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)*)

Contents

CHAPTER 5

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:

$$\text{mean, } \bar{x} = \sum x_i / n$$

$$= \frac{2+6+4+10+8+12+16+14}{8} = 9$$

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

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

$=8+0.5\times2 \ =9$

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

CHAPTER 9

Normal distribution

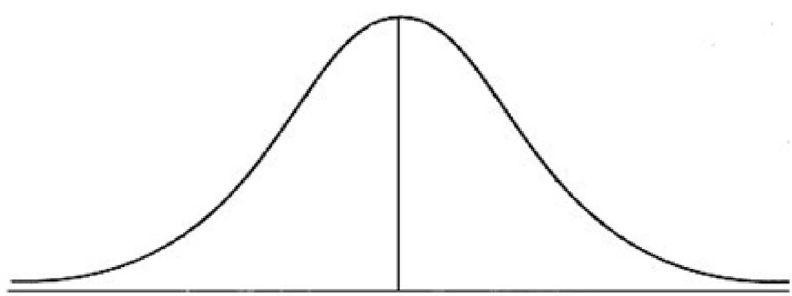
9.1 Introduction

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,

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

Figure 9.1: Normal curve

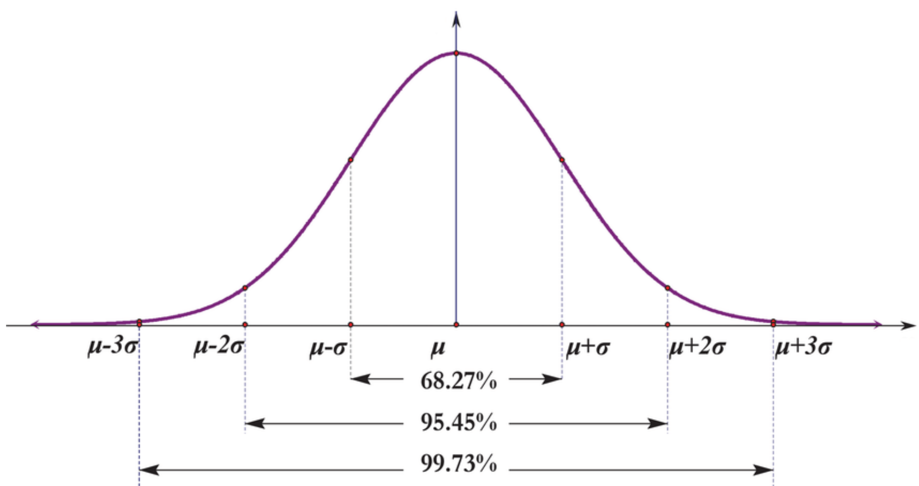


9.2 Properties of a normal distribution/normal curve

- 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

- Unimodel 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 ± 1SD
    - 95.45% of the observations lie between mean ± 2SD
    - 99.73% of the observations lie between mean ± 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.

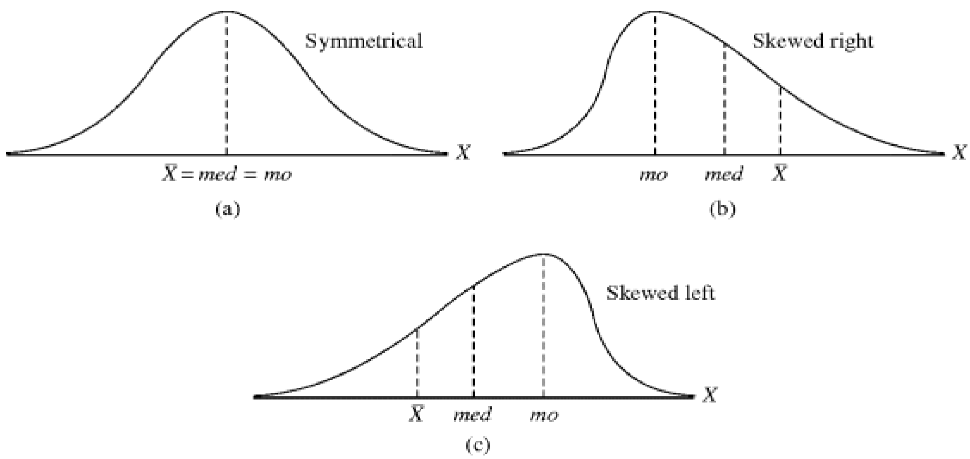
Skewness= (Mean-Mode)/SD or Skewness=3(Mean-Median)/SD

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.

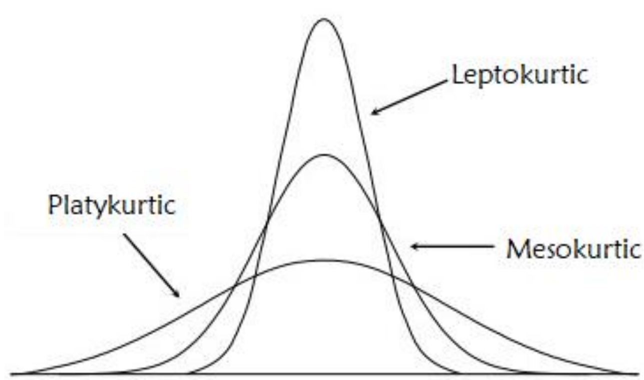
Figure 9.3: Diagrammatic presentation of skewness



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.

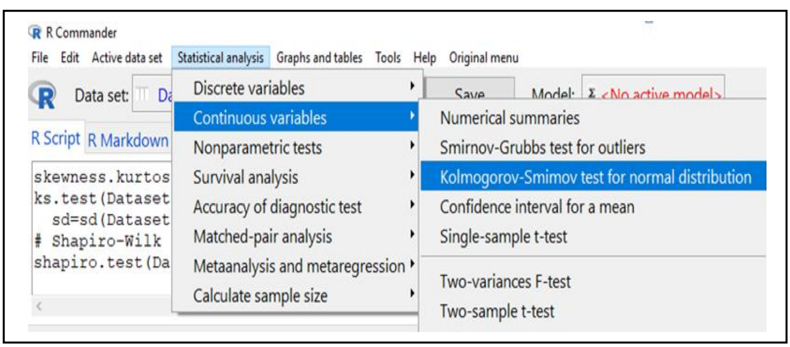
Shapiro-Wilk test: The null-hypothesis of this test is that the population is normally distributed. So if p value is greater than 0.05, data follows normality. Usually applied when sample size is small (n<50).

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.



CHAPTER 11

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 lest. 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:

H_0: There is no significant difference in the average SBP between males and females.

H_1: There is significant difference in the average SBP between males and females.

Calculation of statistic t:

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

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

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

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

$$= 6.52$$

$n_1=8$; $n_2=8$

The test Statistic,

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

$$= \frac{128-148.75}{\sqrt{\frac{(8-1)6.93^2+(8-1)6.52^2}{8+8-2}}\sqrt{\frac{1}{8}+\frac{1}{8}}}$$

$$= -6.169$$

Decision rule:

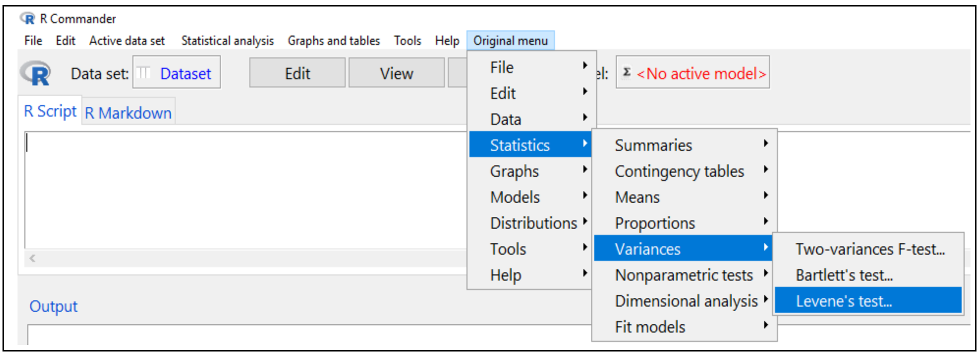
If calculated t value ($|t|$) is greater than table t-value, reject the H_0. Here, $|t|=6.169$ and t table (for d.f (14)) $=2.146$, hence we reject the null hypothesis.

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

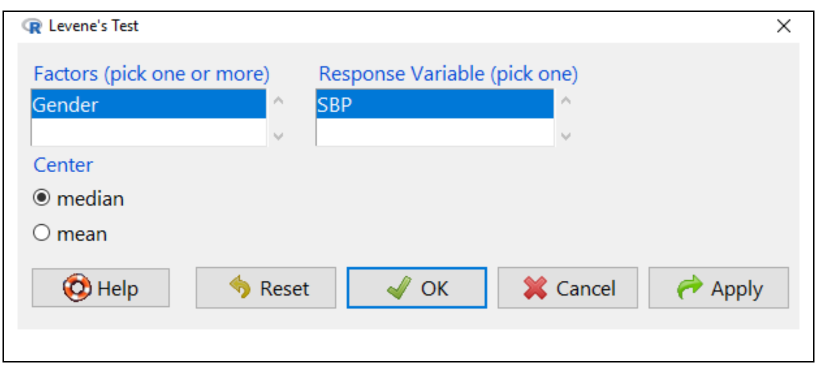
Solution [Using R (EZR) software]:

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*].



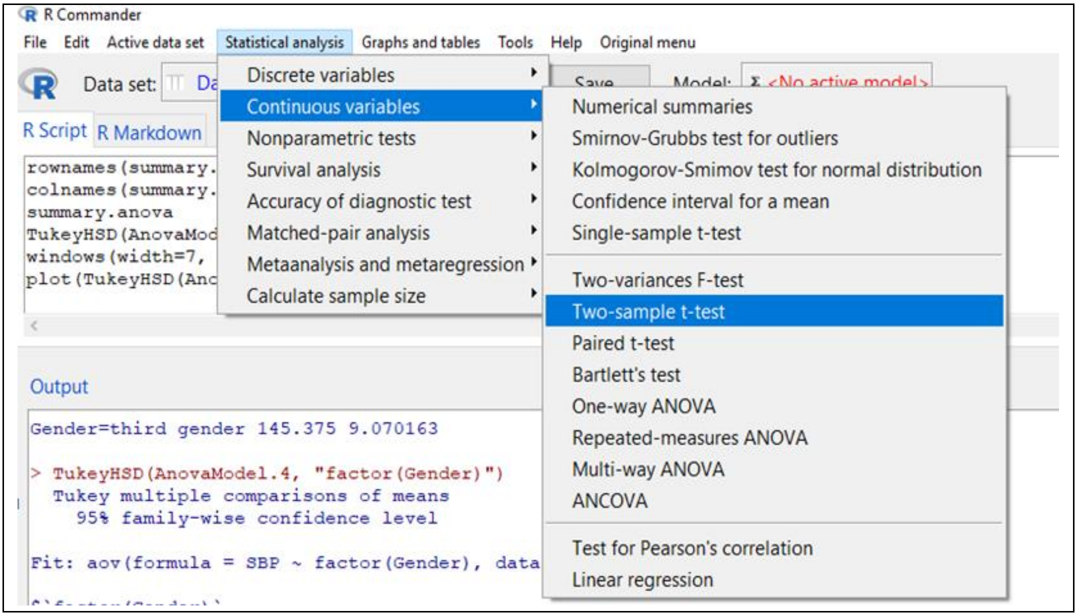
Step 3: Select *Gender* in option 'Factors' and *SBP* in 'Response Variable' and click 'OK'.



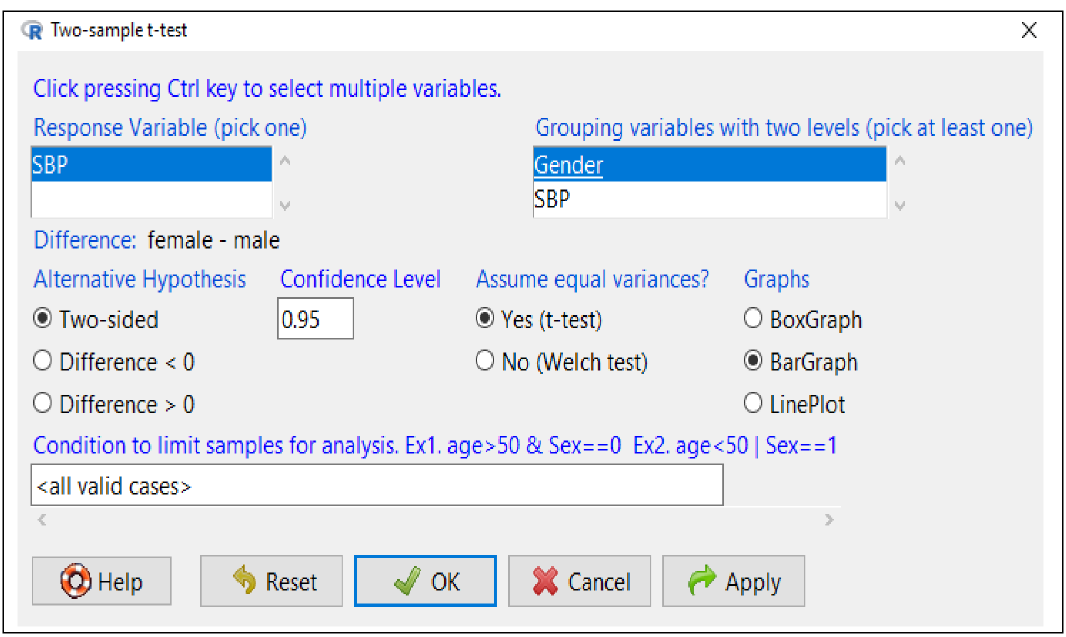
Step 4: The output window displays the result of Levene's test. Here the p value is 0.4314, which implies that the assumption of homogeneity of variance is not violated.

```
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.