 147.

Hence Option D is the answer as it is the closest.

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.
3. Specify the 5th percentile of profit (in Rupees) for the company
4. Which of the two divisions has a larger probability of making a loss in a given year?

Ans:

1. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.

Total mean Profits(x) = 5 + 7 = 12 million dollars

Variance = 32  + 42  = 9 + 16 = 25 Million Dollars

Hence, Standard deviation = 5 Million Dollars

To calculate the range of 95% probability

Let ‘p’ be the lowest point of the range and ‘q’ be the highest point of the profit range.

In a standard distribution, the critical values are approximately +or – 1.96 to contain a 95% range of confidence.

P = 12 – 1.96 \* 5 = 2

Q = 12 + 1.96\* 5 = 22

1. 5th percentile of profit for the company:

Using the Z-score for the 5th percentile (approximately -1.645):

Profit at 5th percentile = Mean - (Z-score \* Standard Deviation)

Profit at 5th percentile = $12 - (1.645 \* 5)

Profit at 5th percentile = $12 - $8.22= $3.78 million

Profit at 5th percentile(in rupees) = $3.78 \* 45 = Approx 170 million rupees

1. Probability of making a loss for each division:

For a normal distribution, Z-score will be used to find the probability.

For Profit1:

Z1 = (0 - 5) / 3 = -1.67

Probability of loss for Profit1 = P(Z < -1.67) = approx.. 0.0475

For Profit2:

Z2 = (0 - 7) / 4 = -1.75

Probability of loss for Profit2 = P(Z < -1.75) = 0.0401

Hence, Hence first type of business is more likely to make a loss.

**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 sample size will depend on population.

False. The sample size should be determined based on statistical considerations such as the desired level of confidence and margin of error. It is not strictly necessary for the sample size to be a fixed percentage of the population size, but it should be large enough to provide meaningful results.

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: False. The sampling frame is a list of all the units (individuals or elements) in the population from which the sample is drawn, not necessarily those that responded. The sampling frame should ideally include all elements from which the sample is to be selected.

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

Ans: True. Larger surveys generally provide a more accurate representation of the population, assuming that the sampling process is random and unbiased. A larger sample size reduces the margin of error and increases the precision of estimates, leading to a more reliable portrayal of the population.

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: randomly chosen 9000 readers of PC magazine.
3. The parameter of interest: satisfaction rating of different brands of electronics, in this case, Kodak Compact digital camera.
4. The sampling frame: readers of PC magazine were chosen as sampling frames.
5. The sample size: 225 out of 9000, 2.5%
6. The sampling design: readers of PC Magazine who participated in the survey and have opinions on electronics brands.
7. Any potential sources of bias or other problems with the survey or sample:

Since participation is voluntary, only readers who are interested in providing feedback may have responded, leading to a potential bias if their opinions differ from those who chose not to participate. The survey may only capture the opinions of PC Magazine readers, which may not be representative of the broader population.

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

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: The 95% confidence interval for the percentage of moviegoers who purchase concessions (30% to 45%) indicates our level of confidence about the range of plausible values for the population proportion. It does not imply that fewer than half of all moviegoers purchase concessions; it provides a range for the population proportion.

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

Ans: False, for a sufficiently large sample size, the sampling distribution of the sample mean will be approximately normally distributed, regardless of the distribution of the population. Therefore, the 95% confidence interval for the population mean can be used even if the sample data are not normally distributed, given a large sample size.

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

Ans: 1/2

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: If the population is up to 40,000 then the sample of 2000 users can be said to be fine. But neither population nor margin of error is known here.

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: Even in this case, Microsoft needs to show for what the Mozilla Forefox browser was used. And needs to be cautious of the error.

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: Incorrect. Confidence level is more like a prediction, which is likely to happen. The shipments could be well below or above 205 and 295 respectively.

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

Ans: Incorrect. Confidence level is more like a prediction, which is likely to happen. The shipments could be well below or above 205 and 295 respectively.

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

Ans: Correct, the range represents highly likely amount of books or inventory, so that amount of inventory could be present.

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. Correct.

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

Ans: Incorrect, confidence level does not represent a range.

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: z-interval is shorter

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

Note: The answer couldn’t be derived without total number of employees and proportion of the new employees to be hired(p value). Hence, it is being assumed that p = 0.5

Following formula was used to solve question number 8 and 9:

Z = 95% confidence = 1.96

p = estimated proportion of employees to be hired = 0.5   
E = Margin of error = 4%

Hence, option A is the answer

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 = 98% confidence = 2.33

p = estimated proportion of employees to be hired(assumed) = 0.5   
E = Margin of error = 4%

n ~ 848

Hence, option C is the answer.

**SET 4**

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



Ans:

1. Are nearly normal? : Plot C
2. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.) Plot B
3. Are skewed (i.e. not symmetric) ? Plot A
4. Have outliers on both sides of the center? Plot D
5. 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:** False. No matter the shape of the population distribution, the sampling distribution of the sample mean shall be nearly normally distributed for sufficiently large sample sizes. As long as the sample size is large enough, the manager can rely on the

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

**Ans:** False**.** The standard error of the sample mean is to be calculated as the standard deviation of the population divided by the square root of the sample size. In formula it is written as: σ/√n. In the above case, the standard deviation of the population (σ) is given as 5 lbs, and the sample size (n) is 25. Therefore, the standard error of the daily average is = 5/√25 = 5/5 = 1 lb, not 1. The statement is false because it incorrectly states the value or unit of the standard error.

1. Top of Form

Note: It is being assumed that 1 means 1 package. If the 1 means 1 lb in the above statement, then the answer shall be True.

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%

**Ans:**

mean (μ) = $50, standard deviation (σ) = $40, n=Sample size = 100

First, we need to find the standard error of the sample mean (SE):

*Std. Error* = *n/ = 40/* = 4

Calculation of z-scores corresponding to the lower and higher limits of the acceptable range:

Z-lower = (45−mean)/std. error = (45−50)/4 = −1.25

Z-higher = (55−mean)/std. error = (55−50)/4= 1.25

Finding probability that the sample mean falls outside this range using the standard normal distribution table/calculator:

*Finding P*(mean<45 or mean>55) means finding *P*(*Z*<−1.25 or *Z*>1.25)

From the standard normal distribution table or calculator, we find:

*P*(*Z*<−1.25) = 0.105

*P*(*Z*>1.25) = 1− *P*(*Z*<1.25) = 1−0.894 =0.106

Therefore, the probability that an investigation will be initiated is:

P(mean<45 or mean>55 ) = *P*(*Z*<−1.25 or *Z*>1.25)

= 0.105 +0.106 = 0.211

In percentage terms, we get approximately 21.1%.

Hence, Option D is the answer

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

previously calculated probability of the sample mean falling outside 45-55 range is approximately 21.1% with sample size of 100 transactions.

Formula to calculate required sample size n, n = (Z\*Std deviation/margin of error)2

Z = 1.96, std. deviation= 40,

margin of error= std. deviation/ = 10

= (1.96\*40/10) = 7.84

n = (1.96 \* 40/7.84)2  = 100

Hence, Sample size should be atleast 100.

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.

Ans: Options A and D are true, while option E is likely to be true subject to conditions.