HYPERGEOMETRIC DISTRIBUTION

Jerzy Szulga

Department of Mathematics and Statistics
Auburn University

MATH 5670-6670 FALL 2019

PROBABILITY I

September 25, 2019

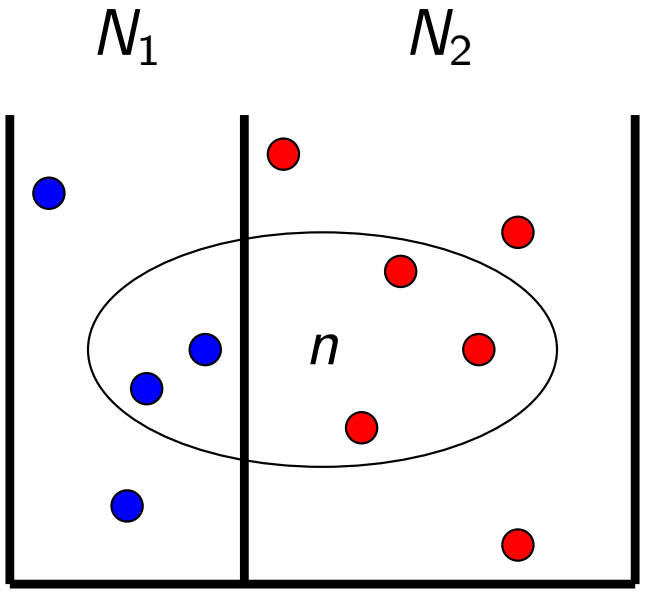
Quasi Bernoulli trials

Return to the example of a medical staff examining potential patients in regard to the presence of a disease.

That is, there are N_1 infected people in a population of size N, and $N_2 = N - N_1$ are healthy. Thus the prevalence of infection is represented by a percentage or fraction $p = N_1/N$. A randomly selected person is examined, and the outcomes are 1 (infected) or 0 (not infected). Upon proper procedure, this person is not examined again. The sampling goes without replacement.

sampling from an urn

Without replacement, with order (one by one) or without (at once).



$N = N_1 + N_2$, here $10 = 4 + 6$

a sample of size n, here $n = 5$

#blue balls $B = x$, here $x = 2$

Check your intuition

Exercise 1. [Russian roulette]. A revolver has one bullet in one of 6 chambers. The cylinder is revolved only once.

Among 6 players, would you prefer to be the first or the last one?

Check your intuition

Exercise 1. [Russian roulette]. A revolver has one bullet in one of 6 chambers. The cylinder is revolved only once.

Among 6 players, would you prefer to be the first or the last one?

Exercise 2. When a card is drawn from a standard deck, then $\mathrm{P}(\mathit{ace}) = \frac{4}{52} = \frac{1}{13}$. Now, cards are drawn without replacement.

What is the probability that the second card is an ace?

The third? The thirteenth? The 52^{nd}?

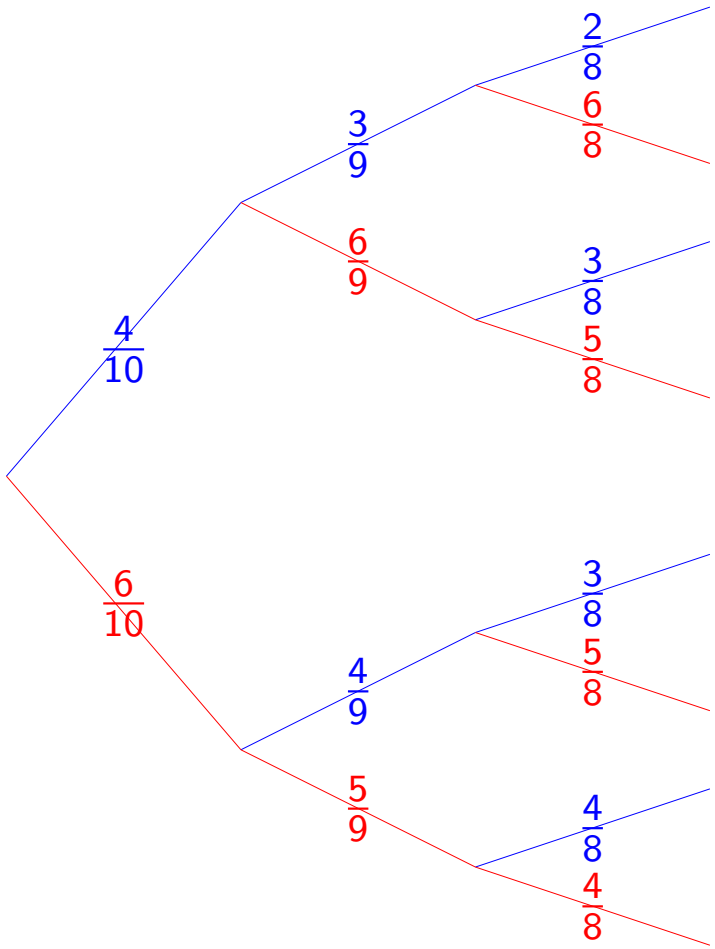
Interchangeability

Let $B_1, \ldots, B_n$ be a quasi-Bernoulli process, representing consecutive sampling from an un urn of N items. There are N_1 items marked by "1" (e.g., blue balls) and $N_2 = N - N_1$ items marked by "0" (e.g., red balls).

For example, $\{ B_1 = 1,\ B_2 = 0,\ B_3 = 1 \}$ means that the first drawn ball is blue, the second is red, and the third is blue again. Here, of course, $N \geq 3$. Surprisingly (or may be not), for any three "colors" $c_1, c_2, c_3 \in \{ 0, 1 \}$ and any permutation (i, j, k) of $(1, 2, 3)$:

$$\mathrm{P}(B_1 = c_1,\ B_2 = c_2,\ B_3 = c_3) = \mathrm{P}(B_1 = c_i,\ B_2 = c_j,\ B_3 = c_k).$$

Checking for $N = 10, N_1 = 4, n = 3$



$$\frac{4\cdot 3\cdot 6}{10\cdot 9\cdot 8} = \mathrm{P}(\mathrm{BBR})$$

$$\frac{4\cdot 6\cdot 3}{10\cdot 9\cdot 8} = \mathrm{P}(\mathrm{BRB})$$

$$\frac{6\cdot 4\cdot 3}{10\cdot 9\cdot 8} = \mathrm{P}(\mathrm{RBB})$$

A "philosphical argument"

Q. Why all permutations do have the same probability?

Q. Why it doesn't matter - probabilistically, not psychologically - who goes first or who goes last in the Russian roulette?

A "philosphical argument"

Q. Why all permutations do have the same probability?

Q. Why it doesn't matter - probabilistically, not psychologically - who goes first or who goes last in the Russian roulette?

A. Because the act of randomness (sampling) has nothing to do with the non-random act of revealing or ordering "1^{st}", "2^{nd}", "3^{rd}", etc.

The latter is just a label.

the counter and the mean

We may mark the items of interest by 1 and the other by 0, yielding Bernoulli RVs B_1, B_2, ... with $p = N_1/N$.

They are **identically distributed** but **dependent**.

The number of ones in an ordered sample of size n

$$S_n = B_1 + \cdots + B_n.$$

(Same as in the binomial.)

By additivity, $\mathrm{E}\, S_n = np$, no need for the pmf.

the second moment

$$\mathrm{E}\, S_n^2 = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right) = \sum_i \mathrm{E}\,(X_i^2) + \sum_{i\neq j} \mathrm{E}\,(X_i X_j),$$

by additivity. Since all permutations have the same distribution and $X_i^2 = X_i$ (squaring 0 or 1), we have

$$\mathrm{E}\,(S_n^2) = n\frac{N_1}{N} + n(n-1)\mathrm{E}\,(X_1 X_2)$$

$X_1 X_2$ is again a Bernoulli 0-1 variable with

$$p' = \mathrm{P}(X_1 X_2 = 1) = \mathrm{P}(X_1 = 1,\, X_2 = 1) = \frac{N_1}{N} \cdot \frac{N_1 - 1}{N - 1}.$$

variance

Plugging in,

$$\mathrm{E}\,(S_n^2) = n\,\frac{N_1}{N} + n(n-1)\,\frac{N_1}{N}\,\frac{N_1-1}{N-1},$$

Hence, a little tedious but amazing algebra entails the variance:

$$\mathrm{Var}(S_n) = \mathrm{E}\,(S_n^2) - (\mathrm{E}\,S_n)^2 = \frac{nN_1N_2(N-n)}{N^2(N-1)}.$$

Again, all computations have been conducted without pmf.

No wonder why the mgf is not displayed. It's too complicated.

pmf

In computation of mean or variance the order or no order is irrelevant since the addition is commutative and the number of ones is their sum.

$$f(x) = \mathrm{P}(S_n = x) = \frac{\binom{N_1}{x} \cdot \binom{N_2}{n-x}}{\binom{N}{n}}, \quad x = 0, \ldots, \min(n, N_1).$$

Why the name "hypergeometric"

It's a semantic fossil. John Wallis, a 17^{th} century British clergyman and mathematician, also the designer of the infinity symbol

$$\infty,$$

called the sequence $1 \cdot 2 \cdot \cdots \cdot n$ **hypergeometric**, quite logically.

Yer the term felt out of use and was replaced by **factorial**, senseless.

Yet, it persisted in the hypergeometric formula which has plenty of factorials (9), or "hypergeometric sequences".

Order

We wonder what is the probability that there are x blue balls among n sampled when they appeared in the specific positions (e.g.,the first, the third, the fifth, etc.).

The positions don't matter. It could be just the first x balls.

$$\mathrm{P}(S_n = x, \text{on specific positions}) = \frac{(N_1)_x\,(N_2)_{n-x}}{(N)_n} = \frac{f(x)}{\binom{n}{x}}.$$

The approximation by binomial

A hypergeometric distribution has three parameters: N, N_1, n.

When $n << N_1$ and $n << N_2$ while N_1/N stabilizes at level p, then the quasi Bernoulli trials become the ideal Bernoulli trials.

That is, the quite cumbersome hypergeometric pmf can be replaced by the much simpler binomial pmf.

This heuristic observation can be confirmed by rather tedious algebra and then routine calculus:

$$N_1 \to \infty, \quad N_2 \to \infty, \quad \frac{N_1}{N} \to p.$$