



HANDBOOK ON BIostatISTICS FOR HEALTH PROFESSIONALS

**DESCRIPTIONS
MANUAL CALCULATIONS
R (EZR) SOFTWARE BASED CALCULATIONS**

First edition

**Kalesh M Karun
Amitha Puranik**

Published by BCC group

Handbook on Biostatistics for Health Professionals

- Descriptions
- Manual calculations
- R (EZR) software based calculations

(FIRST EDITION)

Dr Kalesh M Karun, PhD

Assistant Professor and in-charge

Division of Bio-statistics, MOSC Medical College

(Recognized research centre by Kerala University of Health Sciences)

Kolenchery, Ernakulam, Kerala, India

Ms Amitha Puranik

Assistant Professor

Department of Data Science

Prasanna School of Public Health

Manipal Academy of Higher Education, Manipal, Karnataka, India

(An Institute of Eminence (Status Accorded by MHRD))

Preface

1. Introduction to Biostatistics	1
1.1 Statistics	1
1.2 Branches of Statistics	1
1.3 Biostatistics	3
1.4 Indian statisticians	4
1.5 Applications of Biostatistics	4
2. Types of variables and scales of measurements	6
2.1 Introduction	6
2.2 Variables and its classifications	7
2.2.1 Qualitative (or categorical) variable	7
2.2.2 Quantitative variable	7
2.3 Scales of measurements	8
2.3.1 Properties of measurement scales	8
2.3.2 Types of scales of measurements	9
3. Introduction to R (EZR) software	10
3.1 Introduction	10
3.2 EZR (Menu driven R software)	11
3.3 Data management in EZR	12
3.3.1 Creating and importing data	12
3.3.2 Computing new variable based on existing variables	13
3.3.3 Renaming variable	14
3.3.4 Categorizing continuous variable	15
3.3.5 Converting numeric variable to factor	16
4. Tabular and graphical presentations of data	18
4.1 Introduction	18

4.2 Tabular presentation of qualitative data	18
4.2.1 Frequency distribution	18
4.2.2 Relative frequency	19
4.3 Tabular presentation of quantitative data	19
4.3.1 Frequency distribution	20
4.4 Graphical presentation of data	21
4.4.1 Graphical presentation of qualitative data	22
4.4.2 Graphical presentation of quantitative data	24
5. Measures of central tendency	31
5.1 Introduction	31
5.2 Arithmetic mean	31
5.3 Median	32
5.4 Mode	33
5.5 Geometric mean	34
5.6 Harmonic mean	34
6. Partition values	38
6.1 Introduction	38
6.2 Median	38
6.3 Quartiles	39
6.4 Deciles	40
6.5 Percentiles	41
7. Measures of dispersion/variability	42
7.1 Introduction	42
7.2 Measures of dispersion	42
7.2.1 Range	42
7.2.2 Inter quartile range (IQR)	43
7.2.3 Mean deviation (MD)	44
7.2.4 Standard deviation (SD)	45

8. Probability distributions and statistical inference	51
8.1 Probability	51
8.2 Probability distribution	52
8.2.1 Continuous probability distribution	52
8.2.2 Discrete probability distribution	53
8.3 Statistical inference	53
8.3.1 Point estimation	54
8.3.2 Interval estimation	54
8.3.3 Hypothesis testing	55
9. Normal distribution	57
9.1 Introduction	57
9.2 Properties of a normal distribution/normal curve	57
9.3 Skewness and Kurtosis	58
9.3.1 Skewness	58
9.3.2 Kurtosis	59
9.4 Normality checking	59
9.5 Checking the normality using R (EZR) software	60
10. Introduction to testing of hypothesis	62
10.1 Introduction	62
10.2 Null and alternative hypothesis	63
10.3 Two-tailed and one-tailed tests	63
10.4 Types of errors in hypothesis testing	64
10.5 Test statistic, power of the test and p value	65
10.6 Procedure of hypothesis testing and decision making	66
10.7 Limitations of hypotheses testing	67
11. Parametric testing of hypothesis	68
11.1 Introduction	68
11.2 Parametric tests	68

11.2.1 Independent samples t test	68
11.2.3 Paired sample t test	78
11.2.5 Analysis of covariance (ANCOVA)	85
12. Non parametric testing of hypothesis	89
12.1 Introduction	89
12.2 Non parametric tests	89
12.2.1 Wilcoxon Mann Whitney U test	89
12.2.2 Kruskal–Wallis test	94
12.2.3 Wilcoxon matched-pairs test (or Signed rank test)	99
12.2.4 Friedman ANOVA	102
13. Semi parametric test: Chi-square test	106
13.1 Introduction	106
13.2 Chi-square test for association	107
13.3 Fisher’s exact test	110
13.4 Yates’ correction	112
13.5 Chi-square test for goodness of fit	114
13.6 Binary logistic regression analysis	116
14. Correlation and regression analysis	119
14.1 Introduction	119
14.2 Methods of correlation analysis	119
14.2.1 Scatter diagram method	120
14.2.2 Karl Pearson’s (product moment) coefficient of correlation	121
14.2.3 Spearman’s coefficient of rank correlation	125
14.3 Linear regression analysis	128
15. Validity of diagnostic tests and ROC analysis	134
15.1 Introduction	134
15.2 Validity of the diagnostic test	135
15.3 Likelihood ratios in diagnostic test	137

15.4 ROC analysis	138
16. Measures of agreement	142
16.1 Introduction	142
16.2 Various measures of agreements	142
16.2.1 Kappa statistic	143
16.2.2 Bland-Altman analysis	145
17. Vital and health statistics	149
17.1 Introduction	149
17.2 Sources of vital statistics in India	149
17.2.1 Population Census	149
17.2.2 Civil registration system (CRS)	150
17.2.3 Sample registration system (SRS)	150
17.2.4 Health surveys	150
17.3 Basic concepts and terminologies of vital statistics	151
17.3.1 Ratio	151
17.3.2 Rate	151
17.3.3 Trend	152
17.4 Mortality rates	152
17.4.1 Measures of mortality	152
17.5 Birth rate and fertility rate	155
17.5.1 Measures of birth & fertility	155
17.6 Morbidity rates	156
17.6.1 Measures of morbidity	156
18. Introduction to systematic reviews and meta-analysis	159
18.1 Introduction to systematic reviews	159
18.2 Features of systematic review	159
18.3 Stages of systematic review	160
18.4 Introduction to meta-analysis	163
18.5 Significance of meta-analysis	163
18.6 Statistical methods in meta-analysis	164

18.6.1 Heterogeneity assessment	164
18.6.2 Fixed effect model and random effects model	165
18.6.3 Forest plot	166
18.6.4 Sensitivity analysis	167
18.6.5 Publication bias and funnel plot	168
18.7 Review manager (Revman) software	169
19. Methods of data collection and sampling methods	170
19.1 Introduction to data collection	170
19.2 Methods of collection of primary data	170
19.2.1 Direct personal investigation	170
19.2.2 Indirect personal investigation	171
19.2.3 Investigation through questionnaire	171
19.2.4 Investigation through local correspondents	171
19.2.5 Focus group discussion	171
19.3 Secondary data collection methods	172
19.4 Introduction to sampling methods	172
19.4.1 Steps in sampling process	174
19.4.2 Errors in sampling	174
19.5 Sampling methods	174
19.5.1 Probability (random) sampling methods	175
19.5.2 Non-probability sampling methods	177
20. Introduction to sample size estimation	180
20.1 Introduction	180
20.2 Significance of the optimum sample size	180
20.3 Factors affecting sample size	181
20.4 Sample size estimation procedures	182
20.4.1 Sample size based on estimation of mean	182
20.4.2 Sample size based on estimation of proportion	184
20.4.3 Sample size based on comparison of two means	187
20.4.4 Sample size based on comparison of two proportions	190
21. Design of experiments	194

21.1 Introduction	194
21.2 Comparison in pairs	194
21.2.1 Repeated measures design	195
21.2.2 Cross over design	195
21.3 Completely randomized design (CRD)	197
21.4 Randomized block design (RBD)	197
21.5 Latin square design (LSD)	198
22. Introduction to validity and reliability of rating scales	200
22.1 Introduction	200
22.2 Validity	201
22.2.1 Content validity	201
22.2.2 Face validity	202
22.2.3 Criterion-related validity	202
22.2.4 Construct validity	202
22.3 Reliability	203
22.3.1 Test-retest reliability	203
22.3.2 Internal consistency reliability	203
22.3.3 Inter-rater reliability	205
Statistical Tables	206

Measures of central tendency

5.1 Introduction

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value. This value represents the middle or center of its distribution. There are several measures of central tendencies and choosing the best measure of central tendency depends on the type/nature of data.

The essential properties of a good measure of central tendency are,

- Should be clearly defined
- Should be based on all observations
- Should be amenable for further mathematical treatments
- Should not be affected by extreme values
- Should be easy to calculate and simple to follow

Measures of central tendency are,

1. Arithmetic mean
2. Median
3. Mode
4. Geometric mean
5. Harmonic mean

5.2 Arithmetic mean

Arithmetic mean is the most widely used simple measure of central tendency. Mean is calculated by adding all observations of a variable and dividing by the total number of observations. This is the best descriptive measure for data that are symmetrically (normally) distributed.

$$\bar{x} = \frac{\sum x_i}{n}$$

where,

- x_i is the i^{th} observation
 - n is the sample size
-

Merits:

- Based on all the observations
- Capable of further algebraic treatments
- Stable, doesn't differ much from sample to sample

Demerits:

- Affected by extreme values
- Cannot be calculated for qualitative data

Problem 5.1: Estimate the mean of the following values,

2, 6, 4, 10, 8, 12, 16, 14

Solution:

$$\begin{aligned} \text{mean, } \bar{x} &= \frac{\sum x_i}{n} \\ &= \frac{2+6+4+10+8+12+16+14}{8} = 9 \end{aligned}$$

5.3 Median

The median is the middle value of the distribution when the values are arranged in ascending or descending order. The median divides the distribution into two equal parts, 50% of observations are on either side of the median value. Median is used to summarize the skewed distributions.

Median = Size of $[(n+1)/2]^{\text{th}}$ item

where,

- n is the sample size
-

Merits:

- Not affected by extreme values
- Can be determined by graphical methods

Demerits:

- Not based on all the observations
- Not capable of further algebraic treatments

Problem 5.2: Estimate the median of the following values,

2, 6, 4, 10, 8, 12, 16, 14

Solution:

Values in ascending order: 2, 4, 6, 8, 10, 12, 14, 16

Median = Size of $[(n+1)/2]^{\text{th}}$ item = Size of 4.5th item

$$= 4^{\text{th}} \text{ item} + 0.5(5^{\text{th}} - 4^{\text{th}} \text{ item})$$

$$= 8 + 0.5 \times 2 = 9$$

5.4 Mode

Mode is the most frequently occurring value of the dataset. It is the preferred measure of central location for addressing the value which is the most popular or most common.

Merits:

- Not affected by extreme values
- Can be determined by graphical methods

Demerits:

- Uncertain and vague measure of central tendency
- Not based on all the observations
- Not capable of further algebraic treatments

Problem 5.3: Estimate the mode of the following values,

2, 6, 4, 2, 8, 12, 16, 12, 2

Solution:

Mode = most frequently occurring item
= 2

5.5 Geometric mean

Geometric Mean (GM) is one of the measures of central value which is most common in business and finance. Mainly used when dealing with percentages to calculate growth rates and returns on portfolio of securities.

$$GM = \text{Antilog} \left[\frac{\sum \log x_i}{n} \right]$$

where,

- n is the sample size
- x_i is the i^{th} observation

Problem 5.4: Estimate the GM of the following values,

2, 6, 4, 10, 8, 12, 16, 14



Normal distribution

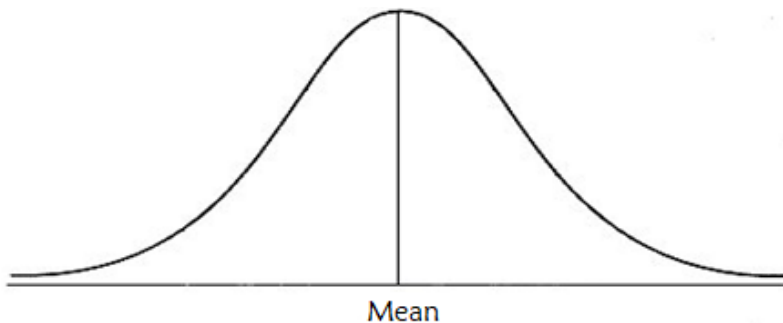
9.1 Introduction

Normal (Gaussian or Gauss or Laplace–Gauss) distribution is a continuous probability distribution in which the distribution is a symmetrical bell-shaped curve. Manufacturing processes and biological/natural occurrences frequently create this type of distribution. Normal curve is characterized by its mean, μ and standard deviation, σ [Fig 9.1].

The probability density function of normal distribution is given as,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ where, } -\infty < x < \infty$$

Figure 9.1: Normal curve

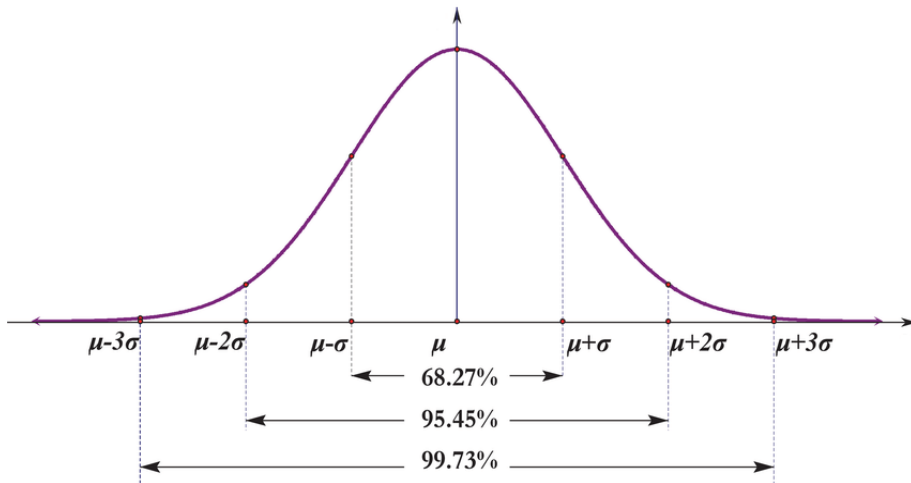


9.2 Properties of a normal distribution/normal curve

- Normal curve is bell-shaped
- Normal curve is symmetric about the mean
- The mean is at the middle and divides the area into halves
- The mean, median, and mode are equal
- The total area under the curve is equal to one
- The normal curve approaches, but never touches, the x-axis and extends up to positive and negative infinity

- Unimodal in nature (only one mode)
- Skewness=0
- Kurtosis=3 (mesokurtic)
- For a normal curve [Fig 9.2],
 - 68.27% of the observations lie between mean \pm 1SD
 - 95.45% of the observations lie between mean \pm 2SD
 - 99.73% of the observations lie between mean \pm 3SD

Figure 9.2: Distribution of observations in a normal curve



9.3 Skewness and Kurtosis

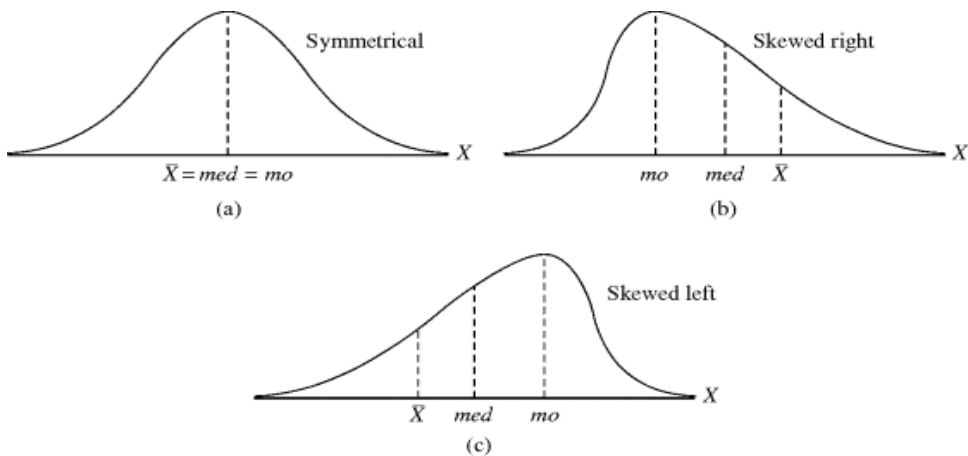
9.3.1 Skewness: Skewness is described as the asymmetry or lack of symmetry of the dataset [Fig 9.3]. A perfectly symmetrical data set will have a skewness of zero. The normal distribution has a skewness of zero.

$$\text{Skewness} = (\text{Mean}-\text{Mode})/\text{SD} \text{ or } \text{Skewness} = 3(\text{Mean}-\text{Median})/\text{SD}$$

There are two types of skewness

- *Positive skewness:* where the distribution is not symmetrical to mean, the curve is shifted more towards the right side. Here, mean > median > mode.
- *Negative skewness:* where the distribution is not symmetrical to mean, the curve is shifted more towards left side. Here, mean < median < mode.

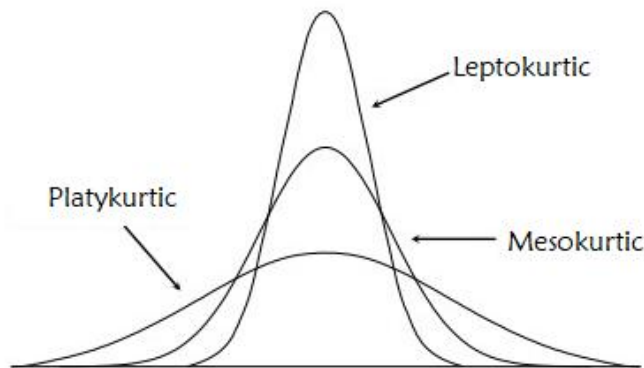
Figure 9.3: Diagrammatic presentation of skewness



9.3.2 Kurtosis: Kurtosis is the degree of peakedness (tallness) of a curve/data distribution (Fig 9.4). There are three types of kurtosis, they are

- Leptokurtic: Sharply peaked (Kurtosis > 3) with fat tails and less dispersed.
- Mesokurtic: Medium peaked (Kurtosis $= 3$), normal curve is mesokurtic.
- Platykurtic: Flattest peak (Kurtosis < 3) and highly dispersed.

Figure 9.4: Diagrammatic presentation of kurtosis



9.4 Normality checking (How to check the normality?)

The three important methods of checking the normality of data are,

Histogram method: Plot a histogram for the data and check whether the curve is symmetrical. If the curve is symmetrical, data/variable follows normality.

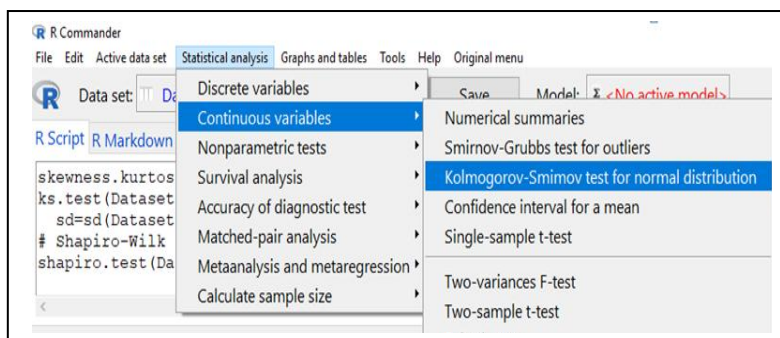
Shapiro-Wilk test: The null-hypothesis of this test is that the population is normally distributed. So if p value is greater than 0.05, data follows normality. Usually applied when sample size is small ($n < 50$).

Kolmogorov Smirnov test (K S test): The null-hypothesis of K S test is that the population is normally distributed. So if p value is greater than 0.05, data follows normality. Usually applied when the sample size is large ($n > 50$).

Thumb rule: If the mean of data set is greater than two times the SD, usually (not always) data follows normality.

9.5 Checking the normality using R (EZR) software

Step 1: Open the dataset in EZR and go to Statistical analysis > Continuous variables > K S test for normal distribution.



Parametric testing of hypothesis

11.1 Introduction

A parametric statistical test makes an assumption about the population parameters i.e. the variable of interest should follow normal distribution. Parametric functions were mentioned by R.A Fisher which created the foundation for modern statistics. Parametric tests are based on a set of assumptions such as,

- Independence: Observations should be independent of each other.
- Normality: Data should be normally distributed (symmetrical).
- Homogeneity of variances: Data from multiple groups should have the same variance.

When these assumptions are not violated, parametric methods will produce more accurate and precise estimates than non-parametric methods, i.e. parametric tests have greater statistical power than non-parametric methods (more information on non-parametric methods is given in chapter 12). Most well-known elementary statistical methods are parametric in nature.

11.2 Parametric tests

The present chapter will focus on most commonly used parametric tests such as,

- Independent samples t test
- Analysis of variance (ANOVA)
- Paired sample t test
- Repeated measures ANOVA (RANOVA)
- Analysis of covariance (ANCOVA)

11.2.1 Independent samples t test: It is a parametric test to compare the average value between two independent groups. This test is also known as an unpaired samples t test.

The test Statistic,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,

- \bar{x}_1, \bar{x}_2 are the mean of groups 1 and 2, respectively.
- σ_1, σ_2 are the SD of groups 1 and 2, respectively.
- n_1 and n_2 are the sample size of groups 1 and 2, respectively.

[The degrees of freedom for this test is given by $n_1 + n_2 - 2$].

Assumptions of independent samples t test

- Samples are randomly selected from normally distributed populations.
- Population variances are equal.

Note: Normality can be checked by means of Histogram, Shapiro-Wilk test or Kolmogorov Smirnov test (K S test) [more details in Chapter 9] and the equality of variance can be checked using Levene's test. The null-hypothesis of Levene's test is that the population variances of various groups are equal. So if p value is greater than 0.05, data satisfies the assumption of homogeneity of variances.

Problem 11.1:

Check whether there is significant difference in the average SBP between males and females based on the data given below (t table = 2.145, for d.f=14)

Males: 120, 122, 124, 132, 136, 138, 130, 122

Females: 138, 144, 147, 160, 148, 149, 150, 154

Solution [Manual calculation]:

State the hypothesis:

H₀: There is no significant difference in the average SBP between males and females.

H₁: There is significant difference in the average SBP between males and females.

Calculation of statistic t:

$$\text{Mean SBP of males } (\bar{x}_1) = \frac{120+122+124+132+136+138+130+122}{8} = 128.00$$

$$\text{Mean SBP of females } (\bar{x}_2) = \frac{138+144+147+160+148+149+150+154}{8} = 148.75$$

$$\text{SD of SBP of males } (\sigma_1) = \sqrt{\sum \left[\frac{(120-128)^2 + (122-128)^2 + \dots + (122-128)^2}{8-1} \right]} = 6.93$$

$$\text{SD of SBP of females } (\sigma_2) = \sqrt{\sum \left[\frac{(138-148.75)^2 + (144-148.75)^2 + \dots + (154-148.75)^2}{8-1} \right]} = 6.52$$

$$n_1=8; \quad n_2=8$$

The test Statistic,

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{128 - 148.75}{\sqrt{\frac{(8-1)6.93^2 + (8-1)6.52^2}{8+8-2}} \sqrt{\frac{1}{8} + \frac{1}{8}}} \\ &= -6.169 \end{aligned}$$

Decision rule:

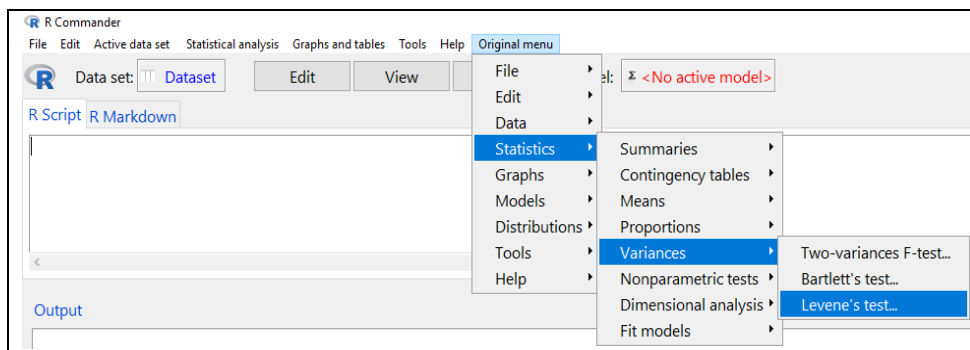
If calculated t value ($|t|$) is greater than table t-value, reject the H₀. Here, $|t| = 6.169$ and t table (for d.f (14)) = 2.146, hence we reject the null hypothesis.

Conclusion: There is a significant difference in the average SBP between males and females.

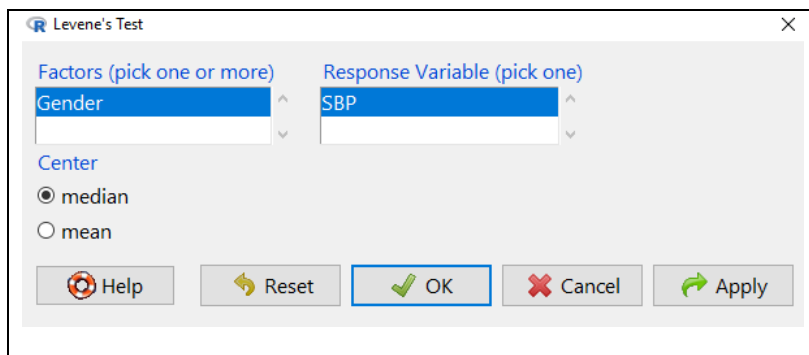
Solution [Using R (EZR) software]:

Step 1: Open the dataset in EZR and go to Statistical analysis > continuous variables > K S test for normal distributions [Note: The tests for checking normality are discussed in chapter 9].

Step 2: Go to Original menu > Statistics > Variances > Levene's test. [Note: This step is to check the assumption of homogeneity of variance].



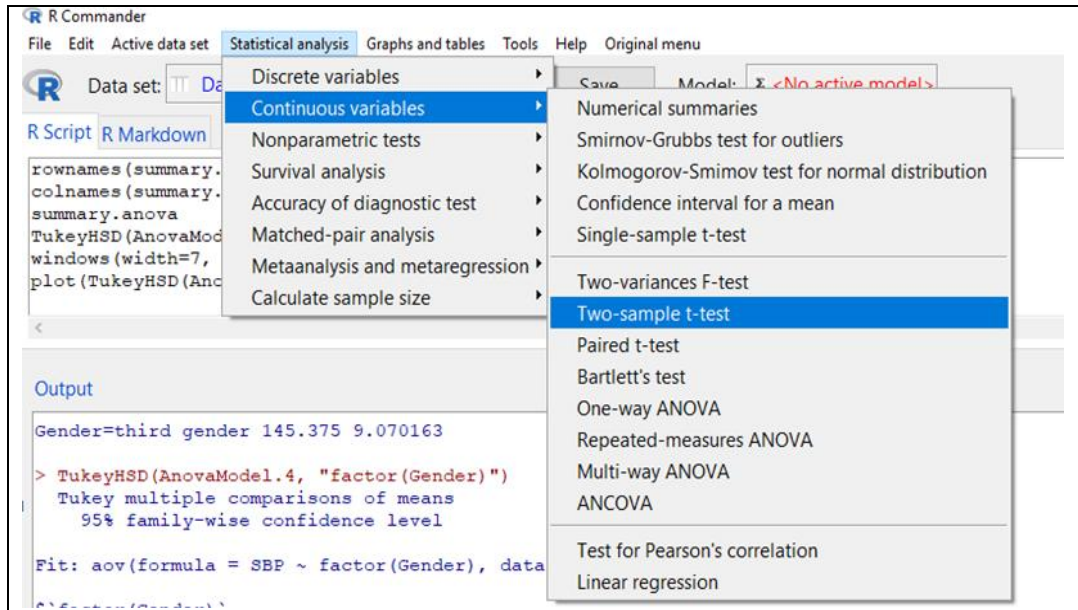
Step 3: Select *Gender* in option 'Factors' and *SBP* in 'Response Variable' and click 'OK'.



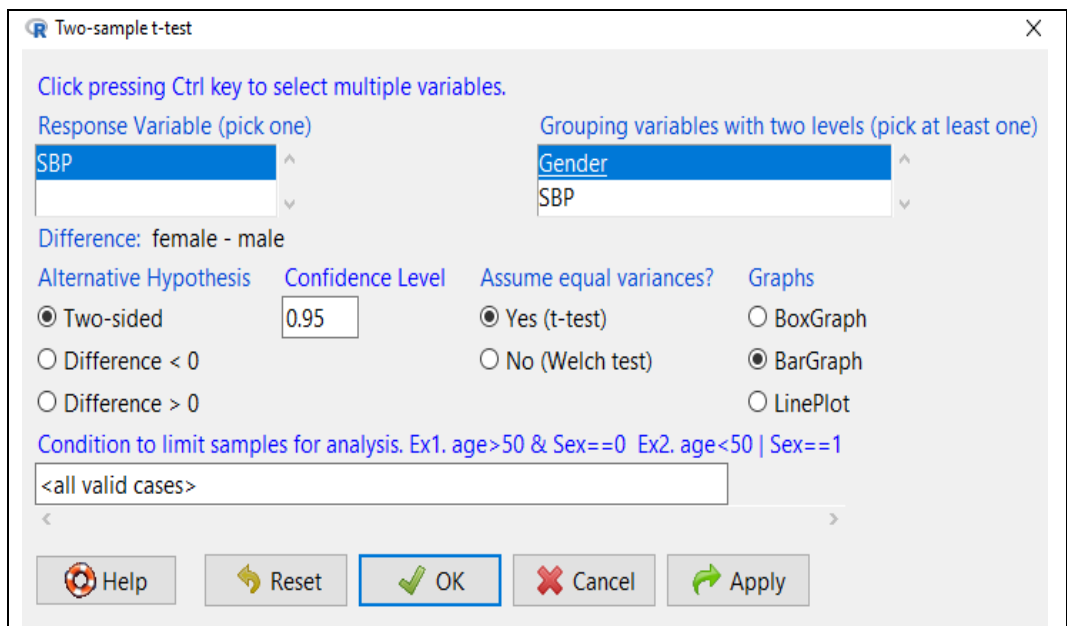
Step 4: The output window displays the result of Levene's test. Here the p value is 0.4314, which implies that the assumption of homogeneity of variance is not violated.

```
Levene's Test for Homogeneity of Variance (center = "median")
  Df F value Pr(>F)
group 1  0.6562 0.4314
  14
```

Step 5: Go to Statistical analysis > Continuous variables > Two-sample t-test.



Step 6: Select *SBP* in the 'Response Variable' and *Gender* in the 'Grouping variable', and click 'OK' [since the assumption of homogeneity of variances is not violated, check 'Yes (t-test)' option under 'Assume equal variances?'].



Step 7: Output window displays independent samples t test results as well as descriptive statistics such as mean and SD of each group [*Note: plot is optional*].

```
Two Sample t-test

data: SBP by factor(Gender)
t = 6.1693, df = 14, p-value = 2.437e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.53621 27.96379
sample estimates:
mean in group female    mean in group male
           148.75                128.00

           mean      sd  p.value
Gender=female 148.75 6.519202 2.44e-05
Gender=male   128.00 6.928203
```

Step 8: Report the results: The mean SBP of males found to be 128 (SD=6.92) and females found to be 148.75 (SD=6.52). It is observed that there is a significant difference in the average SBP between males and females ($p < 0.001$).

Note: If the p value given by the software is very small or if software displays only three digits after the decimal point (i.e. 0.000), then report the p value as $p < 0.001$.

11.2.2 Analysis of Variance (ANOVA): ANOVA is the parametric test to compare the average value between more than two independent groups, that is to test whether the mean of the outcome variable in the different groups is same or not. This test is an extension of the independent samples t test.

