CHAPTER-III

A BRIEF ABOUT EXITING UNIVARIATE STATISTICAL TOOLS IN BIOSTATISTICS
CHAPTER III

A BRIEF ABOUT EXISTING UNIVARIATE STATISTICAL TOOLS IN BIOSTATISTICS

3.1. BASIC CONCEPTS

Error: Error is the collective noun for any departure of the result from the "true" value. Analytical errors can be:

1) Random or unpredictable deviations between replicates, quantified with the "standard deviation".

2) Systematic or predictable regular deviation from the "true" value, quantified as "mean difference" (i.e. the difference between the true value and the mean of replicate determinations).

3) Constant, unrelated to the concentration of the substance analyzed (the analyze).

4) Proportional, i.e. related to the concentration of the analyze.

Accuracy: The "trueness" or the closeness of the analytical result to the "true" value. It is constituted by a combination of random and systematic errors (precision and bias) and cannot be quantified directly. The test result may be a mean of several values. An accurate determination produces a "true" quantitative value, i.e. it is precise and free of bias.
Precision: The closeness with which results of replicate analyses of a sample agree. It is a measure of dispersion or scattering around the mean value and usually expressed in terms of standard deviation, standard error or a range (difference between the highest and the lowest result).

Bias: The consistent deviation of analytical results from the "true" value caused by systematic errors in a procedure. Bias is the opposite but most used measure for "trueness" which is the agreement of the mean of analytical results with the true value, i.e. excluding the contribution of randomness represented in precision. There are several components contributing to bias:

1. **Method bias:** The difference between the (mean) test result obtained from a number of laboratories using the same method and an accepted reference value. The method bias may depend on the analyze level.

2. **Laboratory bias:** The difference between the mean test result from a particular laboratory and the accepted reference value.

3. **Sample bias:** The difference between the mean of replicate test results of a sample and the ("true") value of the target population from which the sample was taken. In practice, for a laboratory this refers mainly to sample preparation, subsampling and weighing techniques. Whether a sample is representative for the population in the field is an extremely important aspect but usually falls outside the responsibility of the laboratory (in some cases laboratories have their own field sampling personnel).
The relationship between these concepts can be expressed in the following equation:

\[ x = \mu + \delta + e \]

where:
- \( x \) is the measured value
- \( \mu \) is the true value
- \( \delta \) is the systematic error bias
- \( e \) is the random error precision

1 - accuracy = 1

Mean:

The average of a set of \( n \) data \( x \),

\[ \bar{x} = \frac{\sum x_i}{n} \]

Standard Deviation:

This is the most commonly used measure of the spread or dispersion of data around the mean. The standard deviation is defined as the square root of the variance (\( V \)). The variance is defined as the sum of the squared deviations from the mean, divided by \( n - 1 \).

Operationally, there are several ways of calculation.

\[ S = \sqrt{\frac{\sum(x, - \bar{x})^2}{n - 1}} \]

or
\[ i = \sqrt{\frac{\sum x_i^2 (\sum x_i^2)}{n - 1}} \]

or

\[ i = \sqrt{\frac{\sum x_i^2 - nx^2}{n - 1}} \]

Relative Standard Deviation Coefficient of Variation:

Although the standard deviation of analytical data may not vary much over limited ranges of such data, it usually depends on the magnitude of such data. The larger the figures, the larger S. Therefore, for comparison of variations (e.g., precision). It is often more convenient to use the relative standard deviation (RSD) than the standard deviation itself. The RSD is expressed as a fraction, but more usually as a percentage and is then called coefficient of variation (CV) often, however, these terms are confused.

\[ \text{PSD} = \frac{S}{X} \times 100\% \]

and variance \( (V) = S^2 \)

3.2. UNIVARIATE STATISTICAL CONCEPTS AND TOOLS:

ARITHMETIC MEAN:

Arithmetic mean of a set of observations is their sum divided by the number of observations. e.g., the arithmetic mean \( \bar{x} \) of \( n \) observations \( x_1, x_2, \ldots, x_n \) is given by.
\[
\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

In the case of frequency distribution \( x_i / f_i, i=1,2,\ldots, n \), where \( f_i \) is the frequency of the variable \( x_i \)

\[
\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{N}
\]

\[
\frac{1}{N} \sum_{i=1}^{N} f_i \left( \frac{1}{N} \sum_{i=1}^{N} f_i = N \right)
\]

In case of grouped or continuous frequency distribution, \( x \) is taken as the mid-value of the corresponding class.

**Standard Deviation or Variance:**

Standard deviation, usually denoted by the greek letter small sigma (\( \sigma \)), is the positive square root of the arithmetic mean of the squares of the deviation of the given values from their arithmetic mean. For the frequency distribution \( x_i/f_i, i=1,2,\ldots,n \),

\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} f_i (x_i - \bar{x})^2}
\]

where \( \bar{x} \) is the arithmetic mean of the distribution and \( \sum f_i = N \).

The square of standard deviation is called the variance and is given by
\[ \sigma^2 = \frac{1}{N} \sum f_i (x_i - \bar{x})^2 \]

Root mean square deviation, denoted by 's' is given by

\[ S = \sqrt{\frac{1}{N} \sum f_i (x_i - A)^2} \]

where A is any arbitrary number.

\[ S^2 \] is called mean square deviation.

**Simple Correlation:**

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, i.e., if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if they constantly deviate in the opposite directions, i.e., if increase (or decrease) is one results in corresponding decrease (or increase) in the other, correlation is said to be diverse or negative correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other. A correlation is a single number that describes the degree of relationship between two variables.
Correlation \( r = \frac{N \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{N \Sigma x^2 - (\Sigma x)^2}} \frac{N \Sigma y^2 - (\Sigma y)^2}{\sqrt{N \Sigma y^2 - (\Sigma y)^2}} \)

r lies between -1 and +1.

If the correlation is negative, we have a negative relationship, if its' positive, the relationship is positive.

**Correlation Coefficient:**

The quantity \( r \), called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the pearson product moment correlation coefficient in honor of its developer Karl pearson. The mathematical formula for computing \( r \) is:

\[
r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2} \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}
\]

Where \( n \) is the number of pairs of data. The value of \( r \) is such that \(-1 \leq r \leq 1\). The + and - signs are used for positive linear correlation and negative correlations, respectively.

**Positive Correlation:** If \( x \) and \( y \) have a strong positive linear correlation, \( r \) is close to +1. An \( r \) value of exactly +1 indicates a perfect positive for positive value indicate a relationship between \( x \) an \( y \) variables such that as values for \( x \) is values for \( y \) also increase.
Negative Correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increases, values for y decrease.

No Correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

Note that r is a dimensionless quantity; that is, it does not depend on the units employed. A perfect correlation of ±1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination, $r^2$ or $R^2$

The coefficient of determination, $r^2$, is useful because it gives the proportion of the variance (fluctation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model / graph.
The Coefficient of determination is the ratio of the explained variation. To the total variation the Coefficient of determination is such that $0 \le \hat{R}^2 \le 1$, and denotes the strength of the linear association between $x$ and $y$.

The Coefficient of determination represents the percent of the data that is closest to the line of best fit.

The Coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

**Simple Linear Regression Model**

Suppose there are $n$ data points $(Y_i, X_i)$, where $i=1, 2, \ldots, n$. The goal is to find the equation of the straight line

$$y = \alpha + \beta x$$

Which would provide a “best” fit for the data points. Here the “best” will be understood as in the least-squares approach. Such a line that minimizes the sum of squared residuals of the linear regression model.

In other words, numbers $\alpha$ and $\beta$ solve the following minimization problem:

$$\text{Find } \min_{\alpha, \beta} Q(\alpha, \beta), \quad \text{where } Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

132
By using either calculus, the geometry of inner product spaces or simply expanding to get a quadratic in $\alpha$ and $\beta$ it can be shown that the values of $\alpha$ and $\beta$ that minimize the objective function $Q$ are:

$$
\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}
= \frac{\text{cov}[x,y]}{\text{var}[x]} = \frac{r_{xy}}{s_x}
$$

$$
\hat{\alpha} = y - \hat{\beta}x,
$$

Where $r_{xy}$ is the sample correlation Coefficient between $X$ and $Y$, $S_x$ is the standard deviation of $X$, and $S_y$ is correspondingly the standard deviation of $y$.

Horizontal bar over a variable means the sample average of that variable.

Eg. $xy = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$

Substituting the above expressions for $\hat{\alpha}$ and $\hat{\beta}$ into

$$
y = \hat{\alpha} + \hat{\beta}x
$$

yields

$$
\frac{y - \bar{y}}{S_y} = \frac{x - \bar{x}}{S_x}
$$

This shows the role $r_{xy}$ plays in the regression line of standardized data points.
Properties:

1) The line goes through the "center of mass" point \((\bar{x}, \bar{y})\)

2) The sum of the residuals is equal to zero, if the model includes a constant:
\[
\sum_{i=1}^{n} \hat{e}_i = 0
\]

3) The linear Combination of the residuals, in which the coefficients are the x-values, is equal to zero:
\[
\sum_{i=1}^{n} x_i \hat{e}_i = 0
\]

4) The estimators \(\hat{\alpha}\) and \(\hat{\beta}\) are unbiased. This requires that we interpret the estimators as random variables and so we have to assume that, for each value of \(x\), the corresponding value of \(Y\) is the generated as a mean response \(\alpha + \beta x\) plus an additional random variable \(\epsilon\) called the errors term. This error term has to be equal to zero on average, for each value of \(x\). under such interpretation, the least-squares estimators \(\hat{\alpha}\) and \(\hat{\beta}\) will themselves be random variables and they will unbiasedly estimate the "true values" \(\alpha\) and \(\beta\).

3.3. UNIVARIATE PROBABILITY DISTRIBUTION

Characteristic function is defined as an expectation value of the function –
\[e^{i(tx)}\]
\[ C(t) = \int e^{itx} f(x) dx \]

Moment generating function is defined as an expectation of \(e^{tx})\):

\[ M(t) = \int e^{tx} f(x) dx \]

Moments can be calculated in the following way obtain derivative of \(M(t)\) and take the value of it at \(t=0\)

\[
E(x^r) = \frac{d^r M(t)}{dt^r} \bigg|_{t=0}
\]

Cumulant generating function is defined as logarithm of the characteristic function

\[ c.g.f = \log (C(t)) \]

**Discrete distributions: Binomial**

Let us assume that we carry out experiment and the result of the experiment can be "success" or "failure" the probability of "success" in one experiment is \(p\). Then probability of failure is \(q=1-p\). We carry out experiments \(n\) times.

Distribution of \(k\) success is binomial:

\[
p(k) = P(X = K) = \binom{n}{k} p^k (1-p)^{n-k} - \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}
\]

**Characteristic function:**
\[ C(t) = (pe(it) + 1 - p)^n \]

Moment generating function:

\[ M(t) = (pe(t) + 1 - p)^n \]

As the number of trials become increases the distribution becomes more symmetric and dense. Mean value is np. Variance is npq = np(1-p)

**Discrete Distributions: Poisson**

When the number of the trials (n) is large and the probability of successes (p) is small and np is finite and tends to \( \lambda \) as \( n \) goes to infinity then the binomial distributions converges to Poisson distribution

\[ p(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots, \lambda > 0 \]

Poisson distribution is used to describe the distributions of an event that occurs rarely (rare events) in a short time period. It is used in counting statistics to describe the number of registered photon.

**Characteristic function is:**

\[ C(t) = e(\lambda(e^{it} - 1)) \]

As \( \lambda \) increases the distribution becomes more and more symmetric.

Expected values is \( \lambda \) and variance is \( \lambda \)
Variance and mean are equal to each other.

**Discrete distributions: Negative Binomial:**

Consider an experiment:

Probability of "success" is $p$ and probability of failure is $q = 1 - p$. We carry out the experiment until $k$-th success. We want to find the probability of $j$ failures before having $k$-th success (It is called sequential sampling. Sampling is carried out until stopping rule – $k$ success – is satisfied). If we have $j$ failures than it means that the number of trials is $k + j$

Last trial was success. Then the probability that we will have $j$ failures is:

$$p(j) = P(x = j) = \binom{k + j - 1}{j} p^k q^j p$$

$$= \binom{k + j}{j} p^k q^j p \quad j = 0, 1, 2, .......$$

It is called negative binomial because coefficients have the same from as those of the terms of the negative binomial series $p^{-k} = (1 - q)^{-k}$

Characteristic function is:

$$C(t) = p^k (1 - qe^{it})^{-k}$$

As the number of required successes increases the distribution becomes more and more symmetric. Mean value is $kp/q$ and variance is $kq(q+1)p$. 

137
Continuous distributions: Uniform:

The simplest form of the continuous distribution is the uniform with density:

\[ f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \]

Cumulative distribution function is:

\[ F(x) = \begin{cases} 0 & x < a \\ x - a & a \leq x \leq b \\ b - a & x > b \end{cases} \]

Continuous distributions: Exponential:

Density of random variable with an exponential distribution has the form:

\[ f(t) = \lambda e^{-\lambda t} \quad 0 < t < \alpha \]

One of the origins of this distribution:

From Poisson type random processes. If probability distribution of \( j(t) \) events occurring during time interval \((0, t)\) is a Poisson with mean value \( \lambda t \) then probability of time elapsing till the first event occurs has the exponential distribution.

Let \( T_r \) denotes time elapsed until \( r \)'th event

\[ P(j(t) < r) = P(T_r > t) \]
Putting \( r = 1 \) we get \( e^{-\lambda t} \). Taking into account that \( P(T_i > t) = 1 - F_i(t) \) and getting its derivative \( w_n \). We arrive to the exponential distribution.

Characteristic function is:

\[
C(u) = \left( 1 - \frac{iu}{\lambda} \right)^{-1}
\]

As lambda becomes larger, fall of the distribution becomes sharper.

Mean value is \( 1/\lambda \) and variance is \( (\lambda+1)/\lambda^2 \).

Continuous distributions: Gamma

Gamma distributions can be considered as a generalisation of the exponential distribution.

It has the form:

\[
f_r(t) = \frac{\lambda^r t^{r-1} e^{-\lambda t}}{(r-1)!}, 0 < t < \infty
\]

It is probability of time \( t \) elapsing before exactly \( r \) events happens.

Characteristic function of this distribution is

\[
C(u) = \left( 1 - \frac{iu}{\lambda} \right)^r
\]

If there are \( r \) independently and identically exponentially distributed random variables then the distribution of their sum is Gamma.
Sometimes for gamma distribution $1/\lambda$ instead of $\lambda$ is written. Implementation in R uses this form. $r$ is called shape and $1/\lambda$ is called scale parameter.

As the shape parameter increases the centre of the distribution shifts to the left and it becomes more symmetric.

Mean value is $r/\lambda$ and variance is $r(\lambda + 1)/\lambda^2$.

**Normal Distributions**

The general formula for the probability density function of the normal distributions is

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu$ is the location parameter and $\sigma$ is the scale parameter.

The case where $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution.

The equation for the standard normal distribution is

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Characteristic function is

$$\phi(t) = e^{i\mu t - \frac{t^2 \sigma^2}{2}}$$
The formula for the hazard function of the normal distributions is

$$ h(x) = \frac{\Phi(x)}{\Phi(-x)} $$

where $\Phi$ is the cumulative distribution function of the standard normal distribution and $\phi$ is the probability density function of the standard normal distribution.

Mean = $\mu$ ; Variance = $\mu$ ; Mode = $\mu$

Range = Infinity in both directions

Standard deviation : The scale parameters $\sigma$

Coefficient of variance: $\sigma / \mu$

Skewness: 0

Kurtosis : 3

The location and scale parameters of the normal distribution can be estimated with the sample mean and sample standard deviation, respectively.
3.4. \textsc{Student’s T- Test Statistic}

Let $x_i$, $(i = 1, 2, \ldots, n)$ be a random sample of size $n$ from a normal population with mean $\mu$ and variance $\sigma^2$. Then Student’s $t$ is defined by the statistic

$$ t = \frac{x - \mu}{s \sqrt{n}} $$

where $x = \frac{1}{n} \sum_{i=1}^{n} x_i$, is the sample mean and

$$ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - x)^2. $$

is an unbiased estimate of the population variance $\sigma^2$, and it follows student’s $t$-distribution with $V = (n-1)df$. with probability density function.

$$ f(t) = \frac{1}{\sqrt{\beta} \Gamma\left(\frac{1}{2}\right)} \left[ 1 + \frac{t^2}{\nu} \right]^{-\frac{\nu+1}{2}} $$

(3.4.2)

Remarks:

1. A statistic $t$ following student’s $t$-distribution with $n$ d.f will be abbreviated as $t \sim t_n$.

2. If we take $V=1$ in (3.4.2), we get
\[ f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \frac{1}{\pi \left(1+t^2\right)} \] 
\[ -\alpha < t < \alpha \quad \therefore V\left(\frac{1}{2}\right) = \sqrt{\pi} \]

Which is the p.d.f of standard Cauchy distribution. Hence, when \( v = 1 \) student's t distribute reduces to Cauchy distribution.

**Derivation of student's t-distribution:**

The expression (1) can be rewritten as

\[ t' = \frac{n(x - \mu)}{s^2} = \frac{n(x - \mu)}{ns^2(n-1)} \quad \therefore ns^2 = (n-1)s^2 \]

\[ \Rightarrow \quad t' = \frac{(x - \mu)}{s^2(n-1)} \cdot \frac{1}{ns'\sigma^2} = \frac{(x - \mu)}{\sigma^2 n} \cdot \frac{1}{ns'\sigma^2} \]

Since \( x_i \), \( i = 1, 2, ..., n \) is a random sample from the normal population with mean \( \mu \) and variance \( \sigma^2 \),

\[ x \sim N\left(\mu, \sigma^2/n\right) \Rightarrow \frac{(x - \mu)}{\sigma\sqrt{n}} \sim N(0, 1) \]

Hence \( \frac{(x - \mu)}{\sigma\sqrt{n}} \), being the square of a standard normal variate is a chi-square variate with 1 d.f.

Also \( \frac{ns^2}{\sigma^2} \) is a \( \chi^2 \)-variate with \( (n-1) \) d.f. [C.f. Theorem]
Further since \( \bar{x} \) and \( S^2 \) are independently distributed [c.f. Theorem], \( \frac{t^2}{n-1} \) being the ratio of two independent \( \chi^2 \)-variates with 1 and \((n-1)\) d.f respectively, is a \( \beta_\left( \frac{1}{2}, \frac{n-1}{2} \right) \) variate and its distribution is given by.

\[
dF(t) = \frac{1}{B\left( \frac{1}{2}, \frac{n-1}{2} \right)} \left( \frac{1}{t^2} \right) \left( \frac{1}{1+t^2} \right)^{\frac{n-1}{2}} d\left( t^2 / v \right), 0 \leq t^2 < \alpha
\]

[Where \( v = (n-1) \)]

\[
\sqrt{v} \left( \frac{1}{2}, \frac{n}{2}, \frac{1}{t^2} \right) = \frac{1}{B\left( \frac{1}{2}, \frac{n}{2} \right)} \left( \frac{1}{1+t^2} \right)^\frac{n-1}{2} dt, \quad -\alpha < t < \alpha
\]

The factor 2 disappearing since the integral from -\( \alpha \) to \( \alpha \) must be unity.

This is the required probability function as given in (2) of student's \( t \)-distribution with \( v = (n-1) \) d.f.

### 3.5. CHI–SQUARE TEST FOR POPULATION VARIANCE

Suppose we want to test if a random sample \( x_i, (i = 1, 2, \ldots, n) \) has been drawn from a normal population with a specified variance \( \sigma^2 = \sigma_0^2 \) (say)

Under the null hypothesis that the population variance is \( \sigma^2 = \sigma_0^2 \), the statistic

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{\sigma_0} \right)^2 = \frac{1}{\sigma_0^2} \left[ \left( \sum_{i=1}^{n} x_i \right)^2 - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n} \right] = nS^2 / \sigma_0^2
\]

(3.5.1)
follows Chi-Square distribution with \((n-1)\) df.

By comparing the calculated value with the tabulated value of \(\chi^2\) for \((n-1)\) df at certain level of significance, (usually 5%) we may retain or reject the null hypothesis.

Remarks: 1) The above test (3.5.1) can be applied only if the population from which sample is drawn is normal.

2) If the sample size \(n\) is large (> 30), then we can use Fisher’s approximation

\[
\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)
\]

i.e. \(Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1)\)

and apply Normal test

3) For a detailed discussion on the significant values, (critical values), for testing \(H_0: \sigma^2 = \sigma_0^2\), against various alternatives:

(i) \(\sigma^2 > \sigma_0^2\)  
(ii) \(\sigma^2 < \sigma_0^2\)  
and  
(iii) \(\sigma^2 \neq \sigma_0^2\).

3.5.1. Chi-Square Test Goodness of Fit

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson ion in 1900 and is known as “Chi – Square test of goodness of fit”. It enables us to find if the deviation of
the experiment from theory is just by chance or is it really due to the inadequacy of the Theory to fit the observed data.

If \( O_i \) (\( i = 1, 2, \ldots, n \)) is a set of observed (experimental) frequencies and \( E_i \) (\( i = 1, 2, \ldots, n \)) is the corresponding set of expected (Theoretical or hypothetical frequencies, then Karl Pearson's Chi-Square, given by

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{(O_i - E_i)^2}{E_i} \right) = \left( \sum_{i=1}^{n} O_i \right) - \left( \sum_{i=1}^{n} E_i \right)
\]

The Chi-Square probability density function is

\[
f(y) = \frac{1}{2^{n/2} \Gamma(n/2)} y^{(n/2)-1} e^{-y/2}
\]

where \( y > 0 \)

\( f(y) = 0 \) for \( y \leq 0 \)

\( \Gamma = \) Gamma function

\( Y \) has a Chi-Square (\( \chi^2 \)) distribution with \( n \) d.f

If \( Y \) has a Chi-Square distribution with \( n \) d.f then its distribution function is

\[
F(y) = P[Y \leq y] = \int f(y) dy
\]

where \( f(y) = \) density function
F(y) is a probability

If F(y) = p where p is a constant, the corresponding value of y associated with p is $\chi^2_{\alpha}$.

This is known as the $p^{th}$ fractile or $p^{th}$ percentage of the Chi-Square distribution with n.d.f.

Mean of $Y$, $\mu_Y = E(Y) = n \cdot \mu$ $\sigma^2_Y = \sigma^2 = E[(Y - \mu)^2] = 2n$

3.6. KRUSKAL WALLIS TEST

Analysis of variance (ANOVA) is a data analysis technique for examining the significance of the factors (independent variables) in a multi factor model. The one factor model can be thought of as a generalization of the two sample t-test. That is, the two sample t-test is a test of the hypothesis that two population means are equal. The one factor ANOVA tests the hypothesis that k population means are equal.

The Kruskal Wallis test can be applied in the one factor ANOVA case. It is a non-parametric test for the situation where the ANOVA normality assumptions may not apply.

Let $n_i (i = 1, 2, \ldots, k)$ represent the sample sizes for each of the k groups (i.e. samples) in the data. Next, rank the combined sample. Then compute $R_i = \text{the sum of the ranks for group } i$, then the Kruskal Wallis test statistic is:
\[ H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \]

This statistic approximate a Chi-Square distribution with \( k-1 \) degrees of freedom if the null hypothesis of equal populations is true. Each of the \( n_i \) should be at least 5 for the approximation to be valid. We reject the null hypothesis of equal population means if the test statistic \( H \) is greater than CHIPPF (\( \text{ALPHA, } k-1 \)).

Where CHIPPF is the Chi-Square percent point function.

\( H_0: \mu_1 = \mu_2 = \ldots = \mu_k \)

\( H_A: \mu_i \neq \mu_j \) for at least one set of \( i \) and \( j \)

Test Statistic: \[ H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \]

Significance Level: \( \alpha \), typically set to 0.05

Critical Region: \( H > \text{CHIPPF} (\alpha, k-1) \) where CHIPPF is the chi-square percent point function.

Conclusion: Reject the null hypothesis if the test statistic is in the critical region.

3.7. FRIEDMAN TEST:

The Friedman test is a non-parametric test for analyzing randomized complete block designs. It is an extension of the sign test when there may be more than two treatments (\( k \geq 2 \)). The Friedman test assumes that there are \( k \) experimental.
The observations are arranged in \( b \) blocks, that is

<table>
<thead>
<tr>
<th>Block</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( X_{11} )</td>
<td>( X_{12} )</td>
<td>\ldots</td>
<td>( X_{1k} )</td>
</tr>
<tr>
<td>2</td>
<td>( X_{21} )</td>
<td>( X_{22} )</td>
<td>\ldots</td>
<td>( X_{2k} )</td>
</tr>
<tr>
<td>3</td>
<td>( X_{31} )</td>
<td>( X_{32} )</td>
<td>\ldots</td>
<td>( X_{3k} )</td>
</tr>
<tr>
<td>( b )</td>
<td>( X_{b1} )</td>
<td>( X_{b2} )</td>
<td>\ldots</td>
<td>( X_{bk} )</td>
</tr>
</tbody>
</table>

Let \( R(X_{ij}) \) be the rank assigned to \( X_{ij} \) within block \( i \) (i.e., ranks within a given row). Average ranks are used in the case of ties. The ranks are summed to obtain

\[
R_i = \sum_{i=1}^{k} R(X_{ij})
\]

Then the Friedman test is

\[ H_0: \text{The treatment effects have identical effects.} \]

\[ H_a: \text{At least one treatment is different from at least one other treatment.} \]

Test Statistic: \[ T_i = \frac{12}{bk(k+1)} \sum_{i=1}^{b} (R_i - b(k+1)/2)^3 \]

149
If there are ties, then

\[
T_i = \frac{(k-1) \sum R - \frac{b(k+1)}{2} \sum A_i - C_i}{2}
\]

where

\[
A_i = \sum \sum (R(x_j))^2
\]

\[
C_i = \frac{bk(k+1)^3}{4}
\]

Note that Conover recommends the statistic

\[
T_j = \frac{(b-1)T_i}{b(k+1) - T_i}
\]

Since it has a more accurate approximate distribution. The \(T^2\) statistic is the two-way analysis of variance statistic computed on the ranks \(R(X_i)\).

Significance level: \(\alpha\)

Critical Region: \(T > F(\alpha, k-1, (b-1)(k-1))\)

where \(F\) is the percent point function of the \(F\)-distribution

\[T_i > \chi^2_{\alpha, k-1}\]

where \(\chi^2\) is the percent point function of the Chi-Square distribution.
The $T_1$ approximation is sometimes poor, so the $T_2$ approximation is typically preferred.

1: Reject the null hypothesis if the test statistic is in the critical region. If the hypothesis of identical treatment effects is rejected, it is often desirable to determine which treatments are different (i.e. multiple comparisons). Treatments $i$ and $j$ are considered different if:

$$|R_i - R_j| > t_{(1-\alpha/2,b-1,k-1)} \sqrt{\frac{2(bA_i - \sum R_i^2)}{(b-1)(k-1)}}$$

Note: The Friedman test is based on the following assumptions:

1) The $b$ rows are mutually independent. That is the results within one block (row) do not affect the results within other blocks.

2) The data can be meaningfully ranked.

3.8. ANOVA TECHNIQUE

The great British statistician R.A. Fisher developed the analysis of variance method in the 1920s.

The heart of this analysis is a significance test, using his $F$ distribution, for detecting differences among a set of population means.

The ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these
estimates due to assignable factors causes with the estimate due to chance factor (causes), the later being known as experimental error or simply error.

Assumptions: For the validity of the F-test in ANOVA, The following assumptions are made:

(i) The observations are independent

(ii) Parent population from which observations are taken is normal and

(iii) Various treatment and environmental effects are additive in nature.

3.8.1. Cochran's Theorem: (Fundamental for ANOVA)

Let \( X_1, X_2, \ldots, X_n \) denote a random sample from normal population \( N(0, \sigma^2) \).

Let the sum of the squares of these values be written in the form

\[
\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k
\]

where \( Q_j \) is a quadratic form in \( X_1, X_2, \ldots, X_n \) with rank (degrees of freedom) \( r_j \), \( j=1, 2, \ldots, k \). Then the random variables \( Q_1, Q_2, \ldots, Q_k \) are mutually independent and \( Q_j/\sigma^2 \) is \( \chi^2 \)-variate with \( r_j \) degrees of freedom if and only if \( \sum r_j = n \).
3.8.2. ANOVA FOR ONE-WAY CLASSIFICATION:

Let us suppose that $N$ observations $x_{ij}$ ($i=1,2,...,k; j=1,2,...,ni$) of a random variable $x$ are grouped, on some basis, into $k$ classes of sizes $n_1,n_2,...,n_k$ respectively, $\left( N = \sum_{i=1}^{k} ni \right)$ as exhibited

<table>
<thead>
<tr>
<th>Means</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
</tr>
<tr>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_{k1}$</td>
<td>$x_{k2}$</td>
</tr>
</tbody>
</table>
The total variation in the observations $x_i$ can be split into the following components:

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation within the classes, i.e. the inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human endeavour and the second type of variation is due to chance causes which are beyond the control of human hand.

The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes. In particular, let us consider the effect of $k$ different rations on the yield in milk of $N$ cows (of the same breed and stock) divided into $k$ classes of sizes $n_1, n_2, \ldots, n_k$ respectively, $N = \sum n_i$.

Here the sources of variation are

(i) Effect of the ration (treatment) $t_i$, $i = 1, 2, \ldots, k$

(ii) Error produced by numerous causes of such magnitude that they are not detected and identified with the knowledge that we have and they together produce a variation of random nature obeying Gaussian (Normal) law of errors.
Mathematical Model: In this case the linear mathematical model will be

\[ x_{ij} = \mu_i + \epsilon_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

where \((i=1,2,\ldots,k; j=1,2,\ldots,n_i)\) \hspace{1cm} (3.8.2.1)

(i) \(x_{ij}\) is the yield from the \(j^{th}\) cow \((j=1,2,\ldots,n_i)\) fed on the \(i^{th}\) ration \((i=1,2,\ldots,k)\)

(ii) \(\mu\) is the general mean effect given by

\[ \mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{N} \] \hspace{1cm} (3.8.2.2)

where \(\mu_i\) is the fixed effect due to the \(i^{th}\) ration, i.e. of there were no treatment differences and no chance causes then the yield of each cow will be \(\mu\).

(iii) \(\alpha_i\) is the effect of the \(i^{th}\) ration given by

\[ \alpha_i = \mu_i - \mu \hspace{0.5cm} (i=1,2,\ldots,k) \] \hspace{1cm} (3.8.2.3)

i.e. the \(i^{th}\) ration increases or decreases the yield by an amount \(\alpha_i\), on using \((3.8.2.2)\) we get

155
\[
\sum_{i=1}^{N} n_{i} \alpha_{i} = \sum_{i=1}^{N} n_{i} (\mu_{i} - \mu) = \sum_{i=1}^{N} n_{i} \mu_{i} - \mu \sum_{i=1}^{N} n_{i} = N \mu - \mu N
\]

(iv) \( E_{ii} \) is the error effects due to chance.

**Assumptions in the model:**

(i) All the observations \( x_{ij} \) are independent.

(ii) Different effects are additive in nature

(iii) \( \epsilon_{ij} \) are i.i.d \( (N, 0, \sigma^{2}_{ij}) \)

Under the third assumption the model (3.8.2.1) becomes

\[
E(x_{ij}) = \mu_{i} = \mu + \alpha_{i} \quad \left( i = 1, 2, \ldots, k \right)
\]

\[
\left( j = 1, 2, \ldots, n_{i} \right)
\]

**Null hypothesis:** we want to test the equality of the population means, i.e., the homogeneity of different rations. Hence null hypothesis is given by

\[ H_{0} : \mu = \mu_{2} = \ldots = \mu_{k} = \mu \quad (3.8.2.4) \]

Which from (3.8.2.4) reduces to

\[ H_{0} : \alpha_{1} = \alpha_{2} = \ldots = \alpha_{k} = 0 \]
### ANOVA TABLE FOR ONE-WAY CLASSIFIED DATA

<table>
<thead>
<tr>
<th>Sources of variation</th>
<th>Sum of Squares</th>
<th>d.f</th>
<th>Mean sum of Squares</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (Ration)</td>
<td>$S_t^2$</td>
<td>k-1</td>
<td>$S_t' = \frac{S_t^2}{k-1}$</td>
<td>$S_t' = F_{k-1, N-k}$</td>
</tr>
<tr>
<td>Error</td>
<td>$S_e^2$</td>
<td>N-k</td>
<td>$S_e' = \frac{S_e^2}{N-k}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$S_t^2$</td>
<td>n-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 3.8.3. ANOVA FOR TWO-WAY CLASSIFICATION

(With one observation Per cell)

Let us consider the case when there are two factors which may affect the variable values $x_{ij}$, e.g., the yield of milk may be affected by differences in treatments, i.e., rations as well as the differences in variety, i.e., breed and stock of the cows. Let us now suppose that the N cows are divided into h different groups or classes according to their breed and stock, each group containing k cows and then let us consider the effect of k treatments (i.e., rations given at random to cows in each group) on the yield of milk.
Let the suffix \( i \) refer to the treatments (rations) and suffix \( j \) refer to the varieties (bread of the cow). Then the yields of milk \( x_{ij} (i = 1, 2, \ldots, k; j = 1, 2, \ldots, h) \) of \( N = h \times k \) cows furnish the data for the comparison of the treatments. The yields may be expressed as variate values in the following \( k \times h \) two-way table.

\[
\begin{array}{cccc}
\text{Means} & \text{Totals} \\
\end{array}
\begin{array}{ccccccc}
x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1h} & x_{\bar{1}} & T_1 \\
x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2h} & x_{\bar{2}} & T_2 \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots \\
x_{k1} & x_{k2} & \cdots & x_{kj} & \cdots & x_{kh} & x_{\bar{k}} & T_k \\
\end{array}
\]

\[
\begin{array}{cccc}
\text{Means} & x_{\bar{1}} & x_{\bar{2}} & \cdots & x_{\bar{j}} & \cdots & x_{\bar{h}} & x \\
\text{Totals} & T_1 & T_2 & \cdots & T_j & \cdots & t_h & G \\
\end{array}
\]

**Mathematical Model**

Let \( x_{ij} \) be the yield from the cow of \( j^{th} \) variety fed on the \( i^{th} \) ration \((i = 1, 2, \ldots, k; j = 1, 2, \ldots, h)\).

Let us suppose that \( x_{ij}, (i = 1, 2, \ldots, k; j = 1, 2, \ldots, h) \) are independent, normally distributed as \( N (\mu_{ij}, \sigma^2) \). The linear mathematical model becomes

\[
E(x_{ij}) = \mu_{ij} \quad \text{ (3.8.3.1)}
\]

\[
x_{ij} = \mu_{ij} + \epsilon_{ij}
\]

or

where \( \epsilon_{ij} \) are i.i.d \( N (0, \sigma^2) \)
\( \mu \) is further split into the following parts:

i) The general mean effect \( \mu \) is given by

\[
\mu = \sum_i \sum_j \mu_{ij} / N
\]  

(3.8.3.2)

ii) The effect \( L_i \) (\( i = 1, 2, \ldots, k \)) due to the \( i^{th} \) ration given by

\[
L_i = \mu_i - \mu
\]

where,

\[
\mu_i = \frac{1}{h} \sum_{j=1}^{h} \mu_{ij} \quad (i = 1, 2, \ldots, k)
\]

Obviously,

\[
\sum_{i=1}^{k} \alpha_i = 0
\]

iii) The effect is \( \beta_j \) (\( j = 1, 2, \ldots, h \)) due to the \( j^{th} \) variety (breed of cow) given by

\[
\beta_j = \mu_j - \mu
\]

where

\[
\mu_j = \frac{1}{k} \sum_{i=1}^{k} \mu_{ij} \quad (j = 1, 2, \ldots, h)
\]

Obviously,

\[
\sum_{j=1}^{h} \beta_j = 0
\]
iv) The interaction effect $\gamma_{ij}$ when the $i^{th}$ level of first factor (rations) and $j^{th}$ level of second factor (breed of cow) occur simultaneously and given by

$$\gamma_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$$

where $\sum_i \gamma_{ij} = 0, \quad \forall \ i = 1, 2, \ldots, k$

and $\sum_j \gamma_{ij} = 0, \quad \forall \ j = 1, 2, \ldots, h$

Thus we have

$$\mu_{ij} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + (\mu_{ij} - \mu_i - \mu_j + \mu)$$

and consequently the model (3.8.3.1) becomes

$$x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$  \hspace{1cm} (3.8.3.3)

where $\epsilon_{ij}$ is the error effect due to chance and

$$\sum_i \alpha_i = 0 = \sum_j \beta_j$$

and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$

for all $i$ \quad for all $j$
As there is only one observation in each cell, the observation corresponding to the $i^{th}$ level of ration and $j^{th}$ level of breed of cow is only one i.e. $x_{ij}$. But we cannot estimate by one value alone. Hence in this case (one observation per cell), the interaction effect $\gamma = 0$ and the model (3.8.3.3) reduces to

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$ \hspace{1cm} (3.8.3.4)

### ANOVA TABLE FOR TWO WAY CLASSIFIED DATA

<table>
<thead>
<tr>
<th>Sources of variation</th>
<th>Sum of squares (S.S)</th>
<th>d.f</th>
<th>M.S.S</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>$S_i^2 = \sum_i h(x_i - \bar{x})^2$</td>
<td>$k - 1$</td>
<td>$S_i^2 = \frac{S_i^2}{(k-1)}$</td>
<td>$F_i = \frac{S_i^2}{S_j^2}$</td>
</tr>
<tr>
<td>Varieties</td>
<td>$S_i^2 = \sum_j h(x_j - \bar{x})^2$</td>
<td>$h - 1$</td>
<td>$S_i^2 = \frac{S_i^2}{(h-1)}$</td>
<td>$F_i = \frac{S_i^2}{S_j^2}$</td>
</tr>
<tr>
<td>Residual (Or Error)</td>
<td>$S_i^2 = \sum_{ij} (x_{ij} - \bar{x}, \bar{x})^2$</td>
<td>$(h-1)(k-1)$</td>
<td>$S_i^2 = \frac{S_i^2}{(h-1)(k-1)}$</td>
<td>$F_i = \frac{S_i^2}{S_j^2}$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum_{ij} (x_{ij} - \bar{x})^2$</td>
<td>$hk - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

161
3.9. MULTIPLE COMPARISONS PROCEDURES:

The terminology "a set of simultaneous 100 \((1-\alpha)\)% confidence intervals" will always refer to the fact that the overall confidence level for a set of contrasts or treatments means is (at least) 100 \((1-\alpha)\)%.

Each of the methods discussed gives confidence intervals of the form:

\[
\sum l_{i} e \left[ \sum l_{i} \pm W \sqrt{\text{Var}(\sum l_{i})} \right] \tag{3.9.1}
\]

where \(W\), which we call the critical coefficient, depends on the method, the number of treatments \(\gamma\), on the number of confidence intervals calculated, and on the number of error of freedom the term

\[\text{nsd} = W \sqrt{\text{Var}(\sum l_{i})} \tag{3.9.2}\]

Which is added and subtracted from the least square estimate in (3.9.1) is called the minimum significant difference, because if the estimate is larger then nsd, the confidence interval excludes zero, and the contrast is significantly different from zero.

3.10 The Least Significant Difference (LSD) Method

Suppose that, following an analysis of variance F test where the null hypothesis is rejected, we wish to test \(H_0: \sum l_{i} e = 0\) against the alternative hypothesis \(H_1: \sum l_{i} e \neq 0\).
For making pairwise comparisons, consider the contrasts of the type \( t_i - t_j \) in which experiments are often interested, are obtainable from \( \sum l_i \) by putting \( l_i = 1 \), \( l_j = -1 \) and zero for the other's. The 100(1-\( \alpha \))% confidence interval for this contrast is

\[
\sum l_i t_i \in \left( \sum l_i t_i \pm t_{edf, \alpha/2} \sqrt{Var(\sum l_i t_i)} \right)
\]

(3.10.1)

Where edf denotes the error degrees of freedom. As we know that the out of a hypothesis test can be deduced from the corresponding confidence interval in the following way. The null hypothesis will be rejected at significance level \( \alpha \) in favour of the two - sided alternative hypothesis if the corresponding confidence interval for \( \sum l_i t_i \) fails to contain 0. The interval fails to contain 0 if the absolute value of \( \sum l_i t_i \) is bigger than \( t_{edf, \alpha/2} \sqrt{Var(\sum l_i t_i)} \). The critical differences or the least significant difference for testing the significance of the difference of two treatment effects say \( t_i - t_j \) is \( lsd = t_{edf, \alpha/2} \sqrt{Var(\sum l_i t_i)} \) where \( t_{edf, \alpha/2} \) is the value of student's t at the level of significance \( \alpha \) and error degree of freedom. If the difference of any two- treatment means is greater than the lsd value, the corresponding treatment effects are significantly different. The above formula is quite general and particular cases can be obtained for different experimental designs. For example, the least significant difference between two treatment effects for a randomized complete block (RCB) design, with V treatments and r
replications is \( t(v-1)(r-1) \), \( \alpha/2\sqrt{2\text{MSE}/r} \). Where \( t(v-1)(r-1) \), \( \alpha/2 \) is the value of students t at the level of significance \( \alpha \) and degree of freedom \( (v-1)(r-1) \).

For a completely randomized design with treatments such that \( i^{th} \) treatment is replicated \( r_i \) times and \( \sum_i r_i = n \), the total number of experimental units, the least significance difference between two treatment effect is

\[
t_{v-1} \frac{\alpha}{2} \sqrt{\text{MSE} \left( \frac{1}{r_i} + \frac{1}{r_j} \right)}
\]

It may be worthwhile mentioning here that the least significant difference method is suitable only for planned pair comparisons. This test is based on individual error rate. However, for those who wish to use it for all possible pair wise comparisons, should apply only after the F-test in the analysis of variance is significant at desired level of significance. This procedure is often referred Fisher's protected lsd. Since F calls for us to accept or reject a hypothesis simultaneously involving means. If we arrange the treatment means in ascending or descending order of their magnitude and keep the, means in one group for which the difference between the smallest and largest mean is less than the lsd, we can identify the homogeneous subsets of treatments.

3.11. DUNCAN'S MULTIPLE RANGE TEST

A widely used procedure for comparing all pairs of means is the multiple range lost developed by Duncan (1955). The application of Duncan's multiple
range test (DMRT) is similar to that of lsd test. DMRT involves the computation of numerical boundaries that allow for the classification of the difference between any two treatment means as significant or non significant. DMRT requires computation of a series of value each corresponding to a specifies set of pair comparisons unlike a single value for all pairwise comparisons in case of lsd. It primarily depends on the standard error of the mean difference as in case of lsd. This can easily be worked out using the estimate of variance of an estimated elementary treatment contrast through the design.

The value of the least significant range

\[ R_p = \frac{t_{\alpha}(p, edf) \times SE_{\bar{x}}}{\sqrt{2}} \]

where \( \alpha \) is the desired significance level, edf is the error degrees of freedom and \( p = 2, \ldots, v \) is one more than the distance in rank between pairs of the treatment means to be compared. If the two treatment means have consecutive rankings, then \( p = 2 \) and for the highest and lowest means it is \( v \). The values of \( t_{\alpha}(p, edf) \) can be obtained from Duncan's table of significant ranges.

Then, the observed differences between means are tested beginning with largest versus smallest, which would be compared with the least significant range \( R_v \). Next, the difference of the largest and the second smallest is computed and compared with the least significant range \( R_{v-1} \). These comparisons are continued until all means have been compared with the largest mean.
Finally, the difference of the second largest mean and the smallest computed and compared against the least significant range \( R_{m,1} \). This process is continued until the differences of all possible \( v(v-1)/2 \) pairs of means have been considered. If an observed difference is greater than the corresponding least significant range, then we conclude that the pair of means in question is significantly different. To prevent contradictions, no differences between a pair of means are considered significant if the two means involved fall between two other means that do not differ significantly.

The confidence intervals of the desired pairwise comparisons is

\[
\sum l_i \in \sum l_i \pm \frac{r_c(p, edf)}{\sqrt{2}} \sqrt{\text{var} \sum l_i}
\]

and least significant range in general is

\[
\text{lsr} = \frac{r_c(p, edf)}{\sqrt{2}} \sqrt{\text{var} \left[ \sum l_i \right]}
\]

In \( m \) simultaneous confidence intervals are calculated for preplanned contrasts, and if each confidence interval has confidence level \( 100(1 - \alpha^*) \% \) then the overall confidence level is greater than or equal to \( 100(1 - m \alpha^*) \%

3.12. BONFERRONI METHOD FOR PREPLANNED COMPARISONS:

In this method the overall Confidence level of \( 100 \times (1 - \alpha^* \)% for \( m \) simultaneous confidence intervals can be ensured by setting \( \alpha^* = \alpha/m \)
Replacing $\alpha$ by $\alpha/m$ in the formula

$$\sum_{i} l_i \in \left[ \sum_{i} l_i, \pm t_{\alpha/2} \sqrt{\text{Var} \left[ \sum_{i} l_i \right]} \right]$$

For an individual confidence interval, we obtain a formula for a set of simultaneous $100(1-\alpha^*)$% confidence intervals for $m$ preplanned contrasts $\sum l_i$, is

$$\sum_{i} l_i \in \left[ \sum_{i} l_i, \pm t_{\alpha/2m} \sqrt{\text{Var} \left[ \sum l_i \right]} \right]$$

Therefore, if the contrast estimate is greater than $t_{\alpha/2m} \sqrt{\text{Var} \left[ \sum l_i \right]}$, the corresponding contrast is significantly different from zero. Therefore, if the contrast estimate is greater than $t_{\alpha/2m} \sqrt{\text{Var} \left[ \sum l_i \right]}$, the corresponding contrast is significantly different from zero.

If can easily be seen that this method is same as that of least significant difference with $\alpha$ in least significant difference to be replaced by $\alpha/m$ since $\alpha/(2m)$ is likely to be a typical value, the percentiles tedf, $\alpha/(2m)$ may need to be obtained. When $m$ is very large, $\alpha/(2m)$ is very small, possibly resulting in extremely wide simultaneously confidence intervals. In this case the Scheffe or Turkey methods would be preferred.
Note that this method can be used only for preplanned contrasts or any m in preplanned estimable contrasts or functions of the parameters. It gives shorter confidence intervals than the other methods if m is small.

3.13. SCHEFFE METHOD OF MULTIPLE COMPARISONS

In the Bonferroni method of multiple comparisons, the major problem is that the m contrasts to be examined must be preplanned and the confidence intervals can become very wide if n is large. Scheffe’s method, on the other hand, provides a set of simultaneous 100(1 - α*)% confidence intervals whose widths are determined only by the number of treatments and number of observations in the experiment. It is not dependent on the number of contrasts are of interest. It utilizes the fact that every possible contrast \( \sum i_j \), can be written as a linear combination of the set of \((v-1)\) treatment versus control contrasts, \( t_1 - t_j, t_j - t_1, \ldots, t_{v-1} - t_1 \), once the experimental data have been collected, it is possible to find a 100(1 - α*)% confidence region for these \( v-1 \) treatment versus control contrasts, it determines bounds for every possible contrast \( \sum i_j \), and, in fact, for any number of contrasts, while the overall confidence level remains fixed. Simultaneous confidence intervals for all the contrast \( \sum i_j \), can be obtained from the general formula

\[
\sum i_j, \in \sum i_j \pm W \sqrt{\text{Var}(\sum i_j)}
\]

by replacing the critical coefficient W by Ws where \( Ws = \sqrt{aFa, edf, \alpha} \) with a as the dimension of the space of linear estimable functions being considered, or
equivalently, a is the number of degrees of freedom associated with the linear estimate functions being considered. The Scheffe's method applies to any m estimate contrasts or functions of the parameters. It gives shorter intervals than Bonferroni methods when m is large and allows data snooping. It can be used for any design.

3.14. TUKEY METHOD FOR ALL PAIRWISE COMPARISONS

Tukey (1953) proposed a method for making all possible pairwise treatment comparisons. The test compares the difference between each pair of treatment effects with appropriate adjustment for multiple testing. This test is also known as Tukey's honestly significant difference test or Tukey's HSD. The confidence intervals obtained using this method are shorter than those obtained from Bonferroni and Scheffe methods.

The simultaneous confidence intervals for all the contrasts of the type \( \sum \alpha \), by replacing the critical coefficient \( W \) by \( W_t = q_{v, edf, \alpha/\sqrt{2}} \) where \( V \) is the number of treatments and edf is the error degree of freedom and values can be seen as the percentile corresponding to a probability level \( \alpha \) in the right hand tail of the studentized range distribution tables.

For the completely randomized design or the one – way analysis of variance model,
\[ \text{Var}(i,-i) = \sqrt{\text{MSE} \left( \frac{1}{\xi} + \frac{1}{\tau_i} \right)} \]

where \( \tau_i \) denotes the replication number of treatment \( i \)
\((i = 1, 2, \ldots, v)\).

Then Tukey's simultaneous confidence intervals for all pairwise comparisons \( r_i - r_j \), \( i \neq j \) with overall confidence level at least \( 100(1 - \alpha^*) \)% is obtained by taking
\[ W_t = q_{n,v} \sqrt{\frac{\alpha}{2}} \text{ and } \text{Var}(i,-i) = \sqrt{\text{MSE} \left( \frac{1}{\xi} + \frac{1}{\tau_i} \right)}. \]

The values of \( q_{n,v}, \alpha^* \) can be seen the studentized range distribution tables.

When the sample sizes are equal \((\tau_i = \tau^*; i = 1, \ldots, v)\) the overall confidence level is exactly \( 100(1 - \alpha^*) \)% when the sample sizes are unequal, the confidence level is at least \( 100(1 - \alpha^*) \)%.

It may be mentioned here that Tukey's method is the best for all pairwise treatment comparisons. It can be used for completely randomized designs, randomized in complete block designs and balanced complete block designs.