Chapter– 8

STATISTICS AND PROBABILITY MATHEMATICS PORTFOLIO

Ch.Id:-ASU/GRF/EB/RAPOET/2022/Ch-08
DOI: https://doi.org/10.52458/9789391842765.2022.eb.grf.asu.ch-08

[1]Ishita Sharma
[1]Apeejay Stya University, Sohna, Gurugram

[2]Dr. Garima Sharma
[2]Apeejay Stya University, Sohna, Gurugram

To most people statistics generally refers to information about an activity or a process whether it be production, population, national income, etc., that is expressed in numbers. Numbers represent quantities and value of commodities produced and sold, prices of products inventories, assets and liabilities, raw materials, customers, incomes and expenses. Records of births, deaths, number of passengers travelled during a year by road and many more also lead to numerical expressions.

HISTORY AND GROWTH

According to a Greek historian, in 1400 B.C. a census of all the lands in Egypt was taken. Similar reports on the ancient Chinese, Greeks and Romanian are also available. People and land were the earliest objects of statistical enquiry.

The word statistics comes from the Italian word 'Statista' (meaning statesman) or the German word statistic which means a political state. It was first used by Professor Gottfried Achenwall, a professor in Marlborough in 1749 to refer to the subject-matter as a whole. He defined statistics as "the political science of the several countries". The word statistics appeared for the first time in the famous book, Elements of Universal Erundition by Baron J.F. Von Bielfeld translated by W. Hooper M.D (3 vols. 1770). The science of statistics is said to have originated from 2 main sources; government records

2.mathematics While a century ago there were some misgivings among scientists as to whether statistics had the right to be recognized as a distinct science, now almost all sciences are statistical.

STATISTICAL TOOLS

Statistical tools help in concluding various things about a set of given data, make decisions for research accordingly. For example a cosmetic company conducts a market survey on how many people of certain age group prefer to use cosmetic products and of what kind, after concluding data, the company can produce products accordingly and/or can change their market strategies and advertising techniques as per brand requirement and consumer demands. This is all possible through various statistical tools given below.

1. Mean
2. Median
3. Mode
4. Standard Deviation
5. Variance
6. Range
7. Skewness
8. Kurtosis
9. Variability
10. Regression Analysis

We will now understand what these tools are and what are they used for?

1. **Mean:** One of the main goals of statistical analysis is to get one single worth that depicts the attributes of the whole mass of cumbersome information. Such value is known as the central value or an average or the expected value of variable. Though average is defined in various ways by various mathematicians it all means the same thing and represents the single value that represents a group of values. Which is of great significance, because that single value tells us so much more about the whole data- sorted or not. since it addresses the whole information, its value lies some in the middle of between the two limits and on account of this mean or average is in some cases alluded to as a measure of central tendency. Compared to the remainder of the information set by being especially small or large in numerical value.

ARITHMETIC MEAN

The most popular method of representing the average single value of the entire data. It is of two types-

a. Weighted

b. Simple

An arithmetic mean is calculated using the following equation:

$$A := \frac{1}{n}\sum_{i=1}^{n} a_i$$

CALCULATION

Simple arithmetic mean

Individual observations	Discrete series	Continuous series
Direct method	Direct method	Direct method
X¯=∑xN	X¯=∑fx/N	X¯=∑fm/N
Short cut method	Short cut method	Short cut method
X¯=A+∑d/N	X¯=A+∑fd/N	X¯=A+ ∑fd/N

A= assumed mean d=(X-A)

N= total number of observations f= frequency

m= mid-point of various classes

Weighted arithmetic mean

A.M gives equal importance to all the observations. But there are cases that need relative importance for different observations. We use W.M in those cases.

XW¯= ∑WX/ ∑W

XW¯ = weighted arithmetic mean X= variable values

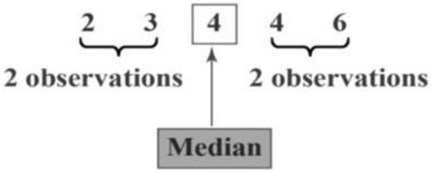
W= weight attached to the variable values

LIMITATIONS OF MEAN

1. The mean cannot be calculated for categorical data, as the values cannot be summed.
2. As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.
3. It also does not give us much insight of the data provided. It considers only

extreme values.

2. **Median:** Median in laymen terms is referred to as the middle value.



Case 1-INDIVIDUAL OBSERVATIONS

1. Arrange the data in ascending or descending order
2. For odd numbered values say 7, add 1 to the total number of values and divide by 2. That valued item would be the median observation.Median = size of 7+1/2 th item= 4th item
3. For even numbered values, the procedure is not as obvious as above. For instance, different values in a group, the median is not really determinable since both the 5th and 6th values are in the center. Thus, we find the value by finding the arithmetic mean of the two middle values- i.e. by adding the two values in the middle and dividing by two.

Median = size of 5th + 6th item /2

Case 2- DISCRETE SERIES

1. Arrange the data in ascending order.
2. Find out the cumulative frequencies.
3. Apply the formula for median N+1/2
4. Now look at the c.f column and find that total which is either equal or next higher to that and determine the value of the variable corresponding to the median value calculated. That gives the value of the median.

Case 3- CONTINUOUS SERIES

Determine the particular class in which the value of median lies. The next step is to find a value which is 50% of the frequencies on one side of it and 50% on other side.

MERITS OF MEDIAN

1. It is useful in case of open-end classes since only the position and not the values of items must be known.
2. Extreme values do not affect the median as strongly as they do the mean.
3. The value of median can be determined graphically whereas value of mean can't be.

Demerits of Medain

1. It is necessary to arrange the data.
2. Since it is positional average, the value is not determined by each and every observation.
3. It is erratic if the number of items is small.
4. The value of median is affected more by sampling fluctuations than the value of the arithmetic mean.

Related Positional Measures

Besides median there are other measures which divide a series into equal parts. These are quartiles, deciles and percentiles. Quartiles (Q) are those values of the variable which divide the total frequency into four equal parts. Deciles (D) divide them into 10 equal parts.

Percentiles (P) divide the total frequencies into 100 equal parts. There are 3 quartiles, 9 deciles and 99 percentiles for a series. They are referred to as measures of dispersion.

3. **Mode:** Mode is the value in a series of observations which occurs with the greatest frequency.

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

l=lower limit of modal class f1=frequency of modal class

f2=frequency of class succeeding modal class f0= frequency of class preceding modal class h=size of intervals

B.DISCRETE SERIES

M.D=∑f|D|/N Steps

a. Calculate the median of the series.
b. Take the deviations of the items from median ignoring signs and denote them by |D|.
c. Multiply these deviations by the respective frequencies and obtain the total ∑f|D|.
d. Divide the total obtained in the previous step by the number of observations. This gives us the value of mean deviation.

4. **Skewness:** Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data

Skp=mean-mode/standard-deviation

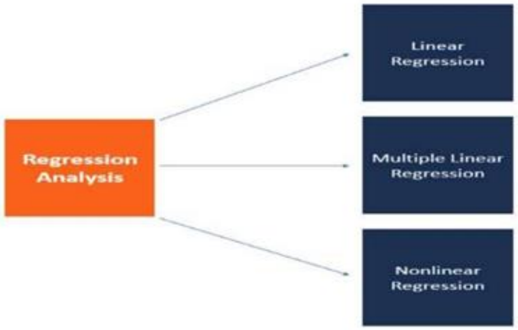
5. **Kurtosis:** For a normal curve value of β2 (measure of kurtosis) is 3 If the value of β2>3 then it is Leptokurtic

   When value of β2<3 then it is platykurtic When value of β2=3 then it is mesokurtic

6. **Variability:** What Is Variability?

   Variability, almost by definition, is the extent to which data points in a statistical distribution or data set diverge—vary—from the average value, as well as the extent to which these data points differ from each other. In financial terms, this is most often applied to the variability of investment returns.

7. **Regression analysis:** Regression analysis determines the extent to which specific factors such as interest rates, the price of a product or service, or particular industries or sectors influence the price fluctuations of an asset. This is depicted in the form of a straight line called linear regression.



8. **STANDARD DEVIATION**

$$\mathrm{SD} = \sqrt{\frac{\sum |x-\mu|^2}{N}}$$

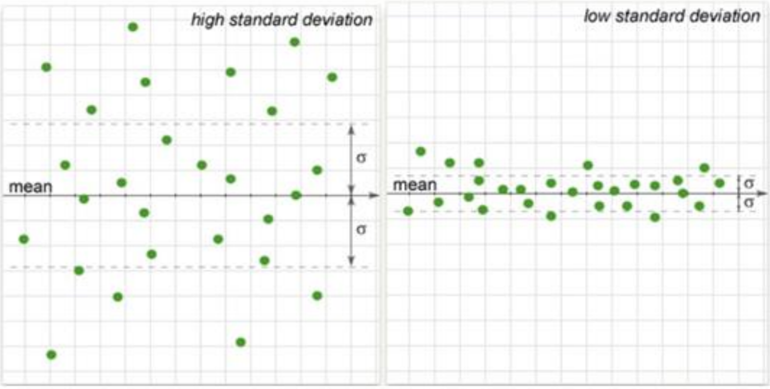
Where μ is the mean of the data Steps to find Standard Deviation are:

Step 1: Find the mean.

Step 2: For each data point, find the square of its distance to the mean.

Step 3: Sum the values from Step 2.

Step 4: Divide by the number of data points. Step 5: Take the square root.



STATISTICAL ANALYSIS OF DATA

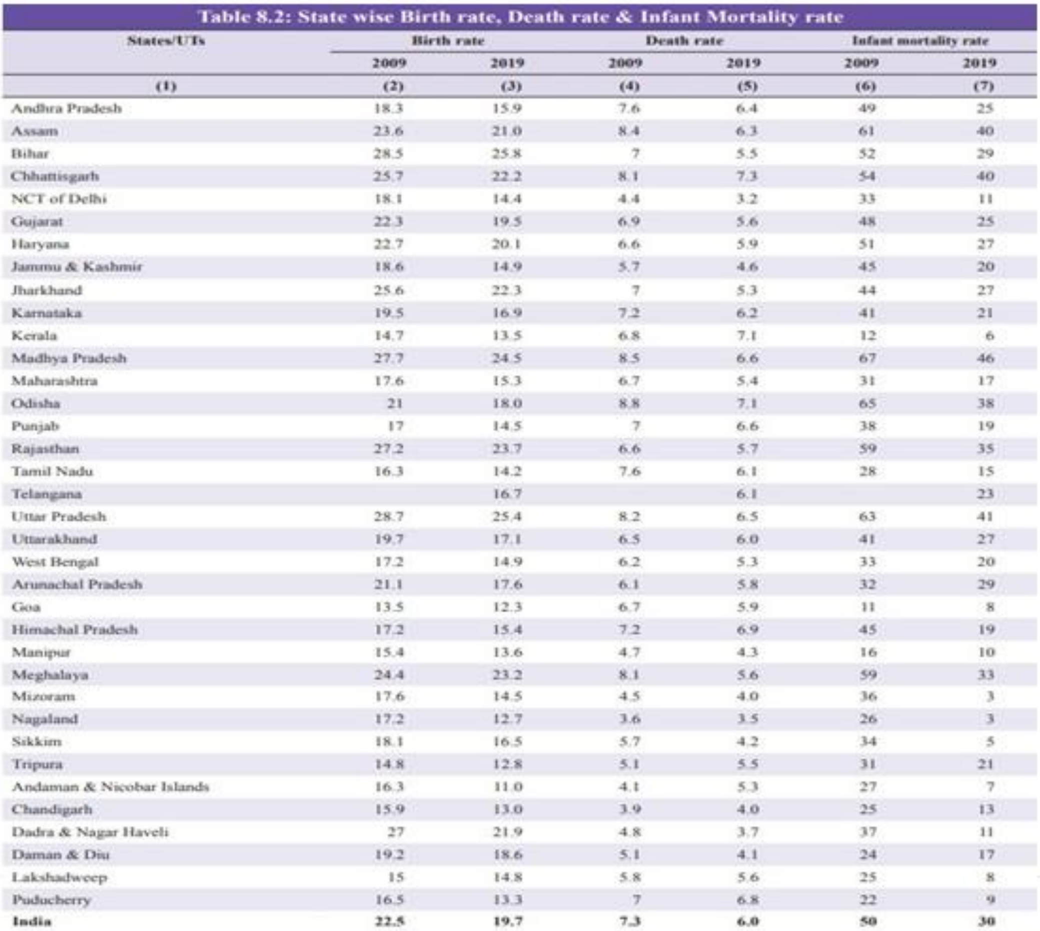
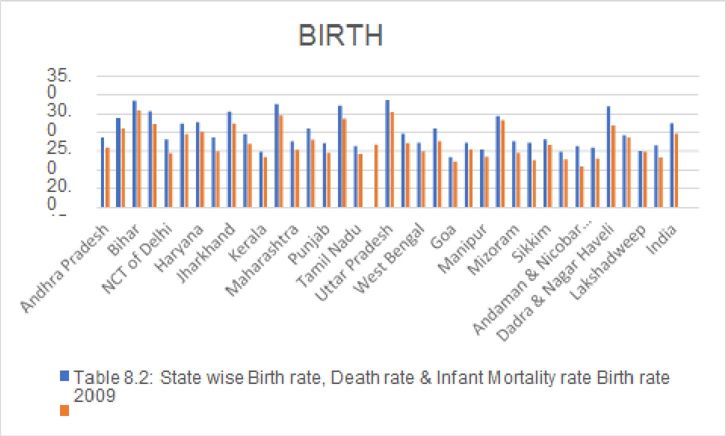


Table 8.2: State wise Birth rate, Death rate & Infant Mortality rate

States/UTs	Birth rate		Death rate		Infant mortality rate	
	2009	2019	2009	2019	2009	2019
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Andhra Pradesh	18.3	15.9	7.6	6.4	49	25
Assam	23.6	21.0	8.4	6.3	61	40
Bihar	28.5	25.8	7	5.5	52	29
Chhattisgarh	25.7	22.2	8.1	7.3	54	40
NCT of Delhi	18.1	14.4	4.4	3.2	33	11
Gujarat	22.3	19.5	6.9	5.6	48	25
Haryana	22.7	20.1	6.6	5.9	51	27
Jammu & Kashmir	18.6	14.9	5.7	4.6	45	20
Jharkhand	25.6	22.3	7	5.3	44	27
Karnataka	19.5	16.9	7.2	6.2	41	21
Kerala	14.7	13.5	6.8	7.1	12	6
Madhya Pradesh	27.7	24.5	8.5	6.6	67	46
Maharashtra	17.6	15.3	6.7	5.4	31	17
Odisha	21	18.0	8.8	7.1	65	38
Punjab	17	14.5	7	6.6	38	19
Rajasthan	27.2	23.7	6.6	5.7	59	35
Tamil Nadu	16.3	14.2	7.6	6.1	28	15
Telangana		16.7		6.1		23
Uttar Pradesh	28.7	25.4	8.2	6.5	63	41
Uttarakhand	19.7	17.1	6.5	6.0	41	27
West Bengal	17.2	14.9	6.2	5.3	33	20
Arunachal Pradesh	21.1	17.6	6.1	5.8	32	29
Goa	13.5	12.3	6.7	5.9	11	8
Himachal Pradesh	17.2	15.4	7.2	6.9	45	19
Manipur	15.4	13.6	4.7	4.3	16	10
Meghalaya	24.4	23.2	8.1	5.6	59	33
Mizoram	17.6	14.5	4.5	4.0	36	3
Nagaland	17.2	12.7	3.6	3.5	26	3
Sikkim	18.1	16.5	5.7	4.2	34	5
Tripura	14.8	12.8	5.1	5.5	31	21
Andaman & Nicobar Islands	16.3	11.0	4.1	5.3	27	7
Chandigarh	15.9	13.0	3.9	4.0	25	13
Dadra & Nagar Haveli	27	21.9	4.8	3.7	37	11
Daman & Diu	19.2	18.6	5.1	4.1	24	17
Lakshadweep	15	14.8	5.8	5.6	25	8
Puducherry	16.5	13.3	7	6.8	22	9
India	**22.5**	**19.7**	**7.3**	**6.0**	**50**	**30**

CONCLUSION

A. Birth rate 2009

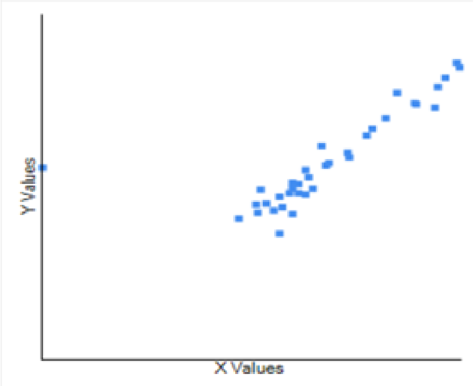


By calculating the above data it is inferred that:

1. Average birth rate in 2009 was 19.422
2. Median being the middle value came out to be 18.2
3. The most frequently occurring value was 17.2 being in west Bengal, himachal Pradesh and Nagaland.
4. Skewness is -0.77 is less than -0.5 then we can say that the sample is negatively skewed.
5. Kurtosis being greater than 3 implies that it is leptokurtic.
6. Standard deviation being 5.436 implies that the data points are close to the mean.
7. Mean deviation came out to be 0.20 implies how far are the values from the mean of the given data set.
8. Range being 28.7 signifies variability of data
9. Variance being 30.4 is a large spread in relation to mean. 2019
    1. Mean in 2019 for birth rate was 17.27
    2. Standard deviation resulted to be 4.12 signifying the data is close to the mean.
    3. Variance is 17 meaning it has greater variability.
    4. Range 14.8 signifies greater variability of data
    5. Skewness 0.611 is less than 0.5 signifies that it is positively skewed
    6. Kurtosis 4.12 is more than 3, concluding it to be leptokurtic.

7. Median 16.2 is the middle value.
8. Mode 14.5, 14.9, since there are 2 modes it is a bimodal set of data.
9. Mean deviation 3.43, signifies the values are not that far from the mean as compared to 2009 data

Correlation between the birth rate of 2009 and 2019 came out to be 0.786 implying that the relationship of one variable with another which implies less effect on each other.



After comparison it is safe to say that the average birth rate has decreased in the10 years.

The range being much greater in 2009 signifies wide spread of data that means the figures for different states must have been varied, meaning some states must have had greater birth rates as compared to some other which have had lesser, the difference between those being larger as compared to that observed in 2019, the data is less spread.

Highest birth rate in 2009 was in Uttar Pradesh whereas in 2019, it was observed to be in Bihar. Similarly concluding for the rest of the data.

2. Death rate 2009

Mean came out to be 6.4 of the data. Median as 6.7

Mode is 14.5 which is the most frequently occurring value in Punjab and Mizoram. Standard deviation is 4.12

Variance 1.96

Range 5.2

Skewness -0.3

Kurtosis 2.2, Positive excess values of kurtosis (>3) indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values.

Mean deviation 6.4

2019

Mean 5.55

Median 5.65

Mode 5.3, 5.6

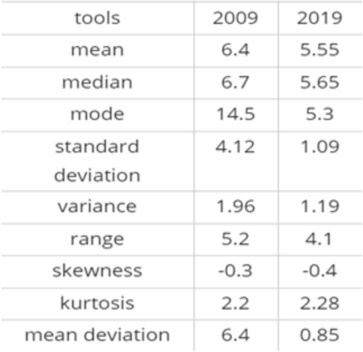
Standard deviation 1.09

Variance 1.19

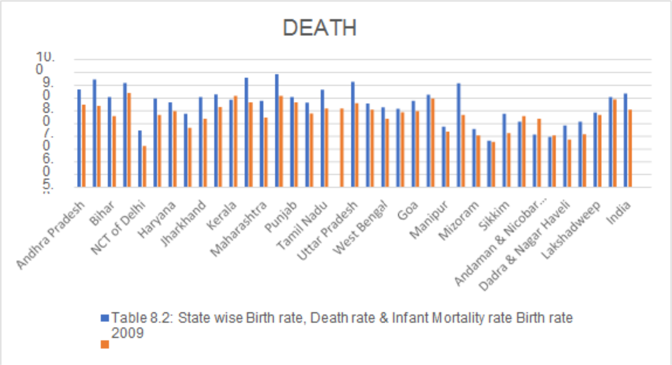
Range 4.1

Skewness -0.4

Kurtosis 2.28, less than 3 indicates the data being platykurtic. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution Mean deviation 0.85 which indicates that the data is not much deviated from the mean.

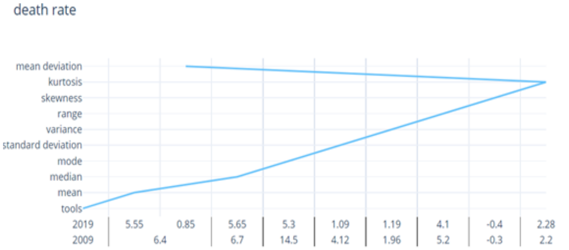


tools	2009	2019
mean	6.4	5.55
median	6.7	5.65
mode	14.5	5.3
standard deviation	4.12	1.09
variance	1.96	1.19
range	5.2	4.1
skewness	-0.3	-0.4
kurtosis	2.2	2.28
mean deviation	6.4	0.85

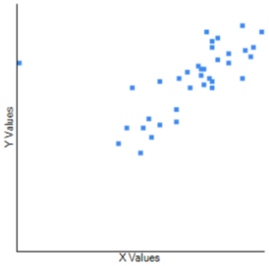


Mean is higher for 2009 which mean the average number of deaths was higher in 2009 than in 2019.

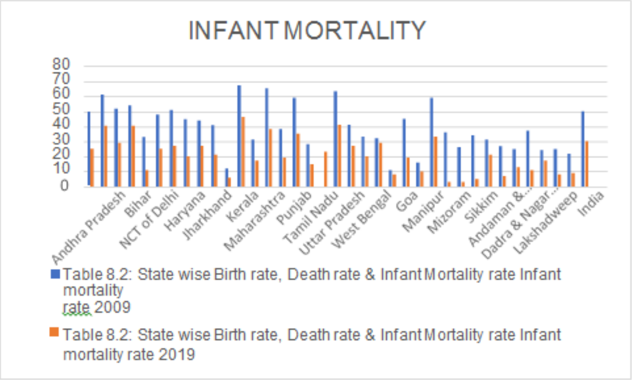
The range for both the data does not have much difference. Skewness for both sets of data is negative, that means the data is negatively skewed, which indicates the mean and the median are both less than the mode.



Highest death rate in 2009 was in Odisha whereas in 2019 it was in Chattisgarh. Correlation 0.6063, which again indicates negative association, that is, as the value of one variable increases, the value of the other variable decreases or simply that the value doesn't affect the other in this case as these are data of two totally different years, nothing connecting them as such.

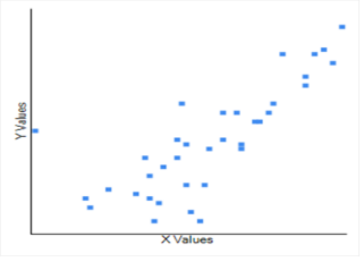


INFANT MORTALITY RATE



tools	2009	2019
mean	39	20.77
median	37	20
mode	25.31	27
standard deviation	15.15	11.85
variance	223.25	140.46
range	56	43
skewness	0.125	0.33
kurtosis	2.11	2.12
mean deviation	12.57	9.56

Infant mortality rate for 2019 is lesser than in 2009 which signifies lesser infant mortality rate average. Kurtosis value for both is less than 3 which indicates it is platykurtic, less peaked value. Skewness is positive for 2009 but negative for 2019. Range is higher for 2009 data which means there was more variability. Correlation was calculated to be 0.7812.



Highest infant mortality rate in 2009 was in Madhya Pradesh whereas lowest was in Goa. In 2019 highest IMR was in Madhya Pradesh and lowest was in Mizoram and Nagaland.

BIBLIOGRAPHY

- *Factors Responsible for High Growth of Population in India - Essay (shareyouressays.com)*
- *ANALYSIS OF BIRTH RATE AND DEATH RATE CASES – EasyProjectMaterials*
- *Kurtosis (Definition, Significance) | 3 Types of Kurtosis (wallstreetmojo.com)*
- *Economic Survey 2022 Pdf Download India |* **आर्थिक सर्वे** *2022 (studydhaba.com)*
- *Kurtosis - Definition, Excess Kurtosis, and Types of Kurtosis (corporatefinanceinstitute.com)*
- *India demographics 2021 - StatisticsTimes.com*
- *Is 0.5 A good correlation coefficient? (findanyanswer.com)*
- *Correlation: Meaning, Types and Its Computation | Statistics (yourarticlelibrary.com)*
- *Birth Rate and Death Rate in India (Statistics) (economicsdiscussion.net)*