Review Paper

A Review on Statistical Tools for Genetic Diversity of Crop Improvement

Fareeha Zafar[1], Aamar Mumtaz[1], *Saif-ul-Malook[1], Muhammad Ubaidullah Aleem[1]

[1]Department of Plant Breeding and Genetics, University of Agriculture Faisalabad
Corresponding author email: aamer3002@.gmail.com saifulmalookpbg@gmail.com

Abstract: Knowledge of genetic diversity and their relationship is very essential for improvement of crops. A large number of statistical tools are available for ascertaining genetic diversity of crops. These tools rely on climate conditions, number of treatments, pedigree data, morphological data and agronomic performance data. For accuracy and un-biasness in estimation of genetic diversity, attention has been needed to sampling strategies, utilization of data on the basis of their strength and weakness and objective determination of genetic relationship. In this review statistical tools commonly used in plant breeding are discussed, also statistics, its use and types are discussed. Statistical tools correlation, factorial, split plot (CRD, RCBD, Split plot) and nested design are discussed.
[Fareeha Zafar, Aamar Mumtaz, Saif-ul-Malook, Muhammad Ubaidullah Aleem. **A Review on Statistical Tools for Genetic Diversity of Crop Improvement.** *Nat Sci* 2015;13(2):83-87]. (ISSN: 1545-0740). http://www.sciencepub.net/nature. 12

Keywords: Correlation, Factorial, Nested design, Split plot

Introduction of Statistics:

Statistics derived from Latin word status.mean a political state. Its original meant useful information to state. So the word statistic mean numerical data systematically arranged.

Some people said that statistic is an important tool that deals with percentages, charts, graph, tables and averages. Some people said that statistics is tools that deal with rules, methods and techniques to research a large number of numerical data. Where other peoples said that statistic making inferences about population on the basis of sample data.

Statistics is an important tool for all types of research and its principle are independent of all subject matte and their procedure are successfully used including Agriculture and biological science. The validity of experimental result depend on accurate data collection method.so statistical tools has important value for plant breeder in breeding programmed (Muhammad, 2014).

Definition of statistics is that it is an important method deal with collection of data, organizing, summarizing, presenting and analysis data, drawing valid conclusion and making final decisions on the basis of that analysis. So these tools used to any experiment for its calculation and observation (Muhammad, 2014).

There are many reasons to use of statistic information such as: Explain things what happened with it, to inform general information, for general comparison, to estimate unknown quantities, for association between factors, regarding future outcomes and predict decision (Acquaah, 2007).

Statistic divided into two types, first is descriptive and second is inferential.

Descriptive method deals with summarizing and description of important numerical data and indicates the spread of observation.

Inferential methods used to describe large group of data as whole called population from as a part of data called sample (Muhammad, 2014).

The world is facing food shortage leading to poverty especially in developing countries due to increasing population. It is a major challenge for breeders to tackle these problems. Improvements in today's crops are mandatory for fulfilling the food requirements of world. Therefore a crop improvements technique remains a major concern for plant breeders in every age. There are some important statistical tools which are used for genetic diversity of crop improvement. For specific and general environment performance several factors affect crop improvements such as weather, soil, climate, biological, edaphic and crop genotype. Crop genotype is the most important factor (Aremu, et al, 2007). Genotypes are composed of different forms including pure and inbred lines hybrids, wildraces, landraces germplasm accessions, cultivars and varieties. Genetic diversity is the diverse and wide origin and genetic background of these crops (Aremu, 2005 and Abbas *et al.*, 2014).

Genetic diversity study is major requirement for success in plan breeding and crop improvement. It is a step wise processes by which existing variation in present germplasm are identified by use of different statistical methods and their combinations (Aremu, 2005). Genetic diversity provides information about quantum of genetic divergence and specific breeding objectives can be achieved through its help (Thompson et al, 1998).

Several researchers including breeders designed different statistical tools for estimating genetic diversity such as correlation, factorial experiment, split plot design, regression, analysis of variance (ANOVA), nested classification. Details of these tools are discussed below.

1. Correlation

Correlation is used to find degree and direction of relationship between two or more variable.so it is also correlate mutual relationship between two or more variables. It is represented by 'r' (Bahmankar *et al.* 2014 and Sarfaraz *et al.*, 2014),

Three types of correlation: First is simple correlation (association between two variables.it is also three type (Bahmankar *et al.* 2014) (a) Phenotypic which is directly observed. It is estimated from phenotypic variance and co variance (b) Genotypic**:** It is inherit and heritable association between two variable. It is estimated from genotypic variance and co variance (c) Environmental**:** It is because of environmental effects and estimated from error variance and co variance.

Second is partial correlation: It is estimation of correlation between two variables by effect of third variable.it is denoted by r12.3.

Third is multiple correlations: It is the correlation of two or more independent variable on dependent variable.it is represented by R.

There are some limitations of correlation studies such as sometime it provides misleading result and sometime correlations become zero between two variables having linear relationship (Bahmankar *et al.* 2014)**.**

2. Factorial experiment

It is the effect of two or more factor. Theses factor are investigated for all possible combination using experimental design. Factorial is a treatment combination not itself experimental design.in this experiment we estimated single factor. Two way factor and three way factor experiments.in single factor experiment factorial set of treatment included. These treatments are of all possible combination from different factors (Bose, 2003).

For example:

Effect of different doses of 4 potassium level on grain yield of wheat, where we compare 3 different varieties wheat. Fertilizer and wheat are called factors and their response variable is yield of wheat (Muhammad 2014).

In single factor experiment: We estimate two separated single experiments one for potassium and 2nd for variety so 4*3=12 treatments. These 12 treatments consist of four level of potassium and three level of variety of wheat crop.

Twelve homogenous experiment plot where avaible.now we apply CRD experiment to estimation of single factorial. If replications are available in given data then we used RCBD. For example in wheat yield 4 replications are available for each variety. so 3*4*4=48. In RCBD experiments we make sub table interaction of variety and fertilizer.

In two way factor experiment: In two way factorial we used two then more variable interaction. For example: 2 fertilizers apply on wheat grain like 4 nitrogen and 4 Zn level for 3 variety of wheat with 4 replication. Both experiment used like CRD without replication and RCBD with replication but mostly used is RCBD. Where we make ZN*N sub-table for their interaction.

In three way factorial: we used more than two variable as well as two way factorial. We make sub tables like V*N, V*Zn, Zn*N and V*ZN*N.

There are three type of effect (Braver, 2003), Simple effect**,** main effect and interaction effect.

Simple effect is the difference between two factors for a certain level of other to measure variations. Main effect is the average effect on simple effect of factor to measure variation among various level of factors and interaction effect is the interaction between different levels of factors to other to measure variations.

It is used to measure relationship among several factors to fine presence and interaction of variables and in any experiment where recombination over a wide range of varieties is used.

3. Split Plot design

Split block design is special kind of incomplete block design used where CRD, RCBD and Latin square is unable to estimate all possible of combination in experimental data.

Split plot design mainly used for factorial experiment in which more design are underlying to measure variations. by this use of method plot are sub divided into sub plots where some or additional factors are applied.

So, the split plot is an experimental design where level of one factor can be applied to large units and the level of other factor to subunits. one factor are assigned to main plot and other factor of subplots assigned within each main plots (Snedecor, 1967 and Mustafa *et al*., 2014ab).

There are three Situations of usage of split plot design:

Firstly it is used when one or more factor requires which have large amount of experimental material in an experimental unit associated to treatments then do other treatments to other factors in experimental units.

Secondly it is used sometimes for additional factors to increase scope.

Thirdly if large number of difference in experiments can be expected among level of factor

and among others (Muhammad, 2014 and Amin *et al*., 2014ab).

For example:

A researcher was interested to compare 3 varieties of rice at 4 seeding rates.an experiments was conducted in RCBD with 4 replication each and with 3 varieties in main plot and4 seeding rates in subplot for plant height at maturity.

4*4*3=48 total number of observation. We make interaction of these factors like S*V, V*R Sub tables, error 1 and error2.

In this experiment addition of restriction on randomization in sub plots makes its necessary to calculate a third error term for the effect of third factor and interaction involved when third factor introduce into subplot by splitting further sub plots of the factor in sub-sub plots.

For example:

It is same as split plot design but their additional third factor of sub plot divided into sub-sub plot equal to level of third factor. Suppose a third factor diseases control with two level is added with 3 variety and 3 fertilazer.We make a model like V*S, V*R, V*N,N*S and V*N*S for error1, error2 and error3 (Muhammad, 2014).

4. Regression:

Regression analysis is a statistical procedure to measure relationships among variables. IN this method has many techniques to analyze and modeling of several variables. Main focus is on the relationship between a dependent variable and one or more independent variables (Acquaah, 2007).

There are ten types of regression for analysis:

(1)Linear regression model is that in which one or more then explanatory variables are used, (2) simple linear regression is that in which single explanatory variable is used, (3) logistic regression is that in which one or more then predictor variables are used, (4) nonlinear regression is depend on one or more independent variables, (5) nonparametric regression is without predictor variable, (6) robust regression is the relationship between dependent and independent variable, (7) stepwise regression is the choice of predictor variable, (8) regression toward the mean is common procedure to use, (9) software regression is the appearance of a bug which is absent in a previous information (10) regression testing is the testing method to find out uncover regression bugs by software (Acquaah, 2007).

5. Analysis of variance

ANOVA is statistical model used to analyze variation among and between groups.

ANOVA is used to estimate genotypic variance, phenotypic variance and environmental variance by use of experimental designs CRD, Latin square design and RCBD (Braver, 2007).

CRD is the experimental design used to measure variation of single or primary factor without need to other. This design used to compare response variable at the level of that primary factor (Acquaah, 2007).

Sov	d.f	s.s	m.s	f.cal	f.tab
Treatment	-	-	-	-	-
Error	-	-	-	-	-
Total	-	-	-	-	-

RCBD is a standard experimental design used to measure any variation where each treatments assigned at random to each block of field data.no same treatment in each block (Acquaah 2007, Mumtaz *et al*., 2014 and Malook *et al*., 2014abcd).

Sov	d.f	s.s	m.s	f.cal	f.tab
Replication	-	-	-	-	-
Treatment	-	-	-	-	-
Error	-	-	-	-	-
Total	-	-	-	-	-

Latin square design: is a design to control and eliminate two source of variation.it is efficient design should not be much used as CRD and RCBD.it have more than one block to allow control two source of variation from error of variation (Acquaah, 2007).

Cross-over design: study to exposure sequence of different treatments. Crossover design is observational design and controlled design. This method used to estimation of parallel and longitudinal sequence. But has two problems like carry over and order effects.

Analysis of co-variance: ANCOVA is type of ANOVA is used to control for potential confounding variables ANCOVA one constant variable in it in which one or more factors present (Kenedal, 1965).

Some assumption of ANCOVA is

(1) Normality of residuals should be normally distributed), (2) homogeneity of variances: the error variances is equal for different treatments, (3) homogeneity of Regression Slopes: The slopes of the different regression lines equivalent, (4) linearity of regression: the regression relationship between the dependent variable and liner variable, (5) Independence of error terms: Uncorrelated (Kendall, 1965).

ANCOVA of XY

SOV	D.F	S.S	M.S	F.cal	F.tab
Replication	-	-	-	-	-
cm.lines	-	-	-	-	-
error	-	-	-	-	-

Incomplete block design: In this experiment should not be all treatment to each block. We uses notation‘t’=no of treatments and define size of block

denoted by 'k' where k is less than **t (k < t).'b'** for number of block size and **'r'** is replication for treatments. Total number of observation is '**N**'.
N=t(r) =b (k).

Block

	1	2	3	4
1	A	A	A	-
2	A	A	-	A
3	A	-	A	A
4	-	A	A	A

Treatments
Example2: b=4.t=4 and r=2.

Block

	1	2	3	4
1	A	-	A	-
2	-	A	-	A
3	A	-	A	-
4	-	A	-	A

Treatments
For example: incidence matrix (where b=4,t=4 and r=3.

6. Nested classification

Nested designs are that each subject has a single score its mean each subject reach only one treatment. In some two way factorial, the level of one factor 'W' is not cross or classified with the other factors 'T', but is "NESTED" with it. The levels of W are different for different levels of T (Muhammad, 2014).

For example: 2 Areas, 4 sites per area, each with 5 replicates where no link from sites on one area to sites on another area. That is, there are 8 sites, not 2.

S1 T	S2 T	S3 T	S4T
Study area T			
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A

S5 W	S6 W	S7 W	S8 W
Control area W			
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A

Where A=replications.

Number of sites (S)/replications is not being equal with each site. Now we using a nested ANOVA not a two-way ANOVA.

Objectives of nested classifications are (1) In this design nested show difference between variability of area and sites and 2nd is showing difference between study and control. (2) If we fail to find variability of site and area then we suggest that is environmental impacts. (3) Difference is only due to area not by sites (Mohammadi and Prasanna, 2003).

Probability distribution method assign to random variables. It depend on two steps: First is probability is always between 0 and 1 and secondly sum of probability of outcome is always 1.

Distribution is the value of random variables can assume their probability make its distribution (Muhammad, 2014).

For example:

X	1	2	3	4	5	6	7	8	sum
P(x)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8	8/8=1

Variance, mean and standard deviation:

In which mean and variance with ungrouped frequency:

Formulas of mean and variance,, where mean denoted by 'μ' and variance is denoted by 'σ^{2}' and standard deviation is $\sqrt{\sigma^2}$ (Mohammadi and Prasanna, 2003).

Mean:
$\sum x/N$ **and**
Variance:
$\sum fx^2-(\sum xf)^2/N/n-1$.
If probability then these formulas is:
Mean: $\mu=\sum x\ p(x)$ and
Variance is $[\sum x^2\ p(x)-(\sum\ x\ p(x)^{2]}$

X	1	2	3	4	5	6	7	8	Sum
P(x)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8	8/8=1
x p(x)	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8	36/8=4.5
x^2p(x)	1/8	4/8	9/8	16/8	25/8	36/8	49/8	64/8	204/8=25.5

Mean: 9/2=4.5
Variance =$25.5-(3.5)^2$ /2/1=19.4
Standard deviation: √variance = √19.4= 4

6. Conclusion:

Statistics is an important science which is very helpful for analyzing data. Statistical tools are very important for any experiment. Each tool has own importance as for estimating degree and nature of relationship correlation is used. If two or more factors of same importance are studied then for estimating their relationship factorial experiment is used. If two or more factor with varying importance are studied then split plot design are used. If relationship has to estimate between dependent and other independent variables then regression analysis is used. If experiment has to done in controlled environment then CRD is used and if two factors are studied in field with one way of variability RCBD is used and if two factors are studied in field with two way of variability then split plot is used. In short every design has his own importance and needed to be studied.

References:

1. Abass H. G., A. Mahmood, Q. Ali, Saif-ul-Malook, M. Waseem and N.H. Khan. 2014. Genetic variability for yield, its components and quality traits in upland cotton (Gossypium hirsutum L.) Nature and Science, 12: 31-35.
2. Acquaah G. (2007). Common statistical methods in plant breeding. IN Principles of Plant Genetics and Breeding. Blackwell Publishing, Australia. pp. 246.
3. Amin W., Saif-ul-malook, S. ashraf and Amir Bibi. 2014b. A review of screening and conventional breeding under different seed priming conditions in sunflower (*Helianthus annus* L.) Nature and Science, 12: 23- 37.
4. Amin,W., Saif-ul-malook, A. Mumtaz, S. ashraf, H. M. ahmad, K. Hafeez[1], M. Sajjad and A. Bibi. 2014a. Combining ability analysis and effect of seed priming on seedling traits in Sunflower (Helianthus annus). Report and Opinion, 6: 19-30.
5. Aremu C.O. 2005. Diversity selection and genotypes Environment interaction in cowpea unpublished Ph.D Thesis. University of Agriculture, Abeokuta, Nigeria. P. 210.
6. Aremu, C.O., Adebayo, M.A., Ariyo O.J., and Adewale B.D. 2007. Classification of genetic diversity and choice of parents for hydridization in cowpea vigna unguiculata (L) walip for humid savanna ecology. African J biotechnol. 6(20): 2333-2339.
7. Bose M. (2003). The Mathematics of Symmetrical Factorial Designs. Resonance. 10:14-19.
8. Braver, S.L., & MacKinnon, D.P. (2003). Levine's guide to SPSS for analysis of variance (2nd Edition). Mahway, NJ: Erlbaum.
9. Kendall, M.G. (1965). A course in multivariate analysis. Charles Griffin & Co., London.
10. Mohammadi S.A. and Prasanna, B.M. (2003). Analysis of Genetic Diversity in Crop Plants Salient Statistical Tools and Considerations Crop Sci. 23: 1235-1248.
11. Muhammad F. (2014). Statistical method data analysis. Ed (2014). Kitab markaz, Lahore, Pakistan.
12. Mumtaz A., Sadaqat, H.A., Saifulmalook, Nazik, A.S. and Ahmad, H.M. 2014. Genetic Behavior of Quality Related Traits in *Brassica rapa* L. Vegetos 27(3): 139-145.
13. Mustafa G., Ehsanullah, Saif-ul-Malook, E. Ahmad,M. Sarfaraz, S. A. Qaisarani andM. K. Shahbaz. 2014. Yield attributes and productivity of various Bt and non Bt cotton varieties in Faisalabad environment. Nature and Science, 12(11): 92-103.
14. Mustafa G., Ehsanullah, Saif-ul-Malook, M. Sarfaraz, M.K. Shahbaz, U.Chopra and Q. Ali.2014. A review of production for various Bt and non Bt cotton varieties in Pakistan. Nature and Science, 12: 81-91.
15. Saif-ul-malook, M. Ahsan and Q. Ali. 2014a. Genetic Variability and Correlation Studies among Morphological Traits of *Zea mays* under Normal and Water Stress Conditions. Persian Gulf Crop Protection. 3(4): 15-24.
16. Saif-ul-malook, M. Ahsan, Q. Ali and A. mumtaz. 2014b. Genetic variability of maize genotypes under water stress and normal conditions. Researcher, 6: 31 – 37.
17. Saif-ul-malook, M. Ahsan, Q. Ali, A. mumtaz. 2014c. Inheritance of yield related traits in maize under normal and drought condition. Nature and Science, 12: 36 – 49.
18. Saif-ul-malook, Qurban ali, Muhammad Ahsan, Aamer Mumtaz and Muhammad Sajjad. 2014d. An overview of conventional breeding for drought tolerance in *Zea mays*. Nature and Science, 12: 7-22.
19. Sarfaraz M., Wasi-Ud-Din, Muhammad Sajjad, Muhammad Wajid, Saif-ul-Malook, Muhammad Khalid Shabaz, Hafiz Mahboob Ahamed and HafizSalman Saeed. 2014. Effect of Planting Time and Nitrogen Levels on various yield Components of Sunflower (*Helianthus annus* L.). Nature and Science, 12(12): 19-28.
20. Snedecor, G.W., and Cochran W.G. (1967). Statistical methods,6th edt. Iowa State University, Ames, IA.
21. Thompson, J.A., Nelson, R.L. and Vodkin, L.O. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci. 38: 1348-1355.

2/8/2015