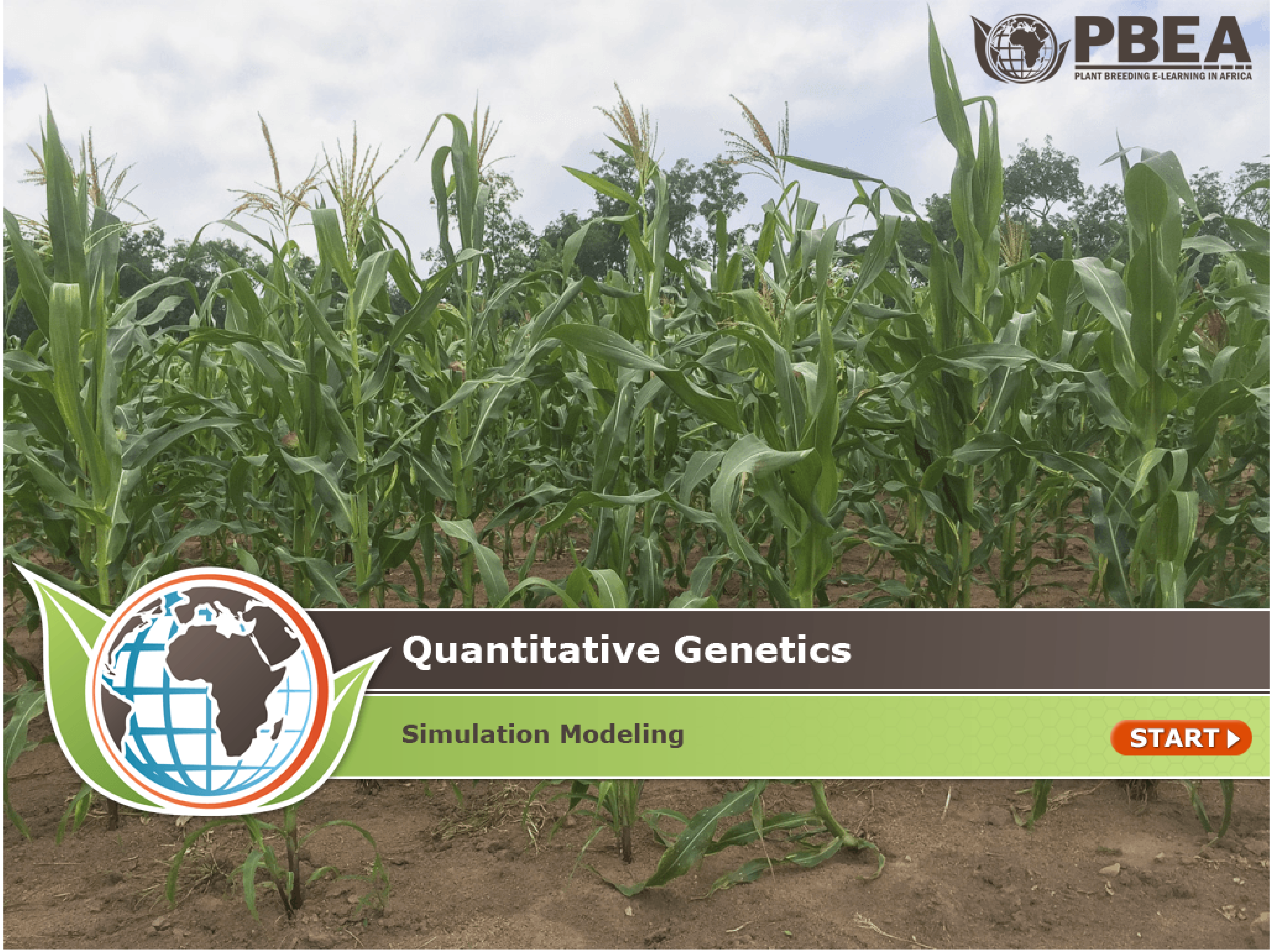


PBEA
PLANT BREEDING E-LEARNING IN AFRICA

Quantitative Genetics

Simulation Modeling

START ▸

Objectives

- Recognize limitations of experimental research
- Translate QG models to simulation models
- Translate simulation models to EXCEL software functions
- Build confidence in use of simulation models

Introduction

Quantitative genetic models are used to represent, describe and quantify the genetic contributions to natural phenomena. These models can be arbitrarily simple, e.g., additive linear models, or complex, e.g., non-additive, non-linear models. R.A. Fisher and Sewell Wright had a decades long debate about which type of model should be considered in the study of natural and artificial selection. Fisher and his disciples argued that the more complex models were not needed. Sewell Wright and his students argued that biology was inherently complex and needed non-linear non-additive models to accurately understand adaptation and evolution. Of course, both were correct and both were wrong. As George Box reminds us, all models are wrong, some are useful. The choice of an appropriate model depends upon the purpose of the research.

Prior to this module, we investigated development of theoretical quantitative genetic models for the purposes of conducting and interpreting analyses of plant breeding experiments. As noted, without the theoretical models, there would be no genetic understanding of the results. Theory provides predictions and predictions are the basis for generating testable hypotheses. In this module we introduce a far more practical justification for theoretical models: With a theoretical model it is possible to simulate many different data analysis techniques and breeding strategies prior to conducting expensive experiments. In other words, natural and artificial systems can be modeled in silico for purposes of predicting unknown outcomes. Many In silico experiments can be compared and the most promising can be used to compare methods or processes. If comparisons are based on objective criteria, such as accuracy, power, precision, efficacy and efficiency, and if the model used for data analyses is the same as the model used for simulating data sets we can make rational decisions about which methods to implement in plant breeding experiments.

History of Simulations

Geneticists first used computers to implement simulation models to evaluate limits to artificial selection (Hill and Robertson, 1968; Bulmer, 1968) in closed breeding populations. By 1988, Oscar Kempthorne, one of R.A. Fisher's disciples pointed out that the classical experimental and algebraic approaches were limited to unrealistic assumptions in breeding and evolutionary systems. Since 1988, plant and animal geneticists and breeders have used simulation models to evaluate the limits to emerging statistical methods (Beavis, 1994) and to choose among selection methods because experimental evaluation of breeding methods is time and resource limited (Podlich and Cooper, 1998). To date, there have been over 15,000 publications in which the terms simulation and breeding occur in the title. Currently there are numerous simulation software packages that have been developed and implemented for public and private research enterprises. Some are quite simple, while others are very flexible and complex. Generally, as the flexibility of the package increases, the learning curve associated with the complexity of the package also increases. Thus, some of these simulation packages require entire courses and years to master. While it is beyond the scope of this module to either advocate or teach any particular simulation package, we will learn how to implement the core quantitative and population genetic models that are part of every useful simulation package.

Core Elements

The core elements of simulation modeling include the genetic architecture of the trait(s), the structure of segregating generations derived from breeding populations and organization of the segregating genomes.

STRUCTURE OF THE BREEDING POPULATION AND SEGREGATING GENERATIONS

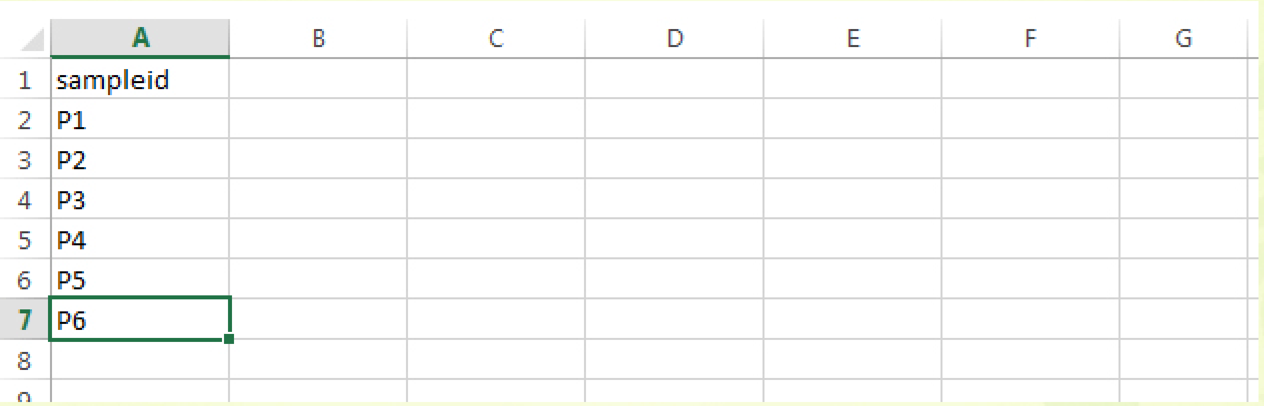
Before we decide on the genetic architecture of the trait, we need to know the structure of the segregating generation derived from the breeding population. In diploid organisms there are usually three genotypes at a SNP locus: {aa, ac, cc} or {tt, tg, gg}. Let's consider a locus with the second triplet, {tt, tg, gg}. If we decide to simulate a random mated population, then each of the three genotypes {tt, tg, gg} will occur at a frequency of p^2, 2pq and q^2. To decide which genotype is going to be assigned to an individual we should obtain a random sample of a number from the Uniform[0,1] distribution. If the random number is in the interval $[0,p^2]$, then we will assign the genotype 'gg' to individual i. If the random number is in the interval $[p^2, p^2+2pq]$, then we will assign the genotype 'tg' to individual i, otherwise we will assign the genotype 'tt' to individual i.

Obtaining a random sample from any distribution will depend on the syntax of the software system we decide to use for simulating the data. Since most students have experience with spreadsheet types of software, we will first learn how to use Excel for simulating SNP genotypes at a single locus in a random mating population, where p (frequency of g) = .3. The frequency of 'gg' genotypes at this locus will be 0.09, the frequency of 'tt' genotypes will be 0.49 and the frequency of heterozygous genotypes will be 0.42. Thus, if we sample a random number from the Uniform distribution in the interval [0, 0.09], then we will assign the genotype 'gg' to individual i; in the interval from (0.09, 0.51] we will assign the genotype 'gt' to individual i; otherwise we will assign the genotype 'tt' to individual i.

Excel-Based Simulation

With these parameters, we will use the Excel functions “IF” and “RAND” in the following steps:

1. Assign a sampleid designator to the first column.

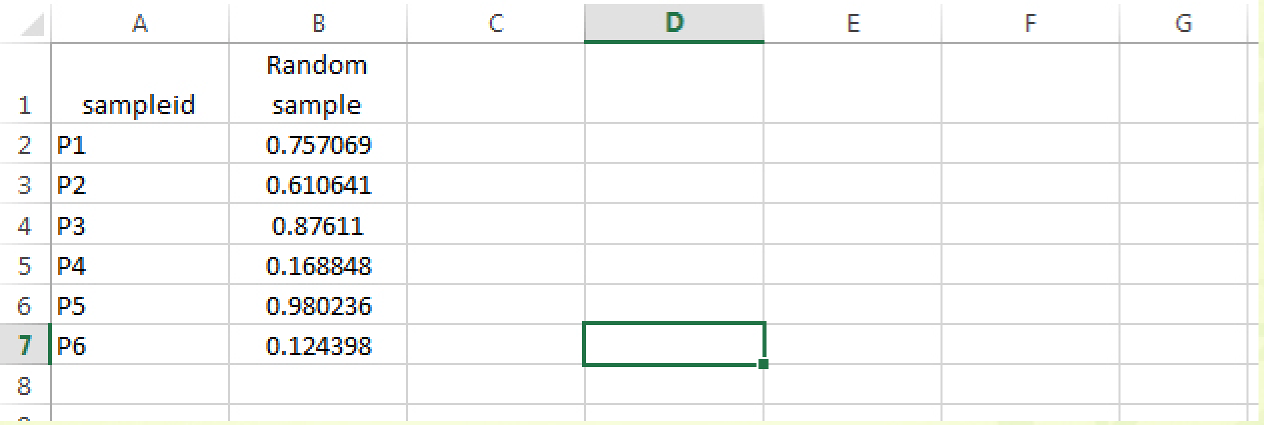


	A	B	C	D	E	F	G
1	sampleid						
2	P1						
3	P2						
4	P3						
5	P4						
6	P5						
7	P6						
8							

Excel-Based Simulation

With these parameters, we will use the Excel functions "IF" and "RAND" in the following steps:

2. Obtain a random number from the Uniform Distribution for each of the sampleid's. Syntax: type `=RAND()` **in cell B1, then drag across all cells in column B.**

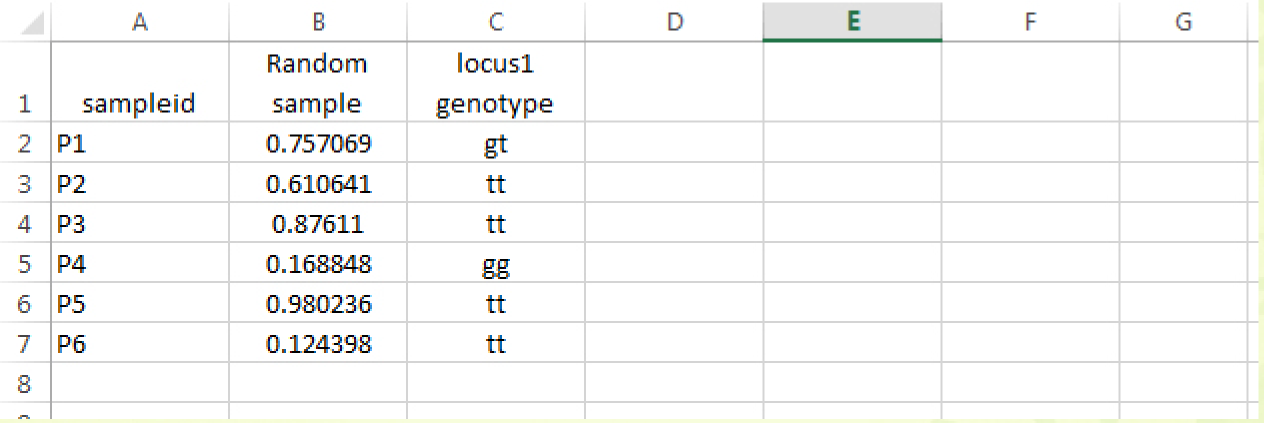


	A	B	C	D	E	F	G
1	sampleid	Random sample					
2	P1	0.757069					
3	P2	0.610641					
4	P3	0.87611					
5	P4	0.168848					
6	P5	0.980236					
7	P6	0.124398					
8							

Excel-Based Simulation

With these parameters, we will use the Excel functions “IF” and “RAND” in the following steps:

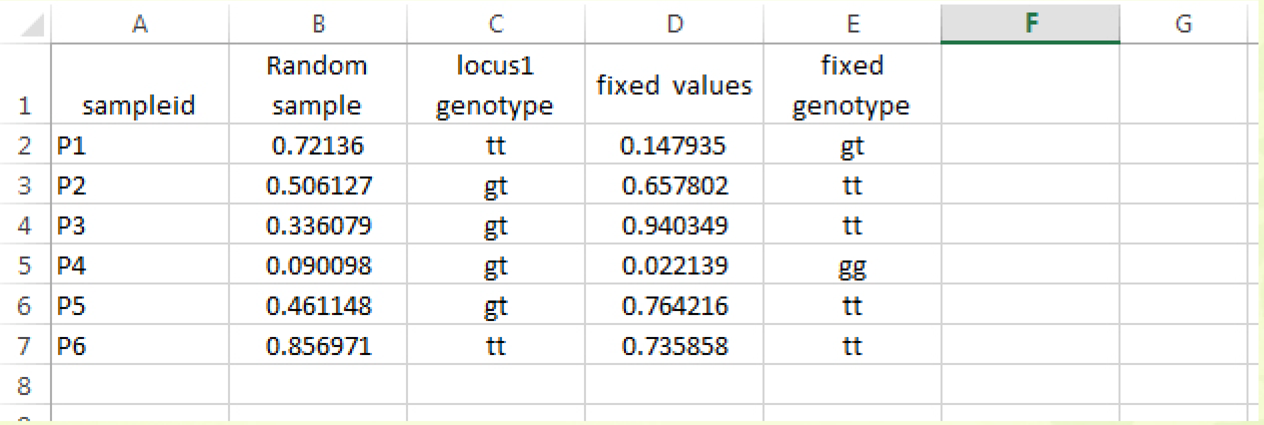
3. Based on the random number assigned to each sampleid, assign a genotype to the locus. Syntax: type `=IF(B3<=0.09,”gg”,IF(B3>0.51,”tt”,”gt”))` **in cell C1, then drag across all relevant cells.**



	A	B	C	D	E	F	G
1	sampleid	Random sample	locus1 genotype				
2	P1	0.757069	gt				
3	P2	0.610641	tt				
4	P3	0.87611	tt				
5	P4	0.168848	gg				
6	P5	0.980236	tt				
7	P6	0.124398	tt				
8							

Excel-Based Simulation

If we do not want the values generated by the RAND function to change as we add functions to the spreadsheet, we should plan to cut and paste a set of the actual values from one of the sampling events into new columns:



	A	B	C	D	E	F	G
1	sampleid	Random sample	locus1 genotype	fixed values	fixed genotype		
2	P1	0.72136	tt	0.147935	gt		
3	P2	0.506127	gt	0.657802	tt		
4	P3	0.336079	gt	0.940349	tt		
5	P4	0.090098	gt	0.022139	gg		
6	P5	0.461148	gt	0.764216	tt		
7	P6	0.856971	tt	0.735858	tt		
8							

Other Population Structures

There are many other possible breeding population structures; some are the result of designed crosses (see the module on mating designs), but most population structures emerge from long term breeding programs in which elite homozygous cultivars are crossed to promising homozygous lines through opportunistic networks of crosses. Simulating genotypes at segregating loci from any mating design or breeding program be obtained in a manner described in the previous paragraph. We need only decide on the frequencies of the genotypes in the segregating populations. For example, the specific case of an F2 generation derived from a cross of two inbred lines, p=q=0.5 and if a random number obtained from U[0,1] is greater than ¼ and less then ¾ then the genotype of individual plant i, will be ‘gt’. We could be interested in the case of recombinant inbred lines derived in the F5 generation of a cross of two inbred lines. In this case, the frequency of homozygotes are now p^2+pqF and q^2+pqF and the frequency of heterozygous lines is 2pq(1-F), where F = .875. Thus if a random number obtained from U[0,1] is less than .4687, then RIL i will be assigned the genotype ‘gg’. It should be obvious that it should be possible to generate mixtures of segregating families from multiple independent of related crosses and simulate genotypes for any particular locus to all individuals in all families, regardless of how the families are derived.

Genetic Architecture of the Trait

Next we need to decide how many loci will influence a trait and whether the alleles at the loci will interact. Let's begin with a single-locus additive quantitative trait, designated P. Further, consider a trait with an average phenotypic expression of 50 units and phenotypic variability in a diploid species that is due to additie genetic variability at a single locus and non-genetic variability. Initially let's plan to let half of the phenotypic variability be due to segregation at the locus and half due to non-genetic sources of variability. The first step is to translate this brief description into a quantitative genetic model, preferably the same model that will be used in eventual analysis of the phenotypic trait:

$$P_{ij} = \mu + G_i + \varepsilon_{ij},$$

where i refers to one of three possible genotypes conferred by two alleles, j refers to one of the repeated samples of the ith genotype, e_{ij} is a non-genetic source of variability and is ~ i.id N(0,σ_{ε}). Thus,

$$\sigma_P^2 = \sigma_G^2 + \sigma_\varepsilon^2$$

Parameter Assignment

The next step is to assign values to each of the parameters in the model. μ is assigned the value of 50 and values from e_{ij} are sampled from a Normal distribution with a mean of zero and a standard deviation of s_e. We decided that we want to simulate data in which

$$\frac{\sigma_G^2}{\sigma_P^2} = 0.5.$$

We can choose any value for σ_p^2, but it is often best to choose a value that is ~ consistent with estimates from field trials for the crop of interest. In this case, let's say our field trials have typically produced estimates of phenotypic variance of ~ 98. Thus, both σ_G^2 and σ_ε^2 are ~ 49. Thus, we can obtain values for ε_{ij} by sampling a Normal distribution with mean = 0 and standard deviation = 7.

Genotypic Values

We also need numeric values for each of the genotypes. Recall from Quantitative Genetic Models Theory, we can assign coded genotypic values to each genotype:

Coded genotypic value of one homozygote (gg) = +a
Coded genotypic value of the other homozygote (tt) = -a
Coded genotypic value of heterozygotes (tg or gt) = d

Since we are simulating an additive genetic model, the genotypic value of the heterozygotes (d) is midway between the two values for the homozygotes, i.e., d=0. Thus,

G_i = a for i = “gg”;
-a for i = “tt”;
and 0 for i = “gt” or “tg”.

Now we need a numeric value for a.

Calculations

Also, recall that

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

where

$$\sigma_A^2 = 2pq\left[a + d\left(q - p\right)\right]^2$$

and

$$\sigma_D^2 = \left(2pqd\right)^2.$$

Since we have decided to simulate d=0,

$$\sigma_G^2 = \sigma_A^2 = 2pqa^2,$$

thus

$$a = \sqrt{\frac{49}{2pq}}$$

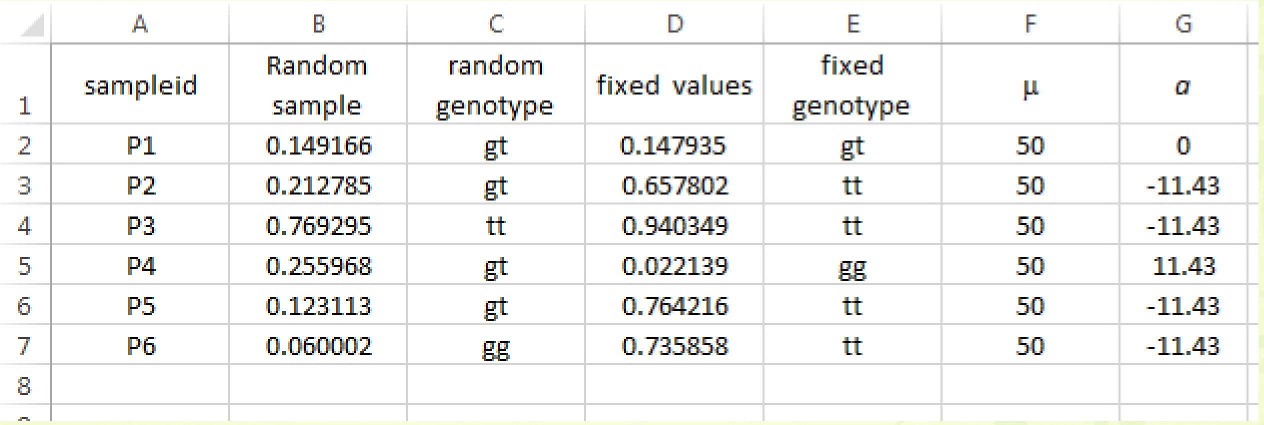
If we assume that the frequency of ‘t’ (or ‘g’) in the population is 1/4, then a reasonable value for

$$a = \sqrt{\frac{49}{2pq}} \sim 11.43$$

Excel Application

Next, let’s translate these values for the parameters into Excel functions.

Syntax for assigning ‘a’: Type `=IF(E3=“gg”,11.43, IF(E3=“tt”,-11.43,0))` in Cell G1, then drag across all relevant cells.



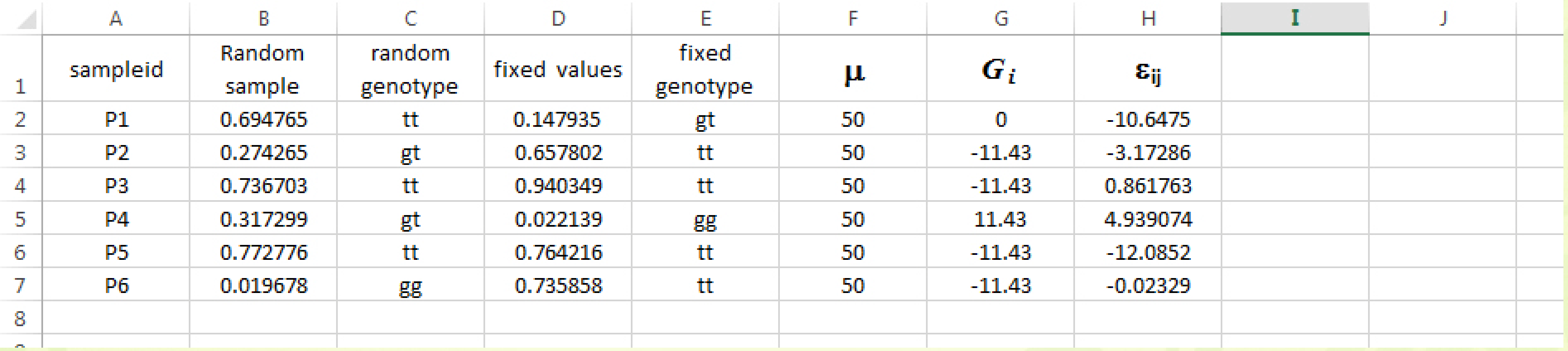
	A	B	C	D	E	F	G
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	*a*
2	P1	0.149166	gt	0.147935	gt	50	0
3	P2	0.212785	gt	0.657802	tt	50	-11.43
4	P3	0.769295	tt	0.940349	tt	50	-11.43
5	P4	0.255968	gt	0.022139	gg	50	11.43
6	P5	0.123113	gt	0.764216	tt	50	-11.43
7	P6	0.060002	gg	0.735858	tt	50	-11.43
8							

In order to fully understand how to sample from a Normal distribution requires knowledge of probability density functions, cumulative density functions and integral calculus that enables the translation between the two. These are topics beyond our current scope, but worth exploring by those who wish to develop their own simulation capabilities.

Normal Distribution Interval

For our immediate purpose the Syntax for obtaining values for e_{ij} by sampling a Normal distribution with mean = 0 and standard deviation = 7 is the following:

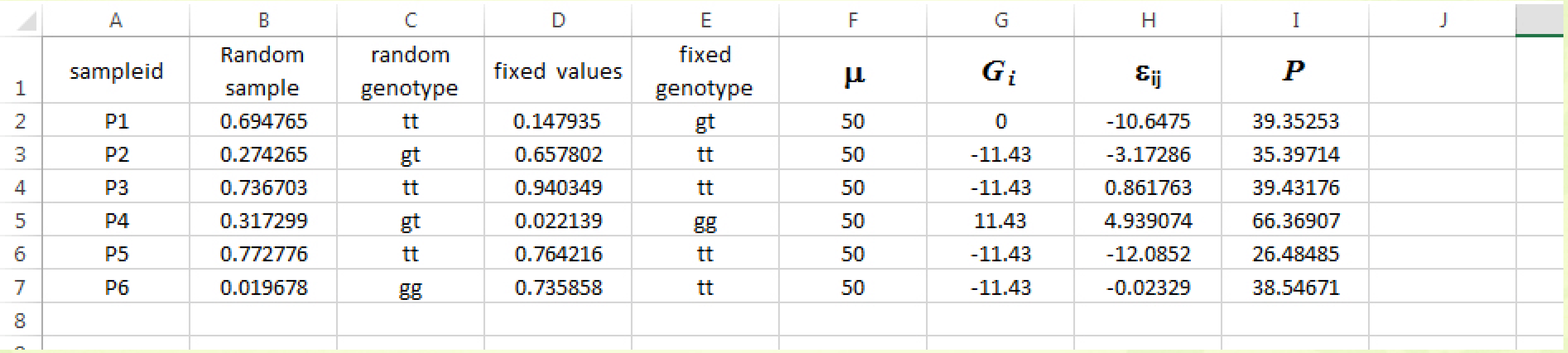
Type `=NORMINV(RAND(),0,7))` in cell H3 and then drag across all relevant cells.



	A	B	C	D	E	F	G	H	I	J
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	G_i	ε_{ij}		
2	P1	0.694765	tt	0.147935	gt	50	0	-10.6475		
3	P2	0.274265	gt	0.657802	tt	50	-11.43	-3.17286		
4	P3	0.736703	tt	0.940349	tt	50	-11.43	0.861763		
5	P4	0.317299	gt	0.022139	gg	50	11.43	4.939074		
6	P5	0.772776	tt	0.764216	tt	50	-11.43	-12.0852		
7	P6	0.019678	gg	0.735858	tt	50	-11.43	-0.02329		
8										

Simulated Phenotypes

We now have values for all of the parameters in the model and need merely sum columns F, G and H to obtain the simulated phenotypes (column I) for each of the sampleids.



	A	B	C	D	E	F	G	H	I	J
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	G_i	ε_{ij}	P	
2	P1	0.694765	tt	0.147935	gt	50	0	-10.6475	39.35253	
3	P2	0.274265	gt	0.657802	tt	50	-11.43	-3.17286	35.39714	
4	P3	0.736703	tt	0.940349	tt	50	-11.43	0.861763	39.43176	
5	P4	0.317299	gt	0.022139	gg	50	11.43	4.939074	66.36907	
6	P5	0.772776	tt	0.764216	tt	50	-11.43	-12.0852	26.48485	
7	P6	0.019678	gg	0.735858	tt	50	-11.43	-0.02329	38.54671	
8										

Keep in mind that if these were field trial data, we would only be able to obtain data found in columns A and I. It should be immediately apparent that column F could be a mean for a particular replication or environment of the sampleids. Thus, it should be possible to simulate data from multiple replicates and multiple environments with different mean values. It should also be apparent that the sampling of e_{ij} could be derived from environments with different plot to plot variability. For example instead of using 7 in the function =NORMINV(RAND(),0,7)), we could designate the standard deviation for some environments to be 14 and thus create a type of GxE that we discussed in Multi Environment Trial modules.

Example Calculations

For the specific case of an F_2 generation derived from a cross of two inbred lines, p=q=0.5,

$$a = \sqrt{\frac{49}{2\left(.25\right)}} = 9.9$$

Alternatively, we could be interested in the case of recombinant inbred lines derived in the F5 generation of a cross of two inbred lines. In this case,

$$\sigma_A^2 = 2pq\left(1+F\right)\left[a+d\left(q-p\right)\right]^2$$

Again, d=0, p=q=.5 , but F = .875 and a reasonable value to simulate for

$$a = \sqrt{\frac{49}{2\left(1+F\right)pq}} \sim 7.24$$

Polygenic Trait Simulation

Let's next simulate a polygenic trait P in which segregation at three loci will contribute additive genotypic values that are responsible for 30% of the phenotypic variability. In this case the phenotype is modeled:

$$P_{ij} = \mu + \sum_{k=1}^{n} G_{i(k)} + \varepsilon_{ij(k)}$$

where i, j, G and ε are as before and k represents each locus, and n is 3. For simplicity, lets assume that segregation at each of the three loci contributes an equal amount to the genotypic variability in an F2 population. Let's refer to these loci as quantitative trait loci (QTL). Using the quantitative genetic models we have already used we learn that

$$a = \sqrt{\frac{\sigma_G^2}{2pq}} / nQTL,$$

for each of the simulated QTL. If we want the total phenotypic variability to be ~98, as before, and the frequency of each of the alleles at all three loci is .5 (as in an F2) then $\sigma_G^2 = \sigma_A^2 = 29$ and a = 2.56 for each of the loci. We would translate this information to the Excel spreadsheet as before, but now the spreadsheet will have three columns for genotypes and three columns for Genotypic values at each of the loci.

QTL Simulations

How would you simulate genotypic effects if you wanted one of the QTL to contribute 75%, a second QTL to contribute 20% and the third to contribute 5% to the total genotypic variability?

For hybrid crops, the segregating progeny are often evaluated in testcross combination. For example, in maize, it is routine to generate doubled haploids (DHs) from a cross of two elite Stiff Stalk homozygous lines. The DH's are then crossed to an elite non-Stiff-Stalk homozygous 'tester'. The resulting sample of Testcrossed DH (TDH) will be evaluated in an initial field trial. Let's simulate this situation for TDH's, grown in a field trial in which the CV for yield is ~ 7% and the mean is ~ 225 bu/ac. In order to simulate TDH's we need to recall that

$$\sigma^2_{A^T} = \frac{1}{2}(1+F)\,\sigma^2(\alpha^T)$$

Because the parents of the DH lines are fully homozygous, we can assume F=1. Thus,

$$\sigma^2_{A^T} = a^2$$

Otherwise the simulations will be generated as before, except we now have a different mean and phenotypic variance.

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Funding provided by:

BILL & MELINDA
GATES foundation

Other collaborating organizations:

AGRA
Growing Africa's Agriculture

CGIAR

Partnering universities:

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

***Quantitative Genetics* Module 9 Co-authors:**
William D. Beavis *(ISU)*

Multimedia Developers:
Gretchen Anderson, Todd Hartnell, and Andy Rohrback (*ISU*)

***Quantitative Genetics* Course Team:**
William D. Beavis *(ISU)*; Richard Akromah, Joseph Sarkodie-Addo, Maxwell Asante *(KNUST)*; Richard Edema,
Paul Gibson, Settumba Mukasa, Patrick Ongom *(MAK)*; John Derera, Pangirayi Tongoona, Hussein Shimelis *(UKZN)*