# Statistical Foundations for AGRO/ANS-931 Fall 2020

This handout serves as the statistical foundations for the population genetics (AGRO/ANS-931) class for Fall Semester 2020 at UNL. The text here derives essentially from previous course notes by Dr.s Rohan Fernando, Jean-Luc Jannink, and Jack Dekkers at Iowa State University. It has been revised and formatted by Dr. Jinliang Yang.

Quantitative genetics dwells primarily on developing theory or statistical models that represent our understanding of the phenomena of interest, and uses that theory to make predictions about how those phenomena will behave under specific circumstances.

**What is a quantitative trait?**

The model that exists to explain observations of `quantitative traits` contains the following components:

- Genetic factors that act in pairs (2 alleles per locus) but that are passed on to progeny individually.
- Many such genetic factors.
- Which of the parent's genetic factors (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.
- Genetic factors sometimes show independent assortment (unlinked loci); sometimes not (linked loci).
- Environmental factors influence the trait.

In order to develop the theory 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. Deducing the theory's consequences 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 to make sure we do those manipulations well and to get a hint at how those rules may relate to the quantitative genetic model.

## Part One:

### Random Variables

In principle, we are interested in the random and non-random processes that determine the value of variables. If the variable of interest is which allele a heterozygous ($Mm$) father passed on to his daughter for a given marker locus, the rule of random segregation indicates that this is a random process. If the variable of interest is the height of the son of a tall woman, some portion of the variable will be non-random (we expect a relatively tall son) and some portion will be random (we don't know exactly what the height will be). Either way, we can identify a random variable with a symbol (say $X^p$ to designate the paternally inherited marker allele, or $Y$ to designate height). Common notation is to use capitals for the name of a variable (e.g. $X$ or $Y$) and regular font to represent the value (or class) of that variable. E.g. $X = x$ indicates the event that variable $X$ has value $x$.

### Sample Space

The sample space is the set of possible values that a random variable can take. So, for example $X^p \in [M, m]$ (i.e., the progeny inherits

either allele $M$ or allele $m$ from its heterozygous $Mm$ father), and $1 < Y < 2.5$ if height is measured in meters. Note that these two example random variables are very different. Random variable $X^p$ can take on just two states (one of the two alleles that the parent has), it is a **categorical variable**, while $Y$ can take on all values between 1 and 2.5, it is a **continuous variable**. Nevertheless, many of the mathematical manipulations we will discuss below can be applied equally to either type variable.

## Probability

We designate the probability of an event A as $Pr(A)$. For example, if the event A is "the daughter received marker allele M from her father" then $Pr(A) = Pr(X^p = M)$. In this case $Pr(A) = 1/2$. The probability function $Pr(\star)$ has certain rules assigned to it, just like, for example multiplication has rules assigned to it. For example, if event A is "any possible event in the sample space of events" then $Pr(A) = 1$. Thus, the probability that $X^p = M$ or $X^p = m$ for a progeny of the heterozygous $Mm$ father is equal to $1/2 + 1/2 = 1$. Intuitively, though, it is most useful to think of the $Pr(A)$ as the chance that event A will happen. If you look at many events (N events, with N very big) and you count $N_A$, the number of times event A happens, then we can interpret $Pr(A)$ as a frequency, i.e. $Pr(A) = N_A/N$. As examples related to the random variables we gave above, if the father is a heterozygote, then Mendel's law of segregation say $Pr(M^p = 0) = Pr(M^p = 1) = 1/2$. For the height Y of the son of a tall woman, we can guess that $Pr(1.5 < Y \leqslant 1.6) < Pr(1.8 < Y \leqslant 1.9)$, that is, the son is less likely to be in a short ten centimeter bracket than a relatively tall ten centimeter bracket.

## Probability Density

The second example leads to the question what is $Pr(Y = 1.8)$? And the answer, oddly, is zero. That is, given that Y can take on an infinite number of values in the range of 1 to 2.5, there is a probability of zero that it will take on any specific value. Intuitively, though, we want to be able to express the idea that the chance that the height will be some tall value is greater than the chance it will be some short value. To do this we define the probability density $f(y) = Pr(y < Y \leqslant y + e)/e$ as $e$ comes increasingly close to zero. This probability density will be useful to discuss random variables that vary continuously (such as the value of a quantitative trait). Using the **probability density function (or pdf)** and integration, we can calculate the probability that Y is contained in a certain bracket as

$$Pr(1.5 < Y \leqslant 1.6) = \int_{1.5}^{1.6} f(y) d_y$$

The most prominent **pdf** that we will use is that of the normal distribution, i.e. the bell-shaped curve.

## Expected Value

The expected value of a random variable is a measure of its location in the sample space, and can be thought of as a mean or an average. It takes slightly different forms depending on whether the variable is categorical or continuous. Consider a categorical variable $X_k$ with sample space $x_1, x_2, ..., x_k$. The expected value of $X$ is:

$$E(X) = \sum_{i=1}^{k} x_i Pr(X = x_i).$$

In general for any function of the random variable we have the formula:

$$E(f(X)) = \sum_{i=1}^{k} f(x_i) Pr(X = x_i) \qquad (1)$$

### Properties of Expectation

Assuming X and Y are random variables and $a$ is a constant (e.g., $a = 5$):

The expectation of a constant is that constant:

$$E(a) = a$$

The expectation of the product of a random variable by a constant is the product of the constant and the expectation of the random variable:

$$E(aX) = aE(X)$$

The expectation of a sum of two random variables is the sum of their expectations.

$$E(X + Y) = E(X) + E(Y)$$

Note that, generally, $E(XY)$ is NOT equal to $E(X)E(Y)$!

$E(XY) = E(X)E(Y)$ ONLY IF X and Y are independent — see later.

**Example 1**

The number of florets per spikelet in oat (= variable X) as affected by a recessive allele that inhibits the development of tertiary kernels (this example is slightly fictitious but serves its purpose). Note that the expected value of a categorical trait may not belong to any of the categories of the trait: the expected value for the number of florets per spikelet is E(X) = 2.75 though any given spikelet obviously has a whole number of florets.

Consider again a categorical variable X with sample space $x_1, x_2, ..., x_k$. There is also a function $g(X)$, and we want the expected value of $g(X)$. The formula has the same form: $E(g(X)) = \sum_{i=1}^{k} g(x_i)Pr(X = x_i)$. Here, $E(g(X))$ means that the expectation is taken over all possible values of variable X, e.g., referring back to Example 1, the expectation of $g(X) = X^2$ is equal to 7.75, as calculated in the last column in Table 1.

## 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:

$$\begin{aligned} Var(Y) &= E(Y - E(Y)^2) \\ &= E(Y^2 - 2YE(Y) + E(Y)^2) \\ &= E(Y^2) - 2E[YE(Y)] + E(Y)^2 \\ &= E(Y^2) - 2E(Y)E(Y) + E(Y)^2 \end{aligned}$$

Thus,

$$Var(Y) = E(Y^2) - E(Y)^2 \quad (2)$$

**Properties of Variance**

Assuming again that $a$ is a constant. The variance of a constant is zero.

$$Var(a) = 0$$

The variance of the product of a variable by a constant is the product of the constant squared and the variable's variance

$$Var(aX) = a^2 Var(X)$$

Looking back at Example 1, the number of florets per spikelet given different genotypes:

$$Var(X) = E(Y^2) - E(Y)^2 = 7.75 - (2.75)^2 = 0.1875$$

# Part Two:

## Joint Probability

The joint probability is the probability for given values of two or more random variables to occur together. The joint probability that random variable $X = x$ and random variable $Y = y$ is denoted $Pr(X = x, Y = y)$.

**Example 2**

Assume a genetic locus $A$ that influences milk yield. The genotypes of specific cows are obtained (= variable $G$) and their milk yield (= variable $MY$) is measured and divided into categories ($MY \leqslant 100; 100 < MY \leqslant 300; MY > 300$). One obtains the joint probability of carrying specific genotypes and belonging to a certain category of milk yield as shown in table 2.

The entries in the body of this table are the joint probabilities. So, for example the joint probability that a cow has genotype $Aa$ and produces more than 300 kg is: $Pr(G = Aa, MY > 300) = 0.16$.

## Marginal probability

Marginal probability is used in Table 2 to show the probabilities of, for example, $MY \leqslant 100$, as the sum across a column of joint probabilities.

| Genotype (G) | Probability (frq) $Pr(X = x_i)$ | Number of florets, $X = x_i$ | $X_i \times Pr(X = x_i)$ | $x_i^2 \times Pr(X = x_i)$ |
|---|---|---|---|---|
| ts / ts | 025 | 3 | 0.75 | 2.25 |
| Ts / ts | 0.50 | 3 | 1.50 | 4.50 |
| Ts / Ts | 0.25 | 2 | 0.50 | 1.00 |
| Sum | 1.00 | - | $E(X) = 2.75$ | $E(X^2) = 7.75$ |

Table 1: The expected number of florets in oat.

| Genotype | Mike Yield (MY) | | | Marginal |
|---|---|---|---|---|
| (G) | $MY \leqslant 100$ | $100 < MY \leqslant 300$ | $MY > 300$ | $Pr(G)$ |
| aa | 0.10 | 0.04 | 0.02 | 0.16 |
| Aa | 0.14 | 0.18 | 0.16 | 0.48 |
| AA | 0.06 | 0.10 | 0.20 | 0.36 |
| Marg. Prob. MY | 0.30 | 0.32 | 0.38 | 1.00 |

Table 2: Joint probability of genotype and milk yield.

That is,

$$\begin{aligned} &Pr(MY \leqslant 100) = Pr(MY \leqslant 100, G = aa) \\ &+ Pr(MY \leqslant 100, G = Aa) \\ &+ Pr(MY \leqslant 100, G = AA) \\ &= 0.10 + 0.14 + 0.6 = 0.30 \end{aligned}$$

What works in the columns for milk yield also works in the rows to get marginal probabilities for genotype.

In general if $E_1, E_2, ..., E_n$ is a **mutually exclusive** and **exhaustive** set of events (i.e. a set of non-overlapping events that includes the complete parameter space for the variables involved), then marginal probabilities for event $A$ can be calculated as the sum of joint probabilities of event A and events $E_i$:

$$Pr(A) = \sum_{i=1}^{n} Pr(E_i, A)$$

In Table 2, for example, events $G = aa$, $G = Aa$, and $G = AA$ are mutually exclusive and exhaustive events and marginal probabilities for $MY$ can be obtained by summing the joint probabilities in a column of Table 2.

## Conditional probability

Intuitively, the conditional probability is the probability of a certain event to occur when you already know that another event is true. Alternately, it is the probability of obtaining a given value for one variable (say, $X = x$), conditional on the fact that the value of another variable (say $Y = y$) value has already been observed. This conditional probability is denoted $Pr(X = x|Y = y)$. First, in order to obtain a given value for $X$ (say $X = x$) while $Y$ has another value (say $Y = y$), both conditions have to hold. So we need the joint probability $Pr(X = x, Y = y)$. Second, because we know that $Y = y$, the parameter space for $X$ is restricted to the subset of events where $Y = y$. All this to help you intuit the definition of conditional probability:

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} \quad (3)$$

In words, the probability of $X = x$ given $Y = y$, is the joint probability of $X = x$ and $Y = y$ divided by the marginal probability of $Y = y$.

Referring back to Table 2, the probability of $Aa$ cows producing more than 300 kg of milk is the probability of $MY > 300$ conditional on $G = Aa$, which is:

$$\begin{aligned} &Pr(MY > 300|G = Aa) \\ &= \frac{Pr(MY > 300, G = Aa)}{Pr(G = Aa)} \\ &= \frac{0.16}{0.48} = 0.333 \end{aligned}$$

One way to interpret this conditional probability is as follows: assuming that we have a total of

100 cows, then on average 48 (=0.48 × 100) will be $Aa$ and of those, on average 16 (=0.16 × 100) will give more than 300 kg. Thus, the proportion of $Aa$ cows that will give more than $300kg = 16/48 = 0.333$.
Note that

$$\begin{aligned}
&\sum_{i=1}^{n} Pr(MY = my_i | G = Aa) \\
&= (Pr(MY \leqslant 100, G = Aa) \\
&+ Pr(100 < MY \leqslant 300, G = Aa) \\
&+ Pr(MY > 300, G = Aa))/Pr(G = Aa) \\
&= \frac{0.14 + 0.18 + 0.16}{0.48} \\
&= \frac{0.48}{0.48} = 1
\end{aligned}$$

Focusing on the numerator of the above equation, with $MY \leqslant 100$, $100 < MY \leqslant 300$, and $MY > 300$ being a mutually exclusive and exhaustive set of events:

$$\sum Pr(MY = my_i, G = Aa) = Pr(G = Aa)$$

## Statistical Independence

Random variable $X$ is statistically independent of $Y$ if the probabilities of obtaining different categories of $X$ are the same irrespective of the value of $Y$. That is,

$$\begin{aligned}
Pr(X = x_i | Y = y_j) &= Pr(X = x_i | Y = y_k) \\
&= Pr(X = x_i)
\end{aligned} \tag{4}$$

for all $i$, $j$, and $k$.

In other words, the conditional probabilities are equal to the marginal probabilities. It follows from the definition of conditional probability that if $X$ is statistically independent of $Y$, the joint probability is equal to the product of their marginal probabilities:

$$Pr(X = x_i, Y = y_j) = Pr(X = x_i) \times Pr(Y = y_j) \tag{5}$$

For the example in Table 2, $MY$ and genotype are NOT independent because, e.g.: $Pr(MY > 300 | G = Aa) = 0.333$ is NOT equal to $Pr(MY > 300) = 0.38$. Also, $Pr(MY > 300, G = Aa) = 0.16$ is NOT equal to the product of the marginal probabilities: $Pr(MY > 300)Pr(G = Aa) = 038 \times 0.48 = 0.1824$.

## Conditional Expectation

The expectation (=mean) for variable $X$ conditional on variable $Y$ being equal to $y$ is:

$$E(X|Y = y) = \sum_{i=1}^{k} x_i Pr(X = x_i | Y = y) \tag{6}$$

and, for continuous variables,

$$E(X|Y = y) = \int_x x f(x|Y = y) dx$$

In the example 1, consider the expectation for the number of florets per spikelet, conditional on the fact that the line carries one Ts allele.

$$\begin{aligned}
&Pr(G = Ts/ts | G \text{ contains } Ts) \\
&= Pr(Ts/ts | Ts) = 2/3
\end{aligned}$$

$$Pr(Ts/Ts | Ts) = 1/3$$

$$E(X|Ts) = 3(2/3) + 2(1/3) = 8/3$$

Note that this expectation is slightly lower than the overall E(X). So, if we know that the line carries one Ts allele, we expect the number of florets per spikelet to be slightly lower than average.

# Part 3:

## Covariance

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

$$\begin{aligned}
Cov(X, Y) &= E([X - E(X)][Y - E(Y)]) \\
&= E(XY) - E(X)E(Y)
\end{aligned} \tag{7}$$

Where:

$$E(XY) = \sum_i \sum_j x_i y_j Pr(X = x_i, Y = y_j)$$

### Properties of Covariance

Assuming again that $a$ is a constant.

- Basic rules:

$$\begin{aligned}
&Cov(X, Y) = Cov(Y, X) \\
&Cov(X, aY) = aCov(X, Y) \\
&Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)
\end{aligned}$$

- The variance of a sum is the sum of variance plus twice the covariance

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

- the covariance of a variable with itself is its variance.

$$Cov(X,X) = Var(X)$$

- If X and Y are independent:

$$E(XY) = E(X)E(Y)$$

So that,

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

The covariance between the genotypic value and the phenotypic value will play a big role in all quantitative genetic inferences. Refer back to Table 1, the number of florets per spikelet, conditional on the oat genotype. In Table 1, the genotypic value for the number of florets per spikelet G is considered the same as the phenotypic value for the number of florets per spikelet P. In that case, the covariance between the genotypic and phenotypic values is equal to the variance of the phenotypic values (0.1875, see above). But consider a slightly more complicated situation in which the environment also contributes to determining the phenotype so that (see table 3):

With this environmental effect on the phenotype, the covariance between genetic and phenotypic values is now:

$$\begin{aligned} Cov(G,P) &= E(GP) - E(G)E(P) \\ &= 6.55 - (2.55)^2 = 0.0475. \end{aligned}$$

Check that for this specific example, $Cov(G,P) = Var(G) = 0.0475$. The variance of phenotype is greater: $Var(P) = 0.2475$. The model that relates phenotype to genotype is: $P = G + E$. where the variable E represents the effect of environment. So, for the first row in Table 3 the $E = 3 - 2.8 = +0.2$. For the second row: $E = 2 - 2.8 = -0.8$. Environmental effects are in the last column of Table 3. Note that the $E(E) = 0$. You can also check that, i.e. environmental effects are independent of genetic effects:

$$\begin{aligned} Cov(G,E) &= 0 \\ Cov(P,E) &= 0.2 \\ Var(E) &= 0.2 \end{aligned}$$

## Correlation

$$\begin{aligned} r_{XY} &= Corr(X,Y) \\ &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \end{aligned}$$

Note that $-1 \leqslant r_{XY} \leqslant 1$.

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

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

For the example of Table3:

$$\begin{aligned} r_{GP} &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{0.0475}{\sqrt{0.0475 \times 0.2475}} \\ &= 0.438 \end{aligned}$$

## 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 of Regression

The regression of $Y$ on $X$:

$$\hat{y} = E(Y|X)$$

This is also called the best predictor of $Y$ given $X$. Regression can be used to define a model:

$$y = \hat{y} + e$$

where $e$ is called the residual.

Relative to the "useful identity for variance" given above, we can now say that the variance of $Y$ is the sum of the variance of the regression of $Y$ on $X$ with the expected variance of the residual.

For quantitative variables, we can define the coefficient of regression of $Y$ on $X$, $b_{YX}$,

| Genotype,T | Prob. | G | P | $Pr(T) \times GP$ | E |
|---|---|---|---|---|---|
| ts / ts | 0.20 | 2.8 | 3 | 1.68 | 0.2 |
| ts / ts | 0.05 | 2.8 | 2 | 0.28 | -0.8 |
| Ts / ts | 0.30 | 2.6 | 3 | 2.34 | 0.4 |
| Ts / ts | 0.20 | 2.6 | 2 | 1.04 | -0.6 |
| Ts / Ts | 0.05 | 2.2 | 3 | 0.33 | 0.8 |
| Ts / Ts | 0.20 | 2.2 | 2 | 0.88 | -0.2 |
| | Expectation: | 2.55 | 2.55 | 6.55 | 0 |

Table 3: Table 3

within the context of the following simple linear regression model:

$$\begin{aligned} y &= \bar{y} + b_{YX}(x - \bar{x}) + e \\ \text{with } \bar{y} &= E(Y) \\ b_{YX} &= Cov(Y, X)/Var(X) \end{aligned}$$

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

$$\begin{aligned} b_{YX} &= Cov(Y, X)/Var(X) \\ &= r_{XY}\frac{\sqrt{Var(Y)Var(X)}}{Var(X)} \\ &= r_{XY}\sqrt{\frac{Var(Y)}{Var(X)}} \end{aligned}$$

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

$$\begin{aligned} b_{YX} &= Cov(Y, X)/Var(X) \\ &= r_{XY}\sqrt{\frac{Var(Y)}{Var(X)}} \end{aligned}$$

Note that these only hold for simple regression with a single independent variable (X). So the resulting prediction model is:

$$\hat{y} = \bar{y} + b_{YX}(x - \bar{x}).$$

For the example of Table 3, suppose we want to predict genotype based on phenotype. We would use the following regression model:

$$\begin{aligned} G =& \bar{G} + b_{GP}(P - \bar{P}) + e \\ \text{with } \bar{G} &= E(G) = E(P) = \bar{P} = 2.55 \end{aligned}$$

The regression coefficient can be computed as:

$$\begin{aligned} b_{GP} &= Cov(G, P)/Var(P) \\ &= 0.0475/0.2475 = 0.192 \end{aligned}$$

Or,

$$\begin{aligned} b_{GP} &= r_{GP}\frac{Var(G)}{Var(P)} \\ &= 0.438\frac{0.475}{0.2475} = 0.192 \end{aligned}$$

So the prediction model is: $\hat{G} = \bar{G} + b_{GP}(P - \bar{P}) = 2.55 + 0.192(P - 2.55)$. Results are in Table 4. The last column in this table shows the prediction error: $\hat{e} = G - \hat{G}$.

**Properties of Regression**

$$\begin{aligned} &E(\hat{Y}) = E[E(Y|X)] = E(Y) \\ &E(e) = E(Y - \hat{Y}) = E(Y) - E(\hat{Y}) = 0 \\ &E(e|X) = E(Y - \hat{Y}|X) \\ &= E(Y|X) - E(\hat{Y}|X) = \hat{Y} - \hat{Y} = 0 \\ &Cov(X, e) = E(Xe) - E(X)E(e) \\ &E(Xe) = 0 \\ &Cov(\hat{Y}, e) = 0 \end{aligned}$$

## Accuracy of prediction

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

$$\begin{aligned} r_{\hat{Y}Y} &= \frac{Cov(\hat{y}, y)}{\sqrt{Var(\hat{y})Var(y)}} \\ &= \frac{Cov(\bar{y} + b_{YX}(x - \bar{x}), y)}{\sqrt{Var(\bar{y} + b_{YX}(x - \bar{x}))Var(y)}} \end{aligned}$$

| Genotype,T | Prob. | G | P | E | $\hat{G}$ | $\hat{e}$ |
|---|---|---|---|---|---|---|
| ts / ts | 0.20 | 2.8 | 3 | 0.2 | 2.636 | 0.164 |
| ts / ts | 0.05 | 2.8 | 2 | -0.8 | 2.444 | 0.356 |
| Ts / ts | 0.30 | 2.6 | 3 | 0.4 | 2.636 | -0.036 |
| Ts / ts | 0.20 | 2.6 | 2 | -0.6 | 2.444 | 0.156 |
| Ts / Ts | 0.05 | 2.2 | 3 | 0.8 | 2.636 | -0.436 |
| Ts / Ts | 0.20 | 2.2 | 2 | -0.2 | 2.444 | -0.244 |
| | Expectation: | 2.55 | 2.55 | 0 | 2.55 | 0.0004 |

Table 4: example

Since $\bar{y}$ and $\bar{x}$ are constants, this simplifies to:

$$\begin{aligned}
&= \frac{Cov(b_{YX}x, y)}{\sqrt{Var(b_{YX}x)Var(y)}} \\
&= \frac{b_{YX}Cov(x, y)}{\sqrt{b_{YX}^2 Var(x)Var(y)}} \\
&= \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} \\
&= r_{XY}
\end{aligned}$$

So the accuracy of a prediction equation based on simple regression is equal to the correlation between the dependent and independent variables.

## Decomposition of variance

Using the below 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 $Cov(X, e) = 0$):

$$\begin{aligned}
Var(Y) &= Var(\bar{y} + b_{YX}(X - \bar{x}) + e) \\
&= b_{YX}^2 Var(X) + Var(e) \\
&= \frac{Cov(Y, X)^2}{Var(X)} + Var(e)
\end{aligned}$$

Note that because

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

the first term can also be written as:

$$\frac{Cov(Y, X)^2}{Var(X)} = \frac{r_{XY}^2 Var(X)Var(Y)}{Var(X)} = r_{XY}^2 Var(Y)$$

This is the variance in $Y$ that is explained by the $X$ through the prediction model. By subtraction we get

$$Var(e) = [1 - r_{XY}^2]Var(Y).$$

This is the unexplained/residual variance.

Thus, variance of Y can be decomposed as:

$$Var(Y) = r_{XY}^2 Var(Y) + (1 - r_{XY}^2)Var(Y)$$

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

$$\begin{aligned}
Var(\hat{y}) &= Var(\bar{y} + b_{YX}(x - \bar{x})) \\
&= b_{YX}^2 Var(X) \\
&= \left(\frac{Cov(Y, X)}{Var(X)}\right)^2 Var(X) \\
&= \frac{Cov(Y, X)^2}{Var(X)} \\
&= \frac{Cov(Y, X)^2}{Var(X)Var(Y)} Var(Y) \\
&= r_{YX}^2 Var(Y)
\end{aligned}$$

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

Considering the example of table 3, the genotypic value G is the expectation of the phenotype P conditional on the genotype T. That is, $G = E(P|T)$. In consequence $E(G) = E(P)$ from property above. You can verify that we used this definition in Table 3 though we had not yet defined genotypic value.

# Useful distributions in Population 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:

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

Its variance is:

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

## 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: $X \sim Binomial(k, p)$, then: From the properties of expectation the *expected value* of X is $kp$: $E(X) = kp$. From the properties of variance the variance of X is $kp(1-p)$: $var(X) = kp(1-p)$. The probability mass function Pr(X = x) is:

$$Pr(X = x) = \binom{k}{s} p^x (1-p)^{k-x}$$

$$\binom{k}{x} = \frac{k!}{x!(k-x)!}$$

where,

$$a! = a \times (a-1)...3 \times 2 \times 1$$