QCB 408 / 508 – Notes on Week 2

Scott Wolf

2020-02-19

Summary

In Week 2 of Foundations of Statistical Genomics, we covered the basic properties of joint, marginal, and conditional distributions and applied them to the example of Hardy-Weinberg equilibrium to see how we can formally model a well known biological model.

- Joint, Marginal, and Conditional Distributions and Their Properties
- Hardy-Weinberg Equilibrium and Non-random Mating

Joint, Marginal, and Conditional Distributions

We know that one can define multiple random variables on the probability space from an experiment. Given this, we must ask how we can explore event spaces spanning multiple random variables – that is how do we *jointly* specify probabilities involving multiple random variables? To probe this, let's first explore the resulting CDFs.

Let X and Y define random variables on the same probability space then the joint CDF can be written as

$$\begin{aligned} F_{X,Y}(x,y) &= \Pr((X \leq x, Y \leq y) \\ &= \Pr(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq y\}) \end{aligned}$$

This directly extends to n random variables $X_1, X_2, \ldots, X_n$. The joint CDF, $F_{X_1,X_2,\ldots,X_n}$ is

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

.

PMFs and PDFs

To deal with pmfs and pdfs for multiple random variables, we can directly extend our single random variable definitions. Let's explore, as before, with the bivariate cases first.

Discrete Case

The joint probability mass function of two discrete random variables, X,Y is given as

$$\begin{aligned} f_{X,Y}(x,y) &= \Pr(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}) \\ &= \Pr(X = x, Y = y) \end{aligned}$$

Which can be generalized to n discrete random variables:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$$

Continuous Case

For the continuous bivariate case we can extend our pdf definition for single distributions to get:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

This definition, just as those above, extends naturally to n variables

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)}{\partial x_1, x_2, \ldots, x_n}$$

Events in Multivariate Spaces

Following directly from the above definitions we can define the probability of an event $A_{x_1} \times \cdots \times A_{x_n} \subseteq \mathbb{R}^n$ as the following generalized formula:

$$\int_{x_1 \in A_{x_1}} \cdots \int_{x_n \in A_{x_n}} dF_{X_n}(x_n) \cdots dF_{X_1}(x_1)$$

Marginal Distributions

In the case of multiple random variables defined on the same probability space, it is natural to ask how we can decompose the joint distribution into individual distributions. This individual distribution within a larger joint space is referred to as the marginal distribution. That is, the marginal distribution gives the distribution of a single variable without reference to other variables. Consider the below probability space where the bottom row and rightmost column represent the marginal distributions for X and Y respectively.

	x_1	x_2	x_3	$\mathbf{f_Y(y)}$
y_1	0.2	0.075	0.15	**0.425**
y_2	0.0375	0.15	0.0375	**0.225**
y_3	0.15	0	0.2	**0.35**
$\mathbf{f_X(x)}$	**0.3875**	**0.225**	**0.3875**	**1**

Again using a bivariate distribution with random variables X and Y, the marginal distribution of some random variable, say X, comes from the Law of Total Probability and is given by

$$f(x) = \sum_{y \in \mathcal{R}_y} f(x,y) \qquad \text{(discrete)}$$

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad \text{(continuous)}$$

Independent Random Variables

If two random variables, say X and Y, are independent, then the outcome of one has no effect on the outcome of the other. In our formal notation, that is:

$$f_{X,Y}(a,b) = f_X(a)f_Y(b)$$

and then factoring of the CDF follows

$$\begin{aligned} F_{X,Y}(a,b) &= \Pr(X \leq a, Y \leq b) \\ &= \Pr(X \leq a) \Pr(Y \leq b) \\ &= F_X(a) F_Y(b) \end{aligned}$$

Conditional Distributions

One of the most important ways for us to deal with multiple random variables defined on the same probability space is to examine how each variable change within subpopulations. Conditional distributions allow us to examine this variation directly.

Note: A good resource for visualizing this can be found at http://setosa.io/conditional/.

To dive into this, consider two random variables X and Y. Our conditional random variable $X|Y \sim F_{X|Y}$ comes from the pmf/pdf given by

$$f(x|y) = \frac{f(x,y)}{f(y)}.$$

Then we have

$$\Pr(X \leq a | Y \leq b) = \frac{\Pr(X \leq a, Y \leq b)}{\Pr(Y \leq b)}$$

which is just

$$F_{X|Y}(a|Y \leq b) = \frac{F_{X,Y}(a,b)}{F_Y(b)}$$

.

Bayes theorem applies to the distributions just as it does to probabilities and gives us:

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

and all of the above notes extend to n random variables.

Moments and Moment Generating Functions for Joint Distributions

The kth conditional moment of a random variable X is

$$\begin{aligned} \mathrm{E}[X^k | Y = y] &= \sum_{x \in R} x^k f(x|y) \\ \mathrm{E}[X^k | Y = y] &= \int_{-\infty}^{\infty} x^k f(x|y) dx \end{aligned}$$

and our variance is

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \end{aligned}$$

and our covariance, which describes how much the two random variables vary together, is given by

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

which is just modification of the variance formula, $\mathrm{Var}(X) = \mathrm{E}[(X - E[X])^2]$.

To get a visual, take look at a figure from https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean.

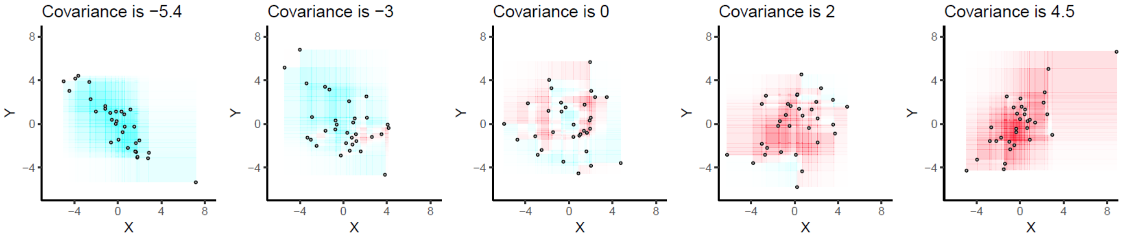

Covariance is −5.4
Covariance is −3
Covariance is 0
Covariance is 2
Covariance is 4.5

Figure 1: Figure Visualizing Covariance from Stack Exchange(User: whuber)

We can also use more general definition which is that covariance, σ_{XY}, is the expected value of a function of X and Y over the space and can be given by

$$\sigma_{XY} = \int\int (x - \mathrm{E}[X])(y - \mathrm{E}[Y])f(x,y)dxdy$$
$$\sum_x\sum_y (x - \mathrm{E}[X])(y - \mathrm{E}[Y])f(x,y)$$

Linear Transforms on Random Variables

Let X be a random variable then we have:

$$\mathrm{E}\left[a + bX\right] = a + b\,\mathrm{E}[X]$$

$$\begin{aligned} E\left[a + bX\right] &= a + b\,\mathrm{E}[X] \\ &= \int (a + bx)f(x)dx \\ &= a\int f(x)dx + b\int xf(x)dxa + b\,\mathrm{E}[X] \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left(a + bX\right) &= \mathrm{E}[(bX + a - b\,\mathrm{E}[X] - a)^2] \\ &= \mathrm{E}[b^2(X - \mathrm{E}[X])^2 \\ &= b^2\,\mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= b^2\,\mathrm{Var}(X) \end{aligned}$$

Law of Total Variance

The law of total variance, also known as the variance decomposition formula, tells us that if we have two random variables, X and Y, on the same probability space then the variance can be decomposed as follows:

$$\mathrm{Var}(X) = \mathrm{Var}(\mathrm{E}[X|Y]) + \mathrm{E}[\mathrm{Var}(X|Y)]$$

This gives us the proportion of variance explained in the second component and the remaining in the first term.

Simple Discrete Joint Probability Table Example using R

Note that this example pulls partially from: https://rstudio-pubs-static.s3.amazonaws.com/209289_9f9ba331cccc4e8f8aabdb9273cc76af.html

For a simplistic example of dealing with discrete joint probability tables in R, see the sample code below.

Consider the probability table below where the columns represent probabilities for three states of Y(say 0,1,2) and the rows represent states(0,1,2 again) from X. Then we have

```
> p <- matrix(c(.1,.04,.06,.06,.15,.1,.12,.22,.15), ncol=3, nrow =3, byrow = T)
> p
     [,1] [,2] [,3]
[1,] 0.10 0.04 0.06
[2,] 0.06 0.15 0.10
[3,] 0.12 0.22 0.15
```

Let's calculate marginal distributions from our probability table!

```
> p_x <- apply(p,1,sum) # for x
> p_x
[1] 0.20 0.31 0.49
>
> p_y <- apply(p,2,sum) # for y
> p_y
[1] 0.28 0.41 0.31
```

We can also directly calculate conditional probabilities from the table! Say we want to find the probability of x_3 given y_2:

```
> p[3,2]/p_y[2]
[1] 0.5365854
```

Now that we've done that, it's obvious that we could continue to calculate the expected value and sum by just assigning values to each state and using the formulas we've learned! Of course this example is a bit simpler than what we would normally run across in practice, but it gives a good example how we would actually deal with a table.

Hardy-Weinberg Equilibrium

The Hardy-Weinberg Theorem gives a framework for understanding Mendelian genetics in the frame of diploid organisms that reproduce sexually. Given the five following assumptions, the theorem tells us that allele frequencies will be in a completely static state, known as Hardy-Weinberg Equilibrium, from generation to generation and that if the allele frequencies are given as p and (1-p) then the expected genotype frequencies are p^2, $2p(1-p)$, and $(1-p)^2$.

1. No non-random mating with respect to the locus in question

2. Infinite population size(that is, no genetic drift!)
3. No migration moves genes in or out of the population
4. No mutation occurs at the locus in question
5. Natural selection is not acting on the locus

For example, considering the the case of some allele with two states A and a. Then let the frequency of A be p and a be (1-p) then the Hardy-Weinberg principle tells us that if the above assumptions are met the frequency of AA is p^2, the frequency of Aa is $2p(1-p)$, and the frequency of aa is $(1-p)^2$. Many other properties of this have been shown including the expansion to more than two alleles and generalization to polyploidy among others.

Now that we have an idea of Hardy-Weinberg Equlibrium in general, let's dive into a proper statistical description of the phenomenon.

Take some random variable describing the state of some SNP in terms of the count of alternative alleles in diploid organism: $X \in 0, 1, 2$. We should also note that p will represent the frequency of the major allele in the following notes.

Then we get $\Pr(X = k) = r_k$ with $k = 0, 1, 2$, and following directly we get $\mathrm{E}[X] = r_1 + 2r_2$ and $\mathrm{E}[\frac{X}{2}] = \frac{r_1}{2} + r_2 = p$.

Then let $Y \in \{0, 1\}$ be the transmitted allele coming from an individual.

$$\Pr(Y = 1|X = k) = \begin{cases} 0 & \text{if k = 0} \\ \frac{1}{2} & \text{if k = 1} \\ 1 & \text{if k = 2} \end{cases} \tag{1}$$

Which across all genotypes gives

$$\begin{aligned} \Pr(Y = 1) &= \sum_{k=0}^{2} \Pr(Y = 1|X = k) \Pr(X = k) \\ &= 0 \cdot r_0 + \frac{1}{2} r_1 + r_2 \\ &= p \end{aligned}$$

Which is simply to say that $Y \sim Bernoulli(p)$.

Consider then Z to describe the following generation genotype which is, of course, the sum of Y_1 and Y_2 which are both described by Bernoulli(p). As mentioned in the first week of class, the sum of Bernoulli's is simply a binomial which gives us

$$Z \sim Binomial(2, p)$$

Following from the above, we have $\mathrm{E}[Z] = 2p = \mathrm{E}[X]$ which tells us that the transmission probability for the following generation is identical to the prior – they are in equilibrium. We can also see that the three genotype frequencies given by the Hardy-Weinberg Theorem come direction from the distribution Binomial(2,p).

Inbreeding

Now that we have dealt with the simplest case of HWE, let's consider what happens when inbreeding occurs. Since inbreeders are related, they are more likely to share alleles than otherwise. This causes an excess of homozygous offspring and a decrease in heterozygosity in the population.

Firstly, we must note that two alleles are said to be identical by descent (IBD) if they are copies from a common ancestor without recombination. Take I to be a random variable indicating whether or not two randomly drawn alleles are IBD. Also Y_1, Y_2, and Y_3 are copies from unrelated ancestors.

For Y_1 and Y_2,

$$\begin{aligned} I = 0 &\implies Y_1, Y_2 \overset{\text{iid}}{\sim} \textit{Bernoulli}(p) \\ I = 1 &\implies Y_1 = Y_2 \sim \textit{Bernoulli}(p) \end{aligned}$$

We also have

$$I \sim \textit{Bernoulli}(f)$$

where f is the inbreeding coefficient.

Then

$$(Z|I = 0) \sim \textit{Binomial}(2, p)$$

$$(Z|I = 1) \sim \textit{Bernoulli}(p)$$

with

$$Y_1, Y_2, Y_3 \overset{\text{iid}}{\sim} \textit{Bernoulli}(p)$$

Quickly this gives us $Z = (Y_1 + Y_2)(1 - I) + 2Y_3 I$.

Assuming there is no inbreeding, we have I = 0 and

$$\Pr(Z = k|I = 0) = \begin{cases} (1-p)^2 & \text{k} = 0 \\ 2p(1-p) & \text{k} = 1 \\ p^2 & \text{k} = 2 \end{cases} \tag{2}$$

and for $I = 1$

$$\Pr(Z = k|I = 0) = \begin{cases} (1-p) & \text{k} = 0 \\ 0 & \text{k} = 1 \\ p & \text{k} = 2 \end{cases} \tag{3}$$

That is, $(Z|I = 1) \sim 2\textit{Bernoulli}(p)$.

To see the actual probabilies of Z, we have the following from the Law of Total Probability.

$$\begin{aligned} \Pr(Z = 0) &= \Pr(Z = 0|I = 0)\Pr(I = 0) + \Pr(Z = 0|I = 1)\Pr(I = 1) \\ &= (1-p)^2)(1-f) + (1-p)f \\ &= (1-p)^2 + p(1-p)f \end{aligned}$$

$$\begin{aligned} \Pr(Z = 1) &= 2p(1-p)(1-f) \\ &= 2p(1-p)(1-f) \end{aligned}$$

which is less than 2p(1-p). That is to say, inbreeding reduces heterozygosity – just as we would suspect!

$$\begin{aligned}\Pr(z=2) &= p^2(1-f)+pf \\ &= p^2+p(1-p)f\end{aligned}$$

Then our expected value comes directly from the above: $\mathrm{E}[Z] = 1 \cdot 2p(1-p)(1-f) + 2 \cdot (p^2 + p(1-p)f)$.

For examining the variance, we can use the Law of Total Variance to get

$$\mathrm{Var}(Z) = \mathrm{E}[\mathrm{Var}(Z|I)] + \mathrm{Var}(\mathrm{E}[Z|I])$$

The conditioned variances come to

$$\mathrm{Var}(Z|I=0) = 2p(1-p)$$

$$\begin{aligned}\mathrm{Var}(Z|I=1) &= \mathrm{Var}(2Y_3) \\ &= 4\,\mathrm{Var}(Y_3) \\ &= 4p(1-p)\end{aligned}$$

$$\mathrm{Var}(Z|I) = 2p(1-p)(1-I) + 4p(1-p)I$$

and then using the expected Value of I, we have

$$\begin{aligned}\mathrm{E}[\mathrm{Var}(Z|I)] &= 2p(1-p)(1-f) + 4p(1-p)f \\ &= 2p(1-p)(1+f)\end{aligned}$$

We can also see that

$$Z|I=0 = Y_1 + Y_2$$

and

$$Z|I=1 = 2Y_3$$

Given that they each have equal probabilities p, we have

$$E[Z|I] = 2p$$

no matter the status of I. That is, inbreeding has no effect on the mean and our variance in the expected value is 0. That is, $\mathrm{Var}(\mathrm{E}[Z|I]) = 0$.

Now that we have dealt with that, let's consider the variance of Z again. We have shown that $\mathrm{E}[\mathrm{Var}(Z|I)] = 2p(1-p)(1+f)$ and $\mathrm{Var}(\mathrm{E}[Z|I]) = 0$. Using the Law of Total Variance, we have our final result:

$$\mathrm{Var}(Z) = 2p(1-p)(1+f)$$

.

Modeling Drift as a Random Variable

Another way to reach the same conclusion found above is to take Q as a continuous distribution describing the frequencies within a population that is divided by some fixation index, f. Then we get Then also $Z|Q \sim Binomial(2, Q)$ and $Z|Q = q \sim Binomial(2, q)$. Note that p is the ancestral allele frequency, and then just as before, we want to find $\mathrm{E}[Z]$ and $\mathrm{Var}(Z)$.

We know that we can model the distribution of Q using the Balding-Nichols model, a reparametarization of the Beta distribution given by

$$BN(p, f) = Beta(\frac{1-f}{f}p, \frac{1-f}{f}(1-p))$$

We should also note that the mean of BN(p,f) is p and the variance is fp(1 − p).

Then directly we get $\mathrm{E}[Z] = \mathrm{E}[\mathrm{E}[Z|Q]] = \mathrm{E}[2Q] = 2p$.

Just as in class, let's start with $\Pr(Z = 2)$. For the third step of this, we need to note again that $\mathrm{Var}(Q) = \mathrm{E}[Q^2] - \mathrm{E}[Q]^2$.

$$\begin{aligned}
\Pr(Z = 2) &= \int \Pr(Z = 2|Q = q)f(q)dq \\
&= \int q^2 f(q)dq \\
&= E[Q^2] \\
&= Var(Q) + E[Q]^2 \\
&= p(1-p)f + p^2
\end{aligned}$$

Next, let's look at $\Pr(Z = 1)$

$$\begin{aligned}
\Pr(Z = 1) &= \int \Pr(Z = 1|Q = q)f(q)dq \\
&= \int 2q(1-q)f(q)dq \\
&= 2\int qf(q)dq - 2\int q^2 f(q)dq \\
&= 2\,\mathrm{E}[Q] - 2\,\mathrm{E}[Q^2] \\
&= 2p - 2(p(1-p)f + p^2) \\
&= 2p(1-p)(1-f)
\end{aligned}$$

and $\Pr(Z = 0)$

$$
\begin{aligned}
\Pr(Z=0) &= \int \Pr(Z=0|Q=q)f(q)dq \\
&= \int (1-q)^2 f(q)dq \\
&= \int (1-2q+q^2)f(q)dq \\
&= \int f(q)dq - 2\int qf(q)dq + \int q^2 f(q)dq \\
&= 1 - 2\,\mathrm{E}[Q] + \mathrm{E}[Q^2] \\
&= p(1-p)f + (1-p)^2
\end{aligned}
$$

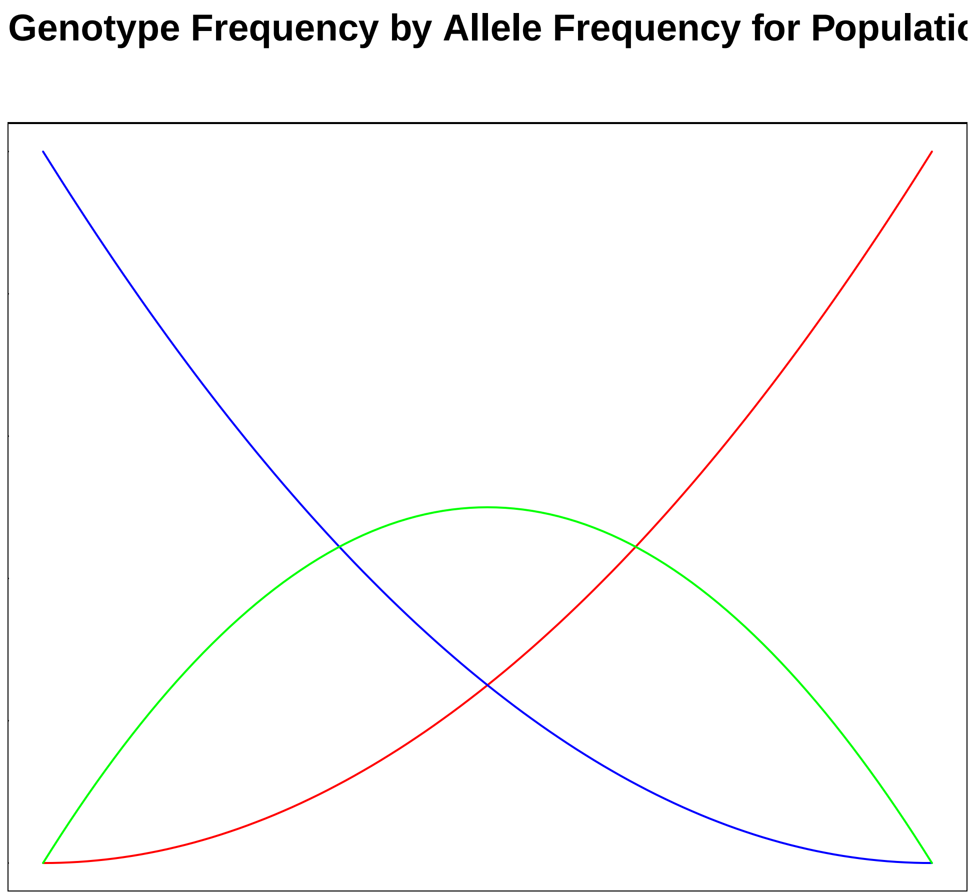
Now that we've dealt with all of that, let's move on to $\mathrm{Var}(Z)$!

$$
\begin{aligned}
\mathrm{Var}(Z) &= \mathrm{E}[\mathrm{Var}(Z|Q)] + \mathrm{Var}(E[Z|Q]) \\
&= \mathrm{E}[2Q(1-Q)] + \mathrm{Var}(2Q) \\
&= E[2Q] - E[2Q^2] + 4p(1-p)f \\
&= 2p - 2E[Q^2] + 4p(1-p)f \\
&= 2p - 2[\mathrm{Var}(Q) + \mathrm{E}[Q]^2] + 4p(1-p)f \\
&= 2p - 2p(1-p)f - 2p^2 + 4p(1-p)f \\
&= 2p(1-p) + 2p(1-p)f \\
&= 2p[p(1-p) + p(1-p)f] \\
&= 2p(1-p)(1+f)
\end{aligned}
$$

HWE in R

To get a better idea of how this works in practice, let's generate some data as if it came from a population in Hardy-Weinberg equilibrium (HWE). Note that this example pulls from the HardyWeinberg from Graffelman et al. The package provides a great framework for exploring HWE.

```
> #Plotting equations
> par(mar = c(0, 0, 5, 0))
> curve(x ^ 2,
+       col = 'red',
+       main = "Expected Genotype Frequency by Allele Frequency for Population in HWE",
+       sub = " (red) AA; (green) Aa; (blue) aa")
> curve((1 - x) ^ 2, col = 'blue', add = T)
> curve(2 * x * (1 - x), col = 'green', add = T)
```

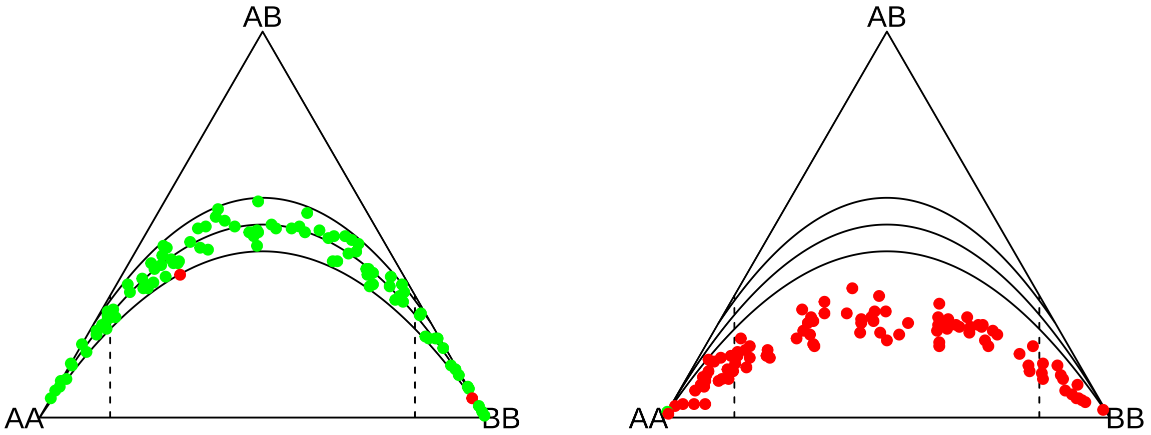


```
>
> #As with 100 populations(written as nm number markers in this case) and sample size 200
> hwe <- HWData(nm = 100,n=200)
> hwe_inbred <- HWData(nm = 100,n=200,f=.5)
```

Now let's make some plots to see how we can visualize HWE.

```
> par(mfrow = c(1, 2))
> par(mar = c(0, 2, 0, 2))
> HWTernaryPlot(hwe)
> HWTernaryPlot(hwe_inbred)
> mtext(
+   "(left) Ternary Plot of Genotypes in HWE (right) Ternary Plot of Genotypes with F_ST = 0.5",
+   line = -2,
+   outer = TRUE,
+   cex = 1
+ )
```

(left) Ternary Plot of Genotypes in HWE (right) Ternary Plot of Genotypes with F_ST = 0.5



Note that in the same way this is done, we can test populations for HWE and see how different levels of inbreeding(and other features!) affect genotype frequencies.

Session Information

```
> sessionInfo()
R version 3.6.1 (2019-07-05)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] HardyWeinberg_1.6.3 Rsolnp_1.16         mice_3.7.0
 [4] lattice_0.20-38     gapminder_0.3.0     forcats_0.4.0
 [7] stringr_1.4.0       dplyr_0.8.4         purrr_0.3.3
[10] readr_1.3.1         tidyr_1.0.2         tibble_2.1.3
[13] ggplot2_3.2.1       tidyverse_1.3.0     knitr_1.28

loaded via a namespace (and not attached):
 [1] mitml_0.3-7      Rcpp_1.0.3       lubridate_1.7.4  assertthat_0.2.1
 [5] digest_0.6.23    truncnorm_1.0-8  pan_1.6          R6_2.4.1
 [9] cellranger_1.1.0 backports_1.1.5  jomo_2.6-10      reprex_0.3.0
[13] evaluate_0.14    httr_1.4.1       pillar_1.4.3     rlang_0.4.4
[17] lazyeval_0.2.2   readxl_1.3.1     rstudioapi_0.11  minqa_1.2.4
[21] nloptr_1.2.1     rpart_4.1-15     Matrix_1.2-18    rmarkdown_2.1
[25] splines_3.6.1    lme4_1.1-21      munsell_0.5.0    broom_0.5.4
[29] compiler_3.6.1   modelr_0.1.5     xfun_0.12        pkgconfig_2.0.3
[33] htmltools_0.4.0  nnet_7.3-12      tidyselect_1.0.0 fansi_0.4.1
[37] crayon_1.3.4     dbplyr_1.4.2     withr_2.1.2      MASS_7.3-51.5
[41] grid_3.6.1       nlme_3.1-143     jsonlite_1.6.1   gtable_0.3.0
[45] lifecycle_0.1.0  DBI_1.1.0        magrittr_1.5     scales_1.1.0
[49] cli_2.0.1        stringi_1.4.5    fs_1.3.1         xml2_1.2.2
[53] generics_0.0.2   vctrs_0.2.2      boot_1.3-24      tools_3.6.1
[57] glue_1.3.1       hms_0.5.3        parallel_3.6.1   survival_3.1-8
[61] yaml_2.2.1       colorspace_1.4-1 rvest_0.3.5      haven_2.2.0
```