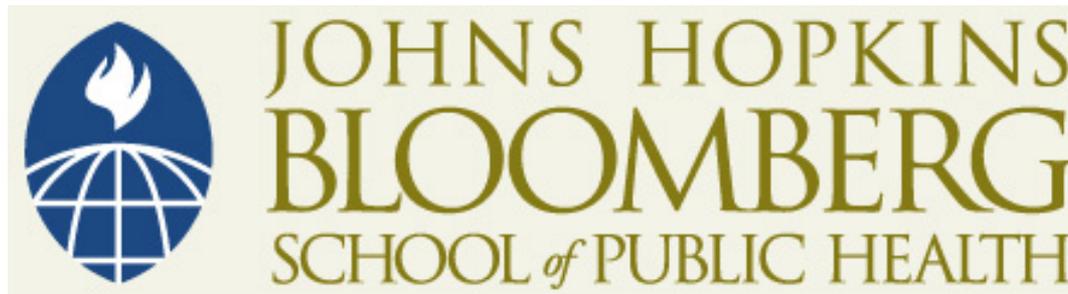


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2008, The Johns Hopkins University and Brian Caffo. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

# Outline

1. Describe multivariate Bernoulli trials
2. Motivate the multinomial distribution

# Multinomial density

- The multinomial distribution is a generalization of the binomial distribution where each trial can take several levels, rather than just two
- To illustrate the multinomial distribution, consider drawing with replacement from an urn with blue, red and yellow balls

$(1, 0, 0)$  for getting a blue ball

$(0, 1, 0)$  for getting a red ball

$(0, 0, 1)$  for getting a yellow ball

## Continued

- Assume that the proportion of blue, red and yellow balls in the urn are  $\pi_1, \pi_2$  and  $\pi_3$
- $\pi_1 + \pi_2 + \pi_3 = 1$

$$P\{X_i = (1, 0, 0)\} = \pi_1$$

$$P\{X_i = (0, 1, 0)\} = \pi_2$$

$$P\{X_i = (0, 0, 1)\} = \pi_3.$$

- Notice that one of the numbers is redundant. For example, if we know that the outcome is neither a blue or a red ball, then it must have been a yellow ball
- We might call this generalization **the multivariate Bernoulli distribution**

## Continued

- Just as a binomial random variable is the sum of iid Bernoulli trials, the multinomial distribution is the sum of iid multivariate Bernoulli trials
- Therefore, in our urn example, if you sum up  $n$  multivariate Bernoullis from this experiment, you get a vector that looks like

$$(n_1, n_2, n_3)$$

where  $n_1$ ,  $n_2$  and  $n_3$  are the number of blue, red and yellow balls (and  $n_1 + n_2 + n_3 = n$ )

# Some properties

- Mass function

$$P\{(N_1, \dots, N_k) = (n_1, \dots, n_k)\} = \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k \pi_i^{n_i}$$

- $E[N_i] = n\pi_i$
- $\text{Var}(N_i) = n\pi_i(1 - \pi_i)$
- $\text{Cov}(N_i, N_j) = -n\pi_i\pi_j$
- The maximum likelihood estimate of  $\pi_i$  is  $N_i/n$ ; that is the sample proportion remains the best estimate of the population proportion.