Statistics Notes

B.Tech. CSE

Gurmukh Singh

Instructor:
Mrs. Neha

Contents

"All models are wrong, but some of them are useful"

~ George Box

1 Measures of Central Tendency

1. Mean
2. Median
3. Mode

1.1 Mean

It is the ratio of sum of all the observations to the total number of observations. let $x_1, x_2, \ldots, x_n$ be all the observations. then:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1.1.1 Properties of Mean

- The sum of deviation of observations from mean is always zero
- the sum of square of deviations of observations is minimum as compared to any other measure.
- suppose there are two sequences:

	Series 1	Series 2
Number of observations	n_1	n_2
mean of the observations	$\overline{x}_1$	$\overline{x}_2$

then

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$$

Problem 1

If there are 5 and 8 number of observations of 2 series with mean 15 and 18, find the combined mean

Solution:
We can get the solution by taking the weighted mean of the two sequences.
so the required mean is :

$$\frac{5 \times 15 + 8 \times 18}{5 + 8}$$
$$= \frac{75 + 144}{13}$$
$$= \frac{219}{13}$$
$$= 16.846154$$

Problem 2

Class	frequency
0-10	3
10-20	5
20-30	7
30-40	4
40-50	1

Solution:
change of origin:

Class	frequency	X	d=X-A	f·d
0-10	3	5	-20	
10-20	5	15	-10	
20-30	7	25	0	
30-40	4	35	10	
40-50	1	45	20	

$$\overline{x} = A + \frac{\sum fd}{n}$$

change of scale

Class	frequency	X	d=X/n	f·d
0-10	3	1	-20	
10-20	5	3	-10	
20-30	7	5	0	
30-40	4	7	10	
40-50	1	9	20	

$$\overline{x} = A + \frac{\sum fd}{n}$$

1.2 Median

Steps to find Median in case of Discrete and continuous data:

1. Arrangement of data
2. if n is odd then the median is the $\frac{n+1}{2}$th term
3. if n is even then the median is the mean of the $\frac{n}{2}$th term and $\frac{n}{2} + 1$th term

Problem 3

find the median for the data :

1. 9,9,10,10,12,13,15
2. 9,9,10,10,12,13,14,15

Solution:

1. 9,9,10,10,12,13,15 has 7 elements. Therefore our median will be the 4th term in the arranged order
   $\therefore Median = 10$

2. 9,9,10,10,12,13,14,15 has 8 elements. Therefore our median will be the mean of the 4th and 5th terms.
   $\therefore Median = \frac{10+12}{2} = 11$

Problem 4

Finding the median of discrete data.

X	f	cf(cumulative frequency)
1	5	5
2	8	13
3	9	22
4	**12**	**34**
5	6	40
6	7	47
7	4	51
Total	51	

find the value of x which has cumulative frequency just greater than $\frac{n}{2}$

In case of continuous data:

$$Median = l + \frac{\left(\frac{n}{2} - cf\right) h}{f}$$

where cf is the cumulative frequency and f is the frequency of the chosen class, h is the class size

1.3 Mode

The observation which occurs the most is called the mode of the data.
In more general terms, the most probable observation in a dataset is the mode of the data.

Problem 5

Find mode for the following data: 10,11,15,18,18,18,15,10,18,20

Problem 6

Find the mean, median and mode for the following data

CI	f
0-10	3
10-20	5
20-30	7
30-40	2
40-50	1
Total	51

How to find the mode for continuous data

1. Find the modal class which is having the maximum frequency.
2. based on that input the values into the following formulae:

$$mode = l + h\left(\frac{f_1 - f_2}{2f_1 - f_0 - f_2}\right)$$

1.4 The interconnection between the measures of central tendency

$$Mode = 3Median - 2Mean$$

1.5 Geometric and Harmonic mean

Def^n :
Geometric mean is defined as the nth root of the product of n observations
Mathematically:

$$GM = \sqrt[n]{\prod_{i=0}^{n} x_i}$$

Problem 7

Find the Geometric Mean for the values 2,4,8

Def^n :
Harmonic mean is defined as the reciprocal of arithemetic mean of the reciprocal of all the observations

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Theorem:
The following inequality is always true:

$$AM \geq GM \geq HM$$

1.6 Histogram

Histogram can also be used to compute the value of mode.

1.7 Ogive

Ogives are used to compute the value of median. They are nothing but graphs of cumulative distribution functions.
The point of intersection of two ogives gives the median.

1.8 Quartiles

Quartiles are the values which divide the dataset into 4 equal parts. These points are called Q_1, Q_2, Q_3.

$$Q_1 = l + \frac{\left(\frac{n}{4} - cf\right) h}{f}$$

$$Q_2 = l + \frac{\left(\frac{n}{2} - cf\right) h}{f}$$

$$Q_3 = l + \frac{\left(\frac{3n}{4} - cf\right) h}{f}$$

1.9 Deciles

Deciles are the values which divide the dataset into 10 equal parts.

1.10 Percentiles

Percentiles are the values which divide the dataset into 100 equal parts.
To find the xth percentile we can use the following formulae:

$$p_x = l + \frac{\left(\frac{xn}{100} - cf\right) h}{f}$$

2 Measures of Spread/Dispersion

Measures of spread are a numerical quantity to signify the variation in the observations.
Dispersion means the scatterment of observations.
There are a number of ways to get a gist of the Dispersion.

1. Range: it is the difference between the maximum and minimum value of the dataset.
2. Quartile Deviation: It is equal to $\frac{Q_3 - Q_1}{2}$
3. Mean Deviation: It is the arithemetic mean of absolute value of deviation of observations from average.

$$\begin{aligned} MD &= \frac{\sum |x - \overline{x}|}{n} (Discrete) \\ &= \frac{\sum f|x - \overline{x}|}{N} (Continuous) \end{aligned}$$

4. Standard deviation: It is the positive squareroot of arithemetic mean of square of deviation of observations from mean. It is denoted by σ. This is applicable when your mean is an integer.

$$\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}(helpful\ when\ mean\ is\ integral)$$

$$\sigma = \sqrt{\frac{1}{n}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}(helpful\ when\ mean\ is\ non-integral\ and\ x\ is\ small)$$

This is an alternate derivation of the formulae $V = E[X^2] - E[X]^2$

$$d = x - a; \sigma = \sqrt{\frac{1}{n}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)}(helpful\ when\ mean\ is\ non-integral\ and\ x\ is\ large)$$

5. Variance: It is the square of standard deviation.

$$V = \sigma^2$$

2.1 Coefficient of variation

It is another way of measuring the spread. It is analogous to standard deviation with respect to the mean.

$$CV = \frac{\sigma}{\overline{X}} \times 100$$

Suppose there are two sequences with n_1 and n_2 observations.
$\overline{x}_1$ and $\overline{x}_2$ are their means.

their combined variance will be:

$$\sigma^2 = \frac{2}{n_2 + n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$$

where $d_i = \overline{x}_i - \overline{x}$. here $\overline{x}$ is the combined mean of the series.

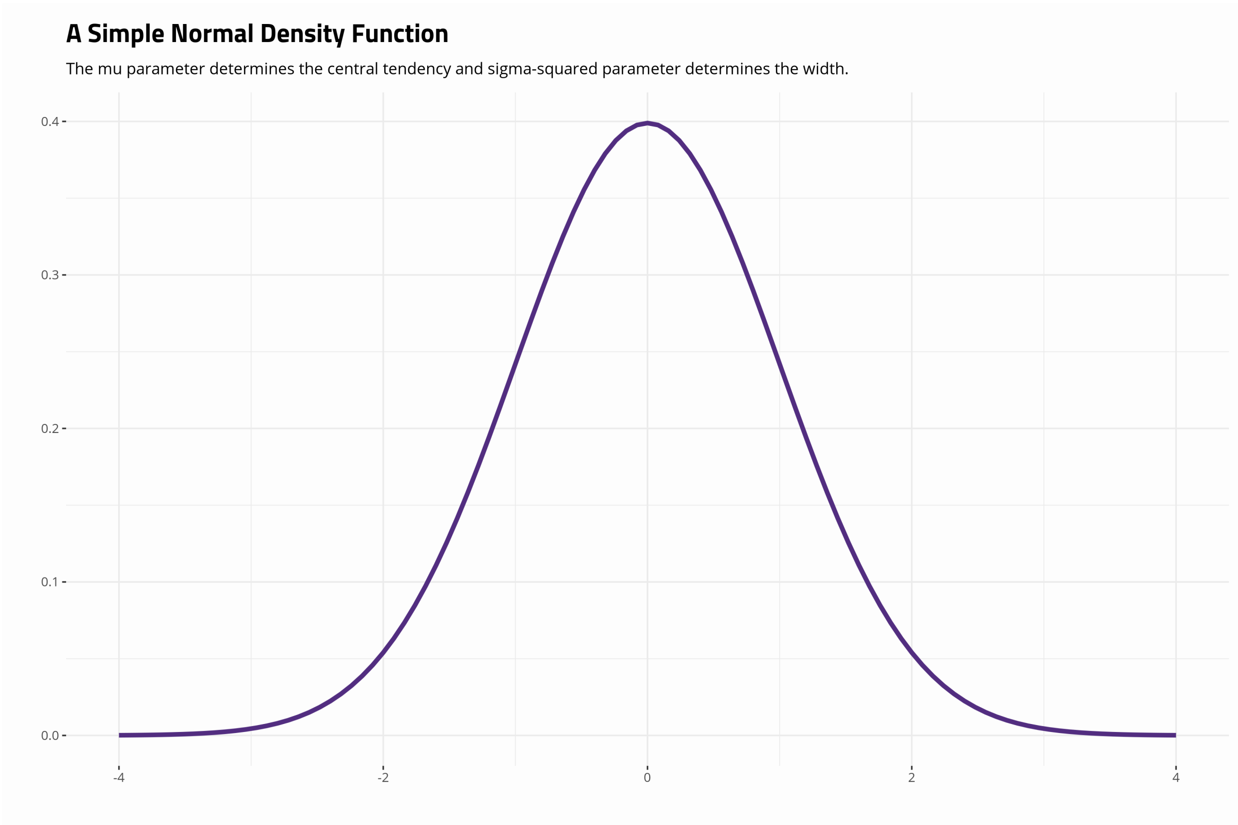
2.2 Skewness and Kurtosis

While studying a distribution we can calculate the measures of central tendency and the measures of spread. However even this information is not enough to determine the behavior of the randow variable distribution. In order to further narrow down in the analysis of the beha vior of the variable we study the skewness of the distribution.

A distribution is called skewed if:

- The Measures of central tendencies do not coincide.
- The curve does not follow gaussian nature.
- The quartiles are not equidistant from the median.

- Some of the positive deviations from median is not equal to the sum of negative deviations from the median.



There can be two types of skewness:

1. Positive skewness (Right skewed)
2. Negative skewness (Left skewed)

$$S_k = Mean - Mode \begin{cases} = 0 & if \ data \ is \ symmetrical \\ > 0 & if \ data \ is \ positively \ skewed \\ < 0 & if \ data \ is \ negatively \ skewed \end{cases}$$

2.2.1 Positive skewness

In this most of the values lie to the right to the peak.
one way to determine this is that $mode > median > mean$

2.2.2 Negative skewness

In this most of the values lie to the left to the peak.
one way to determine this is that $mode < median < mean$

2.2.3 Moments

Arithemetic mean of various powers of deviation of observation from mean.

$$\mu_r = \frac{\sum(x-\overline{x})^r}{n}$$

1. $\mu_1 = 0$
2. $\mu_2 =$ variance
3. $\mu_3 =$ skewness
4. $\mu_4 =$ kurtosis

2.2.4 Coefficient of skewness:

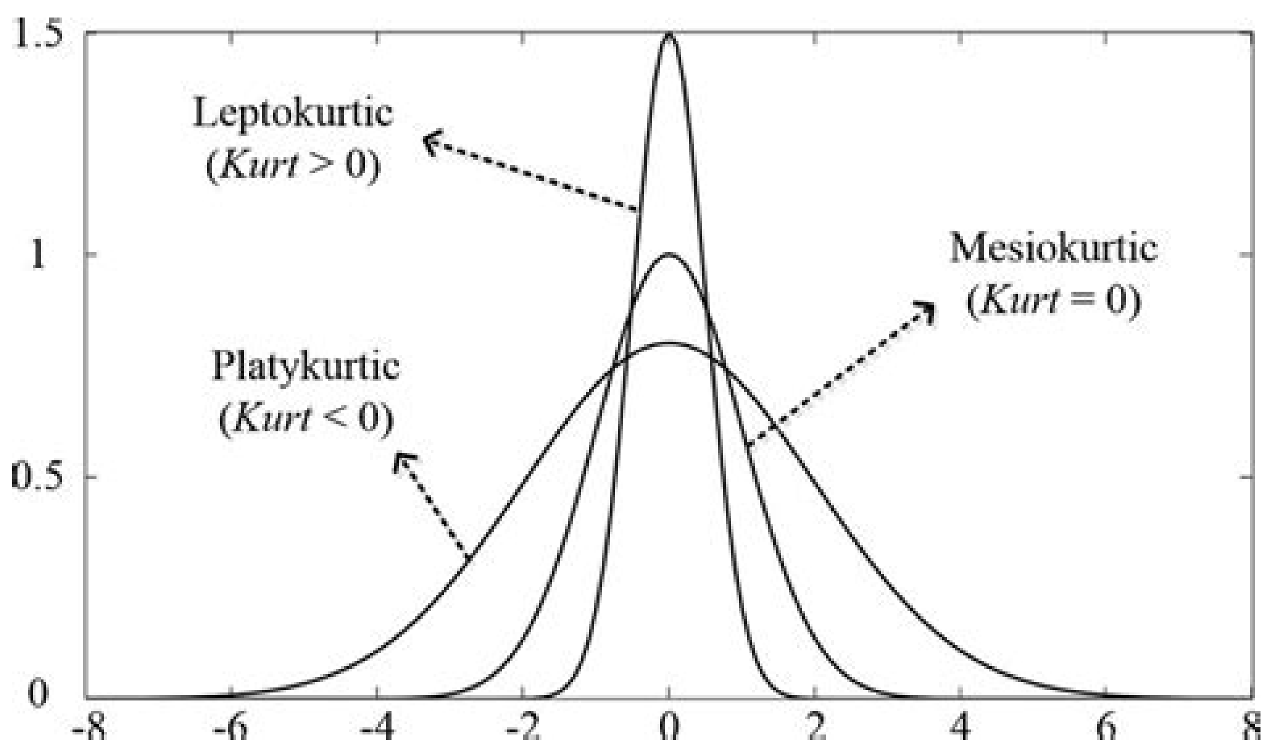
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\gamma_1 = \frac{\mu_3}{(\sqrt{\mu_2})^3}$$

if $\gamma_1 = 0$ the data is symmetrical.
if $\gamma_1 > 0$ the data is right skewed.
if $\gamma_1 < 0$ the data is left skewed.

2.2.5 Kurtosis

It measures the flatness or peakness of the distribution.



$$\beta_2 = \frac{\mu_4}{(\mu_2)^2}$$

$$\gamma_2 = \beta_2 - 3$$

$$Distribution = \begin{cases} Mesokurtic, \ if \ \gamma_2 = 0 \\ Leptokurtic, \ if \ \gamma_2 > 0 \\ Platykurtic, \ if \ \gamma_2 < 0 \end{cases}$$

3 Probability

Probability is the measure of belief that a certain event will occur.

3.1 Prerequisites

Def^n :
Trial:
Suppose an experiment is repeated under identical and homogeneous conditions does not give unique result but may result in one of several possible outcomes. The experiment is known as a random experiment or a Trial.

For example: the toss of a coin.
The outcome of a Trial is known as an **event.**

Def^n :
Exhaustive events:
The set of all the simple events that can occur in a trial are called exhaustive events. The cardinality of the sample space can be found out using basic Combinatorics.

Problem 8

What would be the sample space of the experiment in which 2 coins are tossed together.

Solution:
{HH,HT,TH,TT}

Def^n :
Mutually Exclusive Events:
Events are mutually exclusive if no two or more than 2 events occur simultaneously in the same trial

Mathematically, two events E and F are mutually exclusive $\iff$ $E \cap F = \phi$

Def^n :
Favourable events:
The number of simple events favourable to the happening of an event.

Def^n :
Equally likely events:
This means that events have equal chances of occurence.

Def^n :
Independent events:
Two events are called Independent if the occurence of one event does not influence the occurence of the other.

Mathematically, Two events are independent if

$$P(E \cup F) = P(E) \times P(F)$$

3.2 Definition of Probability

3.2.1 Mathematical or Empirical Probability

The probability of an event E is

$$P(E) = \frac{number\ of\ simple\ events\ favourable\ to\ E}{Total\ number\ of\ simple\ events}$$

Suppose there are total of n number of events and the number of events favourable for a certain event are m. then the probability will be $\frac{m}{n}$.

Then the number of events unfavourable to the same event is $n-m$ and the probability will be $\frac{n-m}{n}$

Axiomatically:

$$P(E) = 1 - P(\overline{E})$$

and

$$0 \leq P(E) \leq 1\ \forall E \subset \Omega$$

3.2.2 Statistical Probability

The probability of an event E is

$$P(E) = \frac{number\ of\ simple\ events\ favourable\ to\ E}{Total\ number\ of\ simple\ events}$$

3.2.3 Axiomatic Probability

Consider a sample space S. The probability is a function that assigns a non-negative value to every event, say A denoted by $P(A)$ is called the probability of an event A if it satisfies the following properties.

1. $P(A) \in [0,1] \forall A \subset S$
2. $P(S) = 1$
3. $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ where each A_i is a mutually exclusive event

3.3 Addition law of Probability

If A and B are two events and are not mutually exclusive, so :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3.4 Multiplication Law of Probability

$$P(A \cup B) = P(A) \cdot P(B|A)$$

if A and B are independent then $P(B|A) = P(B)$ and

$$P(A \cup B) = P(A) \cdot P(B)$$

3.5 Bayes Theorem

If $E_1, E_2, E_3 \ldots E_n$ are n mutually exclusive events with $P(E_i) > 0 \forall i$ then for any event A which is a subset of $\bigcup_{i=1}^{n} E_i$

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{\sum_{i=1}^{n} P(E_i)P(A|E_i)}$$

Here $P(E_i|A)$ is called Posterior Probability.

Problem 8

A class consists of 5 girls and 7 boys. If a comittee of 3 is to be chosen at random what is the probability that

1. 3 boys are selected (Ans = 0.159)
2. exactly 2 girls are selected (Ans = 0.318)

Solution:

1.
2.

3.6 Randow Variables

Axiomatically a Random variable is a variable which takes values at random. A random variable is a variable X that assigns a real number for each and every outcome.

For example, the number of heads in two tosses of a coin.

There are two types of Random variable:

1. Discrete Random Variable:
   A random variable which takes countably many values.
2. Continuous Random Variable:
   A Random variable which takes uncountably many values.

There are two types of probability functios:

1. Probability mass functions:
   The function is said to be a PMF if :
   (a) $0 \leq P(X) \leq 1$
   (b) $\sum_{x \in \Omega} P(X) = 1$
2. Probability density functions:
   It gives a measure of how likely the variable is to lie in the neighbourhood of a point. The function $f_X(x)$ of the numeric values of a continuous randow variable is said to be a PDF if it satisfies:
   (a) $f_X(x) \geq 0 \forall x$
   (b) $\int_{-\infty}^{\infty} f_X(x) = 1$
   (c) $P(a < x < b) = \int_a^b f_X(x)dx$

Problem 9

A coin is tossed 3 number of times. Lets say X be the number of times Head appears.

Solution:

Problem 10

A shipment of 8 microcomputers to a retailer outlet contains 3 defectives. if a school makes a random purchase of 2 computers find the probability distribution for the number of defectives.

Problem 11

In an experiment of tossing 3 coins, obtain the probability distribution of:

1. X denotes the number of heads
2. Y denotes number of head runs
3. Z the number of successive heads
4. $X + Y$
5. XY

3.7 Expectation of Random Variables

Def^n :
if X is a random variable, the expectation of X is denoted as $E(X)$.

$$E(X) = \sum_{x \in \Omega} x f_X(x)$$

$$E(X) = \int_{x \in \Omega} x f_X(x) dx$$

Properties of Expectation;

1. $E(aX) = aE(X)$
2. $E(a) = a$
3. $E(aX + b) = aE(X) + b$

3.8 Variance of a random variable

$$V(X) = E(X^2) - E(X)^2$$

Properties of variance:

1. $V(aX) = a^2 V(X)$
2. $V(a) = 0$
3. $V(aX + b) = a^2 V(X)$

3.9 Standard deviation of a random variable

$$\sigma(X) = \sqrt{V(X)}$$

3.10 Moment generating functions

If X is a random variable, then the moment generating of X, denoted as

$$M_X(t) = \sum_{n} e^{tx} f_X(X)$$

$$M_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx$$