CBS HANDBOOK ON BIOSTATISTICS FOR NURSES

Special Features

- First precise and concise book on Biostatistics
- Includes 50+ examples with solutions for easy understanding of the concepts
- Analytical and Research-based Approach
- Each chapter discussed in quick review format
- Useful for all nursing examinations

CBS Publishers & Distributors Pvt. Ltd.

Mukhmohit Singh
Shveta Saini

CBS Handbook on BIO STATISTICS FOR NURSES

Mukhmohit Singh

MBBS, MD (Community Medicine) PGDHS
Gold Sp Award (2012) ERS
Ex Assistant Professor
Community Medicine and Public Health
Preventive Medicine and Public Health Specialist
Formerly
Epidemiologist – IDSP, UT Administration, Chandigarh
Research Officer, PGIMER, Chandigarh
Member
Indian Public Health Association
Indian Association of Preventive and Social Medicine
Research Society for Study of Diabetes
European Respiratory Society

Shveta Saini

MBBS, MD (Community Medicine)
Ex Assistant Professor
Community Medicine and Public Health
Preventive Medicine and Public Health Specialist
Member
Indian Public Health Association
Indian Association of Preventive and Social Medicine

CBS
Dedicated to Education

CBS Publishers & Distributors Pvt Ltd

• New Delhi • Bengaluru • Chennai • Kochi • Kolkata • Mumbai
• Hyderabad • Nagpur • Patna • Pune • Vijayawada

Preface

It gives us immense pleasure and happiness to present the book — *CBS Handbook on Biostatistics for Nurses*.

For us, medicine is so much relative and dynamic that it changes with an unbelievable pace. The guidelines or cut off values, usefulness of new investigations, newer drugs for intervention and management — these all happen exponentially with each passing day!

Conducting research, Research methodology, Data compilation, and Analyzing the results all have been always an art and an area of exploration with endless limits, innovations and concepts. We are firm believers that it may be understood using strong concepts and foundations, rather than cramming the dry formulas and numerical-based questions.

The basic foundation of this book lies in the fact that concepts should be presented in the simplest manner not just for learning but for retention in long-term memory and as the name goes — for conceptual understanding.

A different approach has been used in the book by giving a touch of basics and hint of the concepts in each chapter. This is followed by solved MCQs and a number of numericals to be solved by the user for better and almost complete understanding. The language and the level of biostatistical understanding have been kept simple to enable the reader to gain a positive sense of satisfaction after finishing this book.

This book essentially starts with the most basic understanding of "WHY" Biostatistics and takes the course of understanding the types of data, measures of central tendency, measures of variation and inferential statistics. The particular sequence has been researched and found to be effective in building a concept of analytical and research-based approach to the field of nursing. Assessment of investigations or screening tools, which will also be useful for most of the nursing examinations has been added to understand Assessment in Research.

As readers, we value your feedback. We also know how important this publication is for many students across the country and abroad. Although, all the efforts have been put in for providing an error-free manuscript and maintaining the quality of content, some lapses cannot be denied.

For any query, difference of opinion or suggestions, please feel free to write to us at dr_mukhmohit5@hotmail.com

All contributions will be duly acknowledged.

Subscribe to our Student Support for Recent advances and updates in PSM:

- **YouTube channel** – https://www.youtube.com/c/DrMukhmohitsinghsCommunityMedicineSimplified
- **Telegram:** https://t.me/mukhmohit01
- **Facebook:** https://www.facebook.com/groups/mukhmohit.community.medicine
- **Website:** https://psmsimplified.com

Mukhmohit Singh

Shveta Saini

Contents

CHAPTER 6

Correlation and Regression

CORRELATION

In technical terms, correlation is the statistical association between two variables. Correlation would help in understanding the association or degree of dependency of a variable on the other variable. It helps in understanding the direction and degree of relationship.

Correlations are used for validity, reliability and hypothesis validation of data or set of data.

Example: The blood pressure is correlated with triglyceride (TG) levels. As the TG levels increase, the blood pressure tends to increase (Fig. 1).

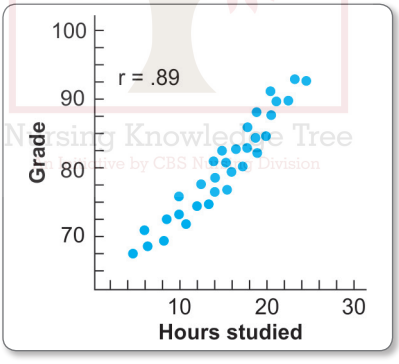


Fig. 1: Correlation

The linear correlation is often represented by the symbol 'r', known as Pearson correlation.

The value of 'r' ranges from –1 to + 1

'r' value = –1 is perfect negative correlation

'r' value = +1 is perfect positive correlation

Coefficient of determination is given by the formula 'r' square. The square of the coefficient (or r^2) is equal to the percent of the variation in one variable that is related to the variation in the other. If r = 0.5, it means that 26% of variation in the outcome variable (or dependent) is due to a unit change in the input (or independent) variable.

Example: From a conceptual exam, the conceptual and logical reasoning score of students was plotted after observing the students for average number of hours the student slept in last 6 nights (Fig. 2). We can see that the correlation curve does not follow a straight line, but a curvilinear approach or non-linear curve is formed.

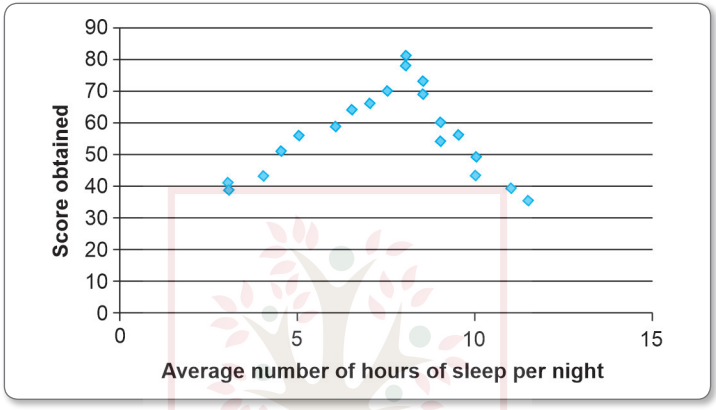


Fig. 2: Curvilinear correlation

REGRESSION

In regression, we predict one value from another variable.

Let's take an example of how does meteorological department tells us the probability of rainfall or the temperature of our city after two months or even after three months. The point is, they are able to tell about the weather based on data from previous many years regarding cloud type, wind velocity, air temperature, humidity and probability of rainfall.

Regression is of two types:

1. Linear regression—for parametric data (quantitative data)
2. Logistic regression—for nonparametric data (qualitative data)

Since all the parameters given in the MCQ are dichotomous data (binomial or two groups) data, the regression we may use is multivariate (-multiple variables) logistic (-nonparametric data) regression.

Solved Multiple Choice Questions

1. A study determined the correlation coefficient (r) = +0.82, between low-density lipoprotein (LDL) levels and mean systolic blood pressure. Which of the following statements is a wrong interpretation of the correlation coefficient observed?

a. Since there is a high correlation, the magnitudes of both the measurements are likely to be close to each other.
b. A patient with a high level of systolic BP is also likely to have a high level of serum LDL.
c. A patient with a low level of systolic BP is also likely to have a low level of serum LDL.
d. About 65% of the variation in systolic blood pressure among the patients can be explained by their serum LDL values.

Ans. (a) Since there is a high correlation, the magnitudes of both the measurements are likely to be close to each other.

As per the research in the MCQ, the LDL and systolic blood pressure (SBP) are correlated with r = 0.82.

Points to remember:

Correlation maybe of two types:

- Pearson correlation (r) – for ungrouped data
- Spearman correlation (ρ) – for ordinal, grouped data

The value of r may range from –1 to +1

Remember:

- If r = –1 = Perfect negative correlation
- If r = +1 = Perfect positive correlation

Coefficient of determination (CoD) = r × r × 100 (expressed in percentage)

CoD shows the percentage variability in the dependent variable, which can be expected because of a unit change in the dependent variable.

Interpretation:

In the given MCQ, r = +0.86, hence it is a strong positive correlation. The CoD would be 0.65 or 65% of the change in blood pressure is due to change in the LDL levels.

2. If we check for the role of variables as: Smoker (Yes/No), Obese (Y/N), Physical activity (Y/N), Hypercholesterolemia (Y/N) and we find the effect of the variables on prediction for development of myocardial infarction.

a. Pearson correlation
b. Dixon's Q test
c. Multivariate logistic regression
d. Curvilinear correlation

Ans. (c) Multivariate Logistic regression

Multiple logistic regression

In regression, we predict one value from another variable.

Let's take an example of how does meteorological department tell us the probability of rainfall or the temperature of our city after two months or even after three months. The point is, they are able to tell about the weather based on data from previous many years regarding cloud type, wind velocity, air temperature, humidity and probability of rainfall.

Now if we wish to take the same example for predicting the probability of cancer depending on stress levels, obesity levels, family history and other factors, we may use some mathematical models to predict the dependent values.

These models are known as regression equations.

Regression is of two types:

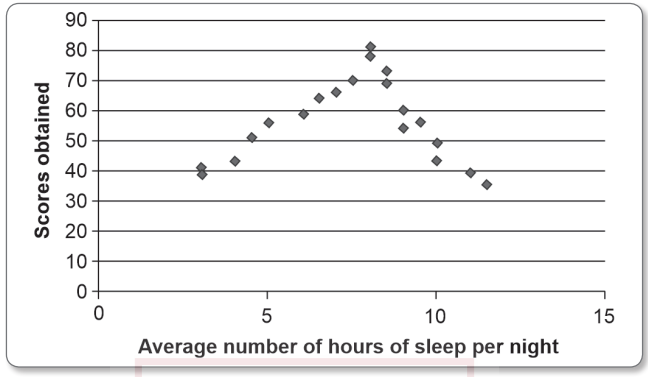
1. Linear regression—for parametric data (quantitative data)
2. Logistic Regression—for nonparametric data (qualitative data)

Since all the parameters given in the MCQ are dichotomous data (binomial or two groups) data, the regression we may use is multivariate (-multiple variables) logistic (-nonparametric data) regression.

Pearson correlation: For linear correlation, ungrouped data, quantitative data.

Dixon's Q test: For identification of outliers.

Curvilinear correlation: It is a type of correlation, where one variable increases along with another variable but later may show a decrease or no change or vice versa.



3. Pearson correlation is denoted by:

a. r b. ρ c. σ d. χ

Ans. (a) r

Remember:

The **Pearson correlation** is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.

The **Point-biserial correlation** is conducted with the Pearson correlation formula except that one of the variables is dichotomous.

Kendall rank correlation is a **nonparametric test** that measures the strength of dependence between two variables.

Spearman rank correlation: Spearman rank correlation is a **non-parametric test** that is used to measure the degree of association between two variables.

4. If the value of correlation coefficient (r) = 0, then the correlation line is:

a. Ninety degree
b. Zero degree
c. Thirty degree
d. Forty-five degree

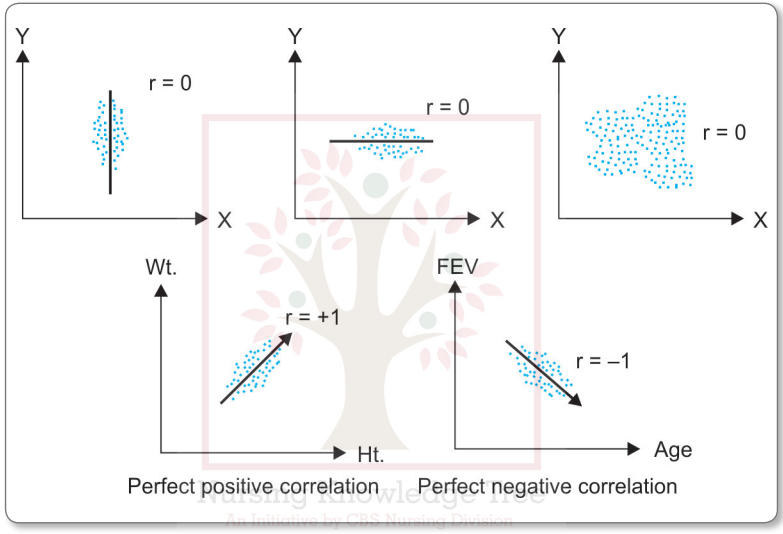
Ans. (a) Ninety degree

If the correlation coefficient (r) = 0, then the correlation line is at 90°.

5. If X and Y are two standard normal parametric variables with correlation coefficient of 'r', then the regression of 'Y' on 'X' is given by:

a. Linear regression
b. Logistic regression
c. Rank correlation
d. Normal correlation

Ans. (a) Linear regression



Linear regression: For parametric or quantitative data.

Logistic regression: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

6. For the regression equation, with 'Y' regressing on variable 'X'. the variable 'X' is:

a. Dependent variable
b. Independent variable
c. Regresses
d. Correlation variable

Ans. (b) Independent variable

For the regression line Y on X, the value of 'Y' will depend on the value of the 'X'. Hence,

X → Y

Y value depends on X.

Y is dependent variable, output variable,

X is the independent variable, input variable

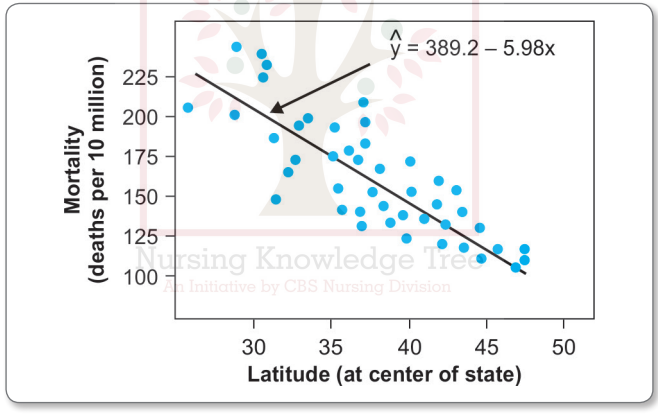
7. Regression equation is also known as:

a. Line of average relationship
b. Prediction equation
c. Estimating equation
d. All of the options

Ans. (d) All of the options

Example: It was seen in a particular state, the burden of skin cancer was increasing, hence, the data was obtained for the pattern of the cases of skin cancer with the latitude of the region.

The data obtained was plotted on a scatter plot as follows:



It is seen that as the latitude decreases (as we move further to north) the skin cancer cases increase in as shown in the above figure. But as we see in the plot, it does increase but is not perfect linear relationship, in the sense that there have been few cases from higher latitudes and higher number of skin cancer (which is contrary to our observation of lower the latitude, higher will be the number for cases of skin cancer from the defined region). Hence, to assess this sort of relationship and 'predict' the cases, burden at a particular independent variable point, is known as regression.

The linear regression equation is:

$$Y = a + bx + e$$

Where, Y is the dependent variable or response value or outcome value.

a = Regression constant

b = Independent variable coefficient, or slope of regression line

x = Independent variable value, predictor value

e = Error estimate

8. If the correlation coefficients of two correlation lines is ±1, then the two lines:

a. Perpendicular to each other
b. Intersect each other
c. There is no correlation
d. Coincide on each other

Ans. (d) Coincide on each other

Correlation:

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between −1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of −1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables.

Hence, in the MCQ, if the correlation coefficient is ±1, the value is either perfect positive or perfect negative and the lines shall coincide.

9. In the regression line, y = a + bx, 'b' represents:

a. Intercept of the line
b. Slope of the line
c. Direction of correlation
d. None of the options

Ans. (b) Slope of the line

The Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

It is given by the formula:

$$Y = a + bX + e$$

- Y is the value of the dependent variable (Y), what is being predicted or explained.

- a or Alpha, a constant; equals the value of Y when the value of X = 0.
- b or Beta, the coefficient of X; the slope of the regression line; how much Y changes for each one-unit change in X. The bigger the 'b', the steeper is the slope.
- X is the value of the independent variable (X), what is predicting or explaining the value of Y.
- e is the error term; the error in predicting the value of Y, given the value of X (it is not displayed in most regression equations).

10. If ranks or ordinal data is used, which of the following correlation maybe used:

a. Line diagram
b. Karl Pearson correlation
c. Spearman's correlation
d. Point correlation

Ans. (c) Spearman's correlation

Refer to Explanation 3

11. A significant positive correlation has been observed between alcohol consumption and the level of systolic blood pressure in men. From this correlation, we may conclude that:

a. There is no association between alcohol consumption and systolic pressure.
b. Men who consume less alcohol are at lower risk for increased systolic pressure.
c. Men who consume less alcohol are at higher risk for increased systolic pressure.
d. The risk of raised systolic pressure would be >2 times with one unit of rise in alcohol intake.

Ans. (b) Men who consume less alcohol are at lower risk for increased systolic pressure.

Positive correlation:

Means that the two variables under consideration show a positive directional association, i.e. if one increases the other increases and vice versa.

Negative correlation:

Means that the two variables under consideration show a negative directional association, i.e. if one increases the other decreases and vice versa.

Association:

Means presence of two variables at a single point of time. The variables, however, may or may not be correlated.

12. A researcher finds that correlation coefficient of cerebral blood flow and severity of psychotic symptoms in schizophrenia is —2.6:

a. The cerebral blood flow accounts for 2.6% of variance in psychotic symptoms.
b. The cerebral blood flow is 2.6 times predictor for severity of psychotic symptoms.
c. The psychotic symptoms are 2.6 times risk in people with altered cerebral blood flow.
d. The correlation coefficient is wrongly reported.

Ans. (d) The correlation coefficient is wrongly reported.

Correlation coefficient (r) ranges from –1 to + 1

1. Correlation coefficient = (–) 1: means negative correlation
2. Correlation coefficient = (+) 1: means positive correlation

Hence, correlation coefficient of 2.6 is wrongly reported.

13. Which of the following is not true about 'correlation'?

a. It indicates degree of association between two characteristics.
b. Correlation coefficient of 1 means that the two variables exhibit linear relationship.
c. Correlation can measure risk.
d. Causation implies correlation.

Ans. (d) Causation implies correlation.

Causation is not given by correlation.

14. What can be true regarding the coefficient of correlation between IMR and economic status?

a. r = +1 b. r = 1 c. r = 0.22 d. r = –0.8

Ans. (d) r = –0.8

High IMR is logically related to low socio-economic status. Hence, will have negative correlation.

15. Test of degree and direction of association between two variables is done by:

a. χ^2 b. Correlation c. Regression d. Z-test

Ans. (b) Correlation

Correlation: Often we wish to know whether there is linear relation between two variables, e.g., height and weight, temperature and pulse, age and vital capacity, etc.

In correlation,
- Plot scatter graphs
- Check for vector of scatter
- The vector of scatter would tell the direction and strength of association

Correlation could be:
- Pearson correlation (r)
- Spearman correlation (ρ)

Correlation provides:
- Direction of association/correlation
- Degree of association/correlation

Regression: If we wish to know in an individual case the value of one variable, knowing the value of the other, we calculate what is known as the regression coefficient of one measurement to the other.

16. A coefficient of correlation value of "r = +0.8" indicates:

a. Strong direct relationship between two variables
b. Strong inverse relationship between two variables
c. Insignificant association between two variables
d. One variable is the cause of the other variable

Ans. (a) Strong direct relationship between two variables

17. Pearson or Spearman coefficient is used for evaluation of:

a. Differences in proportion b. Comparison of >2 means
c. Comparison of variance d. Correlation

Ans. (d) Correlation

18. If we know the value of one variable in an individual and wish to know the value of another variable, we calculate:

a. Coefficient of correlation
b. Coefficient of regression
c. SE of mean
d. Geometric mean

Ans. (b) Coefficient of regression

19. All are features of correlation coefficient except:

a. Risk associates can be predicted
b. Correlated risk to disease
c. Cause effect association cannot be shown
d. Indicates linear relationship.

Ans. (a) Risk associates can be predicted

20. The value of correlation coefficient lies between:

a. 0 – 1
b. –1 to + 1
c. 0 – 100
d. –1 to 100

Ans. (b) –1 to + 1

CBS HANDBOOK ON BIOSTATISTICS FOR NURSES

Salient Features

- CBS Handbook on Biostatistics for Nurses is written in simple, clear and concise manner for conceptual understanding.
- It is based on syllabus prescribed by Indian Nursing Council for nurses.
- A new approach has been adopted to present the concept for better memory retention.
- Chapters are discussed in *quick review format* to be precise but systematic in approach.
- Chapters are followed by *solved MCQs, and numericals* to be solved by the readers for better and complete understanding.
- Concepts are discussed with *analytical and research-based approach* in the field of nursing.
- Useful for most of the nursing examinations.

About the Authors

Mukhmohit Singh, *MBBS, MD (Community Medicine) PGDHS* has been working as a faculty for training students for PSM subject for various entrance exams across India and outside India, with a very high success rate. He has clinical-cum-research experience of working as Medical Officer in corporate hospitals as well as research fellow for projects in PGIMER, Chandigarh. He is associated with Smt. Vidyavati Trust in providing care and innovations in health care. He also has vast array of field and community experience while working as Epidemiologist for UT Administration, Chandigarh and as Assistant Professor (Epidemiology) in the Department of Community Medicine for teaching and training of the undergraduate MBBS students.

The author has been awarded with Gold Sponsorship Award for work on COPD at the prestigious European Respiratory Society. He has also been awarded with the certificates for successful completion of Diploma course for Advancements and Management Plans for Diabetes in India, for Thyroid Disorders and Advance Epidemiology programs. The author has contributed to research in fields of pulmonary medicine, and has various papers published in national and international journals. He is also an active member of various organizations such as ERS, RSSDI, IAPSM, IPHA. He is also an international faculty for Epidemiology and Public Health for teaching medical doctors in USA and Canada for Licensing Exams.

Shveta Saini, *MBBS, MD (Community Medicine)* has been working as Assistant Professor, Community Medicine and Public Health, Preventive Medicine and Public Health Specialist. She is very dynamic and enthusiastic in her field and has been recognized for her active contributions to the field of community medicine. She is a member of several institutions, like—Indian Public Health Association and Indian Association of Preventive and Social Medicine. Her flawless delivery of lecture and clarity in the concepts make her very popular amongst the students. She has also been an author of books like—Review of PGI Chandigarh (PGMEE) 2015-2016 and 2016-2017, Biostatistics Review for PGMEE and PSM Trends for PGMEE, along with Dr Mukhmohit Singh.

CBS Publishers & Distributors Pvt. Ltd.
4819/XI, Prahlad Street, 24 Ansari Road, Daryaganj, New Delhi 110 002, India
E-mail: feedback@cbspd.com **Website:** www.cbspd.com
New Delhi | Bengaluru | Chennai | Kochi | Kolkata | Lucknow | Mumbai | Pune
Hyderabad | Nagpur | Patna | Vijaywada

ISBN: 978-93-90619-10-8