Week 2 QCB 408 / 508 Spring 2020

Continuous rV's (continued)

Normal (Gaussian) distribution

$X \sim \text{Normal}(\mu, \sigma^2)$

$\mathcal{X} = (-\infty, \infty)$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$E[X] = \mu$, $Var(X) = \sigma^2 > 0$

dnorm(x, mean = 2, sd = 3)

Multivariate rV's

Two rV's X and Y

Joint cdf $F(a,b) = Pr(X \leq a, Y \leq b)$

$= Pr(\{\omega : X(\omega) \leq a\} \cap \{\omega : Y(\omega) \leq b\})$

suppose continuous.

$\frac{d^2}{dx\,dy} F(x,y) = f(x,y) \rightarrow$ joint pdf

Suppose discrete

joint pmf $f(x,y) = Pr(X=x, Y=y)$

Marginal distributions

$$f(x) = \sum_{y \in R_y} f(x,y)$$

$$f(x) = \int f(x,y)\,dy$$

See Law of Total Probability.

Independent rv's

If X and Y are independent

then:

$$f(x,y) = f(x)f(y)$$

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

Conditional Distributions

$$\Pr(X \le a \mid Y \le b) = \frac{\Pr(X \le a, Y \le b)}{\Pr(Y \le b)}$$

$\{\omega : X(\omega) \le a\}$ $\{\omega : Y(\omega) \le b\}$

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

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

Bayes Theorem $\quad f(y \mid x) = \dfrac{f(x \mid y)\, f(y)}{f(x)}$.

All of the above extends to

rvs $X_1, X_2, \ldots, X_n$

$$f(x_1, x_2 \mid x_3, x_4, x_5) = \frac{f(x_1, x_2, x_3, x_4, x_5)}{f(x_3, x_4, x_5)}$$

$$f(x_3, x_4, x_5) = \iint f(x_1, x_2, x_3, x_4, x_5)\, dx_1\, dx_2$$

If $X_1, X_2, \ldots, X_n$ are independent then

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

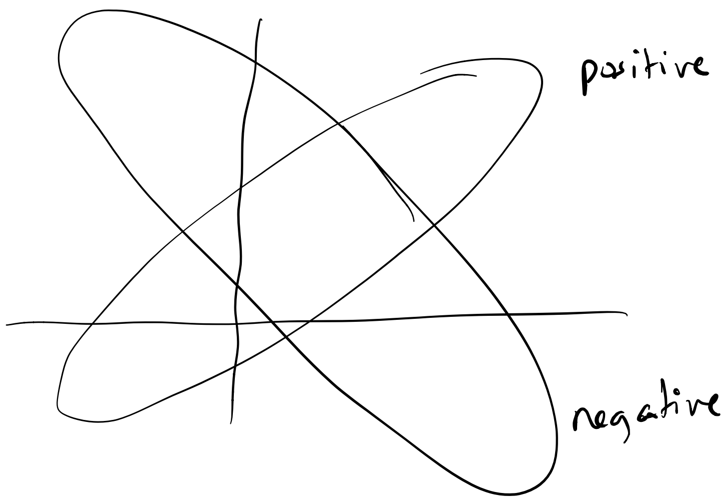
Moments of Joint Distributions

For a single rv, X:

$E[X^k]$ is the k^{th} moment

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

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



First set $E[X], E[Y]$

$$Cov(X,Y) = \int\int (x - E[X])(y-E[Y])\, f(x,y)\, dx\, dy$$

or

$$= \sum_x \sum_y (x - E[X])(y - E[Y])\, \underbrace{f(x,y)}_{Pr(X=x,\, Y=y)}$$

Linear Transformations of rv's

X is a rv

a and b are constants

$$E[a+bX] = a + bE[X]$$

$$Var(a+bX) = b^2\, Var(X)$$

$$E[\underline{a+bX}] = \int (a+bx)\, f(x)\, dx$$

$$= \int a f(x)\, dx + \int bx f(x)\, dx$$

$$= a \underbrace{\int f(x)\, dx}_{=1} + b \underbrace{\int x f(x)\, dx}_{E[X]}$$

$$= a + bE[X]$$

<u>Exercise</u>: $Var(a+bX) = b^2 Var(X)$

Law of Total Variance

Jointly distributed rv's X and Y

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

$E[X|Y] \to ?$

$$E[X|Y=y] = \underbrace{\int x f(x|y)\,dx}_{\text{function of } y}$$

Have a pmf or pdf for y

$$E[X|Y] = \underbrace{\int x f(x|Y)\,dx}_{\text{function of } Y}$$

$$E[\underline{E[X|Y]}] = \int \underbrace{\int x f(x|y)\,dx}_{E[X|Y=y]} f(y)\,dy$$

$$= \int\int x\, \underline{f(x|y) f(y)}\,dx\,dy$$

$$= \int\int x \, f(x,y) \, dx \, dy$$

$$= \int\int x \, f(x,y) \, dy \, dx$$

$$= \int x \underbrace{\int f(x,y) \, dy}_{} \, dx$$

$$= \int x \, f(x) \, dx$$

$$= E[X]$$

$$f(x) = \int f(x,y) \, dy$$

$$Var(X|Y) = E\{(X - E[X|Y])\}^2 | Y]$$

<u>Exercise</u> : Prove Law of Total Variance

$\downarrow$

or read

Hardy-Weinberg Equilibrium

SNP   CC, CT, TT

$$CC \to 0,\quad CT \to 1,\quad TT \to 2$$

Genotype $X \in \{0, 1, 2\}$

$Pr(X=k) = r_k \qquad k = 0, 1, 2$

$E[X] = r_1 + 2r_2 \qquad \underline{E\left[\frac{X}{2}\right] = \frac{r_1}{2} + r_2 \equiv p}$ freq. of T

$Y \in \{0, 1\}$ is the transmitted

$$Pr(Y=1 \mid X=k) = \begin{cases} 0 & k=0 \\ 1/2 & k=1 \\ 1 & k=2 \end{cases}$$

$$Pr(Y=1) = \sum_{k=0}^{2} Pr(Y=1 \mid X=k)\, Pr(X=k)$$

$$= 0 \cdot r_0 + \frac{1}{2} r_1 + r_2$$

$$= p$$

$\Longrightarrow Y \sim \text{Bernoulli}(p)$

$Z \in \{0, 1, 2\}$ is next generation genotype

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

iid = independent, identically distributed

$Z = Y_1 + Y_2$

$\Rightarrow Z \sim \text{Binomial}(2, p)$

$E[Z] = 2p = E[X]$ $\Rightarrow$ transmission probs for Z are identical to X

$\Rightarrow$ equilibrium

$\Rightarrow$ all future generations' genotypes are drawn from Binomial$(2, p)$

$Pr(Z=0) = (1-p)^2$

$Pr(Z=1) = 2p(1-p)$ $\leftarrow$

$Pr(Z=2) = p^2$

Typically $Pr(Z=1) < 2p(1-p)$ when HWE is violated.

Inbreeding

IBD = identical by descent

Two alleles are IBD if they are copies from a common ancestor

Let I be a rv such that it indicates whether two randomly drawn alleles are IBD or not.

Y_1 and Y_2 :

$$I=1 \Rightarrow Y_1 = Y_2 \sim \text{Bernoulli}(p)$$

$$I=0 \Rightarrow Y_1, Y_2 \overset{iid}{\sim} \text{Bernoulli}(p)$$

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

f is the inbreeding coefficient (usually F)

$Z \mid I=0 \sim \text{Binomial}(2, p)$

$Z/2 \mid I=1 \sim \text{Bernoulli}(p)$

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

$Z \mid I=0 \sim \text{Binomial}(2, p)$

$Z \mid I=1 \sim \text{Bernoulli}(p)$

$$Z = (Y_1 + Y_2)(1 - I) + 2Y_3 I$$

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

HWE

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

i.e., $Z \mid I=1 \sim 2 \cdot \text{Bernoulli}(p)$

$$\Pr(Z=0) = \Pr(Z=0 \mid I=0)\Pr(I=0) + \Pr(Z=0 \mid I=1)\Pr(I=1)$$

$$= (1-p)^2(1-f) + (1-p)f$$

$$= \boxed{(1-p)^2 + \underline{p(1-p)f}}$$

$$\Pr(Z=1) = 2p(1-p)(1-f) + 0 \cdot f$$

$$= \boxed{2p(1-p)(1-f)} < 2p(1-p)$$

$$\Pr(Z=2) = p^2(1-f) + pf$$

$$= \boxed{p^2 + \underline{p(1-p)f}}$$

$$E[Z] = E[E[Z \mid I]] =$$

Note: $2p(1-p) = 2p(1-p)(1-f) + 2p(1-p)f$

$$= 2p(1-p)(1-f) + p(1-p)f + p(1-p)f$$

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

$$\mathrm{Var}(Z \mid I=0) = 2p(1-p) \qquad [\text{Binomial}(2,p)]$$

$$\mathrm{Var}(Z \mid I=1) = \mathrm{Var}(2Y_3)$$
$$= 4\,\mathrm{Var}(Y_3)$$
$$= 4p(1-p)$$

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

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

$$Z \mid I=0 = Y_1 + Y_2$$
$$Z \mid I=1 = 2Y_3$$

$$E[Z \mid I=0] = 2p$$
$$E[Z \mid I=1] = 2E[Y_3] = 2p$$

$$E[Z \mid I] = 2p$$

$$\mathrm{Var}(E[Z \mid I]) = 0$$

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

$$f = \frac{\mathrm{Var}(Z) - \mathrm{Var}(Z \mid I=0)}{\mathrm{Var}(Z \mid I=0)}$$

$$f = 1 - \frac{\Pr(Z=1)}{\Pr(Z=1 \mid I=0)}$$

<u>Drift</u>

Make allele frequency random.

Then allow for HWE style mating.

$Z \mid Q \sim \text{Binomial}(2, Q)$

Q is a random variable

$Z \mid Q=q \sim \text{Binomial}(2, q)$ ✓

Q changes according to a distribution

p = ancestral allele frequency

f = fixation index, "inbreeding"

$$Q \sim \text{Beta}\Big(\underbrace{\frac{1-f}{f}p}_{\alpha}, \underbrace{\frac{1-f}{f}(1-p)}_{\beta}\Big)$$

$$= BN(p, f)$$

Balding-Nichols

$$Q \sim BN(p, f)$$

$$E[Q] = p, \quad Var(Q) = p(1-p)f$$

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

$$Pr(Z=2) = \int Pr(Z=2|Q=q) f(q)\, dq$$

$Var(Q) = E[Q^2] - E[Q]^2$

$$= \int q^2 f(q)\, dq$$

$$= E[Q^2]$$

$$= Var(Q) + E[Q]^2$$

$$= p(1-p)f + p^2$$

Verify: $Pr(Z=0) = p(1-p)f + (1-p)^2$

$Pr(Z=1) = 2p(1-p)(1-f)$

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

Model of RNA-seq Data

Assume $i = 1, 2, \ldots, m$ genes
$j = 1, 2, \ldots, n$ observations

Consider a single biological condition or population.

Observe RNA-seq read counts Y_{ij} for gene i in observation j.

Target: True proportion of gene i expression, call it a_i

The proportion of mRNA for gene i in this population is a_i.

$\sum_{i=1}^{m} a_i = 1$, most a_i are small.

Instructive, idealized example to follow.

Step 1. Sample cells and mRNA molecules

Step 2. Sequence mRNA molecules and obtain counts.

Step 1:

M_j is the number of mRNA molecules sampled for observation j

X_{ij} is the number of mRNA molecules sampled for gene i in observation j

Assume completley random sampling of mRNA molecules.

$$\Rightarrow X_{ij} \mid M_j \sim \text{Binomial}(M_j, a_i)$$

M_j is large , a_i is small

$\Longrightarrow$ $X_{ij} \sim \text{Poisson}(M_j a_i)$

$$Var(X_{ij} \mid M_j) = M_j a_i \underbrace{(1-a_i)}_{\approx 1}$$

$\pi_{ij} = \dfrac{X_{ij}}{M_j}$ random proportion of gene i mRNA in observation j

$E[\pi_{ij}] =$ to be continued next week (week 3)...