Week 3 QCB 408/508 Spring 2020

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

$$E[\pi_{ij}] = E\left[\frac{X_{ij}}{M_j}\right]$$

$$= E\left[E\left[\frac{X_{ij}}{M_j} \mid M_j\right]\right]$$

$$= E\left[E\left[\frac{a_i M_j}{M_j} \mid M_j\right]\right]$$

$$= E[a_i] = a_i$$

$$Var(\pi_{ij}) = E[Var(\pi_{ij} \mid M_j)] + \underbrace{Var(\overbrace{E[\pi_{ij} \mid M_j]}^{= a_i})}_{0}$$

$Var(cX) = c^2 Var(X)$

$$Var(\pi_{ij} \mid M_j) = Var\left(\frac{X_{ij}}{M_j} \mid M_j\right)$$

$$= \frac{1}{M_j^2} Var(X_{ij} \mid M_j)$$

$$\approx \frac{1}{M_j^2} a_i M_j = \frac{a_i}{M_j}$$

$$\Rightarrow \operatorname{Var}(\pi_{ij}) \approx E\left[\frac{a_i}{m_j}\right] = a_i E\left[\frac{1}{m_j}\right]$$

$\approx$ "biological variance"

Step 2.

This is assumed (here) to be completely random sampling.

Let D_j be the "read depth" of observation j, which is the total number of reads from observation j.

Note: we observe D_j, say d_j.

Reminder: Y_{ij} RNA-Seq read counts for gene i, obs. j.

$$Y_{ij} \mid \pi_{ij}, d_j \sim \text{Binomial}(d_j, \pi_{ij})$$

$$\overset{\cdot}{\sim} \text{Poisson}(d_j \pi_{ij})$$

d_j large, π_{ij} small

$$E[Y_{ij}] = E\left[E[Y_{ij} \mid \pi_{ij}, d_j]\right]$$

$$= E[\pi_{ij} d_j]$$

$$= d_j E[\pi_{ij}]$$

$$= d_j a_i$$

$$Var(Y_{ij}) = E[Var(Y_{ij} \mid \pi_{ij}, d_j)] + Var(E[Y_{ij} \mid \pi_{ij}, d_j])$$

$$\approx E[\pi_{ij} d_j] + Var(\pi_{ij} d_j)$$

$$= a_i d_j + d_j^2 Var(\pi_{ij})$$

$$= a_i d_j + d_j^2 a_i E\left[\frac{1}{n_j}\right]$$

$$> E[Y_{ij}] \Rightarrow \text{over-dispersed Poisson}$$

Estimate $\hat{\pi}_{ij} = \dfrac{Y_{ij}}{d_j}$, $E[\hat{\pi}_{ij}] = a_i$

$$Var(\hat{\pi}_{ij}) = \frac{1}{d_j^2} Var(Y_{ij})$$

$$= \frac{a_i}{d_j} + Var(\pi_{ij})$$

$$= \frac{a_i}{d_j} + a_i E\left[\frac{1}{n_j}\right]$$

technical variance $\approx \dfrac{a_i}{d_j}$

biological variance $\approx a_i E\left[\dfrac{1}{M_j}\right]$

Consider estimate $\hat{a}_i = \dfrac{\sum_{j=1}^{n} \hat{\pi}_{ij}}{n}$.

$E[\hat{a}_i] = a_i$.

$$\begin{aligned}
Var(\hat{a}_i) &= Var\left(\frac{\sum_{j=1}^{n} \hat{\pi}_{ij}}{n}\right) \\
&= \frac{1}{n^2} Var\left(\sum_{j=1}^{n} \hat{\pi}_{ij}\right) \\
&= \frac{1}{n^2} \sum_{j=1}^{n} Var(\hat{\pi}_{ij}) \\
&= \frac{a_i}{n^2} \sum_{j=1}^{n} \frac{1}{d_j} + \frac{\sum_{j=1}^{n} Var(\pi_{ij})}{n^2}
\end{aligned}$$

Let's assume M_j are i.i.d.

Then $E\left[\dfrac{1}{M_j}\right]$ is the same for all j.

Coefficient of Variation, CV

$$CV = \frac{\sqrt{Var(\pi_{ij})}}{a_i} \Rightarrow \text{biological CV}$$

$$(CV)^2 = \frac{Var(\pi_{ij})}{a_i^2} = \frac{1}{a_i} E\left[\frac{1}{m_j}\right] \equiv \phi_i$$

Let $\mu_{ij} = d_j a_i$ (population mean times obs. read depth)

$$\Rightarrow Var(Y_{ij}) = d_j a_i + d_j^2 Var(\pi_{ij})$$

$$= d_j a_i + (d_j a_i)^2 \frac{Var(\pi_{ij})}{a_i^2}$$

$$= \mu_{ij} + \mu_{ij}^2 \underline{\phi_i}$$

estimated by "borrowing strength" across genes — with similar ϕ_i values

---

Negative Binomial

Bernoulli trials with success p

Y = number of failures before r^{th} success

$Y \sim \text{NegBin}(r, p)$

$$Pr(Y=y) = \binom{r+y-1}{y} p^r (1-p)^y$$

$y = 0, 1, 2, \ldots$

$$E[Y] = r\frac{(1-p)}{p} \quad Var(Y) = \frac{r(1-p)}{p^2}$$

$$\text{Let } \mu = \frac{r(1-p)}{p}$$

$$Var(Y) = \mu + \mu^2 \underbrace{\frac{1}{r}}_{\phi}$$

RNA-Seq data under the above model is therefore sometimes modeled as a Neg Bin.

$$Y_{ij} \sim NegBin(p_{ij}, r_i)$$

$$\text{where } \mu_{ij} = \frac{r_i(1-p_{ij})}{p_{ij}} \text{ and } \phi_i = \frac{1}{r_i}.$$

Compound Gamma-Poisson Distribution

$$Y_{ij} \mid \lambda_{ij} \sim Poisson(\lambda_{ij})$$

$$\lambda_{ij} \sim Gamma(\alpha, \beta)$$

Then Y_{ij} is marginally a Gamma-Poisson r.v.

Neg Bin is a special case of Gamma-Poisson.

Suppose $Y \mid \lambda \sim \text{Poisson}(\lambda)$

$\lambda \sim \text{Gamma}(\alpha, \beta)$

Gamma pdf

$$f(\lambda; \alpha, \beta) = \frac{\lambda^{\beta-1} e^{-\lambda/\alpha}}{\alpha^{\beta}\,\Gamma(\beta)} \qquad \lambda > 0$$

$E[\lambda] = \alpha\beta$, $Var(\lambda) = \alpha^2\beta$

Gamma-Poisson:

$$f(y; \alpha, \beta) = \frac{\Gamma(y+\beta)\,\alpha^{y}}{\Gamma(\beta)\,(1+\alpha)^{\beta+y}\,y!}$$

$E[Y] = \alpha\beta \qquad Var(Y) = \alpha\beta + \alpha^2\beta$

Let $\mu = \alpha\beta$

$Var(Y) = \mu + \mu^2 \cdot \frac{1}{\beta}$

Let's map this back to the 2-step model.

$$Y_{ij} \mid \lambda_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$\lambda_{ij} \sim \text{Gamm}(\alpha, \beta) \rightarrow$$ let's determine what are λ_{ij}, α, and β

λ_{ij} is $\pi_{ij} d_j$, which is random

$E[\pi_{ij} d_j] = a_i d_j$ and

$E[\lambda_{ij}] = \alpha\beta$. so

$$\mu_{ij} = a_i d_j = \alpha\beta$$

$$\text{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}^2 \phi_i,$$

so $\frac{1}{\beta} = \phi_i = \frac{1}{a_i} E\left[\frac{1}{M_j}\right]$.

$$\Rightarrow \beta_{ij} = a_i E\left[\frac{1}{M_j}\right]^{-1}$$

$$\Rightarrow \alpha_{ij} = a_i d_j \times \frac{1}{a_i} E\left[\frac{1}{M_j}\right]$$

$$= d_j \times E\left[\frac{1}{M_j}\right]$$

$$\Rightarrow \lambda_{ij} \sim \text{Gamma}(\alpha_{ij}, \beta_{ij}).$$

$Y \mid \lambda \sim \text{Poisson}(\lambda)$, $\lambda > 0$

λ is r.v. with $E[\lambda]$ and $Var(\lambda)$

$$Var(Y) = E[Var(Y \mid \lambda)] + Var(E[Y \mid \lambda])$$

$$= E[\lambda] + Var(\lambda)$$

Since $\lambda > 0$, $E[\lambda]$ will "appear" in $Var(\lambda)$ $\Rightarrow$ mean-variance relationship

Sums of Random Variables

If X is a rv and a, b are constants, then $E[a + bX] = a + bE[X]$ and $Var(a + bX) = b^2 Var(X)$.

Let $X_1, X_2, \ldots, X_n$ be n rv's. Then

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j)$$

when $X_1, X_2, \ldots, X_n$ are independent,

then $Cov(X_i, X_j) = 0$ for $i \neq j$

so $Var(\sum X_i) = \sum Var(X_i)$.

Let $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Suppose $X_1, X_2, \ldots,$ X_n are independent.

$$E[\overline{X}_n] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underline{E[X_i]}$$

when $E[X_1] = E[X_2] = \ldots = E[X_n] = \theta$

then $E[\overline{X}_n] = \theta$.

$$Var(\overline{X}_n) = \frac{1}{n^2} \sum Var(X_i)$$

If $Var(X_1) = Var(X_2) = \ldots = Var(X_n) = \tau^2$

$$Var(\overline{X}_n) = \frac{\tau^2}{n}.$$

Convergence of rv's

Let $Z_1, Z_2, \ldots$ be a sequence of rv's.

Examples: $Z_n = \bar{X}_n$.

$Z_n \sim \text{Binomial}(n, p)$

① Convergence in Distribution:

$\{Z_n\}$ converges in distribution to rv W

$$Z_n \xrightarrow{D} W \text{ as } n \to \infty \text{ if}$$

$$F_{Z_n}(y) = \Pr(Z_n \le y) \to \Pr(W \le y) = F_W(y)$$

for all $y \in \mathbb{R}$, as $n \to \infty$.

② Convergence in Probability:

$$Z_n \xrightarrow{P} W \text{ as } n \to \infty \text{ if}$$

$\Pr(|Z_n - W| \le \varepsilon) \to 1$ as $n \to \infty$, for $\varepsilon > 0$.

If θ is a fixed number, then we can have $Z_n \xrightarrow{P} \theta$.

③ Almost sure convergence

$\{Z_n\}$ converges "almost surely" (a.s.) or "with probability 1" to W

$$Z_n \xrightarrow{a.s.} W \quad \text{if}$$

$$P_r(\{\omega : |Z_n(\omega) - W(\omega)| \xrightarrow{n\to\infty} 0\}) = 1$$

<u>Strong Law of Large Numbers</u>

Suppose $X_1, X_2, \ldots$ are iid rv's with population mean $E[X_i] = \mu$ where $E[|X_i|] < \infty$. Then

$$\bar{X}_n \xrightarrow{a.s.} \mu, \text{ as } n \to \infty.$$

<u>Central Limit Theorem</u>

Suppose $X_1, X_2, \ldots$ are iid rv's with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.

Then as $n \to \infty$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \text{Normal}(0, \sigma^2)$$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \text{Normal}(0, 1)$$

Note:

$$\text{Var}(\bar{X}_n - \mu) = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

$$\text{Var}\left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}\right) = \frac{1}{\sigma^2/n}\text{Var}(\bar{X}_n) = 1$$

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0, 1).$$

<u>Normal rv's – some useful facts</u>

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$

$E[\bar{X}_n] = \mu$, $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$

$\bar{X}_n \sim \text{Normal}(\mu, \sigma^2/n)$

because

$$X_1 + X_2 + \dots + X_n \sim \text{Normal}(n\mu, n\sigma^2)$$

$$aX_1 + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$$