Iman Noor
Submission Data: 24_06_24

BYTEWISE FELLOW

TASK 09

INTRODUCTION TO STATISTICS AND PROBABILITY

STATISTICS

Definition:

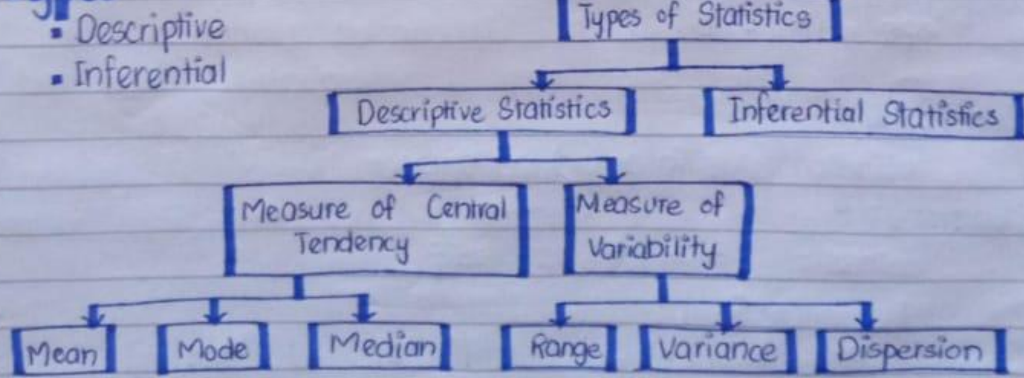
"Statistics is a branch of math focused on collecting, organizing and understanding numerical data."

Applications:

- Weather forecasting
- Stock Market Analysis
- Insurance
- Betting
- Data Science

- It is the study and manipulation of data.

Types:
- Descriptive
- Inferential



Descriptive Statistics:

"It uses data that provides a description of the population either through numerical calculated graphs or tables. It provides graphical summary of data."

Categories:

- Measure of Central Tendency
- Measure of Variability

Measure of Central Tendency:

"It is also known as summary statistics that are used to represent the center point or particular value of a data set or sample set."

Three common measures of central tendency are:

- Mean
- Median
- Mode

(i) Mean:

The arithmetic average of a set of values.

$$\bar{x} = \frac{\Sigma x}{n}$$

Example:

Dataset: $\{2, 4, 6, 8, 10\}$

$$\text{Mean} = \bar{x} = \frac{2+4+6+8+10}{5}$$

$$\text{Mean} = \bar{x} = 5$$

(ii) Median:

Middle value in an ordered dataset.

Example:

Dataset: $\{3, 6, 9, 12, 15\}$

Median $= 9$

n: "Even" Median $= [(n/2)^{th} \text{ term} + (n/2+1)^{th} \text{ term}]/2$

n: "Odd" Median $= (n+1)/2$

Dataset : { 21.3 , 20.8, 19, 15 }
Data in ascending order: 15, 19, 20.8, 21.3
Median : (20.8 + 19)/2 = 39.8/2
Median = 19.9

(iii) Mode :

Value that appears most frequently in dataset.

Example :

Dataset : { 2, 3, 4, 4, 5, 6, 6, 6, 7, 7, 7, 7 }
Mode = 7

Mode = Term with Highest Frequency

Measure of Variability :

"It is also known as measure of dispersion and is used to describe variability in a sample or population."

Three common measure of variability are :

- Range
- Variance
- Dispersion

(i) Range :

It is measure of how to spread apart values in sample set or dataset.

Range = Maximum value − Minimum value

Example :

Dataset : { 4, 8, 6, 5, 3, 7 }
Range = 8 − 3 = 5

Minimum value = 3
Maximum value = 8

(ii) Variance :

Variance measures dataset's spread or dispersion.

- It is calculated by averaging squared deviations from mean.

$$S^2 = \sum_{i=1}^{n} [(X_i - \bar{X})^2 / n]$$

n = total data points
$\bar{X}$ = mean of datapoints
X = individual datapoints

Example :

Dataset : $\{3, 5, 7, 9, 11\}$

$$\bar{x} = \mu = \frac{3+5+7+9+11}{5} = 7$$

Calculating deviation from mean and square it:

- $(3-7)^2 = 16$; • $(5-7)^2 = 4$
- $(7-7)^2 = 0$; • $(9-7)^2 = 4$
- $(11-7)^2 = 16$

Sum of above : $16+4+0+4+16 = 40$

Now,

Population variance : σ^2

Sample variance: S^2

$$S^2 = 40/5$$

$$\boxed{S^2 = 8}$$

Standard deviation :

The square root of variance

$$\sigma = \sqrt{S^2} = \sqrt{8}$$

$$\boxed{\sigma = 2.83}$$

(iii) Dispersion:

It is measure of dispersion of a set of data from its mean.

$$\sigma = \sqrt{(1/n) \sum_{i=1}^{n} (x_i - \mu)^2}$$

- Dispersion is a general term for measures that describe the spread of data points. It can be quantified in multiple ways including range, interquartile range and variance.
- Standard deviation is a specific measure of dispersion that indicates average distance of data points from mean.

Stages :

Five stages of statistics

- Problem
- Plan
- Data
- Analysis
- Conclusion

QUESTION 1:

Find the mean of the following frequency distribution using the step-deviation method.

x	10	30	50	70	90	110
f	135	187	240	273	124	151

SOLUTION:

x	f	$d = x - A$	$t = \frac{x-A}{i} = \frac{x-70}{20}$	ft
10	135	-60	-3	-405
30	187	-40	-2	-374
50	240	-20	-1	-240
A=70	273	0	0	0
90	124	20	1	124
110	151	40	2	302
	$f = 1110$			$ft = -593$

$A = 70$ and $i = 20$

Now,

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times i$$

$$= 70 + (-593/1110 \times 20)$$

$$\bar{x} = 59.32$$

QUESTION 2:

The weights of 45 people in society were recorded, to the nearest kg, as follows:

Wt. (in nearest kg)	46	48	50	52	53	54	55
No. of people	7	5	8	12	10	2	1

Calculate the median weight.

SOLUTION:

Weight (x)	No. of people (f)	Cummulative frequency (cf)
46	7	7
48	5	$7+5=12$
50	8	$12+8=20$
52	12	$20+12=32$
53	10	$32+10=42$
54	2	$42+2=44$
55	1	$44+1=45$

$n=45 \Rightarrow$ "ODD"

Median $= \left(\frac{n+1}{2}\right)^{th}$ term

$= (45+1/2)^{th}$ term

Median $= 23^{rd}$ term

The weight of 23^{rd} person $= 52$ kg.

QUESTION 3:

Find the mode from the following frequency distribution:

Number	7	8	9	10	11	12	13
Frequency	3	8	12	15	14	17	9

SOLUTION:

Since the frequency of number 12 is maximum.

Mode $= 12$

PROBABILITY

Definition:

"Probability is a number that reflects the chance or likelihood that a particular event will occur."

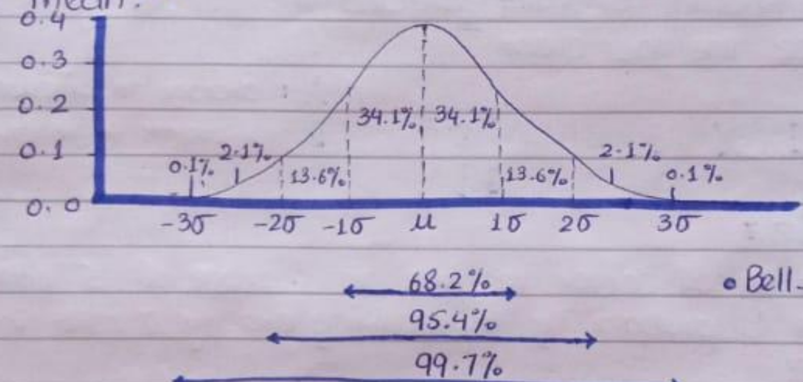
- It ranges from 0 (impossible event) to 1 (certain event)

Probability Distribution:

A function that describes the likelihood of different outcomes in a random experiment.

Normal Distribution:

"A continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data from the mean."



- Bell-shaped curve

Example:

The percentage of students scored between 60 and 80 corresponds to 1σ which covers 68%

Binomial Distribution:

"A discrete probability distribution of the number of successes in a sequence of "n" independent experiments, each asking a yes/no question and each with its own boolean valued outcome"

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n: no. of trials
p: success probability
$1-p$: failure probability

Example:

Suppose you flip a coin 10 times. What is probability of getting exactly 6 heads?

$n = 10$, $p = 0.5$, $k = 6$:

$$P(X=6) = \binom{10}{6}(0.5)^6(0.5)^4 = \frac{10!}{6!(10-6)!} \times (0.5)^{10}$$

$$P(X=6) = 0.205$$

Poisson Distribution:

"A discrete probability distribution expressing the probability of a given number of events occuring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event."

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ = average no. of events

k = k events in interval.

Example:

Suppose a call center receives an average of 5 calls per hour. What is probability of receiving exactly 7 calls in the next hour?

$\lambda = 5$, $k = 7$;

$$P(X=7) = \frac{5^7 e^{-5}}{7!}$$

$$P(X=7) = 0.104$$

Uniform Probability:

"A probability distribution where all outcomes are equally likely. Each n value has an equal probability of $1/n$."

Types:

- Discrete : Each of a finite no. of values
- Continuous : All values in continuous range

$$f(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b$$

- Mean : $\frac{a+b}{2}$
- Variance : $\frac{(b-a)^2}{12}$

Example:

Suppose the waiting time for a bus is uniformly distributed between 0 and 20 minutes. What is probability that you will wait less than 5 minutes?

$a = 0$, $b = 20$, $x = 5$:

$$P(X<5) = \frac{5-0}{20-0} = \frac{5}{20} = 0.25$$

QUESTION 4:

The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of $4.11 and a standard deviation of $1.37. What percentage of customers will spend less than $3.00 on concessions?

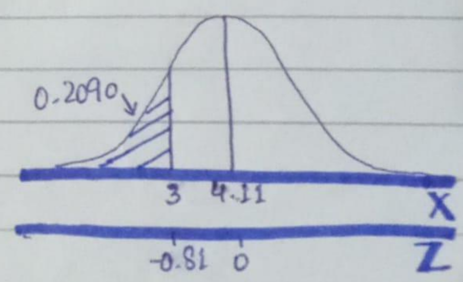
SOLUTION:

$\mu = 4.11$, $\sigma = 1.37$

$P(X<3.00) = P(Z<-0.81)$

$= 0.2090 \; \boxed{= 20.9\%}$

$Z = \frac{3.00-4.11}{1.37} = -0.81021$



QUESTION 5:

What spending amount corresponds to the top 87th percentile?

SOLUTION:

$\mu = 4.11$, $\sigma = 1.37$

$P(X>?) = 0.87 \Rightarrow P(Z>?) = 0.87$

$P(Z>?) = 0.87 \Rightarrow P(Z<?) = 1-0.87$

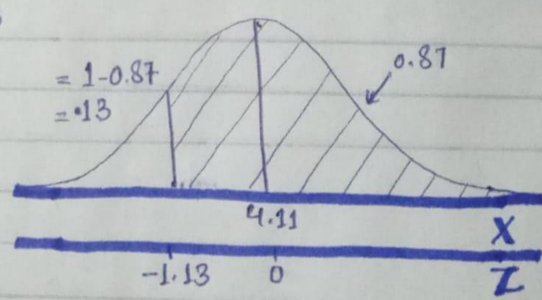
$= 0.13$

$Z \Rightarrow -1.13$

$X = 4.11 - 1.5481$

$X = 2.5619$

$X = \$2.56$



QUESTION 6:

The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

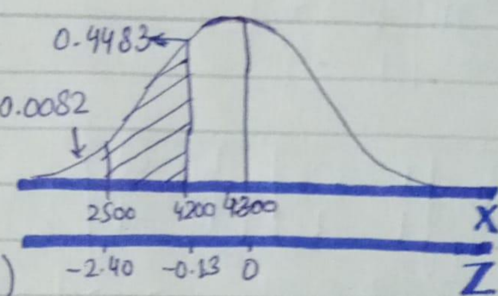
SOLUTION:

$$z = \frac{2500-4300}{750} = -2.40$$

$$z = \frac{4200-4300}{750} = -0.133$$

$\mu = 4300$, $\sigma = 750$

$$P(2500<X<4200) = P(-2.40<Z<-0.13)$$
$$= P(Z<-0.13) - P(Z<-2.40)$$
$$= 0.4483 - 0.0082$$
$$P(-2.40<Z<-0.13) = 0.4401$$



QUESTION 7:

The data in Table 5.3.1 are 55 smiling times, in seconds, of an eight-week-old baby.

Table 5.3.1

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

SOLUTION:

Sample mean = 11.49 , Sample standard deviation = 6.23

Theoretical mean = $\mu = \frac{a+b}{2} = \frac{0+23}{2} = 11.50$ seconds

Theoretical σ = $\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(23-0)^2}{12}} = 6.64$ seconds

Theoretical mean and σ are close to the sample mean and σ.

QUESTION 8:

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are equally likely. State the values of *a* and *b*. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

Table 5.3.2

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

SOLUTION:

$a = 0 \quad \Rightarrow \quad \mu = 0+14/2 = 7$ passengers

$b = 14 \quad\quad \sigma = \sqrt{\frac{(14)^2}{12}} = 4.04$ passengers

QUESTION 9:

A coin is tossed12 times. What is the probability of getting exactly 7 heads?

SOLUTION:

$n = 12$, $p = 1/2$, $1-p = 1-1/2 = 1/2$

$P(r) = {}^nC^r \, p^r (1-p)^{n-r} \qquad r = 7$

$P(7) = {}^{12}C_7 \, p^7 (1-p)^{12-7}$

$= 792 (1/2)^7 (1/2)^5$

$= 792 (1/2)^{12}$

$= 792 (1/4096)$

$P(7) = 0.193$ → Probability of getting exactly 7 heads.

QUESTION 10:

The probability that a person can achieve a target is 3/4. The count of tries is 5. What is the probability that he will attain the target at least thrice?

SOLUTION:

$p = 3/4$, $q = 1/4$, $n = 5$

$$P(X) = {}^{n}C_{x} \cdot p^{x}(1-p)^{n-x}$$

$= P(X=3) + P(X=4) + P(X=5)$

$= {}^{5}C_{3}(3/4)^{3}(1/4)^{2} + {}^{5}C_{4}(3/4)^{4}(1/4)^{1} + {}^{5}C_{5}(3/4)^{5}$

$= 459/512$

The probability of person that will attain target at least twice is $459/512 = 0.896$

QUESTION 11:

As only 3 students came to attend the class today, find the probability for exactly 4 students to attend the classes tomorrow.

SOLUTION:

$\lambda = 3$ (average rate of value)

$x = 4$ (poisson random variable)

$$P(X = x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \qquad \therefore \text{ also } = \frac{\lambda^{k} e^{-\lambda}}{k!}$$

$P(X = 4) = 3^{4} e^{-3}/4!$

$P(X = 4) = 0.168031$

QUESTION 12:

Scenario: A data center experiences an average of 10 server failures per month. We want to determine the following:

1. The probability of exactly 15 server failures in a given month.
2. The probability of having at most 5 server failures in a month.
3. The expected number of server failures over the next six months.

SOLUTION:

(1) $\lambda = 10, \quad k = 15$

$$P(X = 15) = \frac{10^{15} e^{-10}}{15!} = 0.03472$$ (exactly 15 server failure)

(2) $$P(X \leq 5) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$$

$$= \frac{10^0 e^{-10}}{0!} + \frac{10^1 e^{-10}}{1!} + \frac{10^2 e^{-10}}{2!} + \frac{10^3 e^{-10}}{3!} + \frac{10^4 e^{-10}}{4!} + \frac{10^5 e^{-10}}{5!}$$

$$= 4.5\times10^{-5} + 4.539\times10^{-4} + 2.269\times10^{-3} + 7.566\times10^{-3} + 1.89\times10^{-2} + 3.7\times10^{-2}$$

$$P(X \leq 5) = 0.0527$$ (at most 5 server failure)

(3) For one month $\lambda = 10$

For **six** months, expected no. of server failures :

$$E[X_{6 \text{ months}}] = 6 \times \lambda = 6\times10 = 60$$