# CS 107, Probability, Spring 2020
# Lecture 25

Michael Poghosyan
mpoghosyan@aua.am

AUA

03 April 2020

# Content

- Some important Discrete rv.s
  - Negative Binomial Distribution
  - Hypergeometric Distribution

# So far ...

So far

- We defined and studied the general theory and properties of Discrete rv.s
- We defined and studied the general theory and properties of Continuous rv.s
- We introduced and studied some important Discrete Distributions:
  - Discrete Uniform Distribution;
  - Bernoulli Distribution;
  - Binomial Distribution;
  - Geometric Distribution;
  - Poisson Distribution

# Negative Binomial Distribution

# Negative Binomial Distribution

**Distribution Name:** *NBinom*, *NegBinom*;
**Parameters:** $m, p$ $(m \in \mathbb{N}, p \in (0,1])$

## Negative Binomial Distribution

We say that the r.v. $X$ has a Negative Binomial Distribution with parameters $m$ and $p$, and we will write $X \sim NBinom(m,p)$, if the range of $X$ is $\{m, m+1, m+2, ...\}$, and the PMF of $X$ is

$$\mathbb{P}(X=k) = \binom{k-1}{m-1} \cdot p^m \cdot (1-p)^{k-m}, \qquad k = m, m+1, m+2, ...$$

that is,

| Val's of $X$ | $m$ | $m+1$ | $m+2$ | ... | $k$ |
|---|---|---|---|---|---|
| $\mathbb{P}(X=x)$ | $p^m$ | $mp^m(1-p)^1$ | $\binom{m+1}{m-1}p^m(1-p)^2$ | ... | $\binom{k-1}{m-1}p^m(1-p)^{k-m}$ |

# Negative Binomial Distribution

**Distribution Name:** *NBinom*, *NegBinom*;
**Parameters:** $m, p \ (m \in \mathbb{N}, p \in (0,1])$

## Negative Binomial Distribution

We say that the r.v. $X$ has a Negative Binomial Distribution with parameters $m$ and $p$, and we will write $X \sim NBinom(m,p)$, if the range of $X$ is $\{m, m+1, m+2, ...\}$, and the PMF of $X$ is

$$\mathbb{P}(X=k) = \binom{k-1}{m-1} \cdot p^m \cdot (1-p)^{k-m}, \qquad k = m, m+1, m+2, ...$$

that is,

| Val's of $X$ | $m$ | $m+1$ | $m+2$ | ... | $k$ |
|---|---|---|---|---|---|
| $\mathbb{P}(X=x)$ | $p^m$ | $mp^m(1-p)^1$ | $\binom{m+1}{m-1}p^m(1-p)^2$ | ... | $\binom{k-1}{m-1}p^m(1-p)^{k-m}$ |

**Exercise:** Use $Calc2 \cup RA$ to prove that $\sum_{k=m}^{\infty} \mathbb{P}(X=k) = 1$.

# Negative Binomial Distribution

**Usage:** Here we are doing independent *Bernoulli*$(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes.

# Negative Binomial Distribution

**Usage:** Here we are doing independent *Bernoulli*$(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes. So:

- we have a simple Experiment, and an event in that Experiment, called Success;

# Negative Binomial Distribution

**Usage:** Here we are doing independent *Bernoulli*$(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes. So:

- we have a simple Experiment, and an event in that Experiment, called Success;
- the probability of Success is $p$, $p \in (0, 1]$;

# Negative Binomial Distribution

**Usage:** Here we are doing independent *Bernoulli*$(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes. So:

- we have a simple Experiment, and an event in that Experiment, called Success;
- the probability of Success is $p$, $p \in (0, 1]$;
- we repeat independently this Experiment until achieving the $m$-th Success;

# Negative Binomial Distribution

**Usage:** Here we are doing independent *Bernoulli*$(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes. So:

- we have a simple Experiment, and an event in that Experiment, called Success;
- the probability of Success is $p$, $p \in (0, 1]$;
- we repeat independently this Experiment until achieving the $m$-th Success;
- our Negative Binomial r.v. $X$ is showing how many times we have repeated our Experiment (no. of trials).

# Negative Binomial Distribution

**Usage:** Here we are doing independent $Bernoulli(p)$ trials (i.e., trials with Success/Failures), and $X$, our Negative Binomial r.v. is showing how many times we need to do that trials to have the first $m$ Successes. So:

- we have a simple Experiment, and an event in that Experiment, called Success;
- the probability of Success is $p$, $p \in (0, 1]$;
- we repeat independently this Experiment until achieving the $m$-th Success;
- our Negative Binomial r.v. $X$ is showing how many times we have repeated our Experiment (no. of trials).

**Note:** Clearly, the Geometric Distribution is the particular case of the Negative Binomial Distribution, with $m = 1$:

$$Geom(p) = NBinom(1, p).$$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m, p)$: for $k \geq m$,
$\mathbb{P}(X = k) =$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m, p)$: for $k \geq m$,
$\mathbb{P}(X = k) =$

$$= \mathbb{P}(m - 1 \text{ Successes in } k - 1 \text{ trials and } k\text{-th one was Success}) =$$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m, p)$: for $k \geq m$,

$\mathbb{P}(X = k) =$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials and } k\text{-th one was Success}) =$$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials}) \cdot \mathbb{P}(k\text{-th one was Success}) =$$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m, p)$: for $k \geq m$,

$\mathbb{P}(X = k) =$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials and } k\text{-th one was Success}) =$$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials}) \cdot \mathbb{P}(k\text{-th one was Success}) =$$

$$= \binom{k-1}{m-1} \cdot p^{m-1} \cdot (1-p)^{k-m} \cdot p =$$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m, p)$: for $k \geq m$,
$\mathbb{P}(X = k) =$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials and } k\text{-th one was Success}) =$$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials}) \cdot \mathbb{P}(k\text{-th one was Success}) =$$

$$= \binom{k-1}{m-1} \cdot p^{m-1} \cdot (1-p)^{k-m} \cdot p = \binom{k-1}{m-1} \cdot p^m \cdot (1-p)^{k-m}$$

# Negative Binomial Distribution

Let's explain how to get the PMF of $X \sim NBinom(m,p)$: for $k \geq m$,

$\mathbb{P}(X = k) =$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials and } k\text{-th one was Success}) =$$

$$= \mathbb{P}(m-1 \text{ Successes in } k-1 \text{ trials}) \cdot \mathbb{P}(k\text{-th one was Success}) =$$

$$= \binom{k-1}{m-1} \cdot p^{m-1} \cdot (1-p)^{k-m} \cdot p = \binom{k-1}{m-1} \cdot p^m \cdot (1-p)^{k-m}$$

**Note:** The Negative Binomial r.v. can be viewed as the waiting time (in the sense of the number of trials) for the $m$-th success.

# Negative Binomial Distribution

**Example 25.1:** Assume $X \sim NBinom(2, 0.3)$. Explain what is representing/counting $X$.

# Negative Binomial Distribution

**Example 25.1:** Assume $X \sim NBinom(2, 0.3)$. Explain what is representing/counting $X$.

**Example 25.2:** Assume $X \sim NBinom(4, 0.5)$.

a. Calculate $\mathbb{P}(X = 3)$;
b. Calculate $\mathbb{P}(X = 5)$;
c. Calculate $\mathbb{P}(X > 5)$.

# Negative Binomial Distribution

**Example 25.3:** Assume we have a box full of 10 white and 5 red balls, and I am picking a ball at random, one at a time, with replacements, until having a red ball 3 times. What is the probability that I will pick a ball $10$ times?

**Solution:** OTB

# Negative Binomial Distribution

**Example 25.4:** Assume that the probability that the child born will be a boy child is $0.5$. A recently married couple is planning to have 3 boys, and stop having children after the 3rd boy.

a. What is the probability that they will have $6$ children?
b. What is the probability that they will have more than $6$ children?

# Negative Binomial Distribution

**Example 25.5:** Generate in **R** 10 Negative Binomial distributed random numbers, with parameters $m = 3$ and $p = 0.4$. Explain what that numbers are representing.

**Note:** Please note that in **R**, the $NBinom(m, p)$ is calculating *the number of failures before having exactly m successes*. So, in order to get a $NBinom(m, p)$ random number in our definition sense, you need to add $m$ to that **R**-generated number.

# Negative Binomial Distribution

**Example 25.6:** Assume $X \sim NBinom(3, 0.6)$.

a. Calculate, in **R**, $\mathbb{P}(X = 7)$;
b. Calculate, in **R**, $\mathbb{P}(X \text{ is even})$;
c. Plot, in **R**, the PMF of $X$.

# Hypergeometric Distribution

# Hypergeometric Distribution

**Distribution Name:** *Hyper*, *HyperGeom*;
**Parameters:** $N, m, n$ $(N, m, n \in \mathbb{N},\ m \leq N, n \leq N)$

## Hypergeometric Distribution

Assume we have a box full of $N$ balls, from which $m$ are white and $N - m$ are black. We choose $n$ balls from that box at random, **without replacements**, and the order is not important.

# Hypergeometric Distribution

**Distribution Name:** *Hyper*, *HyperGeom*;
**Parameters:** $N, m, n$ $(N, m, n \in \mathbb{N},\ m \leq N, n \leq N)$

## Hypergeometric Distribution

Assume we have a box full of $N$ balls, from which $m$ are white and $N - m$ are black. We choose $n$ balls from that box at random, **without replacements**, and the order is not important. Let $X$ denote the number of white balls. Then

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}, \qquad k = 0, 1, 2, ..., n.$$

# Hypergeometric Distribution

**Distribution Name:** *Hyper*, *HyperGeom*;
**Parameters:** $N, m, n$ $(N, m, n \in \mathbb{N},\ m \le N, n \le N)$

## Hypergeometric Distribution

Assume we have a box full of $N$ balls, from which $m$ are white and $N - m$ are black. We choose $n$ balls from that box at random, **without replacements**, and the order is not important. Let $X$ denote the number of white balls. Then

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}, \qquad k = 0, 1, 2, ..., n.$$

This r.v. $X$ is called Hypergeometric r.v. with parameters $N, m, n$, and this is denoted by $X \sim Hyper(N, m, n)$.

# Hypergeometric Distribution

**Note:** In order to have the above definition work for any $k$, we need to assume that $n \leq m$ and $n - k \leq N - m$.

# Hypergeometric Distribution

**Note:** In order to have the above definition work for any $k$, we need to assume that $n \leq m$ and $n - k \leq N - m$. Another way of guaranteeing that we can calculate the above probabilities is to define binomial coefficients

$$\binom{s}{t} = 0, \qquad \textit{if} \qquad t < 0 \quad \textit{or} \quad s < t.$$

# Hypergeometric Distribution

**Note:** In order to have the above definition work for any $k$, we need to assume that $n \leq m$ and $n - k \leq N - m$. Another way of guaranteeing that we can calculate the above probabilities is to define binomial coefficients

$$\binom{s}{t} = 0, \qquad \textit{if} \qquad t < 0 \quad \textit{or} \quad s < t.$$

**Note:** Sometimes, in the literature, you will find the definition of a Hypergeometric r.v. in the form $X \sim Hyper(m_w, m_b, n)$, where $m_w$ is the number of white balls, $m_b$ is the number of black balls in the box (this is the case also in **R**). So, you can take $N = m_w + m_b$, and we will obtain the same as our definition.

# Hypergeometric Distribution

**Note:** In order to have the above definition work for any $k$, we need to assume that $n \leq m$ and $n - k \leq N - m$. Another way of guaranteeing that we can calculate the above probabilities is to define binomial coefficients

$$\binom{s}{t} = 0, \qquad \textit{if} \qquad t < 0 \quad \textit{or} \quad s < t.$$

**Note:** Sometimes, in the literature, you will find the definition of a Hypergeometric r.v. in the form $X \sim Hyper(m_w, m_b, n)$, where $m_w$ is the number of white balls, $m_b$ is the number of black balls in the box (this is the case also in **R**). So, you can take $N = m_w + m_b$, and we will obtain the same as our definition. And sometimes you can find the definition of the Hypergeometric r.v. in the form $X \sim Hyper(N, p, n)$, where $p$ is the proportion of white balls. In this case we can just take $m = p \cdot N$ to reduce to our model.

# Hypergeometric Distribution

**Exercise:** Prove that

$$\sum_{k=0}^{n} \mathbb{P}(X = k) = 1.$$

**Hint:** You can use the binomial expansion of

$$(1+x)^m \cdot (1+x)^{N-m} = (1+x)^N.$$

# Hypergeometric Distribution

**Example 25.7:** Assume $X \sim \mathit{Hyper}(20, 8, 5)$.

a. Explain what $X$ is showing/counting;
b. Explain what $X = 2$ means;
c. Calculate $\mathbb{P}(X = 2)$.

# Hypergeometric Distribution

**Example 25.8:** Student knows the answer to 20 questions out of the 50 total. For the exam, student needs to take at random 3 questions, and answer to at least 2 questions to pass the exam.

a. What is the probability that our student will pass the exam?
b. Now assume our student knows the answer for $m$ questions. Plot, in **R**, the graph of probabilities that our student will pass the exam vs $m$.

# Hypergeometric Distribution

**Example 25.9:** A supermarket is buying light bulbs in boxes. Each box is containing 300 light bulbs. The factory, producing that bulbs, claims that the no. of produced defectives is not exceeding $2\%$. To check this claim, supermarket manager is making a test: he is taking a box at random, choosing 20 light bulbs out of that box, and checking how many of them are working. The test results in 3 defectives. Does this support the claim of the producer?

# HyGeo Distribution: Capture - Recapture Method

Now, about one application of the Hypergeometric Distribution.

# HyGeo Distribution: Capture - Recapture Method

Now, about one application of the Hypergeometric Distribution. Have you ever wondered how to calculate the number of fishes in a lake? Or estimate the number of street dogs in some area?

# HyGeo Distribution: Capture - Recapture Method

Now, about one application of the Hypergeometric Distribution. Have you ever wondered how to calculate the number of fishes in a lake? Or estimate the number of street dogs in some area?

One of the methods is called the Capture-Recapture or the Mark and Recapture methods, a method, which is used in ecology to estimate the animal population size when it is not possible to count every individual. This method is used also in Epidemiology to estimate the size of persons having some disease.

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown.

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown. We want to get (estimate!) $N$ without counting every individual.

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown. We want to get (estimate!) $N$ without counting every individual. For that, the Capture-Recapture method suggest the following:

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown. We want to get (estimate!) $N$ without counting every individual. For that, the Capture-Recapture method suggest the following:

- We capture $m$ individuals, mark (tag) them in some way, and then released back into the population;

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown. We want to get (estimate!) $N$ without counting every individual. For that, the Capture-Recapture method suggest the following:

- We capture $m$ individuals, mark (tag) them in some way, and then released back into the population;
- At a later time, assuming that that marked individuals are well mixed in the population, we capture again $n$ individuals from the population.

# HyGeo Distribution: Capture - Recapture Method

The two-sample version of the method is the following: assume the size of the population is $N$, and it is unknown. We want to get (estimate!) $N$ without counting every individual. For that, the Capture-Recapture method suggest the following:

- We capture $m$ individuals, mark (tag) them in some way, and then released back into the population;
- At a later time, assuming that that marked individuals are well mixed in the population, we capture again $n$ individuals from the population.
- Assume $k$ is the number of marked individuals in our second sample of size $n$. Then the most probable value of $N$ is
$$\widehat{N} = \left[\frac{m \cdot n}{k}\right].$$

To get this result, we consider the number $X$ of possible marked individuals in the second sample, with the size $n$: clearly,

$$X \sim Hyper(N, m, n),$$

To get this result, we consider the number $X$ of possible marked individuals in the second sample, with the size $n$: clearly,

$$X \sim Hyper(N, m, n),$$

and the Probability of having exactly $k$ marked individuals is

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

# HypGeo Distribution: Capture - Recapture Method

To get this result, we consider the number $X$ of possible marked individuals in the second sample, with the size $n$: clearly,

$$X \sim Hyper(N, m, n),$$

and the Probability of having exactly $k$ marked individuals is

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Now here, we do not know the value of $N$, but we have $k$.

# HypGeo Distribution: Capture - Recapture Method

To get this result, we consider the number $X$ of possible marked individuals in the second sample, with the size $n$: clearly,

$$X \sim Hyper(N, m, n),$$

and the Probability of having exactly $k$ marked individuals is

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Now here, we do not know the value of $N$, but we have $k$. We think in the following way: which is the most probable value of $N$, i.e., for which value of $N$, this probability will be the maximum?

To find that maximum, we consider the sequence $\mathbb{P}(X = k)$ for different values of $N$: let

$$a_N = \mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

# HypGeo Distribution: Capture - Recapture Method

To find that maximum, we consider the sequence $\mathbb{P}(X = k)$ for different values of $N$: let

$$a_N = \mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Now, $a_N$ is increasing up to some value, and then decreasing:

# HypGeo Distribution: Capture - Recapture Method

To find that maximum, we consider the sequence $\mathbb{P}(X = k)$ for different values of $N$: let

$$a_N = \mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Now, $a_N$ is increasing up to some value, and then decreasing: to show that, we calculate

$$\frac{a_N}{a_{N-1}} = \frac{(N - m) \cdot (N - n)}{N \cdot (N - m - n + k)} = \frac{N^2 - (n + m)N + mn}{N^2 - (n + m)N + kN}$$

# HypGeo Distribution: Capture - Recapture Method

To find that maximum, we consider the sequence $\mathbb{P}(X = k)$ for different values of $N$: let

$$a_N = \mathbb{P}(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Now, $a_N$ is increasing up to some value, and then decreasing: to show that, we calculate

$$\frac{a_N}{a_{N-1}} = \frac{(N-m)\cdot(N-n)}{N\cdot(N-m-n+k)} = \frac{N^2-(n+m)N+mn}{N^2-(n+m)N+kN}$$

Now, clearly,

$$\frac{a_N}{a_{N-1}} > 1, \quad \mathit{if}\, kN < mn, \qquad \mathit{and} \qquad \frac{a_N}{a_{N-1}} < 1, \quad \mathit{if}\, kN > mn.$$

So $a_N$ is increasing for $N < \dfrac{m \cdot n}{k}$, and is decreasing for $N > \dfrac{m \cdot n}{k}$, so the maximum value of $a_N$, i.e., for $\mathbb{P}(X = k)$, will be for

$$\widehat{N} = \left[\frac{m \cdot n}{k}\right].$$

# HypGeo Distribution: Capture - Recapture Method

So $a_N$ is increasing for $N < \dfrac{m \cdot n}{k}$, and is decreasing for $N > \dfrac{m \cdot n}{k}$, so the maximum value of $a_N$, i.e., for $\mathbb{P}(X = k)$, will be for

$$\widehat{N} = \left[\frac{m \cdot n}{k}\right].$$

**Interpretation:** In the second sample, among $n$ individuals, there were $k$ marked, so the proportion of the marked individuals in that sample is $p = \frac{k}{n}$.

# HypGeo Distribution: Capture - Recapture Method

So $a_N$ is increasing for $N < \dfrac{m \cdot n}{k}$, and is decreasing for $N > \dfrac{m \cdot n}{k}$, so the maximum value of $a_N$, i.e., for $\mathbb{P}(X = k)$, will be for

$$\widehat{N} = \left[\frac{m \cdot n}{k}\right].$$

**Interpretation:** In the second sample, among $n$ individuals, there were $k$ marked, so the proportion of the marked individuals in that sample is $p = \frac{k}{n}$. Now, if we assume that marked species were well-mixed in the population, we need to have the same proportion in the whole population, and since we had $N$ species (population size), with $m$ marked, we need to have $p = \frac{m}{N}$.

So $a_N$ is increasing for $N < \dfrac{m \cdot n}{k}$, and is decreasing for $N > \dfrac{m \cdot n}{k}$, so the maximum value of $a_N$, i.e., for $\mathbb{P}(X = k)$, will be for

$$\widehat{N} = \left[\frac{m \cdot n}{k}\right].$$

**Interpretation:** In the second sample, among $n$ individuals, there were $k$ marked, so the proportion of the marked individuals in that sample is $p = \frac{k}{n}$. Now, if we assume that marked species were well-mixed in the population, we need to have the same proportion in the whole population, and since we had $N$ species (population size), with $m$ marked, we need to have $p = \frac{m}{N}$. Hence,

$$\frac{m}{N} = \frac{k}{n}, \qquad so \qquad N = \frac{m \cdot n}{k}.$$

**Example 25.10:** Assume scientists catch 200 fishes in a lake, mark them, and return them into a lake. On a later date, they catch another 200 fishes, and find out that 10 are marked. Estimate the number of fishes in the lake.

**Solution:** OTB

# HypGeo Distribution: Important Remark

Very important, in terms of applications, is the following result:

# HypGeo Distribution: Important Remark

Very important, in terms of applications, is the following result:

If $X \sim Hyper(N, m, n)$ is a Hypergeometric r.v., and if $N \to +\infty$, and $m = p \cdot N$, i.e., if $p$ is the probability (proportion) of having a white ball, then $X$ tends to the Binomial r.v. $Y \sim Binom(n, p)$, in the sense that

$$\mathbb{P}(X = k) \to \mathbb{P}(Y = k), \qquad k = 0, ..., n.$$

**Proof:** OTB (you can safely skip the proof!)

# HypGeo Distribution: Important Remark

Very important, in terms of applications, is the following result:

If $X \sim Hyper(N, m, n)$ is a Hypergeometric r.v., and if $N \to +\infty$, and $m = p \cdot N$, i.e., if $p$ is the probability (proportion) of having a white ball, then $X$ tends to the Binomial r.v. $Y \sim Binom(n, p)$, in the sense that

$$\mathbb{P}(X = k) \to \mathbb{P}(Y = k), \qquad k = 0, ..., n.$$

**Proof:** OTB (you can safely skip the proof!)

**Corollary:** So if $N$ is large, $p = \frac{m}{N}$, and $X \sim Hyper(N, m, n)$, then

$$\mathbb{P}(X = k) \approx \mathbb{P}(Y = k), \qquad k = 0, ..., n,$$

where $Y \sim Binom(n, p)$.

# HypGeo Distribution: Important Remark

**Remark:** Statisticians use this result when doing sampling.

# HypGeo Distribution: Important Remark

**Remark:** Statisticians use this result when doing sampling. Usually, we do a sampling **without repetitions** (say, I am asking persons randomly about their age, or choosing at random 50 phone numbers to call and to get the answer about they prefer watching TV or reading probability books),

# HypGeo Distribution: Important Remark

**Remark:** Statisticians use this result when doing sampling. Usually, we do a sampling **without repetitions** (say, I am asking persons randomly about their age, or choosing at random 50 phone numbers to call and to get the answer about they prefer watching TV or reading probability books), but the theory is mostly developed and works for the model **with replacements**, when the results are **independent**.

# HypGeo Distribution: Important Remark

**Remark:** Statisticians use this result when doing sampling. Usually, we do a sampling **without repetitions** (say, I am asking persons randomly about their age, or choosing at random 50 phone numbers to call and to get the answer about they prefer watching TV or reading probability books), but the theory is mostly developed and works for the model **with replacements**, when the results are **independent**.

The above result shows that one can approximate the results of without-repetitions-sampling (non-independence case) by with-replacements-sampling, independent sampling model results, if the population is large enough.

# HypGeo Distribution: Sampling

**Example 25.11:** Assume that in some supermarket, $2\%$ of customers are buying Mer Sareri Matsoun. We stand at the doors of the supermarket, and are randomly asking 150 customers, as they leave, if they have bought MSM. What is the probability that more than 5 persons have bought MSM?

**Solution:** OTB

**Remark:** One can also define the Negative Hypergeometric distribution, in the following way:

# HypGeo Distribution: Remark

**Remark:** One can also define the Negative Hypergeometric distribution, in the following way: assume again that we have $N$ balls in a box, with $m$ white balls, and $N-m$ black ones.

# HypGeo Distribution: Remark

**Remark:** One can also define the Negative Hypergeometric distribution, in the following way: assume again that we have $N$ balls in a box, with $m$ white balls, and $N-m$ black ones. We draw a ball from the box, one at a time, without replacements. We are interested in how many draws we need to do to get exactly $k$ white balls.

# HypGeo Distribution: Remark

**Remark:** One can also define the Negative Hypergeometric distribution, in the following way: assume again that we have $N$ balls in a box, with $m$ white balls, and $N-m$ black ones. We draw a ball from the box, one at a time, without replacements. We are interested in how many draws we need to do to get exactly $k$ white balls. If $X$ is the number of draws (trials), then the distribution of $X$ is called Negative Hypergeometric distribution (with given parameters $N, m, k$).