

(SHORT COMMUNICATION)



## Biostatistics: Simple practical way to use Microsoft excel to find mean, median, mode, standard deviation and correlation

Tapan Kumar Mahato\*

*Department of Pharmaceutical Chemistry, B. Pharmacy College Rampura, At Rampura, PO Kakanpur, Taluka Godhra, District: Panchmahal, Gujarat 388713, India.*

World Journal of Biology Pharmacy and Health Sciences, 2023, 13(03), 150–155

Publication history: Received on 04 February 2023; revised on 15 March 2023; accepted on 17 March 2023

Article DOI: <https://doi.org/10.30574/wjbphs.2023.13.3.0128>

### Abstract

India is among the top five countries in the pharmaceutical sector. Software is used more frequently because it completes the work in very short time in terms of analyzing data, calculations etc. Computers and cutting-edge software are increasingly used in research. Application of mathematics to biological systems is known as biostatistics (living things). Because biostatistics is employed in both experimental and observational investigations, it is a subject worth studying in the pharmaceutical sector, medical and paramedical (nursing, pharmacy, etc.) schools. Use of Pen and paper for measurements and calculations take far too long to complete. Software is the option to reduce the time. There is a wide variety of statistical software available today that produces findings quickly. Some examples of statistical software are Minitab, SPSS, R online, and Microsoft Excel. Microsoft excel is free and widely accessible software to calculate central tendency, dispersion and correlation. In this article we will discuss the use of Microsoft excel to find mean, median, mode, standard deviation and correlation of the given data.

**Keywords:** Mean; Median; Dispersion; Standard deviation; Central tendency; Karl pearson's coefficient of correlation

### 1. Introduction

Data collection, analysis, interpretation, and presentation are all topics covered by statistics. The objective of statistics is to understand the data, not to conduct multiple computations using formulas.

#### 1.1. Biostatistics

The application of statistics to the biological or medical sciences is known as biostatistics. Biostatistics is credited to Francis Galton as its father. Correlation is a statistical term he invented. It is employed when dealing with statistics in the fields of biology, medicine, nursing, pharmacy, and public health, among other health sciences. Depending on whether applications are in the health sciences (Biostatistics) or in broader biology (biometry), such as agriculture, ecology, or wildlife biology, biostatistics may be distinguished from biometry. The biologist can derive general laws from small samples and comprehend the nature of variability with the use of statistics.

#### 1.2. The central tendency (Mean, Median, and Mode)

It is a statistical metric that establishes a single value that precisely characterizes the distribution's centre and represents the complete range of scores.

#### 1.3. Mean

Arithmetic Mean is a value created by dividing the total number of observations by the total number of observations.

\*Corresponding author: Tapan Kumar Mahato

#### **1.4. Median**

The variable's median positional value divides the distribution into two equal parts: one half includes all values larger than or equal to the median value, while the other portion includes all values less than or equal to it. It could be regarded as the "middle" value for a data set.

#### **1.5. Mode**

A measurement with a reasonably high concentration is considered to be in the mode when it occurs the most frequently in a set of observations. The value that appears the most frequently in a collection of measurement of values is indicated by the symbol  $M_o$ . In other words, it is the value that is used the most in a particular set.

#### **1.6. Dispersion**

The term "dispersion" describes how the objects differ from one another and from the average. The more a series' products vary from one another, the more dispersion there will be. A.L. Bowley claims that dispersion is a measure of the items' variety.

#### **1.7. Standard deviation**

Karl Pearson first developed the idea of standard deviation in 1894. The most practical and often used measure of dispersion among populations is the standard deviation. The Greek letter sigma ( $\sigma$ ) is used to represent it. The standard deviation considers the value of each observation, just like the mean deviation does.

#### **1.8. Karl Pearson's coefficient of correlation**

The correlation coefficient  $[r]$ , often known as Pearson's correlation coefficient, was developed by Karl Pearson. Alternatively it is called as the Product Moment Correlation Coefficient. Together with the Scatter Diagram and Spearman's Rank Correlation, it is one of the three most effective and widely used techniques for determining the degree of correlation. In order to quantify the strength of the linear relationship between X and Y, the Karl Pearson correlation coefficient method is used. Such a correlation's coefficient is denoted by the letter "r".

---

## **2. Statistical tools**

### **2.1. Microsoft excel**

Excel is a popular statistics application that may be used to examine manually calculated answers to homework problems as well as to grasp statistical principles. It gives practice using Excel for basic statistical analysis and for presenting data summaries. This covers tabulating data, creating pivot tables, and creating graphics—basic data management. From version 5 in 1993, it has been the most popular spreadsheet programme on this platform. The Microsoft Office suite includes Excel.

### **2.2. SPSS (Statistical Package for Social Sciences)**

With just one click, the highly interactive software SPSS can carry out extremely complicated data manipulation and analysis. Researchers of all stripes utilize it for sophisticated statistical data analyses. For the management and statistical analysis of social science data, the SPSS software package was developed. It was created by IBM Corporation employees Norman H. Nie and C. Hadlai Hull and released by SPSS Inc. in 1968. IBM later purchased SPSS Inc. in 2009. To create tabulated reports, charts and plots of distributions and trends, summarize statistics, and carry out extensive statistical analysis, SPSS can accept data from practically any form of file.

### **2.3. Minitab**

At Pennsylvania State University in the United States, researchers Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner created Minitab, a potent statistical programme, in 1972. It started out as OMNITAB 80's lighter iteration. The majority of Minitab's uses are in statistical analysis and research. Accuracy in analysis, dependability of outcomes, and quicker speed are all advantages of using Minitab for statistical analyses. With the use of its sophisticated graphs, charts, and other exploratory tools, it aids in exploratory data analysis. Most analyses benefit from the menu system. It can open numerous file kinds, including text files, HTML files, and Excel worksheets.

## 2.4. R Online

Ross Ihaka and Robert Gentleman from the University of Auckland in New Zealand created the free software and computer language R in 1993. For the creation of statistical software and data analysis, statisticians frequently utilize R. It was developed using the S language.

## 2.5. Question

Here are listed the ages and weights of the first 10 patients consulted on Monday in a hospital outpatient department (OPD). With the provided data, calculate the mean, median, mode, standard deviation, and Karl Pearson's correlation coefficient.

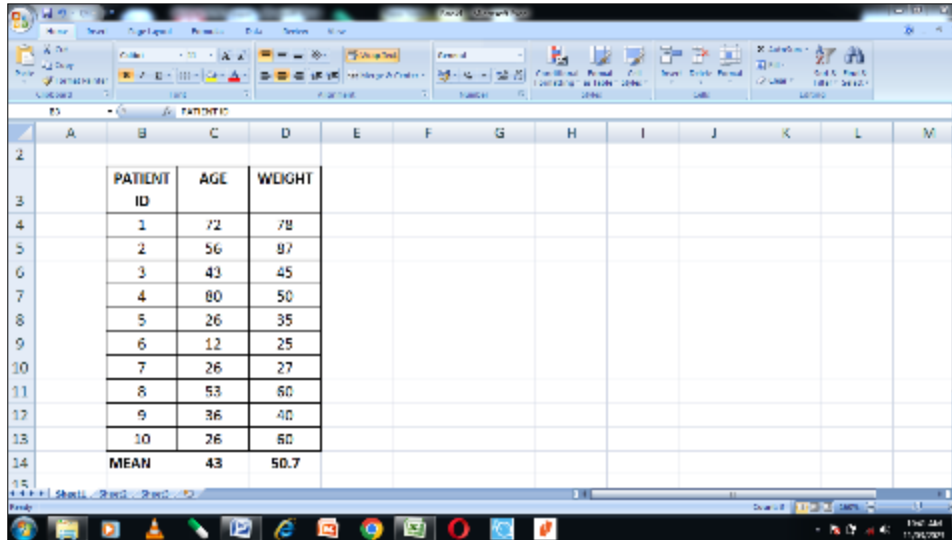
**Table 1** The ages and weights of the 10 patients consulted on Monday in a hospital outpatient department (OPD)

Patient ID	Age (In years)	Weight (In kg)
1	72	78
2	56	87
3	43	45
4	80	50
5	26	35
6	12	25
7	26	27
8	53	60
9	36	40
10	26	60
Mean	?	?
Median	?	?
Mode	?	?
SD	?	?
Correlation	?	

## 2.6. Solution

### 2.6.1. Calculation of mean

- Open an Excel spreadsheet and enter the data in the rows and columns supplied.
- Align the cells as needed.
- Put mean at the bottom of the patient ID column.
- In the next box, type =AVERAGE(select all data from the age column) and press enter on the keyboard. In that box, the calculated mean of age will be presented.
- Pick that box and move the point to the next empty weight column box.
- The calculated weight mean will be presented in that box.
- Answer = 43 years (Age) AND 50.7 kg (Weight)

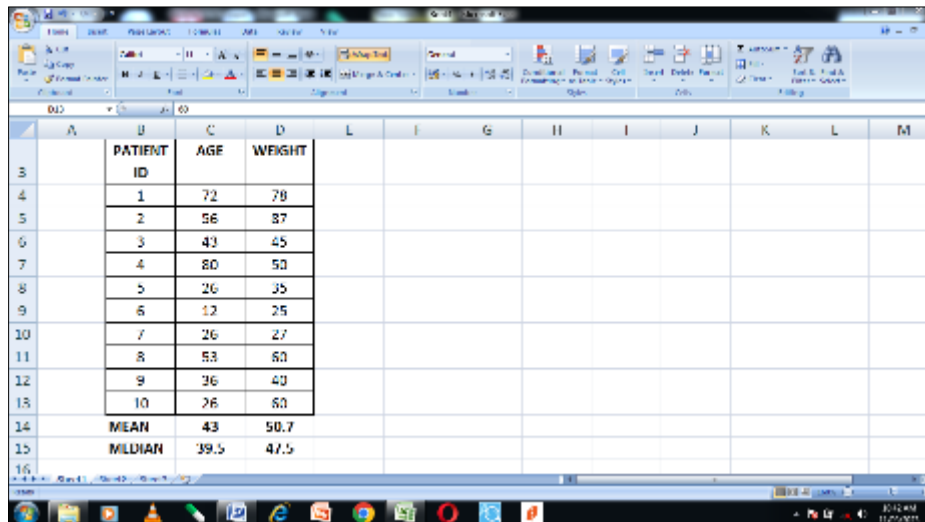


vii. Mean = 43 (Age) & 50.7 (Weight)

**Figure 1** Calculation of mean using excel spreadsheet

### 2.6.2. Calculation of median

- Launch Excel and enter the supplied data in rows and columns.
- If necessary, align the cells.
- Indicate the median at the base of the patient ID column.
- In the next box, type =MEDIAN(select all the values in the age column), and then hit the enter key on your keyboard. The calculated age median will be shown in that box.
- Choose that box and move the point to the weight column's following empty box.
- The calculated weight median will be shown in that box.
- Answer: 39.5 years (Age) and 47.5 kg (Weight)



vii. Answer: 39.5 (Age) and 47.5 (Weight)

**Figure 2** Calculation of median using excel spreadsheet

### 2.6.3. Calculation of mode

- Open an Excel spreadsheet and enter the data in the rows and columns supplied.
- Align the cells as needed.
- Write mode at the bottom of the patient ID column.

- Choose the next box and type =MODE(select all values from the age column) and hit enter on the keyboard. In that box, the calculated mode of age will be displayed.
- Pick that box and move the point to the next empty weight column box.
- The calculated mode of weight will be shown in that box.
- Answers = 26 years (Age) & 60 kg (Weight)

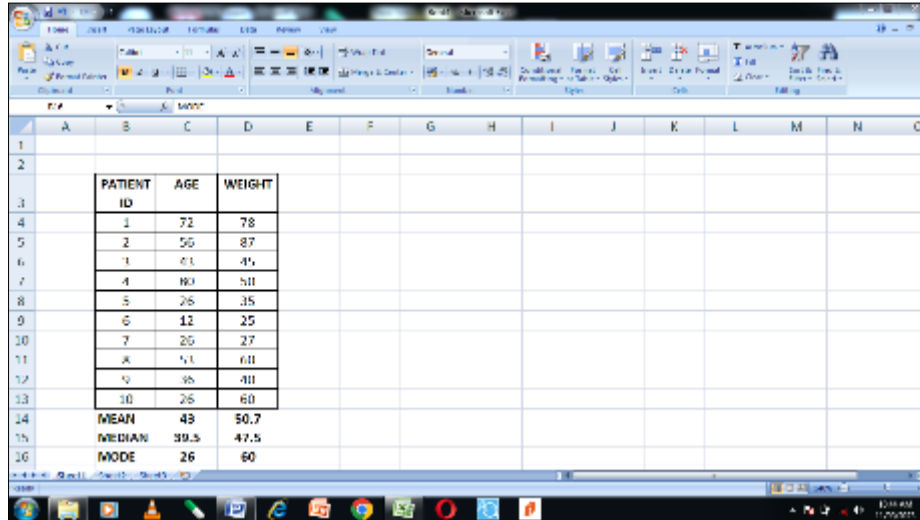


Figure 3 Calculation of mode using excel spreadsheet

2.6.4. Calculation of standard deviation (SD)

- Launch Excel and enter the supplied data in rows and columns.
- If necessary, align the cells.
- Add SD to the patient ID column at the bottom.
- In the next box, type =STDEVA(select all the information in the age column), and hit the Enter key on your keyboard.
- The calculated standard deviation of age will be shown in that box.
- Choose that box and move the point to the weight column's following empty box.
- The calculated SD of weight will be shown in that box in item
- Answer: 22 years (Age) and 20.6884 kg (Weight)

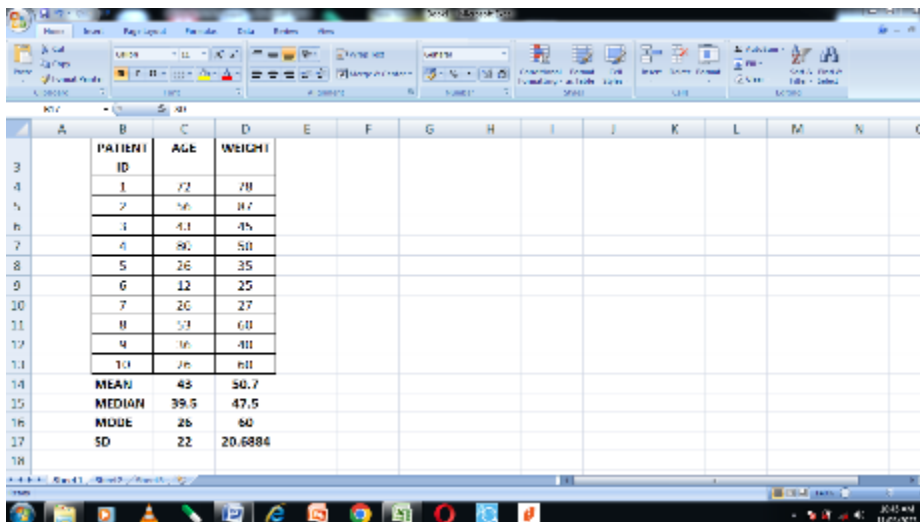


Figure 4 Calculation of standard deviation using excel spreadsheet

### 2.6.5. Calculation of Karl Pearson's coefficient of correlation

- Launch Excel and enter the supplied data in rows and columns.
- If necessary, align the cells (ii).
- In the patient ID column, at the bottom, write correlation.
- Go on to the following empty box and choose the  $f_x$  option listed at the top of the excel page.
- A dialogue box will show up when you type CORREL into the function's field and press the GO button. The CORREL option will be shown; choose it.
- A dialogue window will once more appear, asking for arrays 1 and 2. For array 1, select all data of age column.
- For array 2, choose all of the weight column's data. Hit "OK."
- In that box, the calculated correlation between age and weight will be visible.
- Correlation is 0.66255 in the answer. From the obtained answer, it can be concluded that age and weight have a reasonably high link with one another (moderately strong correlation).

	PATIENT ID	AGE	WEIGHT
3			
4	1	72	78
5	2	55	67
6	3	65	65
7	4	80	50
8	5	26	35
9	6	12	25
10	7	26	27
11	8	55	60
12	9	15	20
13	10	26	60
14	MEAN	48	50.7
15	MEDIAN	39.5	47.5
16	MODE	26	60
17	SD	22	20.6884294
18	CORRELATION		0.66255

Figure 5: Calculation of Karl Pearson's coefficient of correlation using excel spreadsheet

### 3. Conclusion

Because of their benefits such as simplicity, time savings, and reduced labor, technologies are being used more and more frequently. If a person learns how to manage and use software, he can complete the task accurately and in a short amount of time. Biostatistics is a vital component of research. It is popular and widely used to quantify central tendency using the terms mean, median and mode as well as to measure dispersion using the term standard deviation. In order to calculate this with pen and paper, it takes a lot of time. Although there are several statistical software accessible including SPSS, R Online and Minitab, Microsoft Excel is the most popular choice since it is free and easy to use, which saves time, labor and money. Microsoft Excel is an easy and calculates biostatistical problems in a straightforward manner like central tendency, dispersion and correlation.

### Compliance with ethical standards

#### Acknowledgments

I would like to hearty thanks the authors of the text book mentioned in the reference because of whom I was able to write this article.

### Reference

- [1] Singh G, Mahato TK. Biostatistics and Research Methodology. 1<sup>st</sup> ed. Pune: Technical publications; 2022.