Bernoulli, Binomial, Trinomial, Multinomial

Bernoulli RV

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

$P(X=1)=p$

$f(x)=p^x(1-p)^{1-x}$

$$\mu_x = E(X) = \sum_{x=0}^{1} x\, f(x)$$

$$\sigma_x^2 = Var(X) = E\left(X-\mu_x\right)^2$$

Bernoulli, Binomial, Trinomial, Multinomial

Binomial

X = # Successes in n Bernoulli trials (i.e., counts in 1 of 2 allelic/genotypic categories)

$X \in \{0,n\}$

$$f(x)=P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$$
$$=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

$\mu_x = E(X) = np$

$\sigma_x^2 = Var(X) = np(1-p)$

- Note: *X/n* is the observed sample proportion, $\hat{p}$ (which is the MLE for *p*)

Bernoulli, Binomial, Trinomial, Multinomial

$$f(x_1,x_2,x_3)=P(X_1=x_1,X_2=x_2,X_3=x_3)=\frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}$$

$$f(x_1,x_2)=\frac{n!}{x_1!x_2!(n-x_1-x_2)!}p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2}$$

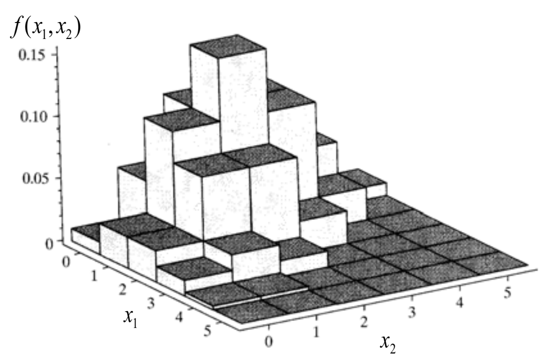

Figure 4.1-6: Trinomial distribution, $p_1 = 1/5$, $p_2 = 2/5$, and $n = 5$

Bernoulli, Binomial, Trinomial, Multinomial

Multinomial (*m* categories)

$x_1 + \ldots + x_m = n$

$p_1 + \ldots + p_m = 1$

$$f(x_1,\ldots,x_m)=\frac{n!}{x_1!\cdots x_m!}p_1^{x_1}\cdots p_m^{x_m}$$

$$f(n_1,\ldots,n_m)=\frac{n!}{n_1!\cdots n_m!}p_1^{n_1}\cdots p_m^{n_m}$$

Trinomial (3 categories)

$x_{AA} + x_{Aa} + x_{aa} = n$

$p_{AA} + p_{Aa} + p_{aa} = 1$

$$f(x_{AA},x_{Aa},x_{aa})=\frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}p_{AA}^{x_{AA}}p_{Aa}^{x_{Aa}}p_{aa}^{x_{aa}}$$

$$f(n_{AA},n_{Aa},n_{aa})=\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}p_{AA}^{n_{AA}}p_{Aa}^{n_{Aa}}p_{aa}^{n_{aa}}$$

Marginal Distributions

$$f(x_{AA}, x_{Aa}, x_{aa}) = \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!} p_{AA}^{x_{AA}} p_{Aa}^{x_{Aa}} p_{aa}^{x_{aa}}$$

- Summing over one (or more) of the categories gives a marginal distribution for the remaining count(s).
- For the trinomial, collapsing two categories gives the Binomial Distribution – with category *i* versus *"other"*.

$$f(x_{AA}, x_{Aa+aa}) = \frac{n!}{x_{AA}!x_{Aa+aa}!} p_{AA}^{x_{AA}} p_{Aa+aa}^{x_{Aa+aa}}$$
$$= \frac{n!}{x_{AA}!(n-x_{AA})!} p_{AA}^{x_{AA}} \left(1-p_{AA}\right)^{n-x_{AA}}$$

Multinomial Moments (Expected Cell Counts & Var)

- Since marginal counts for each cell (vs. "other") are Binomial, we have

$$E(X_i) = np_i$$
$$Var(X_i) = np_i(1-p_i)$$
$$Var(\hat{p}_i) = Var\left(\frac{X_i}{n}\right) = \frac{p_i(1-p_i)}{n}$$

Maximum Likelihood Estimation (Binomial)

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \rightarrow \quad L(p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

- The maximum likelihood estimate (MLE) is the value of the parameter that maximizes the likelihood (or the log-likelihood - which is easier to work with).

$$l(p) = \ln\left[L(p)\right] = \ln\left[n!\right] - \ln[x!(n-x)!] + x\ln[p] + (n-x)\ln[1-p]$$

$$\frac{d\,l(p)}{dp} = \frac{x}{p} - \frac{(n-x)}{1-p} = 0 \quad \rightarrow \quad \hat{p} = \frac{x}{n} \text{ is the MLE for } p$$

Maximum Likelihood Estimation (Trinomial)

$$f(x_1, x_2, x_3 \mid p_1, p_2, p_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \quad \rightarrow \quad L(p_1, p_2, p_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

- In order to maximize the (log-) likelihood, we need to take into account the constraint on the frequencies ($p_1 + p_2 + p_3 = 1$) and use Lagrange multipliers
  → maximize the Lagrangian function $H(p_1, p_2, p_3, \lambda)$, which incorporates the constraint

$$H(p_1, p_2, p_3, \lambda) = \ln\left[L(p)\right] + \lambda\left(1 - \sum_i p_i\right)$$

$$\frac{\partial H}{\partial p_1} = 0,\ \frac{\partial H}{\partial p_2} = 0,\ \frac{\partial H}{\partial p_3} = 0,\ \frac{\partial H}{\partial \lambda} = 0 \quad \rightarrow \quad \text{a set of equations to solve for } \hat{p}_i$$

$$\rightarrow \quad \hat{p}_i = \frac{x_i}{n}$$

The Delta Method for var[*f(X)*]

- It is based on Taylor series expansions for the mean and variance of a function of a Random Variable.
- It is necessary because there is no general theoretical expression for the variance of most functions of a mean (e.g., the inverse of a mean, or the ratio of two means) and other statistics.
- Estimators for many genetic parameters involve ratios of multinomial frequencies, meaning that exact expressions for variances can not be found.

Taylor Series Expansion of a function *f(x)*

- Taylor Series expansion of ***f(·)*** about the value ***x=a***

$$f(x) \overset{(T)}{=} f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

$$f(x) \overset{(T)}{\approx} f(a) + f'(a)(x-a)$$

- The Delta Method uses a TS expansion with $a=\mu_x$, the mean of ***X***

$$y = f(x) \overset{(T)}{\approx} f(\mu_x) + f'(\mu_x)(x-\mu_x)$$

The Delta Method for var[*f(X)*]

$$y \overset{(T)}{=} f(\mu_x) + f'(\mu_x)(x-\mu_x) + \underbrace{\frac{f''(\mu_x)}{2!}(x-\mu_x)^2 + \cdots + \frac{f^{(n)}}{n!}(\mu_x)(x-\mu_x)^n}$$

$$E(Y) = E[f(X)] = E[f(\mu_x) + f'(\mu_x)(x-\mu_x) + H.O.T.]$$

$$Var(Y) = Var[f(X)] = E[Y - E(Y)]^2$$

The Delta Method for var[*f(X)*]

$$Var(Y) = Var[f(X)] \approx [f'(\mu_x)]^2\, Var(X)$$

Suppose $Y = 1/X$. Then $f(x) = 1/x$ and $f'(x) = -1/x^2$, so that:

$$\mathrm{Var}(Y) \approx \left[-\frac{1}{\mu_x^2}\right]^2 \mathrm{Var}(X) = \frac{\sigma_x^2}{\mu_x^4}.$$

The Delta Method

- For a function of a single variable, we use a TS expansion about $\boldsymbol{x=\mu_x}$, the mean of the RV $\boldsymbol{X}$

$$y = f(x) \overset{(T)}{\approx} f(\mu_x) + f'(\mu_x)(x-\mu_x)$$

to get

$$Var(Y) = Var\left[f(X)\right] \approx \left[f'(\mu_x)\right]^2 Var(X)$$

- For $T=f(x_1,x_2)$, we use a TS expansion about the point $(\boldsymbol{x,y})=(\boldsymbol{\mu_{x1}}, \boldsymbol{\mu_{x2}})$

$$T = f(x_1,x_2) \overset{(T)}{\approx} f(\mu_{x_1},\mu_{x_2}) + \frac{\partial f}{\partial x_1}(x_1-\mu_{x_1}) + \frac{\partial f}{\partial x_2}(x_2-\mu_{x_2})$$

to get

$$Var(T) \approx \left[\frac{\partial f}{\partial x_1}\right]^2 Var(X_1) + \left[\frac{\partial f}{\partial x_2}\right]^2 Var(X_2) + 2\frac{\partial f}{\partial x_1}\frac{\partial f}{\partial x_2} Cov(X_1,X_2)$$

The Delta Method for var[T=$f(X_1, X_2,\ldots)$]

- For 2 dimensions [$T=f(x_1,x_2)$]

$$Var(T) \approx \left[\frac{\partial f}{\partial x_1}\right]^2 Var(X_1) + \left[\frac{\partial f}{\partial x_2}\right]^2 Var(X_2) + 2\frac{\partial f}{\partial x_1}\frac{\partial f}{\partial x_2} Cov(X_1,X_2)$$

- For k dimensions

Let $T = f(x_1,\ldots,x_k)$

$$Var(T) \approx \sum_i \left(\frac{\partial T}{\partial x_i}\right)^2 Var(x_i) + \sum_i \sum_{j\neq i} \frac{\partial T}{\partial x_i}\frac{\partial T}{\partial x_j} Cov(x_i,x_j)$$

Multinomial Covariances

- The covariance of counts (or proportions) is the expected value of the product of deviations of the counts from their means:

$$\begin{aligned} Cov(X_i,X_j) &= E\left\{\left[(X_i - E(X_i)\right]\left[(X_j - E(X_j)\right]\right\} \\ &= E(X_i,X_j) - E(X_i)E(X_j) \\ &= -np_ip_j \end{aligned}$$

$$Cov(\hat{p}_i,\hat{p}_j) = -\frac{1}{n}p_ip_j$$

The Delta Method for var[$f(X,Y)$]

Two-Variable Taylor Series Expansion: Suppose now we have random variables X, Y. A Taylor series expansion of $f(x,y)$ about the values (x_0,y_0) is given by:

$$f(x,y) = f(x_0,y_0) + \left.\frac{\partial f(x,y)}{\partial x}\right|_{(x_0,y_0)}(x-x_0) + \left.\frac{\partial f(x,y)}{\partial y}\right|_{(x_0,y_0)}(y-y_0) + \begin{pmatrix}\text{2nd and higher} \\ \text{order terms}\end{pmatrix}$$

Suppose $f(x,y) = \frac{y}{x}$. Then: $\frac{\partial f(x,y)}{\partial x} = \frac{-y}{x^2}$, $\frac{\partial f(x,y)}{\partial y} = \frac{1}{x}$

$$\Longrightarrow \quad f(x,y) = \frac{y}{x} \approx \frac{\mu_y}{\mu_x} + \frac{-\mu_y}{\mu_x^2}(x-\mu_x) + \frac{1}{\mu_x}(y-\mu_y)$$

$$Var\left(\frac{Y}{X}\right) \approx Var\left[\frac{\mu_y}{\mu_x} + \frac{-\mu_y}{\mu_x^2}(X-\mu_x) + \frac{1}{\mu_x}(Y-\mu_y)\right]$$