QCB 408 / 508 – Notes on Week 2

Avinash Boppana

2020-03-01

Summary

The topics for QCB 408/508 Week 2 begin with a review of important distributions and random variables, such as the Normal, Multivariate, and Conditional distributions, as well as independent random variables. Proximately, generalized properties for statistical distributions and random variables are covered, detailing Bayes Theorem, Moments, and Linear Transformations. Next, a thorough, worked-out proof is provided to supplement the Law of Total Variance. The latter half of the notes discuss Hardy-Weinberg Equilibrium (HWE), including random variable models for both HWE and violations of HWE (inbreeding and drift). Proofs and simulations are provided to enhance one's understanding of the topics discussed.

- Random Variables
- Hardy-Weinberg Equilibrium

Random Variables

[1] Distributions

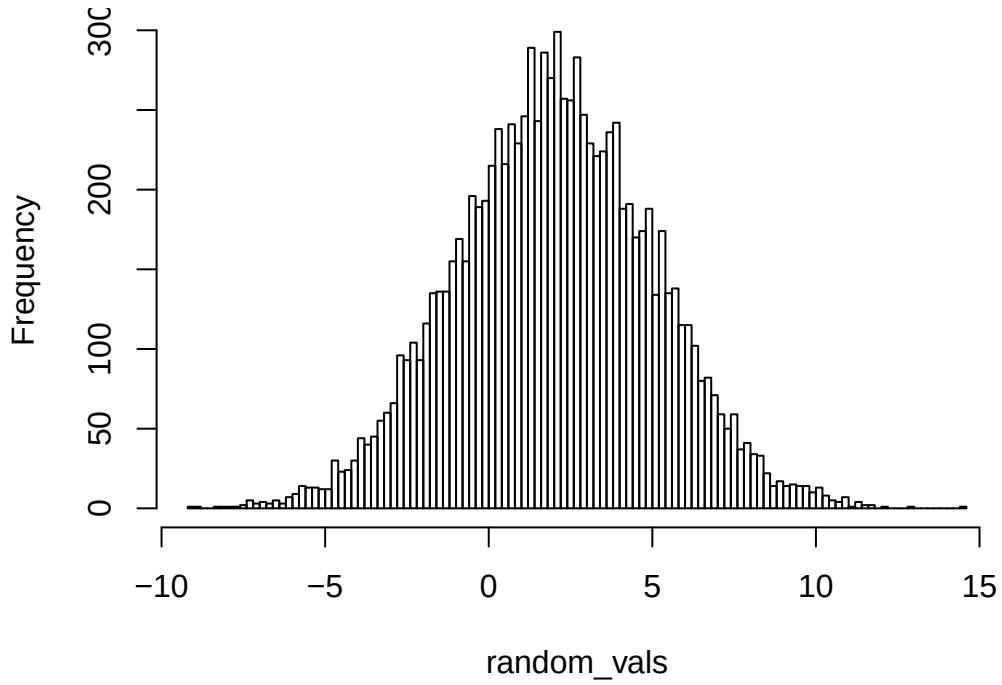
Normal (Gaussian) Distribution:

- **Parameters:** $X \sim Normal(\mu, \sigma^2)$
- **Range:** $R = (-\infty, \infty)$
- **Function:** $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- **Expected Value and Variance:** $E[X] = \mu, Var(X) = \sigma^2 > 0$
- **R-Example:** $dnorm(x, mean = 2, s.d. = 3)$, where $s.d. = \sqrt{\sigma^2}$

```
> dnorm(3.3, 2, 3)
[1] 0.1210635
```

- Below is a simulation of the Normal Distribution

```
> random_vals = rnorm(10000, mean= 2, sd = 3)
> hist(random_vals, n=100, main="")
```



- The bell-shaped distribution of the simulation is indicative of a Normal distribution

Multivariate rv's:

- Say we have two rvs X and Y
- The joint cdf is $F(a,b) = Pr(X \leq a, Y \leq B) = Pr(\{w : X(w) \leq a\} \cap \{w : Y(w) \leq b\})$
- **If the rvs are continuous:** $\frac{\partial^2}{\partial x \partial y} F(x,y) = f(x,y)$ or the joint pdf
- **If the rv's are discrete:** joint pmf $f(x,y) = Pr(X = x, Y = y)$
- **Marginal Distribution (discrete):** $f(x) = \sum_{y \in R(y)} f(x,y)$
- **Marginal Distribution (continuous):** $f(x) = \int f(x,y)dy$

Independent rv's:

- If X and Y are independent, then: $f(x,y) = f(x)f(y)$
- $F_{X,Y}(a,b) = Pr(X \leq a, Y \leq b) = F_X(a) * F_Y(b) = Pr(X \leq a) * Pr(X \leq b)$

Conditional Distributions:

- $Pr(X \leq a | Y \leq b) = \frac{Pr(X \leq a, Y \leq b)}{Pr(Y \leq b)}$
- Keep in mind that: $X \leq a \rightarrow \{w : X(w \leq a\}$ and $Y \leq b \rightarrow \{w : Y(w \leq b\}$
- $F_{X|Y}(a|Y \leq b) = \frac{F_{X,Y}(a,b)}{F_y(b)}$, where $F_{X|Y}$ is the conditional cdf
- $f(x|y) = \frac{f(x,y)}{f(y)}$

[2] Properties of Distributions

Bayes Theorem: $f(y|x) = \frac{f(x|y)f(y)}{f(x)}$

- All of the above extends to more than just two random variables as well
- **Ex (if discrete):** $f(x_1, x_2 | x_3, x_4, x_5) = \frac{f(x_1, x_2, x_3, x_4, x_5)}{f(x_3, x_4, x_5)}$
- **Ex (if continuous):** $f(x_3, x_4, x_5 | x_1, x_2) = \int \int f(x_1, x_2, x_3, x_4, x_5) dx_1 dx_2$
- If $X_1, X_2, ...X_3$ are independent, then

- $f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i)$

Moments of Joint Distribution:

- For a single rv X: $E[X^k]$ is the kth moment
- $Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$
- $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$
- **Note:** A positive covariance means X,Y are either both above or both below their means, while a negative covariance means the random variables are on opposite sides of their means
- **Calculating Covariance after finding marginal expected values** $E[X], E[Y]$
- **Continuous:** $\int\int (x - E[X])(y - E[Y])f(x, y)dxdy$
- **Discrete:** $\sum_x \sum_y (x - E[X])(y - E[Y])f(x, y)$

Linear Transformations of rv's:

- X is a rv, and a and b are constants:

1. $E[a + bX] = a + bE[X]$
2. $Var(a + bX) = b^2 Var(X)$

- **Proof for 1:**

$$E[a + bX] = \int (a + bx)f(x)dx = \int af(x)dx + \int bxf(x)dx$$

$$= a\int f(x)dx + b\int xf(x)dx = aE[1] + bE[X] = a + bE[X]$$

- **Worked out Proof for 2**:

$Var(a+bX) = E\left[(a + bX) - (E[a + bX])^2\right] = E[(a+bX)^2] - (E[a+bX])^2$ [Through equivalent variance relation]

$$= \int (a + bx)^2 f(x)dx - \left(\int (a + bx)f(x)dx\right)\left(\int (a + bx)f(x)dx\right)$$

$$= \int (a^2 + 2abx + b^2x^2)f(x)dx - (a + bE[X])^2$$

$$= \int a^2 f(x)dx + \int 2abxf(x)dx + \int b^2x^2 f(x)dx - \left(a^2 + 2abE[X] + b^2(E[X])^2\right)$$

$$= E[a^2] + E[2abX] + E[b^2X^2] - a^2 - 2abE[X] - b^2(E[X])^2$$

$$= a^2 + 2abE[X] + b^2E[X^2] - a^2 - 2abE[X] - b^2(E[X])^2 = b^2E[X^2] - b^2(E[X])^2$$

$$= b^2\left(E[X^2] - (E[X])^2\right) = b^2 Var(X)$$

Law of Total Variance (for jointly distributed rvs X and Y):

- $Var(X) = Var(E[X|Y]) + E[Var(X|Y)]$
- The terms represent the variance explained and not explained by the model, respectively
- **All of model-fitting can be understood through the law of total variance**
- $E[X|Y]$ essentially means that X is a **function of Y**
- $E[X|Y = y] = \int xf(x|y)dx$,
- **Note:** The integral contains "little x" b/c we are considering all instances of r.v. X
- The following exercise will be a necessary step in proving the Law of Total Variance

$$E[E[X|Y]] \stackrel{?}{=} E[X]$$

$$E[E[X|Y]] = \int\int xf(x|y)dxf(y)d(y)$$

$$= \int\int xf(x|y)f(y)dxdy = \int\int xf(x,y)dxdy = \int\int xf(x,y)dydx$$

$$= \int x \int f(x,y)dydx = \int xf(x)dx, \text{[By simplifying marginal dist. of x]}$$

$$= E[X]$$

Worked Out Proof for Law of Total Variance

$$Var(X|Y) = E\left[(X|Y - E[X|Y])^2\right] \qquad \text{[By definition of variance]}$$

$$= E[X^2|Y] - (E[X|Y])^2 \qquad \text{[Common equivalent form of variance]}$$

$$E[Var(X|Y)] = E\left[E[X^2|Y] - (E[X|Y])^2\right] \qquad \text{[Taking expected value of both sides]}$$

$$= E\left[E[X^2|Y]\right] - E\left[(E[X|Y])^2\right] = E[X^2] - E\left[(E[X|Y])^2\right] \qquad \text{[Using the exercise from above]}$$

$$= E[X^2] - E\left[(E[X|Y])^2\right] + (E[X])^2 - (E[X])^2 = \left(E[X^2] - (E[X])^2\right) - \left(E\left[(E[X|Y])^2\right] - (E[X])^2\right)$$

$$= \left(E[X^2] - (E[X])^2\right) - \left(E\left[(E[X|Y])^2\right] - (E\left[E[X|Y]\right])^2\right) \qquad \text{[Using exercise from above, opposite direction]}$$

$$= Var(X) - Var\left(E[X|Y]\right) \qquad \text{[Using the same equivlence for variance, opposite direction]}$$

$$E[Var(X|Y)] = Var(X) - Var\left(E[X|Y]\right)$$

$$Var(X) = E[Var(X|Y)] + Var\left(E[X|Y]\right)$$

Hardy-Weinberg Equilibrium (HWE)

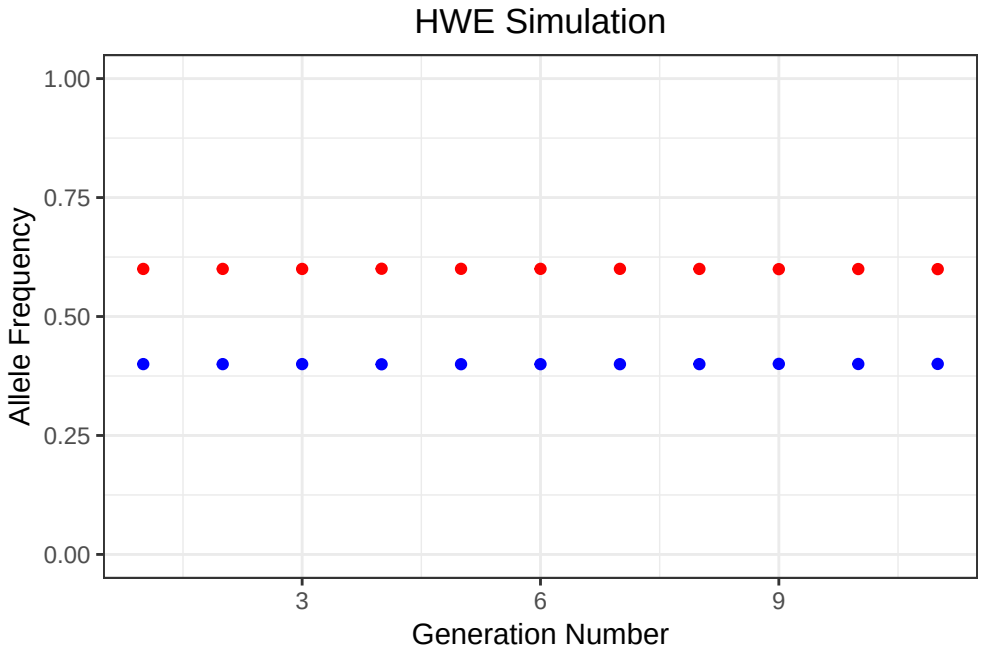
- This is a population system with infinite size, random mating, no drift, no mutations, etc.
- The following is a rv-modeling approach to representing HWE
- **Note:** An inifinite population size ensures that we can work w/ probability distributions rather than keeping track of individuals

[1] Modeling Hardy-Weinberg Equilibrium

- Let the distribution of SNPs at a particular locus $\in \{CC, CT, TT\}$
- Let the genotype rv X represent the # of T alleles $\rightarrow X \in \{0, 1, 2\}$
- $Pr(X = k) = r_k, k = 0, 1, 2$
- $E[X] = 0 * r_0 + 1 * r_1 + 2 * r_2 = r_1 + 2r_2$
- $E\left[\frac{X}{2}\right] = \frac{r_1}{2} + r_2 = p$,where p is the freq. of T at the specific site
- Let the transmitted allele $Y \in \{0, 1\}$, then $Pr(Y = 1|X = k) = \begin{cases} 0, k = 0 \\ \frac{1}{2}, k = 1 \\ 1, k = 2 \end{cases}$
- $Pr(Y = 1) = \sum_{k=0}^{2} Pr(Y = 1|X = k)Pr(X = k) = 0 * r_0 + \frac{1}{2} + r_2 = p$
- Given that $Pr(Y = 1) = 1$, it follows that $Pr(Y = 0) = 1 - p$, hence Y can be modeled as $Y \sim Bernoulli(p)$
- **We can extend this model to future generations:** Let rv $Z \in \{0, 1, 2\}$ represent the genotype of the next generation
- **Note:** Assuming random selection of an individual from the next generation allows us to model Z with a random variable
- $Y_1, Y_2 \overset{iid}{\sim} Bernoulli(p)$
- $Z = Y_1 + Y_2$, and since Z is the sum of Bernoulli rvs, it can be modeled as $Z \sim Binomial(2, p)$
- $E[Z] = 2p = E[X]$, showing that the transmission probabilities of genotypes for Z are identical to X
- **These indentical transmission probabilities are representative of an equilibrium!**
- All genotypes in future generations are drawn from $Z \sim Binomial(2, p)$
- $\begin{cases} Pr(Z = 0) = (1 - p)^2 \\ Pr(Z = 1) = 2p(1 - p) \\ Pr(Z = 2) = p^2 \end{cases}$
- Below is a simulation of Hardy-Weinberg Equilibrium:

```
>
> #Defining CC = 0, CT= 1, TT = 2
>
> #Initial Population Size
> pop_size = 20000
>
> #Initial Allele Frequencies
> c_freq = 0.4
> t_freq = 0.6
>
> #Initial Allele Counts
> c_count = pop_size * 2 * c_freq
> t_count = pop_size * 2 * t_freq
>
> #Number of generations to simulate
> n_gen = 10
>
```

```
> c_freq_history = c(c_freq)
> t_freq_history = c(t_freq)
>
> for(i in 1:n_gen)
+ {
+   #Randomly selects number of new offspring from previous generations mating
+   new_births = sample((0*pop_size):(1*pop_size), 1)
+
+   #KEY: Transmission of genotypes to offspring is ~Binom(2, t_freq)
+   new_genotypes = rbinom(n = new_births, size = 2, prob = t_freq)
+
+   #Updating Counts and Frequencies
+   new_c_count = 2 * table(new_genotypes)['0'] + 1 * table(new_genotypes)['1']
+   new_t_count = 1 * table(new_genotypes)['1'] + 2 * table(new_genotypes)['2']
+   c_count = c_count + new_c_count
+   t_count = t_count + new_t_count
+
+   pop_size = pop_size + new_births
+
+   new_c_freq = c_count / (2 * pop_size)
+   new_t_freq = t_count / (2 * pop_size)
+   c_freq_history = append(c_freq_history, new_c_freq)
+   t_freq_history = append(t_freq_history, new_t_freq)
+
+   t_freq = new_t_freq
+   c_freq = new_t_freq
+ }
>
> allele_freq_history = data.frame(c_freq_history, t_freq_history)
>
> ggplot() +
+   geom_point(data = allele_freq_history, aes(x = 1:length(t_freq_history),
+                                              y = t_freq_history), color = 'red') +
+   geom_point(data = allele_freq_history, aes(x = 1:length(c_freq_history),
+                                              y = c_freq_history), color = 'blue') +
+   ylim(0,1) + xlab('Generation Number') + ylab('Allele Frequency') +
+   ggtitle('HWE Simulation')
```



[2] Modeling Violations of Hardy-Weinberg Equilibrium

- To preface, HWE is typically violated when $Pr(Z = 1) < 2p(1 - p)$
- Most random processes in population dynamics lead to a decrease in heterozygotes

Inbreeding:

- Due to the limitations from finite population sizes, varying degrees of inbreeding occur
- Essentially, inbreeding describes the extent of the presence of commmon ancestors between genetic material (DNA) in the present-day population
- **IBD:** Identical by decent
- Two alleles are IBD if both are copies from a common ancestor
- Let I be an r.v. that indicates whether 2 randomly drawn alleles are IBD or not:
- $\begin{cases} I = 1 \rightarrow Y_1 = Y_2 \sim Bernoulli(p) \\ I = 0 \rightarrow Y_1, Y_2 \overset{iid}{\sim} Bernoulli(p) \end{cases}$
- **Note:** We only need one random variable, as we are effectively only drawing 1 allele from the population
- $I \sim Bernoulli(f)$, where f is the inbreeding coefficient from a population
- $Z|I = 0 \sim Binomial(2, p)$
- $\frac{Z}{2}|I = 1 \sim Bernoulli(p)$, the $\frac{1}{2}$ coefficient allows us to create a Bernoulli rv with values 0 and 2, rather than 0 and 1
- Let $Y_1, Y_2, Y_3 \overset{iid}{\sim} Bernoulli(p)$
- $Z|I = 0 \sim Binomial(2, p)$
- $Z|I = 1 \sim Bernoulli(p)$
- $Z = (Y_1 + Y_2)(1 - I) + 2Y_3I$, where an inbred allele leads to a value of 0 or 2

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

$Pr(Z = 0) = Pr(Z = 0|I = 0)Pr(I = 0)+Pr(Z = 0|I = 1)Pr(I = 1) = (1-p)^2(1-f)+(1-p)f = (1-p)^2+p(1-p)f$

$Pr(Z = 1) = Pr(Z = 1|I = 0)Pr(I = 0)+Pr(Z = 1|I = 1)Pr(I = 1) = 2p(1-p)(1-f)+0(f) = 2p(1-p)(1-f)$

- **Note:** $2p(1-p)(1-f) \leq 2p(1-p)$, indicating that the reduction is distributed to the hetero-zygotes

$$Pr(Z = 2) = Pr(Z = 2|I = 0)Pr(I = 0) + Pr(Z = 2|I = 1)Pr(I = 1) = p^2(1-f) + pf$$

Below are the derivations for the variance and expected value of Z:

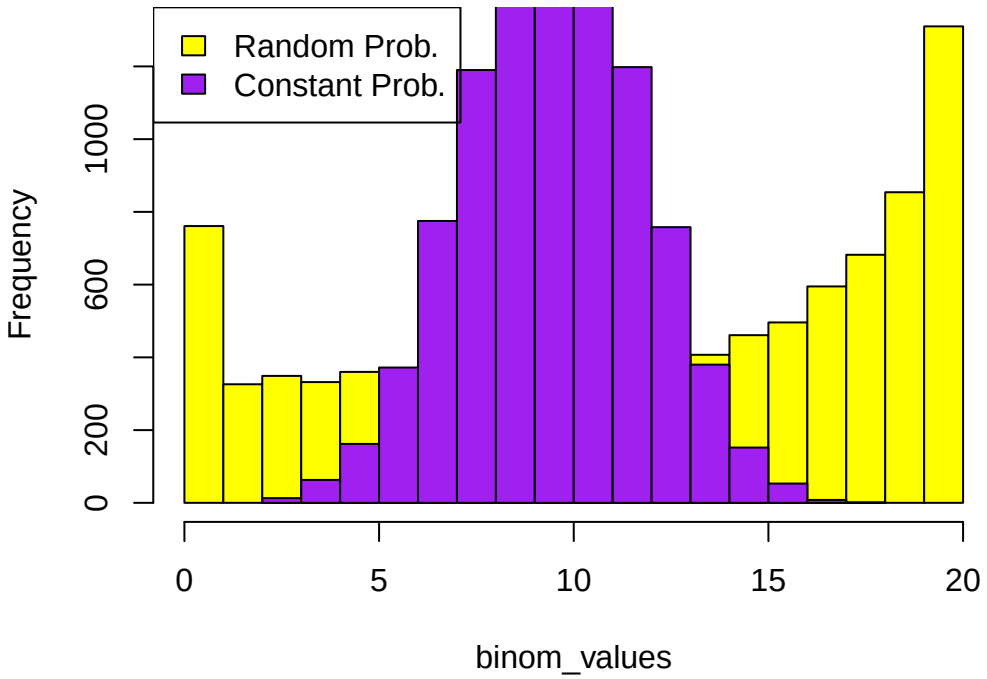
$$Var(Z) = E[Var(Z|I)] + Var(E[Z|I])$$
$$Var(Z|I = 0) = 2p(1-p) \qquad [Binomial(2, p)]$$
$$Var(Z|I = 1) = Var(2Y_3) = 4Var(Y_3) = 4p(1-p)$$
$$Var(Z|I) = 2p(1-p)(1-I) + 4p(1-p)I$$
$$E[Var(Z|I)] = 2p(1-p)(1-f) + 4p(1-p)f = 2p(1-p)(1+f)$$
$$Z|(I = 0) = Y_1 + Y_2$$
$$Z|(I = 1) = 2Y_3$$
$$E[Z|I = 0] = 2p$$
$$E[Z|I = 1] = 2E[Y_3] = 2p$$

- The above two lines indicate that the expected values are the same regardless of inbreeding or not
- Hence, $E[Z|I] = 2p \rightarrow E\left[E[Z|I] = E[2p] = 2p\right.$
- $E[Z] = E[E[Z|I]] = 2p$
- **Result:** $E[Z] = 2p$
- And, $Var(E[Z|I]) = 0$, as $E[Z|I]$ is a constant
- **Result:** $Var(Z) = 2p(1-p)(1+f) + 0 = 2p(1-p)(1+f)$
- $f = \frac{Var(Z)-Var(Z|I=0)}{Var(Z|I=0)}$, (proportion of variance explained by population structure)
- $f = 1 - \frac{Pr(Z=1)}{Pr(Z=1|I=0)}$

Drift:

- To model drift, we need to make the allele freq. random, and then allow for HWE-Style mating
- $Z|Q \sim Binomial(2, Q)$, where Q is a r.v.
- **Better Notation:** $Z|Q = q \sim Binomial(2, q)$
- Below is a simulation of a binomial random variable with a random variable probability parameter (determined by the beta distribution), compared to a binomial random variable with a constant probability parameter

```
>
> #Initialized values for calculating parameters of beta random variable
> p = 0.6
> f = 0.4
>
> #Applying transformations to the values to determine beta r.v. parameters
> alpha_val = p * (1 - f) / f
> beta_val = (1 - p) * (1 - f) / f
>
> #Determining binom parameter probabilites from alpha and beta
> q_values = rbeta(n=10000, shape1 = alpha_val, shape2 = beta_val)
>
> #Initializing 1 constant binom parameter probability for comparision
> q_constant = 0.5
>
> #Determining binom values utilizing the probabilities generated by the beta dist.
> binom_values = unlist(lapply(q_values, function(x) rbinom(n=1, size = 20, x)))
>
> #Generating binom_values for constant probability
> binom_values_const = rbinom(n=10000, size = 20, q_constant)
>
> #Plotting Histograms of Binomial Values
> hist(binom_values, xlim = c(0,20), col="yellow", main = "")
> hist(binom_values_const, add=T, col="purple")
> legend("topleft", c("Random Prob.", "Constant Prob."), fill = c("yellow", "purple"))
```



- $p =$ ancestral allele frequency
- $f =$ fixation index (inbreeding)
- $Q \sim Beta\left(\frac{1-f}{f}p, \frac{1-f}{f}(1-p)\right)$, also known as the Balding–Nichols model
- $Q \sim BN(p, f)$
- $E[Q] = p$, $Var(Q) = p(1-p)f$
- $E[Z] = E[E[Z|Q]] = E[2Q] = 2p$

$$Pr(Z=2) = \int Pr(Z=2|Q=q)f(q)dq = \int q^2 f(q)dq$$
$$= E[Q^2] = Var(Q) + E[Q]^2 = p(1-p)f + p^2$$

- **Worked Out Derivations for** $Pr(Z=0)$ and $Pr(Z=1)$

$$Pr(Z=0) = \int Pr(Z=0|Q=q)f(q)dq$$
$$= \int (1-q)^2 f(q)dq = \int (1-2q+q^2)f(q)dq = \int f(q)dq - 2\int qf(q)dq + \int q^2 f(q)dq$$
$$= E[1] - 2E[Q] + E[Q^2] = 1 - 2(p) + Var(Q) + (E[Q^2])^2 = 1 - 2p + p(1-p)f + (p)^2$$
$$= p(1-p)f + 1 - 2p + p^2 = p(1-p)f + (1-p)^2$$
$$Pr(Z=0) = p(1-p)f + (1-p)^2$$

$$Pr(Z=1) = \int Pr(Z=1|Q=q)f(q)dq$$

Note: The coefficient of 2 accounts for the fact that two outcomes yield heterozygosity

$$= \int 2q(1-q)f(q)dq = \int 2(q-q^2)f(q)dq = \int 2qf(q)dq - \int 2q^2 f(q)dq$$
$$= 2E[Q] - 2E[Q^2] = 2p - 2(Var(Q) + (E[Q])^2) = 2p - 2p(1-p)f - 2(p)^2$$
$$= 2p - 2pf - 2p^2 f - 2p^2 = 2p(1-f) - 2p^2(f-1)$$
$$= 2p(1-f) + 2p^2(1-f) = (2p+2p^2)(1-f) = 2p(1+p)(1-f)$$
$$Pr(Z=1) = 2p(1+p)(1-f)$$

- Derivation of variance below

$$Var(Z) = E[Var(Z|Q)] + Var(E[Z|Q]) = E[2Q(1-Q)] + Var(2Q)$$
$$= E[2Q] - E[2Q^2] + 4p(1-p)f = 2p - 2E[Q^2] + 4p(1-p)f$$
$$= 2p - 2[Var(Q) + E(Q)^2] + 4p(1-p)f = 2p - 2p(1-p)f - 2p^2 + 4p(1-p)f$$
$$= 2p(1-p) + 2p(1-p)f = 2[p(1-p) + p(1-p)f] = 2p(1-p)(1+f)$$

Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  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] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.1
 [5] purrr_0.3.2     readr_1.3.1     tidyr_0.8.3     tibble_2.1.1
 [9] ggplot2_3.1.1   tidyverse_1.2.1 knitr_1.22

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1       cellranger_1.1.0 pillar_1.4.0     compiler_3.6.0
 [5] plyr_1.8.4       tools_3.6.0      digest_0.6.18    lubridate_1.7.4
 [9] jsonlite_1.6     evaluate_0.13    nlme_3.1-140     gtable_0.3.0
[13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.3.4      cli_1.1.0
[17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.0      xfun_0.7
[21] withr_2.1.2      xml2_1.2.0       httr_1.4.0       hms_0.4.2
[25] generics_0.0.2   grid_3.6.0       tidyselect_0.2.5 glue_1.3.1
[29] R6_2.4.0         readxl_1.3.1     rmarkdown_1.12   modelr_0.1.4
[33] magrittr_1.5     backports_1.1.4  scales_1.0.0     htmltools_0.3.6
[37] rvest_0.3.3      assertthat_0.2.1 colorspace_1.4-1 labeling_0.3
[41] stringi_1.4.3    lazyeval_0.2.2   munsell_0.5.0    broom_0.5.2
[45] crayon_1.3.4
```