STAT 315: Binomial, Multinomial,Hypergeometric

Luc Rey-Bellet

University of Massachusetts Amherst

luc@math.umass.edu

November 4, 2025

Variance of a sum

If

$$Y = X_1 + X_2 + \cdots + X_n$$

then

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

and

$$\begin{aligned} V[Y] &= \sum_{i=1}^{n} V[X_i] + 2 \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} V[X_i] + \sum_{i > j} \mathrm{Cov}(X_i, X_j) \end{aligned}$$

Sampling without replacement: binomial

Suppose Y binomial with parameters n, p. Then

$$Y = X_1 + \cdots + X_n \,, \quad \text{where} \quad X_i = \begin{cases} 1 & i^{th} \text{ trial} = \text{success} \\ 0 & i^{th} \text{ trial} = \text{failure} \end{cases}$$

We have

$$E[X_i] = p, \quad V[X_i] = p(1-p)$$

The X_i are independent and so $\mathrm{Cov}(X_i, X_j) = 0$ and thus

$$\text{Mean: } E[Y] = np \,, \quad \text{Variance: } V[Y] = np(1-p)$$

Sampling without replacement: hypergeometric

Sample n balls out of N balls with r red balls and $N-r$ green balls.

$$Y = X_1 + \cdots + X_n \quad \text{where} \quad X_i = \begin{cases} 1 & i^{th} \text{ ball} = \text{red} \\ 0 & i^{th} \text{ ball} = \text{green} \end{cases}$$

The X_i are not independent but they are identically distributed. It does not matter how we order the n balls we sample!

We have

$$V[Y] = \sum_{i=1}^{n} V[X_i] + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

$$P(X_1 = 1) = p = \frac{r}{N} \implies V(X_1) = \frac{r}{N}\left(1 - \frac{r}{N}\right)$$

$$P(X_1 = 1, X_2 = 1) = \frac{r}{N}\frac{r-1}{N-1} \quad \implies E(X_1, X_2) = \frac{r}{N}\frac{r-1}{N-1}$$

and so

$$\mathrm{Cov}(X_1, X_2) = \frac{r}{N}\frac{r-1}{N-1} - \frac{r}{N}\frac{r}{N} = -\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}$$

that is X_1 and X_2 are negatively correlated.
By symmetry $\mathrm{Cov}(X_i, X_j)$ are all equal $(i \neq j)$ and so we find

$$V[Y] = n\frac{r}{N}\left(1 - \frac{r}{N}\right) - n(n-1)\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}$$

$$V(Y) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$$ Variance of Hypergeometric

Note as $N \to \infty$ if we assume $\frac{r}{N} \to p$ then $\frac{N-n}{N-1} \to 1$ and

$$V[Y] \to np(1-p)$$

The Multinomial Distribution

Multinomial

- We perform n independent trials, each with k possible outcomes $C_1, C_2, \ldots, C_k$.
- Each outcome C_i occurs with probability p_i, where $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$.
- Let X_i be the number of times outcome C_i occurs.

Definition:

$$(Y_1, Y_2, \ldots, Y_k) \sim \mathrm{Multinomial}(n; p_1, p_2, \ldots, p_k)$$

PDF:

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}, \quad \sum_{i=1}^k n_i = n.$$

Examples

Example 1: Rolling a fair die

Roll a fair die $n = 10$ times:

$$(X_1, X_2, X_3, X_4, X_5, X_6) \sim \mathrm{Multinomial}\Big(10; \tfrac{1}{6}, \ldots, \tfrac{1}{6}\Big)$$

$$P(X = (2, 1, 3, 0, 2, 2)) = \frac{10!}{2!1!3!0!2!2!} \left(\tfrac{1}{6}\right)^{10}.$$

Each X_i counts how many times face i appears.

Example 2: Survey on preferred transport mode

20 people choose: Car (0.5), Bus (0.3), Bike (0.2):

$$(X_{\text{car}}, X_{\text{bus}}, X_{\text{bike}}) \sim \mathrm{Multinomial}(20; 0.5, 0.3, 0.2)$$

$$P(X = (10, 6, 4)) = \frac{20!}{10!6!4!}(0.5)^{10}(0.3)^6(0.2)^4.$$

Interpretation: counts across categories follow a multinomial law.

Properties of the multinomial

Mean, Variance, Covariance

Mean and variance: $E[Y_i] = np_i\,, \quad V(Y_i) = np_i(1-p_i)$.

Covariance: $\mathrm{Cov}(X_i, X_j) = -np_ip_j$

We write

$$Y_i = X_{i,1} + \cdots + X_{i,n} \quad \text{where} \quad X_{i,l} = \begin{cases} 1 & l^{th} \text{ trial} = C_i \\ 0 & l^{th}\text{trial} = \text{something else} \end{cases}$$

Since the trials are independent with $P(X_{i,l} = 1) = p_i$ we find, like for a binomial random variable, $E[Y_i] = np_i$ and $V(Y_k) = np_i(1-p_i)$.

Using that the trials are independent we find

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \sum_{l,m=1}^{n} \mathrm{Cov}(X_{i,l}, X_{j,m}) = \sum_{l}^{n} \mathrm{Cov}(X_{i,l}, X_{j,l}) \\ &= \sum_{l=1}^{n} E[X_{i,l}X_{j,l}] - E[X_{i,l}]E[X_{j,l}] = -np_ip_j \end{aligned}$$

since the product $X_{i,l}X_{j,l}$ is always 0 for $i \neq j$.