**Name: Vibin Ravikumar Batch ID: DSWDMCOP 260422H**

# Basic Statistics (Module – 4 (Part – 1))

Q1) Calculate probability from the given dataset for the below cases

Data\_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

1. P(MPG>38)

**33/81=0.407**

1. P(MPG<40)

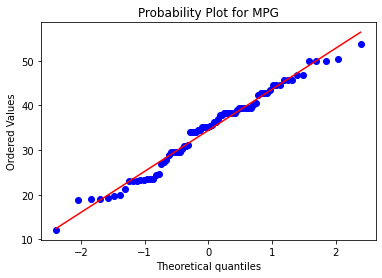
**61/81=0.75**

1. P (20<MPG<50)

**69/81=0.85**

Q2) Check whether the data follows normal distribution

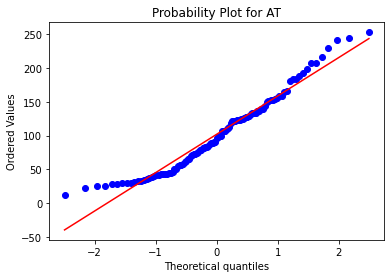
1. Check whether the MPG of Cars follows Normal Distribution Dataset: Cars.csv

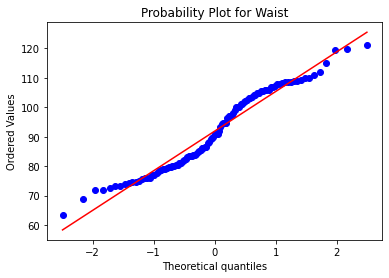
Ans: **Variable MPG of Cars follows Normal Distribution.** 

1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**Ans: Both AT and Waist variables of wc-at dataset do not follow normal distribution.**





Q3) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Ans:**

|  |  |  |
| --- | --- | --- |
| S.No | Confidence Interval | Z-Scores |
| 1 | 90% | ±1.6448 |
| 2 | 94% | ±2.053 |
| 3 | 60% | ±0.8416 |

Q4) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Ans**

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Confidence Interval | Degree of Freedoms | t-Scores |
| 1 | 95% | 24 | ±2.064 |
| 2 | 96% | 24 | ±2.17 |
| 3 | 99% | 24 | ±2.8 |

Q5**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

**Ans:** Using the t-statistics and degree of freedom we can calculate the probability using python

Formula for t-statistics is given below:

**t = [ x -** **μ ] / [ s / sqrt( n ) ]**

x-sample average=260

μ-Population average=270

s-sample standard deviation=90

n=number of samples=18

t = ( 260- 270 ) / [ 90 / sqrt( 18) ]

t = - 0.471

Using the t-statistics value=-0.471 and the degree of freedom=n-1=17 we can find the probability in python with the below code.



**32.2%** is the probability that 18 randomly selected bulbs would have an average life of more than 260 days.

Q6) 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?

A. 0.3875

B. 0.2676

C. 0.5

D. 0.6987

**Ans:**  = 45 minutes ,  = 8 minutes

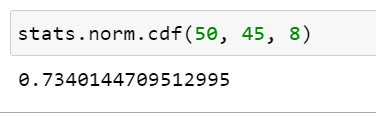
Allotted time is 1 hour=60 minutes, Delay to begin work= 10 minutes.

Time left to complete the work= 50 minutes.

P(X>=50) is to be calculated.

We can find P(X<=50) using scipy.stats library in python.

**stats.norm.cdf (X, ,  )**

****

P(X<=50)=0.734

P(X>=50)=1- P(X<=50)=1-0.734=**0.26**

**Thus Option B, 0.2676 is the correct answer.**

Q7) 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.
2. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

**Ans:**  = 38,  =6

Need to find.

P(X>=44)=?

P(38<X<44)=?

P(X<30)=?

P(X<=44)=0.841



P(X>=44)=1-.841=0.159=**15.9%**

We know that 38 is the average so below 38 has half the probability

So P(X<=38)=0.5

P(38<X<44)=1- P(X>=44)- P(X<=38)=1-0.159-0.5=0.341=**34.1%**

**From the above details it is clear that Statement A is False. As Number of employees older than 44 is 15.9% whereas the number of employees between 38 and 44 is 34.1%**

P(X<30)



P(X<30)=0.0912

Number of employees below 30=0.0912\* Total No of employees.

=0.0912\* 400=36.48 approx **36 employees**

**From the above details it is clear that Statement B is True.**

Q8) If X1 ~ N(μ, σ2) and X2 ~ N(μ, σ2) are iid normal random variables, then what is the

difference between 2 X1 and X1 + X2? Discuss both their distributions and parameters.

**Ans:**

From the properties of **normal random variables**,

if  and  are two independent identically distributed random variables then

* the **sum** of normal random variables is given by

,

* and the **difference** of normal random variables is given by



* When  , the **product** of X is given by



* When  , the **linear combination** of X and Y is given by



Find 2X1:

**2X1****~ N(****2μ, 22σ2) ~N(2μ, 4σ2 )**

Find X1 + X2 :

**X1 + X2** **~ N(****μ + μ ,** **σ2 + σ2) ~N(2μ, 2σ2 )**

Find Difference between 2X1 and X1+X2:

**2X1 - (X1+X2) ~ N(μ - μ , 4σ2 + 2σ2) ~N(0, 6σ2 )**

**From the above we can see that the mean of 2X1 and X1+X2 are same but the variance of 2X1 is double of X1+X2.**

**As the mean is same we can say that both are identical but are independently distributed as the variance is different.**

Q9) Let X ~ N(100, 20^2) its (100, 20 square).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.

A.

90.5, 105.9

B. 80.2, 119.8 C.

22, 78

D. 48.5, 151.5

E. 90.1, 109.9

**Ans:** μ=100, σ=20

**Z=(X-** **μ) /** **σ therefore to find X we can use X=Z\* σ + μ**

Z score for 99% confidence is ±2.575 using python or the z-score table.

X= (Z0.005\* 20) + 100 , (Z0.995\* 20) +100

= (-2.575\* 20) + 100 , (2.575\* 20) + 100

= **48.5, 151.5**

**Therefore Option D is the answer.**

Q10) Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions Profit1 ~ N(5, 3^2) and Profit2 ~ N(7, 4^2) 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.
2. Specify the 5th percentile of profit (in Rupees) for the company
3. Which of the two divisions has a larger probability of making a loss in a given year?

**Ans:** $1 = Rs. 45

Profit1 ~ N(5, 3^2)

Profit2 ~ N(7, 4^2)

So companies profit=P= Profit1 + Profit2

P~ N(5+7, 32 42) ~N(12, 52)

1. **Z=(X- μ) / σ therefore to find X we can use X=Z\* σ + μ**

Z score for 95% probability is ±1.96 using python or the z-score table.

X= (Z0.025\* 5) + 12 , (Z0.975\* 5) +12

= (-1.96\* 5) + 12 , (1.96\* 5) + 12

= 2.2 , 21.8 in million dollars

**= 99 , 981 in million rupees.**

**A - Rupee range between 99 to 981 million rupees contains 95% probability for the annual profit of the company.**

1. Z- score value for 5th Percentile= -1.645

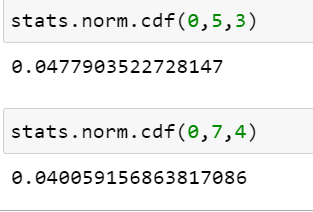
X= Z\* σ + μ = (-1.645\* 5)+12 = 3.8 million dollars =3.8\* 45= 171 million rupees.

**B- 5th percentile of profit (in Rupees) for the company is 171 million rupees.**

1. Loss means Profit is less or equal to zero

So for profit1 X=0, σ =3, μ=5 using python to find the probability.

And for profit2 X=0, σ =4, μ=7 using python to find the probability.



C- From the above detail it is clear that Division 1 has a higher probability of making a loss in a given year.

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. For each assignment the solution should be submitted in the below format
3. Research and Perform all possible steps for obtaining solution
4. For Basic Statistics explanation of the solutions should be documented in black and white along with the codes.

One must follow these guidelines as well:

* 1. Be thorough with the concepts of Probability, Central Limit Theorem and Perform the calculation stepwise
  2. For True/False Questions, explanation is must.
  3. R & Python code for Univariate Analysis (histogram, box plot, bar plots etc.) for data distribution to be attached

1. All the codes (executable programs) should execute without errors