STATISTICS

Module 2 *of*
MATHEMATICS AND STATISTICS

Marina Santacroce

marina.santacroce@unibocconi.it
Department of Decision Sciences, Bocconi University

L8. Discrete random variables: Bernoulli, geometric, binomial, Poisson and hypergeometric

Outline

Bernoulli Random Variable

Assume an outcome of an experiment can be classified as either "success" or "failure" and that the probability a success occurs is p.

The random variable $X = \mathbb{1}_{\{success\}}$, where $X = 1$ corresponds to a success and $X = 0$ corresponds to a failure, is a discrete random variable taking on the values $\{0, 1\}$

- $\mathbb{P}(X = 1) = p(1) = p$
- $\mathbb{P}(X = 0) = p(0) = 1 - p.$

Def. Such a random variable is called a Bernoulli random variable with parameter p

$$X \sim Be(p), \qquad 0 < p < 1$$

with mass probability function

- $p(0) = 1 - p, \qquad p(1) = p.$
- $\boxed{\mathbb{E}(X) = p}$ and $\boxed{\mathrm{Var}(X) = p(1 - p)}$.
- $\phi_X(t) = \mathbb{E}[e^{tX}] = 1 - p(1 - e^t).$

Binomial Random Variable

Consider a sequence of n independent success/failure experiments, each of which yields success with probability p.

Let X be the "*number of successes*", then X is a discrete random variable which takes on the values $\{0, 1, \dots, n\}$ with probability

- $\mathbb{P}(X = k)$= p(k) = $\binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, \dots, n\}$.

Def. The random variable X is said to be a binomial random variable with parameters n and p,

$$X \sim Bi(n, p) \qquad n \in \mathbb{N} \qquad \mathit{and} \qquad p \in (0; 1)$$

The mass probability function is

- $p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \dots n\}$.
- $\boxed{\mathbb{E}(X) = np}$ and $\boxed{\mathrm{Var}(X) = np(1-p)}$.
- $\phi_X(t) = \mathbb{E}[e^{tX}] = \left(1 - p(1 - e^t)\right)^n$.

Rem. If $\{X_i, i : 1, \dots, n\}$ are the n Bernoulli random variables (X_i corresponding to the $i - th$ trial) and $\mathbb{P}(X_i = 1) = p$, since trials are independent, then

$$X = \sum_{i:1}^{n} X_i \qquad X \sim Bi(n, p)$$

Example: Blood test

A large number, N, of people are subjected to a blood test. This can be administered in two ways:

(1) Each person can be tested separately, in this case N tests are required,

(2) the blood samples of k persons can be pooled and analyzed together. If this test is negative, one test is enough for the k people. If the test is positive, each of the k persons must be tested separately, and in all, $k + 1$ tests are required for the k people.

Assume that the probability p that a test is positive is the same for all people and that these events are independent.

(a) Find the probability that the test for a pooled sample of k people will be positive.

(b) What is the expected value of the number Y of tests necessary under plan (2)? (Assume that N is divisible by k.)

Poisson Random Variable

Def. A discrete random variable X taking on the values 0, 1, 2, ... is said to be a Poisson random variable with parameter λ,

$$X \sim \mathit{Poisson}(\lambda) \qquad \lambda \in \mathbb{R}^+$$

if its mass probability function is

- $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k \in \{0, 1, \dots\}.$

This is indeed a probability distribution as

$$\sum_{k:0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$$

- $\boxed{\mathbb{E}(X) = \lambda}$ and $\boxed{\mathrm{Var}(X) = \lambda}$.
- $\phi_X(t) = \mathbb{E}[e^{tX}] = e^{-\lambda(1-e^t)}$.

Poisson: the *law of rare events*

Prop. Let $\{X_n, n > \lambda\}$, $\lambda > 0$, be a sequence of binomial random variables $X_n \sim Bi(n, p_n = \frac{\lambda}{n})$. [1] Then

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k : 0, 1, \ldots.$$

◇ Alternatively, check the convergence of the corresponding MGF, i.e.

$$\lim_{n\to\infty} \phi_{X_n}(t) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}(1-e^t)\right)^n = e^{-\lambda(1-e^t)}.$$

Practically it means that the number of successes obtained in a *large* number of independent trials, each with probability of success p can be approximated by a Poisson r.v. of parameter $\lambda = np$, if np is *moderate*.

Poisson distribution has a wide range of different applications. Some examples that are Poisson distributed to a good approximation are: the number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry, the number of phone calls at a call center per minute, the number of misprints in a page (or chapter) of a book, ...

[1] It holds also when $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$

Example: Murders in London

Each year the London Metropolitan Police record around 160 murders, and this has been stable for the last 5 years. Each of these murders is an individual crime that cannot be predicted. It may appear strange, but this very randomness means that the overall pattern of murders is in some ways quite predictable.

Assuming the number of daily murders are Poisson distributed, compute the probability of having

a) *a peaceful day*: no murders during a day,

b) *a bloody day*: 3 or more murders in one day,

c) *a peaceful week*: a week without any murder.

Geometric Random Variable

Consider again independent trials, each with success probability p. These trials are performed until a success occurs.

Let X be the number of trials required. Then X takes on the values $\{1, 2, \dots\}$ with probability

- $p(k) = \mathbb{P}(X = k) = p(1-p)^{k-1}, \quad k \in \{1, \dots\}.$

Def. This random variable is said to be geometric with parameter p,

$$X \sim Ge(p) \qquad p \in (0;1)$$

◇ Verify that $p(k)$ is a mass probability function, i.e. $\sum_{k:1}^{\infty} p(k) = 1$.

- $\boxed{\mathbb{E}(X) = \frac{1}{p}}$ and $\boxed{\mathrm{Var}(X) = \frac{(1-p)}{p^2}}$.
- $\phi_X(t) = \mathbb{E}[e^{tX}] = \frac{pe^t}{1-e^t(1-p)}$.

◇ Find the law of the random variable denoting the *trial of the r-th success* $\Rightarrow$ negative binomial(r, p).

Example: Collecting pictures

Assume that, every time your little brother buys a box of Wheaties, he receives a picture of one of the n players of the Canadian Hockey team.

Let X_k be the number of additional boxes he has to buy, after he has obtained $k-1$ different pictures, in order to obtain the next new picture. Thus $X_1 = 1$, X_2 is the number of boxes bought after this to obtain a picture different from the 1st pictured obtained, and so forth.

We assume that Wheaties does not favor any player, i.e. the probability of finding player i equals the probability of finding player j and is $1/n$.

a) What is the distribution of X_k?

b) What is the expectation of Y, where Y is the total number of boxes you must buy to get the n different players?

In order to prove part b) use $\Rightarrow$ $\quad \mathbb{E}(Y) = \sum_{k:1}^{n} \mathbb{E}(X_k)$

Hypergeometric Random Variable

Consider a barrel or urn containing N balls of which m are white and $N - m$ are black. We take a simple random sample (i.e. without replacement) of size n and denote by X, the *number of white balls in the sample*.

Def. The Hypergeometric distribution is the distribution of X under this sampling scheme. It is a discrete random variables which takes on the values $\{\max(0, n-(N-m)), \ldots, \min(n, m)\}$ with probability

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

Let us introduce $p = m/N$ then one can prove

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

Suppose that N is very large compared to n, it seems reasonable that sampling without replacement is not too much different than sampling with replacement. It can indeed be shown that the hypergeometric distribution can be well approximated by the binomial of parameters $p = m/N$ and n.

Example: Survey

Suppose that as part of a survey, 7 houses are sampled at random from a street of 40 houses in which 5 contain families whose family income puts them below the poverty line.

What is the probability that:

(a) None of the 5 families are sampled?

(b) 4 of them are sampled?

(c) No more than 2 are sampled?

(d) At least 3 are sampled?