**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 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



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: 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

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

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

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: 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



Answer the following three questions based on the histogram above.

1. 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 .

1. Comment on the skewness of the dataset.

Ans: Right-Skewed. Mean>Median>Mode

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: 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.

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:  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

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 |

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. 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

1. 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

1. 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$

1. 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 ** = 45 minutes and ** = 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?

b.2676

1. The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean ** = 38 and Standard deviation ** =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.

FALSE , comparing probabilities

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

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.

Sol-

Iid – independent and identical

|  |  |
| --- | --- |
| 2X1 | X1+X2 |
| *2.X1* ~ *N*(2.μ , 4.σ2) | *X1 + X2*  ~*N*(2.μ , 2.σ2) |

Hence there is difference between X1+X2 and 2X1

As we can observe that variance is twice in case 2.X then X+Y

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.

d.48.5, 151.5

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
   1. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company. Sol - 16 to 91 crore rupees
   2. Specify the 5th percentile of profit (in Rupees) for the company Cumulative sum = 12.45cr RS

|  |  |
| --- | --- |
|  | profit(cr) |
| -3 | -13.5 |
| -2 | 9 |
| -1 | 31.5 |
| 0 | 54 |
| 1 | 76.5 |
| 2 | 99 |
| 3 | 121.5 |

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

Total LOSS = 22.5cr Total LOSS = 27 cr

|  |  |
| --- | --- |
| division 1  (CR) | division 2  (CR) |
| -18 | -22.5 |
| -4.5 | -4.5 |
| 9 | 13.5 |
| 22.5 | 31.5 |
| 36 | 49.5 |
| 49.5 | 67.5 |
| 63 | 85.5 |
| 157.5 | 220.5 |

**Topics: Confidence Intervals**

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 in order to produce representative results.

FALSE , larger the sample size more effective is the result for population

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

TRUE

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

TRUE

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:

|  |  |
| --- | --- |
| A. The population | all of readers |
| B. The parameter of interest | satisfaction from different brands |
| C. The sampling frame | 9000 readers |
| D. The sample size | 225 readers |
| E. The sampling design | Rating scale |
| F. bias or other problems | NON-RESPONSIVE BIAS |

1. For each of the following statements, indicate whether it is True/False. If false, explain why.
   1. If the 95% confidence interval for the average purchase of customers at a department store is $50 to $110, then $100 is a plausible value for the population mean at this level of confidence. FALSE, its range is +30 and -30 hence mean 80$
   2. If the 95% confidence interval for the number of moviegoers who purchase concessions(snacks) is 30% to 45%, this means that fewer than half of all movie goers purchase concessions.

TRUE

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

TRUE

1. What are the chances that *X*  ** ?

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

NO , SHARE IS MORE THAN 5% , one tail t test

* 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?

TRUE , as Microsoft claim goes along population mean

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.

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

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

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

1. We can be 95% confident that the range 160 to 340 holds the population mean*.*
2. Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ =s?

### The z-interval is shorter

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

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.

Let standard deviation = 0.5

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)?

### A. 600

B. 400

C. 550

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?

A. 1000

B. 757

### C. 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.) => B and D
2. Are skewed (i.e. not symmetric)? =>A,B and D
3. Have outliers on both sides of the center? => A and 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 are 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

: 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.

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

Ans: True :

As SE(Standard Error) = sample standard deviation / Square root of (number of sample)

SE = 5 / (25)^1/2

SE = 1

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: 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.

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: D

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

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: E

Standard error = sigma / (n)^0.5

= standard deviation / (sample size)^0.5

= 120 / (40000)^0.5

= 0.6