Jitendra Kumar · Aditya Pratap
Shiv Kumar *Editors*

Phenomics in Crop Plants: Trends, Options and Limitations

Springer

Phenomics in Crop Plants: Trends, Options and Limitations

Jitendra Kumar • Aditya Pratap
Shiv Kumar
Editors

Phenomics in Crop Plants: Trends, Options and Limitations

Springer

Editors
Jitendra Kumar
Division of Crop Improvement
ICAR – Indian Institute of Pulses Research
Kalyanpur, Kanpur, Uttar Pradesh, India

Aditya Pratap
Division of Crop Improvement
ICAR – Indian Institute of Pulses Research
Kalyanpur, Kanpur, Uttar Pradesh, India

Shiv Kumar
International Center for Agricultural Research in the Dry Areas
Rabat – Instituts, Rabat, Morocco

ISBN 978-81-322-2225-5 ISBN 978-81-322-2226-2 (eBook)
DOI 10.1007/978-81-322-2226-2

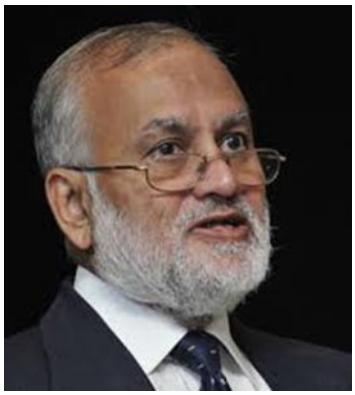
Library of Congress Control Number: 2015931204

Springer New Delhi Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer (India) Pvt. Ltd. is part of Springer Science+Business Media (www.springer.com)

Foreword



Growing world population is expected to cause a "perfect storm" of food, energy and water shortages by 2030 as demand for food and energy will jump by 50 % and for fresh water by 30 %, as the population tops at 8.3 billion. The overarching challenge before the policy makers and agricultural scientists is how to ensure food and nutrition security for an ever-increasing population from limited and fast depleting resources under climate change scenario, especially in countries like India where sizeable population is still suffering from the triple burden of malnutrition. To meet the future demand of agricultural production, we need to develop more productive and nutritious varieties of agricultural crops which incorporate both high intrinsic yield potential and resilience under climatic stresses. This requires discovery and deployment of superior but complex traits from the vast germplasm resources being held in various gene banks to agronomically superior varieties efficiently and precisely.

Traits of breeders' interest such as grain yield, plant growth and resistance to biotic and abiotic stresses are complex as these are controlled by many genes of minor effects and highly influenced by environmental factors and their multi-dimensional interactions. In the past, plant breeders were successful in selecting desirable varieties empirically on the basis of visual observations, more so for qualitative traits; but empirical selection remains elusive and less effective for traits essential for meeting the current challenges such as underground, physiological and biochemical traits. Recent advances in genomics have created enormous genomic resources in several crop species which have the potential to increase harvestable yield

manifolds. However, available gene sequences and molecular markers could not be mainstreamed in crop improvement programs mainly due to the lack of precise phenotyping data. Therefore, it is imperative to phenotype the available germplasm precisely and efficiently in various crop species.

The current knowledge and voluminous information generated on phenotyping tools and techniques available in literature need to be consolidated so that researchers and scholars have access to such vast knowledge at one place. The present book, *Phenomics of Crop Plants: Trends, Options and Limitations*, which is a meticulously edited volume, is an attempt in this direction to bring together information on precision phenotyping under controlled versus natural environments, digital and image based phenotyping, phenomics of biotic and abiotic stresses and functional traits, and precision nutrient management. This book also covers experimental designs and biometrical approaches suitable for precision phenotyping of complex traits, and how phenomics can help to harness potentiality of genomics. Various chapters in this book have been contributed by renowned scientists whose research contributions are acknowledged globally. I am quite hopeful that the information contained in this book will boost research efforts of plant scientists to bring about a major breakthrough in agricultural production and will serve as a resource material for those who are involved in teaching and research in agricultural crops. I congratulate the editors Jitendra Kumar, Aditya Pratap and Shiv Kumar for bringing out this book timely on such an important and emerging aspect and hope that it would be widely read by scholars and researchers.

Secretary, DARE and Director General, ICAR S. Ayyappan
Krishi Bhavan, New Delhi, India

Preface

It has been estimated that agricultural production must be doubled by 2050 in order to meet the predicted demand of growing world population. Achieving this goal poses a serious challenge to plant breeders as the current agricultural production growth rate of 1.3 % per annum is below the population growth rate. In the recent past, research advances have been made in the development of genomic tools and techniques which have the potential to increase the rate of genetic improvement. The whole genome studies have the potential to greatly facilitate genetic dissection of complex traits such as yield and stress tolerance by using technological advances in genotyping and sequencing. However, successful application of genomics towards the genetic improvement of crop plants depends upon our ability of precision phenotyping of these complex traits. Low cost and high-throughput genotyping has paved the way for the development of large mapping populations and diversity panels of thousands of recombinant inbred lines. These genetic resources require precise phenotyping. Marker-assisted recurrent selection (MARS) and genome-wide selection require phenotypic data, although conceptually selections are made on the basis of genetic information. A single phenotyping cycle is used to identify markers for subsequent selection through generations. In transgenic studies also, phenotyping is necessary for identification of promising events. Molecular breeding populations sometimes include up to 5,000 lines and their accurate characterization simultaneously is a challenging task. Also phenotyping of such complex traits are labor intensive, and many other interesting traits involved in biological processes are currently not suitable for genetic mapping due to the lack of approach to efficient and reliable measurement. The success in development of improved varieties relies on the ability to identify the best genetic variation for advancement. Because breeding is essentially a numbers game, more crosses and environments are required to identify superior variation with greater probability. Therefore, plant breeders want to phenotype a large number of lines rapidly and accurately to identify the best progeny. Advances in phenotyping are essential to capitalize on the developments in conventional, molecular, and transgenic breeding and ensure genetic improvement of crops for future food security.

In recent years, there has been increased interest in development of high-throughput phenotyping tools and techniques for screening of agronomic, physiological, and biochemical traits expressing especially under biotic and

abiotic stresses. These techniques have become much more advanced and have now entered the era of high-throughput field phenotyping. Several phenotyping platforms have been developed around the world, which are fully automated facilities in greenhouses or growth chambers with robotics, precise environmental control, and remote sensing techniques to assess plant growth and performance. Consequently, voluminous literature has been generated on different aspects of phenotyping which is scattered in numerous journals and books. However, no single publication is available to provide a comprehensive insight into this literature with a focus on phenomics of crop plants. This book, *Phenomics of Crop Plants: Trends, Options and Limitations*, is an attempt in this direction to bring together various high throughput, advanced phenotyping tools, techniques and platforms for directed genetic improvement in crop plants.

The present book comprises 19 chapters contributed by renowned scientists in their fields of expertise. The first chapter presents an overview on the recent developments in phenotyping. The second chapter deals with traits that require precise phenotyping. Chapter 3 discusses various issues related to phenotyping under controlled and natural environments while the subsequent three chapters (Chaps. 4, 5, and 6) deal with the imaging tools in phenotyping agronomic and physiological traits in crop plants. Chapters 7, 8, and 9 focus on phenotyping tools available for heat and drought related traits and soil problems. Chapter 10 deals with screening methods for diseases and possibility of using the recent developments in the field of phenomics. The subsequent three chapters (Chaps. 11, 12, and 13) discuss the advances in phenotyping of functional traits, role of florescence approaches for understanding the functional traits of photosynthesis and use of NMR in identification of subcellular structural and metabolic challenges. The next two chapters are on precision nutrient management and identification of nutritional and anti-nutritional factors of seeds (Chaps. 14 and 15). The subsequent two chapters (Chaps. 16 and 17) discuss the role of experimental designs for precision phenotyping and use of biometrical approaches in data analysis of the complex traits. As vast amount of genomic resources are now available in several crop plants, precision phenotyping can harness the potentiality of these genomic resources for accelerating the genetic improvement through mainstreaming them in the ongoing breeding programs. Therefore, the next two chapters (Chaps. 18 and 19) deal with how the available genomic resources can be utilized in a better way by using the available phenomics platforms worldwide for precise phenotyping of agronomic and physiological traits. Each chapter of this book has focused on the current trends, available options for phenotyping the target traits and limitations in their use for phenomics of crop plants.

The review of entire published work was neither possible in a single volume nor was the aim of this book. However, the contributors of individual chapters have provided exhaustive list of references on significant work done so far on different aspects of phenomics. Keeping in view the scope of the book, a little overlap in the subject is possible albeit all chapters have been dealt in depth by various experts. We are extremely grateful to all the authors

who despite being busy with their research and academics completed their chapters with a professional approach and great care.

We are highly indebted to Dr. S. Ayyappan, Secretary, Department of Agricultural Research and Education (DARE), Government of India, and Director General, Indian Council of Agricultural research (ICAR); and Dr. Mahmoud Solh, Director General, International Centre for Agricultural Research in the Dry Areas (ICARDA) for encouragement and inspiration in bringing out this publication. We are also thankful to Prof. Swapan Datta, Deputy Director General (Crop Science), ICAR; Dr. Maarten van Ginkel, Deputy Director General (Research), ICARDA; Dr. Michael Baum, Director of BIGM, ICARDA and Dr. B. B. Singh, Assistant Director General (Oilseed and Pulses), ICAR, for providing support and state-of-the-art facilities to carry out research on pulses. Dr. N. P. Singh, the present Director and Dr. N. Nadarajan, Ex-Director of IIPR, Dr. S. K. Chaturvedi, Head, Crop Improvement Division, IIPR, have been the source of encouragement for the present endeavor. Several people have rendered invaluable help in bringing this publication to life and they deserve our heartfelt appreciation and gratitude: Dr. Sanjeev Gupta, Project Coordinator, MULLaRP, IIPR, for technical comments and scientists of Crop Improvement Division, IIPR, for their valuable technical inputs during the course of editing the chapters; Mr. Ramesh Chandra, Senior Technical Assistant; Mr. Rohit Kant, Miss Nupur Malviya and Rakhi Tomar, Senior Research Fellows, for helping in compilation of references and voluminous correspondence, and Springer International for bringing the book through printing process with a thorough professional approach. Last but not least, our kids Neha, Gun and Puranjay and our better halves, Mrs. Renu Rani, Dr. Rakhi Gupta and Dr. Pankaj Rani Agrawal, deserve special thanks for their unstinting help, patience and emotional support during the course of this book.

Kanpur, Uttar Pradesh, India — Jitendra Kumar
Kanpur, Uttar Pradesh, India — Aditya Pratap
Rabat, Morocco — Shiv Kumar

Contents

Contributors

Partha Sarathi Basu Division of Basic Sciences, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Vanya Bawa Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Lucas Busemeyer COALA – Competence Center of Applied Agricultural Engineering, University of Applied Sciences Osnabrück, Osnabrück, Germany

Christopher L. Butts USDA-ARS National Peanut Research Laboratory, Dawson, GA, USA

Charles Y. Chen Department of Crop, Soil and Environmental Sciences, Auburn University, Auburn, AL, USA

Ying Long Chen Soil Science and Plant Nutrition, School of Earth and Environment, and The UWA Institute of Agriculture, The University of Western Australia, Crawley, WA, Australia

Neeraj Choudhary Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Phat M. Dang USDA-ARS National Peanut Research Laboratory, Dawson, GA, USA

Hannah Dee Department of Computer Science, Aberystwyth University, Aberystwyth, UK

Rekha Sapru Dhar Plant Biotechnology, Indian Institute of Integrative Medicine Research, Jammu, Jammu and Kashmir, India

Ivica Djalovic Department for Maize, Institute of Field and Vegetable Crops, Novi Sad, Serbia

Khaled El-Shama'a International Center for Agricultural Research in the Dry Areas (ICARDA), Amman, Jordan

Priyanka Gupta International Centre for Agricultural Research in the Dry Areas, Rabat – Instituts, Rabat, Morocco

Jerry L. Hatfield National Laboratory for Agriculture and the Environment, Ames, IA, USA

Shalu Jain Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

Henk Jalink Wageningen Campus, Greenhouse Horticulture, Wageningen, The Netherlands

Casey R. Johnson Food Chemistry and Analysis, Nutrient Interactions, North Dakota State University, Fargo, ND, USA

Rohit Kant Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Pradeep Kumar Katiyar Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Seid Ahmed Kemal International Center for Agricultural Research in the Dry Areas (ICARDA) – Ethiopia Office, C/o International Livestock Research Institute, Addis Ababa, Ethiopia

Irshad Ahmad Khandy Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Ajay Kumar Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

Jitendra Kumar Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Shiv Kumar International Center for Agricultural Research in the Dry Areas, Rabat – Instituts, Rabat, Morocco

Sumita Kumari Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu, (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Huihui Li Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, and CIMMYT China Office, Chinese Academy of Agricultural Sciences, Beijing, China

Nupur Malviya Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Rebecca McGee USDA Agriculture Research Service, Grain Legume Genetics and Physiology Research Unit, Johnson Hall, Washington State University, Pullman, WA, USA

Suhel Mehandi Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Reyazul Rouf Mir Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Vankat Raman Pandey Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Somanagouda Patil International Centre for Agricultural Research in the Dry Areas, Rabat – Instituts, Rabat, Morocco

Gerrit Polder Wageningen University and Research Centre, Wageningen, The Netherlands

Priyanka Porwal Division of Basic Sciences, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Aditya Pratap Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Tony Pridmore Center for Plant Integrative Biology, University of Nottingham, Nottingham, UK

Karthika Rajendran International Centre for Agricultural Research in the Dry Areas, Rabat – Instituts, Rabat, Morocco

Zed Rengel Soil Science and Plant Nutrition, School of Earth and Environment, and The UWA Institute of Agriculture, The University of Western Australia, Crawley, WA, Australia

David Rousseau Université de Lyon, CREATIS, CNRS UMR5220, INSERM U1044, Université de Lyon 1, INSA-Lyon, Villeurbanne, France

Arno Ruckelshausen COALA – Competence Center of Applied Agricultural Engineering, University of Applied Sciences Osnabrück, Osnabrück, Germany

Bikram Singh Division of Plant Breeding & Genetics, Sher-e-Kashmir University of Agricultural Sciences & Technology of Jammu (SKUAST-J), Chatha, Jammu, Jammu and Kashmir, India

Jagdish Singh Division of Basic Sciences, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Murari Singh International Center for Agricultural Research in the Dry Areas (ICARDA), Amman, Jordan

Parul Singh Division of Basic Sciences, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Parvez Sofi Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir (SKUAST-K), Kashmir, Jammu and Kashmir, India

Mudit Srivastava Division of Basic Sciences, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Dil Thavarajah Vegetable/Pulse Quality and Nutrition, Clemson University, Department of Agricultural & Environmental Sciences, Clemson, SC, USA

Pushparajah Thavarajah Food Chemistry and Analysis, Nutrient Interactions, North Dakota State University, Fargo, ND, USA

Cengiz Toker Department of Field Crops, Faculty of Agriculture, Akdeniz University, Antalya, Turkey

Rakhi Tomar Division of Crop Improvement, ICAR – Indian Institute of Pulses Research, Kalyanpur, Kanpur, Uttar Pradesh, India

Bulent Uzun Department of Field Crops, Faculty of Agriculture, Akdeniz University, Antalya, Turkey

Gerie W.A.M. van der Heijden Dupont Pioneer, Johnston, IA, USA

Wageningen University and Research Centre, Wageningen, The Netherlands

Rob van der Schoor Wageningen Campus, Greenhouse Horticulture, Wageningen, The Netherlands

Jiankang Wang Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, and CIMMYT China Office, Chinese Academy of Agricultural Sciences, Beijing, China

Ming Li Wang USDA-ARS Plant Genetic Resources Conservation Unit, Griffin, GA, USA

Aijaz Wani Department of Botany, University of Kashmir, Srinagar, Jammu and Kashmir, India

Engin Yol Department of Field Crops, Faculty of Agriculture, Akdeniz University, Antalya, Turkey

About the Editors

Dr. Jitendra Kumar, born in 1973, is presently working as a Senior Scientist in the Division of Crop Improvement at Indian Institute of Pulses Research, Kanpur. He has an excellent research career throughout. He secured Gold Medal during masters' programme and pursued his Ph.D. in Genetics and Plant Breeding from G. B. Pant University of Agriculture and Technology, Pantnagar, India. He was awarded CSIR-Research Associateship during 2003–2005 for postdoctoral studies at the Institute of Integrative Medicine, Jammu (India). He has more than 14 years of research experience in genetic improvement using both conventional and molecular marker-assisted breeding approaches on various crops including sunflower, medicinal and aromatic, cereal and pulse crops. He has done work on development of SSR markers, identification of QTLs for preharvest sprouting and high grain protein content and marker-assisted breeding in wheat for pyramiding the preharvest sprouting tolerance and high grain protein content and leaf rust resistance and developed a number of lines. During this period, he undertook study tours of several countries including Austria, Syria, Bangladesh, Nepal, Lebanon and Canada. His research interests include conventional and molecular breeding, QTL analysis and marker-assisted selection for crop improvement. He has more than 100 publications including research and review articles in reputed national and international journals, book chapters, meeting reports, popular articles, and bulletins. He has also co-edited three books including *Biology and Breeding of Food Legumes* published by CABI, Oxfordshire, *Alien Gene Transfer in Crop Plants: Innovations, Methods and Risk Assessment* and *Alien Gene Transfer in Crop Plants: Achievement and Impacts* both published by Springer, New York, USA. He has developed high-yielding varieties (IPL 316 and IPL 526) of lentil and several others are in the pipeline. His current priorities include involvement of molecular marker technology in conventional lentil breeding programme for making genetic improvement towards biotic and abiotic stresses.

Dr. Aditya Pratap, born on October 18, 1976, is currently working as a Senior Scientist (Plant Breeding) in the Crop Improvement Division, Indian Institute of Pulses Research, Kanpur. He obtained his Master's and Ph.D. degrees in Plant Breeding and Genetics from CSK Himachal Pradesh Agricultural University, Palampur, India, in 1999 and 2003. Holding a brilliant academic and service record, he has been associated with crop research since last 10 years. He has worked on genetic improvement of crop plants including wheat,

triticale, rapeseed-mustard, chickpea and *Vigna* species and has been instrumental in the development of haploidy breeding protocol in cereals through chromosome elimination technique. He has been associated with the development and release of five crop varieties including two in rapeseed-mustard (RSPT-2 and RSPR03), two in green gram (IPM 02-14 and IPM 02-3) and one in facultative winter wheat (DH 114). He has developed two extra early mungbean genotypes (IPM 205-7 and IPM 409-4 (48 days maturity)) besides being instrumental in establishing prebreeding garden of rapeseed-mustard at SKUAST-Jammu and of pulses at IIPR, Kanpur. Presently, he is working on genetic improvement of green gram (*Vigna radiata*) through distant hybridization aided by conventional and biotechnological tools. His research interests include distant hybridization, doubled haploidy breeding, plant tissue culture, and molecular breeding. To his credit, he has about 120 publications which include research papers published in high-impact journals as well as reviews/chapters for best international publishers including Springer, Academic Press and CRC. He has published four books entitled, *Haploidy Breeding in Triticale and Triticale X Wheat Hybrids: Comparison of Anther Culture and Chromosome Elimination Techniques* by Lambert Academic Publishing, Germany; *Biology and Breeding of Food Legumes* published by CABI, Oxfordshire; *Alien Gene Transfer in Crop Plants: Innovations, Methods and Risk Assessment* and *Alien Gene Transfer in Crop Plants: Achievements and Impacts*, both published by Springer, New York. He is also a recipient of the prestigious Norman E. Borlaug International Agricultural Science and Technology Fellowship. He is an acknowledged speaker and has several awards to his credit.

Dr. Shiv Kumar is Food Legumes Coordinator and works as Lentil and Grasspea Breeder at the International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat platform, Morocco. Before joining the present position, he served the Indian Council of Agricultural Research as a Plant Breeder for 18 years. He also served the International Crops Research Institute for Semi-Arid Tropics as Post Doctoral Fellow and worked on basic studies in chickpea breeding and genetics between 1991 and 1993. His post-doctoral work has led to identification of extra early photo-thermo insensitive genotypes in chickpea which have been used as donors across the globe. Dr. Kumar has over 25 years of research experience on basic and applied aspects of breeding rice and pulses including chickpea, grasspea, *Vigna* crops and lentil. He has been associated in the development of 28 varieties of pulse crops and one variety of rice. He also identified useful new germplasm for use in breeding program of rice, lentil, chickpea, grasspea, mungbean and urdbean. He has to his credit more than 300 articles including 110 research papers in refereed journals, 52 book chapters, 6 books, 7 technical bulletins and 2 training manuals. He also received a number of academic distinctions and awards including Rockefeller Fellowship, Best Scientist Award from IIPR for the years 2005 and 2008, and Best Research Team Award from MULLaRP of ICAR in 2008. His research interests include pre-breeding activities, genetic enhancement through conventional and marker-assisted breeding and biometrical genetics.

Biometrical Approaches for Analysis of Phenotypic Data of Complex Traits

17

Huihui Li and Jiankang Wang

Abstract

Phenotype (or phenotypic value) is the performance of a trait in interest, which can be observed in the field and then used in estimating the unknown genotypic value (or the phenotypic mean). In this chapter, we introduced statistical approaches to analyze three types of phenotypic observation, i.e., (1) replicated observations of one genotype in one environment, (2) replicated observations of multiple genotypes in one environment, and (3) replicated observations of multiple genotypes in multiple environments. The principle of analysis of variance (ANOVA) was applied on each kind of phenotypic data. From the results of ANOVA, we can further estimate genotypic value, genetic effects, variance components, heritability, etc., which can be further used in genetic studies and breeding applications. In the end, we present a computer tool implemented in the integrated genetic software QTL IciMapping, which includes the biometrical approaches introduced in this chapter and can be readily used in phenotyping complex traits.

Keywords

Phenotype • Analysis of variance (ANOVA) • Genotype • Genetic variance • Heritability

17.1 Introduction

For making genetic improvement, plant breeders collect huge amount of phenotypic data on various populations. The phenotype of traits particularly quantitative traits is controlled by genotype and environments, and thus raw phenotypic data measured for various complex traits include the combined effect of both the genotypic value (G) and the environmental deviation (E): $P = G + E$. However, for making genetic improvement in trait, genotypic value is more important than the phenotypic value that is the combined effect of all the genetic effects, including nuclear genes, mitochondrial genes, and interactions

H. Li • J. Wang (✉)
Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, and CIMMYT China Office, Chinese Academy of Agricultural Sciences, No. 12 Zhongguancun South Street, Beijing 100081, China
e-mail: jkwang@cgiar.org

J. Kumar et al. (eds.), *Phenomics in Crop Plants: Trends, Options and Limitations*,
DOI 10.1007/978-81-322-2226-2_17,

between the genes. Therefore it is essential to know the contribution of heritable variation in total phenotypic variation of a quantitatively inherited trait. For this purpose, different statistical approaches have been used to study the inheritance of quantitative traits which is known as biometrical genetics. Therefore, this chapter has described the different approaches developed in biometrical genetics for analysis of phenotypic data for finding out the genotypic value of traits.

17.2 Basic Statistics Theory

17.2.1 Random Variable and Normal Distribution

If a random variable X has the following probability density, X is stated to have a normal distribution with mean μ and variance σ^2, where μ and σ^2 are known constants or unknown but estimable parameters:

$$f(X=x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{17.1}$$

The function given in Eq. (17.1) is called the probability density (PDF), and the random variable X is normally denoted by $X \sim N(\mu, \sigma^2)$. When the normal distribution has a mean of 0 and a variance of 1, the distribution is also called a standard normal distribution, denoted by $N(0, 1)$. For any normal distribution, $X \sim N(\mu, \sigma^2)$, $(X\text{-}\mu)/\sigma \sim N(0, 1)$, where σ is the square root of variance σ^2.

Assuming a random variable X has the PDF $f(x)$ and the possible value of X is any real number, mean (also called expectation) and variance of the random variable X are defined as follows:

$$E(X)=\int_{-\infty}^{+\infty} x\ f(x)dx, \tag{17.2}$$

$$V(X)=\int_{-\infty}^{+\infty}[x-E(x)]^2 f(x)dx = \int_{-\infty}^{+\infty}x^2 f(x)dx-[E(x)]^2 \tag{17.3}$$

The two statistics defined in the above two equations are also called mean and variance of the distribution of the random variable X. For a normal distribution $X \sim N(\mu, \sigma^2)$, it can be proved that $E(X) = \mu$ and $V(X) = \sigma^2$. This is the reason why the normal distribution $N(\mu, \sigma^2)$ is stated to have a mean of μ and a variance of σ^2. Or equally, if $X \sim N(\mu, \sigma^2)$, the random variable X is stated to have a mean of μ and a variance of σ^2.

17.2.2 Distributions Derived from the Standard Normal Distribution

From the standard normal distribution, we can define other distributions commonly used in statistical inference. If random variables $X_1, X_2, \ldots, X_n$ are independent and identical to the standard normal distribution $N(0, 1)$, the sum square of these variables is defined to follow a χ^2(chi--square) distribution with the degree of freedom of n, i.e.,

$$Y=X_1^2+X_2^2+\cdots X_n^2 \sim \chi^2(n) \tag{17.4}$$

Random variable $X \sim N(0, 1)$, random variable $Y \sim \chi^2(n)$, and the two variables are independent, and then $\frac{X}{\sqrt{\frac{1}{n}Y}}$ is defined to follow a t distribution with the degree of freedom of n, i.e.,

$$\frac{X}{\sqrt{\frac{1}{n}Y}} \sim t(n) \tag{17.5}$$

Figure 17.1 shows PDFs of the standard normal distribution $N(0, 1)$ and some t and χ^2 distributions. As $N(0, 1)$, t distribution is symmetrical. But, t distribution has much longer tails compared with $N(0, 1)$. With the increase in the degree of freedom, t distribution can quickly approach to the standard normal distribution (left of Fig. 17.1). In practice, when the degree of freedom is greater than 30, t distribution will be viewed to be the standard normal distribution. χ^2 distributions can only have positive values, by definition. With the increase in the degree of freedom, χ^2 distribution becomes much flatter

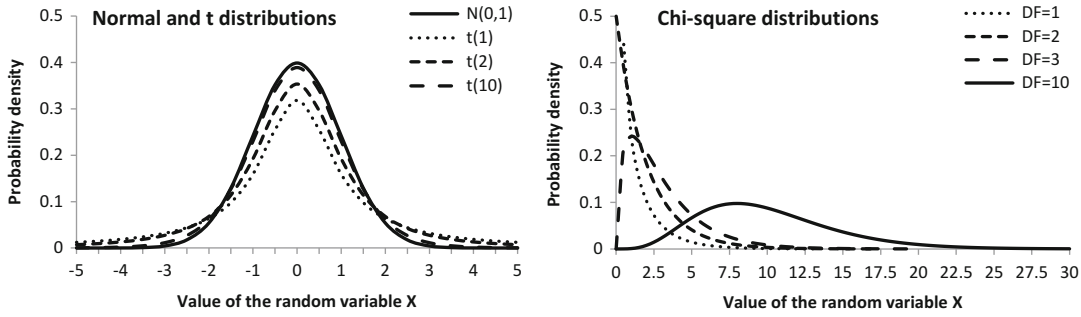


Fig. 17.1 Probability density functions (PDF) of the standard normal distribution $N(0, 1)$ and some t and χ^2 distributions

and has much longer tail to the right side (right of Fig. 17.1).

Assuming that two random variables $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$ are independent, $\frac{\frac{1}{n_1}Y_1}{\frac{1}{n_2}Y_2}$ is defined to follow an F distribution with the two degrees of freedom of n_1 and n_2, i.e.,

$$\frac{\frac{1}{n_1}Y_1}{\frac{1}{n_2}Y_2} \sim F(n_1, n_2) \tag{17.6}$$

17.3 One Genotype in One Environment

17.3.1 Mean and Variance of a Phenotypic Distribution of Trait in Interest

It is assumed that we can repeatedly observe the phenotype (P) of a given genotype for a trait in interest in a given environment, and each observation is independent. True genotypic value in the environment is represented by G, which is an unknown parameter. Error in phenotypic measurement is a normally distributed random variable, having a mean of 0 and a variance of ${\sigma_\varepsilon}^2$. Error variance is unknown as well or has been estimated from previous experiments. Therefore, the observation will be independently and normally distributed around the true genotypic value G, and variance of the phenotypic distribution will be equal to the error variance ${\sigma_\varepsilon}^2$ in the environment. In statistics, we say these phenotypic observations have independently identical distributions (iid), that is,

$$P_k \sim N\left(G, \sigma_\varepsilon^2\right), \ (k = 1, 2, \ldots, r \text{ for replication}), \text{ iid} \tag{17.7}$$

In practice, we do not know the true genotypic value G and the true error variance ${\sigma_\varepsilon}^2$. However, they can be estimated from the replicated phenotypic observations P_k ($k = 1, 2, \ldots, r$), as the observation contains the information about these true values. This can be seen more clearly when distribution model (17.7) is represented in the following equivalent linear model:

$$\begin{aligned} P_k &= G + \varepsilon_k, \varepsilon_k \\ &\sim N\left(0, \sigma_\varepsilon^2\right), \ (k = 1, 2, \ldots, r), \text{ iid} \end{aligned} \tag{17.8}$$

Using the replicated observations, we can calculate the sample mean (represented by $\overline{P}$) and sample variance (represented by MS_ε) and use them as the estimates of the unknown G and error variance, respectively. Sample mean and the estimate of the genotypic value and their distributions are

$$\begin{aligned} \hat{G} &= \overline{P} = \frac{1}{r}\sum_k P_k, \text{ and } \hat{G} = \overline{P} \\ &\sim N\left(G, \ \frac{\sigma_\varepsilon^2}{r}\right) \end{aligned} \tag{17.9}$$

So, expectation of the estimated effect $\hat{G}$ is equal to the true genotypic effect. In statistics, we say the sample mean is an unbiased estimate of the

true genotypic effect. Variance of the estimate $\hat{G}$ is $\frac{1}{r}$ of the error variance. So, more observations give more precise estimate of the true effect. With the increase in sample size, the sample mean will asymptotically approach to the true genotypic value G. For this reason, the true genotypic value G is also called phenotypic mean. In statistics, it can also be proved that, among all possible unbiased linear combinations of observations ($k = 1, 2, \ldots, r$), the estimate given in Eq. (17.9) has the least variance. So the sample mean given in Eq. (17.9) is also called the best linear unbiased estimate (BLUE) of the phenotypic mean.

Sample variance (represented by MS_ε) and the estimate of the error variance (represented by $\hat{\sigma}_\varepsilon^2$) are

$$\hat{\sigma}_\varepsilon^2 = \mathrm{MS}_\varepsilon = \frac{1}{r-1}\sum_k \left(P_k - \overline{P}\right)^2 \tag{17.10}$$

For each observation, $(P_k\text{-}\overline{P})$ is the deviation of the observation from the sample mean, which can be used to measure random error effect in the observation. The sum square (SS) of each deviation is represented by SS_ε, i.e.,

$$\mathrm{SS}_\varepsilon = \sum_k \left(P_k - \overline{P}\right)^2 \tag{17.11}$$

Under the assumptions in distribution model (17.7) or equally in linear model (17.8), we can prove the following relationship between the expectations of SS_ε and ${\sigma_\varepsilon}^2$:

$$\begin{aligned} E(\mathrm{SS}_\varepsilon) &= E\sum_k \left(P_k - \overline{P}\right)^2 = E\sum_k \left[(P_k - G) - \left(\overline{P} - G\right)\right]^2 \\ &= \sum_k E(P_k - G)^2 - 2E\left[\left(\overline{P} - G\right)\sum_k (P_k - G)\right] + rE\left(\overline{P} - G\right)^2 \\ &= \sum_k E(P_k - G)^2 - 2E\left[\left(\overline{P} - G\right)r\left(\overline{P} - G\right)\right] + rE\left(\overline{P} - G\right)^2 \\ &= \sum_k E(P_k - G)^2 - rE\left(\overline{P} - G\right)^2 \\ &= r\sigma_\varepsilon^2 - r\frac{\sigma_\varepsilon^2}{r} = (r-1)\sigma_\varepsilon^2 \end{aligned} \tag{17.12}$$

The coefficient $(r - 1)$ before error variance in Eq. (17.12) is called the degree of freedom of the error effects. From Eqs. (17.10) and (17.11), it can be easily seen that

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 = \mathrm{MS}_\varepsilon &= \frac{\mathrm{SS}_\varepsilon}{r-1}, \quad E\left(\hat{\sigma}_\varepsilon^2\right) \\ &= E(\mathrm{MS}_\varepsilon) = \sigma_\varepsilon^2 \end{aligned} \tag{17.13}$$

Therefore, sample variance is an unbiased estimate of the unknown error variance.

17.3.2 An Example on Plant Height in Four Genetic Populations

Table 17.1 gives observations of plant height (cm) in two inbred lines A and B and their F_1 and F_2 populations. Using Eqs. (17.9) and (17.10), we can estimate that inbred A has the a mean height of 160 cm, inbred B of 103 cm, F_1 hybrid of 149 cm, and F_2 population of 140 cm. Mean plant height in F_1 or F_2 population is between the two inbred parents. The two parents and their F_1 have similar variance, but variance of F_2 population is much greater. The larger variance in F_2 indicates the presence of genetic variance in plant height. If we can assume random error effects are homogeneous in the three non-segregating populations, we can combine the three sum squares to have one estimate of the error variance, i.e.,

Table 17.1 Plant height (cm) in two inbred lines and their F_1 and F_2 population. There are 10, 10, 10, and 30 observations in the four genetic populations

Population	Individual plant height (cm)	Sample mean	DF	SS	MS (=sample variance)
Inbred A	155, 161, 150, 164, 165, 161, 160, 158, 166, 164	160.40	9	222.40	24.71
Inbred B	97, 109, 92, 103, 109, 104, 98, 106, 102, 110	103.00	9	314.00	34.89
F_1	156, 148, 140, 150, 148, 147, 146, 155, 148, 150	148.80	9	183.60	20.40
F_2	89, 157, 149, 169, 123, 158, 151, 83, 167, 154, 152, 167, 116, 146, 97, 147, 162, 159, 111, 143, 144, 124, 137, 156, 80, 169, 157, 152, 157, 116	140.00	29	20074.00	692.21

$$\begin{aligned} \mathrm{SS}_T &= \mathrm{SS}_{P_1} + \mathrm{SS}_{P_2} + \mathrm{SS}_{F_1} = 720, \\ \mathrm{DF}_T &= \mathrm{DF}_{P_1} + \mathrm{DF}_{P_2} + \mathrm{DF}_{F_1} \\ &= 27, \quad \hat{\sigma}_\varepsilon^2 = \frac{\mathrm{SS}_T}{\mathrm{DF}_T} = 26.67 \end{aligned} \tag{17.14}$$

Assuming the height of inbred B has the normal distribution $N(100, 30)$, based on the observed height in Table 17.1. Therefore, as a random variable, the height has a mean of 100 and a variance of 30. Figure 17.2 shows the distribution curves of sample means for several sample sizes. It is clear that larger sample size results in smaller variance in the sample mean. Each curve in Fig. 17.2 represents how the sample mean will be distributed if we can repeat the sampling procedure infinitely. In practice, we normally have one set of samples. In Table 17.1, we only have one set of 10 phenotypic height values. Therefore, that is no guarantee that the sample mean of inbred B (i.e., 103 cm) is equal to the true genotypic height.

The unbiased sample mean to the true value, as given in Eq. (17.9), is a statistical property from the large number of sampling. The same is true for the sample variance. In statistics, the unknown parameters can be estimated from samples drawn from their population. But this does not indicate that the estimated value will be equal to the unknown parameter. Instead, each sample drawn from a population in interest is viewed as a random variable. Any estimate from a set of samples is also a random variable. Unknown parameters are normally assumed to be constants. Therefore, it does not make sense to say that a random variable is equal to a constant.

But, to know the distribution of a sample, statistics is good enough for conducting statistical inference and test. Say, we can tell how likely the true height is located in a given interval, how likely the true height is different from another genotype or genetic population, and so on. For example, using Table 17.1, we can tell the probability that the true height of inbred B is from 95 to 105 cm. We can calculate the significance probability between inbred A and inbred B, where t distribution will be used. We can test whether the genetic variance of F_2 population is significant, where F distribution will be used.

In addition, if we can assume random errors in F_2 populations have equal variance as estimated in Eq. (17.14), we are able to estimate the genetic variance of F_2 by subtracting the error variance from the phenotypic variance. That is,

$$\hat{\sigma}_G^2 = \hat{\sigma}_P^2 - \hat{\sigma}_\varepsilon^2 = 692.21 - 26.67 = 665.54$$

And the heritability in the broad sense in the F_2 population can be estimated,

$$H = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_P^2} = \frac{665.54}{692.21} = 0.96$$

17.3.3 Calculating Sample Mean and Sample Variance from Frequency Data

In many cases, the raw data with large sample size are grouped and the frequency of each group is given instead. Table 17.2 shows the number of samples falling in each group represented by the mid-group value of ear length (cm) in four genetic populations (East 1911). Taking inbred

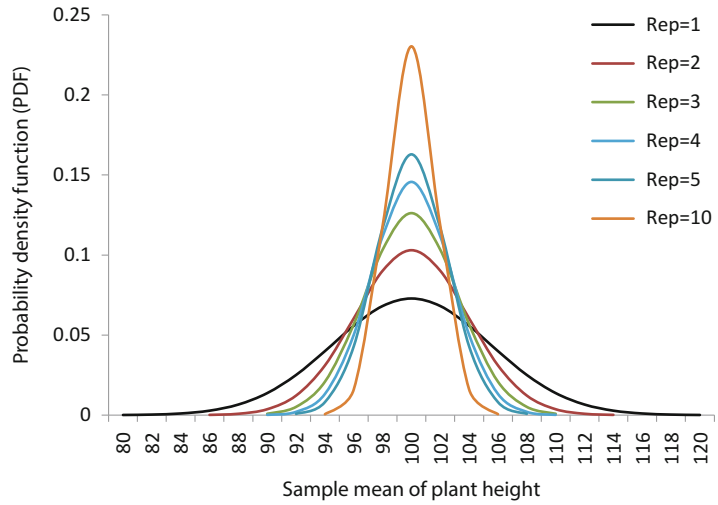


Fig. 17.2 Distribution of the sample mean of inbred B plant height. As a random variable, the height of inbred B is assumed to be normally distributed, having a mean of 100 and a variance of 30

Table 17.2 Frequency of ear length (cm) in four genetic populations

Ear length	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Size
Inbred A	4	21	24	8	–	–	–	–	–	–	–	–	–	–	–	–	–	57
Inbred B	–	–	–	–	–	–	–	–	3	11	12	15	26	15	10	7	2	101
F_1	–	–	–	–	1	12	12	14	17	9	4	–	–	–	–	–	–	69
F_2	–	–	4	5	22	56	80	145	129	91	63	27	17	6	1	–	–	646

Adapted from East (1911)

F_1 is the hybrid between the two inbred lines, and F_2 is the selfing generation of the F_1 hybrid

A as an example, among the 57 ears, 4 have ear length between 4.5 and 5.5, 21 have ear length between 5.5 and 6.5, 24 have ear length between 6.5 and 7.5, and 8 have ear length between 7.5 and 8.5. Let x_k be the mid-value of the kth group and $f_k = \frac{n_k}{n}$ be the relative frequency, estimates of the population mean and variance are, therefore,

$$\begin{aligned}\hat{\mu} &= \sum_k f_k x_k, \quad \text{and } \hat{\sigma}_\varepsilon^2 \\ &= \sum_k f_k x_k^2 - \hat{\mu}^2 \end{aligned} \tag{17.15}$$

One may find that the sample mean and variance given in Eq. (17.15) are similar to the distribution mean and variance given in Eqs. (17.2) and (17.3). From the above equation, means and variances of the four populations can be estimated. That is,

$$\overline{P}_A = 6.63, \ \overline{P}_B = 16.80, \ \overline{P}_{F_1} = 12.12, \ \overline{P}_{F_2} = 12.68$$
$$\hat{\sigma}_A^2 = 0.65, \ \hat{\sigma}_B^2 = 3.53, \ \hat{\sigma}_{F_1}^2 = 2.28, \ \hat{\sigma}_{F_2}^2 = 3.97$$

Assuming that errors have the same variance in these populations, we may use average of variances in the three non-segregating populations to estimate the true error variance and estimate the genetic variance and heritability in the F_2 population. That is,

$$\hat{\sigma}_\varepsilon^2 = 2.15, \ \hat{\sigma}_G^2 = \hat{\sigma}_{F_2}^2 - \hat{\sigma}_\varepsilon^2 = 1.82, \ H = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_{F_2}^2}$$
$$= \frac{1.82}{3.97} = 0.46$$

In addition, if the multifactorial hypothesis in classical quantitative genetics is applicable, we could estimate the number of loci (l) affecting ear length in these populations from the Castle-Wright formula. That is,

$$l = \frac{\left(\overline{P}_A - \overline{P}_B\right)^2}{8\left(\hat{\sigma}^2_{F_2} - \hat{\sigma}^2_{\varepsilon}\right)} \approx 7 \tag{17.16}$$

The multifactorial hypothesis mentioned above is fundamental in classical quantitative genetics. Major content in the hypothesis is that quantitative traits are controlled by a large number of Mendelian genes having smaller effects and can be easily modified by environments. In addition to the multifactorial hypothesis, when calculating the number of loci affecting ear length in Eq. (17.16), we also assume that the genes have equal additive effect on ear length, inbred A has all the alleles reducing the ear length, and inbred B has all the alleles increasing the ear length.

17.4 Multiple Genotypes in One Environment

17.4.1 Assumptions and Models

It is assumed that we make the field phenotyping experiment with r replications for a set of g genotypes in a given environment. The phenotypic means are represented by μ_i ($i = 1, 2, \ldots, g$), which are unknown parameters. Error effects are normally distributed with a mean of 0 and an unknown variance ${\sigma_\varepsilon}^2$. Randomization of genotypes in the field will assure that the observations are independent. So the observed phenotype for the ith genotype and kth replication is

$$P_{ik} \sim N\left(\mu_i, \sigma^2_\varepsilon\right), (i = 1, 2, \ldots, g;\ \ k = 1, 2, \ldots, r) \tag{17.17}$$

Therefore, we are having g normal distributions, corresponding to the g genotypes. The populations may have different means, but they should have the equal variance, which is actually the random error variance.

Given the phenotypic means of g genotypes, we can define an overall phenotypic mean, that is, $\mu \hat{=} \frac{1}{g}\sum_i \mu_i$. By defining the deviation of each phenotypic mean to the overall mean as the genotypic effect, represented by G_i, we can have the following linear model:

$$P_{ik} = \mu_i + \varepsilon_{ik} = \mu + G_i + \varepsilon_{ik},$$
$$\text{where } \mu \hat{=} \frac{1}{g}\sum_i \mu_i, \text{and } \varepsilon_{ik} \sim N\left(0,\ \sigma^2_\varepsilon\right)\ (i = 1,\ 2,\ \ldots, g; k = 1,\ 2,\ \ldots, r) \text{ and iid} \tag{17.18}$$

Genetic variance can be defined from the g phenotypic means:

$$\sigma^2_G \hat{=} \frac{1}{g-1}\sum_i G_i^2 \tag{17.19}$$

17.4.2 Estimation of Genotypic Effect and Genetic Variance

As we could see, there are two major purposes when making field experiment. The first one is to estimate the phenotypic means of a set of genotypes, as defined in distribution model (17.17) or in linear model (17.18). Based on the estimation of phenotypic means, we can conduct further genetic study, say QTL mapping. In the perspective of breeding, we can decide which genotypes have better performance and should be selected and advanced to the next season or which genotypes should be grown in this environment. The second one is to estimate error variance as defined in linear model (17.18) and genetic variance as defined in Eq. (17.19). From the estimation of the two variances, we can estimate heritability, which has been seen in the previous section.

Now we will show how the genotypic effects $G_i (i = 1,\ 2, \ldots, g)$, genetic variance ${\sigma_G}^2$, and error variance ${\sigma_\varepsilon}^2$ can be estimated by using observations P_{ik} ($i = 1, 2, \ldots, g$; $k = 1, 2, \ldots, r$). First, we define the overall sample mean ($\overline{P}..$), i.e., the mean across the g genotypes and the r replications. Second, we

define sample mean for each genotype ($\overline{P}_i.$), i.e., the mean across the r replications for each genotype. Based on the distribution model (17.17) of each observation, we have

$$\overline{P}.. = \frac{1}{gr}\sum_{j,k} P_{jk} \sim N\left(\mu, \frac{\sigma_\varepsilon^2}{gr}\right), \text{ and } \overline{P}_i. = \frac{1}{r}\sum_{k} P_{ik} \sim N\left(\mu_i, \frac{\sigma_\varepsilon^2}{r}\right)$$

By defining the overall mean and sample mean of each genotype, the phenotype can be decomposed as:

$$P_{ik} = \overline{P}.. + \left(\overline{P}_i. - \overline{P}..\right) + \left(P_{ik} - \overline{P}_i.\right) \text{ or equally} \quad (17.20)$$

$$P_{ik} - \overline{P}.. = \left(\overline{P}_i. - \overline{P}..\right) + \left(P_{ik} - \overline{P}_i.\right) \quad (17.21)$$

In Eq. (17.20), the first term is the overall sample mean. The second term is the deviation of the sample mean of each genotype to the overall mean. The third term is the residual deviation. The three terms in Eq. (17.18) can be used to estimate the three parameters defined in the distribution model (17.18). That is,

$$\hat{\mu} = \overline{P}.., \hat{G}_i = \left(\overline{P}_i. - \overline{P}..\right), \hat{\varepsilon}_{ik} = \left(P_{ik} - \overline{P}_i.\right) \quad (17.22)$$

Given the observed phenotypic values P_{ik} $(i = 1, 2, \ldots, g; k = 1, 2, \ldots, r)$, we can use Eq. (17.20) to estimate the overall mean, genotypic effect, and residual effect (if we want). To estimate genotypic variance and error variance, we have to consider the sum of the squared deviations. Total sum square (SS_T) is defined from the left side of Eq. (17.21). Total sum square can be further decomposed into two parts, which are represented by SS_G and SS_ε, corresponding to the two terms in the right side of Eq. (17.21). That is,

$$SS_T = \sum_{i,k}\left(\overline{P}_{ik} - \overline{P}..\right)^2 = \sum_{i,k}\left[\left(\overline{P}_{ik} - \overline{P}_i.\right) + \left(\overline{P}_i. - \overline{P}..\right)\right]^2 = \sum_{i,k}\left(\overline{P}_{ik} - \overline{P}_i.\right)^2 + r\sum_{i}\left(\overline{P}_i. - \overline{P}..\right)^2 = SS_\varepsilon + SS_G$$

We have a total of $g \times r$ independent observations. Total sum square (SS_T) has a degree of freedom of $gr - 1$. The one degree of freedom can be understood as being used in the estimation of the overall sample mean. Without the overall sample mean, we are unable to estimate the deviations on the left side of Eq. (17.21). Sum square of the estimated genotypic effects, i.e., SS_G, has a degree of freedom of $g - 1$. There are g estimated genotypic effects (Eq. 17.22), but the sum of these effects is equal to 0. So, the degree of freedom of $g - 1$ can be understood as the number of independent estimated genotypic effects. There are gr estimated residual effects (Eq. 17.20), but they are not completely independent. The degree of freedom of $g(r - 1)$ can also be understood as the number of independent estimated residual effects. Of course, it can also be found by subtracting $g - 1$ from the total degree of freedom $gr - 1$.

Mean square (MS) is defined as the sum square divided by its degree of freedom. That is,

$$MS_G = \frac{SS_G}{g - 1}, \text{ and } MS_\varepsilon = \frac{SS_\varepsilon}{g(r - 1)}$$

Intuitively, mean square of estimated genotypic effects reflects the magnitude of genotypic variance defined in Eq. (17.19). Mean square of the residual effects reflects the magnitude of error variance defined in distribution model Eq. (17.18). In statistics, we can prove

Table 17.3 ANOVA of single-environmental phenotyping trials of multiple genotypes

Source of variation	Degree of freedom (DF)	Sum square (SS)	Mean square (MS)	Expected mean square (EMS)
Genotype	$g-1$	SS_G	MS_G	$\sigma_\varepsilon^2 + r\sigma_G^2$
Error	$g(r-1)$	SS_ε	MS_ε	${\sigma_\varepsilon}^2$
Total	$gr-1$	SS_T		

$$E(SS_G) = (g-1)\sigma_\varepsilon^2 + (g-1)r\sigma_G^2,\ \ E(SS_\varepsilon) = g(r-1)\sigma_\varepsilon^2 \quad (17.23)$$

Therefore,

$$E(MS_G) = \sigma_\varepsilon^2 + r\sigma_G^2,\ \ E(MS_\varepsilon) = \sigma_\varepsilon^2 \quad (17.24)$$

From Eq. (17.24), we can see that the expectation of MS_ε is equal to the error variance and therefore is the unbiased estimate of error variance. In addition to genetic variance, error variance is also included in the expectation of MS_G. Therefore, $\frac{1}{r}MS_G$ cannot be an unbiased estimate for genotypic variance. Instead, we can have the following unbiased estimates for error variance and genotypic variance:

$$\hat{\sigma}_\varepsilon^2 = MS_\varepsilon,\ \ \hat{\sigma}_G^2 = \frac{1}{r}(MS_G - MS_\varepsilon) \quad (17.25)$$

The above procedure is called analysis of variance (ANOVA) in statistics and can be summarized in Table 17.3. An F-statistic can be constructed to test the significance of the genotypic variation compared with error, i.e.,

$$F = \frac{MS_G}{MS_\varepsilon} \sim F[g-1,\ g(r-1)] \quad (17.26)$$

In many cases, each replication of the g genotypes may be arranged in one relatively homogeneous block. Variation between blocks can occur. The use of block is another important concept in experimental design, which can reduce the random error variance and improve the precision when comparing genotypes. In this case, the deviation of the block mean to the overall sample mean estimates the block effect (represented by B_k), i.e.,

$$\hat{B}_k = (\overline{P}_{\cdot k} - \overline{P}_{\cdot\cdot}),\quad (k = 1,\ 2,\ \ldots, r) \quad (17.27)$$

And linear model (17.21) becomes

$$P_{ik} - \overline{P}_{\cdot\cdot} = (\overline{P}_{\cdot k} - \overline{P}_{\cdot\cdot}) + (\overline{P}_{i\cdot} - \overline{P}_{\cdot\cdot}) + (P_{ik} - \overline{P}_{i\cdot} - \overline{P}_{\cdot k} + \overline{P}_{\cdot\cdot}) \quad (17.28)$$

Similarly, ANOVA can be done based on the above model. It can be seen that including the block effect will not affect the estimation of genotypic effects and the genotypic variance but will affect the estimation of residual effects and the error variance. When the block effect is significant, estimated error variance will be lower than that from linear model (17.21). The reduced error variance allows more precise comparison of phenotypic means. In practice, other options are to estimate the block effect using Eq. (17.27), adjust the raw data by the block effect, and apply linear model (17.21) on the adjusted phenotypic observations.

17.4.3 Estimation of Heritability in the Broad Sense

As represented by the linear model (17.18), in single environment, phenotype of a quantitative trait for a given genotype or line or family can be decomposed into three parts: (1) overall mean across genotypes and replications, (2) genotypic effect of the specific genotype, and (3) random residual error. That is,

$$P = \mu + G + \varepsilon \quad (17.29)$$

where overall mean and genotypic effect are assumed to be unknown parameters, and residual error is assumed to be random variable. When residual error has a normal distribution, the phenotypic variance ${\sigma_P}^2$ is equal to the sum of genotypic variance ${\sigma_G}^2$ and error variance ${\sigma_\varepsilon}^2$. In session 3.2, we have seen that ANOVA can acquire the unbiased estimates of genotypic variance and error variance. Therefore, we can have

Table 17.4 Yield performance of ten maize inbred lines in three replications

Genotype	Replication			Phenotypic mean $(\overline{P}_{i\cdot})$	Estimated genotypic effect $(\hat{G}_i)$
	I	II	III		
RIL1	2.56	2.66	2.43	2.550	−0.247
RIL2	2.66	2.50	2.75	2.637	−0.160
RIL3	2.93	2.97	2.70	2.867	0.070
RIL4	2.57	2.21	1.80	2.193	−0.604
RIL5	3.06	2.61	2.72	2.797	0.000
RIL6	1.94	2.16	2.14	2.080	−0.717
RIL7	1.85	1.69	2.25	1.930	−0.867
RIL8	3.83	3.58	3.80	3.737	0.940
RIL9	4.32	4.12	4.14	4.193	1.396
RIL10	2.81	3.33	2.81	2.983	0.186
Block mean $(\overline{P}_{\cdot k})$	2.853	2.783	2.754	$\overline{P}_{ik} = 2.797$	
Estimated block effect $(\hat{B}_k)$	0.056	−0.014	−0.043		

the unbiased estimate of phenotypic variance as follows:

$$\sigma_P^2 = \sigma_G^2 + \sigma_\varepsilon^2 \tag{17.30}$$

In quantitative genetics, proportion of genetic variance over phenotypic variance is defined as the heritability in the broad sense, represented by H^2, i.e.,

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\varepsilon^2} \tag{17.31}$$

Therefore, applying the estimates of genotypic and error variances in Eq. (17.31), we can estimate the heritability in the broad sense of quantitative traits.

The heritability estimated by the formula (17.31) is based on single observations. In many cases, genetic analysis is based on phenotypic mean across replications, i.e., $\overline{P}_{i\cdot}$. In this case, we may want to estimate the heritability based on $\overline{P}_{i\cdot}$, and Eqs. (17.29), (17.30), and (17.31) become

$$\overline{P} = \mu + G + \overline{\varepsilon} \tag{17.32}$$

$$\sigma_{\overline{P}}^2 = \sigma_G^2 + \frac{1}{r}\sigma_\varepsilon^2 \tag{17.33}$$

$$H^2 = \frac{\sigma_G^2}{\sigma_{\overline{P}}^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{r}\sigma_\varepsilon^2} \tag{17.34}$$

Obviously, a higher heritability will be achieved, when phenotypic mean is used.

17.4.4 An Example on Yield of Ten Maize Inbred Lines in One Environment

To investigate the yield performance of ten maize inbred lines, randomized block design (RBD) is used, where each replication of the ten lines is arranged in one homogenous field block. Yield is measured on each plot, and raw data is given in Table 17.4. By raw data, we first calculate the mean for each row and mean for each column. The mean for each row is the phenotypic mean of the three replications of each inbred, and the mean for each column is the block mean. The result is respectively given in column 5 and row 13 in Table 17.4. Then we calculate the overall mean, i.e., $\overline{P}_{ik} = 2.80$, in Table 17.4. Finally we can calculate the genotypic effect and the block effect. Genotypic effect is the deviation of the phenotypic mean to the overall mean, and block effect is the deviation of the block mean to the overall mean. The result is respectively given in column 6 and row 14 in Table 17.4.

From the estimated block effect in the last row in Table 17.4, we can see that block effect may not be important. If we can ignore the block effect and use linear model (17.20), the ANOVA result is given in Table 17.5. It can be seen that the ten inbred lines show significant difference on yield. From the two mean squares, the error variance is estimated at 0.0473, and the genotypic variance is estimated at 0.4941. From

Table 17.5 Yield performance of ten maize inbred lines in three replications

Source	DF	SS	MS	Estimated variance	*F* value	*P* value
Genotype	9	13.768	1.530	0.494	32.346	0.000
Error	20	0.946	0.047	0.047		
Total	29	14.714				
R^2 (%)	93.571					
H^2 per plot	0.913					
H^2 per mean	0.969					

estimated variances, heritability in the plot level is estimated at 91.27 %, and heritability in the phenotypic mean level is estimated at 96.91 %.

17.5 Multiple Genotypes in Multiple Environments

17.5.1 Assumptions and Models

It is assumed that we make the field phenotyping experiment with r replications for a set of g genotypes in a set of e environments. The phenotypic means of the g genotypes in e environments are represented by $\mu_{ij}(i = 1, 2, \ldots, g; j = 1, 2, \ldots, e)$, which are unknown parameters. Error effects are normally distributed with a mean of 0 and an unknown variance ${\sigma_\varepsilon}^2$. So the observed phenotype for the ith genotype, jth environment, and kth replication is

$$P_{ijk} \sim N\left(\mu_{ij},\ \sigma_\varepsilon^2\right), (i = 1, 2, \ldots, g; j = 1, 2, \ldots, e; k = 1, 2, \ldots, r) \tag{17.35}$$

Therefore, we are handling a total of $g \times e$ normal distributions, corresponding to the g genotypes and e environments. The populations may have different means, but they should have the equal variance, which is actually the random error variance across environments. However, unequal error variances may occur if environments are highly heterogeneous. The unequal error variances between environments will be discussed in session 4.5.

Given the phenotypic means of g genotypes and e environments, we can define an overall phenotypic mean $\mu \hat{=} \frac{1}{ge}\sum_{i,j}\mu_{ij}$, phenotypic mean across environments $\overline{\mu}_{i\cdot} \hat{=} \frac{1}{e}\sum_{j}\mu_{ij}$, and environmental mean across genotypes $\overline{\mu}_{\cdot j} = \frac{1}{g}\sum_{i}\mu_{ij}$. We then define the genotypic effect (G_i) as the deviation of each phenotypic mean to the overall mean, environmental effect (E_j) as the deviation of each environmental mean to the overall mean, and genotype by environment interaction (GE_{ij}) as follows:

$$G_i \hat{=} (\overline{\mu}_{i\cdot} - \mu),\ \ E_j \hat{=} \left(\overline{\mu}_{\cdot j} - \mu\right), GE_{ij} \hat{=} \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \mu \tag{17.36}$$

Therefore, we can have the following linear model of phenotypic observations:

$$\begin{aligned} &P_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{ijk}, \\ &\varepsilon_{ijk} \sim N\left(0,\ \sigma_\varepsilon^2\right)\ (i = 1, 2, \ldots, g; j = 1, 2, \ldots, e; k = 1, 2, \ldots, r)\ \text{and iid} \end{aligned} \tag{17.37}$$

Variances corresponding to the three kinds of effects indicated in Eq. (17.36) can be defined as well, i.e.,

$$\sigma_G^2 \hat{=} \frac{1}{g-1}\sum_i G_i^2, \sigma_E^2 \hat{=} \frac{1}{e-1}\sum_j E_j^2, \sigma_{GE}^2 \hat{=} \frac{1}{(g-1)(e-1)}\sum_{i,j} GE_{ij}^2 \tag{17.38}$$

17.5.2 Estimation of Effects and Variances

The purpose of multi-environmental trials is to estimate the effects defined in distribution model (17.37) and variances defined in Eq. (17.38), so as to compare the performance of genotypes across environments. Now we will show how the genotypic effect G_i $(i = 1, 2, \ldots, g)$, environmental effect E_j $(j = 1, 2, \ldots, e)$, and interaction effect GE_{ij} $(i = 1, 2, \ldots, g; j = 1, 2, \ldots, e)$ can be estimated from observations P_{ijk} $(i = 1, 2, \ldots, g; j = 1, 2, \ldots, e; k = 1, 2, \ldots, r)$. First, we define the overall sample mean $(\overline{P}\ldots)$, i.e., the mean across the g genotypes, e environments, and r replications. Second, we define sample mean for each genotype and environment $(\overline{P}_{ij\cdot})$, i.e., the mean across the r replications for each genotype and each environment. Third, we define sample mean for each genotype $(\overline{P}_{i\cdot\cdot})$, i.e., the mean across the e environments and r replications for each genotype. Forth, we define sample mean for each environment $(\overline{P}_{\cdot j\cdot})$, i.e., the mean across the g genotypes and r replications for each environment. By calculating the above sample means, the deviation of phenotype to the overall sample mean can be decomposed as the following linear model:

$$P_{ijk} - \overline{P}\ldots = (\overline{P}_{i\cdot\cdot} - \overline{P}\ldots) + (\overline{P}_{\cdot j\cdot} - \overline{P}\ldots) + (\overline{P}_{ij\cdot} - \overline{P}_{i\cdot\cdot} - \overline{P}_{\cdot j\cdot} + \overline{P}\ldots) + (P_{ijk} - \overline{P}_{ij\cdot}) \tag{17.39}$$

The left side of Eq. (17.39) is the deviation of each phenotypic observation to the overall sample mean. On the right side of Eq. (17.39), the first term is the deviation of the genotypic sample mean to the overall mean, which can be used to estimate the genotypic effect defined in Eq. (17.36) or linear model (17.37). The second term is the deviation of the environmental sample mean to the overall mean, which can be used to estimate the environmental effect. The third term quantifies the interaction between genotype and environment, and the last term quantifies the residual random effect. So the effects defined in Eq. (17.36) or linear model (17.37) can be estimated as:

$$\begin{aligned} \hat{\mu} &= \overline{P}\ldots, \hat{G}_i = (\overline{P}_{i\cdot\cdot} - \overline{P}_{\cdot\cdot}), \hat{E}_j = (\overline{P}_{\cdot j\cdot} - \overline{P}\ldots), \\ \hat{GE}_{ik} &= (\overline{P}_{ij\cdot} - \overline{P}_{i\cdot\cdot} - \overline{P}_{\cdot j\cdot} + \overline{P}\ldots) \end{aligned} \tag{17.40}$$

Total sum square (SS_T) corresponds to the deviation on the left side of the model (17.39). Sum square of genotype (SS_G) corresponds to the first term on the right side of the model (17.39). Sum square of environment (SS_E) corresponds to the second term on the right side of the model (17.39). Sum square of interaction (SS_{GE}) corresponds to the third term on the right side of the model (17.39). Sum square of error (SS_ε) corresponds to the forth term on the right side of the model (17.39). That is,

$$\begin{aligned} SS_T &= \sum_{i,j,k} \left(\overline{P}_{ijk} - \overline{P}\ldots\right)^2 = SS_G + SS_E + SS_{GE} + SS_\varepsilon, \\ SS_G &= er\sum_i \left(\overline{P}_{i\cdot\cdot} - \overline{P}\ldots\right)^2, SS_E = gr\sum_j \left(\overline{P}_{\cdot j\cdot} - \overline{P}\ldots\right)^2, \\ SS_{GE} &= r\sum_{i,j} \left(\overline{P}_{ij\cdot} - \overline{P}_{i\cdot\cdot} - \overline{P}_{\cdot j\cdot} + \overline{P}\ldots\right)^2, SS_\varepsilon = \sum_{i,j,k} \left(P_{ijk} - \overline{P}_{ij\cdot}\right)^2 \end{aligned} \tag{17.41}$$

We have a total of $g \times e \times r$ independent observations. Total sum square (SS_T) has a degree of freedom of $ger - 1$. The lost one degree of freedom can be understood as being used in the estimation of the overall sample mean. Sum square of the estimated genotypic effects, i.e., SS_G, has a degree of freedom of $g - 1$, which is equal to the number of independent estimated genotypic effects. Sum square of the estimated environmental effects, i.e., SS_E, has a degree of freedom of $e - 1$, which is equal to the number of independent estimated environmental effects. Sum square of the estimated interaction effects, i.e., SS_{GE}, has a degree of freedom of $(g - 1)(e - 1)$, which is equal to the number of independent estimated interaction effects. There are $g \times e \times r$ estimated residual effects (Eq. 17.39), but they are not completely independent. The degree of freedom of $ge(r - 1)$ can also be understood as the number of independent estimated residual effects. Of course, it can also be found by subtracting $g - 1$, $e - 1$, and $(g - 1)(e - 1)$ from the total degree of freedom $ger - 1$.

Mean square (MS) is defined as the sum square divided by its degree of freedom. That is,

$$\begin{aligned} MS_G &= \frac{SS_G}{g-1}, MS_E = \frac{SS_E}{e-1}, MS_{GE} = \frac{SS_{GE}}{(g-1)(e-1)}, \text{and} \\ MS_\varepsilon &= \frac{SS_\varepsilon}{ge(r-1)} \end{aligned} \tag{17.42}$$

Intuitively, mean square of estimated genotypic effects reflects the magnitude of genotypic variance defined in Eq. (17.38). Mean square of estimated environmental effects reflects the magnitude of environmental variance. Mean square of estimated interaction effects reflects the magnitude of interaction variance. Mean square of the residual effects reflects the magnitude of error variance. In statistics, we can prove

$$\begin{aligned} E(SS_G) &= (g-1)\sigma_\varepsilon^2 + (g-1)er\sigma_G^2, \\ E(SS_E) &= (e-1)\sigma_\varepsilon^2 + g(e-1)r\sigma_E^2, \\ E(SS_{GE}) &= (g-1)(e-1)\sigma_\varepsilon^2 + (g-1)(e-1)r\sigma_{GE}^2, \text{and} \\ E(SS_\varepsilon) &= ge(r-1)\sigma_\varepsilon^2 \end{aligned}$$

Therefore,

$$E(MS_G) = \sigma_\varepsilon^2 + er\sigma_G^2,\ E(MS_E) = \sigma_\varepsilon^2 + gr\sigma_E^2,\ E(MS_{GE}) = \sigma_\varepsilon^2 + r\sigma_{GE}^2,\ \text{and}\ E(MS_\varepsilon) = \sigma_\varepsilon^2 \tag{17.43}$$

From Eq. (17.43), we can see that the expectation of MS_ε is equal to the error variance and therefore is the unbiased estimate of error variance. In addition to genetic variance, error variance is also included in the expectations of MS_G, MS_E, and MS_{GE}, respectively. After some

Table 17.6 ANOVA of multi-environmental phenotyping trials of multiple genotypes

Source of variation	Degree of freedom (DF)	Sum square (SS)	Mean square (MS)	Expected mean square (EMS)
Genotype	$g-1$	SS_G	MS_G	$\sigma_\varepsilon^2 + er\sigma_G^2$
Environment	$e-1$	SS_E	MS_E	$\sigma_\varepsilon^2 + gr\sigma_E^2$
Interaction	$(g-1)(e-1)$	SS_{GE}	MS_{GE}	$\sigma_\varepsilon^2 + r\sigma_{GE}^2$
Error	$ge(r-1)$	SS_ε	MS_ε	σ_ε^2
Total	$ger-1$	SS_T		

algebra manipulations, we can have the following unbiased estimates for the four variances defined in model 17.37 and Eq. (17.38):

$$\hat{\sigma}_G^2 = \frac{1}{er}(\mathrm{MS}_G - \mathrm{MS}_\varepsilon),\ \ \hat{\sigma}_E^2 = \frac{1}{gr}(\mathrm{MS}_E - \mathrm{MS}_\varepsilon),$$
$$\hat{\sigma}_{GE}^2 = \frac{1}{r}(\mathrm{MS}_{GE} - \mathrm{MS}_\varepsilon),\ \text{ and }\ \ \hat{\sigma}_\varepsilon^2 = \mathrm{MS}_\varepsilon \tag{17.44}$$

The above procedure can be summarized in Table 17.6. The following three F-statistics can be calculated to test the significance of the genotypic variation, environmental variation, and interaction variation compared with error variance, respectively:

$$F_G = \frac{\mathrm{MS}_G}{\mathrm{MS}_\varepsilon} \sim F[g-1,\ ge(r-1),]$$
$$F_E = \frac{\mathrm{MS}_E}{\mathrm{MS}_\varepsilon} \sim F[e-1,\ ge(r-1)],\ \text{ and}$$
$$F_{GE} = \frac{\mathrm{MS}_{GE}}{\mathrm{MS}_\varepsilon} \sim F[(g-1)(e-1),\ ge(r-1)]$$

When each replication of the g genotypes is arranged in one relatively homogeneous block in each environment, the block effect in each environment can also be estimated and included in the linear model (17.39). Same as single-environmental trials, the use of block can reduce the random error variance and improve the precision when comparing genotypes. In this case, the deviation of the block mean to the environmental sample mean estimates the block effect (represented by $B_{k(j)}$), i.e.,

$$\hat{B}_{k(j)} = \left(\overline{P}_{\cdot jk} - \overline{P}_{\cdot j\cdot}\right) \quad (j = 1, 2, \ldots, e; k = 1, 2, \ldots, r) \tag{17.45}$$

It should be noted that the block effects have to be defined in each environment. It does not make any sense to talk about the block effects across environments, as field blocks in one environment are totally different from blocks in other environments. Block is not a factor across environment, and block effects are nested in each environment. When the block effects are included, linear model (17.39) becomes

$$P_{ijk} - \overline{P}\ldots = \left(\overline{P}_{\cdot jk} - \overline{P}_{\cdot j\cdot}\right) + \left(\overline{P}_{i\cdot\cdot} - \overline{P}\ldots\right) + \left(\overline{P}_{\cdot j\cdot} - \overline{P}\ldots\right) + \left(\overline{P}_{ij\cdot} - \overline{P}_{i\cdot\cdot} - \overline{P}_{\cdot j\cdot} + \overline{P}\ldots\right)$$
$$+\left(P_{ijk} - \overline{P}_{\cdot jk} + \overline{P}_{\cdot j\cdot} - \overline{P}_{ij\cdot}\right) \tag{17.46}$$

Therefore, ANOVA can be done based on the above model. It can be seen from linear model (17.46) that including the block effect will not affect the estimation of genotypic effects, environmental effects, and interaction effects and will not affect the estimation of genotypic variance, environmental variance, and interaction variance, either. However, it will affect the estimation of residual effects and therefore the error variance. When the block effect is significant, estimated error variance will be lower than that from linear model (17.39). The reduced error variance allows more precise comparison of phenotypic means. In practice, other options are to estimate the block effect using Eq. (17.45), adjust the raw data by the block effect, and apply linear model (17.39) on the adjusted phenotypic observations.

17.5.3 Estimation of Heritability in the Broad Sense

As represented by the linear model (17.37), in multiple environments, phenotype of a quantitative trait for a given genotype or line or family can be decomposed into five parts: (1) overall mean across genotypes, environments, and replications; (2) genotypic effect of the specific genotype; (3) environmental effect of the specific environment; (4) genotype by environment interaction effect; and (5) random residual error. That is,

$$P = \mu + G + E + GE + \varepsilon \tag{17.47}$$

where overall mean, genotypic effects, environmental effects, and interaction effects are assumed to be unknown parameters (or fixed effects), and residual error is assumed to be a random variable. When random error has a normal distribution, the phenotypic variance ${\sigma_P}^2$ is equal to the sum of genotypic variance ${\sigma_G}^2$, environmental variance ${\sigma_E}^2$, interaction variance ${\sigma_{GE}}^2$, and error variance ${\sigma_\varepsilon}^2$. In session 4.2, we have seen that ANOVA can give the unbiased estimates of those variances. Therefore, we can have the unbiased estimate of phenotypic variance as follows:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2$$

In genetics, we are more concerned about the genetic variance and genotype by environment interaction. So environmental variance is normally excluded from phenotypic variance, i.e.,

$$\sigma_P^2 = \sigma_G^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2 \tag{17.48}$$

Similar to single-environmental trials, proportion of genetic variance over phenotypic variance is defined as the heritability in the broad sense, represented by H^2, i.e.,

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2} \tag{17.49}$$

Therefore, in applying the estimates of genotypic, interaction, and error variances in Eq. (17.49), we can estimate the heritability in the broad sense of quantitative traits.

The heritability estimated by the formula (17.49) is based on single observations. When genetic analysis is based on phenotypic mean across environments and replications, i.e., $\overline{P}_{i\cdot\cdot}$, Eqs. (17.47), (17.48), and (17.49) become

$$\overline{P} = \mu + G + \overline{\varepsilon} \tag{17.50}$$

$$\sigma_{\overline{P}}^2 = \sigma_G^2 + \frac{1}{er}\sigma_\varepsilon^2 \tag{17.51}$$

$$H^2 = \frac{\sigma_G^2}{\sigma_{\overline{P}}^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{er}\sigma_\varepsilon^2} \tag{17.52}$$

17.5.4 An Example on Yield of Ten Maize Inbred Lines in Three Environments

Ten maize recombination inbred lines (RILs) were evaluated across three environments, and the randomized block design (RBD) was used in each environment. In each environment, three replications of the ten lines are arranged in three homogenous field blocks. Yield was measured on each plot of replication, and raw data is given in Table 17.7. Please be noted that Table 17.4 actually shows the results from environment I. Let's first ignore the issue of heterogeneous environments and assume that the three environments have equal error variance. The issue of heterogeneous environments will be discussed in session 8.5.

Row mean represents the phenotypic mean of each RIL, which is shown in the second last column in Table 17.7. Column mean across the three replications represents each environmental mean, which is shown in the fourth last row in Table 17.7. Mean across the ten RILs, three environments, and three replications is the overall mean, which is equal to 3.13 (Table 17.7). The deviation of phenotypic mean to the overall mean is the genotypic effect, which is shown in the last column in Table 17.7. The deviation of environmental mean to the overall mean is the environmental effect, which is shown in the last third row in Table 17.7. Obviously, the ten genotypic effects have a sum of 0 and so have the three environmental effects. The interaction effects can be calculated by the formula (17.40) (not shown). Interaction effects have a total number

Table 17.7 Yield performance of ten maize inbred lines in three replications and three environments

	Environment I			Environment II			Environment III				
Genotype	Rep I	Rep II	Rep III	Rep I	Rep II	Rep III	Rep I	Rep II	Rep III	Phenotypic mean $(\overline{P}_{i\cdot})$	Genotypic effect $(\hat{G}_i)$
RIL1	2.56	2.66	2.43	2.25	2.34	2.25	4.09	4.19	4.01	2.976	−0.151
RIL2	2.66	2.50	2.75	2.61	2.21	2.75	1.35	2.12	2.81	2.418	−0.709
RIL3	2.93	2.97	2.70	2.79	3.11	3.06	3.89	3.20	3.15	3.089	−0.038
RIL4	2.57	2.21	1.80	4.01	3.69	3.23	3.83	4.73	4.33	3.378	0.251
RIL5	3.06	2.61	2.72	3.60	3.11	3.15	3.66	3.71	3.89	3.279	0.152
RIL6	1.94	2.16	2.14	2.16	2.93	1.71	4.27	3.17	4.32	2.756	−0.371
RIL7	1.85	1.69	2.25	2.39	2.03	2.57	3.73	3.65	1.88	2.449	−0.678
RIL8	3.83	3.58	3.80	3.74	5.04	4.46	3.33	3.30	3.38	3.829	0.702
RIL9	4.32	4.12	4.14	4.91	4.32	3.96	4.06	3.90	4.29	4.224	1.097
RIL10	2.81	3.33	2.81	3.69	3.15	3.51	2.89	1.55	2.12	2.873	−0.254
Environmental mean $(\overline{P}_{\cdot j\cdot})$	2.797			3.158			3.427			$\overline{P}\ldots = 3.127$	
Environmental effect $(\hat{E}_j)$	−0.330			0.031			0.300				
Block mean $(\overline{P}_{\cdot jk})$	2.853	2.783	2.754	3.215	3.193	3.065	3.510	3.352	3.418		
Block effect $(\hat{B}_{k(j)})$	0.056	−0.014	−0.043	0.057	0.035	−0.093	0.083	−0.075	−0.009		

of $g \times e$, with a sum of 0. Additionally, the g effects for each environment have a sum of 0, the e effects for each genotype have a sum of 0, and therefore the number of independent effects is equal to the degree of freedom of $(g-1) \times (e-1)$.

Column mean across the ten genotypes represents the block mean, which is shown in the second last row in Table 17.7. The deviation of block mean to the environmental sample mean is the block effect, which is shown in the last row in Table 17.7. So we have block effects for the three replications and the three environments. Obviously, the three block effects in each environment have a sum of 0.

When block effects are also ignored, Table 17.8 shows the combined ANOVA across the three environments. Actually, SS of genotype is equal to $e \times r$ times of the sum of the squared genotypic effects. SS of environment is equal to $g \times r$ times of the sum of the squared environmental effects. SS of interaction is equal to r times of the sum of the squared interaction effects. It can be seen from Table 17.8 that the ten RILs show significant difference on the yield. In addition, environmental effects and interaction effects are highly significant as well. From the four mean squares, error variance is estimated at 0.165, environmental variance at 0.094, genotypic variance at 0.312, and interaction variance at 0.220. From estimated variances, heritability in the plot level is estimated at 44.8 %, and heritability in the phenotypic mean level is estimated at 94.5 %.

17.5.5 Estimation of Genotypic Value in Heterogeneous Environments

In multiple-environmental trials, it is generally assumed that different genotypes in a specific environment have the same error variance. This assumption may be unrealistic when environmental conditions are heterogeneous or when the data span a long time period. Several sources of heterogeneous variances are identified to make the environments heterogeneous, including temperature, water, soil, pest, etc. There are several useful tests of the homogeneity of variance assumption. Here we show how to use Bartlett's test to check if variances are homogenized. Let $\hat{\sigma}^2_{\varepsilon_j}$ and df_{ε_j} be error variance for the jth environment and its degree of freedom, respectively, and then null hypothesis and alternative hypothesis are

$$H_0 : \sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \cdots = \sigma^2_{\varepsilon_e}, \text{ and}$$
$$H_A : \text{at least two of } \sigma^2_{\varepsilon_1}, \sigma^2_{\varepsilon_2}, \ldots \text{ and } \sigma^2_{\varepsilon_e} \text{ are not equal.}$$

Under null hypothesis, the combined error variance ${\sigma_\varepsilon}^2$ can be obtained by individual error variances of the e environments, that is,

$$\sigma^2_\varepsilon = \frac{1}{\sum_j df_{\varepsilon_j}} \sum_j df_{\varepsilon_j} \times \sigma^2_{\varepsilon_i} \tag{17.53}$$

Bartlett's statistics approximately follows χ^2 distribution with degree of freedom $e-1$, that is,

$$\chi^2 = \left(\sum_j df_{\varepsilon_j}\right) \ln\left(\sigma^2_\varepsilon\right) - \sum_j df_{\varepsilon_j} \times \ln\left(\sigma^2_{\varepsilon_i}\right) \sim \chi^2(e-1) \tag{17.54}$$

Under heterogeneous environments, the mean performance of one genotype is assumed to be μ, error variance in the jth environment is $\sigma^2_{\varepsilon_j}$, and P_j is its phenotypic value in the jth environment. Therefore, the linear model is

Table 17.8 ANOVA of the multi-environmental trial shown in Table 17.7

Source	DF	SS	MS	Variance	*F* value	*P* value
Environment	2	5.996	2.998	0.094	18.219	0.000
Genotype	9	26.753	2.973	0.312	18.064	0.000
Interaction	18	21.686	1.205	0.220	7.322	0.000
Error	60	9.873	0.165	0.165		
Total	89	64.308				
R^2 (%)	84.647					
LSD ($P = 0.05$)	1.087					
LSD ($P = 0.01$)	1.489					
H^2 per plot	0.448					
H^2 per mean	0.945					

$P_i = \mu + \varepsilon_j, \varepsilon_j \sim N\left(0,\ \sigma^2_{\varepsilon_j}\right) (j = 1, 2, \ldots, e)$ and independent

In this case, the simple mean $\overline{P}. = \frac{1}{e}\sum_j P_j$ is still an unbiased estimation of μ but not the best one. That is to say, there are other estimates having smaller variance than the simple mean. By calculating the variance of the linear combination of P_j ($j = 1, 2, \ldots, e$), BLUE of μ can be found as

$$\hat{\mu} = \sum_j w_j P_j, \quad \text{where}$$

$$w_j = crdisplaystyle \frac{\frac{1}{\sigma^2_{\varepsilon_j}}}{\frac{1}{\sigma^2_{\varepsilon_1}} + \frac{1}{\sigma^2_{\varepsilon_2}} + \cdots + \frac{1}{\sigma^2_{\varepsilon_e}}} \tag{17.55}$$

The variance of $\hat{\mu} = \sum_j w_j P_j$ can be found as

$$V(\hat{\mu}) = \frac{1}{\frac{1}{\sigma^2_{\varepsilon_1}} + \frac{1}{\sigma^2_{\varepsilon_2}} + \cdots + \frac{1}{\sigma^2_{\varepsilon_e}}}. \tag{17.56}$$

The variance given in Eq. (17.56) is the least among all possible unbiased linear combinations of P_j ($j = 1, 2, \ldots, e$). If and only if environments are homogeneous, $V(\hat{\mu})$ and $V\left(\overline{P}.\right)$ are equal. That is to say, when environments are heterogeneous, weighted mean is a better estimate than simple mean, in the sense of least variance. When an environment has smaller error variance, a higher weight should be given for this environment.

To illustrate the effectiveness of weighted mean, we assume that there are two environments and $\sigma^2_{\varepsilon_2} = s\sigma^2_{\varepsilon_1}$. Then, we have

$$V(\hat{\mu}) = \frac{s}{1+s}\sigma^2_{\varepsilon_1},\ V\left(\overline{P}\right) = \frac{1}{4}(1+s)\sigma^2_{\varepsilon_1},\ \text{and}\ \frac{V(\hat{\mu})}{V\left(\overline{P}\right)} = \frac{4s}{(1+s)^2}$$

We can clearly see the ratio of variance of weighted mean and simple mean in Fig. 17.3. When $s = 1$, that is, $\sigma^2_{\varepsilon_2} = \sigma^2_{\varepsilon_1}$, $V(\hat{\mu}) = V\left(\overline{P}\right)$. When $s \neq 1$, $V(\hat{\mu})$ is always smaller than $V\left(\overline{P}\right)$.

$\sigma^2_{\varepsilon_1} \neq 0$ and $\sigma^2_{\varepsilon_2} = 0$ ($s = 0$) are an extreme case, which indicates that the second environment does not have any error. In this case, the observation in the second environment is equal to the phenotypic mean. The error variance in the first environment is nonzero, which indicates that the observation in the first environment may be deviated from the phenotypic mean. In this case, the observation in the second environment is the best estimate of μ. Including observations from the first environment may cause deviation from the phenotypic mean.

$\sigma^2_{\varepsilon_1} \neq 0$ and $\sigma^2_{\varepsilon_2} = \infty$ represent another extreme case. Observations in the second environment have nothing to do with μ. In this case, observation in the second environment is complete random error, which does not contain any useful information about the phenotypic mean to be estimated. Including observations from the second environment may cause more deviation from the phenotypic mean. Thus observations in the first environment are the best estimate of μ.

Now, we revisit the data in Table 17.7. Table 17.9 summarizes the ANOVA results in the three environments. The three error variances were estimated respectively at 0.0473, 0.1525,

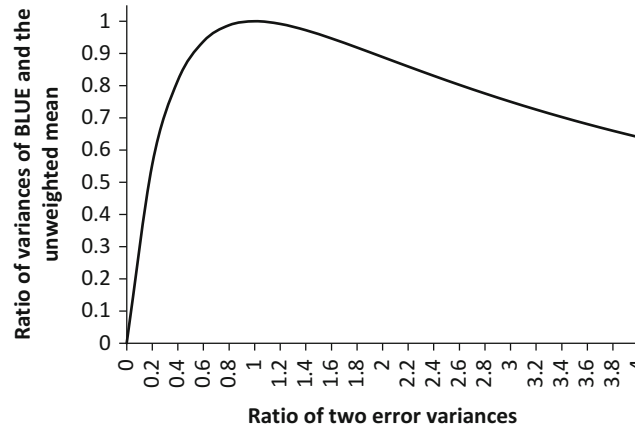


Fig. 17.3 The variance ratio of BLUE and the simple unweighted mean

Table 17.9 ANOVA in each environment using data in Table 17.7

	Mean square				Estimated variance		Heritability	
Environment	Genotype	Error	*F* value	DF of error	Genotype	Error	Per plot	Per mean
I	1.530	0.047	32.346	20	0.494	0.047	0.913	0.969
II	2.044	0.153	13.404	20	0.630	0.153	0.805	0.925
III	1.809	0.294	6.154	20	0.505	0.294	0.632	0.838

and 0.2939, with the same degree of freedom of 20. If the null hypothesis $H_0 : \sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \sigma^2_{\varepsilon_3}$ is true, the combined error variance is estimated $\hat{\sigma}^2_{\varepsilon} = 0.1646$ by Eq. (17.53). The Bartlett's χ^2 statistic defined in Eq. (17.54) has a value of 14.86, and its degree of freedom is 2. Hence the significance probability can be found at $P = 0.0003$, which is highly significant. The high significance from the χ^2 test indicates the heterogeneity among the three environments. In theory, it is not appropriate to conduct the combined ANOVA as Table 17.8, when the environments are heterogeneous. Instead, ANOVA should be conducted for each environment, as shown in Table 17.9.

For comparison, Table 17.10 gives simple means and BLUE (or weighted means) of the ten RILs and ranks of the ten RILs from simple mean and BLUE. Weights in BLUE are 0.68, 0.21, and 0.11 for the three environments in calculating BLUE (last row in Table 17.10). Environment I has the least error variance of 0.047 (Table 17.9) and therefore has the highest weight in BLUE. Environment III has the largest error variance of 0.294 (Table 17.9) and therefore has the lowest weight in BLUE. The use of weighted mean does not change the ranks of the two top RILs but gives quite different ranks for other RILs. As indicated before, BLUE has the least variance compared to any other unbiased linear estimates. Considering the highly heterogeneous environments, BLUE given in Table 17.10 is expected to be much closer to the true phenotypic mean of each RIL and therefore should be used in further genetic studies, such as QTL mapping.

The sample mean across the three replications of each RIL has a variance at 0.016 in environment I, 0.051 in environment II, and 0.098 in environment III (Table 17.10), which is equal to the estimated error variance divided by the number of replications. Therefore, variance of the simple mean and variance of BLUE and their standard error (SE) are

Table 17.10 Comparison of simple mean and BLUE and using data in Table 17.7

	Environment						
Genotype	I	II	III	Simple mean	Rank	BLUE	Rank
RIL1	2.55	2.28	4.10	2.976	6	2.662	7
RIL2	2.64	2.52	2.09	2.418	10	2.555	8
RIL3	2.87	2.99	3.41	3.089	5	2.954	5
RIL4	2.19	3.64	4.30	3.378	3	2.726	6
RIL5	2.80	3.29	3.75	3.279	4	3.007	3
RIL6	2.08	2.27	3.92	2.756	8	2.321	9
RIL7	1.93	2.33	3.09	2.449	9	2.141	10
RIL8	3.74	4.41	3.34	3.829	2	3.838	2
RIL9	4.19	4.40	4.08	4.224	1	4.222	1
RIL10	2.98	3.45	2.19	2.873	7	2.993	4
Variance	0.016	0.051	0.098				
Weight	0.680	0.211	0.109				

$$V(\hat{\mu}) = \frac{1}{\frac{1}{0.016} + \frac{1}{0.051} + \frac{1}{0.098}} = 0.011,$$

$$\mathrm{SE}(\hat{\mu}) = \sqrt{0.011} = 0.104, V(\overline{P})$$

$$= \frac{1}{3} \times 0.016 + \frac{1}{3} \times 0.051 + \frac{1}{3} \times 0.098$$

$$= 0.055, \quad \mathrm{SE}(\overline{P}) = \sqrt{0.055} = 0.234$$

Obviously, the weighted mean has much smaller variance and SE compared with the unweighted mean. The error variance estimated above can be used in significance test between the ten RILs, as normally done in ANOVA.

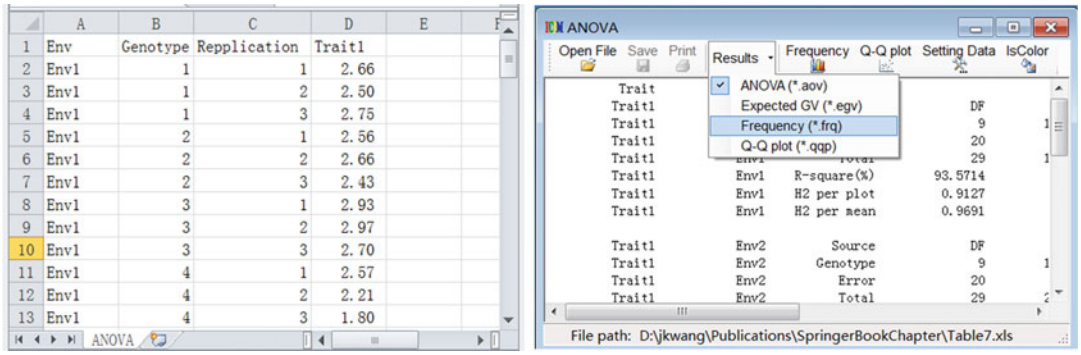
17.6 A Computer Tool for Analyzing Multi-environmental Trials

QTL IciMapping (freely available from www.isbreeding.net) was an integrated software for linkage map construction and QTL mapping. For multi-environmental trials, a tool called ANOVA is implemented in the software to estimate the genetic variance and heritability in broad sense from phenotypic data. The data format can be in CSV, XLS, or XLSX. If in format of XLS or XLSX, the sheet name must be "ANOVA" (left of Fig. 17.4). In sheet ANOVA, the first row is for Environment, the second row is for Genotype, the third column is for Replication, the forth column is for the first trait, the fifth column is for the second trait, and so on (left of Fig. 17.4). The first three columns can be either number or string. Columns for traits must be numbers. Missing trait values were denoted as −100.00.

From output, the users can find standard ANOVA tables for each environment and combined analysis across environments for each trait in AOV file, and the expected genotypic values for each environment and combined analysis for each trait in EGV file. Besides, frequency histogram (FRQ file) and Q-Q plot (QQP file) can be shown by selecting corresponding menus for raw phenotype and expected genotype per replication, per trait, and per environment (Right of Fig. 17.4).

17.6.1 Objectives in Phenotyping Complex Traits

Genotypic value can be estimated by marker loci or by known quantitative trait loci. But even if marker loci or quantitative trait loci are not analyzed, the relative magnitudes of additive, dominance, and epistatic effects across unknown loci can be estimated from an analysis of the phenotypic value. In classic quantitative genetics, the number of genes controlling the trait of interest can be estimated by Castle-Wright formula (Sect. 17.2.3). The segregation analysis can



	A	B	C	D	E
1	Env	Genotype	Repplication	Trait1	
2	Env1	1	1	2.66	
3	Env1	1	2	2.50	
4	Env1	1	3	2.75	
5	Env1	2	1	2.56	
6	Env1	2	2	2.66	
7	Env1	2	3	2.43	
8	Env1	3	1	2.93	
9	Env1	3	2	2.97	
10	Env1	3	3	2.70	
11	Env1	4	1	2.57	
12	Env1	4	2	2.21	
13	Env1	4	3	1.80	

Trait			
Trait1			DF
Trait1			9
Trait1			20
Trait1			29
Trait1	Env1	R-square(%)	93.5714
Trait1	Env1	H2 per plot	0.9127
Trait1	Env1	H2 per mean	0.9691
Trait1	Env2	Source	DF
Trait1	Env2	Genotype	9
Trait1	Env2	Error	20
Trait1	Env2	Total	29

Fig. 17.4 Multi-environmental phenotyping data in Excel (*left*) and the interface of the ANOVA computing tool (*right*)

be used to estimate genetic parameters of the variation of a quantitative trait, including additive and dominance effects, additive and dominance variance, and heritability for both major genes and polygenes (Gai and Wang 1998). Therefore, only by phenotypic data, we can distinguish the effects of major genes from polygenes and/or environments, which is important for understanding the expression of a major gene in relation to its genetic background and for predicting the segregation of a cross in breeding.

17.6.2 The Three Basic Principles of Field Experimental Design in Phenotyping Complex Traits

In field experiments, environmental condition may vary from one stage of the experiment to the next. Some factors may not be possible to isolate from others, thus forcing the investigation of several factors jointly. On top of that, measurement errors may introduce unwanted error into the system. Therefore, precautionary measures need to be taken. To design the experiment in a better way, R. A. Fisher has enumerated three principles of experimental designs (Fisher 1926): (1) the principle of local control, (2) the principle of randomization, and (3) the principle of replication. These are discussed in details in Chap. 16 and briefly described below.

The principle of local control eliminates the variability caused by extraneous factors can be measured. This means that we should plan the experiment in a manner that we can perform a three-way ANOVA, in which the total variability of the data is divided into four components attributed to treatments (genotype in our case), environments (in our case), extraneous factor (e.g., soil fertility), and experimental error. In other words, in each environment, we first divide the field into several homogeneous parts, known as blocks, and then each such block is divided into parts equal to the number of genotypes. In general, blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of a three-way ANOVA (Eq. 17.46 in Sect. 17.4.2). For increasing the statistical accuracy of the experiment, the principle of replication is required in which the experiment is repeated more than once. Thus, each treatment is applied in many experimental units instead of one. For example, when considering multiple genotypes in one environment (Sect. 17.3), more replications give more precise estimate of the true genotype (Eq. 17.25). Sometimes the entire experiment can be repeated several times for better results. It should be remembered that replication is introduced in order to increase the precision of a study, that is to say, to increase the accuracy with which the main effects and interactions can be estimated. Finally, principle of randomization provides protection against the effects of extraneous factors by randomization, and each treatment has equal opportunity to get a

place where soil fertility is good or bad. For instance, if we grow one variety of maize, say, in the first half of the parts of a field and the other variety is grown in the other half, then it is just possible that the soil fertility may be different in the first half in comparison to the other half. If this is so, our results would not be realistic. In such a situation, we may assign the variety of rice to be grown in different parts of the field on the basis of some random sampling technique. Through the application of the principle of randomization, we can have a better estimate of the well-known experimental error.

17.6.3 Quality Control of Phenotype

Phenotypic values are usually measured in multiple environments/locations/years, each with several replications. Measurement procedure is fraught with sources of potential error, which may arise in the observer, in the plant, or in the overall application of the technique. Therefore quality assurance and quality control of phenotype are essentially required to insure the success of genetic study. Generally, quality control of phenotype should be conducted right after the phenotypic values are out of the measurements. Maximum value, minimum value, mean, variance, histogram, variance of replications, heritability in broad sense, etc. are the easiest and most effective statistics to evaluate the raw phenotypes.

- Outliers (or unusual values) are values that lie outside the usual range of phenotype of the trait of interest. They can seriously affect the results of analyses. There are two aspects in dealing with outliers, identifying them and dealing with them. There are formal tests for detecting outliers (Miller 1993; Sokal and Rohlf 1995), but they can be easily highlighted by the distribution of phenotype, that is, histogram plot. Outliers would make either maximum value or minimum value to be anomalous and make variance abnormally large. Once we identify outliers, we should first check to make sure they are not a mistake, such as an error typing in your data or in writing values down. They often show up as impossible values, for example, −5 cm for plant height. If you can classify an outlier as a mistake, it should be deleted. If you have no reason to suspect an outlier as being a mistake, you can do the genetic study without the outlier to see how much they influence the outcome of the analysis. If the conclusions are altered, then you should try and determine why those values are so different. Perhaps there was contamination during pollen; plants were from different subpopulation, etc.
- The shape of the distribution of phenotype can be examined by plotting a histogram. Is the distribution symmetrical or skewed? Is it unimodal or multimodal? We may find a priori like biological or physiological reasons to explain this distribution. In some cases, genetic models for the trait of interest can be estimated by the distributions (Gai and Wang 1998; Wang et al. 2001).
- The large variance of replications should be argued. Replication demonstrates the results to be reproducible, at least under the current experimental conditions. Large variance of replications indicated that the repetitiveness of the experiments is poor, the precision for estimates of genotype mean is low, and the experimental error variance is large, which may cause the low heritability in broad sense.
- Heritability in broad sense is low. Heritability is a concept that summarizes how much of the variation in a trait is due to variation in genetic factors. We have demonstrated how to estimate heritability in broad sense by phenotypic values in previous sections. A low heritability means that of all observed variations, a small proportion is caused by variation in genotypes (Visscher et al. 2008). That is to say, the heritability of a group of individuals with relatively similar heredities is relatively low, and the phenotype of an individual is not a good predictor of the genotype. In many gene-mapping experiments, the probability of detecting a gene of large effect increases with heritability (Bradford and Famula 1984; Oliver et al. 2005; Weedon et al. 2007). Therefore, low heritability implies that the follow-up genetic study may be not efficient.

17.6.4 Fixed Effect or Random Effect

Analysis based on phenotypic values in Sects. 17.2, 17.3, 17.4, and 17.5 is under assumption that all interested effects, including genotypic effect, environmental effect, genotypic by environment interaction effect, and block effect, are fixed effects. It is natural to ask when and which explanatory variables (also called independent variables) to give random effects. Conceptually, effects of variables might be treated as random if we can think of the levels of the variable that we included in the study as a sample drawn from some larger population of levels that could (in principle) have been selected. Practically, one key difference between fixed and random effects is in the kind of information we want from the analysis of the effects.

In the case of fixed effects, we are usually interested in making explicit comparisons of one level against another. For example, we would want to compare the yield mean in Beijing to that in New York in an experiment. If explicit comparison of the levels of a variable against one another is the goal of the research, then the levels of the variable are usually treated as "fixed." If, on the other hand, our primary interest is in the effects of other variables or treatments across the levels of a factor (e.g., the effect of block on yield, across genotypes from three environments), that is to say, we assumed that the block effect varies randomly within the population of environments, and the researcher is interested to test and estimate the variance of these random effects across this population. Then the block variable might be treated as a "random" effect. In this chapter, we assumed that all the related effects are fixed.

17.6.5 Conclusion

Phenotypic variability may be caused by genotype and environmental factors. Therefore plant geneticists are interested to dissect phenotype and to identify the genes that play important roles in the inheritance of phenotype. They try to explain the role of those genes in relation to one another and in relation to the environment. Genetic mapping correlates the phenotype with the genotype of genetic markers, which are expected to be located close to the genes (or genomic regions) of interest. The relationships between the phenotype and gene of interest can be weakened if phenotypic variability is underestimated or overestimated. Therefore, traits should be measured reproducibly on a large number of samples, and biometrical techniques can be useful to determine the true genotypic value. This can help to locate, enumerate, and annotate genes and to assign known or putative biochemical functions. However, only about two-thirds of all genes have an assigned biochemical function, and only a fraction of those are associated with a phenotype. Therefore some efforts are needed on phenomics (Bochner 2003) in order to accelerate the pace of the discovery of genes using advanced tools and techniques of biometrics.

Acknowledgment This research was supported by the HarvestPlus Challenge Program of CGIAR.

References

Bochner BR (2003) New technologies to assess genotype–phenotype relationships. Nat Rev Genet 4:309–314

Bradford GE, Famula TR (1984) Evidence for a major gene for rapid postweaning growth in mice. Genet Res 44:293–308

East EM (1911) A Mendelian interpretation of variation that is apparently continuous. Am Nat 44:65–82

Fisher RA (1926) The arrangement of field experiments. J Minist Agric Great Brit 33:503–513

Gai J, Wang J (1998) Identification and estimation of a QTL model and its effects. Theor Appl Genet 97:1162–1168

Miller JN (1993) Outliers in experimental data and their treatment. Analyst 118:455–461

Oliver F, Christians JK, Liu X, Rhind S, Verma V, Davison C, Brown SDM, Denny P, Keightle PD (2005) Regulatory variation at glypican 3 underlies a major growth QTL in mice. PLoS Biol 3:e135

Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn. WHFreeman, New York

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era – concepts and misconceptions. Nat Rev Genet 9:255–266

Wang J, Podlich DW, Cooper M, DeLacy IH (2001) Power of the joint segregation analysis method for testing mixed major-gene and polygene inheritance models of quantitative traits. Theor Appl Genet 103:804–816

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B, Zeggini E, Lango H, Lyssenko V, Timpson NJ, Burtt NP, Rayner NW, Saxena R, Ardlie K, Tobias JH, Ness AR, Ring SM, Palmer CN, Morris AD, Peltonen L, Salomaa V, Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, Davey Smith G, Groop LC, Hattersley AT, McCarthy MI, Hirschhorn JN, Frayling TM (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39:1245–1250