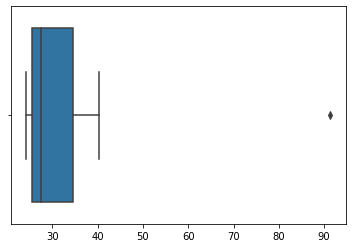
**Topics: Descriptive Statistics and Probability**

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% |

1. The only one out layer are present in above upper extreme or upper limit



1. The mean of the data is (Mu)=33.80
2. The standard deviation of data is (sigma)=17.4541
3. The variance of the data is (sigma\*\*2)=304.646



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.
2. The inter quartile range(IQR) =Q3-Q1

=12-5

=7

1. What can we say about the skewness of this dataset?
2. The Skewness of the dataset is Right Skewness

Because number of observations are present on right side

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?

A)the new box-plot be affected in 0, 0.5, 1.0, 1.5, 2.0, 2.5



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?
2. The mode of the dataset is present in between 5 to 10

Because the no of observations concentrated on 5 to 10

(ii)Comment on the skewness of the dataset.

1. The skewness of the dataset is Right skewness

No of observations are presented on present in Right side

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.

1. The above Histogram and box-plot both Observations presented in the Right side

The histogram plot is used to easy to identify the Skewness and density

The boxplot is used to identify the outlayers,IQR, and Quartile values .

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.)
2. 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)

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

P(1)=0.0245

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?
2. most likely monetary outcome of the business venture  is 2000  $

as it has maximum probability = 0.3

1. Is the venture likely to be successful? Explain

|  |  |  |
| --- | --- | --- |
| E(x) | P(x) | E(X)P(X) |
| -2,000 | 0.1 | -200 |
| -1000 | 0.1 | -100 |
| 0 | 0.2 | 0 |
| 1000 | 0.2 | 200 |
| 2000 | 0.3 | 600 |
| 3000 | 0.1 | 300 |
| totall |  | 800 |

Expected value =  ∑E(X)P(X)  = 800

1. What is the long-term average earning of business ventures of this kind? Explain
2. long-term average earning of business ventures  = 800 $

The long term is calculated by totall earnings

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure
2. Expected value is + ve   = 800 $

**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
6. The time required for servicing transmissions is normally distributed with *μ* = 45 minutes

*σ* = 8 minutes.

x=50(x is the time taken to complete the work)

**FIND Z SCORE**

Z=(x-mu)/sigma

=(50-45)/8

=0.625

**Find probability**

P(X<=50) = P(Z<= 0.625)

=0.734014

P(X>50) =1-P(Z<=0.625)

=0.2659

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.

Sol: Mean (mu) = 38

Std (sigma) =6

1. More employees at the processing center are older than 44 than between 38 and 44.
2. Probability of employers greater than 44

P(x>44)=1-P(x<=44)

Z=(x-mu)/sigma

=(44-38)/6

=1.0

P(x<=44)=P(Z<=1)

=0.841344 = 84.13%

P(x>44)= 1-P(Z<=1) = 15.86%

(OR)

More employees at the processing center are older than 44 is 100-84.14 = 15.86%

Probability of employers between 38 & 44 is P(x<=44)-P(x>=38)

P(x>=38) = 1-P(x<=38)

P(x<=44)-P(x>=38)=0.8413-0.5

=34.13%

More employees at the processing center are older than 44 than between 38 and 44 is **TRUE**

1. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.
2. Probability employees less than 30=P(x<30)

Z calculate value is -1.333

Z table value =0.9176

So the no of employees with probability 0.0917

The under 30 age = 400\*0.917

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

X + Y ∼ N (µ1 + µ2, σ1^2 + σ2^2 )

X − Y ∼ N (µ1 − µ2, σ1^2 - σ2^2 )

Z is linear combination of X and Y , then Z ∼ N(aµ1 + bµ2, a^2σ1^2 + b^2σ2^2 )

|  |  |
| --- | --- |
| 2X1~ N(2 u,4 σ^2) and | |
|  | X1+X2 ~ N(µ + µ, σ^2 + σ^2 ) ~ N(2 u, 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
7. p(a<x<b) = 0.99

Mean(mu) =100

standard Deviation(sigma^2) = 20

**The Values find in python**

From scipy import stats

From scipy .stats import norm

stats**.**norm**.**interval(0.99,100,20)

(48.48,151.51)

Option 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. Profit1 ~ N(5, 32) and Profit2 ~ N(7, 42)

that $1 = Rs. 45

Profit1

Mu=5

Sigma^2=3^2=9

Profit2

Mu=7

Sigma^2=4^2=16

Mean profits from two different divisions of a company = Mean1 + Mean2

=5+7

=12\*45=540 Million

Variance of profits from two different divisions of a company = sqrt( SD1^2 + SD2^2)

= sqrt(16+9)

=5\*45 = 225.0 Million

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

Sol) **import** numpy **as** np

**from** scipy **import** stats

**from** scipy.stats **import** norm

stats**.**norm**.**interval(0.95,540,225)

Range is Rs (99.00810347848784, 980.9918965215122) in Millions

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

Sol) To compute 5th Percentile, we use the formula X=μ + Zσ

To find 5% of z score = stats.norm.ppf(0.05)

Z= -1.644

X=540+(-1.644)225.0 = 170

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

Sol)Probability of Division 1 making a loss P(X<0)

stats**.**norm**.**cdf(0,5,3) = 0.047

Probability of Division 2 making a loss P(X<0)

stats**.**norm**.**cdf(0,7,4) = 0.040

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

The sampling frame refers to a list of an item which responds to the question and not the ones which do not respond to the questions.

1. Larger surveys convey a more accurate impression of the population than smaller surveys.
2. TRUE
3. *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:
4. The population

Sol) P = x/n => 225/900=0.025

1. The parameter of interest

Sol) Sample size, average scale

1. The sampling frame

Sol) 9000

1. The sample size

Sol) 225

1. The sampling design

Sol)

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

Sol) the results is unreliable

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.
3. TRUE
4. 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.
5. TRUE
6. The 95% Confidence-Interval for *μ* only applies if the sample data are nearly normally distributed.
7. False
8. What are the chances that ?
9. ¼
10. ½
11. ¾
12. 1
13. Option A
14. 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.

A) Null Hypothesis (H0)

P >= 0.05(Accept Null Hypoyhesis): means that Mozilla has more than or equal

to 5% share of the market}

Alternate Hypothesis (H1)

p < 5%(Accept Alternative Hypothesis): means that Mozilla has a less than 5%

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?

n = sample of users = 2,000

that the Mozilla Firefox browser launched in 2004 had grabbed a 4.6% share of the market = 0.046

95% of the confidence interval is 1.96

calculate Z score = z\_scores=(0.046-0.05)/(np.sqrt((0.05\*(1-0.05))/2000))

z\_scores = -0.820782681668124

p\_ value=1-stats.norm.cdf(abs(z\_scores))

p\_ value = 0.20588

P Value > 0.20588

Accept Null Hypothesis (H0): means that Mozilla has more than or equal

to 5% share of the market

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?

Accept Null Hypothesis (H0): means that Mozilla has more than or equal

to 5% share of the market

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

A) **Given:**

the ?f=95% confidence interval for the size of the shipment was ?f=250%2B45%2C250-45

books.

**Find:**

We have to find the which of the following interpretations of this interval are

correct.

**Solution:**

Given that,

the ?f=95% confidence interval for the size of the shipment was ?f=250%20%2B%2045%2C250-45 books.

?f=295%20%2C%20205

Option A is correct

The ?f=95% confidence interval for the size of shipment was books.

All shipments are not between ?f=205 and ?f=295 books. Because due to ?f=95% confidence interval for the size of shipment not all shipments.

The known statement is not correct.

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
6. Option A

difference between mean of distribution and data points in standard deviation .

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
6. Margin of Error(ME) = 4%=0.04

95 % of confidence of interval = 95%=0.95

95% of confidence interval in Z Score = 1.96

P Value(P^) =0.5

Q Value (P^)=0.5

Formula ( ME) = Z\*Sqrt(P^ \* Q ^)/n

0.04 =1.96\*Sqrt(0.05\* 0.05)/n

n=(1.96)^2\* 0.5\*0.5/0.04^2

n=0.9604/0.0016

=600

**Option 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
6. 95 % of confidence of interval = 98%=0.98

95% of confidence interval in Z Score = 2.326

P Value(P^) =0.5

Q Value (P^)=0.5

Formula ( ME) = Z\*Sqrt(P^ \* Q ^)/n

0.04 =2.326\*Sqrt(0.05\* 0.05)/n

n=(2.326)^2\* 0.5\*0.5/0.04^2

n=845.35

**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. Answer C
4. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.)
5. Answer B
6. Are skewed (i.e. not symmetric) ?
7. Answer B,C,D
8. Have outliers on both sides of the center?
9. Answer A



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. Sample (n) = 25

Mu=22 lbs

Sigma=5lbs

1. Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that weights of individual packages are normally distributed.
2. TRUR
3. The standard error of the daily average SE() = 1.
4. TRUE
5. Standard error equal to standard deviation divided by square root of sample size =

5/sqrt(25) =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%
7. Sample(n) = 100

Average(Mu)=$50

Standard deviation(sd) = $40

As no. of samples is more than 30, we can consider it normal distribution.

the auditors decide to not initiate further investigations if the mean transaction amount of the sample is between $45 and $55 = P(45<X<55)

further investigation =1-P(45<X<55)

Calculating Z at 45 is Z= X-mU/sigma sigma= 40/sqrt(n)=40/(100)\*\*0.5

Z=45-50/4 =4

Z= -5/4 = -1.25

Calculating Z at 55 is Z= X-mU/sigma sigma= 40/sqrt(n)=40/(100)\*\*0.5

Z=55-50/4 =4

Z=5/4 = 1.25

For No investigation P(45<X<55) using z\_scores = P(X<55)-P(X<45)

= 0.7887004526662893

For investigation P(45<X<55) using z\_scores = 1 - P(X<55)-P(X<45)

=1- 0.7887004526662893

= 0.21130000000000004 = 21.3%

1. Option C
2. 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.
3. 144
4. 150
5. 196
6. 250
7. Not enough information

A)For 5%, z has to be +/-1.96  
 So 1.96=(5)\*sqrt(n)/40  
 sqrt(n)=15.68  
 n=245.86 or 246

Option E(Not enough information)

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?

Sol)The SEM is sd/sqrt(n)=120/sqrt(40000)=0.6

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

Sol) SD will not be 120 of scores in any one sample, especially since we don't know the

sample size.

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

Sol) SD of mean across several samples will also not be 120. It will be less

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

Sol) it couldn't be less or more**.**

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

Sol)This is certainly possible

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

Sol) The SEM will be 0.60,The mean would have an expected value of 720, but in calculations, the SEM is 0.6

Option E correct