AccelerateAI

Data Science Global Bootcamp

Probability & Statistics: Assignment 03

Probability

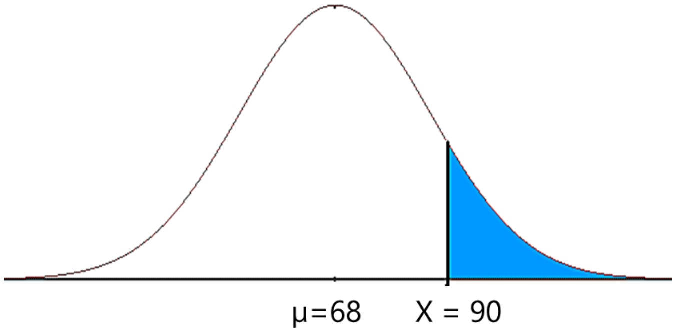
Q1: MyGrocery.com, an online grocery store, makes a claim to deliver orders within 90 minutes. Based on past data, it was found that the average time to deliver is 68 min with standard deviation of 14 minutes, and follows a normal distribution.

1) What proportion of orders is delivered after 90 minutes?
2) What should be the promised delivery time, if the target is to deliver at least 99% orders before that time?

Solution:

The delivery time follows a normal distribution with mean = 68 min and s.d. = 14 mins.

Proportion of orders that are delivered after 90 mins, will be given by the shaded region shown below.



P(X > 90) = 1 – P(X ≤ 90) = 1 – CDF(X=90)

= 1 - st.norm.cdf(x=90, loc=68, scale=14)

= 0.058

About 6% of orders are delivered after 90 minutes.

2) We have to find x such that P(X<= x) = 0.99

x = Inv(CDF(0.99) = st.norm.ppf(0.99, loc=68, scale=14) = 100.5

The promised delivery time should be 100 minutes, if 99% orders have to arrive before that.

Q2. The flight delay times of planes leaving San Diego airport in California are monitored (on time flights or early departures are not included, hence no negative times). The delay in departures of 6 flights are noted. These delay times are 5.5, 10.5, 13, 22.5, 45, 55 minutes.

A family of 5, is delayed to the airport by 25 minutes. What is the probability that they will catch their flight? (Use a t-distribution with 5df)

Solution:

Sample Mean delay = (5.5 + 10.5 + 13 + 22.5 + 45 + 55) / 6 = 25.25

Sample Std Dev delay = 20.21

The family will catch the flight if the delay is greater than 25 minutes.

Probability (delay > 25) = 1 – P (delay ≤ 25) = 1 – t.cdf(x=25, mean=25.25, sd=20.21, df=5)

= 0.495

There is close to 50% chance that the family will catch their flight.

Q3. The number of miles traveled by a given car before its transmission ceases to function is governed by the exponential distribution with mean 100, 000 miles. What is the probability that a car's transmission will fail during its first 50,000 miles of operation?

Solution:

The probability (CDF) of failure for an exponential distribution is given by:

$$P(X \le x) = 1 - e^{-\lambda x}$$

Here rate of failure is in per unit distance (not time)

$$\lambda = \frac{1}{100{,}000}$$

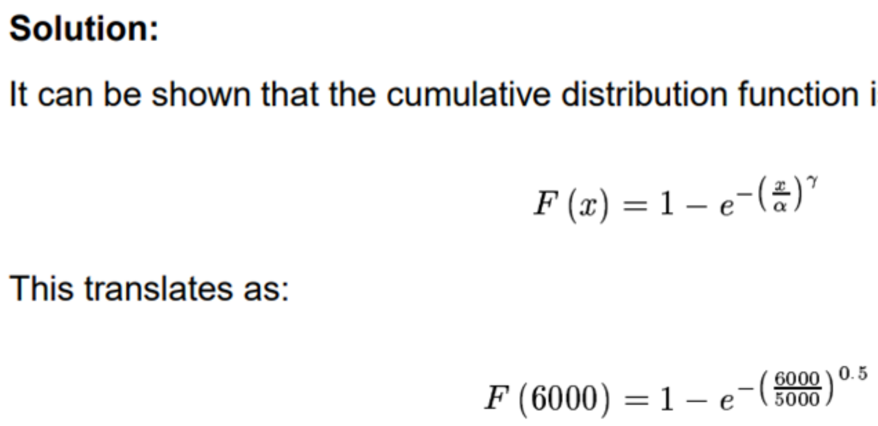
x = 50,000

=> $P(X \le x) = 1 - e^{-\frac{50{,}000}{100{,}000}}$ = 0.393

The probability of transmission failure in first 50K miles is about 40%.

Q4. The time to failure (in hours) of a bearing in a mechanical shaft is satisfactorily modeled as a Weibull random variable with α =5000 and gamma = 0.5.

1) Determine the probability that a bearing lasts fewer than 6000 hours
2) What is the mean time to failure?



Solution:

It can be shown that the cumulative distribution function is given by

$$F\left(x\right) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\gamma}}$$

This translates as:

$$F\left(6000\right) = 1 - e^{-\left(\frac{6000}{5000}\right)^{0.5}}$$

$$=> \mathrm{F}(6000) = 0.666$$

Mean time to failure:

$T_{avg.}$ = 5000* Tau(1+ 1/gamma) = 5,000* 2! = 10,000 hours

#Tau – the Gamma function

Tau(3) = 2!

Q5. Identify the data type of the following as ***nominal, ordinal, interval or ratio scale:***

S No	Description	Data Type
1	GMAT Score	Interval
2	Blood Type	Nominal
3	Credit Score	Interval
4	Temperature (Celsius)	Interval
5	Temperature (Kelvin)	Ratio
6	Zip code	Nominal
7	Shoe Size (UK)	Ordinal
8	Birth Year	Interval
9	Military Rank	Ordinal
10	Tax rate	Ratio

Q6. A population has a mean of μ=35 and a standard deviation of σ=5. After 3 is added to every score in the population, what are the new values for the mean and standard deviation?

Solution:

When we add 3 to every score, the population has an increase in mean by 3 which is self-explanatory, hence μ will be 38.

However, it will have no impact on the σ which will remain as 5.

Further explanation:

When we Add a constant - The median, mean, and quartiles will be changed by adding a constant to each value. However, the range, interquartile range, standard deviation and variance will remain the same.

When we Multiply with a constant - Multiplying every value by a constant, however, will multiply the mean, median, quartiles, range, interquartile range, and standard deviation by that constant, and multiply the variance (which is simply the square of the standard deviation) by the square of that constant.

Q7. A consulting firm is analyzing the expenses for its employees during a meetup event. A sample of 80 professionals are surveyed and the average amount spent by them on travel and beverages is $593.84. The sample standard deviation is approximately $369.34

Construct a 95% confidence interval for the population mean amount of money spent during the event.

Solution:

$\mu = \bar{x} \pm [Z_{\alpha/2} * \sigma / \sqrt{n}\ ]$

For 95% confidence interval, we can use $Z_{\alpha/2} = 1.96$

$\mu = 593.84 \pm [1.96 * 369.34 / \sqrt{80}\ ]$

$= 593.84 \pm [1.96 * 369.34 / 8.944]$

$= 593.84 \pm [1.96 * 41.294]$

$= 593.84 \pm\ 80.93$

$= (512.91, 674.77)$

Q8. As an asset manager in the large global Investment Banking firm, you are tasked to analyze the portfolio of your customers and help them predict risk. For the distribution of the dataset you have, the Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13 and mean is 59.2. Find the mode and median of the distribution.

Solution:

We have, Karl Pearson coefficient of skewness $S_k = 0.64$, standard deviation $\sigma = 13$ and Mean $\mu = 59.2$

Therefore, by using formulae S_k = (Mean – Mode) / σ , you get the following:

0.64 = (59.2 – Mode) / 13

=> 8.32 = 59.2 – Mode

=> Mode = 50.88

Mode = 3 Median – 2 Mean

=> 50.88 = 3 Median – 2 * 59.2

=> Median = 56.42

Q9 One of the IPL franchises is formulating strategy for the next auction during team selection. A player is to be selected for their team. The choice is between player A and player B given below, on the basis of their past 5 batting performances.

A	**25**	**85**	**40**	**80**	**120**
B	**50**	**70**	**65**	**45**	**80**

Which player should they choose if they want

- **a higher run-getter**
- **a more consistent and reliable batter into the team**

Answer for both these options with explanation.

i) Average of Batsman A is higher than that of Batsman B, so he should be selected if we want a high scorer.

ii) The Batsman B is more reliable than Batsman A. This is because the coefficient of variation of Batsman A is higher than that of Batsman B

Solution:

Yes, CV (Coefficient of Variation) = (SD/Mean) * 100 should be looked into for understanding variability as well.

The definition of CV - The coefficient of variation (CV) is a statistical measure of the relative dispersion of data points in a data series around the mean.

CV is also known as Relative Std dev. (RSD)

For A,

Mean = 70

SD = 33.91

CV = (33.91/70)*100 = 48.4%

For B,

Mean = 62

SD = 12.88

CV = (12.88/62)*100 = 20.8%

A is better run-getter.

B is more consistent.

Q10 In a manufacturing setup, assume that two machines produce a part of the product which are on average 10 inches long. A sample of 11 parts are selected from each machine.

Machine A: 6, 8, 8, 10, 10, 10, 10, 10, 12, 12, 14.

Machine B: 6, 6, 6, 8, 8, 10, 12, 12, 14, 14, 14.

Which machine is better?

Solution:

The SD of Machine A = 2.08

The SD of Machine B = 3.19

Smaller the sample SD better the machine (consistent).

So, Machine A is better.

Q11. Examples related to various sampling methods are provided below. Identify the odd one out.

a) **Asking volunteers at the mall**
b) **Pilot testing of a new product launch**
c) **Find brand of smartphone preferred by people based on either gender or age or socio-economic status**
d) **Pick all students and split them by grade and select random from each grade**

Answer: Option d

Option a is an example of Convenience sampling

Option b is an example of Purposive sampling

Option c is an example of Quota sampling

Option d is an example of Stratified Random sampling

Option d is probabilistic where as other 3 are non-probabilistic sampling methods.