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

**Name: Upadhyay Sachin Naresh**

**Batch ID: Data science\_08032021**

Q1) 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)
2. P(MPG<40)
3. P (20<MPG<50)

Answer:

cars =read.csv(file.choose())

MPG <- cars$MPG

sample(MPG)

a=subset(MPG,MPG>38)

show(a)

#ans = 33/81

b=subset(MPG,MPG<40)

show(b)

#ans = 57/81

c=subset(MPG,MPG>20 & MPG <50)

show(c)

#ans = 65/81

Q2) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution Dataset: Cars.csv
2. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

ANSWER:

#Check whether the data follows normal distribution

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

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

cars=pd.read\_csv('C:/Users/usach/Desktop/confidence/Cars.csv')

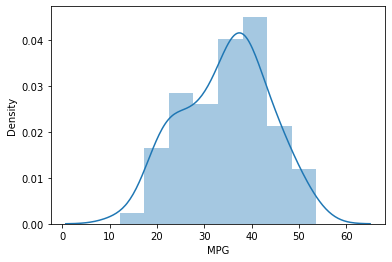
cars

sns.distplot(cars.MPG, label='Cars-MPG')

plt.xlabel('MPG')

plt.ylabel('Density')

plt.legend();



cars.MPG.mean()

Out: 34.422075728024666

cars.MPG.median()

Out: 35.15272697

Inference: MPG of Cars does follow normal distribution approximately

(as mean and median are approx. same)

b)Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist)

from wc-at data set follows Normal Distribution Dataset: wc-at.csv

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

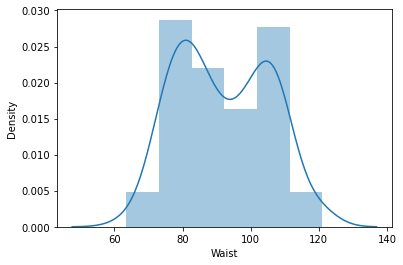
wcat=pd.read\_csv('C:/Users/usach/Desktop/confidence/wc-at.csv')

wcat

# plotting distribution for Waist Circumference (Waist)

sns.distplot(wcat.Waist)

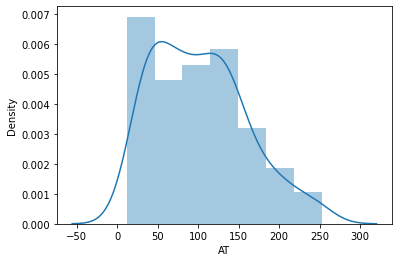
plt.ylabel('density');



# plotting distribution for Adipose Tissue (AT)

sns.distplot(wcat.AT)

plt.ylabel('density');



# WC

wcat.Waist.mean() , wcat.Waist.median()

Out: (91.90183486238533, 90.8)

# AT

wcat.AT.mean() , wcat.AT.median()

Out: (101.89403669724771, 96.54)

Inference: Both the Adipose Tissue (AT) and Waist Circumference (Waist) data set do follow the normal distribution approximately (as mean and median of both the data are approximately same)

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

ANSWER:

from scipy import stats

from scipy.stats import norm

Z-score of 90% confidence interval

stats.norm.ppf(0.95)

out:1.6448536269514722

# Z-score of 94% confidence interval

stats.norm.ppf(0.97)

out: 1.8807936081512509

# Z-score of 60% confidence interval

stats.norm.ppf(0.8)

out: 0.8416212335729143

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

ANSWER:

from scipy import stats

from scipy.stats import norm

# t scores of 95% confidence interval for sample size of 25

stats.t.ppf(0.975,24) # df = n-1 = 24

out:2.0638985616280205

# t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

out: 2.1715446760080677

# t scores of 99% confidence interval for sample size of 25

stats.t.ppf(0.995,24)

out:2.796939504772804

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

Answer:

**To find:**

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

**Solution:**

t - statistics for the data is given as follows:



x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18









t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 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

Answer: B

We have a normal distribution with = 45 and = 8.0. Let X be the amount of time it takes to complete the repair on a customer's car. To finish in one hour you must have X ≤ 50 so the question is to find Pr(X > 50).

Pr(X > 50) = 1 - Pr(X ≤ 50).

Z = (X - )/ = (X - 45)/8.0

Thus the question can be answered by using the normal table to find

Pr(X ≤ 50) = Pr(Z ≤ (50 - 45)/8.0) = Pr(Z ≤ 0.625)=73.4%

Probability that the service manager will not meet his demand will be = 100-73.4 = 26.6% or 0.2676

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.

Answer:

We have a normal distribution with = 38 and = 6. Let X be the number of employees. So according to question

a)Probabilty of employees greater than age of 44= Pr(X>44)

Pr(X > 44) = 1 - Pr(X ≤ 44).

Z = (X - )/ = (X - 38)/6

Thus the question can be answered by using the normal table to find

Pr(X ≤ 44) = Pr(Z ≤ (44 - 38)/6) = Pr(Z ≤ 1)=84.1345%

Probabilty that the employee will be greater than age of 44 = 100-84.1345=15.86%

So the probability of number of employees between 38-44 years of age = Pr(X<44)-0.5=84.1345-0.5= 34.1345%

**Therefore the statement that “More employees at the processing center are older than 44 than between 38 and 44” is TRUE.**

b) Probabilty of employees less than age of 30 = Pr(X<30).

Z = (X - )/ = (30 - 38)/6

Thus the question can be answered by using the normal table to find

Pr(X ≤ 30) = Pr(Z ≤ (30 - 38)/6) = Pr(Z ≤ -1.333)=9.12%

So the number of employees with probability 0.912 of them being under age 30 = 0.0912\*400=36.48( or 36 employees).

**Therefore the statement B of the question is also 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.

Answer:

Given two x1 and x2: following the normal distribution

Mean and standard deviations is the parameter of normal disribution

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

Similarly if Z = aX + bY , where X and Y are as defined above, i.e Z is linear combination of X and Y , then Z ∼ N(aµ1 + bµ2, a^2σ1^2 + b^2σ2^2 ).

Therefore in the question

2X1~ N(2 u,4 σ^2) and

X1+X2 ~ N(µ + µ, σ^2 + σ^2 ) ~ N(2 u, 2σ^2 )

2X1-(X1+X2) = N( 4µ,6 σ^2)

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

Answer: D

Given mu =100, sigma =20

Probability inside the a and b is 99%

Level of significance = 0.005%

Now z(0.005) = -2.57

Zscore = a-mu/sigma

-2.57 = a- 100/20

a = -51.4+100

b = 48.5

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?

Answer:

a) Rs 603.68

b) Rs 476.33

c) first division

**Step-by-step explanation:**

Total profit = profit 1 + profit 2 = P

hence P ~ N(12,74) = 

A ) Specifying a Rupee range ( centered on the mean ) that contains 95% probability for annual profit of the company

$13.41 = Rs 603.68

B) specifying the 5th percentile of profit

p = $10.59 ≈ Rs 476.33

C) The division that has a larger probability of making a loss in a given year is the first division

**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