Statistical Foundation of Quantitative Genetics

Jack Dekkers

Dept. Animal Science, Iowa State University

Quantitative genetics is the study of continuous traits (such as height or weight) and its underlying mechanisms. It is based on extending the principles of Mendelian and populations genetics to quantitative traits.

Falconer and Mackay:
"Quantitative genetics theory consists of the deduction of the consequences of Mendelian inheritance when extended to the properties of populations and to the simultaneous segregation of genes at many loci."

For the purposes of this class: Quantitative genetics = A set of concepts based on the theory of inheritance that help us understand and dissect the genetic basis of quantitative traits and predict what the consequences of different breeding choices will be and therefore allow us to make decisions that lead to the most desirable outcomes.

Quantitative traits
Quantitative genetics covers all traits that are determined by many genes.

- Continuous traits are quantitative traits with a continuous phenotypic range. They are usually polygenic, and may also have a significant environmental influence.
- Traits whose phenotypes are expressed in whole numbers, such as number of offspring, or number of bristles on a fruit fly. These traits can be either treated as approximately continuous traits or as threshold traits.
- Some qualitative traits can be treated as if they have an underlying quantitative basis, expressed as a threshold trait (or multiple thresholds). E.g. diseases that are controlled by multiple traits but for which phenotype is observed as healthy/diseased.

"…genetics is meant to explain two apparently antithetical observations – that organisms resemble their parents and differ from their parents. That is, genetics deals with both the problem of heredity and the problem of variation." Lewontin, 1974.

Francis Galton (1822-1911): regression toward mediocracy – progeny of parents with extreme phenotypes tend to be closer to average.

The modern synthesis of Quantitative Genetics was founded by R.A. Fisher, Sewall Wright, and J.B.S. Haldane, based on evolutionary concepts and population genetics, and aimed to predict the response to selection given data on the phenotype and relationships of individuals.

Some Basic Quantitative Genetic Concepts and Models

Quantitative genetics dwells primarily on developing theory or mathematical models that represent our understanding of phenomena of interest, and uses that theory to make predictions about how those phenomena will behave under specific circumstances. The model that exists to explain observations of quantitative traits contains the following components:

- Loci that carry alleles that affect phenotype – so-called quantitative trait loci or QTL
- Many such quantitative trait loci
- Alleles at QTL that act in pairs (2 alleles per locus) but that are passed on to progeny individually
- Which of the parent's alleles are passed on to progeny occurs at random (i.e. a random one of the pair of alleles that a parent has at a locus is passed on to a given progeny), which introduces variability among progeny
- Loci that affect phenotype sometimes show independent assortment (unlinked loci); sometimes not (linked loci)
- Environmental factors influence the trait

In order to develop the quantitative genetic theory and models and to deduce its consequences or predictions it might make, quantitative geneticists have translated these concepts and their behavior into mathematical and statistical terms/models. The most basic model of quantitative genetics is that the phenotypic value (P) of an individual is the combined effect of the individual's genotypic value (G) and the environmental deviation (E):

$$P = \mu + G + E \qquad \text{where } \mu \text{ is the trait mean}$$

G is the combined effect of all the genes that affect the trait.
E is the combined effect of all environmental effects that affect the phenotype of the individual.

The simplest model to describe inheritance of a quantitative trait (under a lot of assumptions that will covered later), is that the genotypic value of the offspring can be expressed in terms of the genotypic values of its sire (s) and dam (d), based on the fact that half of the genes that the offspring have come from each parent:

$$G_o = \tfrac{1}{2} G_s + \tfrac{1}{2} G_d + RA_s + RA_d$$

Here the terms RA_s and RA_d are random assortment or Mendelian sampling terms, which reflect that parents pass on a random half of their alleles (i.e. a random one of two alleles at each locus).
Developing these quantitative genetic models and deducing their consequences, e.g. the consequences of natural or artificial selection on the trait and the population, then involves manipulating the mathematical terms, that is doing algebra and even a little calculus sometimes (!). Quantitative geneticists were really pioneers in this type of mathematical treatment of biological phenomena and as a result the early growth of quantitative genetics was almost synonymous with the early growth of statistics. Indeed, R.A. Fisher is hailed as a founder of quantitative genetics but also of analysis of variance and randomization procedures in statistics. The early geneticists Galton and Pearson originated the concepts of regression and correlation. Anyway, the upshot for us here is that we will be deeply involved with the mathematical manipulation and statistical evaluation of our representations of the basic quantitative genetic model. We will review some of the rules of probability and statistics, such as variance, covariance, correlation and regression, and will give a hint at how they may relate to the quantitative genetic model.

Variance

The variance of a random variable is a measure of the spread of a variable over the sample space. Intuitively, we want to know how far we can expect the value of a given variable on average to be from its expected value. That is, we want to know something about the average deviation of the

random variable from its expected location. The way to obtain a variance is to find the average of the squared deviation from the mean:

$\text{var}(Y) = E\{[Y - \mu_Y]^2\}$ where $\mu_Y = E(Y)$

$= E\{Y^2 - 2Y\mu_Y + \mu_Y{}^2\} = E(Y^2) - 2E[Y\mu_Y] + \mu_Y{}^2 = E(Y^2) - 2\mu_Y\mu_Y + \mu_Y{}^2$

Thus: $\mathbf{var(Y) = } E\{[Y - \mu_Y]^2\} = \boldsymbol{E(Y^2) - \mu_Y{}^2}$

Note from your statistics class that when we have a sample of N observations for a random variable X (instead of frequencies of the variable attaining certain values), the variance of the sample can be computed as: $\text{var}(X) = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$ or as $\text{var}(X) = \dfrac{\sum_{i=1}^{N} x_i{}^2}{N} - \bar{x}$ where $\bar{x}$ is the average of X

Realizing that taking the average is sample equivalent to taking the expectation of a variable, note that these equations are similar to the equations for variances based on expectations, as given above.

Covariance

The covariance between variables X and Y quantifies the (linear) relationship or dependence between X and Y based on the extent to which they "co-vary".

$\text{Cov}(X, Y) = \text{E}\{[X - \mu_X][Y - \mu_Y]\}$

$= \text{E}(XY) - \mu_X\mu_Y$ where $\text{E}(XY) = \Sigma_i\Sigma_j\, x_i y_j \Pr(X = x_i, Y = y_j)$

Example: The covariance between the genotypic value and the phenotypic value will play a big role in quantitative genetic inferences.

Properties of Variance and Covariance

Assuming again that a is a constant:

Var(a) = 0 The variance of a constant is zero

Var(aX) = a^2Var(X) The variance of the product of a variable by a constant is the product of the constant squared and the variable's variance

Cov(X,Y) = Cov(Y,X)

Cov(X,aY) = aCov(X,Y)

Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)

Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) The variance of a sum is the sum of variance plus twice the covariance

Generalizing the equation for Var(X+Y) to the sum of many variables:

$\text{Var}(\Sigma_i X_i) = \Sigma_i \text{Var}(X_i) + 2\Sigma_{(i<j)}\text{Cov}(X_i, X_j)$ If X's are independent ➔ $\text{Var}(\Sigma_i X_i) = \Sigma_i \text{Var}(X_i)$

Also: **Var(X-Y)** = Var[X+(-Y)] = Var(X) + Var(-1*Y) + 2Cov[X,(-1*Y)] =

$= \text{Var}(X) + (-1)^2\text{*Var}(Y) + 2\text{*}(-1)\text{*Cov}(X,Y) =$

$= \mathbf{Var(X) + Var(Y) - 2Cov(X,Y)}$

Cov(X, X) $= \text{E}(XX) - \text{E}(X)\text{E}(X)$

$= \text{E}(X^2) - [\text{E}(X)]^2$

= Var(X) ➔ the covariance of a variable with itself is its variance

If X and Y are independent: $E(XY) = \Sigma_i \Sigma_j x_i y_j \Pr(X = x_i, Y = y_j)$

$= \Sigma_i \Sigma_j x_i y_j \Pr(X = x_i)\Pr(Y = y_j)$

$= [\Sigma_i x_i \Pr(X = x_i)] [\Sigma_j y_j \Pr(Y = y_j)]$

$= E(X)E(Y)$

So that $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$

Correlation

The correlation measures the (linear) relationship between two variables on a standardized scale, by dividing their covariance by the product of their standard deviations:

$$r_{XY} = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Note that: $-1 \le r_{XY} \le 1$

Based on rearrangement of the correlation equation, we get the following expression for the covariance, which we also frequently use:

$$\mathrm{Cov}(X,Y) = r_{XY}\sqrt{Var(X)Var(Y)}$$

Variance of a Mean

If the observations that make up the mean are independent and identically distributed (iid), then Var(Mean) = Var(individual observation)/n = σ^2/n

If observations that make up the mean are correlated, with correlation = *t*. This is referred to as the intra-class correlation for that 'class' of observations. then the mean can be split into a part that is consistent for each observation and a part that is unique for each observation. The variance of the consistent bit is what makes observations co-vary and is equal to $t\sigma^2$. The rest of the variance, i.e. $(1\text{-}t)\sigma^2$ is due to the unique bits. And, since the unique bits are iid, the variance of the average of the unique bits is $(1\text{-}t)\sigma^2/n$. Thus, the variance of the mean of n observations with intraclass correlation t is: $\mathrm{Var(Mean)} = t\sigma^2 + (1\text{-}t)\sigma^2/n = \frac{1+(n-1)t}{n}\sigma^2$

Regression

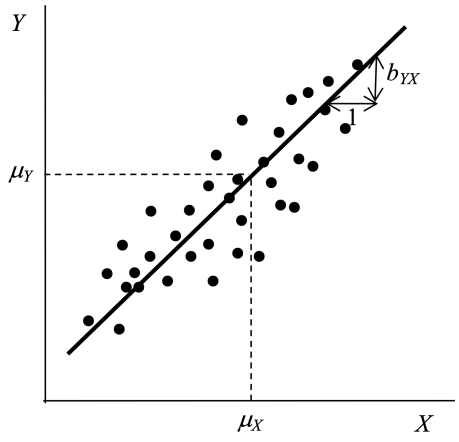
A repeated theme in quantitative genetics is the estimation of quantities associated with individuals or parameters associated with populations when those quantities or parameters are themselves not directly observable. The most obvious example is the desire to estimate an individual's genotypic value for a trait when the only information we have available derives from the individual's phenotype. Regression is used for this kind of estimation.

Definition: The regression of *Y* on *X* is the expected value of Y conditional on having a certain value for variable X: $\hat{y} = E(Y|X)$

This is also called the ***best*** *(linear)* ***predictor*** of *Y* given *X*.

Regression can be used to define a model: $y = \hat{y} + e$

where e is the residual, which is the deviation of the observed value for Y from its expected value conditional on X.



For quantitative variables, the predicted value for Y can be derived using linear regression:

$$\hat{y} = E(Y|X) = \mu_Y + b_{YX}(x-\mu_X)$$

with $\mu_Y = E(Y)$ $\mu_X = E(X)$

b_{YX} = coefficient of regression of Y on X = expected change in Y per 1 unit increase in X

Given data, b_{YX} can be derived by fitting the following linear regression model:

$$y = \mu_Y + b_{YX}(x-\mu_X) + e$$

Using least squares (see Lynch & Walsh p39), b_{YX} can be derived to be equal to:

$$b_{YX} = \mathrm{Cov}(Y,X)/\mathrm{Var}(X)$$

Note that b_{YX} can also be expressed in terms of the correlation coefficient:

$$b_{YX} = \mathrm{Cov}(Y,X)/\mathrm{Var}(X) = r_{XY}\sqrt{Var(Y)Var(X)} / \mathrm{Var}(X) = r_{XY}\sqrt{\frac{Var(Y)}{Var(X)}}$$

So the important equations to remember for the regression coefficient are:

$$\boxed{b_{YX} = \frac{Cov(Y,X)}{Var(X)} = r_{XY}\sqrt{\frac{Var(Y)}{Var(X)}}}$$

Note that these only hold for simple regression with a single independent variable (X).

Properties of Regression

1. The average of predicted values is equal to the average of Y's: $\mathbf{E(\hat{Y}) = E(Y) = \mu_Y}$

   $$E(\hat{y}) = E[\mu_Y + b_{YX}(x-\mu_X)] = E(\mu_Y) + E[b_{YX}(x-\mu_X)] = \mu_Y + b_{YX}[E(x)-\mu_X] = \mu_Y$$

   This also implies that the regression line always passes through the mean of both X and Y; substituting μ_X for x into the prediction equation gives $\hat{y} = \mu_Y$

2. The average value of the residual is zero: $\mathbf{E(e) = 0}$.

   $E(e) = E(Y - \hat{Y})$ *from regression model*
   $= E(Y) - E(\hat{Y})$ *property of expectation*
   $= 0$ *from property 1 above*

3. The expectation of the residual is zero for all values of X: $\mathbf{E(e|X) = 0}$

   $E(e|X) = E(Y - \hat{Y}|X)$ *from regression model*
   $= E(Y|X) - E(\hat{Y}|X)$ *property of expectation*
   $= \hat{Y} - \hat{Y} = 0$ *by definition of regression*

   This implies that predictions of Y are on average equal to the true Y across the range of possible values for X.

4. ***Accuracy of prediction* = Corr($\hat{Y}$,*Y*) = $r_{\hat{y}y}$**

The accuracy of the prediction equation is equal to the correlation of $\hat{y}$ with its true value *y*. We can derive accuracy as:

$$\text{Accuracy} = r_{\hat{y}y} = \frac{\text{Cov}(\hat{y}, y)}{\sqrt{\text{Var}(\hat{y})\text{Var}(y)}} = \frac{\text{Cov}(\mu_Y + b_{YX}(x - \mu_X), y)}{\sqrt{\text{Var}(\mu_Y + b_{YX}(x - \mu_X))\text{Var}(y)}}$$

Since μ_Y and μ_X are constants, this simplifies to:

$$\text{Accuracy} = \frac{\text{Cov}(b_{YX}x, y)}{\sqrt{\text{Var}(b_{YX}x)\text{Var}(y)}} = \frac{b_{YX}\,\text{Cov}(x, y)}{\sqrt{b_{YX}{}^2\text{Var}(x)\text{Var}(y)}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \text{r}_{XY}$$

So the accuracy of a prediction equation based on simple (= 1-variable) regression is equal to the correlation between the dependent and independent variables.

5. ***Decomposition of variance in Y into that explained by the prediction and unexplained variance***

Using the above equation, we can also show that the variance of *Y* is the sum of the variance explained by the regression on *X* and residual variance (note that Cov(*X,e*)=0):

$$\text{Var}(y) = \text{Var}(\mu_Y + b_{YX}(x\text{-}\mu_X) + e) = b_{YX}^2\text{Var}(x) + \text{Var}(e) = [\text{Cov}(y,x)]^2/\text{Var}(x) + \text{Var}(e)$$

Note that because $\text{Cov}(y,x) = \text{r}_{XY}\sqrt{Var(x)Var(y)}$ the first term can also be written as:

$$[\text{Cov}(y,x)]^2/\text{Var}(x) = \text{r}_{XY}{}^2\,\text{Var}(x)\,\text{Var}(y) / \text{Var}(x) = \text{r}_{XY}{}^2\,\text{Var}(y) = r_{\hat{y}y}^2\,\text{Var}(y)$$

This is the variance in *Y* that is explained by the *X* through the prediction model

By subtraction we get $\text{Var}(e) = [1\text{-}\,\text{r}_{XY}^2]\,\text{Var}(y)$. This is the unexplained/residual variance.

Thus, variance of *Y* can be decomposed as: $\text{Var}(y) = \text{r}_{XY}^2\text{Var}(y) + [1\text{-}\,\text{r}_{XY}^2]\,\text{Var}(y)$

Note that the variance of predicted values is equal to the explained variance:

$$\text{Var}(\hat{y}) = \text{Var}[\mu_Y + b_{YX}(x\text{-}\mu_X)] = b_{YX}^2\text{Var}(x) = \left\{\frac{Cov(y,x)}{Var(x)}\right\}^2 \text{Var}(x) =$$

$$= \frac{[Cov(y,x)]^2}{Var(x)} = \frac{[Cov(y,x)]^2}{Var(x)Var(y)}Var(y) = r_{YX}^2\text{Var}(y)$$

So the variance of predicted values is equal to the variance explained by the model, which depends on the correlation between *Y* and *X*.

The above equations apply when prediction is based on one variable (*x*), in which case $r_{\hat{y}y} = r_{XY}$.

In general, prediction can be based on multiple *x*'s = multiple regression. In that case the partitioning of variance is: $\text{Var}(y) = \text{Var}(\hat{y}) + \text{Var}(e) = r_{\hat{y}y}^2\,\text{Var}(y) + [1\text{-}\,r_{\hat{y}y}^2]\,\text{Var}(y)$

6. Residuals are uncorrelated with the predictor variable, *X*: **Cov(*X*,*e*) = 0**

$$\begin{aligned}\text{Cov}(x, e) &= \text{Cov}[x, y - \hat{y}] = \text{Cov}[x, y - (\mu_Y + b_{YX}(x\text{-}\mu_X))] = \\ &= \text{Cov}(x, y) - \text{Cov}(x,\mu_Y) - b_{YX}\,\text{Cov}(x, x) - b_{YX}\,\text{Cov}(x,\mu_X) \\ &= \text{Cov}(x, y) - 0 - b_{YX}\,\text{Var}(x) - 0 \\ &= \text{Cov}(x, y) - \frac{Cov(y,x)}{Var(x)}\text{Var}(x) = 0\end{aligned}$$

7. Residuals are uncorrelated with the predictions: **Cov($\hat{y}$, *e*) = 0**

$$\text{Cov}(\hat{y}, e) = \text{Cov}[\hat{y}, y - \hat{y}] = \text{Cov}(\hat{y}, y) - Var(\hat{y}) =$$

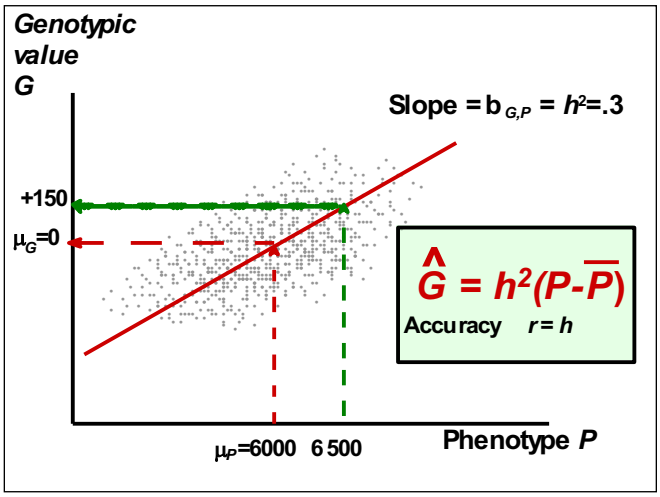
$$= \mathrm{Cov}(x, y) - \mathrm{Cov}(x,\mu_Y) - b_{YX}\,\mathrm{Cov}(x, x) - b_{YX}\,\mathrm{Cov}(x,\mu_X)$$
$$= \mathrm{Cov}(x, y) - \quad 0 \quad - b_{YX}\,\mathrm{Var}(x) \quad - \quad 0$$
$$= \mathrm{Cov}(x, y) - \frac{Cov(y,x)}{Var(x)}\mathrm{Var}(x) \quad = 0$$

Properties 6 and 7 imply that all information on Y that is contained in X is captured in the predicted values, as the residual is uncorrelated to both X and the predicted values.

Some Applications to Quantitative Genetic Theory

The standard quantitative genetics model equation for the observed phenotype of an individual i for a quantitative trait (P_i) is that it is the sum of the effect of genetics (the genotypic value G_i) and the effect of environment (E_i): $P_i = \mu_P + G_i + E_i$

In practice, we only observe phenotype and cannot directly observe G_i or E_i. However, ***if*** we could observe both P_i and G_i for a group of individuals, we could plot genotypic against phenotypic values, as in the figure below.
Using such a hypothetical plot, or model, and statistics such as correlation, covariance, variance, and regression, we can specify important population parameters such as heritability (h^2) and make a number of inferences or predictions, such as predicting an individual's genotypic value or 'breeding value' from its observed phenotype:



1) Covariance and correlation between phenotypic and genotypic values:

Based on $P_i = \mu_P + G_i + E_i$

$$\mathrm{Cov}(P,G) = \mathrm{Cov}(\mu + G + E, G) =$$
$$= \mathrm{Cov}(G,G) + \mathrm{Cov}(G, E) = \mathrm{Var}(G)$$

The last step assumes that $\mathrm{Cov}(G, E) = 0$, i.e. that the environment that an individual receives is independent of its genotypic value. The result of this covariance, $\mathrm{Var}(G)$, which is often denoted σ_G^2, is the ***genetic variance*** in the population, i.e. the variance of genotypic values of individuals in a population. This in contrast to the phenotypic variance, $\mathrm{Var}(P)$, often denoted σ_P^2, which is the variance of phenotypic values of individuals in a population.

Then, the correlation between phenotypic and genotypic values can be derived as:

$$r_{P,G} = \frac{Cov(G,P)}{\sqrt{Var(G)Var(P)}} = \frac{\sigma_G^2}{\sqrt{\sigma_G^2\sigma_P^2}} = \frac{\sigma_G}{\sigma_P}$$

Thus, the correlation between genotypic and phenotypic values of individuals in a population is equal to the ratio of the genetic and phenotypic standard deviations for the trait.
The square of this correlation, therefore, is equal to the ratio of the genetic and phenotypic variances, or to the proportion of phenotypic variance that is genetic. This proportion is also defined as the ***heritability*** of the trait (= h^2). Thus: $(r_{P,G})^2 = = \frac{\sigma_G^2}{\sigma_P^2} = h^2$

2) Regression of genotypic on phenotypic values:

Using the above model and referring to the figure, we can also set up a regression equation between the genotypic and phenotypic values to predict *G*:

$G_i = \mu_G + b_{G,P}(P_i - \mu_P) + e_i$ where $b_{G,P}$ is the coefficient of regression of *G* on *P*.

This regression coefficient can be derived as: $b_{G,P} = \frac{Cov(G,P)}{Var(P)} = \frac{\sigma_G^2}{\sigma_P^2} = h^2$

Thus, the slope of the regression of genotypic on phenotypic values is equal to heritability

3) Prediction of genotypic values:

The above regression model can be used to predict an individual's genotypic value based on it's observed phenotype, using the following prediction equation:

$$\hat{G}_i = \mu_G + h^2(P_i - \mu_P)$$

In practice, we often set μ_G to zero, because we're primarily interested in ranking individuals in a population. Thus: $\hat{G}_i = h^2(P_i - \mu_P)$

As an ***example*** (see figure), assume a dairy cow produces 6500 kg milk, which is its phenotypic value (P_i). The mean production of the herd she is in is 6000 kg (= μ_P).

Milk production is a trait with an (assumed known) heritability of 0.3, a phenotypic standard deviation of 1200 kg (σ_P=1200). Using $h^2 = \frac{\sigma_G^2}{\sigma_P^2}$ and, thus, $\sigma_G^2 = h^2\sigma_P^2$, the genetic standard deviation for milk yield is equal to $\sigma_G = h\sigma_P = \sqrt{0.3} * 1200 = 657.3$ kg

Then, this cow's genotypic value can be predicted to be:

$$\hat{G}_i = h^2(P_i - \mu_P) = 0.3\ (6500\text{-}6000) = +150 \text{ kg}$$

So this cow's genotypic value is expected to be 150 kg greater than the average in this herd.

We can also attach an accuracy to this prediction, based on the previously derived result that the correlation between predicted and true values based on linear regression is equal to the correlation to the dependent (*Y*) and independent (*X*) variables: $r_{\hat{G},G} = r_{G,P} = h$

Thus, when predicting an individual's genotypic value based on its phenotypic value, the accuracy of this prediction will be equal to the square-root of heritability of the trait.

When we predict genotypic values for all individuals in a population in this manner, and take the variance of these predicted values, we expect this variance to be equal to (based on property 5):

$$\text{Var}(\hat{y}) = r_{YX}^2 \text{Var}(y)$$

which in this case simplifies to: $\text{Var}(\hat{G}) = h^2\sigma_G^2 = h^4\sigma_P^2$

And, using property 5 above, the variance of prediction errors ($e_i = G - \hat{G}$) is equal to:

$$\text{Var}(e) = [1 - r_{XY}^2]\ \text{Var}(y)$$

which in this case simplifies to: $\text{Var}(e) = (1-h^2)\sigma_G^2$

For the example, the variance of predicted values is: $\text{Var}(\hat{G}) = h^2\sigma_G^2 = 0.3*657.3^2 = 129600$

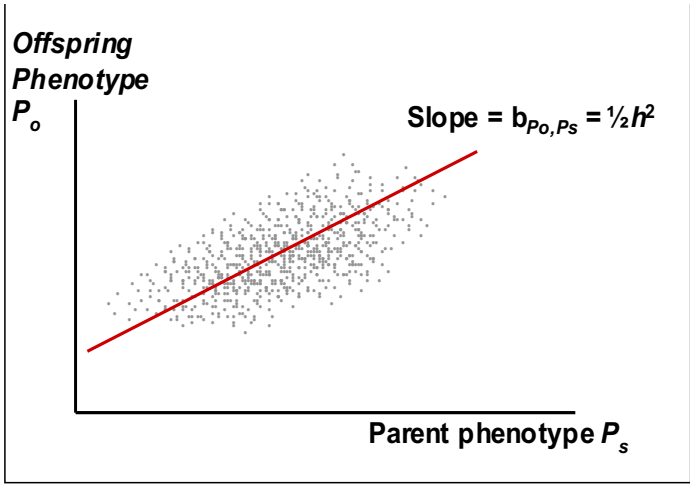
and the variance of prediction errors is: $\text{Var}(e) = (1-h^2)\sigma_G^2 = 0.7*657.3^2 = 302430$

Note that these two variances sum to the genetic variance: $129600 + 302430 = 43203 = 657.3^2$

Based on $\text{Var}(e) = 302430$ kg^2 we can also add a confidence interval to our prediction (see later).

4) Regression of offspring phenotype on parent phenotype

One of the problems with predicting genotypic values, as described above, is that it requires you to know the heritability of the trait. Luckily, we can also get estimates of heritability for a trait from phenotypic data. We do this by observing how similar the phenotype of offspring is to that of their parents; if these are very similar, we expect the trait to be more heritable.



When we have phenotypes observed on offspring and their sires, we can estimate heritability by regressing the phenotype of the offspring on that of their parents, as illustrated below:

The regression model is:

$$P_o = \mu_o + b_{PoPs}(P_s - \mu_s) + e$$

The regression coefficient can be derived as:

$$b_{PoPs} = \frac{Cov(P_o, P_s)}{Var(P_s)} = \frac{Cov(G_o + E_o, G_s + E_s)}{\sigma_P^2}$$

$$= \frac{Cov(G_o, G_s) + Cov(G_o, E_s) + Cov(E_o, G_s) + Cov(E_o, E_s)}{\sigma_P^2} = \frac{Cov(G_o, G_s)}{\sigma_P^2}$$

The last step assumes that the environment that the offspring progeny received is independent of the phenotype of the sire (a sometimes strong assumption), making the last 3 covariance terms 0.

To derive the covariance between the genotypic value of offspring and that of their sire, we can express the genotypic value of the offspring in terms of the genotypic values of its sire (s) and dam (d), based on the fact that half of the genes that the offspring have come from each parent:

$$G_o = ½\ G_s + ½\ G_d + RA_s + RA_d$$

Here the terms RA_s and RA_d are random assortment or Mendelian sampling terms, which reflect that parents pass on a random half of their alleles (i.e. a random one of two alleles at each locus). Using this genetic model (which has quite a number of assumptions, which well be covered later), we can continue our derivation as:

$$Cov(G_o, G_s) = Cov(½G_s + ½G_d + RA_s + RA_d, G_s) =$$

$$= Cov(½G_s, G_s) + Cov(½G_d, G_s) + Cov(RA_s, G_s) + Cov(RA_d, G_s)$$

Assuming random mating and the fact that Mendelian sampling terms are independent (see later), the last three covariance terms are zero, resulting in:

$$b_{PoPs} = \frac{Cov(G_o, G_s)}{\sigma_P^2} = \frac{Cov(½\ G_s, G_s)}{\sigma_P^2} = \frac{½\ Cov(G_s, G_s)}{\sigma_P^2} = \frac{½\sigma_G^2}{\sigma_P^2} = ½h^2$$

Thus, heritability of a trait can be estimated based on phenotypes of relatives, by measuring the degree of resemblance between relatives, using statistics such as linear regression. More on this later.

Some Distributions useful in Population and Quantitative Genetics

Bernoulli distribution.

Named after the mathematician Daniel Bernoulli, 1700-1782. A Bernoulli random variable is characterized by one parameter, that is typically designated p and is sometimes called the "probability of success". The random variable can have one of two values: 1 with probability p and 0 with probability $1 - p$.

If Y is a Bernoulli random variable with probability p, its expectation is:

$$E(Y) = \sum_{i=1}^{2} y_i \Pr(Y = y_i) = 0(1-p) + 1(p) = p$$

Its variance is

$$\begin{aligned} \text{var}(Y) &= E(Y^2) - E(Y)^2 \\ &= [0^2(1-p) + 1^2(p)] - p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

The Bernoulli distribution is used in population and quantitative genetics in relation to the presence or inheritance of alleles at a locus. For example, for a locus with two possible alleles, A and a, and with the frequency of allele A in the population denoted by p, then the process of drawing one allele at this locus from a population can be specified by a Bernouilli distribution by specifying a variable Y that is equal to 1 if allele A is drawn and equal to 0 if allele a is drawn.

Binomial Distribution

The Binomial distribution is based on the Bernoulli distribution. A binomial random variable is the sum of k independent Bernoulli random variables all with parameter p. The binomial is therefore characterized by two parameters, k and p and can have integer values from 0 to k. If X is binomially distributed with k trials and p probability of success: $X \sim \text{Binomial}(k, p)$, then:

From the properties of expectation of a sum, the **expected value** of X is kp: $E(X) = kp$.
From the properties of variance of a sum of independent variables, the **variance** of X is $kp(1-p)$: $\text{var}(X) = kp(1-p)$

The probability density function $\Pr(X = x)$ is

$$\Pr(X = x) = \binom{k}{x} p^x (1-p)^{k-x} \text{ where } \binom{k}{x} = \frac{k!}{x!(k-x)!} \text{ and } \quad a! = 1*2*3*\ldots*a$$

When considering population or quantitative genetics, the Binomial Distribution could correspond to the process of randomly drawing k alleles at a locus from a population.

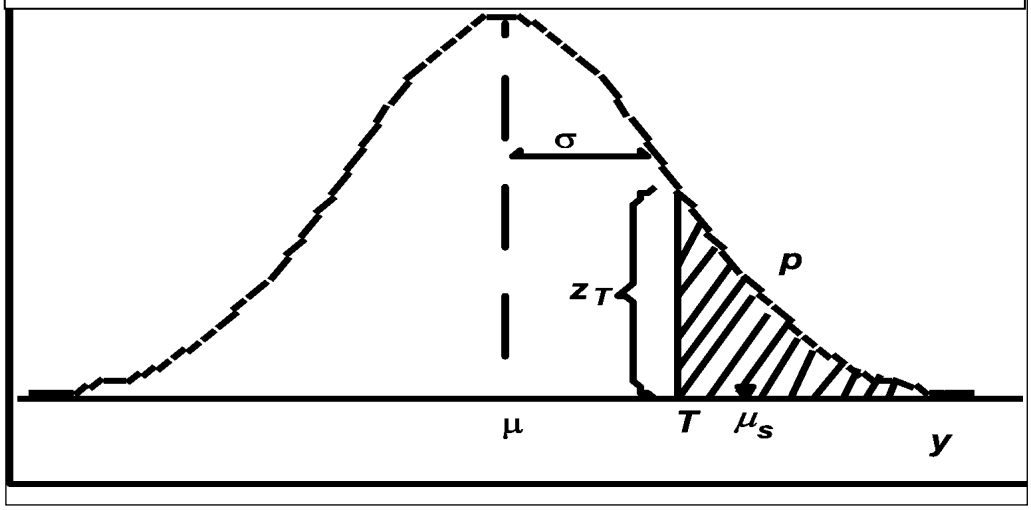
Normal or Gaussian distribution.

This is perhaps the most important distribution in quantitative genetics, as phenotypes for most quantitative traits approximately follow a normal distribution, or can be transformed to follow a normal distribution. This is a property of the fact that phenotype is the sum of many genetic factors and of many environmental factors. Following the Central Limit Theorem of statistics, this is expected to result in a Normal distribution, even if the distribution of variables that are included in the sum is not Normal.

The probability distribution function for a variable y that has a Normal distribution with mean μ and standard deviation σ, denoted by $y \sim N(\mu, \sigma^2)$ is:

$$\Pr(y) = z = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]}$$

Normal distribution $N(\mu,\sigma^2)$ for variable y with mean μ and st.dev. σ.
Truncated Normal distribution for fraction p selected, truncation point T,
ordinate height z_T, and mean of selected group μ_S .



It is often useful to work with the *Standard* Normal distribution, which has mean zero and standard deviation 1: N(0,1) Any Normally distributed variable can be 'standardized' to a variable that follows N(0,1) by subtracting the mean and dividing by the standard deviation:

If $y \sim N(\mu,\sigma^2)$ then $y' = (y-\mu)/\sigma$ follows N(0,1)

Truncated Normal distribution.

In plant and animal breeding, we often are interested in using individuals with the highest phenotype for breeding. If phenotype (y) is Normally distributed ($y \sim N(\mu,\sigma^2)$) then it is of interest to know something about the distribution of phenotypes of the selected individuals. This is the Truncated Normal distribution, as illustrated in the figure above:

Selecting a proportion p of individuals from a population based on phenotype (y) is equivalent to truncating the Normal distribution at a truncation point T, such that a fraction p falls above the truncation point.

The mean phenotype for the selected individuals is denoted by μ_S (see Figure).

The difference between the mean of the selected individuals over that of all individuals is called the selection differential: $S = \mu_S - \mu$