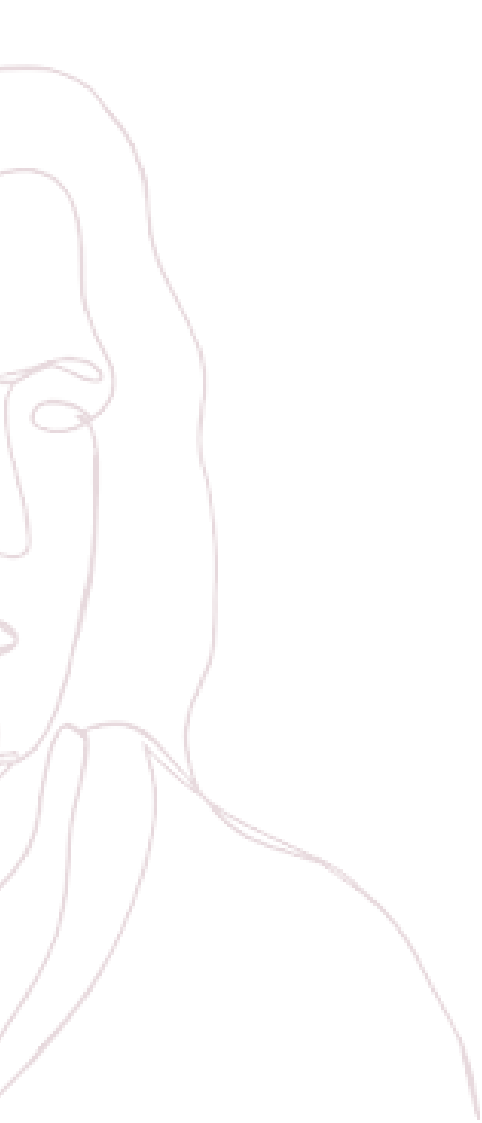




Bernoulli, Binomial and Hypergeometric Distributions

Statistics I · Lecture 9

Today

- Bernoulli and Binomial distributions, and the tables
- Hypergeometric distribution, and when it becomes a Binomial
- Practice, from  to 

Disclaimer

These slides are a free translation and adaptation from the slide deck for Estatística I by Prof. Sandra Custódio and Prof. Teresa Ferreira from the Lisbon Accounting and Business School, Polytechnic University of Lisbon.

Types of random variables

- We studied a couple of ways in which we could categorize the random variables.
- One of these classifications was *discrete* and *continuous* random variables.
- We will start by focusing in the first family of *r.v.s*

Types of random variables

- According to the type of event we are measuring, we can do even better, and cluster together random variables that behave similarly.
- We will define *families* of **statistical distributions**. Knowing to which statistical distribution our random variable belongs, will endow us with a set of tools to deal with them adequately.

Bernoulli and Binomial Distributions

Bernoulli experiment (or trial)

The **Bernoulli experiment** consists in a random experiment with the following features:

- Only two events: A and A^c , for example *success* and *failure*.
 $\Omega = \{A, A^c\}$
- The success happens with probability p and failure with probability $q = 1 - p$:

$$P(A) = p \quad P(A^c) = q = 1 - p$$

Bernoulli Process

This process has even further these three characteristics:

- Only two possible outcomes in each draw
- $P(A)$ remains constant thorough the experiment
- Repeated trials are independent, *i.e.* what happened before does not affect what will happen in the future.

This is process has a binary outcome: Success (1) - Failure (0).

Bernoulli Distribution

The random variable X follows a Bernoulli distribution with parameter $p \in [0, 1]$, $X \sim \text{Ber}(1, p)$ if its probability function is given by:

$$P(X = x) = f_X(x) = \begin{cases} p^x(1 - p)^{1-x} & x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

It is trivial to check that $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

Bernoulli Distribution

The first two moments of the distribution (mean and variance) are given by:

- $E[X] = p$
- $V[X] = p(1 - p)$

In summary, Bernoulli's tries to model what is the probability of success or failure in a single independent trial.

Binomial Distribution

Binomial Distribution

This distribution, is a **generalization** of Bernoulli's.

In this case, it might be better to start from the summary to stress the difference with Bernoulli's:

The Binomial distribution answers what is the probability that, after n independent fail-success trials, in any order, for example, you succeeded $m \leq n$ times.

Binomial Distribution

A random variable X follows a Binomial distribution, $X \sim \text{Bin}(n, p)$ if its probability function $P(X = x)$ is:

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

Note that if we change n or p we would get another binomial distribution, i.e. $B(n, p) \neq B(m, r)$ if either $n \neq m$ or $p \neq r$.

Bernoulli is a Binomial distribution with $n = 1$.

Side note about binomials

Binomial distribution

Computing this probability can become very laborious, even for small values for n . This can be solved thanks to:

1. Computers and advanced calculators
2. Tables for values for n from 1 to 20, and for p from 0.05 to 0.5. Note that if $p > 0.5$ you can just

$$X \sim \text{Bin}(x, n, p) \leftrightarrow \hat{X} \sim \text{Bin}(n - x, n, 1 - p)$$

Binomial distribution

The first two moments of the distribution are found as:

$$1. E[X] = \mu_X = \sum x \binom{n}{x} p^x (1-p)^{n-x} = np$$

$$2. V[X] = E[(X - \mu)^2] = \\ = \sum x^2 \binom{n}{x} p^x (1-p)^{n-x} - \mu^2 = np(1-p)$$

Additivity Theorem of the Binomial Distribution

Let k independent random variables X_i with $i = 1, 2, \dots, k$, where $X_i \sim \text{Bin}(x_i, n_i, p)$, then

$$\text{if } S_k = X_1 + X_2 + \dots + X_k = \sum_{i=1}^k X_i$$

$$S_k \sim \text{Bin} \left(s_k, n = \sum_{i=1}^k n_i, p \right)$$

Example

Using the [tables](#) for the binomial distribution, or the calculator, find the value of the following probabilities:

1. $P(X \leq 3)$ if $X \sim \text{Bin}(9, 0.4)$
2. $P(2 \leq X \leq 6)$ if $X \sim \text{Bin}(7, 0.3)$
3. If $X \sim \text{Bin}(10, p)$ and $P(X \leq 5) = 0.8338$, what is p ?

►  Answer

Example

The SNS (Portugal's National Health Service) has said that the probability of an individual carrying a specific gene, that triggers a certain disease is 0.1. In the city of Porto, 10 individuals were randomly selected for a study, while 8 individuals were randomly chosen in Lisbon for the same purpose. Assess the veracity of the following sentence:

The probability of, in Lisbon, 2 individuals carrying the gene, knowing that between Porto and Lisbon only 5 individuals were found to carry that gene, is 0.2941.

Example

Let $X \sim \text{Bin}(10, 0.1)$ (for Porto) and $Y \sim \text{Bin}(8, 0.1)$ (for Lisbon). Then $X + Y \sim \text{Bin}(18, 0.1)$.

$$\begin{aligned} P(Y = 2 | X + Y = 5) &= \frac{P(Y = 2, X + Y = 5)}{P(X + Y = 5)} \\ &= \frac{P(Y = 2, X = 3)}{P(X + Y = 5)} \\ &= \frac{P(Y = 2)P(X = 3)}{P(X + Y = 5)} \end{aligned}$$

Example

Using the [tables](#) we have $P(Y = 2) = 0.1488$, $P(X = 3) = 0.0574$, $P(X + Y = 5) = 0.0218$, and therefore $P(Y = 2|X + Y = 5) = 0.392$. ❌ False

? Bernoulli and Binomial · Question 1

If $X \sim \text{Bin}(n, p)$, then $E[X]$ and $V[X]$ are:

A. np and $np(1 - p)$

B. np and np

C. p and $p(1 - p)$

D. np^2 and np

✓ **A.** A Binomial is a sum of n independent Bernoulli trials, each with mean p and variance $p(1 - p)$.

? Bernoulli and Binomial · Question 2

A Binomial model requires:

A. draws without replacement

B. a continuous support

C. n independent trials with a constant p

D. a known λ

 **C.** Without replacement the probability changes from draw to draw, and that is the Hypergeometric, not the Binomial.



Bernoulli and Binomial · Question 3

A machine produces defective items with probability $p = 0.1$, independently.

In a sample of $n = 10$ items, compute $P(X = 0)$, $P(X \leq 1)$, $E[X]$ and $V[X]$.



Bernoulli and Binomial · Solution

$$P(X = 0) = \binom{10}{0} 0.1^0 0.9^{10} = 0.9^{10} \approx 0.3487$$

$$P(X = 1) = \binom{10}{1} 0.1 \times 0.9^9 \approx 0.3874, \text{ so } P(X \leq 1) \approx 0.7361$$

$$E[X] = np = 1 \cdot V[X] = np(1 - p) = 0.9$$

Hypergeometric Distribution

Hypergeometric Distribution

Suppose you have a population with size N , from where M elements have some characteristic, and therefore $N - M$ do not. From that sample, we draw randomly a sample of size n , without replacement, and we would like to know, from this sample, how many have the characteristic we mentioned.

Given we are not replacing the elements we put in our sample, the events here are not independent!

Hypergeometric Distribution

In this case the r.v. X would be how many elements from our sample have the referred characteristic. $X \sim \text{Hypergeometric}(N, M, n)$ if its probability distribution function $P(X = x)$ is:

$$f_X(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$

where $\max(0, M + n - N) \leq x \leq \min(M, n)$

Hypergeometric Distribution

The first two moments of the distribution are found as:

- $E[X] = \mu_X = n \frac{M}{N} = np$
- $V[X] = npq \frac{N-n}{N-1}$ note here that because we do not replace elements, we get a smaller variance.

Hypergeometric vs Binomial

- **Binomial**
 - Replacement
 - Same population every draw
 - Independent draws
- **Hypergeometric**
 - No replacement
 - Population changes every time
 - Dependent draws

Hypergeometric $\rightarrow$ Binomial

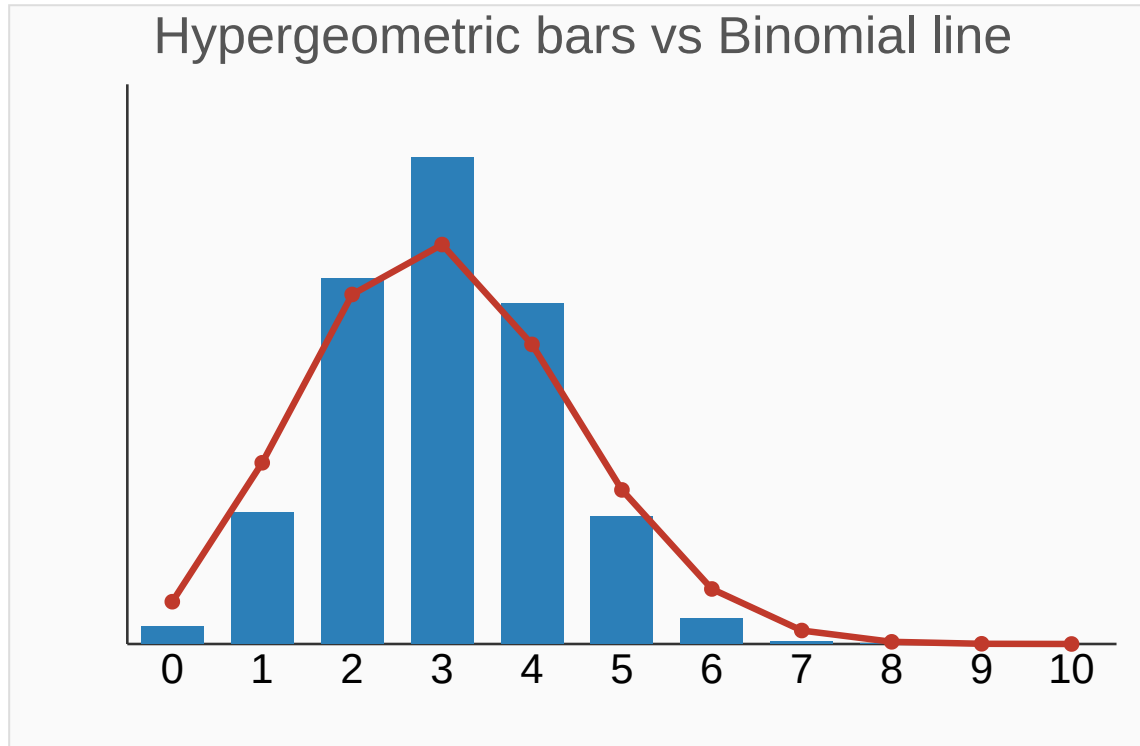
Let $X \sim \text{Hypergeometric}(N, M, n)$. If $N \gg n$ then the change in the population is very small, and then the draws approximate independent draws (population is very similar after a draw). As a rule of thumb if $\frac{n}{N} \leq 0.1$ or if the sample size is less than 10% of the population, we can approximate the Hypergeometric with the Binomial.

$$X \sim \text{Hypergeom}(N, M, n) \leftrightarrow X \approx \text{Bin} \left(n, p = \frac{M}{N} \right)$$

The smaller $\frac{n}{N}$ the better the approximation.

How big must the population be?

Bars are the exact $\text{Hypergeom}(N, M, n)$, drawing **without** replacement. The red outline is $\text{Bin}(n, p = M/N)$, drawing **with** replacement. The share M/N is held at 0.3 as N grows.



Population $N = 30$



Sample $n = 10$



$M = 9$, $p = M/N = 0.300$

$n/N = 0.333$

biggest gap = 0.0583

Rule of thumb: too small a population

needs $n/N \leq 0.1$

What the chart showed

With $N = 30$ and $n = 10$ you are removing a third of the population as you go, so each draw visibly changes the odds for the next one and the two distributions part company.

Push N past 100 with the same sample and the bars settle onto the red line: taking 10 out of 400 barely disturbs the population, so **draws without replacement start behaving like independent draws**.

That is all $\frac{n}{N} \leq 0.1$ is saying. It is not a statement about N being large, it is about the **sample being small relative to it**.

? Hypergeometric · Question 1

The Hypergeometric differs from the Binomial because the draws are:

- A. independent
 - B. infinite in number
 - C. continuous
 - D. without replacement, so p changes at each draw
- ☒ **D.** Removing an item changes the composition of what is left, so the draws are not independent.

? Hypergeometric · Question 2

Approximating a Hypergeometric by a Binomial is safe when:

A. N is large

B. $\frac{n}{N} \leq 0.1$

C. $M = N$

D. $n > \frac{N}{2}$

 **B.** The rule of thumb is about the sample being **small relative to the population**, not about N being large on its own.



Hypergeometric · Question 3

A box holds $N = 20$ items, of which $M = 5$ are defective. You draw $n = 4$ without replacement.

Compute $P(X = 1)$, and say whether the Binomial approximation would be appropriate here.



Hypergeometric · Solution

$$P(X = 1) = \frac{\binom{5}{1} \binom{15}{3}}{\binom{20}{4}} = \frac{5 \times 455}{4845} = \frac{2275}{4845} \approx 0.4696$$

Here $\frac{n}{N} = \frac{4}{20} = 0.2 > 0.1$, so the Binomial approximation is **not** appropriate. ❌

Practice

Practice

The exercises go from easy to hard:

-  warm-up: one formula, one step
-  standard: two or three steps
-  exam level: this is what you will find in the midterm
-  stretch: you need to combine several things

Keep the [Binomial table](#) open. Remember the table gives $F(x) = P(X \leq x)$, not $P(X = x)$.

Practice · Exercise 1

Let $X \sim \text{Bin}(8, 0.35)$. Using the table, find:

(a) $P(X \leq 2)$

(b) $P(X = 4)$

(c) $P(X \geq 6)$

(a) Direct inspection: $F(2) = 0.4278$

(b) $P(X = 4) = F(4) - F(3) = 0.8939 - 0.7064 = 0.1875$

(c) $P(X \geq 6) = 1 - P(X \leq 5) = 1 - F(5) = 1 - 0.9747 = 0.0253$

Practice · Exercise 2

For a given company, the probability of a receipt  being filled incorrectly is 0.2. You pick randomly 10 receipts.

Find the probability that:

- (a)** exactly 5 receipts contain irregularities;
- (b)** at most 3 receipts contain irregularities;
- (c)** at least one receipt was filled incorrectly.

$X \sim \text{Bin}(10, 0.2)$ counts the irregular receipts.

(a) $P(X = 5) = F(5) - F(4) = 0.9936 - 0.9672 = 0.0264$ **(b)**

$P(X \leq 3) = 0.8791$ **(c)** $P(X \geq 1) = 1 - F(0) = 1 - 0.1074 = 0.8926$

Practice · Exercise 3

Let $X \sim \text{Bin}(10, 0.7)$. Find $P(X \geq 8)$.

The table stops at $p = 0.5$... 

If X counts the successes, $Y = 10 - X$ counts the failures, and $Y \sim \text{Bin}(10, 0.3)$.

$$P(X \geq 8) = P(10 - Y \geq 8) = P(Y \leq 2) = 0.3828$$

Practice · Exercise 4

In the supermarket *SweetDrop* 🛒, the probability that a customer spends more than 25 euro is 0.1.

(a) In a group of n customers, 45 are expected to spend more than 25 euro. What is n ?

(b) From a group of 20 customers, 10 bought products of the *SweetDrop* brand. We choose 5 customers, without replacement. Find the probability that exactly 2 of them bought the brand.

(a) $E[X] = np = 0.1n = 45 \Rightarrow n = 450$

(b) Without replacement, and a small population: $X \sim \text{Hypergeometric}(20, 10, 5)$

$$P(X = 2) = \frac{\binom{10}{2} \binom{10}{3}}{\binom{20}{5}} = \frac{45 \times 120}{15504} \approx 0.3483$$

Practice · Exercise 5

Machine 1 produces 6 pieces per hour and machine 2 produces 4. Each piece is defective with probability 0.1, independently of everything else.

Find the probability that in one hour, between both machines, **at most one** defective piece is produced.

$X_1 \sim \text{Bin}(6, 0.1)$ and $X_2 \sim \text{Bin}(4, 0.1)$, independent, with the **same** p .

By the additivity theorem, $X_1 + X_2 \sim \text{Bin}(10, 0.1)$, and from the table:

$$P(X_1 + X_2 \leq 1) = 0.7361$$

If the machines had a different p , we could not add them like this.

Practice · Exercise 6

An auditor  receives a folder with 50 invoices, 6 of which contain errors. She checks 5 invoices, chosen randomly, without replacement.

- (a)** Find the probability that she finds at least one invoice with errors.
- (b)** Compute the same probability with the Binomial approximation. Is the approximation good here?



Practice · Exercise 6 · Solution

(a) $X \sim \text{Hypergeometric}(50, 6, 5)$:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{6}{0} \binom{44}{5}}{\binom{50}{5}} \approx 1 - 0.5126 = 0.4874$$

(b) With $p = \frac{6}{50} = 0.12$, $X \approx \text{Bin}(5, 0.12)$ and
 $P(X \geq 1) \approx 1 - 0.88^5 \approx 0.4723$

$\frac{n}{N} = \frac{5}{50} = 0.1$, right at the limit. The error is of 1.5 percentage points, which is not huge, but not negligible either.

Practice · Exercise 7

Each receipt is filled incorrectly with probability 0.2.

How many receipts must you inspect, at least, to have a probability of at least 0.9 of finding one or more incorrect receipts?

$X \sim \text{Bin}(n, 0.2)$ and we want $P(X \geq 1) = 1 - 0.8^n \geq 0.9$

$$0.8^n \leq 0.1 \Leftrightarrow n \geq \frac{\ln 0.1}{\ln 0.8} \approx 10.32$$

So $n = 11$ receipts. Note the inequality flips, because $\ln 0.8 < 0$.

● Practice · Exercise 8

Let $X \sim \text{Bin}(n, p)$ with $E[X] = 4$ and $V[X] = 2.4$.

(a) Find n and p , and then $P(X \leq 3)$.

(b) True or false: if $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(n, \theta)$ are independent, then $X + Y \sim \text{Bin}(n, p + \theta)$.

(a) $\frac{V[X]}{E[X]} = \frac{np(1-p)}{np} = 1 - p = 0.6$, so $p = 0.4$ and $n = \frac{4}{0.4} = 10$. From the table, $P(X \leq 3) = 0.3823$

(b) ✗ False. Additivity needs the **same** p , and then the n add up: $\text{Bin}(n_1 + n_2, p)$. With $p + \theta$ we could even get a “probability” above 1.



Homework

Problem set **3.1**, Questions 1 to 5.

Bibliography

- Murteira, B.; Silva Ribeiro, C.; Andrade e Silva, J. & Pimenta, C., Introdução à Estatística, Escolar Editora, McGraw-Hill, 2010
- Paulino, C. D. & Branco, J. A. (2005). Exercícios de Probabilidade e Estatística. Escolar Editora
- Pimenta, F., Andrade e Silva, J.; Silva Ribeiro, C. & Murteira, B., Introdução à Estatística, 3ª Edição, Escolar Editora, 2015

Appendix

Binomials

Back to Binomial Distribution

Binomial Table

[Back to exercise](#)

Distribution Function

$n = 1$

x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$n = 2$

x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$n = 3$

x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
---	------	------	------	------	------	------	------	------	------	------