**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**: from boxplot we conclude that, 91.36 % is the outlier for given data.

Mean of the Measure is: 33.27133333333333

Standard deviation of the Measure is: 16.370812590976932

Variance of the Measure is: 268.00350488888887



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**: IQR (Inter quartile range) = 12-5

= 7

this represents the range which contains 50% of the data points and it have 1 outlier

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

**Ans**: from the above boxplot we can say the data is shifted towards the left side Hence it is positively skewed/right skewed distribution. And More than 50% of the data is between 7-12

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**: if the value 25 is replaced by 2.5 then outlier will be removed from the data, and median is shifts, it will reduce the right skewness of the data



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

**Ans:**   The mode is always one of the numbers in data. We need to have actual points/data to get exact value of mode. the mode can lie between 4 to 8 because there are many values in this range but this is just an assumption. The two bars of the same height does not indicate mode every time.

1. Comment on the skewness of the dataset.

**Ans:** Most of data lies on left side and tail is at right side so it’s right/ positively skewed distribution and also have outlier.

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:** From histogram and boxplot we can say the value of the outlier is 25, and both the datasets shows the data is right skewed or positively skewed distribution. Median in boxplot and Mode in histogram. Histogram provides the frequency distribution so we can see how many times each data point is occurring however boxplot provides the quantile distribution i.e.50% data lies between 5 and 12. Boxplot provides whisker length to identify outliers, no information from 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 us consider the probability of 1 call misdirected out of 200 as event A.

Probability of occurring of event A = 1/ 200

P(A) = 1/200

Probability of having at least one successful call will be, 1-P(A) = 1-(1/200) = 1-0.005 = 0.995

As every event is independent of other event the probability will be, 1-(0.995)^5

= 1-0.975

= 0.025

= 2.5% 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:** The most likely monetary outcome of the business venture is 2000 because it has high probability 0.3

1. Is the venture likely to be successful? Explain

**Ans:** Success can be defined in multiple ways but based on the provided data, we can say venture is successful if x is positive. Therefore if x=1000,2000,3000 respective probability is, 0.2+0.3+0.1 = 0.6 which is greater than 0.5. Hence 60% chance of the venture would be successful.

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

**Ans:** let us consider similar business ventures of this type whose distribution of the returns is similar to this venture. In that case we say that the expected value of returns to this particular venture is the required average.

Long term average earning of business venture is =x\*P(x)

= (-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:** A good measure to evaluate the risk would be variance and standard deviation of the variable x.

Var = 3500000

Std = 1870.83

The large value of standard deviation of $1870 is considered along with the average returns of %800 indicates that this venture is highly risky.

**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: B) 0.2676**

μ=45min, σ=8min

μ=10min after = 45+10=55

Car will be ready in 1hr=60min(x)

Z = x- μ/σ

= 60-55/8

= 0.625

1 = stats.norm.cdf (z)

1 = stats.norm.cdf (0.625)

= 0.2659

P value for z score is 0.2659 approx to 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.**

68% of the data falls within one standard deviation of the mean (µ + *σ* ).

Here, µ = 38

*σ* = 6

Z score = (X- µ)/ *σ*

Z score for 44  = (44 - 38)/6  = 1

In python- stats.norm.cdf(1) =>  84.13 %

=> People above 44 age = 100 - 84.13 =  15.87%

Given 400 clerical employes, that is calculating 15.87% of 400=336.52 out of 400

=400 - 336.52

=63.49≈63(approx)

Z score for 38  = (38 - 38)/6 = 0

In python- stats.norm.cdf(0) => 50%

Hence People between 38 & 44  age = 84.13 - 50 = 34.13 % ≈  137 out of 400

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

**Ans:** **True.**

Z score for 30  = (30 - 38)/6 =  -1.33  =  9.15  %   ≈ 36 out of 400

(Z = (X - µ) / ϭ

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

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:** 2X1 is simply a larger scale version of the random variable X, if X1 is normally distributed then 2X1 is also normally distributed.

X1 and X2 are normal distributed, the associated sums and random samples are exactly normal, with appropriate parameters.

(As both are independent normal random variables, X1+X2 is normal with *N*(μ1+ μ2, σ21+ σ22). And 2X1 will just scale the normal distribution by 2 times)

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

**Ans:** Since we need to find out the values of a and b, which are symmetric about the mean, such that the probability of random variable taking a value between them is 0.99, we have to work out in reverse order.

The probability of getting value between a and b is: 0.99.

Therefore, probability of getting value outside a and b is: 1-0.99 = 0.01.

α = 1-0.99 / 2 = 0.005

Using Z-table, for probability 0.005 the Z value is: -2.57

Z = (X - µ) / ϭ

Rearranging formula to find the value of X:

X = (Z \* ϭ) + µ

For Z(-0.005) : Z(-0.005)\*20+100 = -(-2.57)\*20+100 = 151.4

For Z(0.005) : Z(0.005)\*20+100 = (-2.57)\*20+100 = 48.6

**OR**

**Here, we need range of 99% data which lies between 3rd standard deviation of the mean.**

**Here µ = 100, ϭ = 20**

**From empirical rule, µ ± 3 ϭ = 100 ± 3\*20 = (100-60,100+60) = (40,160)**

**Option D is right.**

**OR**

Using Python,

(p(a<x<b) = 0.99, µ=100, σ = 20

To Find:

Identify symmetric values for the standard normal distribution such that the area enclosed is 0.99

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

Two values symmetric about mean for the given standard normal distribution are [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
2. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.

**Ans: Range is Rs (99.00810347848784, 980.9918965215122) in Millions**

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

**Ans: 5th percentile of profit (in Million Rupees) is 202.0**

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

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

**Ans: 0.047**

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

**Ans: 0.040**

**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 depends upon the size (n) of the sample. The sample size should have at least 30 observations.

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 items in the target population from which the sample is selected.

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

**Ans: True.**

Large sample size will result in less standard deviation compared to small sample size. Hence we can say that larger sample is more accurate,

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:** Readers if the magazine = 9000

1. The parameter of interest

**Ans:** Rating of camera (7.5)

1. The sampling frame

**Ans:** All readers of the issue where the survey which was included 9000 .

1. The sample size

**Ans:** 225

1. The sampling design

**Ans:** Voluntary response

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

**Ans:** It is possible that only those who were particularly pleased or only who are displeased with the product participated in the survey which can makes the results 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.

**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: FALSE**

We have evidence in that direction but we cannot confirm 100% based on this data. We have to consider the values out of this range (i.e. more than 95% confidance interval).

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

**Ans: False**

With a large enough​ sample, the central limit theorem implies a normal sampling distribution regardless of the distribution of the data.

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

**Ans: ½**

This is pure assumption. There is a 50% chance that the sample mean( X greater than the population mean(µ).

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?
3. 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?

**(Question solved in jupyter file.)**

**Ans:** In this case, we have data on the entire population and the sample value accurately reflects the population number. Thus we can conclude that the share is less than 5%.

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**

The interval of (205,295) is for 95% confidence not for 100%.

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

**Ans: Incorrect**

The interval doesn’t describe individual shipments.

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

**Ans: Correct**

95% of intervals created in this way contain the true population mean.

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: Incorrect**

The interval doesn’t describe the mean of another sample.

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

**Ans:** **Incorrect**

The interval doesn’t correspond to a 95% confidence level.

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:** **The 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

**Ans: 600**

P=0.50

Z=1.960 for 95% confidence interval

N=(z/M)2 (p)(1-p)

N=(1.960/0.04)2 (0.5)(1-0.5)

= 600.25

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

**Ans: 848**

Z = 2.576 for 98% confidence interval

0.04 = 2.326 \*√ 0.5∗0.5 / n

n = (2.3262)2 ∗ 0.5 ∗ 0.5 / (0.04)2

= 1.3525 / 0.0016

= 845.35 ≈ 848 C

**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 :** Plot C is the Normal/ linear.

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

**Ans:** Plot D is Bimodal Distribution.

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

**Ans:** Plot A is right skewed and not symmetric.

1. Have outliers on both sides of the center?

**Ans:** Plot B contains outliers on both the sides.



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:** **False**, there is no need to confirm weights of every individual packages for sample distribution

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

**Ans: True**, Standard error = σ/sqrt(n)

= 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%

**Ans: 21.1%**

N=100

Std=40

Standard error= σ/sqrt(n)=40/100=4

Amount transaction between=45 to 55

P(45<x<55)=0.7887

1-0.7887=0.2113 = 21.1%

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**

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