Basic Statistics and Probability

Chapter 4: Discrete Random Variables

- Two Types of Random Variables
- Probability Distributions for Discrete Random Variables
- Expected Values of Discrete Random Variables
- The Binomial Random Variable
- The Poisson Random Variable

A random variable (r.v.) is a variable that takes numerical values associated to the outcomes of a certain random experiment, where one (and only one) numerical value is assigned to each sample point.

Examples:

- Toss a coin and define the r.v.

$$X = \begin{cases} 1 & \text{if head} \\ 0 & \text{if tail} \end{cases}$$

- Consider the r.v. $Y =$ "Number of social networking services used by a person chosen at random"
- Define the r.v. $Z =$ "Height (in m) of a 10-year-old child chosen at random"

The aim is to know the probability, the chances, that X will take certain values: which values of the r.v. are more frequent and which are less likely.

Two Types of Random Variables

1. If the number of possible values of a r.v. are countable, that is, the values can be listed (although they may be infinite), then we say that the r.v. is discrete.

- Number of eggs per nest
- Numbers of patients per day coming to see a certain physician
- Number of coin tosses till the first occurrence of "Tails"

2. If the possible values of a r.v. are all the numbers in an interval (limited or unlimited) we say that the variable is continuous.

- Length of each egg
- Blood pressure of each patient
- Height of each individual
- Temperature

Probability Distributions for Discrete RV's

To characterize completely the probability distribution of a discrete r.v. we have to specify all the values the variable can take and the probability with which the variable assumes that value.

Example 4.1: Tossing two coins

Toss two coins and define the r.v. X = "Number of heads obtained". The possible values of X are 0, 1 and 2. The probability distribution of X specifies how the probability is distributed over those values

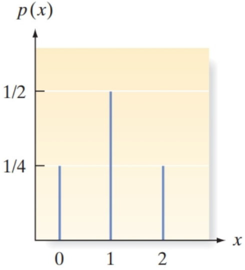
$$P\{X=0\} = P(\mathrm{TT}) = \tfrac{1}{4}$$
$$P\{X=1\} = P(\mathrm{TH}) + P(\mathrm{HT}) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}$$
$$P\{X=2\} = P(\mathrm{HH}) = \tfrac{1}{4}$$

Usually, the probability distribution of a discrete r.v. X is characterized via its (probability) mass function, which gives for each possible value x of X, the probability that X is equal to x:

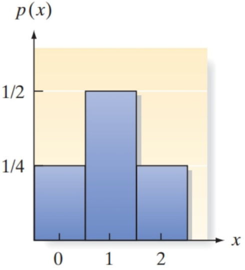
$$x \mapsto p(x) = P\{X = x\}.$$

Example 4.1: Tossing two coins

$p(0) = 1/4$
$p(1) = 1/2$
$p(2) = 1/4$



a. Point representation of $p(x)$



b. Histogram representation of $p(x)$

Example 4.2: Insurance policy

An insurance company sells a \$10,000 one-year term insurance policy at an annual premium of \$290. Actuarial tables show that the probability of death during the next year for a person of the typical customers age, sex, health, etc., is .001.

Gain x	Sample Point	Probability
\$290	Customer lives	.999
−\$9,710	Customer dies	.001

Properties of a general mass function:

$$p(x) \geq 0 \quad \text{for all } x \qquad\qquad \sum_x p(x) = 1$$

The population mean, the expected value or the expectation of a random variable X is a measure of cental tendency of the probability distribution of X.

The expectation of a discrete r.v. X is defined as

$$\mu = E(X) = \sum_{x} x\, p(x).$$

Example 4.1: Tossing two coins

Example 4.2: Insurance policy

What is the expected gain (amount of money made by the company) for a policy of this type?

The population variance of a r.v. X is the expected squared distance between X and its mean μ:

$$\sigma^2 = V(X) = E[(X-\mu)^2] = \sum_x (x-\mu)^2 p(x).$$

It holds that $\sigma^2 = E(X^2) - \mu^2 = \sum_x x^2 p(x) - \mu^2$.

Example 4.1: Tossing two coins

Example 4.2: Insurance policy

The standard deviation (s.d.) of a r.v. X is the square root of its variance:

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}.$$

Example 4.1: Tossing two coins

Example 4.2: Insurance policy

Bernoulli Distribution

A Bernoulli experiment or Bernoulli trial is a random experiment with only two (mutually exclusive) possible results: success (S) and failure (F), with $P(\mathrm{S}) = p$ and $P(\mathrm{F}) = 1 - p$.

Example 4.3: Toss one coin. Take S = Head and F = Tail.

Example 4.4: A couple, each of them with a recessive gene (blue) and a dominant one (brown) for the eyes colour, have a child. We codify S = Brown-eyed child and F = Blue-eyed child.

Example 4.5: In a campaign for early detection of diabetes among volunteers, an oral glucose tolerance test measures blood glucose after not eating for at least 8 hours and 2 hours after drinking a glucose-containing beverage. If the glucose level is above 200 mg/dl, the individual is classified as a potential diabetic. If not, the individual is considered healthy. We codify S = "Potential diabetic" with $p = 0.03$.

The Bernoulli distribution is that of the r.v.

$$X = \begin{cases} 1 & \text{if the Bernoulli trial yields success} \\ 0 & \text{if failure is obtained} \end{cases}$$

We denote it $X \sim \text{Bernoulli}(p)$. The mass function is

The expectation and variance are

$$E(X) = p \qquad \text{and} \qquad V(X) = p(1-p).$$

Bernoulli experiments give rise to many other probability distributions, such as the binomial, or the geometric.

Binomial Distribution

We repeat n independent Bernoulli trials, with $P(\mathrm{S}) = p$ in each trial. The binomial distribution $B(n, p)$ is the probability distribution of the r.v. $X =$ "number of successes in the n trials".

The mass function is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n,$$

where $\quad \binom{n}{x} = \frac{n!}{x!(n-x)!} \quad$ and $\quad n! = n(n-1)\cdots 3 \cdot 2 \cdot 1.$

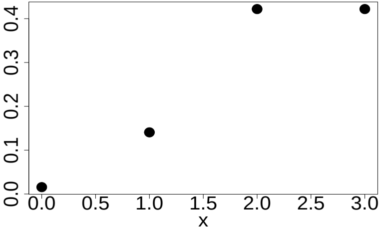
The expectation and variance are

$$E(X) = np \qquad \text{y} \qquad V(X) = np(1-p).$$

Remark: X can be expressed as $X = \sum_{i=1}^{n} Z_i$, where $Z_i \sim \text{Bernoulli}(p)$ for $i = 1, \ldots, n$.

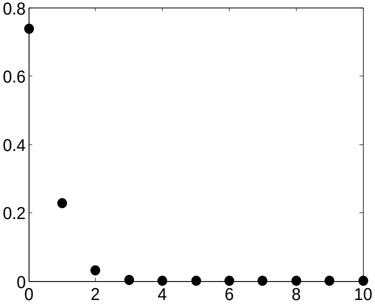
Example 4.4: Colour of eyes in a child

If the couple has three children in common, which is the mass function of the r.v. $X =$ "number of children with brown eyes"?



Example 4.5: Early detection of diabetes

The test is carried out on 10 volunteers and we define the r.v. $X =$ "number of potential diabetics among those 10".



What is the probability that there is more than one potential diabetic among the 10 observed volunteers?

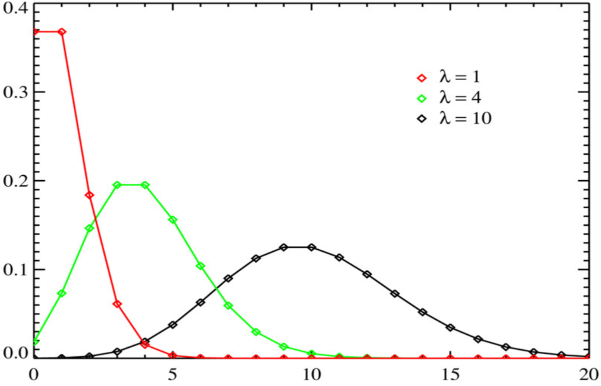
Example 4.6: Colorectal Cancer and Gene Mutation

In a certain country the probability that a person who has suffered colorectal cancer has a mutation in gene p53 is 60%. A sample of 5 patients with colorectal cancer is taken. Compute the probability that at most one of them has the mutated gene. Which is the expected number of patients, among those 5, who will have the mutated gene? Which is the variance?

The r.v X follows a Poisson distribution with parameter $\lambda > 0$, and we denote it by $X \sim \text{Poisson}(\lambda)$, if its mass function is

$$p(x) = e^{-\lambda}\,\frac{\lambda^x}{x!} \qquad \text{for } x = 0, 1, 2, \ldots$$

Then $E(X) = \lambda = V(X)$.



The Poisson distribution is the limit of the binomial in the following sense: $B(n, p) \longrightarrow Poisson(\lambda)$ when $n \to \infty$, $p \to 0$ and $np \to \lambda$ (*Law of Rare Events*).

In practice, if $X \sim B(n, p)$ with $n \geq 30$, $p \leq 0.1$ and $np \leq 10$, then

$$P\{X = k\} \simeq P\{Y = k\},$$

where $Y \sim \text{Poisson}(\lambda)$ and $\lambda = np$.

Example 4.5: Early detection of diabetes

The test is tried on 100 volunteers. What is the probability that at most 3 of them are potential diabetics?

The Poisson distribution is frequently used as a probabilistic model for the number of independent events (arrivals, accidents, calls,...) taking place in a time or space unit, when the rate or frequency of those events (that is, the mean number of those events per time or space unit) is constant.

- Number of traffic accidents per month at a busy intersection.
- Number of mutations in a fragment of DNA of a specified length after a certain dose of radiation.
- Number of misprints per page in a book.
- Number of nuclear disintegrations per unit time in a radioactive material.
- Number of excitatory postsynaptic potentials received by the dendritic tree of a neuron in one minute.
- Number of death claims per day received by an insurance company.

More examples in `www.wikigenes.org/e/mesh/e/5842.html`