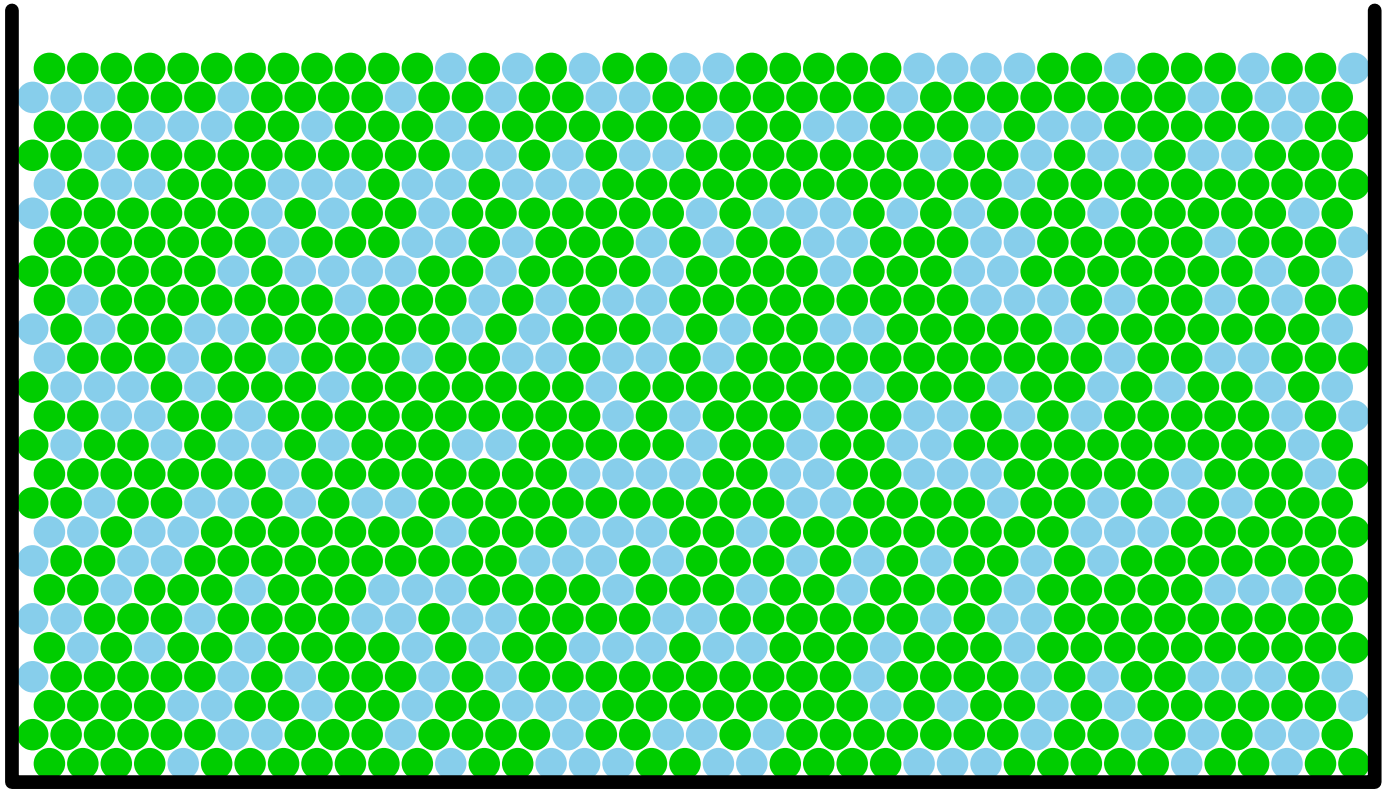
3.2 Hypergeometric Distribution
3.5, 3.9 Mean and Variance

Prof. Tesler

Math 186
Winter 2017

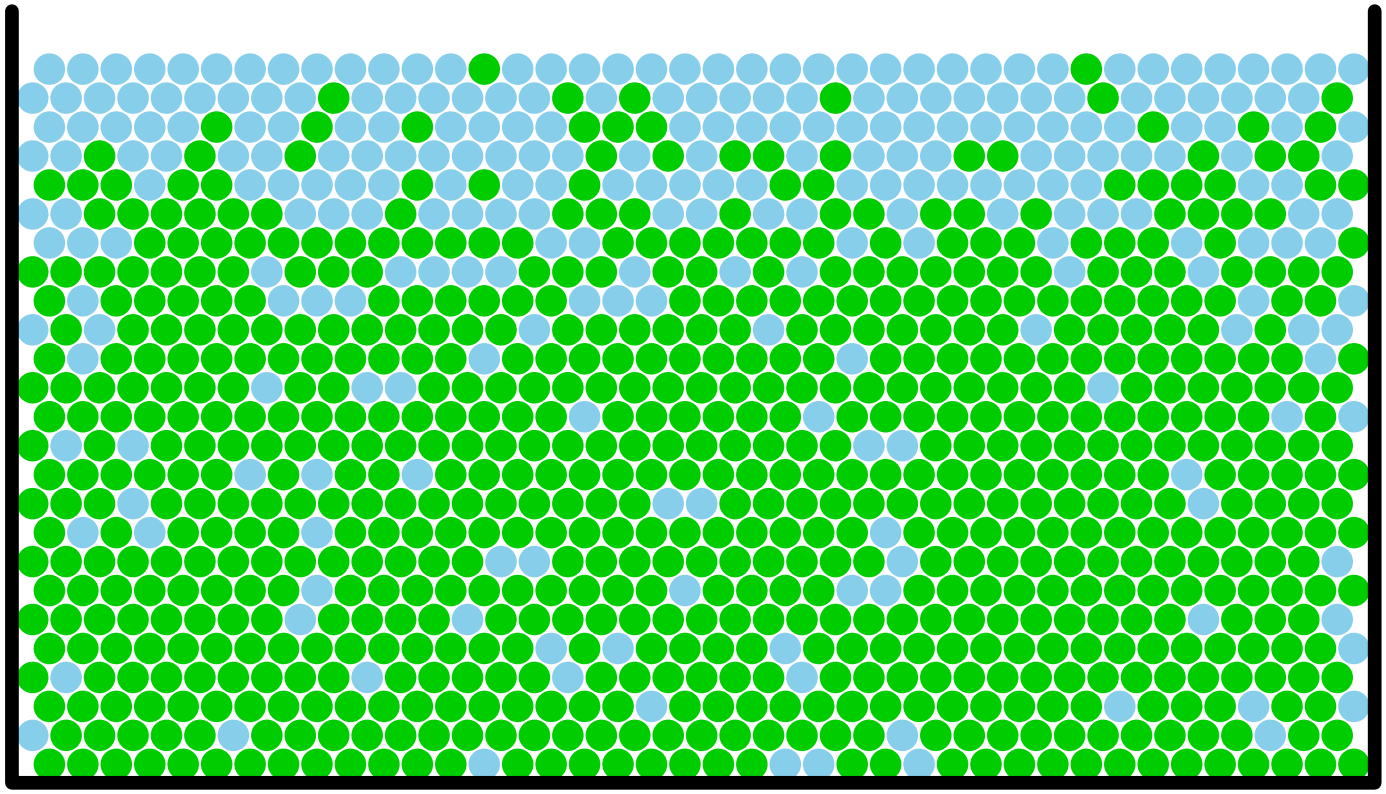
Sampling from an urn



- An urn has 1000 balls: 700 green, 300 blue.
- Pick a ball at random. The probability it's green is

$$p = 700/1000 = 0.7.$$

Sampling from an urn



- An urn has 1000 balls: 700 green, 300 blue.
- The urn needs to be well-mixed. Here, if you pick from the top, the chance of blue is much higher than in the total population.

Sampling with and without replacement

A urn has 1000 balls: 700 green, 300 blue.

Sampling with replacement

- Pick one of the 1000 balls. Record color (green or blue).
  Put it back in the urn and shake it up.
  Again pick one of the **1000** balls and record color.
  Repeat n times.
- On each draw, the probability of green is $700/1000$.
- The # green balls drawn has a binomial distribution, $p = \frac{700}{1000} = .7$

Sampling without replacement

- Pick one of the 1000 balls, record color, and set it aside.
  Pick one of the remaining **999** balls, record color, set it aside.
  Pick one of the remaining **998** balls, record color, set it aside.
  Repeat n times, never re-using the same ball.
- Equivalently, take n balls all at once and count them by color.
- The # green balls drawn has a *hypergeometric distribution*.

Sampling with and without replacement

A urn has 1000 balls: 700 green, 300 blue.
A sample of 7 balls is drawn.
What is the probability that it has 3 green balls and 4 blue balls?

Sampling with replacement

- Each draw has the same probability to be green: $p = \frac{700}{1000} = 0.7$
- $P(\text{3 green \& 4 blue}) = \binom{7}{3} p^3 (1-p)^4 = \binom{7}{3} (0.7)^3 (0.3)^4 = 0.0972405$

Sampling without replacement

- # samples with 3 green balls and 4 blue balls: $\binom{700}{3} \cdot \binom{300}{4}$
- # samples of size 7: $\binom{1000}{7}$
- $P(\text{3 green and 4 blue}) =$

$$\frac{\text{\# samples with 3 green and 4 blue}}{\text{\# samples of size 7}} = \frac{\binom{700}{3}\binom{300}{4}}{\binom{1000}{7}} \approx 0.0969179$$

Hypergeometric distribution

Exact distribution for sampling without replacement

Notation

Population (full urn)	Sample
N balls	n balls
K green	k green
$N-K$ blue	$n-k$ blue
$p = K/N$	$\hat{p} = k/n$

Hypergeometric distribution (for sampling w/o replacement)

- Draw n balls without replacement.
- Let random variable X be the number of green balls drawn.
- Its pdf is given by the *hypergeometric distribution*

$$P(X=k) = \binom{K}{k}\binom{N-K}{n-k} \Big/ \binom{N}{n}$$

- $E(X) = np$ and $\mathrm{Var}(X) = \frac{np(1-p)(N-n)}{(N-1)}$.

Sampling without replacement (2nd method)

A urn has 1000 balls: 700 green (G), 300 blue (B).

What is the probability to draw 7 balls in the order GBBGBGB?

- $P(1^{\text{st}} \text{ is G}) = 700/1000$
- $P(2^{\text{nd}} \text{ is B} | 1^{\text{st}} \text{ is G}) = 300/999$
- $P(3^{\text{rd}} \text{ is B} | \text{first two are G,B}) = 299/998$

$$P(\text{GBBGBGB}) = \frac{700}{1000} \cdot \frac{300}{999} \cdot \frac{299}{998} \cdot \frac{699}{997} \cdot \frac{298}{996} \cdot \frac{698}{995} \cdot \frac{297}{994}$$
$$= \frac{(700 \cdot 699 \cdot 698)(300 \cdot 299 \cdot 298 \cdot 297)}{1000 \cdot 999 \cdot 998 \cdot 997 \cdot 996 \cdot 995 \cdot 994}$$

Probability a sample of size 7 has 3 green and 4 blue

- Each sequence of 3 G's and 4 B's has that same probability; numerator factors are in a different order, but the result is equal.
- Adding probabilities of all $\binom{7}{3}$ sequences of 3 G's and 4 B's gives

$$P(3 \text{ G's \& } 4 \text{ B's}) = \binom{7}{3} \frac{(700 \cdot 699 \cdot 698)(300 \cdot 299 \cdot 298 \cdot 297)}{1000 \cdot 999 \cdot 998 \cdot 997 \cdot 996 \cdot 995 \cdot 994}$$

Sampling without replacement

Equivalence of both methods, and approximation by binomial distribution

A urn has 1000 balls: 700 green, 300 blue.
A sample of 7 balls is drawn, without replacement.
What is the probability that it has 3 green balls and 4 blue balls?

Probability with hypergeometric distribution (1st method)

$$= \frac{\binom{700}{3}\binom{300}{4}}{\binom{1000}{7}} = \frac{\frac{700\cdot 699\cdot 698}{3!}\cdot\frac{300\cdot 299\cdot 298\cdot 297}{4!}}{\frac{1000\cdot 999\cdot 998\cdot 997\cdot 996\cdot 995\cdot 994}{7!}}$$

$$= \frac{7!}{3!\,4!}\cdot\frac{700\cdot 699\cdot 698}{1000\cdot 999\cdot 998}\cdot\frac{300\cdot 299\cdot 298\cdot 297}{997\cdot 996\cdot 995\cdot 994}$$

2nd method to compute hypergeometric distribution

$$\approx \binom{7}{3}(700/1000)^3(300/1000)^4$$

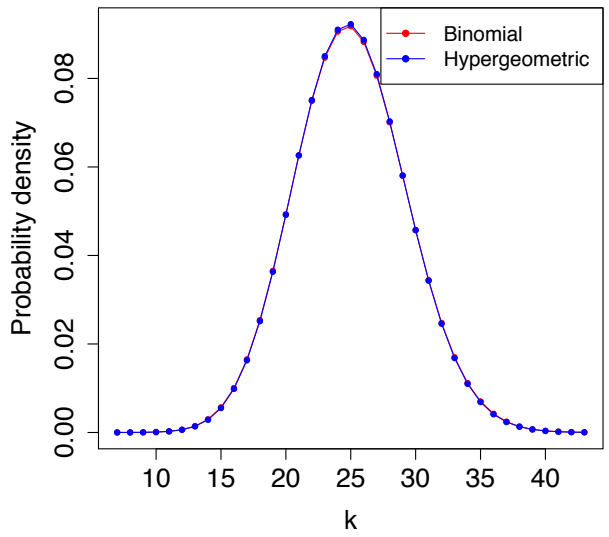
Probability with binomial distribution

If the numbers of green, blue, and total balls in the sample are much smaller than in the urn, the hypergeometric pdf $\approx$ the binomial pdf.

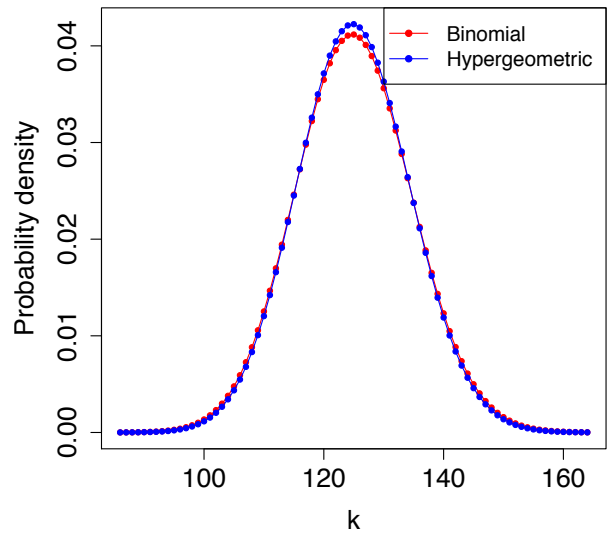
Hypergeometric distribution vs. Binomial distribution

$p = 0.25$, Population size $N = 10000$

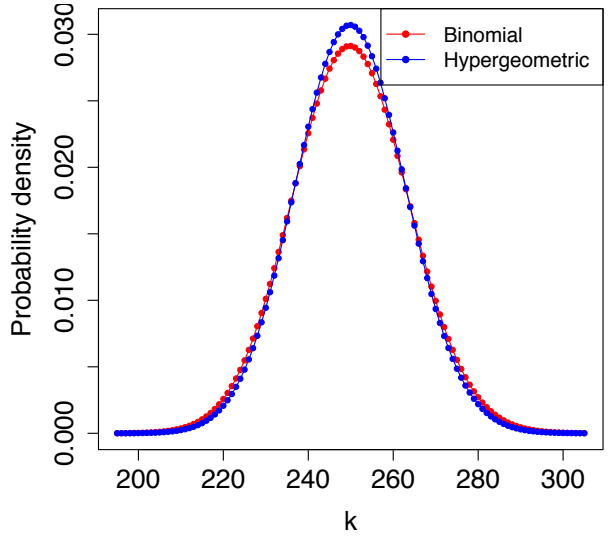
trials $n = 100$, $n/N = 1\%$



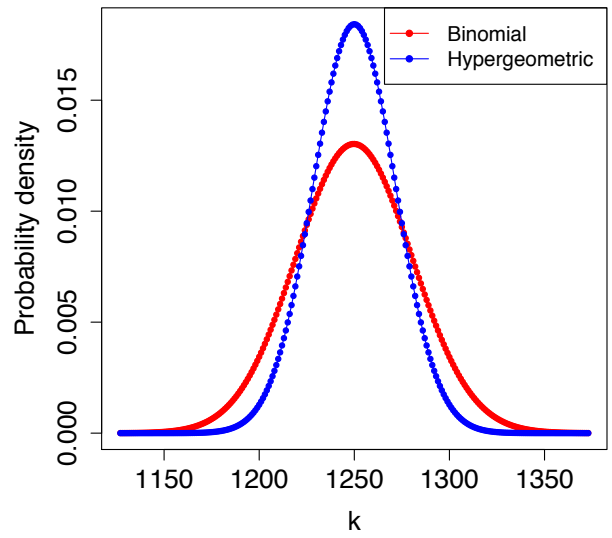
trials $n = 500$, $n/N = 5\%$



trials $n = 1000$, $n/N = 10\%$

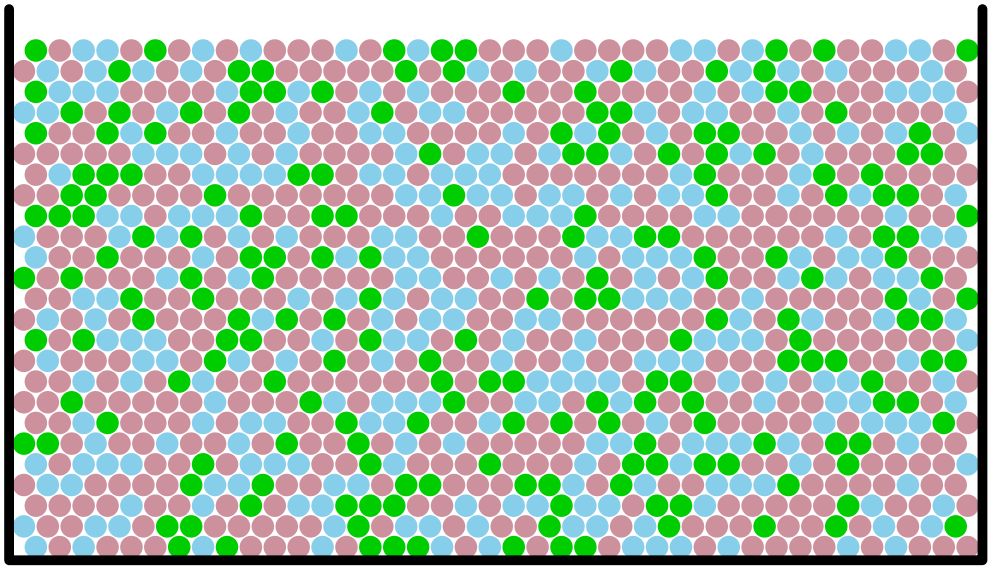


trials $n = 5000$, $n/N = 50\%$



Multihypergeometric distribution

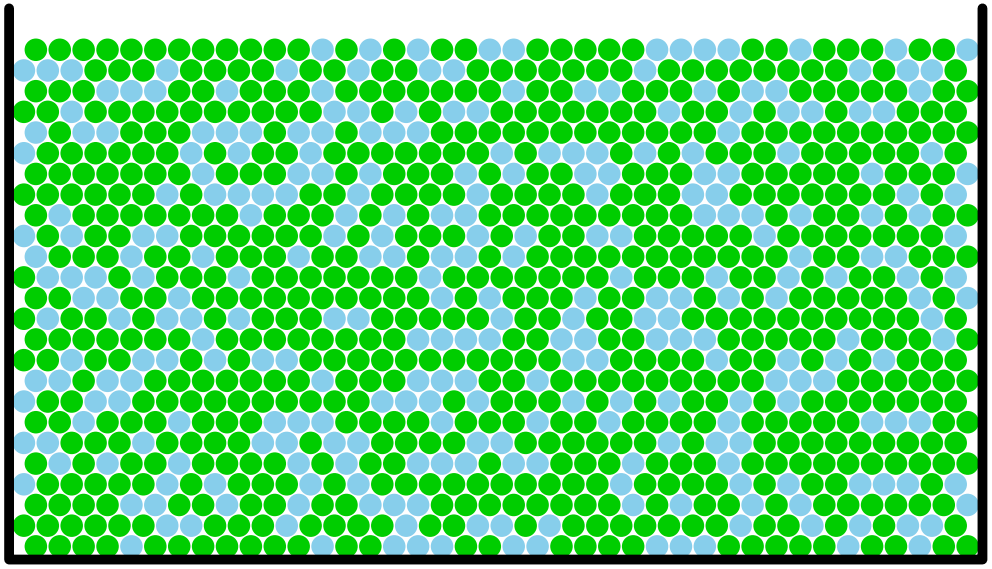
Sampling without replacement from an urn with multiple colors
Illustrated for 3 colors, but works for any number of colors



- An urn has 1000 balls: 500 red, 300 blue, 200 green.
- Draw a sample of 10 balls without replacement.
- What is the probability it has 2 red, 3 blue, and 5 green balls?

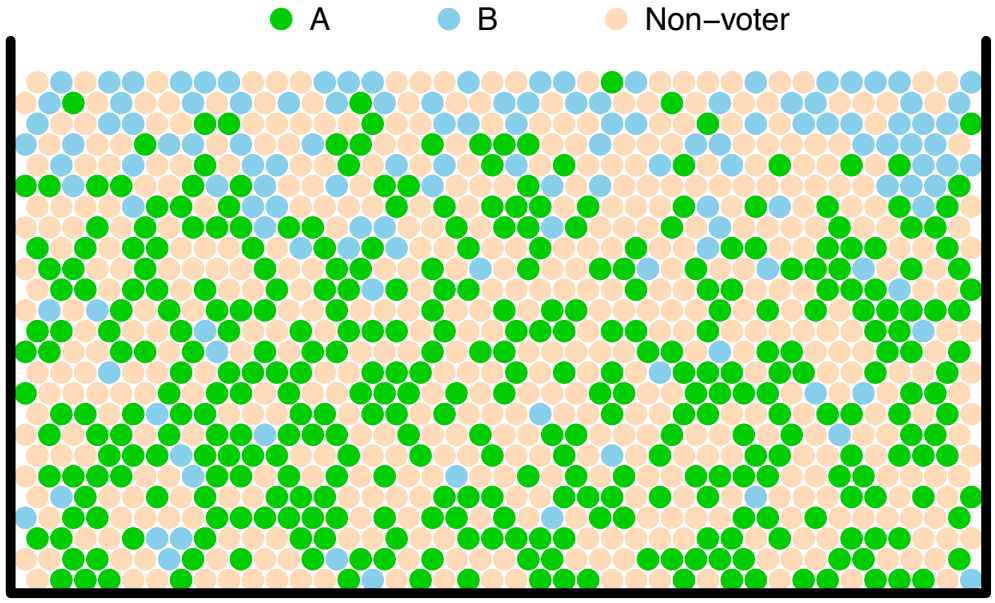
$$\frac{\text{\# samples with 2 red, 3 blue, 5 green}}{\text{\# samples of size 10}} = \frac{\binom{500}{2}\binom{300}{3}\binom{200}{5}}{\binom{1000}{10}} \approx 0.00535$$

Election polls



- A poll is taken before an election to estimate the fraction voting for each option.
- Should sample w/o replacement (hypergeometric distribution), to avoid polling the same person twice.
- If the sample size is much smaller than the population size, can approximate by binomial distribution.

Election polls



Complications in using balls in an urn to model a poll include:

- Sample may have non-voters (light color above).
- Sample may not be representative.
  Polling based on geography, landlines, cellphones, etc. may give different proportions in sample than in population.
  Above: more B's than A's at the top, but more A's than B's overall.
- Respondents may not reply, may not tell the truth, may change their minds by the time of the election, . . .

3.5 Expected value of hypergeometric distribution

- Let $p = K/N$ be the fraction of balls in the urn that are green.
- Draw a sample of n balls without replacement.
- For $i = 1, \ldots, n$, let $X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ ball is green;} \\ 0 & \text{otherwise.} \end{cases}$
- The total number of green balls in the sample is $X = X_1 + \cdots + X_n$.
- The X_i's are identically distributed, but dependent. For each ball:

$$P(X_i = 1) = K/N = p$$
$$P(X_i = 0) = 1 - p \qquad E(X_i) = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = p$$

This does not use info about the other balls.
It is ***not*** a conditional probability, such as $P(X_3 = 1 | X_1 = a, X_2 = b)$.

- $E(X) = E(X_1) + \cdots + E(X_n) = \underbrace{p + p + \cdots + p}_{n} = np = nK/N.$
- We calculated $E(X)$ this way for the binomial distribution too!
  Dependence between X_i's isn't an issue for $E(X)$, but is for $\mathrm{Var}(X)$.

3.9 Variance of hypergeometric distribution

- X_i's are dependent, so variance isn't additive. Instead, use:

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \leqslant i < j \leqslant n} \mathrm{Cov}(X_i, X_j) \ .$$

(Generalizing $\mathrm{Var}(U+V) = \mathrm{Var}(U) + \mathrm{Var}(V) + 2\,\mathrm{Cov}(U,V)$.)

- $\mathrm{Var}(X_i) = E(X_i^2) - (E(X_i))^2$:
  - Since $X_i = 0$ or 1, we have $X_i^2 = X_i$. Thus, $E(X_i^2) = E(X_i)$, so

$$\mathrm{Var}(X_i) = E(X_i) - (E(X_i))^2 = p - p^2 = \frac{K}{N} - \frac{K^2}{N^2} = \frac{K(N-K)}{N^2}.$$

- For $i < j$, $\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$:
  - $X_i X_j = \begin{cases} 1 & \text{if } X_i = X_j = 1; \\ 0 & \text{otherwise.} \end{cases}$ so $P(X_i X_j{=}1) = P(X_i{=}X_j{=}1) = \frac{K(K-1)}{N(N-1)}$.
  - $E(X_i X_j) = 1P(X_i X_j = 1) + 0P(X_i X_j = 0) = P(X_i X_j = 1) = \frac{K(K-1)}{N(N-1)}$.
  - $\mathrm{Cov}(X_i, X_j) = \frac{K(K-1)}{N(N-1)} - (K/N)^2 = \frac{K(K-N)}{N^2(N-1)}$.

Variance of hypergeometric distribution

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{Var}(X_1 + \cdots + X_n) \\
&= \underbrace{\sum_{i=1}^{n}}_{n \text{ terms}} \underbrace{\operatorname{Var}(X_i)}_{=\frac{K(N-K)}{N^2}} + 2 \underbrace{\sum_{1 \leqslant i < j \leqslant n}}_{\binom{n}{2} \text{ terms}} \underbrace{\operatorname{Cov}(X_i, X_j)}_{=\frac{K(K-N)}{N^2(N-1)}} \\
&= \frac{n\,K(N-K)}{N^2} + \frac{2n(n-1)}{2} \frac{K(K-N)}{N^2(N-1)} \\
&= \frac{n\,K(N-K)}{N^2} \left(1 - \frac{n-1}{N-1}\right) = \boxed{\frac{\boldsymbol{n\,K(N-K)(N-n)}}{\boldsymbol{N^2(N-1)}}}
\end{aligned}$$

Using $p = K/N$, substitute $K = Np$ to rewrite this as

$$\operatorname{Var}(X) = \boxed{\boldsymbol{np(1-p)\frac{N-n}{N-1}}}$$