**Review of Quantitative Methods used in Plant Breeding**

This module is a review of materials that will be needed for understanding information covered in this introductory course. **Students should cover the information in this module before the class is taught.** It is important for students to honestly self-evaluate their understanding of the basic concepts covered in this module before they continue with the remainder of the course materials.

Topics and objectives covered in this module:

*Plant Breeding Basics*

Categorize plant breeding activities within the framework of genetic improvement, cultivar development and product placement.

*Trait Measures*

Demonstrate ability to distinguish among the various types of phenotypic and genotypic traits that are assessed routinely in a plant breeding program.

*Exploratory Data Analysis*

* Demonstrate understanding on descriptive and inferential statistics
* Demonstrate ability to conduct and interpret exploratory data analyses
* Distinguish parameters from estimators and estimates
* Estimate means, in both balanced and unbalanced data sets
* Estimate variances, covariances and correlations
* Download and install R and R Studio.
* Conduct exploratory data analyses (EDA) on data from a simple Completely Randomized Design (CRD).
* Interpret results from EDA.
* Conduct Analysis of Variance (ANOVA) on data from a CRD.

*Hypothesis Tests*

Interpret types of errors that can be made from testing various kinds of hypotheses.

*Analysis of Variance*

Demonstrate ability to conduct and interpret Analysis of Variance.

Demonstrate ability to conduct and interpret regression analyses.

Demonstrate ability to conduct and interpret Analysis of Covariance

**Lesson Map**

* Plant breeding
  + Types of plant breeding projects
* Quantitative methods
  + Types of Measurements
  + Principles of Experimental Design
  + Important Field Plot Designs
* Models
  + Data Models
  + Phenotype Models
* Exploratory Data Analyses
  + Estimation
    - Parameters, estimators, estimates
    - Means,
    - Variance
      * Components of variances
    - Covariance
    - Regression
  + Prediction
* Analysis of Variance
  + Linear Models
  + Expected Mean Squares
  + With covariates
* Mixed Model Equations
  + Shrinkage and Prediction
  + BLUE’s and BLUPs
* Decisions Using Statistical Inference
  + Hypothesis tests
  + Types of decision errors
  + Significance thresholds
  + Decision metrics
* Appendix 1: Matrix Algebra
  + Operational rules
    - Some simple hand calculations
    - Some simple EXCEL calculations
* Appendix 2: Computational Considerations
* References

**Plant Breeding**

Historically plant breeding has been defined as the art and science of the genetic improvement of domesticated plants. While plant breeders use advanced technologies and scientific knowledge to change, modify and shape the ability of plants to provide useful products, is plant breeding is a science? In other words: Are there any fundamental theorems of plant breeding that can be falsified with experimental evidence (Popper 1959) ? If not, then it is difficult to classify plant breeding as a science. Plant breeding is a decision making discipline that uses the scientific method to help make decisions, so perhaps the following definition is better suited:

*Plant Breeding consists of decision making activities designed to improve the genetic potential of plant species to produce products that are useful for humans.*

This definition implies that a decision maker designs and applies a process to a population of plants resulting in genetic changes that are valued because they confer desirable characteristics for humans. Current breeding programs are the result of thousands of years of refinements that have been implemented through considerable trial and error. Plant breeding processes are constrained by limited resources, technologies and the reproductive biology of the species. Thus, plant breeding may be better considered as the engineering counterpart to plant biology.

*Other Definitions*

* Art of plant breeding: … the ability to discern fundamental differences of importance in available plant materials and to select and increase the more desirable types…(Hayes and Immer 1942)
* “Plant breeding, broadly defined, is the art and science of improving the genetic pattern of plants in relation to their economic use.”(Fehr 1991; Smith 1966).
* “Plant breeding is the science, art, and business of improving plants for human benefit.”(Bernardo 2010).

**Types of Plant Breeding Projects**

As with engineering projects plant breeding projects should have a set of ***measurable goals*** based on the intended product (outcome). Explicitly the first step in designing a breeding project is to provide specifications for the desired outcomes. Outcomes are usually defined as ***cultivars*** with improved characteristics such as 5% greater yield, complete disease resistance, member of maturity group, etc. Once the specifications are defined using measurable attributes, ***processes*** and ***projects*** can be ***designed*** to produce plants with the desired attributes.

Develop replicable lines

It is important to understand distinctions among types of plant breeding projects: genetic improvement, cultivar development, trait introgression and product placement. The distinctions among these types of projects are nuanced aspects of plant breeding programs, yet the distinctions are critical for specifying models used in data analyses and decision-making.

Create Useful Genetic Variability

Regional Trials

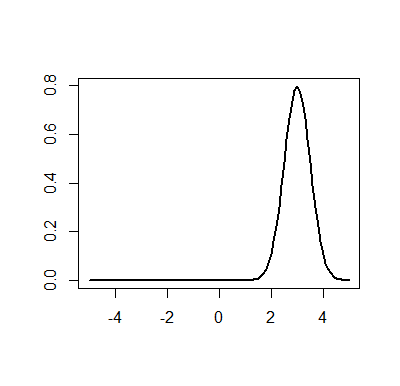
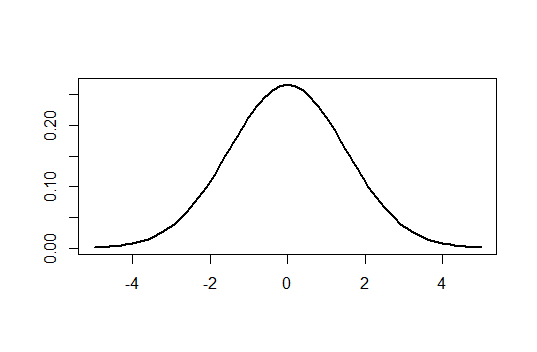
Advanced Yield

Trials

Small

Strip Trials

Commercialization



Select lines to cross

Preliminary Trial

Strip

Trials

Develop replicable progeny

**Figure 0.1**

The primary goal of a ***genetic improvement*** (red) projects is to improve the genetic potential of the ***breeding population***. Typically this is accomplished through a recurrent cycle of creating replicable groups of genotypes, such as clones, lines, hybrids, synthetics, followed by identifying and selecting those with desirable characteristics to cross in a breeding nursery. Realized genetic gain, is a measureable metric that can be used to determine if the goal has been met.

Selection among and within ***segregating families*** of pure-line varieties, synthetics, hybrids, or clones is accomplished with phenotypic assays of field plots in single and Multi-Environment Trials (METs) as well as genotypic assays of molecular markers that are associated with desirable traits. Data analyses will include analyses of binary traits with binomial and multinomial models and quantitative traits with mixed linear models. In the early stages of field trials, environments are modeled as fixed (nuisance) effect parameters, while replicable genotypes are modeled as random effects.

The primary goal of a ***cultivar development*** project (blue filters) is to identify replicable groups that have potential to be grown by farmers throughout a targeted population of environments, a.k.a. market segment. Thus, in a cultivar development project replicable genotypic units sampled from segregating populations will be evaluated for quantitative traits in multi-environment trials (METs). Analyses of data from advanced METs also use mixed linear models, although for the advanced trials cultivars are modeled as fixed effects while the environments are modeled as random effects.

The goals of ***product placement*** projects are to identify the best combinations of cultivars, agronomic management and field environments to maximize profitability for the farmer. In a product placement project agronomic management practices as well as developed cultivars represent designed treatments applied to field plots. These are often organized in hierarchical (split plot) experimental designs. Thus, the parameters of a mixed linear model associated with agronomic practices as well as cultivars will be modeled as fixed effects, while various levels of residual variability associated with split plot experimental units will be modeled as random effects.

For this introductory course on Quantitative Genetics we will focus primarily on genetic improvement, a little bit on cultivar development projects and no time will be spent on product placement projects.

Conceptually, genetic improvement consists of a simple two-step, iterative, decision making process: 1) select pairs of parents to cross and 2) evaluate their segregating progeny to provide metrics that can be used to select the next generation of parents. Operational implementation of genetic improvements for any crop species requires far more detail. For example (Byrum et al. 2016) identified at least 200 binary decisions that need to be made in a soybean variety development program.

The details of any particular breeding program will likely consist of many activities that are organized based on project goals, budget and reproductive biology. Plant breeding projects historically have been developed ‘backwards’, i.e., with the designed product, goals and constraints in mind. If the objectives and constraints are clearly stated, they can be translated into mathematical functions that can be used to find optimal solutions to the trade-offs that will be required.

**A Review of Quantitative Methods.**

Quantitative analytic methods provide metrics that enable plant breeders to make better decisions.

**Types of Measurements**

Quantitative genetics provides genetic models to explain and predict changes in quantitative traits over generations of crossing and selection. Recall traits can be evaluated on ***categorical*** or ***continuous*** scales. If the trait of interest is evaluated based on some quality (examples include disease resistance, flower color, presence/absence of a molecular marker, developmental phase, etc.) then it is considered a categorical trait. There are three further distinctions of categorical scales:

**Binary** consist of only two categories such as resistant and susceptible. For example presence or absence of a SNP allele.

**Nomina**l consist of unordered categories. For example, viral disease vectors might be categorized as insects, fungi or bacteria;

**Ordinal** consist of categorical data where the order is important. For example, disease symptoms might be classified as none, low, intermediate and severe.

Binary, nominal and ordinal data are typically analyzed using General**ized** Linear Models. Such models require that we model the error structures using Poisson or Negative Binomial distributions and are beyond the scope of introductory quantitative methods and genetics (see McCullagh and Nelder, 1989 or Christensen, 1997 for descriptions of General**ized** Linear Models). It is important to remember that it is not advisable to apply General Linear Models to categorical responses

There are two distinctions of traits that are evaluated on **quantitative** scales:

**Discrete** data occur when there are gaps between possible values. These type of data usually involve counting. Examples include flowers per plant, number of seeds per pod, number of transcripts per sample, etc.

**Continuous** data can be measured with instruments and are only limited by the precision of the measuring technology. Examples include plant height, yield per unit of land, seed weight, seed size, protein content, etc.

In the context of measurement, **Precision** refers to the level of detail in the scale of the measurement. **Accuracy** refers to whether the measurement represents the true value.

**Principles of Experimental Design**

Just as design is the most important decision affecting the type of breeding project, design of experiments used for obtaining data to help make decisions is critical and should be determined based on the project objectives long before replicable genotypic groups are placed in the field. In other words the experimental objectives need to be aligned with project objectives and clearly stated before designing the experiment. Often, these objectives are stated as testable hypotheses. Once the objectives are clearly stated, the design of the experiment(s) need to clearly define **experimental units**, **treatment structure** and **design structure**. Subsequently, the principles of **randomization** and **replication** need to be applied when assigning treatments to homogeneous groups of experimental units.

Experimental designs consist of design structures, treatment structures, and allocation of these structures to experimental units. Experimental units for field breeders usually consist of plots of land, or greenhouse pots. The primary treatment designs of interest involve allocation of replicable genotypes to the experimental units. The development of replicable genotypes is accomplished primarily through reproductive biology, although with the emergence of biotechnologies, such as protoplast fusion, tissue culture and various transgenic technologies, there are many ways to develop replicable genotypic treatments. Would you consider treatments from these biotechnologies as fixed or random effects? Why? Experimental units can be split in both time and space, resulting in the ability to apply treatment and design structures to different sized experimental units.

Design principles in allocation of treatments (genotypes) to experimental units include **Randomization, Replication** and **Blocking**. These are principles rather than rigid rules. As such they provide flexibility in designing experiments to draw inferences about biological questions. Assuming that these principles are applied appropriately, experimental data can be used for obtaining unbiased estimates of treatment effects, variances, covariances and predictions of breeding values.

**Important Field Plot Designs**

Typical design structures utilized by plant breeders include Randomized Complete Block, Incomplete Block and Augmented Designs. **Completely randomized designs** are often used to help the novice learn concepts such as randomization and replication of treatments that are to be applied to **experimental units** under homogeneous conditions. An **experimental unit** is defined as the basic unit to which a treatment will be applied. A **sampling unit** is defined as a discrete representative from a population of interest. However, because homogeneous conditions are very rare, especially for large experimental units or large numbers of treatments, the completely randomized design merely provides a motivation for blocking homogeneous experimental (and sampling) units. If a complete set of treatments can be randomly assigned to all of the experimental units of a homogenous block, then the design is known as a **randomized complete block design** (RCBD). RCBDs are often employed by agronomists responsible for product placement projects (Figure 0.1).

In the early stages of field trials used for genetic improvement and cultivar development, the number of replicable genotypes (treatments) can consist of many thousands. Even if the experimental units are tiny plots, homogeneous growing conditions will not exist across all experimental units. Yet, the breeder has to make decisions about which replicable genotypes to select without the confounding influence of variability that exists across field plots that are used to evaluate thousands of replicable genotypes. **Incomplete block designs** were developed for plant breeding projects where the goal is to make comparisons among all genotypes (treatments) grown in different blocks (Yates 1936).

**Alpha lattices** are a type of partially balanced lattice that are used extensively by plant breeders in early stage field trials because available seed for each genotype is sufficient for only a few replicates, In the alpha lattice design each replicate will consist of a set of incomplete blocks that contain all treatments (replicable genotypes). The idea with alpha lattices is to distribute the replicable genotypes among the incomplete blocks so that all possible pairs of lines occur within incomplete blocks at equal frequencies. Thus, effects associated with each incomplete block are not included in the estimated residual variance while precision for comparing genotypes within incomplete blocks is maximized.

For example, consider a field trial consisting of large experimental units used to evaluate 24 oat varieties. (John and Williams 1995) evaluated these oat varieties in three replicates consisting of six incomplete blocks with four experimental units per block.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Replication 1** | | | | | | **Replication 2** | | | | | | **Replication 3** | | | | | |
| **B1** | **B2** | **B3** | **B4** | **B5** | **B6** | **B7** | **B8** | **B9** | **B10** | **B11** | **B12** | **B13** | **B14** | **B15** | **B16** | **B17** | **B18** |
| 11 | 21 | 23 | 13 | 17 | 6 | 8 | 24 | 12 | 5 | 2 | 19 | 11 | 2 | 17 | 12 | 21 | 3 |
| 4 | 10 | 14 | 3 | 15 | 12 | 20 | 15 | 11 | 9 | 18 | 7 | 1 | 15 | 18 | 14 | 22 | 5 |
| 5 | 20 | 16 | 19 | 7 | 24 | 14 | 3 | 21 | 10 | 13 | 6 | 14 | 9 | 4 | 10 | 16 | 20 |
| 11 | 2 | 18 | 8 | 1 | 9 | 4 | 23 | 17 | 1 | 22 | 16 | 19 | 8 | 6 | 23 | 24 | 7 |

Notice that the incomplete blocks are nested within the replicates.

**Augmented designs** refer to designs in which a set of checks are added to all of the incomplete blocks (Federer 1961; Federer 1975). Each incomplete block consists of a subset of experimental genotypes plus the set of checks. Within a replicate experimental genotypes are randomly assigned to only one incomplete block, while a complete set of checks are included in all incomplete blocks. The replicated checks can then be used to estimate block effects which can be used to adjust the values for the experimental lines that occur in the same block. There is greater cost associated with augmented designs relative to alpha lattice designs because the addition of replicated checks requires more experimental units. It is possible to estimate block effects by assuming that the sample of experimental lines is random and thus any variability among blocks is due to non-genetic sources, e.g., (Lado et al. 2013). While, it seems obvious that inclusion of checks in incomplete blocks is wasteful, there are other reasons that plant breeders include checks in incomplete blocks. Can you name a few other reasons?

**Models**

Models are representations or abstractions of reality. Some models can be very useful, e.g., prediction of phenotypes, even if they are not accurate. Most often predictive models are in the form of mathematical functions. Also, there are models for organizing data, analyses, processes and systems. Yes, breeding systems and genetic processes can be represented as sets of mathematical equations. Historically the subject of designing an optimal breeding system has been approached through *ad hoc* management activities that are evaluated through trial and error. In the 21st Century design and development of plant breeding systems will be treated with the same rigor that engineers use to design optimal manufacturing or transportation systems.

**Data Models**

Even if it were possible to record data without error, as soon as we evaluate a trait and record the value on a living organism, we lose information. The challenge is to develop a data model that will minimize recording errors and loss of information.

**What is Data Modeling**?

* Data modeling is the process of defining data requirements needed to support decisions.
* Data modeling is used to assure standard, consistent and predictable management of data as a resource for making decisions.
* Data models support data and decision systems by providing definitions and formats.

If the data are modeled consistently throughout a plant breeding program then compatibility of data can be achieved. If a single data structure is used to store and access data then multiple data analyses can share data.

**Example of steps for modeling data in a plant breeding project**

* Determine the trait metrics that will be used to make decisions.
* Outline the plant breeding process.
* Determine the experimental or sampling units that will be evaluated at each step in the process
* Determine the number of experimental or sampling units that will be evaluated
* Characterize the experimental and sampling units that will be evaluated.

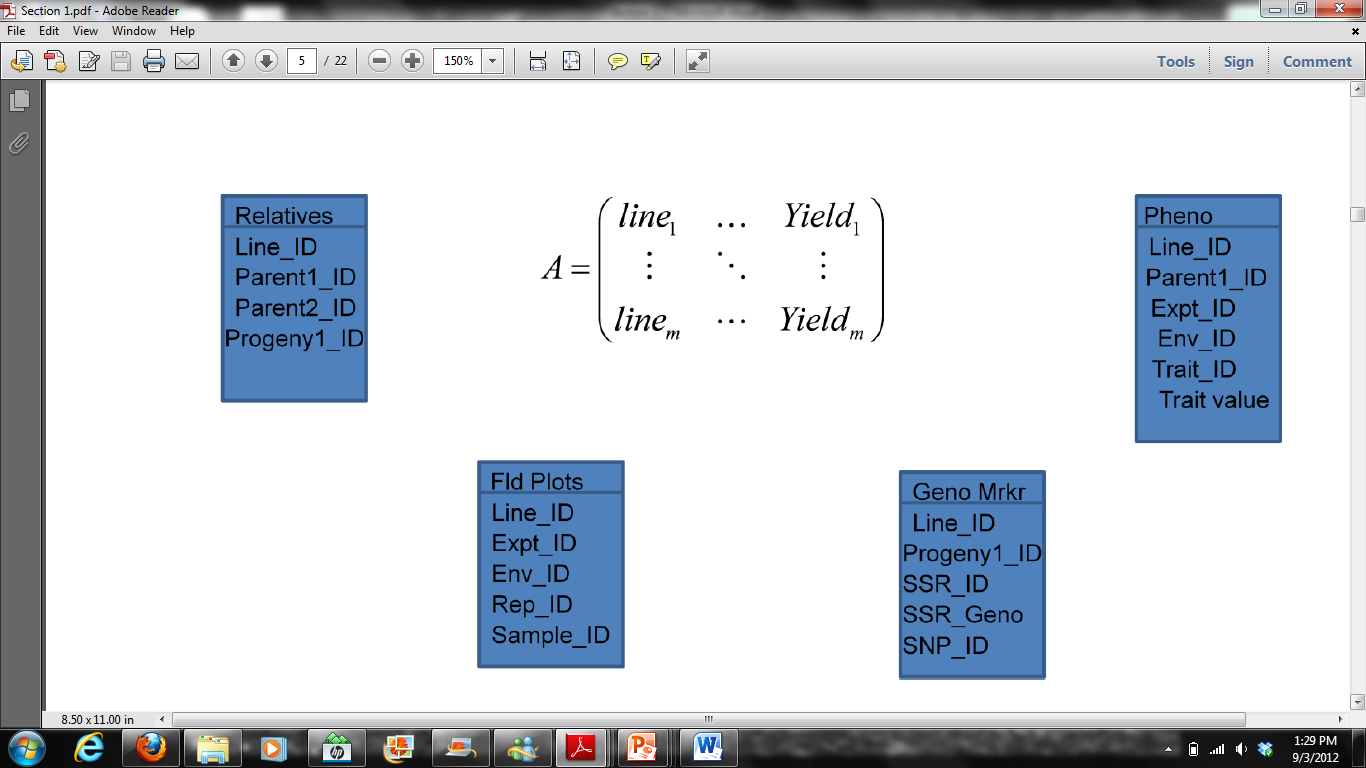
In quantitative genetics we evaluate responses (traits) of experimental or sampling units on continuous scales, e.g., grain yield, plant height, harvest index, etc. Note that a measurement taken on a continuous, i.e., quantitative scale, is not the same as a continuously measured trait. **Continuously measured traits**, such as grain fill, transpiration, disease progression or gene expression are measured continuously over the growth and development of an organism. Historically, evaluation of continuously changing traits have been too labor intensive to justify their expense. The emergence of ‘phenomics’ using image processing will overcome the limitations of acquiring the data. However, the need to store and manage continuously measured traits using phenomic technologies is going to require novel data models and storage capabilities.

*Data models address the need to organize data for subsequent analyses.*

A simple data model consists of a Row x Column matrix, where all experimental or sampling units are represented in rows and the evaluated characteristics or attributes for each unit are recorded in the columns:



While the ***A***(r x c) matrix is sufficient for small research projects, it is inadequate and cumbersome for breeding programs consisting of multiple types of evaluation trials at multiple stages of development. For such programs relational data bases are designed to optimize the ability to search and prepare data for analyses using statistical and genetic models (Figure 1.2). Further, unless data in an *A*(r x c) matrix is disseminated through “read only” access, there is potential for alteration of originally recorded data. Thus, the use of excel files, too commonly used to store experimental data in an *A*(r x c) matrix, can create serious ethical issues. While such issues do not disappear with relational databases, relational databases enable more effective protection of data as originally recorded. Recently, a publicly available database designed for organizing data from plant breeding projects has been developed. Known as the Breeding Management System, it is part of the Integrated Breeding Platform designed and developed by the Generation Challenge Program of the Consultative Group of International Agricultural Research centers.



**Figure 0.2**

While the development of relational databases is outside of the scope for this course, it is important to note that plant breeders routinely work with database developers to design, implement and curate relational databases.

Test your understanding with ALA 0.1 and ALA 0.2

**Phenotype Models**

For the most part, plant breeders rely on linear models to represent measured traits.

A general (not general*ized*) linear model for the phenotype, can be denoted:

where ***Y****i* represents the phenotype of individual *i* and ***e****i* represents residual variability (or lack of precision) in the measurement of the phenotype of individual *i*. We often assume that the variability associated with each measurement, ***e****i*, are distributed as random identical and independent Normal variables (denoted ~iid N()). This simple model is typically associated with the hypothesis that the only source of variability is that due to chance (noise). We can extend the simple model to include genetic and environmental sources of variability:

Test your understanding with ALA 0.3

**Exploratory Data Analyses**

Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, scatter plots and simple descriptive statistics such as the range (maximum, minimum), quartiles, correlations, and coefficients of variation. These are known as exploratory data analysis (EDA) techniques and can be used to identify data errors and provide preliminary inferences about the structure of the data prior to conducting analyses for decision making. However, prior to conducting EDA, the phenotype should be ***modeled*** using the parameters defined by the experimental and sampling designs.

**Estimation.**

*Statistical Parameters* are quantities that are used to describe central tendencies and dispersion characteristics of populations. Parameters are determined by models used to represent the traits of interest. Parameters of interest in population and quantitative genetics include frequencies, means, variances and covariances.

Because populations often consist of very large (potentially infinite) numbers of members it is usually impossible to determine values for the parameters. Instead ***estimates*** of the parameters are determined from samples. The rule, i.e. ***algorithm***, by which an estimate of a parameter is calculated is known as an ***estimator*.**  For example, the algorithm for calculating a sample average given by

,

which provides an ***estimator*** of the mean. And the calculated value, e.g., 132.38, obtained from 25 (=*n*) sampled measurements, X*i* from a population would be an ***estimate*** of the population mean.

Calculating arithmetic means, either simple or weighted within-group estimates, represents a common approach to summarizing and comparing groups. Data from most agronomic experiments include multiple treatments (or samples) and sources of variability. Further, the numbers of observations per treatment often are not equal; even if an experiment is designed to provide balance, experimental units are often lost during the execution of an experiment. Indeed, most data sets come from experiments that have multiple effects of interest and are not balanced. In such situations, the arithmetic mean for a group may not accurately reflect the "typical" response for that group because the arithmetic mean may be biased by unequal weighting among multiple sources of variability. The calculation of least square means, ***emmeans*** now **estimated marginal means (*emmeans)*** was developed for such situations. In effect, *emmeans* are within-group means appropriately adjusted for the other sources of variability. The adjustments made by *emmeans* are meant to provide estimates as though the data were obtained from a balanced design. When an experiment is **balanced, arithmetic averages and *emmeans* are the same.**

|  |  |  |
| --- | --- | --- |
| Table 0.1 | | |
| Cultivar | Location | Yj,k |
| A | Ames | 17, 28, 19, 21, 19 |
| A | Sutherland | 43, 30, 39, 44, 44 |
| A | Castana | -, -, 16, -, - |
| B | Ames | 21,21 ,-, 24,25 |
| B | Sutherland | 39,45,42,47, - |
| B | Castana | -, 19,22, -, 16 |
| C | Ames | 22,30, -,33,31 |
| C | Sutherland | 46, -, -, -, - |
| C | Castana | 25,31,25,33,29 |

|  |  |  |
| --- | --- | --- |
| Table 0.2 | | |
| Cultivar | N | Average |
| A | 11 | 29.1 |
| B | 11 | 29.2 |
| C | 10 | 30.5 |

Consider a data set consisting of 3 cultivars evaluated at each of 3 locations (Table 0.1). Despite exercising best agronomic practices, note that some plots at some locations did not produce phenotypic values.

The estimated means and number of observations for each cultivar indicate that there is very little difference among the cultivars, although cultivar C appears to have the highest yield (Table 0.2).

|  |  |
| --- | --- |
| Table 0.3 | |
| Cultivar | *emmean* |
| A | 25.6 |
| B | 28.3 |
| C | 34.4 |

A closer investigation of the data reveals that the means are unequally weighted by location effects. Recalculating the *emmeans* for the cultivars indicates more distinctive differences among the cultivars, once the differences among environments were taken into account (Table 0.3).

**Estimation of Variance** and standard deviation

If we model a trait value *Y* as:

Then the variance of the population consisting of individuals, *i* = 1,2,3 …. N is:

The square root of the variance is known as the standard deviation. Since it is not possible to evaluate a population of a crop species (think about it), we usually take a sample of individuals representing the population, i = 1,2,3 … *n*, where *n* << N. The estimator of the sample variance from a sample of *n* valuesis:



The sample standard deviation is the square root of the sample variance.

**Estimation of Covariance**

The covariance is a measure of the joint variation between two variables. Let us refer to one trait as *X* and a second trait as *Y*. We can model *Y* as before and we can model *X* in a similar manner i.e.,

**,**

The covariance of *X* and *Y* is



Again, it is not possible to evaluate a population, so we usually take a sample of individuals representing the population, i = 1,2,3 … *n*, where *n* << N. So the estimator of a sample covariance is:



**Estimation of Correlation**

Linear correlation is a *descriptive statistic* that quantifies the strength and direction of a linear relationship between two continuous variables. The Pearson Correlation Coefficient, usually designated , determines how close to linear the change in one variable X will be associated with a change in a second continuous variable Y. As a population parameter  is determined:

,

where  is the covariance between variables X and Y and  and are the standard deviations for the variables X and Y, respectively. The covariance between variables X and Y, is

, the joint mean for X and Y minus the product of the mean of X and the mean of Y.

 may take any value between plus and minus one. The sign of  (+ , -) defines the direction of the relationship. A positive relationship means that a positive change in one variable is associated with a corresponding positive change in the other, while a negative relationship is associated with a negative change in the other variable. The numerical value of *r* describes the strength of the relationship. Correlation coefficients of  +1.0 or -1.0 indicate perfect linear relationships. If  = 0.0 then there is an absence of a linear relationship. A correlation coefficient of 0.50 indicates a stronger degree of linear relationship than one of ** =0.40.

TRY THIS: Describe pairs of continuous variables measured on plants that are likely to be linearly associated.

Correlation is an often misused descriptive statistic. A correlation of zero does not mean that there is no association between the two variables. There can be non-linear associations that will not be detected with . Thus it always a good idea to plot the data using a **scatter plot**. Also correlations can be spurious. For example, a positive relationship between the number of sheep in the United States and the number of golf courses does not mean that sheep numbers have increased because there are more golf courses. Both variables are likely to be related to an underlying trend of increasing population in the U.S. Many things can be correlated, but it is the physical or biological relationship that gives a correlation relevance. Correlation only states the degree of linear association (not cause and effect) between the two variables.

A straightforward way to visualize relationships between pairs of continuous variable is through the use of scatter plots. Usually, the dependent variable is plotted on the vertical axis of the plot while the other variable is plotted on the horizontal axis. Such a plot can provide visual evidence of a linear relationship between the variables.

Sampled data can be used to estimate , usually denoted , involves estimating the co-variance of two variables, and estimating the standard deviations of the two continuous variables X and Y:



The numerator is the sum of cross products of xy and measures the combined distances of all points from the average of the two variables (x̄,ȳ). The more closely X and Y are related, the greater this value will be. The denominator is the product of the square roots of the sums of squared deviation of X and Y. The product of these two roots quantifies how much X and Y vary independently of each other.

Test your understanding and ability to conduct EDA using R with ALA 0.4.

If you have not downloaded and installed R see the following references:

Introduction to R.docx

<http://www.r-project.org/>

<https://www.lynda.com/R-tutorials/Up-Running-R/120612-2.html>

**Analysis of Variance**

The ANOVA has been the primary tool for testing hypotheses about parameters in models. The ANOVA was originally developed and introduced for analyses of quantitative genetic questions by R.A. Fisher. Since its introduction, the assumptions underlying the ANOVA have guided development of sophisticated experimental designs, and with increasing computational capabilities the ANOVA has evolved to provide estimates of variance components. In an introductory Quantitative Methods course the ANOVA is usually obtained using **least squares estimators** that are applied to balanced data sets. Remember an estimator is an algorithm, i.e., a set of instructions used to compute estimates of the parameters of a model.

**Linear models**

Let us imagine that we have two plant accessions that have been collected and reside in a germplasm repository. We wish to evaluate whether these two accessions are unique with respect to yield. Assume that we have 10 plots available for purposes of testing the null hypothesis that there is no difference in their yield. Also, assume that we have enough seed to plant 200 seeds in each plot. Let’s next assume that the 10 plots consist of two-row plots that are arranged in a 5x2 grid consisting of five ranges with 2 plots per row. We can randomly assign seed from each accession to the 10 plots. This would represent a Completely Random Design (CRD). Can you explain why? Prior to execution of the experiment, we want to model the phenotypic data using a linear function. In this case we would model the phenotypic data using:

** (1)**

where *Yij* is the yield of plot *i,j* *******i* represents the mean of accession *i* evaluated across all *j* replicates and *****i,j* ~ i.i.d. N(0,It is important to get in the habit of recognizing whether the parameters of the model are considered random or fixed effects. In this first model, since we selected the two accessions, rather than sampled them from some population, we should consider them to be fixed effects. The parameter *****i,j* represents the residual variability that is based on a sample of plots (experimental units) to which the treatments will be assigned, so *i,j*is considered a random effect (Table 0.4).

|  |  |
| --- | --- |
| Table 0.4 | |
| PI 1 (bu/ac) | PI 2 (bu/ac) |
| 27 | 30 |
| 31 | 29 |
| 35 | 32 |
| 34 | 32 |
| 28 | 31 |

Is this a tidy or messy data model? Which organization is needed to create boxplots, histograms and an ANOVA with R? EXCEL?

Next, let’s say that we evaluate the plots for yield (bushels per acre) as well as stand counts (plants per plot) at the time of harvest. The resulting data might look something like (Table 0.5)

Is this a tidy or messy data model? Which is needed to create boxplots, histograms and an ANOVA using R? EXCEL?

|  |  |  |  |
| --- | --- | --- | --- |
| Table 0.5 | | | |
| PI accession 1 | | PI accession 2 | |
| (bu/ac) | (plants/plot) | (bu/ac) | (plants/plot) |
| 27 | 91 | 30 | 102 |
| 31 | 122 | 29 | 89 |
| 35 | 143 | 32 | 139 |
| 34 | 145 | 32 | 147 |
| 28 | 110 | 31 | 112 |

Suppose that there is a known gradient for some soil factor (moisture, organic matter, fertility, etc.) across the ranges. In order to remove the effect of the gradient on our comparisons between the two accessions we should ‘block’ each range as a factor in our model. Let us further assume that we block the accession ‘treatments’ into five blocks consisting of two plots each. If we randomly group pairs of the accessions into 5 sets, next randomly assign each set to a range and third randomly assign each accession within a set to the plots within ranges, we will have a randomized complete block design (RCBD) that can be modeled as

*Yij =* + b*j +* P*i +* *ij* **(2)**

where the definition of parameters is the same as the CRD model, but with the added term for a blocking factor. Table 2.3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 0.6 | | | | |
|  | PI accession 1 | | PI accession 2 | |
| Block | (bu/ac) | (plants/plot) | (bu/ac) | (plants/plot) |
| 1 | 27 | 91 | 30 | 102 |
| 2 | 31 | 122 | 29 | 89 |
| 3 | 35 | 143 | 32 | 139 |
| 4 | 34 | 145 | 32 | 147 |
| 5 | 28 | 110 | 31 | 112 |

**Variance Components**

If we extend our simple model to include genetic and environmental sources of variability:

(1.9)

then, noting that ****** is a constant and applying some algebra we can show that the Variance of *Y* is

If we further assume that genotype and environment are independent and that there is no genotype x environment interaction:

The Variance of Y is equal to the sum of the **variance components** V(G), V(E) and V(e).

A question to consider is whether the parameters of the linear model represent ***fixed***or ***random*** effects, because this determination will affect how we estimate variance components and inferences about relative contributions to the overall phenotypic variability. This determination depends on the ***inference space***to which results are going to be applied. Fixed effects denote components of the linear model with levels that are deliberately arranged by the experimenter. Inferences in fixed effect models are restricted to the set of conditions that the experimenter has chosen, whereas random effect models provide inferences for a population from which a sample is drawn.

As a practical matter, it is hard to justify designating a parameter as a random effect if the parameter space is not sampled well. Consider environments, for example. Since we cannot control the weather, it is tempting to designate environments as random effects, however drawing inferences to a targeted population of environments (TPE) will be difficult if we sample a small number of environments, say less than 40.

Because the inference space of interest for genetic improvement depends on random samples of genotypes obtained from a conceptually large breeding population, we do not consider genotypes as fixed effects until the genotypes have been selected. At the same time it is a rare experimental design that does not include a fixed effect. Often random effects, such as environments are classified as fixed effects in *mixed models* (more on this topic later).

**Expected Mean Squares**

The output in ANOVA tables produced by **least squares estimators** cannot be interpreted without understanding the expected sources of variability represented by the ANOVA Mean Squares. This is also known as the **expected mean squares** (EMS). In the case of balanced field plot designs with only a few sources of variation the expected mean squares are easily determined. If a particular design involves many sources of random and fixed factors, students have found the approach of (Lorenzen and Anderson 1993) to be useful.

1. Write the terms of the model with associated subscripts down the left side of the page. Across the top write the single letter subscripts (i,j,k, etc.). Above each subscript place either F or R if the factor associated with that transcript is fixed or random. Above that place the number of levels associated with that subscript (I,J,K, etc.).
2. Enter a 1 in every slot where the subscript at the top is contained within brackets in the term at the left.
3. Enter a 0 in every slot where the subscript at the top is fixed and also contained in the term as the left. Enter a 1 in every slot where the subscript at the top is random and also contained in the terms at the left.
4. Fill in the remaining slots with the number of levels at the top of each column.
5. To compute the Expected Mean Squares (EMS) for a given term having df > 0, start at the bottom and work up. Only consider terms whose indices include all the indices in the term whose EMS you are deriving. Compute the coefficient of this term by covering the columns corresponding to the indices in the term whose EMS you are deriving and multiplying the values in the remaining columns. If there is a 0 column that is not covered, this term need not be written in the EMS. A factor is considered fixed and denoted with a ɸ only if all of its indices are fixed. Otherwise it is considered random and denoted by the appropriate term.

Notice that this algorithm can be used to compute EMS for all terms in the model, including those that have zero df. A term that has zero df has no expected mean squares. For this reason, we will not compute EMS for terms having zero df even though such terms are in the algorithm to make the EMS of the other terms come out right. Note that this simple algorithm for determining the EMS in an AOV assumes that the data are balanced, i.e., each of the sources of variability (model parameters) have data for all levels, i, j, and k.

To illustrate, let us consider a slightly more complex, but typical RCBD design used by plant breeders to evaluate many genotypes grown in many Blocks with several environments for purposes of identifying and discarding poor performing genotypes in a cultivar development project. The phenotype *Y* for this typical field trial will be something like:

** (3)**

Factors:

Factor E – Fixed

Factor G – Random

Blocks – Random

Step 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | E | G | B | EMS |
| F | R | R |
| i | j | k |
| Ei |  |  |  |  |
| B(E)k/i |  |  |  |  |
| Gj |  |  |  |  |
| GEij |  |  |  |  |
| k/ij |  |  |  |  |

Step 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | E | G | B | EMS |
| F | R | R |
| i | j | k |
| Ei |  |  |  |  |
| B(E)k/i |  |  | 1 |  |
| Gj |  | 1 |  |  |
| GEij | 1 | 1 |  |  |
| k/ij | 1 | 1 | 1 |  |

Step 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | E | G | B | EMS |
| F | R | R |
| i | k | j |
| Ei | 0 |  |  |  |
| B(E)k/i | 0 |  | 1 |  |
| Gj |  | 1 |  |  |
| GEij | 1 | 1 |  |  |
| k/ij | 1 | 1 | 1 |  |

Step 4:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | E | G | B | EMS |
| F | F | R |
| i | J | k |
| Ei | 0 | G | R |  |
| B(E)k/i | 0 | G | 1 |  |
| Gj | E | 1 | R |  |
| GEij | 1 | 1 | R |  |
| (ij)k | 1 | 1 | 1 |  |

Step 5: Table 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | E | G | B | EMS | MS |
| F | F | R |
| i | k | j |
| Ei | 0 | R | G | GB(E)+ |  |
| B(E)k/i | 0 | G | 1 | GB(E) |  |
| Gj | E | 1 | R | RGE + REG |  |
| GEij | 1 | 1 | R | RGE |  |
| (ij)k | 1 | 1 | 1 |  |  |

If we conduct an ANOVA of yield using model (1) for a CRD, will result in an ANOVA table that looks something like,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **df** | **MS** | **F** | **Prob** |
| PI | 1 |  |  |  |
| Residual | 8 |  |  |  |

will be created.

What are the Expected Mean Squares for this simple ANOVA table?

If we conduct an ANOVA for yield or Germ for a CRD model, we will generate a table that looks something like:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **df** | **MS** | **F** | **Prob** |
| PI | 1 |  |  |  |
| Residual | 7 |  |  |  |

What are the Expected Mean Squares?

In model **(2)** the PI accessions are selected so we should consider them to be fixed effect parameters. Although the block parameter represents a sample of 5 of many possible blocks in the field trial there are only a few blocks that represent a ‘nuisance’ source of variability, so we can treat them as a fixed effect, while the parameter *i,j* represents the residual or error in the model which is based on a sample of plots to which experimental units are assigned. Thus *i,j*is considered a random effect where *i,j* ~ i.i.d. N(0, and the model is considered a mixed linear model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **df** | **MS** | **F** | **Prob** |
| Block | 4 |  |  |  |
| Accession | 1 |  |  |  |
| Residual | 4 |  |  |  |

What are the Expected Mean Squares for this ANOVA table?

Evaluate your understanding and ability to conduct ANOVA and obtain estimates of variance components from the expected mean squares using R with ALA 0.5.

**Regression and prediction.**

While correlation attempts to establish a linear relationship between two variables, regression techniques try to determine a predictive relationship. Regression is the foundation of methods used for Genomic Prediction. Linear regression attempts to model the relationship between a dependent quantitative variable Y (e.g., yield per unit of land) and one or more independent quantitative variables (e.g., breeding values of lines) denoted Xas a General Linear Model (GLM). In a GLM the response or dependent variable is modeled using a linear function of independent or explanatory variables. There are five basic assumptions made about the relationship between a response variable Y and an explanatory variable X.

1. All Y values are from independent experimental or sample units.
2. Each value of X has a known fixed value, i.e., it is measured without error.
3. For each value of X, the possible Y values are distributed as normal random variables.
4. The normal distribution for Y values corresponding to a particular value of X has a mean that lies on a line,

**(4)**

where: : is the intercept and represents the mean of the Y values when X=0 and is the slope of the line. represents the change in the values of Y per unit increase in X.

1. The distribution of *Y* values corresponding to a particular value of X and has standard deviation . The standard deviation is usually assumed to be the same for all values of X so that we may write. Violation of the last assumption is typical in plant breeding data and development of methods to account for unequal variances is an important area of research.

Suppose we have *n* observations of a response variable *Y* and an explanatory variable X: (X1,Y1*), . . . , (*Xn,Yn*)*, the model can be rewritten as:

**(5)**

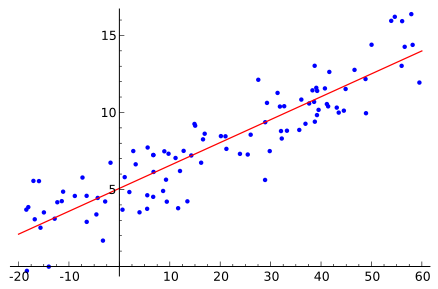
for *i* = 1, . . . , *n* experimental units. are assumed to be independent normal random variables with mean 0 and standard deviation .

The estimators for parameter and and are

Least squares predictors of the Y*i*  values are:

The residual vector of parameter ***ei*** (*e1*, . . . , *en*) can be estimated, with :

Notice that provides a predicted value (Figure 7.1) and the predicted value is “shrunken” relative to the actual observed values (deviations from the line).



**Figure 0.3**

Imagine that the *xi* values are an index such as the sum of all allelic values (+1 or -1) at quantitative trait loci throughout the genomes of homozygous diploid lines. Some lines could have 60 positive allelic values and no negative allelic values, while other cultivars could have a genotypic index of -20 (e.g., Figure 0.3). If the positive allelic values are associated with high phenotypic values, such as in the figure, then we will have a predictive relationship that can enable the plant breeder to predict phenotypes without having to grow all lines. The better the relationship between the genotypic index and the phenotype (less variability around the line), the better the ability to predict. This concept provides a foundation for what is widely referred to as Genomic Prediction.

TRY THIS: Copy data from Table 0.5 into EXCEL conduct an ANOVA for the relationship between yield and plants per plot. Compare the EXCEL results with analyses using the same model in R.

**Analysis of variance with covariates**

AOV with covariates is typically applied when there is a need to adjust results for variables that cannot be controlled by the experimenter. For example imagine that we have two germplasm accessions, and we wish to evaluate whether these have different yield values. Also, imagine that germination rates for each is different but unknown, especially under field conditions. We could decide to over-plant each plot and reduce the number of plants per plot to a constant number equal to a stand count that is typical for current agronomic practices. However, such an approach will be labor intensive and no more informative than adjusting plot yields for stand counts.

Assume that we have 10 plots available for purposes of testing the null hypothesis that there is no difference in yield between accessions. Also, assume that we have enough seed to plant 200 seeds in each plot, although current agronomic practices are more closely aligned with stands of about 125 plants per plot. Let us next assume that the 10 plots are arranged in a 5x2 grid consisting of five ranges with 2 plots per row. We suspect a gradient for some soil factor (moisture, organic matter, fertility, etc.) across the ranges. In order to remove the effect of the gradient on our comparisons between the two lines we should probably ‘block’ each range as a factor in our model. If we randomly group the accessions as pairs in each of 5 sets, next randomly assign each set to a range and third randomly assign each accession within a set to the plots within ranges, we will have a RCBD. At the time of harvest we evaluate the plots for yield (bushels per acre) as well as stand counts (plants per plot).

The resulting data are arranged in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PI accession 1 | | PI accession 2 | |
| Block | (bu/ac) | (plants/plot) | (bu/ac) | (plants/plot) |
| 1 | 27 | 91 | 30 | 102 |
| 2 | 31 | 122 | 29 | 89 |
| 3 | 35 | 143 | 32 | 139 |
| 4 | 34 | 145 | 32 | 147 |
| 5 | 28 | 110 | 31 | 112 |

If we use model (2) for yield where *Yij* is the yield of plot *i,j* *i* represents the mean of accession I, **b*j*** represents the *j*th block in which each pair of accessions are grown and *ij* ~ i.i.d. N(0,the resulting analysis revealed that the variability between accessions is not much greater than the residual variability. We might interpret this to mean that there is no difference in yield. However, our real interest is in whether there is a difference between the accessions adjusted for stand counts. A more appropriate model for the question of interest is:

***Yij =* *i* X *+* P*i* *+*******bj* ij*** . **(6)**

The model has two intercepts, denoted **P***i* for each of the accessions, and two slopes denoted *i* , for each of the accessions. The model also has a fixed effect nuisance parameters denoted by b*j* and *i,j* ~ i.i.d. N(0,. The resulting analyses is known as Analysis of Covariance can be thought of as an approach that takes advantage of both regression and ANOVA of factors, i.e., an AOC model includes parameters representing both regression and factor variables. The result of the estimation procedure will enable us to evaluate whether the accessions are equal at various stand counts of interest. In other words it will be possible to adjust yield values to various stand counts of interest. As a matter of ethics in science, the variable stand count needs to be modeled prior to conducting the field trial.

If we conduct an ANOVA of yield using model (6), list the sources of variability in the resulting ANOVA table

**Decisions from Statistical Inference**

In addition to estimation and prediction, statistical inference consists of ***hypothesis tests*** that are used to interpret the data we obtain from sampling and designed experiments. For decision makers, the primary purpose of statistical inference is to quantify the probabilities of correct and incorrect decisions.

Hypotheses are questions about parameters in models. For example, “Is the average value for a trait different than zero?” is a question about whether the parameter µ, in the model yi=  + i, has nonzero value. Formally, the hypothesis is written as  
and is called the null hypothesis, while   
is called an alternative hypothesis.

A test statistic is used to quantify the probability of obtaining the actual data from the experiment if the null hypothesis is true. For this simple hypothesis, the value of the test statistic should be close to zero if the null hypothesis is true and far from zero if the alternative hypothesis is true. Notice that in all models there is a parameter, *i*, included to indicate that there is some variability in the data that cannot be ascribed to the other parameters in the model. It is entirely possible that the variability in the data is due entirely to *i* and that an estimate of  or any other parameter in the model, is no different than a random number.

How often will an estimate of a parameter, e.g.,  in this model, be different from zero when Ho is true? We can answer this question by rerunning an experiment in which we know the parameter = 0 a million times, generate a histogram of the resulting distribution and then see how often (relative to 1 million) an estimated mean is equal to or more extreme than our experimental estimate. This is the frequency associated with finding our estimated value or a more extreme value when Ho is true.

The good news is that we don’t have to conduct a million such experiments because someone else has already determined the distribution for the case when the parameter = 0, is true. The frequency value associated with a test statistic as extreme or more extreme than the one observed from the experiment is often referred to as a ‘p’ value. The smaller the pvalue, the more comfortable we should be in rejecting the null hypothesis in favor of an alternative hypothesis. Keep in mind that we can be wrong in making a decision to accept the alternative hypothesis. In fact we are admitting that such a decision will be incorrect at a frequency = p.

Consider another simple example where we hypothesize that two genotypes have the same mean for some trait of interest. The difference between two genotypes is tested by: , where and are the true genotypic effects on the trait of interest. Whether or not a decision based on observed data is correct depends on the true value of the difference between the means.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3.1 Possible outcomes from the hypothesis that equals zero | | | |
| Decision based on empirical data | True Situation | | |
|  |  |  |
|  | Correct decision | Type I error | Type III error |
|  | Type II error | Correct decision | Type II error |
|  | Type III error | Type I error | Correct decision |

Columns: indicate the three possible truth (unobserved). Rows: indicate the three possible decisions made on the basis of estimates from measured data.

**Types of decision errors**

A Type I error is committed if the null hypothesis is rejected when it is true ( and the null hypothesis is rejected). A Type II error is committed if the null hypothesis is not rejected when it should be (). A Type III error occurs if the first decision is made when the third decision should have been made. This error also occurs if the third decision was made when the first decision was correct. Type III errors are sometimes called reverse decisions.

**Significance thresholds**

Decision makers often set a threshold, denoted by α, for committing Type I errors. The choice of α can be fixed at any desirable value between zero and one. Unfortunately, many decision makers use a value of α without thinking about the consequences. If our data provide a p value that is smaller (or larger) than α, then why not report the p value instead?

A Type III error rate, γ, is the frequency of incorrect reverse decisions and is always less than even for the smallest magnitudes of the standardized true difference, where is the parameter value of the standard error of the mean difference. Representative values of γ are shown in Table 3.2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3.2 Type III error rates, γ when a significant t-test  is based on 40 df. | | | | |
| Standardized true difference | Significance Level (α) | | | |
| 0.05 | 0.10 | 0.20 | 0.40 |
| 0.3 | 0.0127 | 0.0271 | 0.0584 | 0.1283 |
| 0.9 | 0.0026 | 0.0068 | 0.0167 | 0.0438 |
| 1.5 | 0.0005 | 0.0014 | 0.0039 | 0.0119 |
| 2.1 | 0.0001 | 0.0002 | 0.0008 | 0.0026 |
| 2.7 | 0.0000 | 0.0000 | 0.0001 | 0.0005 |

Last, consider the error that is committed if a null hypothesis is not rejected when it should be. This is also known as a Type II error and the probability of this type of error is denoted by β. It is the frequency of failure to detect real differences and is also affected by both the choice of α and the magnitude of the true difference (Table 3.3).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3.3. Type II error rates, β, or the frequencies of failure to detect differences when the test of significance is based on 40 df. | | | | |
| Standardized true difference | Significance Level (α) | | | |
| 0.05 | 0.10 | 0.20 | 0.40 |
| 0.3 | 0.941 | 0.886 | 0.781 | 0.579 |
| 0.9 | 0.863 | 0.774 | 0.639 | 0.437 |
| 1.5 | 0.697 | 0.571 | 0.419 | 0.248 |
| 2.1 | 0.469 | 0.340 | 0.214 | 0.107 |
| 2.7 | 0.251 | 0.158 | 0.085 | 0.035 |

Notice that is not equal to 1. The power of the test is = 1- and is denoted thus  +  = 1 The power of a test is the probability of rejecting the null hypothesis when it should be rejected. It can be increased by decreasing either the value of α or decreasing the value of by increasing the number of replications per treatment or by improving the experimental design.

#### **Additional decision metrics**

#### The most common and important decisions are those in which newly created and replicable genotypes are either retained (selected) or discarded. Retention (discard) decisions usually are based on statistical estimates of values for quantitative traits. As soon as any set of criteria are used to make a binary decision (retain or discard), the criteria are recognized as members of a binary classifier model. For binary classifier models **decision accuracy** is the proportion of correctly retained and correctly discarded relative to the total number of possible decisions:

***Decision Accuracy*** = ****

For binary decisions, accuracy is composed of two components: the proportion of correctly retained and the proportion of correctly discarded. The former is defined as the decision **sensitivity** and the latter is defined as the decision **specificity**.

#### Sensitivity: determines the probability of correct retention. It is calculated as

***Sensitivity*** = ****

#### Specificity: is also called the true negative rate and it determines the proportion of correct discards. It is calculated as

***Specificity*** = ****

Accuracy, sensitivity and specificity as well as other metrics can be determined from a confusion table.

|  |  |  |  |
| --- | --- | --- | --- |
| N=400 | **Predict discard** | **Predict retain** |  |
| **discard** | True discard  300 | False retain 30 | 330 |
| **retain** | False discard  20 | True retain 50 | 70 |
|  | 320 | 80 |  |

* Accuracy = (300+50)/400
* Misclassification rate = (30+20)/400 = (1-accuracy)
* Sensitivity = 50/70 = True retention rate
* False discard rate = 20/70= (1-sensitivity)
* Specificity = 300/330 = true retention rate
* False retention rate: 30/330 = (1-specificity)
* Precision: 50/80
* Prevalence: true rate of retention = 70/400

The confusion table applies to the discrimination threshold for each classifier model. If the discrimination threshold of a classifier model is changed, the relationships between true retention and false retention rates will likewise change. ***Receiver Operating Characteristic*** (ROC) curves are used to summarize the many confusion tables that might be needed to represent the trade-offs between sensitivity and specificity for the many possible discrimination thresholds. Explicitly the ROC curve is a plot of true retention rates against the false retention rates.

**References:**

Bernardo R (1996) Best Linear Unbiased Prediction of Maize single-cross performance Crop Sci 36:50-56

Bernardo R (2020) Breeding for Quantitative Traits in Plants. 3rd ed. Stemma Press, Woodbury, MN

Byrum J et al. (2016) Advanced Analytics for Agricultural Product Development Interfaces 46:5-17 doi:10.1287/inte.2015.0823

Federer WT (1961) Augmented designs with one-way elimination of heterogeneity Biometrics 17:447-473

Federer WT (1975) On augmented designs Biometrics 31:29-35

Fehr WR (1991) Principles of Cultivar Development vol 1 Theory and Technique. McMillan,

Hayes HK, Immer FR (1942) Methods of plant breeding. . McGraw-Hill, New York, NY

Henderson CR (1963) Selection index and expected genetic advance. In: Hanson WD, Robinson HF (eds) Statistical Genetics and Plant Breeding, vol 982. National Academy of Sciences-National Research Council, Washington, D.C., pp 141-163

John JA, Williams ER (1995) Alpha lattice design of spring oats. In: Cyclic and computer generated designs. Chapman and Hall, London, p 146

Lado B et al. (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data G3 (Bethesda) 3:2105-2114 doi:10.1534/g3.113.007807

Lorenzen TJ, Anderson VL (1993) Design of Experiments: a no-name approach vol 139. Statistics, textbooks and monographs. M. Dekker, New York, NY

Meuwissen THE, B.J. Hayes, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps Genetics 157:1819-1829

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize Crop Sci 49:1165-1176

Popper K (1959) The Logic of Scientific Discovery. Routledge, 270 Madison Ave. New York, NY 10016

Smith DC (1966) Plant breeding - Development and success. In: Frey KJ (ed) Plant Breeding. Iowa State University Press, Ames, Iowa, pp 3-54

Xu S (2003) Estimating Polygenic Effects Using Markers of the Entire Genome Genetics 163:789-801

Yates F (1936) A new method for arranging variety trials involving a large number of varieties Journal of Agricultural Science 26:424-455