3 Statistics and Modeling

Mukhtar Ahmed

Abstract

This chapter describes the application of statistical concepts with illustration about statistical models, probability, normal distribution, and analysis of variance (ANOVA). Statistical analysis is an important action process in research that deals with data. It follows well-defined, systematic, and mathematical procedures and rules. Data is information obtained to answer questions related to how much, how many, how long, how fast and how related. Statistics main objective is the analysis of data from generated experiment, but how should this data be collected to address our research questions and what should be our experimental design? Thus, in order to address question of interest clearly and efficiently, we need to organize experiment accurately so that we can have right type and amount of data. This is only possible using experimental design which has been elaborated in this chapter. The designs discussed here are completely randomized design (CRD), randomized complete block design (RCBD), Latin square design, nested and split plot design, strip-plot/split-block design, and split-split plot design. Similarly, factorial experiments have been discussed in detail with description about the interaction. The concept about fractional factorial design, multivariate analysis of variance (MANOVA), and analysis of covariance (ANCOVA) has been presented. Principal component analysis which is the method of multivariate statistics and used to check variation and patterns in a data set was also presented. It is easy way to visualize and explore data. The relationship between one or more variables to generate model which could be used for the prediction analysis has been discussed using concept of regression. Finally, association between two or more variables was presented using correlation. At the end different analytical

M. Ahmed (✉)
Department of Agricultural Research for Northern Sweden, Swedish University of Agricultural Sciences, Umeå, Sweden

Department of Agronomy, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan
e-mail: mukhtar.ahmed@slu.se; ahmadmukhtar@uaar.edu.pk

M. Ahmed (ed.), *Systems Modeling*, https://doi.org/10.1007/978-981-15-4728-7_3

tools/software were listed which can be used to do different kind of statistical analysis.

Keywords

Statistics · Probability · Normal distribution · Analysis of variance · Experimental designs · Factorial experiments · Regression · Correlation

3.1 Basic Statistics

Statistics is the science (pure and applied) dealing with creating, developing, and applying techniques to evaluate uncertainty of inductive inferences. It helps to answer the question about different hypothesis. It can model the role of chance in our experiments in a quantitative way and gives estimates with errors. Propagation of error in input values could also be determined by the statistics. History of statistics goes back to the experience of gambling (seventeenth century) which leads to the concept of probability. Afterwards concepts of normal curve/normal curve of error were introduced. Charles Darwin (1809–1882) work was largely biostatistical in nature. Karle Pearson (1857–1936) founded the journal *Biometrika* and school of statistics. Pearson was mainly concerned with large data, and his student W. S. Gosset (Pseudonym, Student) (1876–1937) presented Student's *t-test* which is a basic tool of statistician and experimenters throughout the globe. Genichi Taguchi (1924–2012) promoted the use of experimental designs.

Observations in the form of numbers are very important to perform different kind of statistical analysis. In case of crop production, observation can be phenology, leaf area, crop biomass, and yield. These numbers then constitute data, and its common characteristics include variability or variation. Variables may be quantitative or qualitative. Observations on quantitative variables may be further classified as discrete or continuous. Furthermore, probability of occurrence of value such as blondeness may be measured by probability function or probability density function (PDF). Chance and random variable terms are generally used for the variables possessing PDF. Population is all possible values of a variable, while part of population is called a sample. The concept of randomness is used to have true representative data sample from the population. Collected data could be characterized using tables, charts (pie chart, bars, etc.), and pictures (histogram). Afterwards data are presented in frequency tables, and measure of central tendency is used to locate center. This can help to find measure of spreading of the observation. Mean or average (μ) is the most common method to use the measure of central tendency. In case of dice, μ can be calculated by using following equation

$$\mu = \frac{1+2+3+4+5+6}{6} = 3\frac{1}{2} \tag{3.1}$$

If a sample is taken from the population having four observation, then $\overline{Y}$ (sample mean) for the four observation (3, 5,7,9) is

$$\overline{Y} = \frac{3+5+7+9}{4} = 6 \tag{3.2}$$

This can be further symbolized by

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{4} \tag{3.3}$$

where $Y_1 = \textit{value of first observation},\ Y_2 = \textit{value of second observation},\ \ Y_3 = \textit{value of third observation},\ \ \textit{and}\ Y_4 = \textit{value of fourth observation}$. For the *n*th observations, Y_i is used to represent the *i*th observation and ⍰Y is given by

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4 + \ldots + Y_i + \ldots. + Y_n}{n} \tag{3.4}$$

This equation can be further shortened to

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{3.5}$$

Difference between observations (Y_i) and sample mean $\left(\overline{Y}\right)$ is called sample deviation (Y_i– $\overline{Y}$), and its sum is equal to zero $\sum\left(Y_i - \overline{Y}\right) = 0$.

For the different number of observations, it's better to use weights that depend on the number of observations in each mean called weighted mean. A weighted mean is defined as follows:

$$\overline{Y}_w = \frac{\sum w_i Y_i}{\sum w_i} \tag{3.6}$$

Another term supplement to the mean is median and it is value for which 50% of the observations lie on each side. However, if values are even, then median is average of the two middle values, e.g., 3, 6, 8, and 11 median is 7 (6 + 8)/2. If data is nonsymmetrical in that case, mean and median could be different, and data might be skewed in one direction; thus arithmetic mean may not be a good criteria to measure central value. Mode (most frequent value) is another measure to calculate central tendency. Central tendency provides summary about the data but does not provide information about variation. Standard deviation or variance or square root $(Y_i\ -\ \mu)^2$ is used to measure variation or dispersion from the mean. It can be represented by two symbols: (i) σ^2 (sigma square for the population) and (ii) S^2 (sample). Population variance is defined as sum of squared deviations divided with total number, and it can be elaborated by the following equation if we intent to sample this population with replacement:

$$\sigma^2 = \frac{(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + (Y_3 - \mu)^2 + \ldots + (Y_N - \mu)^2}{N} \tag{3.7}$$

$$= \frac{\sum_i (Y_i - \mu)^2}{N} \tag{3.8}$$

However, when sampling is without replacement, then divisor is $N-1$, and it could be represented by the equation as follows:

$$S^2 = \frac{(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + (Y_3 - \mu)^2 + \ldots + (Y_N - \mu)^2}{N - 1} \tag{3.9}$$

$$= \frac{\sum_i (Y_i - \mu)^2}{N - 1} \tag{3.10}$$

The sample variance/mean square can be computed by using following formulas:

$$s^2 = \frac{\left(Y_1 - \overline{Y}\right)^2 + \left(Y_2 - \overline{Y}\right)^2 + \left(Y_3 - \overline{Y}\right)^2 + \ldots + \left(Y_N - \overline{Y}\right)^2}{n - 1} \tag{3.11}$$

$$s^2 = \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n - 1} \tag{3.12}$$

$$(n - 1)s^2 = \sum\nolimits_i \left(Y_i - \overline{Y}\right)^2 \tag{3.13}$$

$s^2 = SS$ (sum of squares). For example, for the numbers 3, 5, 7, and 9, the SS is

$$\begin{aligned}&(3 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 + (9 - 6)^2 = (-3)^2 + (-1)^2 + (1)^2 + (3)^2\\ &\quad = 9 + 1 + 1 + 9 = 20\end{aligned}$$

The variance for this data set will be 20/3 = 6.66, and the square root of the sample variance is called the standard deviation (s). For the above example, it can be calculated by the following method:

$$s = \sqrt{\frac{20}{3}} = 2.58$$

Thus Eq. (3.12) can be represented as follows:

$$SS = \sum\nolimits_i \left(Y_i - \overline{Y}\right)^2 \tag{3.14}$$

This Eq. (3.14) could be further modified to a computing formula as follow:

$$\sum\nolimits_i \left(Y_i - \overline{Y}\right)^2 = \sum_i Y_i^2 - {\left(\sum\nolimits_i Y_i\right)^2}/{n} \tag{3.15}$$

Table 3.1 Data set for the validation of sum of squares equation

	Y_i	Y_i^2	$Y_i - \overline{Y}$	$\lvert Y_i - \overline{Y} \rvert$	$(Y_i - \overline{Y})^2$
	3	9	3–6 = −3	3	9
	5	25	5–6 = −1	1	1
	7	49	7–6 = 1	1	1
	9	81	9–6 = 3	3	9
$\sum_i$:	24	164	**0**	**8**	**20**
$\overline{Y}$	6				

The term $\left(\sum_i Y_i\right)^2/_n$ is called the correction factor (CF) or correction term or adjustment for the mean. The Eq. (3.15) could be easily validated by using following data set in the Table 3.1.

Thus, $SS = \sum_i \left(Y_i - \overline{Y}\right)^2 = 20$ and by the $\sum\limits_i Y_i^2 - \left(\sum_i Y_i\right)^2/_n = 164 - \frac{(24)^2}{4} =$ 20 (Table 3.1). Another term which is generally used is called degree of freedom (df) (number of values in the calculation that are free to vary), and it is equal to $n-1$. The absolute mean deviation or average deviation is calculated as:

$$\text{Average deviation or Absolute mean deviation} = \frac{\sum_i \left|Y_i - \overline{Y}\right|}{n} \tag{3.16}$$

The absolute mean deviation or average deviation for the values 3, 5, 7, and 9 is 2 as vertical bars tell us consider all deviations as positive. The variance of the population $\left(\sigma^2{}_{\overline{Y}}\right)$ of $\overline{Y}$ can be calculated by the following equation:

$$\sigma^2{}_{\overline{Y}} = \frac{\sigma^2}{n} \tag{3.17}$$

However, $\sigma_{\overline{Y}}$ for the population can be computed by the following expression:

$$\sigma_{\overline{Y}} = \sqrt{\frac{\sigma^2}{n}} \tag{3.18}$$

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} \tag{3.19}$$

Standard deviation of sample mean is called standard error (SE). Variance for the sample $(s^2{}_{\overline{Y}})$ can be calculated by the following equations:

$$s^2{}_{\overline{Y}} = \frac{s^2}{n} \tag{3.20}$$

$$\mathrm{SE}_{\overline{Y}} = \sqrt{\frac{s^2}{n}} \tag{3.21}$$

Table 3.2 Example of the data set for the calculation of above concepts

Number of observations = i	Yield (kg ha^{-1}) = Y_i	$\overline{Y}$= Mean	$Y_i - \overline{Y}$	
1	1500	1536		−36
2	1850	1536	314	
3	1300	1536		−236
4	1730	1536	194	
5	1300	1536		−236
Total	7680		508	−508

$$\mathrm{SE}_{\overline{Y}} = \frac{s}{\sqrt{n}} \tag{3.22}$$

SE can be calculated by using following equation for the numbers 3, 5, 7, and 9 as used above to calculate standard deviation.

$$\mathrm{SE} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{6.66}{4}} = \sqrt{1.66} = 1.29$$

Variation can also be measured using coefficient of variability (CV) or relative standard deviation (RSD) which is widely used as a well-known indicator as described in Table 3.2. It is a measure of relative variability. It is the ratio of standard deviation (σ) to the mean (μ) and can be calculated by the following expression:

$$\text{coefficient of variation (CV)} = \frac{\sigma}{\mu} \tag{3.23}$$

$$\overline{Y} = \frac{\sum Y_i}{5} = \frac{7680}{5} = 1536\,\mathrm{kg\ ha^{-1}}$$

$$s^2 = \frac{\sum {Y_i}^2 - \left(\sum Y_i\right)^2/5}{4} = \frac{12,045,400 - (7680)^2/5}{4} = 62,230$$

$$s = \sqrt{62,230} = 249.45\,\mathrm{kg\ ha^{-1}}$$

$${s^2}_{\overline{Y}} = \frac{s^2}{5} = \frac{62,230}{5} = 12,446$$

$$\mathrm{SE}_{\overline{Y}} = \sqrt{\frac{s^2}{5}} = \sqrt{\frac{62,230}{5} = 12,446} = 111.56\,\mathrm{kg\ ha^{-1}}$$

$$\mathrm{CV} = \frac{249.45}{1536} \times 100 = 16\%$$

3.2 Statistical Models

A model is an abstract representation of a system in a quantitative way. It is a way of describing a real system in mathematical functions or diagrams. It can also be used to represent the simplification in different process trying to represent biological systems. A model can summarize factors affecting different process in a system. Mathematical models use different notation and expressions from mathematics to describe process, while statistical model is a mathematical model that allows variability in the process. This variability might be due to the number of reasons such as sampling, biological, and inaccuracies in measurements or due to the influential variables being omitted from the model. Thus, statistical models have potential to measure uncertainty associated with it. Statistical models come in the category of empirical models where principle of correlation was used to build a simple equation to describe relationship with different explanatory variables. Furthermore, if the explanatory variables are in numbers (quantitative), they were referred as variates, while if they are qualitative, then they were considered as factors and distinct groups as factor levels. For example, qualitative trait height can be classified as short, medium, or tall. Linear models are most importantly used statistical model.

3.3 The Linear Additive Model

Natural phenomenon in science such as earth rotation could be explained by the models. Linear additive model (LAM) is a commonly used model to describe the observation which has mean and error. Assumption for the application of this model includes that population of Y should be selected at random as well as errors are at random. This model could be used to make inferences about population means and variance. The simple LAM could be represented by the following equation:

$$Y_i = \mu + \varepsilon_i$$

where $\mu =$ mean and $\varepsilon_i =$ sampling error.

The sampling error for the population having mean zero could be calculated by the following procedure in which sample from the population is drawn in a random manner. The steps include

$$\overline{Y} = \frac{\sum_i Y_i}{n} = \frac{\sum_i (\mu + \varepsilon_i)}{n} = \mu + \frac{\sum_i \varepsilon_i}{n}$$

For random sampling the equation will be $= \frac{\left(\sum_i \varepsilon_i\right)}{n}$, and it is expected to be smaller as sample size increases and positive and negative epsilon will cancel. Generally variance of mean of large samples are usually small. Epsilon could be calculated by using $\left(Y_i - \overline{Y}\right)$.

3.4 Probability

Probability is a numerical description of how likely an event is to occur or how likely it is that a proposition is true. Probability is a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. The best example for understanding probability is flipping a coin: There are two possible outcomes—heads (H) or tails (T). What's the probability of the coin landing on heads? We can find out using the equation

$$\text{probability of head } P_H = \frac{1}{2}$$

or

$$\text{Probability of an event} = \frac{\text{number of ways it can happen}}{\text{total number of outcomes}}$$

Similarly, in case of dice rolling, there are six different outcomes (1, 2, 3, 4, 5, and 6), and probability of getting a one will be:

$$P_1 = \frac{1}{6}$$

The probability of getting 1 or 6 can be calculated by following way:

$$P_{1 \text{ or } 6} = \frac{2}{6} = \frac{1}{3}$$

The probability of rolling an even number (2, 4, and 6) will be:

$$P_{2,4 \text{ or } 6} = \frac{3}{6} = \frac{1}{2}$$

For many experiments there are only two possible outcomes, for example, a tossed coin falls heads or tails or student fail or pass or plant could be tall or short. Such outcomes are referred as binomial, and sample space will consist of two points only. Thus, sample space is made up of sample points (represented with E and, if event does not occur, represented with $-E$ or $\bar{E}$ or $\not{E}$) as shown in the following Fig. 3.1. Probability associated with each value of the random variable is called as

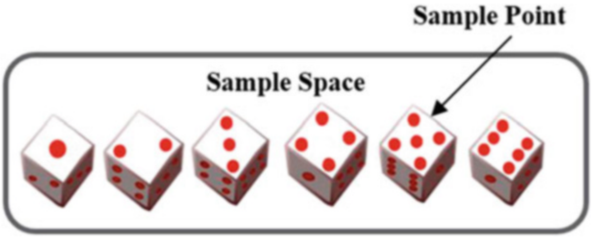


Fig. 3.1 Illustration of sample space and sample point

binomial probability function or binomial distribution. Formula that can gives the probability associated with each chance event e.g. for a fair coin if we consider $Y = 0$ for tail and $Y = 1$ for head will be:

$$P_{Y=Y_i} = \frac{1}{2} \ Y_i = 0 \text{ and } 1$$

For tossing a fair dice, probability distribution would be:

$$P_{Y=Y_i} = \frac{1}{6} \ Y_i = 1, 2, 3, 4, 5 \text{ and } 6$$

Ten thousand random digit tables are a very large sample for a population, and probability distribution for this table would be

$$P_{Y=Y_i} = \frac{1}{10} \ Y_i = 0, 1, 2, 3, 4, 5 \ldots 9$$

If we consider only odd and even numbers, then we can relate ten thousand random digit tables with $P_{Y=Y_i} = \frac{1}{2} \ Y_i = 0$ and 1, $P_{Y=Y_i} = \frac{1}{6} Y_i = 1$, 2, 3, 4, 5 and 6 and $P_{Y=Y_i} = \frac{1}{10} Y_i = 0$, 1, 2, 3, 4, 5. . .9, but it would not be binomial now, it will be multinomial. Probabilities of binomial distribution in single statement can be elaborated by generating single equation. Consider an experiment that contains n independent trials. Let $P_E = P_1 = p$ then $P_{\text{⍰}E} = P_0 = 1 - p$ as we know that $p = \frac{\textit{number of successes}}{\textit{total number of events (Successes+Failures)}}$ and probability of an event (E_i) *l*ies between 0 and 1 $(0 \leq P_{E_i} \leq 1)$ and sum of the probabilities of events in a mutually exclusive set is 1 $\left(\sum_i P_{E_i} = 1 \right)$. Five tosses of coins could result in (0, 0, 1, 1, 0), that is, two tails followed by two heads and final tail. Since trial is independent, thus probability of this outcome can be found by multiplying probabilities in each stage, i.e., $(1-p)(1-p)pp(1-p) = p^2(1 - p)^3$. If $p = 0.5$ then $(0.5)^5 = 0.03125$ or 3 % . The random variable Y associates a unique value with each sample point, e.g., for sample vector (0, 0, 1, 1, 0), we have $Y = 2$, and there are possibilities of 10 sequences with $Y = 2$. Thus $Y = 2$ is $10p^2(1 - p)^3$. The equatin which can be used to calculate this value directly will be:

$$\binom{n}{Y} = \frac{n!}{Y!(n - Y)!}$$

where $n! = n$ factorial $= n(n-1)(n-2)$. . .0.1. Thus, for $Y = 2$, i.e., two 1 s in $n = 5$ trials, the equation would be:

$$\begin{pmatrix}5\\2\end{pmatrix} = \frac{5.4.3.2.1}{2.1.3.2.1} = 10$$

One formula which can be used to count sample points with the same Y and one that assigns probability to each sample point in the binomial probability distribution can be represented as:

$P(Y = Y_i|n) = \begin{pmatrix}n\\Y_i\end{pmatrix} p^{Y_i}(1-p)^{n-Y_i}$ (In this equation the probability that the random variable Y takes the particular value Y_i in a random experiment with n trials). For the coin above illustration, this equation will be:

$$P(Y = 2|5) = \begin{pmatrix}5\\2\end{pmatrix}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3$$

The mean and variance of a random variable with a binomial distribution could be calculated by using following equations:

$$\text{Mean} : \mu = np$$

$$\text{Variance} : \sigma^2 = np(1-p)$$

3.5 Normal Distribution

Normal distribution is the most important widely used probability distribution as it fits with many natural processes such as heights, blood pressure, IQ score, and measurement error. It is also called as bell curve or Gaussian distribution. It is a standard reference for probability-related problems. The normal distribution has two parameters, i.e., mean (μ) and standard deviation (σ) (Fig. 3.2). The characteristics of normal distributions are as follows: (i) X lies between $-\infty$ and ∞ ($-\infty \leq X \leq \infty$); (ii) symmetric; (iii) normal density function rule, $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$; (iv) 2/3 of the most cases lies with one σ of μ, i.e., $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6826$; and (iv) 95% of cases lies two σ of μ, i.e., $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$.

3.6 Comparison of Means

Statistical concepts are used everywhere in daily life, e.g., while purchasing honey bottle from market, it may be labelled as 500 g, but to confirm this claim, we need to take random sample from the population. We could report the probability of obtaining a sample at least this uncommon if true mean is 500 g. This can be the problem of hypothesis testing. In such cases testing is done by using Student's t-test or F-Test. If means are more than two, the analysis of variance (ANOVA) F-test is to

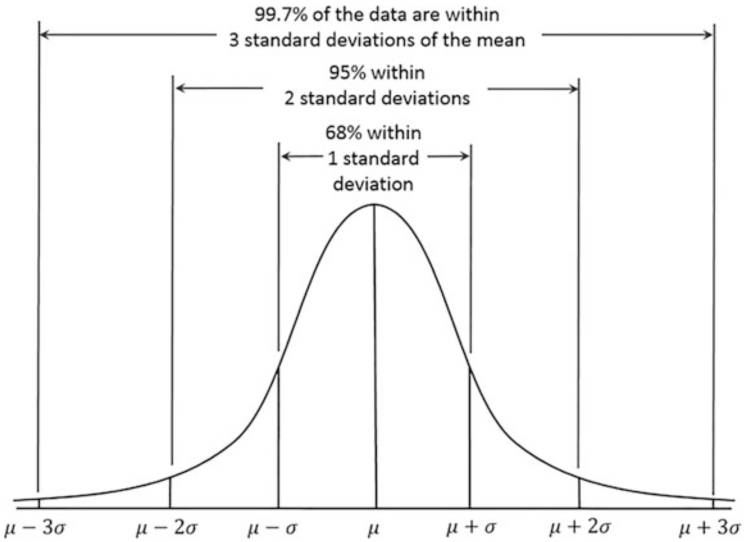


Fig. 3.2 Normal distribution curve

be used. Thus, sample size should be considered while selecting a test. Hypothesis test and confidence interval (CI) are interlinked. The formula to apply Student's *t*-test is

$$t = \frac{\overline{Y} - \mu}{S_{\overline{Y}}}$$

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

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

For the data having two means, *t*-test equation will be:

$$t = \frac{\overline{Y_1} - \overline{Y_2}}{S_{\overline{Y_1}} - S_{\overline{Y_2}}}$$

where $\overline{Y} = \text{sample mean}$, s is the sample standard deviation, and n is the sample size.

Consider a null hypothesis $H_o : \mu = \mu_o$ and alternative hypothesis $H_1 : \mu \neq \mu_o$, if t exceeds critical value $t_{0.025}$, then H_o is rejected, but if null hypothesis is true and still, it has been rejected and is called type I error. However, if H_1 is true and we accept H_o anyway, this type of error is called type II error.

3.7 Analysis of Variance (ANOVA)

It is an undeniable fact that agronomic research resulted to the improved quality of life and sustainability of the planet earth. The principles and procedures of analysis of variance (ANOVA) have been considered as fundamental tools in all agronomic research. ANOVA is an established statistical procedure that can be used to test the hypothesis by partitioning the sources of variation (SOV), variance components estimation, explanation and reduction of residual variation, and determination of the significance of effects. ANOVA history of application in agronomic field research and plant breeding trials goes back to the early twentieth century in which the main goal of research work was to have a better understanding of the effects of treatments, e.g., fertilizer, cultivars, planting dates, soil amendments, and their interactions. Earlier, trials main focus was on yield and thus to have better scientific understanding of the effects of treatments and guidance to the farmers; ANOVA was used widely. ANOVA helped in the early twentieth century to have good credibility of field agronomic trials. Furthermore, significant differences between treatment and check plots could be evaluated by ANOVA; however, there were issues between years as random effects of years could not be replicated (Loughin 2006). Fisher was a pioneer in the introduction of ANOVA, and he applied this concept in the 1920s on long-term wheat yield experiments (>half century) in response to the soil amendments (Fisher 1921). Fisher used ANOVA to disentangle large variability in average yield from other changes and evaluate significant difference between treatments. The basis of ANOVA was described as the variance (mean σ of variate from its mean thus square of its standard deviation) produced by all the causes at once in an operation is the sum of the values produced by each cause individually. Thus, with ANOVA we can partition the total variation into separate and independent SOV. To implement ANOVA accurately, it is important that treatment plots (experimental units) must be replicated and randomized. The basic assumptions to apply ANOVA are (i) Treatments and environment effects are additive and (ii) Experimental errors are random, independently and normally distributed about zero mean and with a common variance. Fisher in his experimental design work documented that the systematic arrangement of treatments resulted in the biased estimates of treatment averages, overestimation, and underestimation of error variation and correlated errors. Thus, replication is needed to estimate experimental error and randomization to have correct probability or level of significance. Generally, ANOVA divides total variation into two independent sources: (i) variation among treatments and (ii) variation within treatments (experimental error/residual error/error mean square/error variance). After considering that data is normally and independently distributed, F-ratio $\left(F = {}^{\text{variation between sample means}}\!/_{\text{variation within the samples}}\right)$ is used to test the null hypothesis that treatment means are equal or not. One-way ANOVA example could be best way to understand this ratio. Firstly, ANOVA was used for the fixed effect models (Model I, specific treatments or level of treatments of interest) but later used also for the random effect models (Model II). Afterwards it has been proposed that ANOVA should also be used for the mixed effect models (both fixed

Table 3.3 One-way analysis of variance with equal replication

SOV	df	Sum of squares (SS)	Mean squares (MS)	*F*
Treatments	$t-1$	$r\sum_i(\overline{X}_i.-\overline{X}..)^2=\sum_i\frac{X_i{}^2}{r}-\frac{X^2..}{rt}$	$\frac{\text{SS}_{\text{treatments}}}{\text{df}_{\text{treatments}}}$	$\frac{\text{MS}_{\text{treatments}}}{\text{MS}_{\text{error}}}$
Error	$t(r-1)$	$\sum_{i,j}(\overline{X}_{ij}-\overline{X}.)^2$	$\frac{\text{SS}_{\text{error}}}{\text{df}_{\text{error}}}$	
Total	$rt-1$	$\sum_{i,j}(\overline{X}_{ij}-\overline{X}..)^2=\sum_{i,j}X_{ij}{}^2-\frac{X^2..}{rt}$		

Table 3.4 Analysis of variance in randomized complete block

SOV	df	Sum of squares (SS)	Mean squares (MS)	*F*
Blocks	$r-1$	$t\sum_j(\overline{X}_{.j}-\overline{X}..)^2=\frac{\sum_j X^2{}_{.j}}{t}-C$	$\frac{\text{SS}_{\text{blocks}}}{\text{df}_{\text{blocks}}}$	
Treatments	$t-1$	$r\sum_i(\overline{X}_i.-\overline{X}..)^2=\sum_i\frac{X_i{}^2}{r}-C$	$\frac{\text{SS}_{\text{treatments}}}{\text{df}_{\text{treatments}}}$	$\frac{\text{MS}_{\text{treatments}}}{\text{MS}_{\text{error}}}$
Error	$(r-1)(t-1)$	$\sum_{i,j}(X_{ij}-\overline{X}_{.J}-\overline{X}_i.+\overline{X}..)^2$ $=\text{SS}_{\text{total}}-\text{SS}_{\text{blocks}}-\text{SS}_{\text{treatments}}$	$\frac{\text{SS}_{\text{error}}}{\text{df}_{\text{error}}}$	
Total	$rt-1$	$\sum_{i,j}(X_{ij}-\overline{X}..)^2=\sum_{i,j}X_{ij}{}^2-C$		

and random treatment factors) (Gbur et al. 2012; West and Galecki 2012). The importance of mixed effect models was shown in some of experiments where use of fixed model instead of mixed models resulted to the misleading results (Acutis et al. 2012; Bolker et al. 2009; Moore and Dixon 2015; Yang 2010). Fisher's ANOVA is the most frequently used method to determine if differences among means are significant or not. His preference was to declare significance when $P \leq 0.05$ (P value) by considering F table also. The components of ANOVA include sources of variations (SOV), degrees of freedom, sum of squares, mean squares, F values, and P values (Tables 3.3, 3.4 and 3.5). The ANOVA importance and applications in different earlier work have been presented in Table 3.6. Meantime as Fisher was working on his ANOVA framework, Neyman and Pearson presented the concept of type of errors (type I (true null hypothesis rejection) and type II errors (failing to reject false null hypothesis)) (McIntosh 2015).

3.7.1 Calculation of the *F*-Test

F-ratio calculation for one-way ANOVA is possible by using following equations and is reported in the representative Table 3.7.

Table 3.5 Analysis of variance for Latin square

SOV	df	Sum of squares (SS)	Mean squares (MS)	F
Rows	$r-1$	$r\sum_{i}\left(\overline{X}_{i}.-\overline{X}..\right)^{2}=\frac{\sum_{i}X^{2}{}_{i}.}{r}-C$	$\frac{\mathrm{SS}_{\mathrm{blocks}}}{\mathrm{df}_{\mathrm{blocks}}}$	
Columns	$r-1$	$r\sum_{j}\left(\overline{X}_{j}.-\overline{X}..\right)^{2}=\frac{\sum_{j}X^{2}{}_{j}.}{r}-C$		
Treatments	$r-1$	$r\sum_{t}\left(\overline{X}_{t}-\overline{X}..\right)^{2}=\sum_{t}\frac{X_{t}{}^{2}}{r}-C$	$\frac{\mathrm{SS}_{\mathrm{treatments}}}{\mathrm{df}_{\mathrm{treatments}}}$	$\frac{\mathrm{MS}_{\mathrm{treatments}}}{\mathrm{MS}_{\mathrm{error}}}$
Error	$(r-1)$ $(r-2)$	$\sum_{i,j}\left(X_{ij}-\overline{X}_{i}.-\overline{X}._{j}-\overline{X}\,\overline{X}..\right)^{2}$ $=\mathrm{SS}_{\mathrm{total}}-\mathrm{SS}_{\mathrm{blocks}}-\mathrm{SS}_{\mathrm{treatments}}$	$\frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{df}_{\mathrm{error}}}$	
Total	$rt-1$	$\sum_{i,j}\left(X_{ij}-\overline{X}..\right)^{2}=\sum_{i,j}X_{ij}{}^{2}-C$		

Table 3.6 ANOVA importance and applications in different earlier work

S. no	Applications	References
1.	Statistical guidelines for authors	Nature Publishing Group (2005) and (2013a, b)
2.	Raising of data analysis standards	McNutt (2014)
3.	Improvement in the accuracy of the statistical analyses	Acutis et al. (2012)
4.	ANOVA is a commonly used technique, but selection of factors as fixed or random can be complex	Bennington and Thayne (1994)
5.	Mixed model analysis	Yang (2010)
6.	Inclusion/exclusion of fixed by random effects in mixed model	Blouin et al. (2011)
7.	Analysis of combined experiments	McIntosh (1983)
8.	Combined experiment analysis	Moore and Dixon (2015)
9.	Choice of models	Lencina et al. (2005)
10.	Mixed models controversy	Nelder (2008)
11.	Accurate selection of analysis	Nelder and Lane (1995)
12.	Mixed models controversy	Voss (1999)
13.	Two-way factorial ANOVA with mixed effects and interactions	Wang and DeVogel (2019)
14.	ANOVA to show relationship between sources of variation (SOV) and terms in the general linear model (GLM)	Gomez and Gomez (1984)
15.	Explanation of statistical ideas	Mead (2017)
16.	Tests of significance	Snedecor (1942)
17.	Application of statistics principles and procedures	Steel and Torrie (1980)
18.	SAS application in experimental design and analysis	Lawson (2010)

Table 3.7 Representative table for *F*-test calculation

SOV	SS	Df	MS	*F*
Factor of interest (between groups)	$\mathrm{SS}_B = \sum \left(\overline{x_j} - \bar{x}\right)^2 = \sum_j n_j \left(x_j^{\,2}\right) = \frac{\left(\sum x_{ij}\right)^2}{n}$	$\mathrm{df}_B = j - 1$	$\mathrm{MS}_B = \frac{\mathrm{SS}_B}{\mathrm{df}_B}$	$F = \frac{\mathrm{MS}_B}{\mathrm{MS}_W}$
Error (within groups)	$\mathrm{SS}_W = \sum_j \sum_i \left(x_{ij} - \bar{x}_j\right)^2$	$\mathrm{df}_W = (n - 1) - (j - 1)$	$\mathrm{MS}_W = \frac{\mathrm{SS}_W}{\mathrm{df}_W}$	
Total	$\mathrm{SS}_T = \sum \left(x_{ij} - \bar{x}\right)^2 = \sum \left(x_{ij}^2\right) - \frac{\left(\sum x_{ij}\right)^2}{n}$	$\mathrm{df} = n - 1$		

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1}$$

where σ^2 = vraince, x_i = observation, $\overline{\mathrm{x}}$ = sample population mean, and n = obsevtaion number.

Sum of squares (SS) in ANOVA is sum of the squared deviations of observation from the mean. Total sum of squares ($\mathrm{SS_T}$) can be calculated by using following equation:

$$\mathrm{SS_T} = \sum \left(x_{ij} - \overline{x}\right)^2$$

where x_{ij} = ith observation in the jth group. The formulae can be rewritten as:

$$\mathrm{SS_T} = \sum \left(x_{ij} - \overline{x}\right)^2 = \sum \left(x_{ij}^2\right) - \frac{\left(\sum x_{ij}\right)^2}{n}$$

The total SS between group ($\mathrm{SS_B}$) and within group ($\mathrm{SS_w}$) can be calculated by using following equations:

$$\mathrm{SS_B} = \sum \left(\overline{x_j} - \overline{x}\right)^2 = \sum_j n_j \left({x_j}^2\right) = \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\mathrm{SS_W} = \sum_j \sum_i \left(x_{ij} - \overline{x}_j\right)^2$$

Total SS in the model can be calculated by following equation which can be further used to get $\mathrm{SS_w}$:

$$\mathrm{SS_T} = \mathrm{SS_B} + \mathrm{SS_W}$$

$$\mathrm{SS_W} = \mathrm{SS_{TT}} + \mathrm{SS_B}$$

The mean square (MS) (mean of entire sample population or average squared deviation of observation from grand mean) is calculated next which is sum of squares ($\mathrm{SS_T}$) by the total number of *degrees of freedom* (df) or n–1. The mean square between groups ($\mathrm{MS_B}$) can be calculated by using following equation:

$$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{df_B}}$$

Finally, R ratio is calculated by using following equation:

$$F = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$$

3.8 Experimental Design and Its Principles

New knowledge can be easily obtained by careful planning, analysis, and interpretation of data. Designing of an efficient experiment needs consultation with statistician as they can help to have appropriate design which can enable researchers to have unbiased estimates of treatment means and experimental error. An experiment is planned inquiry to obtain new facts or to confirm earlier findings. Experiments are generally designed to answer the questions or test the hypothesis. Before designing an experiment, it is important that objectives of the experiment should be clear. The unit of material or place where one application of treatment is applied is called experimental unit or experimental plot. Variation is the characteristics of all experimental material and experimental error is used to measure the variation among experimental unit. Variation could be due to number of reasons. It can be due to inherent variability or lack of uniformity in the physical conduction of experiment. Replication is another important component of experimental design. The main functions of replication are to (i) estimate experimental error, (ii) improve precision of the experiment by minimizing standard deviation of treatments, (iii) control error variance, and (iv) increase the scope of inference of the experiment. Error in the experiments could be controlled by the selection of appropriate experimental design, use of parallel observations, and choice of size and shape of the experimental units. Furthermore, unbiased estimate of experimental error is possible by the application of randomization.

3.8.1 Completely Randomized Design (CRD)

Completely randomized design is used when experimental units are homogeneous and less to be gained by putting them into blocks due to similarity of response. For example, variety trial in greenhouse will be subjected to CRD because of uniformity of soil. Similarly, laboratory experiments where it's easy to control variability and experimental units are homogenous; CRD is used. The advantages of CRD are as follows: number of replicates can vary from treatment to treatment, and loss of information due to missing data is small. The precision of experiment is high due to maximum degree of freedom (df) for estimating experimental error. In this design treatments are assigned at random so that each experimental unit receives same chance of getting treatment. The randomization procedure and layout for the pot experiment having four treatments (A, B, C, and D) replicated four times have following steps:

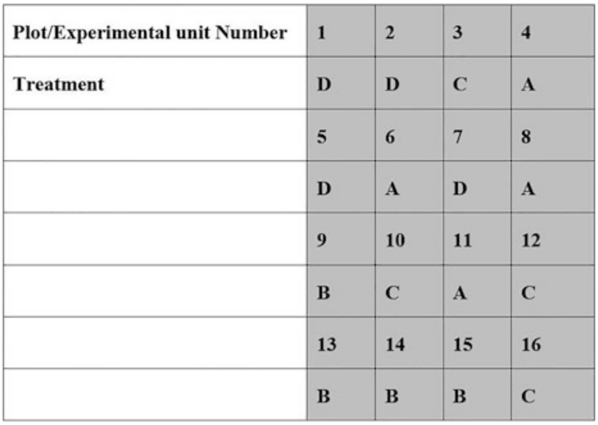
1. Determination of total number of plots or experimental unit (n): Determine the total number of plots or experimental unit by multiplying treatments (t) with the number of replications (R); $n = Rt = 4 \times 4 = 16$. However, if replications are not the same, then "n" can be calculated by getting sum of the replications of each treatment.
2. Assigning of plot number

Table 3.8 Random ranking of experimental unit

Random number	Experimental unit	Ranking	Treatments
0.07	1	4	A
0.842	2	15	B
0.502	3	10	C
0.174	4	5	D
0.426	5	8	A
0.699	6	14	B
0.926	7	16	C
0.039	8	2	D
0.244	9	6	A
0.663	10	13	B
0.045	11	3	C
0.305	12	7	D
0.503	13	11	A
0.429	14	9	B
0.583	15	12	C
0.025	16	1	D

Table 3.9 Group numbers based on random numbers ranking

Treatments	Group number	Ranks in the group			
A	1	4	8	6	11
B	2	15	14	13	9
C	3	10	16	3	12
D	4	5	2	7	1



Plot/Experimental unit Number	1	2	3	4
Treatment	D	D	C	A
	5	6	7	8
	D	A	D	A
	9	10	11	12
	B	C	A	C
	13	14	15	16
	B	B	B	C

Fig. 3.3 A layout of completely randomized design with four treatments (A, B, C, and D) replicated four times

3. Assigning of treatments into plots using random number method and further its ranking as shown in the Table 3.8. Afterwards group number assigned based on random number ranking (Table 3.9) and treatments was placed in the experimental units as shown in the layout (Fig. 3.3).

In order to have ANOVA for the treatments mentioned in Table 3.10, we need to obtain $X_i.$ and $\sum_j X^2 ij$ as mentioned in Table 3.10 (points 1 and 2). Afterwards each treatment total is squared and divided by $r = 5$ to get $(X_{i\cdot})^2/_r$ named as treatments sum of square. Correction factor (CF) is calculated afterwards by dividing total sum of squares of all observations with total numbers (rt). The equation to calculate CF is:

$$\mathrm{CF} = \frac{X^2..}{rt} = \frac{\left(\sum_{i,j} X_{ij}\right)^2}{rt} = \frac{(670.6)^2}{(5)(6)} = 14,990.15$$

$$\mathrm{SS\ (total)} = \sum_{i,j} X^2 ij - \mathrm{CF} = 16,093.56 - 14,990.15 = 1103.41$$

$$\begin{aligned}\mathrm{SS\,(treatment)\,(between\ \ or\ \ among\ \ groups)} &= \frac{X^2 1. + \cdots + X^2 t.}{r} - \mathrm{CF}\\ &= \frac{(148.1)^2 + (132.8)^2 + \cdots + (100.9)^2}{5}\\ &= \frac{7,788,008.00}{5} - 14,990.15\\ &= 15,576.02 - 14,990.15\\ &= 585.87\end{aligned}$$

The sum of squares (SS) among individuals is called within group SS, residual SS, error SS, or discrepancy SS, and it can be obtained by following equation:

$$\begin{aligned}\mathrm{SS}_{\mathrm{error}} &= \mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Treatment}}\\ &= 1103.41 - 585.87\\ &= 517.54\end{aligned}$$

The error SS ($\mathrm{SS}_{\mathrm{error}}$) can also be calculated by pooling the within treatments SS as shown below:

$$\begin{aligned}\mathrm{SS}_{\mathrm{error}} &= \sum_i \left(\sum_j X^2 ij - \frac{X^2 i.}{r}\right)\\ &= \left(4593.45 - \frac{148.1^2}{5}\right) + \left(3623.34 - \frac{132.8^2}{5}\right) + \left(1980.28 - \frac{95.8^2}{5}\right)\\ &\quad + \left(2406.37 - \frac{109.3^2}{5}\right) + \left(1435.61 - \frac{83.7^2}{5}\right) + \left(2054.51 - \frac{100.9^2}{5}\right)\\ &= 517.54\end{aligned}$$

Table 3.10 Nitrogen contents of Lucerne plants inoculated with *Rhizobium trifolii* strains (RTS) (mg)

Calculation	RTS1	RTS2	RTS3	RTS4	RTS5	Composite	Total
1. $\sum_j X_{ij} = X_i.$	17.4	18.1	19.1	21.2	15.3	19.3	
	33.7	27.8	22.4	24	17.4	22.4	
	29.0	30.9	12.1	23.5	14.8	22.1	
	33.0	28.2	14.9	21.8	14.6	19.9	
	35.0	27.8	27.3	18.8	21.6	17.2	
	148.1	132.8	95.8	109.3	83.7	100.9	$670.6 = X..$
2. $\sum_j X^2 ij$	4593.45	3623.34	1980.28	2406.37	1435.61	2054.51	16093.56
3. $(X_i.)^2/_r$	4386.72	3527.17	1835.53	2389.30	1401.14	2036.16	15,576.02
4. $\sum_j (X_{ij} - \overline{X}_i.)^2$	206.73	96.17	144.75	17.07	34.47	18.35	517.54
5. $\overline{X}_i.$	29.62	26.56	19.16	21.86	16.74	20.18	$22.4 =$ mean

Note: $X_{ij} = j$th observation on the ith treatment ($i = 1, 2, 3 \ldots .t$ and $j = 1, 2, \ldots .. r.$), $X_i =$ individual observation, and $X_i. =$ sum of all observation for the ith treatment

Table 3.11 Analysis of variance for data of Table 3.10

SOV	df	SS	Mean squares (MS)	$F_{\text{calulated}}$	$F_{\text{tablulated}}$
Among treatments	6–1 = 5	585.87	$\frac{585.87}{5}$ $= 117.17$	$\frac{117.17}{21.56} = 5.43^{**}$ Since $F_{\text{cal}} > F_{\text{tab}}$ at 0.05 and 0.01 thus there are highly significant (∗∗) differences among treatments	2.62 (0.05) 3.90 (0.01)
Error	6 (5–1) = 24	517.54	$\frac{517.54}{24}$ $= 21.56$		
Total	(5)(6)–1 = 29	1103.41			

These generated numerical results are presented in an AONVA (Table 3.11), and it shows that there is significant difference among treatments. The standard error of treatment mean ($\text{SE}_{\overline{X}}$) and differences between treatment, CV, and least significance difference (LSD) are calculated by using the following equations:

$$\text{SE}_{\overline{X}} = \sqrt{\frac{s^2}{r}} = \sqrt{\frac{21.56}{5}\text{mg}} = \sqrt{4.312} = 2.07\ \text{mg}$$

$$SE_{\overline{X}_{i}.-\overline{X}_{i}.\cdot.} = \sqrt{\frac{2s^2}{r}} = \sqrt{\frac{2(21.56)}{5}} = \sqrt{\frac{43.12}{5}} = \sqrt{8.62} = 2.93\ \text{mg}$$

$$\text{CV (Coefficient of variability)} = \frac{\sqrt{S^2}}{\overline{X}} \times 100 = \frac{\sqrt{21.56}}{22.4} \times 100 = \frac{4.64}{22.4} \times 100$$
$$= 20.7\%$$

$$\text{LSD} = t_{\alpha/2} S_{\overline{X}i.-\overline{X}i.\cdot.} = t_{\alpha/2} S\sqrt{\frac{2}{r}}\ (\text{for equal } r)$$

$$\text{LSD}_{0.05} = t_{0.025} S_{\overline{X}i.-\overline{X}i.\cdot.} = 2.064\sqrt{\frac{2(21.56)}{5}} = 2.064\sqrt{8.62} = 2.064 \times 2.93$$
$$= 6.06\ \text{mg}$$

$$\text{LSD}_{0.01} = t_{0.005} S_{\overline{X}i.-\overline{X}i.\cdot.} = 2.797\sqrt{\frac{2(21.56)}{5}} = 8.21\ \text{mg}$$

The observed differences are $\overline{X}1. - \overline{X}2. = 29.62–26.56 = 3.06$; $\overline{X}3. - \overline{X}4. = 19.16–21.86 = -2.7$; and $\overline{X}5. - \overline{X}6. = 16.74–20.18 = -3.44$. Now rank the means from the smallest to largest as shown below:

RTS1	RTS2	RTS3	RTS4	RTS5	Composite
29.62 (6)	26.56 (5)	19.16 (2)	21.86 (4)	16.74 (1)	20.18 (3)

Next is to calculate the difference and test significance level using LSD test at 5%.

6–1 = 29.62–16.74 = 12.88 > 6.06 = significant
6–2 = 29.62–19.16 = 10.46 > 6.06 = significant
6–3 = 29.62–20.18 = 9.44 > 6.06 = significant
6–4 = 29.62–21.86 = 7.76 > 6.06 = significant
6–5 = 29.62–26.56 = 3.06 < 6.06 = nonsignificant
5–1 = 26.56–16.74 = 9.82 > 6.06 = significant
5–2 = 26.56–19.16 = 7.4 > 6.06 = significant
5–3 = 26.56–20.18 = 6.38 > 6.06 = significant
5–4 = 26.56–21.86 = 4.70 < 6.06 = nonsignificant
4–1 = 21.86–16.74 = 5.12 < 6.06 = nonsignificant
4–2 = 21.86–19.16 = 2.70 < 6.06 = nonsignificant
4–3 = 21.86–20.18 = 1.68 < 6.06 = nonsignificant
3–1 = 20.18–16.74 = 3.44 < 6.06 = nonsignificant
3–2 = 20.18–19.16 = 1.02 < 6.06 = nonsignificant
2–1 = 19.16–16.74 = 2.42 < 6.06 = nonsignificant

3.8.2 Randomized Complete Block Design (RCBD)

The randomized complete block design (RCBD) is one of the most widely used designs in an agronomic field research. In this design experimental unit can be meaningfully grouped, and number of units in a group is equal to the number of treatments. These groups are called block or replication. The objective to have groups in blocks is to minimize error and ensure that observed differences will be due to treatments only. The RCBD has more advantages than the CRD due to blocking and further randomization which results to the more precision. The main purpose of blocking is to have higher accuracy by minimizing the experimental error due to the known sources of variation (SOV) among the experimental units. Grouping is done in such a way that variability within each block is minimized, while among block it is maximized. Variation within a block will be part of the experimental error; thus blocking is most effective when experimental area has a predictable pattern of variability. An ideal known SOV which can be used as basis for the blocking includes soil heterogeneity in nitrogen fertilizer experiments or varietal trials at multiple sites or sowing date experiments.

Thus, basis of blocking depends on the main SOV. The size and shape of blocks are selected in such a way so that there should be maximum variability among blocks. To do blocking, firstly, identify the gradient and do blocking vertical to the gradients, and if gradient occurs in two directions (one strong and other weak), then consider that gradient which is stronger, e.g., in case of fertility gradient. If fertility gradient is strong on both sides and perpendicular to each other, then use square blocks and choose Latin square design as elaborated by Gomez and Gomez (1980). Furthermore, whenever blocking is done, blocks identity and purpose should be clear. Similarly, if SOV is beyond the control, then ensure that such variation occurs among blocks as compared to within blocks. For example, in case of application of herbicides or data collection which might not be possible to complete in one day. In such scenario, it is recommended that it should be completed firstly for all plots of the same block. In this way, variation due to collection of data by multiple observers or application of treatments in more than one day becomes part of block variation and excluded from the experimental error. Following steps should be followed to design layout for RCBD.

1. Division of experimental area into "*R*" equal blocks (*R* = replications). The experimental area is divided into four blocks as shown in Fig. 3.4.
2. Subdivision of blocks into experimental plots based on number of treatments. For example, here if we suppose there are six treatments, i.e. A, B, C, D, E, and F, then divide each block into six subplots and assign each treatment into subplot using the random numbers (Fig. 3.5).
3. Repetition of step 2 for the remaining blocks (Fig. 3.6).

Let's apply the concept of RCBD on the data provided in Table 3.12 to generate ANOVA table and see significant difference among different oil contents of different canola cultivars. Step 1 includes arranging of raw data in ways as shown in Table 3.4. Calculate $\sum X^2$ and treatment ($X_{i.}$) and blocks ($X_{.j}$) totals, i. e. , $\sum_j X^2{}_{ij}$; $i = 1, 2...t$, and $\sum_i X^2{}_{ij}; j = 1, 2...r$. Step 2 is to calculate sum of squares using following formulas:

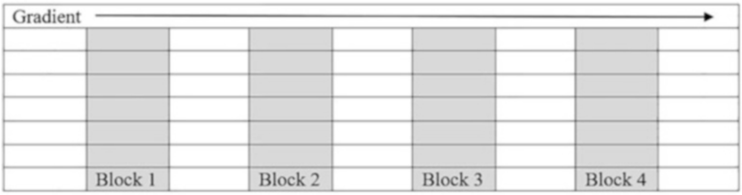
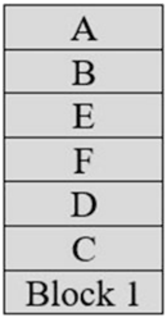
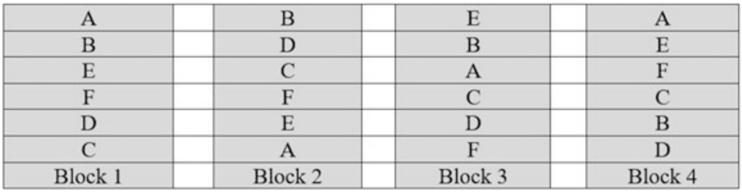


Fig. 3.4 Layout for the RCBD (division of experimental area into four blocks)



A
B
E
F
D
C
Block 1

Fig. 3.5 Subdivision of blocks into experimental plots based on number of treatments and randomization of treatments (A, B, C, D, E, and F)



A		B		E		A
B		D		B		E
E		C		A		F
F		F		C		C
D		E		D		B
C		A		F		D
Block 1		Block 2		Block 3		Block 4

Fig. 3.6 A randomized layout for the RCBD (six treatments and four replications)

$$\text{Correction factor} = \text{CF} = \frac{Y^2_{..}}{rt} = \frac{(1085.5)^2}{24} = \frac{(1085.5)^2}{24} = \frac{1,178,310.25}{24}$$
$$= 49,096.26$$

$$\text{SS}_{\text{total}} = \sum_{i,j} X^2 ij - \text{CF}$$

$$\text{SS}_{\text{total}} = 49,150.77 - 49,096.26 = 54.51$$

$$\text{SS}_{\text{block}} = \frac{\sum_j Y^2_{.j}}{t} - \text{CF}$$

$$\text{SS}_{\text{block}} = \frac{(269.8)^2 + (268.8)^2 + (274.2)^2 + (272.7)^2}{6} - 49,096.26$$

Table 3.12 Oil content (%) data of different canola cultivars with analysis of variance table

Cultivars		**Block**				**Treatments totals**		
		1	2	3	4	$Y_{i.}$	$\sum_j X^2{}_{ij}$	$\overline{Y}_i$
Can1		44.1	45.6	45.7	43.8	179.2	8031.1	44.8
Can2		43.0	41.6	44.6	46.8	176.0	7759.0	44.0
Can3		46.5	46.3	46.7	46.1	185.6	8612.0	46.4
Can4		44.1	43.7	44.2	42.8	174.8	7640.0	43.7
Can5		46.0	44.6	45.6	46.8	183.0	8374.8	45.8
Can6		46.1	47.0	47.4	46.4	186.9	8733.9	46.7
Block totals	$X_{.j}$	269.8	268.8	274.2	272.7	1085.5		
	$\sum_i X^2{}_{ij}$	12,142.08	12,061.46	12,538.3	12,408.93		49,150.77	

Analysis of variance (ANOVA)

SOV	**df**	**Sum of squares (SS)**	**Mean squares (MS)**	***F***
Blocks	$r-1 = 4-1 = 3$	3.14	1.05	
Treatments	$t-1 = 6-1 = 5$	31.65	6.33	4.83**
Error	$(r-1)(t-1) = 15$	19.72	1.31	
Total	$rt-1 = 24-1 = 23$	54.51		

** $P < 0.05$

$$\mathrm{SS}_{\mathrm{block}} = 49,099.4 - 49,096.26 = 3.14$$

$$\mathrm{SS}_{\mathrm{treatment}} = \frac{\sum_{i} Y^{2}{}_{i.}}{r} - \mathrm{CF}$$

$$\mathrm{SS}_{\mathrm{treatment}} = \frac{(179.2)^{2} + (176.0)^{2} + (185.6)^{2} + (174.8)^{2} + (183.0)^{2} + (186.9)^{2}}{4} - 49,096.26$$

$$\mathrm{SS}_{\mathrm{treatment}} = \frac{196,511.70}{4} - 49,096.26$$

$$\mathrm{SS}_{\mathrm{treatment}} = 49,127.91 - 49,096.26 = 31.65$$

$$\mathrm{SS}_{\mathrm{error}} = \mathrm{SS}_{\mathrm{total}} - \mathrm{SS}_{\mathrm{block}} - \mathrm{SS}_{\mathrm{treatment}}$$

$$\mathrm{SS}_{\mathrm{error}} = 54.51 - 3.14 - 31.65 = 19.72$$

3.8.3 Missing Values Estimation

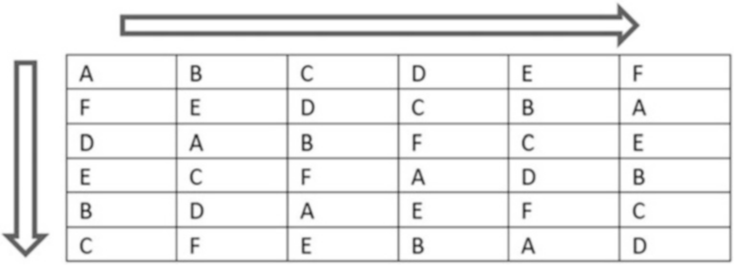
Sometimes due to poor germination or due to climatic conditions, etc., data might be missing from the experimental unit. This missing data can be calculated by using following equation:

$$y = \frac{rB_o + tT_o - G_o}{(r-1)(t-1)}$$

where y = missing value estimation; t = number of treatments; r = number of replications; B_o = replication total that contains missing value; T_o = treatments total that contains missing value; and G_o = total of all observed values.

3.8.4 Latin Square Design

Treatments are arranged in rows and columns in Latin square design. Treatments (t) are repeated "t" times in such a way that t appear exactly one time in each column and row and denoted by Roman characters, thus called as Latin square design. The main purpose of this design is to reduce systematic error due to columns and rows (treatments) ($n \times$ n). The advantage in the use of this design is in the field experiment where two major SOVs exist, e.g., in case of soil difference in two directions, this design will help to remove variation. The disadvantage of this design is that number of rows, columns, and treatments should be equal. Latin square design for six treatments, i.e., A, B, C, D, E, and F, will be like as shown in Fig. 3.7. Analysis of



A	B	C	D	E	F
F	E	D	C	B	A
D	A	B	F	C	E
E	C	F	A	D	B
B	D	A	E	F	C
C	F	E	B	A	D

Fig. 3.7 Layout for Latin square design

variance for an $r \times r$ (6 × 6) Latin square data set oil yield (kg ha^{-1}) of canola cultivars is given in Table 3.13. The calculation involves following steps:

1. Calculation of row totals ($X_i.$), column totals ($X._j$), treatment totals (X_t), and grand total ($Y..$). Similarly, calculate $\sum_j X^2{}_{ij}$ and $\sum_i X^2{}_{ij}$ for each value of rows and columns (Table 3.13).
2. Calculation of correction factor and sum of squares (SS):

$$\mathrm{CF} = \frac{X^2..}{r^2} = \frac{(40,380)^2}{6^2} = 452,92,900$$

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i,j} X^2{}_{ij} - \mathrm{CF} = 459,82,806 - 452,92,900 = 689,906$$

$$\begin{aligned}\mathrm{SS}_{\mathrm{row}} &= \frac{\sum_i X^2{}_{i.}}{r} - \mathrm{CF} \\ &= \frac{(6669)^2 + (6732)^2 + (6781)^2 + (6757)^2 + (6718)^2 + (6723)^2}{6} \\ &- 452,92,900 = 452,94,108 - 45,292,900 = 1208\end{aligned}$$

$$\begin{aligned}\mathrm{SS}_{\mathrm{column}} &= \frac{\sum_j X^2{}_{.j}}{r} - \mathrm{CF} \\ &= \frac{(6592)^2 + (6839)^2 + (6750)^2 + (6749)^2 + (6680)^2 + (6770)^2}{6} \\ &- 452,92,900 = 452,98,864 - 452,92,900 = 5964\end{aligned}$$

Table 3.13 Oil yield (kg ha^{-1}) of different canola cultivars with analysis of variance table under Latin square design

Rows	**Columns**						**Row totals**	
	1	2	3	4	5	6	$X_{i.}$	$\sum_j X^2 ij$
1	1329 (A)	950 (B)	980 (C)	1130 (D)	1060 (E)	1220 (F)	6669	7,518,041
2	1237 (F)	1070 (E)	1150 (D)	990 (C)	965 (B)	1320 (A)	6732	7,651,294
3	1126 (D)	1380 (A)	970 (B)	1240 (F)	985 (C)	1080 (E)	6781	7,787,401
4	1022 (E)	990 (C)	1250 (F)	1370 (A)	1140 (D)	985 (B)	6757	7,733,809
5	923 (B)	1170 (D)	1350 (A)	1040 (E)	1230 (F)	1005 (C)	6718	7,647,854
6	955 (C)	1279 (F)	1050 (E)	979 (B)	1300 (A)	1160 (D)	6723	7,644,407
Column totals $X_{.j}$ $\sum_i X^2 ij$	6592	6839	6750	6749	6680	6770	$\sum_{i,j} X_{ij}$ 40,380	$\sum_{i,j} X^2 ij$
	7,372,724	7,936,641	7,711,300	7,711,541	7,527,550	7,723,050		45,982,806
Cultivar totals and means								
	A	**B**	**C**	**D**	**E**	**F**		
Total = X_t	8049	5772	5905	6876	6322	7456		
Mean = $\overline{X}_t$	1341.5	962	984.1667	1146	1053.667	1242.667		
Analysis of variance								
SOV	**df**	**SS**	**MS**	***F***				
Rows	$r-1=5$	1208	241.6	0.66				
Columns	$r-1=5$	5964	1192.8	3.29				
Cultivars	$r-1=5$	675,501	135,100.2	373.61** (highly significant difference among cultivars for oil yield since *F* calculated is greater than *F* tabulated at 1 and 5%)				
Error	$(r-1)(r-2)=20$	7233	361.6					
Total	$r^2-1=35$	689,906	19,711.6					

$$\begin{aligned}\mathrm{SS}_{\text{treatment}} &= \frac{\sum_t X^2{}_t}{r} - \mathrm{CF}\\ &= \frac{(8049)^2 + (5772)^2 + (5905)^2 + (6876)^2 + (6322)^2 + (7456)^2}{6}\\ &- 452,92,900 = 459,68,401 - 452,92,900 = 675,501\end{aligned}$$

$$\begin{aligned}\mathrm{SS}_{\text{error}} &= \mathrm{SS}_{\text{total}} - \mathrm{SS}_{\text{row}} - \mathrm{SS}_{\text{column}} - \mathrm{SS}_{\text{treatment}}\\ &= 689,906 - 1208 - 5964 - 675,501 = 7233\end{aligned}$$

$$\text{Standard error of treatment means} = S_{\overline{X}} = \sqrt{\frac{S^2}{r}} = \sqrt{\frac{361.6}{6}} = 7.76 \text{ kg}$$

$$\text{Sample standard error of difference between two treatment means} = S_{\overline{X}i-\overline{X}it}$$

$$= \sqrt{\frac{2S^2}{r}} = \sqrt{\frac{2S^2}{r}} = 10.97 \text{ kg}$$

3.8.5 Factorial Experiments

Factorial experiments consist of number of factors as treatment with all possible combinations with different levels of equal importance. For example, an experiment involves temperature as treatment (factor) will have different levels of temperature. Similarly, if silicon (Si) fertilization is used as factor in pot experiment, several levels will be used to evaluate the experiment. For example, if we use two sources of Si (potassium silicate and sodium silicate) each at two different concentrations, it will be referred as a 2 × 2 or 2^2 factorial experiment. The possible combinations of two levels in each of the two factors will be four as shown in Table 3.14. Similarly, if Si fertilization experiment is conducted by using only potassium silicate with its two levels (no application as Si_0 and 200 mg L^{-1} of potassium silicate as Si_{200}) under

Table 3.14 2 × 2 or 2^2 factorial treatment combinations

Treatment combinations		
Treatment number	**Source (factor A)**	**Concentrations (factor B)**
1	Potassium silicate	100 mg L^{-1}
2	Potassium silicate	200 mg L^{-1}
3	Sodium silicate	100 mg L^{-1}
4	Sodium silicate	200 mg L^{-1}
Treatment number	**Water regimes**	**Concentrations**
1	W+	Si_o
2	W−	Si_{200}
3	W+	Si_o
4	W−	Si_{200}

Table 3.15 Symbolic representation of 3 × 3 or 3^2 factorial treatment combinations

Factors	A			
B	Levels	a_0	a_1	a_2
	b_0	a_0b_0	a_1b_0	a_2b_0
	b_1	a_0b_1	a_1b_1	a_2b_1
	b_2	a_0b_2	a_1b_2	a_2b_2

Table 3.16 Shoot dry weight (g) of sorghum plant under different silicon source as factor A and silicon concentration as factor B to illustrate simple effects, main effects, and interactions

Factor	***A* = Si source (case I)**				
B = Si concentrations	Level	a_1	a_2	Mean	a_2−a_1 (simple effects)
	b_1	32.13	34.13	33.13	2
	b_2	38.13	44.13	41.13	6
	Mean	35.13	39.13	37.13	4 (main effect)
	b_2−b_1 (simple effects)	6	10	8 (main effect)	
Factor	***A* = Si source (case II)**				
B = Si concentrations	Level	a_1	a_2	Mean	a_2−a_1 (simple effects)
	b_1	34.13	37.13	35.63	3
	b_2	43.13	33.13	38.13	−10
	Mean	38.63	35.13	36.88	−3.5 (main effect)
	b_2−b_1 (simple effects)	9	−4	2.5 (main effect)	
Factor	***A* = Si source (case III)**				
B = Si concentrations	Level	a_1	a_2	Mean	a_2−a_1 (simple effects)
	b_1	30.13	32.13	31.13	2
	b_2	38.13	40.13	39.13	2
	Mean	34.13	36.13	35.13	2 (main effect)
	b_2−b_1 (simple effects)	8	8	8 (main effect)	

two water regimes, i.e., with water (W+) and without water (W−), the design should be factorial with 2 × 2 or 2^2 as shown in Table 3.14. In factorial experiment, term *level* represents several treatments within any factor. The capital letters are used to represent factors, while levels (treatment combinations and means) were represented with small letters and numerical subscripts, e.g., a_1b_2 may refer to treatment combination consists of first level of A and second level of factor B with the mean of corresponding treatment. The df and SS for the variance among four treatment means in a 2^2 can be divided into single df and SS. Symbolic representation of 3 × 3 or 3^2 factorial treatment combinations has been shown in Table 3.15. The principles involved in the partitioning can be elaborated by Table 3.16. The four differences a_2−a_1 at each level of B and b_2−b_1 at each level of A are called simple effects. Average of simple effects is called main effect denoted by capital letters, e.g., A and B. The A and B for 2^2 factorial experiment can be calculated by using following equations:

$$A = \frac{1}{2}\,[(a_2b_2 - a_1b_2) + (a_2b_2 - a_1b_1)] = \frac{1}{2}\,[(a_2b_2 + a_2b_1) - (a_1b_2 + a_1b_1)]$$

$$B = \frac{1}{2}\,[(a_2b_2 - a_2b_1) + (a_1b_2 - a_1b_1)] = \frac{1}{2}\,[(a_2b_2 + a_1b_2) - (a_2b_1 + a_1b_1)]$$

Main effects in factorial experiment are averaged in number of ways same as other treatment. Different conditions might prevail within blocks and among blocks for factorial experiment in RCBD, and Latin square design thus in Table 3.16 factor A is replicated within every block as it is present at both levels for each level of factor B. In case of factorially arrangement treatment, hypothesis that is usually tested is "there is no interaction among factors." Data presented in Table 3.16 have shown that simple effects under I and II for Si sources (A) and concentrations (B) are different, while for III the simple effects for A and B as well as main effect are the same. The differential response obtained between the simple effects of a factor is called interaction as seen in cases I and II of Table 3.16. However, interaction is not present in case III of Table 3.16. This is the major advantage of application of factorial experiment as it provides information about the interaction between factors. The interaction of A and B can be defined by using following equations:

$$AB = \frac{1}{2}\,[(a_2b_2 - a_1b_2) - (a_2b_1 - a_1b_1)] = \frac{1}{2}\,[(a_2b_2 + a_1b_1) - (a_1b_2 + a_2b_1)]$$

The interaction for the data in Table 3.16:

$$AB = \frac{1}{2}\,(6 - 2) = 2\ (\text{simple effects of } A \text{ for Case I})$$

$$AB = \frac{1}{2}\,(10 - 6) = 2\ (\text{simple effects of } B \text{ for Case I})$$

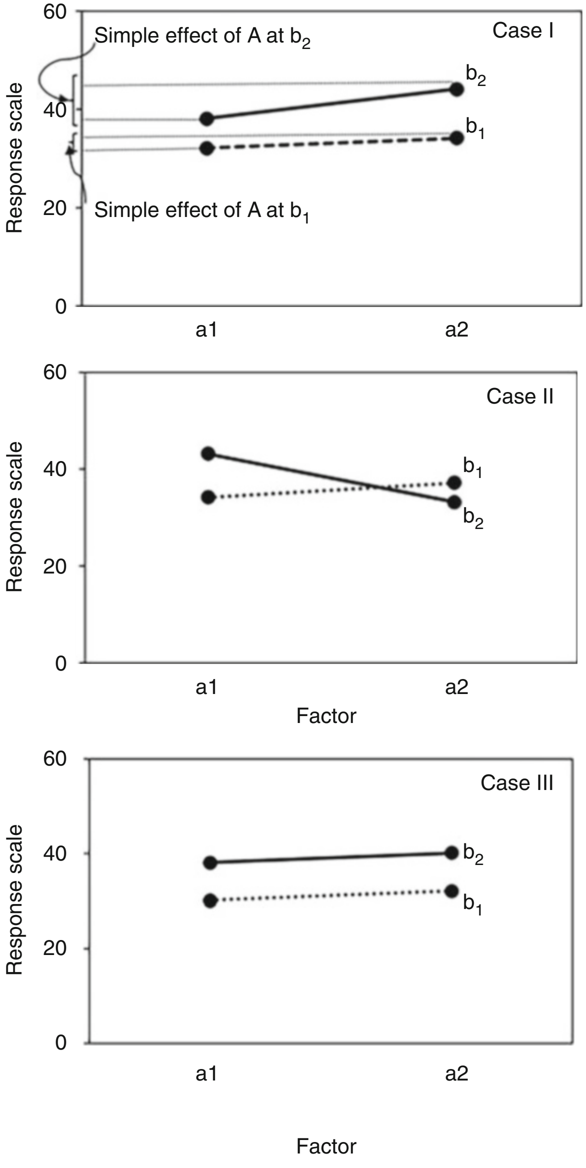
The interaction for case II in Table 3.16:

$$AB = \frac{1}{2}\,[(33.13 - 43.13) - (37.13 - 34.13)]$$

$$AB = \frac{1}{2}\,[33.13 - 43.13 - 37.13 + 34.13]$$

$$AB = \frac{1}{2}\,[-13]$$

Fig. 3.8 Graphical illustration of interaction



$$AB = -6.5$$

The interaction for case III in Table 3.16:

Table 3.17 Three factor ($3 \times 2 \times 3$ or $3^2 \times 2$) factorial experiments

Factor C (sorghum cultivars)	Factor B (Si fertilizer)	Factor A (locations)		
		a_1	a_2	a_3
c_1	b_1	$a_1b_1c_1$	$a_2b_1c_1$	$a_3b_1c_1$
	b_2	$a_1b_2c_1$	$a_2b_2c_1$	$a_3b_2c_1$
c_2	b_1	$a_1b_1c_2$	$a_2b_1c_2$	$a_3b_1c_2$
	b_2	$a_1b_2c_2$	$a_2b_2c_2$	$a_3b_2c_2$
c_3	b_1	$a_1b_1c_3$	$a_2b_1c_3$	$a_3b_1c_3$
	b_2	$a_1b_2c_3$	$a_2b_2c_3$	$a_3b_2c_3$

Table 3.18 Analysis of variance table for $3^2 \times 2$ factorial experiment in RCBD

SOV	df	SS	MS	F
Replication	$r-1=2$	61.65	30.83	8.96
A = locations	$a-1=2$	687.75	343.88	99.96**
B = Si fertilizer	$b-1=1$	149.25	149.25	43.39**
C = sorghum cultivars	$c-1=2$	1438.93	719.46	209.15**
AB	$(a-1)(b-1)=2$	2.47	1.24	0.36
AC	$(a-1)(c-1)=4$	6.35	1.59	0.46
BC	$(b-1)(c-1)=2$	1.38	0.69	0.20
ABC	$(a-1)(b-1)(c-1)=4$	0.024	0.006	0.001744
Error	$(r-1)(abc-1)=34$	116.98	3.44	
Total	$abcr-1=53$	2464.78		

** P < 0.05

$$AB = \frac{1}{2}\,[(40.13 - 38.13) - (32.13 - 30.13)]$$

$$AB = \frac{1}{2}\,[40.13 - 38.13 - 32.13 + 30.13]$$

$$AB = \frac{1}{2}\,[0] = 0\ (\text{no intearction})$$

Interaction concept is further elaborated by using graph as shown in Fig. 3.8. It should be noted that presence or absence of main effects does not tell anything about interaction presences or absence and vice versa. If interaction is nonsignificant, we can conclude that factors act independently. However, if interaction is large and significant, then main effects have little meaning. For large factorial experiments, it has been suggested to use confounded designs as described by Das and Giri (1979).

Factorial experiment other case includes e.g. if we have actor A as three locations and factor B as Si fertilizer with two levels, while factor C consists of three sorghum cultivars; such kind of factorial experiment will be referred as $3 \times 2 \times 3$ or $3^2 \times 2$ (Table 3.17).

ANOVA calculation for the 3 × 3 × 2 or $3^2 \times 2$ factorial experiments involves following steps with results presented in ANOVA Table 3.18:

1. Calculation of correction factor, total sum of square, block SS, treatment SS and error SS

$$\text{Correction factor} = \text{CF} = \frac{X^2_{\cdot\cdot}}{rabc} = \frac{(2903)^2}{54} = 156,038.77$$

$$\text{SS}_{\text{total}} = \sum_{i,j,k,r} X^2{}_{ijkr} - \text{CF} = 158,503.56 - 156,038.77 = 2464.78$$

$$\text{SS}_{\text{replication}} = \frac{\sum_{k=1}^{r} R^2{}_k}{abc} - \text{CF}$$

$$\text{SS}_{\text{repliaction}} = \frac{(968)^2 + (983)^2 + (953)^2}{18} - 156,038.77$$
$$= 156,100.43 - 156,038.77 = 61.65$$

$$\text{SS}_{\text{treatment}} = \frac{\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{c} Tr^2{}_{ijk}}{R} - \text{CF}$$

$$\text{SS}_{\text{treatment}} = \frac{(187)^2 + \ldots + (134)^2}{3} - 156,038.77 = 158,324.90 - 156,038.77$$
$$= 2286.15$$

$$\text{SS}_{\text{error}} = \text{SS}_{\text{total}} - \text{SS}_{\text{repliaction}} - \text{SS}_{\text{treatment}} = 2464.78 - 61.65 - 2286.15 = 116.98$$

2. Partitioning of treatments sum of squares into main effects and interactions

$$\text{SS}_A = \frac{\sum_j (a_j)^2}{rbc} - \text{CF}$$

$$\text{SS}_A = \frac{(1053)^2 + (952)^2 + (898)^2}{18} - 156,038.77 = 156,726.5 - 156,038.77$$
$$= 687.75$$

$$\text{SS}_B = \frac{\sum_k (b_k)^2}{rac} - \text{CF}$$

$$\mathrm{SS}_B = \frac{(1406)^2 + (1496)^2}{27} - 156{,}038.77 = 156{,}188 - 156{,}038.77 = 149.25$$

$$\mathrm{SS}_C = \frac{\sum_i (c_i)^2}{rab} - \mathrm{CF}$$

$$\mathrm{SS}_C = \frac{(1067)^2 + (992)^2 + (843)^2}{18} - 156{,}038.77 = 157{,}477.7 - 156{,}038.77$$
$$= 1438.93$$

$$\mathrm{SS}_{AB} = \frac{\sum_{j,k} \left(a_j b_k\right)^2}{rc} - \mathrm{CF} - (\mathrm{SS}_A + \mathrm{SS}_B)$$

$$\mathrm{SS}_{AB} = \frac{(509)^2 + (544)^2 + (462)^2 + (490)^2 + (435)^2 + (462)^2}{9}$$
$$- 156{,}038.77 - 687.75 - 149.25 = 2.47$$

$$\mathrm{SS}_{AC} = \frac{\sum_{j,i} \left(a_j c_i\right)^2}{rb} - \mathrm{CF} - (\mathrm{SS}_A + \mathrm{SS}_C)$$

$$\mathrm{SS}_{AC} = \frac{(387)^2 + (360)^2 + (306)^2 + (350)^2 + (326)^2 + (277)^2 + (330)^2 + (307)^2 + (261)^2}{6}$$
$$- 156{,}038.77 - (687.75 + 1438.93) = 6.35$$

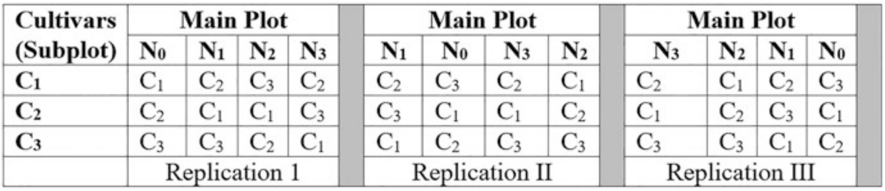
$$\mathrm{SS}_{BC} = \frac{\sum_{k,i} (b_k c_i)^2}{ra} - CF - (\mathrm{SS}_B + \mathrm{SS}_C)$$

$$\mathrm{SS}_{BC} = \frac{(517)^2 + (550)^2 + (481)^2 + (512)^2 + (409)^2 + (435)^2}{9}$$
$$- 156{,}038.77 - (149.25 + 1438.93) = 1.38$$

$$\mathrm{SS}_{ABC} = \frac{\sum_{i,j,k} \left(a_j b_k c_i\right)^2}{r} - \mathrm{CF} - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_C - \mathrm{SS}_{AB} - \mathrm{SS}_{AC} - \mathrm{SS}_{BC}$$

$$\mathrm{SS}_{ABC} = \frac{(187)^2 + \ldots (134)^2}{3} - 156{,}038.77$$
$$- (\ \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_C + \mathrm{SS}_{AB} + \mathrm{SS}_{AC} + \mathrm{SS}_{BC})$$

$$\mathrm{SS}_{ABC} = \frac{(187)^2 + \ldots (134)^2}{3} - 156{,}038.77 - 2286.3 = 0.024$$



Cultivars	Main Plot				Main Plot				Main Plot			
(Subplot)	N_0	N_1	N_2	N_3	N_1	N_0	N_3	N_2	N_3	N_2	N_1	N_0
C_1	C_1	C_2	C_3	C_2	C_2	C_3	C_2	C_1	C_2	C_1	C_2	C_3
C_2	C_2	C_1	C_1	C_3	C_3	C_1	C_1	C_2	C_1	C_2	C_3	C_1
C_3	C_3	C_3	C_2	C_1	C_1	C_2	C_3	C_3	C_3	C_3	C_1	C_2
	Replication 1				Replication II				Replication III			

Fig. 3.9 Layout for the split plot design

3.8.6 Fractional Factorial Design

Fractional factorial design is used when large number of factors needs to be tested. In this case, only fraction of total number of treatments is going to be tested based upon the systematic selection.

3.8.7 Nested and Split Plot Design

Nested and split plot experiments are multifactor experiments. Split plot design is used for factorial experiment with a principle that whole plots are divided into subplots or subunits. The factors which need more importance, greater precision, and smaller experimental material and expected to exhibit smaller differences are placed in the subunits. Consider an experiment to test factor *A* (nitrogen fertilizer) at four levels of RCBD and second factor *B* (sorghum cultivars) at three levels which can be placed by dividing each A units into subunits. Thus, layout for the split plot includes factor *A* which will be in the main plot while factor *B* in the subplot as shown in Fig. 3.9.

Layout design steps for the split plot includes (i) Division of experimental area into three blocks or replication with further division into four main plots for the nitrogen fertilizer application (ii) Two separate randomization is needed, firstly for the main plot (*N* treatments) and then for the subplots (cultivars). Split plot design in figure showed that size of the main plot is "*c*" times greater than subplot. Since in this experiment $c = 3$ (cultivars in subplot), thus the size of main plot is three times greater than subplot. However, each main plot treatment is tested, e.g., 3 times, while subplot treatment will be tested 12 times which leads to more precision in subplot treatments as compared to the main plot. Partitioning of degree of freedom for the split plot design under different arrangements has been presented in Table 3.19.

3.8.8 Strip Plot/Split-Block Design

Experiments in which both factors (e.g., *A* and *B* with multiple levels of *a* and *b*) require larger plot area strip plot design are used. In this design, whole area is divided into "*a*" horizontal and "*b*" vertical strips. One level of factor *A* is applied in

Table 3.19 Degree of freedom for split plot design under different arrangements

Completely randomized (r replications)		RCBD		Latin square	
SOV	df	SOV	df	SOV	df
Main unit or main plot					
				Rows	$a-1$
		Blocks	$r-1$	Columns	$a-1$
A	$a-1$	A	$a-1$	A	$a-1$
Error (a)	$a(r-1)$	Error (a)	$(a-1)(r-1)$	Error (a)	$(a-1)(a-2)$
Total	$ar-1$	Total	$ar-1$	Total	a^2-1
Subunit or subplot					
B	$b-1$	B	$b-1$	B	$b-1$
AB	$(a-1)(b-1)$	AB	$(a-1)(b-1)$	AB	$(a-1)(b-1)$
Error (b)	$a(r-1)(b-1)$	Error (b)	$a(r-1)(b-1)$	Error (b)	$a(a-1)(b-1)$
Subtotal	$ar(b-1)$	Subtotal	$ar(b-1)$	Subtotal	$a^2(b-1)$
Total	$abr-1$	Total	$abr-1$	Total	a^2b-1

Table 3.20 Analysis of variance for split-split plot design

SOV	df
(Main plot)	
Block	$r-1$
Factor A	$a-1$
Whole plot error	$(r-1)(a-1)$
(Subplots)	
Factor B	$b-1$
$A \times B$	$(a-1)(b-1)$
Subplot error	$a(r-1)(b-1)$
(Sub-subplots)	
Factor C	$c-1$
$A \times C$	$(a-1)(c-1)$
$B \times C$	$(b-1)(c-1)$
$A \times B \times C$	$(a-1)(b-1)(c-1)$
Sub-subplot error	$ab(r-1)(c-1)$
Total	$(rabc)-1$

horizontal strips while level of B in vertical strips. Strip plot main difference from split plot is to have second factor as strip.

3.8.9 Split-Split Plot Design

Split-split plot designs are applicable when there are three-factor factorial experiments with factor A assign to whole plots while factor B to subplot and factor C to sub-subplot. The ANOVA for split-split plot design with r blocks, a levels of factor A, b levels of factor B, and c levels of factor C has been shown in Table 3.20.

3.8.10 MANOVA (Multivariate Analysis of Variance)

Multivariate analysis of variance (MANOVA) is ANOVA with several dependent variables. It tests the difference in two or more vectors of means, e.g., evaluation of student's improvements in Physics and Chemistry using different syllabus. In this case, response variable (students' improvements) is altered by the observer manipulation of the independent variables. The assumptions to use MANOVA are:

1. The dependent variable should be normally distributed.
2. Linear relationship among all pairs of dependent variables.
3. Homogeneity of variances.

3.9 ANCOVA (Analysis of Covariance)

Analysis of covariance (ANCOVA) uses concepts of both analysis of variance and regression, and it is used when one independent variable is not at predetermined level. The uses of ANCOVA includes (i) increase of precision and control of error, (ii) estimation of missing data, (iii) adjustment of treatment means of dependent variables for corresponding independent variables, (iv) assistance in the data interpretation, and (v) partitioning of total covariance into parts.

3.10 Principal Component Analysis (PCA)

Principal component analysis is the method of multivariate statistics used to check variation and patterns in a data set. It is an easy way to visualize and explore data (Ahmed et al. 2020). Consider a data in two dimensions first (e.g., height and weight). The data can be plotted using scatter plot, but if we want to see variation, we must use PCA with new coordinate system. The axes don't have any physical meaning. Thus, PCA is a statistical procedure that uses orthogonal transformation to convert set of observation of correlated variables into values of linearly uncorrelated variables. It is the most common form of factor analysis applied to analyze interrelationship among variables (Fig. 3.10). The main objective of PCA is to cluster variables into manageable groups. These groups are known as the components (factors). Steps involved for the PCA are:

1. Standardization of the data $\left(z = {}^{\text{Variable value}-\text{Mean}}\!/_{\text{Standard deviation}}\right)$
2. Computing the covariance matrix (identification of correlation and dependence among features in a data set)
3. Eigenvectors and eigenvalues calculation
4. Commuting the principal components
5. Reducing the dimension of data set

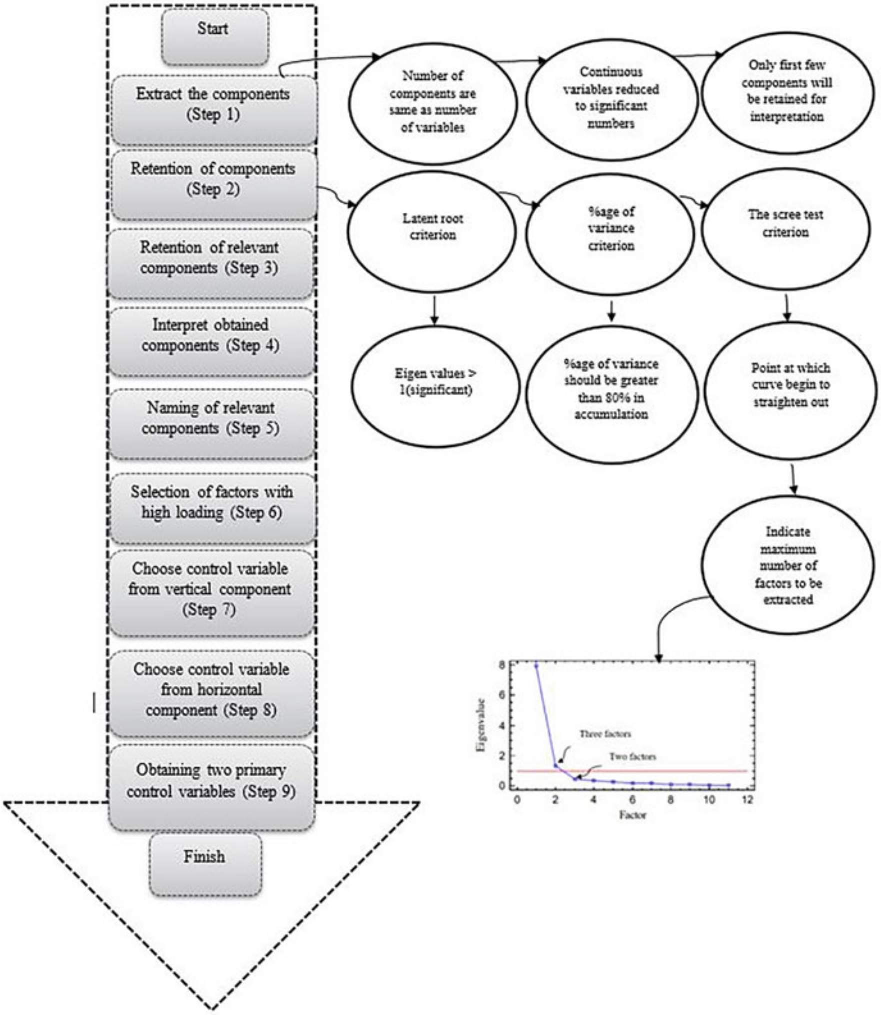

Start
Extract the components (Step 1)
Retention of components (Step 2)
Retention of relevant components (Step 3)
Interpret obtained components (Step 4)
Naming of relevant components (Step 5)
Selection of factors with high loading (Step 6)
Choose control variable from vertical component (Step 7)
Choose control variable from horizontal component (Step 8)
Obtaining two primary control variables (Step 9)
Finish
Number of components are same as number of variables
Continuous variables reduced to significant numbers
Only first few components will be retained for interpretation
Latent root criterion
%age of variance criterion
The scree test criterion
Eigen values > 1(significant)
%age of variance should be greater than 80% in accumulation
Point at which curve begin to straighten out
Indicate maximum number of factors to be extracted
Eigenvalue
Factor
Three factors
Two factors
0
2
4
6
8
10
12

Fig. 3.10 PCA flow diagram

3.11 Regression

Consider a random sample of n observations in which Y values are determined from the corresponding X values, i.e., (X_1, Y_1), (X_2, Y_2), (X_3, Y_3). . . . (X_n, Y_n). In this case, Y is a dependent variable while X is an independent variable. First descriptive technique which can be used to determine the relationship between X and Y is the scatter diagram. This diagram is drawn by plotting the X and Y in Cartesian coordinates. The plotting pattern of points obtained between variables tells the relationship which can be either linear or nonlinear (Fig. 3.11). If relationship is

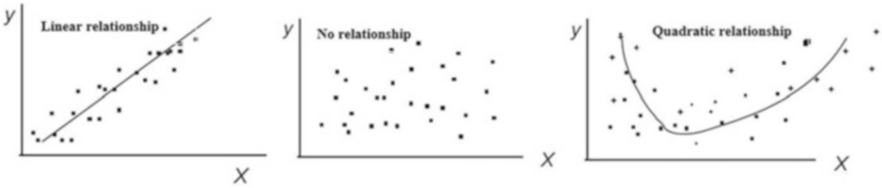


Fig. 3.11 Scatter plot to show relationship between two variables X and Y

linear, then we need to fit model that fits with the given data. Mathematically, the relation between X and Y can be elaborated by the following equation:

$$Y \propto X$$

This shows that there is relationship present between the two variables and drawn straight line between the points can serve as moving average of the Y values. The equation of straight line can be:

$$Y = a + bX$$

Any point (X, Y) on this line has a X coordinate (abscissa) and a Y coordinate (ordinate) whose values satisfy this equation. When $X = 0$ or minimum, $Y = a$ (intercept, value of Y X is minimum or zero). When intercept (a) is zero, the line passes through the origin. A unit change in Y due to unit change in X is called slope of the line and represented with b. Thus $b = \frac{\Delta Y}{\Delta X} = \frac{\text{Unit change in } Y}{\text{Unit Change in } X}$. If b is positive, both values increase or decrease together, but if b is negative, then one value increases while other decreases. This is an example of simple linear regression equation (Ahmed et al. 2011). However, if we increase number of X variables called as predictor variable (X_1 to X_n) against Y, it will be called multiple linear regression. The form of equation for the multiple linear regression will be:

$$Y = a + \beta_o X_1 + \beta_1 X_2 + \beta_2 X_3 + \ldots \beta_n X_n + \varepsilon$$

where $X_1 \ldots X_n =$ independent non-random variable; $\beta_0, \beta_1, \beta_2 \ldots \beta_n =$ slope; and $\varepsilon =$ random varible represnting error term and genearlly equal to zero.

Let's consider the data set presented in Table 3.21 to describe the method of least square in order to fit a straight line and calculate simple regression equation and coefficient of determination (R^2). The calculation involves determination of SS_{xx}, SS_{xy}, $\overline{X}$, $\overline{Y}$, and β_1 as shown in the following equations:

$$\text{SS}_{xx} = \sum\nolimits_{i=1}^{n} X^2{}_i - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = 639 - \frac{(45)^2}{10} = 436.5$$

Table 3.21 Data set to illustrate method of least squares to fit a straight line

X_i	Y_i	X_iY_i	X_i^2
−2	−7	14	4
0	−3	0	0
4	3	12	16
−4	−9	36	16
7	8	56	49
8	11	88	64
10	15	150	100
13	23	299	169
14	25	350	196
−5	−11	55	25
$\sum X_i$=45	$\sum Y_i$=55	$\sum X_iY_i$=1060	$\sum X_i^2$=639

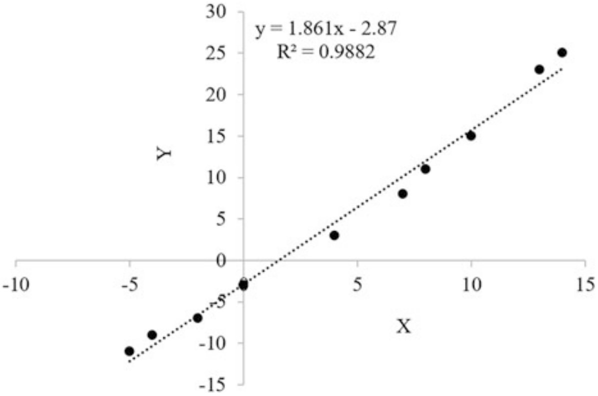


Fig. 3.12 Simple linear regression line with regression equation and coefficient of determination (R^2)

$$\mathrm{SS}_{xy} = \sum\nolimits_{i=1}^{n} X_iY_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 1060 - \frac{(45)(55)}{10} = 812.5$$

$$\overline{X} = 4.5 \text{ and } \overline{Y} = 5.5.$$

$$\beta_1 = \frac{\mathrm{SS}_{XX}}{\mathrm{SS}_X} = \frac{812.5}{436.5} = 1.86$$

and

$$\overline{Y} = a + \beta_1\overline{X}$$

$$a = \overline{Y} - \beta_1\overline{X} = 5.5 - (1.86)(4.5) = 5.5 - 8.37 = -2.87.$$

Hence simple regression equation for this data is:

Table 3.22 ANOVA table for simple regression

SOV	df	Sum of squares (SS)	Mean squares (MS)	F
Regression (model)	1	$\mathrm{SS}_R=\sum_{i=1}^{n}(\widehat{\boldsymbol{y}}_i-\overline{\boldsymbol{y}})^2$	$\frac{\mathrm{SS}_R}{\mathrm{df}_R}$	$\frac{\mathrm{MS}_R}{\mathrm{MS}_{\mathrm{error}}}$
Error (residuals)	$n-2$	$\mathrm{SS}_E=\sum_{i=1}^{n}(\boldsymbol{y}_i-\widehat{\boldsymbol{y}}_i)^2$	$\frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{df}_{\mathrm{error}}}$	
Total	$n-1$	$\mathrm{SS}_T=\sum_{i=1}^{n}(\boldsymbol{y}_i-\overline{\boldsymbol{y}})^2$		

$$\widehat{Y}=a+\beta_1X=-2.87+(1.86)\mathrm{X}.$$

The plot for this least square line is shown in Fig. 3.12. The quality of this fit can be measured quantitatively by using coefficient of determination (R^2). The equation for R^2 calculation is:

$$R^2=\frac{\mathrm{SS}_{yy}-\mathrm{SS}_{\mathrm{error}}}{\mathrm{SS}_{yy}}=1-\frac{\sum\limits_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

$$\mathrm{SS}_{\mathrm{error}}=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2=\sum_{i=1}^{n}\Bigg(y_i-(a+\beta_1X)^2=\sum_{i=1}^{n}(y_i-a-\beta_1X)^2$$

Other approach which could be used to test hypothesis is use of ANOVA table as presented in earlier section. The ANOVA table for regression analysis is presented in Table 3.22. Furthermore, application of concept of multiple linear stepwise regression models has been elaborated using spring wheat grain yield data with respective R^2 (Table 3.23).

3.12 Correlation

Correlation is used to measure intensity or degree of association between variables. It is the same as covariance. It is a bivariate statistical technique. The simple linear correlation coefficient or simple correlation (total correlation and product-moment correlation) is sued for descriptive purposes and can be calculated by using following equations:

$$\boldsymbol{r}=\frac{\sum(X-\overline{X})(Y-\overline{Y})/_{n-1}}{\sqrt{\sum(X-\overline{X})^2/_{n-1}}\sqrt{\sum(Y-\overline{Y})^2/_{n-1}}}$$

Table 3.23 Multiple linear stepwise regression models for spring wheat grain yield with environmental variables (E = environments (2008–09 and 2009–10), PW = planting windows, SR1 = solar radiation at anthesis, SR1 = solar radiation at maturity, T1 = mean average temperature at anthesis, T2 = mean average temperature at anthesis, PTQ1 = photothermal quotient at anthesis, PTQ2 = photothermal quotient at maturity) using stepwise method developed to predict wheat grain yield under changing climate

Regression models	
GY = β0 + β1X1	***R^2***
GY = 4495.18–872.517∗E	69.13
GY = 4115.66–309.75∗PW	69.28
GY = −1516.48 + 3.95491∗SR1	92.63
GY = −1575.85 + 2.57179∗SR2	93.43
GY = 3284.76–5.97019∗T1	51.02
GY = 4542.37–57.863∗T2	52.15
GY = −3814.55 + 49.8777∗PTQ1	87.77
GY = −2582.32 + 31.8736∗PTQ2	93.34
GY = β0 + β1X1 + β2X2	***R^2***
GY = 5424.43–872.517∗E − 309.75∗PW	87.41
GY = −2440.37 + 224.118∗E + 4.44915∗SR1	93.18
GY = −2366.93 + 193.912∗E + 2.84192∗SR2	93.86
GY = 5196.35–901.852∗E − 39.8932∗T1	69.85
GY = 1594.13–1225.4∗E + 146.385∗T2	73.54
GY = −2465.19 − 302.166∗E + 43.4934∗PTQ1	89.34
GY = −3268.13 + 149.999∗E + 34.4196∗PTQ2	93.61
GY = −836.161 − 80.6277∗PW + 3.5862∗SR1	93.51
GY = −820.636 − 92.5138∗PW + 2.31383∗SR2	94.63
GY = 1995.01–443.928∗PW + 153.171∗T1	76.77
GY = 5347.72–308.175∗PW − 52.7775∗T2	70.24
GY = −3453.92 − 26.2988∗PW + 47.8705∗PTQ1	87.84
GY = −2399.35 − 16.9782∗PW + 31.144∗PTQ2	93.38
GY = −1667.97 + 1.65868∗SR1 + 1.55638∗SR2	94.14
GY = −1241.88 + 3.96355∗SR1–17.2937∗T1	92.77
GY = −4931.17 + 27.4934∗SR1–519.678∗T2	93.91
GY = −1411.9 + 4.0845∗SR1–1.84298∗PTQ1	92.64
GY = −2262.19 + 1.62354∗SR1 + 19.4378∗PTQ2	93.91
GY = −945.844 + 2.61657∗SR2–43.2789∗T1	94.29
GY = −2612.93 + 2.64662∗SR2 + 38.3422∗T2	93.90
GY = −1808.8 + 2.39643∗SR2 + 3.97299∗PTQ1	93.46
GY = −2133.2 + 1.35065∗SR2 + 15.5732∗PTQ2	93.97
GY = 4550.48–0.599316∗T1–57.7877∗T2	52.15
GY = −3680.6 − 8.23152∗T1 + 49.8894∗PTQ1	87.80
GY = −2804.74 + 12.603∗T1 + 31.9554∗PTQ2	93.42
GY = −4063.73 + 8.96602∗T2 + 50.156∗PTQ1	87.80
GY = −4661.87 + 71.4527∗T2 + 34.112∗PTQ2	94.89
GY = −2572.68 − 0.241837∗PTQ1 + 32.0078∗PTQ2	93.34
GY = β0 + β1X1 + β2X2 + β3X3	***R^2***

(continued)

Table 3.23 (continued)

Regression models	
GY = −1121.59 + 49.1956∗E − 70.8354∗PW + 3.73947∗SR1	93.52
GY = −740.097 − 14.575∗E − 95.0961∗PW + 2.28633∗SR2	94.63
GY = 3783.86–791.812∗E − 405.893∗PW + 109.752∗T1	91.10
GY = 2365.39–1246.31∗E − 314.376∗PW + 155.057∗T2	92.36
GY = 305.873–494.994∗E − 139.285∗PW + 28.7889∗PTQ1	90.73
GY = −4950.36 + 342.09∗E + 74.6∗PW + 40.8858∗PTQ2	93.81
GY = −2778.85 + 266.679∗E + 2.0718∗SR1 + 1.67498∗SR2	94.90
GY = −2210.96 + 209.746∗E + 4.42281∗SR1–10.7162∗T1	93.23
GY = −3512.62 + 0.220823∗E + 4.23482∗SR1 + 70.9633∗T2	94.16
GY = −1752.91 + 564.722∗E + 7.79189∗SR1–36.8568∗PTQ1	94.12
GY = −3220.65 + 230.332∗E + 2.10348∗SR1 + 19.6712∗PTQ2	94.49
GY = −1660.51 + 164.599∗E + 2.8428∗SR2–40.3133∗T1	94.59
GY = −2686.81 + 105.353∗E + 2.7677∗SR2 + 25.1835∗T2	93.97
GY = −2248.81 + 223.19∗E + 3.06154∗SR2–4.05167∗PTQ1	93.88
GY = −2985.38 + 203.511∗E + 1.58608∗SR2 + 16.1864∗PTQ2	94.44
GY = 2279.55–1352.58∗E − 73.2781∗T1 + 176.787∗T2	75.78
GY = −2050.44 − 322.027∗E − 20.0369∗T1 + 43.1024∗PTQ1	89.52
GY = −3836.52 + 190.605∗E + 21.6873∗T1 + 35.2497∗PTQ2	93.81
GY = −3818.53 − 545.453∗E + 87.7865∗T2 + 41.0788∗PTQ1	90.87
GY = −4546.3 − 96.9147∗E + 82.7066∗T2 + 32.8195∗PTQ2	94.97
GY = −3189.3 + 212.106∗E − 9.10147∗PTQ1 + 40.5279∗PTQ2	93.72
GY = −981.556 − 81.4573∗PW + 1.27236∗SR1 + 1.56575∗SR2	95.03
GY = −959.051 − 111.806∗PW + 3.43148∗SR1 + 24.3069∗T1	93.65
GY = −2758.03 − 61.055∗PW + 3.92173∗SR1 + 62.4805∗T2	94.64
GY = 881.792–127.627∗PW + 5.00859∗SR1–23.2851∗PTQ1	94.09
GY = −1734.58 − 43.2957∗PW + 1.93291∗SR1 + 15.2077∗PTQ2	94.11
GY = −772.826 − 75.9091∗PW + 2.37316∗SR2–12.596∗T1	94.67
GY = −1696.5 − 86.1818∗PW + 2.39096∗SR2 + 30.4708∗T2	94.92
GY = 365.25–128.985∗PW + 2.88071∗SR2–15.1477∗PTQ1	94.97
GY = −765.837 − 95.1082∗PW + 2.38026∗SR2–0.939409∗PTQ2	94.63
GY = 3529.35–449.248∗PW + 161.654∗T1–70.7574∗T2	78.47
GY = −3435.79 − 31.2956∗PW + 3.09644∗T1 + 47.4847∗PTQ1	87.85
GY = −2497.58 − 63.0856∗PW + 33.7223∗T1 + 29.3817∗PTQ2	93.66
GY = −3669.07 − 24.3883∗PW + 6.79871∗T2 + 48.2274∗PTQ1	87.86
GY = −4910.46 + 15.9228∗PW + 74.0978∗T2 + 34.879∗PTQ2	94.92
GY = −2318.27 − 18.7935∗PW − 1.54335∗PTQ1 + 31.923∗PTQ2	93.38
GY = −1128.05 + 1.28052∗SR1 + 1.82477∗SR2–35.6472∗T1	94.68
GY = −3287.84 + 2.29884∗SR1 + 1.27881∗SR2 + 58.5742∗T2	95.13
GY = −1166.99 + 2.14204∗SR1 + 1.65779∗SR2–9.00197∗PTQ1	94.26
GY = −1975.86 + 1.25311∗SR1 + 1.08073∗SR2 + 9.23233∗PTQ2	94.29
GY = −4391.49 + 1.78814∗SR1 + 74.2775∗T2 + 20.5039∗PTQ2	95.58
GY = −1715.88 + 2.1342∗SR1–11.1793∗PTQ1 + 21.734∗PTQ2	94.10
GY = −2132.17 + 2.71418∗SR2–49.035∗T1 + 46.9579∗T2	94.98

(continued)

Table 3.23 (continued)

Regression models	
GY = −860.582 + 2.6733∗SR2–44.0303∗T1–1.2676∗PTQ1	94.29
GY = −987.816 + 2.5613∗SR2–42.0901∗T1 + 0.689252∗PTQ2	94.29
GY = −2736.16 + 2.53799∗SR2 + 37.6309∗T2 + 2.42991∗PTQ1	93.91
GY = −4176.79 + 0.754416∗SR2 + 63.405∗T2 + 24.7552∗PTQ2	95.07
GY = −2003.58 + 1.38361∗SR2–2.97666∗PTQ1 + 16.8284∗PTQ2	93.99
GY = −3948.13 − 9.1929∗T1 + 10.189∗T2 + 50.2071∗PTQ1	87.84
GY = −4766.43 + 7.30566∗T1 + 70.6152∗T2 + 34.1332∗PTQ2	94.92
GY = −4497.06 + 81.5281∗T2–11.4913∗PTQ1 + 40.8087∗PTQ2	95.12
GY = β0 + β1X1 + β2X2 + β3X3 + β4X4	***R*²**
GY = −1717.12 + 125.93∗E − 56.4191∗PW + 1.58618∗SR1 + 1.61888∗SR2	95.12
GY = −668.395 − 53.9363∗E − 128.193∗PW + 3.2354∗SR1 + 28.713∗T1	93.66
GY = −1479.49 − 418.451∗E − 132.643∗PW + 2.81869∗SR1 + 99.8435∗T2	95.20
GY = −117.991 + 382.545∗E − 84.8503∗PW + 7.21027∗SR1–39.8168∗PTQ1	94.61
GY = −4412.45 + 362.263∗E + 52.7493∗PW + 2.00147∗SR1 + 24.9586∗PTQ2	94.59
GY = −1029.59 + 50.4408∗E − 59.3438∗PW + 2.49561∗SR2–18.383∗T1	94.68
GY = −818.424 − 375.193∗E − 144.014∗PW + 1.78818∗SR2 + 72.0508∗T2	95.36
GY = 211.56 + 35.3623∗E − 124.003∗PW + 2.96738∗SR2–15.6806∗PTQ1	94.98
GY = 17.1993–84.1002∗E − 125.078∗PW + 2.60741∗SR2–6.39615∗PTQ2	94.65
GY = 1809.0–1117.23∗E − 380.863∗PW + 76.9233∗T1 + 124.976∗T2	93.99
GY = 1727.51–626.402∗E − 285.99∗PW + 72.3236∗T1 + 14.7125∗PTQ1	91.54
GY = −4499.41 + 276.811∗E + 40.3019∗PW + 12.304∗T1 + 38.3838∗PTQ2	93.83
GY = −345.954 − 932.419∗E − 203.84∗PW + 125.59∗T2 + 18.5194∗PTQ1	93.55
GY = −4041.76 − 179.645∗E − 25.9982∗PW + 87.9945∗T2 + 30.4637∗PTQ2	94.98
GY = −5519.16 + 522.302∗E + 105.491∗PW − 14.7528∗PTQ1 + 53.4642∗PTQ2	94.07
GY = −2201.79 + 233.107∗E + 1.71356∗SR1 + 1.87739∗SR2–28.8663∗T1	95.25
GY = −3381.18 + 122.646∗E + 2.32382∗SR1 + 1.40491∗SR2 + 43.4754∗T2	95.23
GY = −1594.43 + 1002.58∗E + 7.31229∗SR1 + 2.8626∗SR2–76.3668∗PTQ1	98.07
GY = −3001.23 + 259.019∗E + 1.72496∗SR1 + 1.27873∗SR2 + 7.62509∗PTQ2	95.00
GY = −2118.36 − 13.347∗E + 2.69947∗SR2–49.4896∗T1 + 48.7048∗T2	94.98
GY = −1298.82 + 231.246∗E + 3.36756∗SR2–44.8499∗T1–9.67945∗PTQ1	94.74
GY = −1996.51 + 173.71∗E + 2.46492∗SR2–31.7531∗T1 + 4.86799∗PTQ2	94.61
GY = −3250.19 − 640.28∗E − 41.8047∗T1 + 107.051∗T2 + 39.7331∗PTQ1	91.58
GY = −4587.11 − 89.3775∗E + 2.22363∗T1 + 81.5764∗T2 + 32.9265∗PTQ2	94.97
GY = −4475.23 − 28.2616∗E + 84.084∗T2–10.6634∗PTQ1 + 39.9494∗PTQ2	95.12
GY = −955.007 − 73.689∗PW + 1.24475∗SR1 + 1.61059∗SR2–6.07485∗T1	95.04
GY = −2462.56 − 66.0303∗PW + 1.87942∗SR1 + 1.33248∗SR2 + 48.8522∗T2	95.69
GY = 1674.45–155.41∗PW + 2.88299∗SR1 + 1.98577∗SR2–36.528∗PTQ1	96.35
GY = −312.752 − 112.05∗PW + 1.66867∗SR1 + 2.20423∗SR2–12.3245∗PTQ2	95.17
GY = −1843.34 − 44.634∗PW + 2.55566∗SR2–30.0858∗T1 + 39.5518∗T2	95.09
GY = 363.669–128.524∗PW + 2.88078∗SR2–0.28732∗T1–15.1136∗PTQ1	94.97
GY = −408.71 − 81.9211∗PW + 2.81537∗SR2–20.0906∗T1–5.75433∗PTQ2	94.69
GY = −3670.93 − 24.1757∗PW − 0.12556∗T1 + 6.83431∗T2 + 48.2449∗PTQ1	87.86
GY = −4874.7 + 10.215∗PW + 3.72791∗T1 + 72.7223∗T2 + 34.6149∗PTQ2	94.92

(continued)

Table 3.23 (continued)

Regression models	
GY = −4602.61 + 6.44894∗PW + 82.3013∗T2–11.1513∗PTQ1 + 40.9213∗PTQ2	95.12
GY = −2789.6 + 1.91952∗SR1 + 1.55887∗SR2–39.4595∗T1 + 62.1691∗T2	95.80
GY = −396.329 + 1.89706∗SR1 + 1.98872∗SR2–39.1942∗T1–12.183∗PTQ1	94.91
GY = −480.841 + 1.59121∗SR1 + 2.54324∗SR2–53.3785∗T1–11.3543∗PTQ2	94.77
GY = −1893.98 + 2.90465∗SR2–51.719∗T1 + 48.6524∗T2–4.17775∗PTQ1	95.01
GY = −3391.81 + 1.46283∗SR2–23.0213∗T1 + 58.4869∗T2 + 15.9022∗PTQ2	95.16
GY = −3981.32 + 0.791952∗SR2 + 73.4787∗T2–11.9459∗PTQ1 + 31.2513∗PTQ2	95.31
Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5	***R*²**
GY = −2348.68 + 258.764∗E + 12.2235∗PW + 1.76716∗SR1 + 1.9187∗SR2–33.0254∗T1	95.25
GY = −1792.68 − 234.418∗E − 105.354∗PW + 1.58191∗SR1 + 1.12343∗SR2 + 71.9212∗T2	95.84
GY = −184.511 + 838.766∗E − 73.2942∗PW + 6.81694∗SR1 + 2.82043∗SR2–78.3416∗PTQ1	98.43
GY = −289.506 − 2.53508∗E − 112.94∗PW + 1.66736∗SR1 + 2.21121∗SR2–12.48∗PTQ2	95.17
GY = −2589.5 − 27.4556∗PW + 1.85404∗SR1 + 1.50077∗SR2–28.13∗T1 + 57.0945∗T2	95.84
GY = 2202.57–231.216∗PW + 3.51711∗SR1 + 1.77738∗SR2 + 43.6298∗T1–46.4137∗PTQ1	96.65
GY = 245.489–93.855∗PW + 1.78851∗SR1 + 2.83208∗SR2–29.5739∗T1–20.2299∗PTQ2	95.30
GY = −746.366 − 94.1724∗PW + 3.00058∗SR2–17.8032∗T1 + 36.8552∗T2–13.6174∗PTQ1	95.34
GY = −3009.02 − 20.8459∗PW + 1.63506∗SR2–19.2907∗T1 + 52.7591∗T2 + 12.7727∗PTQ2	95.18
GY = −4266.05 − 32.9476∗PW + 23.4955∗T1 + 76.5957∗T2–15.1799∗PTQ1 + 41.4392∗PTQ2	95.21
GY = −1819.0 + 3.50132∗SR1 + 1.81621∗SR2–48.6106∗T1 + 85.463∗T2–26.5257∗PTQ1	96.70
GY = −3031.46 + 1.8465∗SR1 + 1.33771∗SR2–34.3125∗T1 + 64.033∗T2 + 3.36904∗PTQ2	95.80
GY = −3461.14 + 1.27339∗SR2–15.7506∗T1 + 69.2434∗T2–10.9138∗PTQ1 + 24.633∗PTQ2	95.32
Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6	***R*²**
GY = −2115.22 − 148.483∗E − 67.5486∗PW + 1.67558∗SR1 + 1.30211∗SR2–17.0565∗T1 + 68.4621∗T2	95.88
GY = −517.057 + 898.256∗E − 38.6623∗PW + 6.85498∗SR1 + 2.95882∗SR2–16.5802∗T1–77.5504∗PTQ1	98.47
GY = −738.307 + 118.975∗E − 48.5922∗PW + 1.87334∗SR1 + 2.62473∗SR2–35.251∗T1–14.4497∗PTQ2	95.32
GY = −1178.05 + 899.602∗E + 6.8455∗SR1 + 2.96831∗SR2–31.9094∗T1 + 13.5518∗T2–74.1903∗PTQ1	98.46
GY = −3028.67 + 21.7114∗E + 1.86749∗SR1 + 1.38125∗SR2–34.1386∗T1 + 61.0605∗T2 + 2.95928∗PTQ2	95.80

(continued)

Table 3.23 (continued)

Regression models	
GY = −3486.29 + 92.9008∗E + 1.41235∗SR2–12.406∗T1 + 59.2894∗T2–13.9926∗PTQ1 + 25.9524∗PTQ2	95.38
GY = 594.397–196.044∗PW + 4.63831∗SR1 + 1.66246∗SR2 + 23.0033∗T1 + 72.8577∗T2–53.4338∗PTQ1	97.91
GY = −2481.28 − 29.6402∗PW + 1.8767∗SR1 + 1.58055∗SR2–29.1928∗T1 + 55.9794∗T2–1.28573∗PTQ2	95.84
GY = −2227.94 − 68.9646∗PW + 1.75248∗SR2 + 0.0719391∗T1 + 55.4433∗T2–16.1384∗PTQ1 + 18.4592∗PTQ2	95.50
GY = −2930.65 + 3.35027∗SR1 + 0.709417∗SR2–23.3052∗T1 + 98.4384∗T2–30.3258∗PTQ1 + 17.4223∗PTQ2	96.86
Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6 + β7X7	R^2
GY = −522.117 + 734.207∗E − 62.8549∗PW + 6.59519∗SR1 + 2.7072∗SR2–12.0194∗T1 + 22.7315∗T2–74.0542∗PTQ1	99.53
GY = −1202.83 − 216.846∗E − 100.018∗PW + 1.73874∗SR1 + 1.72226∗SR2–18.7737∗T1 + 66.5448∗T2–8.24543∗PTQ2	96.90
GY = 905.486–202.513∗PW + 4.70728∗SR1 + 1.88784∗SR2 + 20.0973∗T1 + 69.7402∗T2–53.5304∗PTQ1–3.62754∗PTQ2	98.92
Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6 + β7X7 + β8X8	R^2
GY = −3670.28 + 1097.09∗E + 56.959∗PW + 6.98681∗SR1 + 1.34258∗SR2–5.07598∗T1 + 23.9757∗T2–83.4396∗PTQ1 + 30.2752∗PTQ2	99.78

$$r = \frac{\sum\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\right)}{\sqrt{\sum\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^2 \sum\left(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\right)^2}}$$

Correlation coefficient ranges from +1 to −1. If $r = +1$, then it shows positive covariance, while if $r = -1$, it means negative correlation, and if $r = 0$, it means no correlation at all. Correlation measures co-relation a joint property of two variables, while regression deals with the change of one variable in relation to change of another variable. In correlation, random pair of observation was obtained, while in regression, only the dependent variable needs to be randomly and normally distributed. The application of concept of correlation has been illustrated in Fig. 3.13 (Ahmed 2011).

3.13 Analytical Tools/Software

Analytical tools which can be used for the statistical analysis are listed below:

1. R
2. SAS
3. Sigma plot

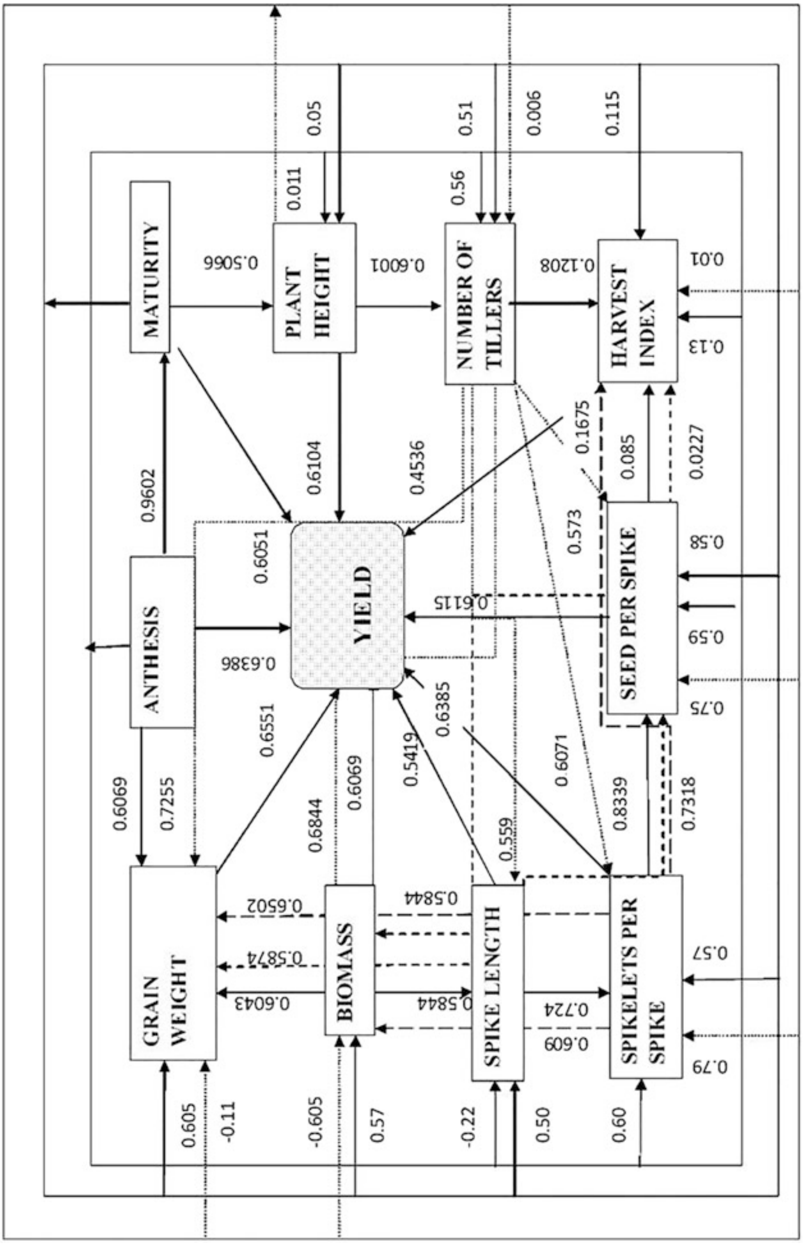


Fig. 3.13 Correlation analysis between spring wheat yield and yield components

4. Stat graphics
5. Minitab
6. SPSS
7. MS Excel
8. MATLAB
9. GraphPad Prism
10. GenStat
11. SigmaStat
12. Stata
13. Statistica

References

Acutis M, Scaglia B, Confalonieri R (2012) Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. Eur J Agron 43:129–135. https://doi.org/10.1016/j.eja.2012.06.006

Ahmed M (2011) Climatic resilience of wheat using simulation modeling in Pothwar. PhD thesis. Arid Agriculture University, Rawalpindi

Ahmed M, Hassan FU, Aslam MA, Akram MN, Akmal M (2011) Regression model for the study of sole and cumulative effect of temperature and solar radiation on wheat yield. Afr J Biotechnol 10(45):9114–9121. https://doi.org/10.5897/AJB11.1318

Ahmed K, Shabbir G, Ahmed M, Shah KN (2020) Phenotyping for drought resistance in bread wheat using physiological and biochemical traits. Sci Total Environ 729:139082. https://doi.org/10.1016/j.scitotenv.2020.139082

Bennington CC, Thayne WV (1994) Use and misuse of mixed model analysis of variance in ecological studies. Ecology 75(3):717–722. https://doi.org/10.2307/1941729

Blouin DC, Webster EP, Bond JA (2011) On the analysis of combined experiments. Weed Technol 25(1):165–169

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24(3):127–135. https://doi.org/10.1016/j.tree.2008.10.008

Das MN, Giri NC (1979) Design and analysis of experiments. Wiley Eastern, New Delhi, 295 p

Fisher RA (1921) Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. J Agric Sci 11(2):107–135. https://doi.org/10.1017/S0021859600003750

Gbur EE, Stroup WW, KS MC, Durham S, Young LJ, Christman M, West M, Kramer M (eds) (2012) Analysis of generalized linear mixed models in the agricultural and natural resources sciences. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison. https://doi.org/10.2134/2012.generalized-linear-mixed-models.frontmatter

Gomez KA, Gomez AA (1984) Statistical procedures for agricultural research. Wiley, New York, 680 p

Lawson J (2010) Design and analysis of experiments with SAS. Chapman and Hall/CRC

Lencina VB, Singer JM, Stanek EJ III (2005) Much ado about nothing: the mixed models controversy revisited. Int Stat Rev 73(1):9–20. https://doi.org/10.1111/j.1751-5823.2005.tb00248.x

Loughin TM (2006) Improved experimental design and analysis for long-term experiments this work was done while the author was on faculty in the Department of Statistics at Kansas State University. Crop Sci 46(6):2492–2502. https://doi.org/10.2135/cropsci2006.04.0271

McIntosh MS (1983) Analysis of combined Experiments1. Agron J 75(1):153–155. https://doi.org/10.2134/agronj1983.00021962007500010041x

McIntosh MS (2015) Can analysis of variance be more significant? Agron J 107(2):706–717. https://doi.org/10.2134/agronj14.0177

McNutt M (2014) Raising the bar. Science 345(6192):9–9. https://doi.org/10.1126/science.1257891

Mead R (2017) Statistical methods in agriculture and experimental biology. Chapman and Hall/CRC

Moore KJ, Dixon PM (2015) Analysis of combined experiments revisited. Agron J 107 (2):763–771. https://doi.org/10.2134/agronj13.0485

Nature Publishing Group (2005) Statistically significant. Nat Med 11(1):1–1. https://doi.org/10.1038/nm0105-1

Nature Publishing Group (2013a) Nature. Medicine 19(5):508–508. https://doi.org/10.1038/nm0513-508

Nature Publishing Group (2013b) Reporting life sciences research. Nature Publishing Group, London. http://www.nature.com/authors/policies/reporting.pdf

Nelder JA (2008) What is the mixed-models controversy? Int Stat Rev 76(1):134–135. https://doi.org/10.1111/j.1751-5823.2007.00022_1.x

Nelder JA, Lane PW (1995) The computer analysis of factorial experiments: in memoriam—Frank Yates. Am Stat 49(4):382–385. https://doi.org/10.1080/00031305.1995.10476189

Snedecor GW (1942) The use of tests of significance in an agricultural experiment station. J Am Stat Assoc 37(219):383–386. https://doi.org/10.2307/2279007

Steel R, Torrie J (1980) Principles and procedures of statistics, 2nd edn. McGraw-Hill Book Co., New York

Voss DT (1999) Resolving the mixed models controversy. Am Stat 53(4):352–356. https://doi.org/10.2307/2686056

Wang T, DeVogel N (2019) A revisit to two-way factorial ANOVA with mixed effects and interactions. Commun Stat Theory Method:1–18. https://doi.org/10.1080/03610926.2019.1604961

West BT, Galecki AT (2012) An overview of current software procedures for fitting linear mixed models. Am Stat 65(4):274–282. https://doi.org/10.1198/tas.2011.11077

Yang RC (2010) Towards understanding and use of mixed-model analysis of agricultural experiments. Can J Plant Sci 90(5):605–627. https://doi.org/10.4141/CJPS10049