Lecture 2 - Random Variables

Jan Bouda

FI MU

April 10, 2009

Random variable - motivation

- In many situation outcomes of a random experiment are numbers.
- In other situations we want to assign to each outcome a number (in addition to probability).
- It may e.g. quantify financial or energetic cost of a particular outcome.
- We will define the random variable to to develop methods for studying random experiments with outcomes that can be described numerically.
- E.g. in case of Bernoulli trials we may be interested only in number of 'successes' and not in actual sequence of 'successes' and 'failures'.
- Almost all real probabilistic computation is done using random variables.

Random variable - definition

A random variable is a rule that assigns a numerical value to each outcome of an experiment.

Definition

A **random variable** $\mathbf{X}$ on a sample space $\mathcal{S}$ is a function $\mathbf{X} : \mathcal{S} \rightarrow \mathbb{R}$ that assigns a real number $\mathbf{X}(s)$ to each sample point $s \in \mathcal{S}$.

We define the **image** of a random variable $\mathbf{X}$ as the set $\mathbf{Im}(\mathbf{X}) = \{\mathbf{X}(s) | s \in \mathcal{S}\}$. This definition is similar to image of any other function.

Random variable - definition

A random variable partitions the sample space into a set of mutually exclusive and collectively exhaustive events. For a random variable $\mathbf{X}$ and a real number x we define the event $A_x =$ "$\mathbf{X} = x$" (sometimes called the **inverse image** of the set $\{x\}$) to be the set of all events from $\mathcal{S}$ to which $\mathbf{X}$ assigns the value x

$$A_x = \{s \in \mathcal{S} | \mathbf{X}(s) = x\}.$$

Whenever you are not sure what some operation with random variable means, always recall the basic definition of the random variable. In example, the statement $\mathbf{X} \leq \mathbf{Y}$ means that $\mathbf{X}$ and $\mathbf{Y}$ are defined on the same sample space $\mathcal{S}$ and for every sample point $s \in \mathcal{S}$ it holds that $\mathbf{X}(s) \leq \mathbf{Y}(s)$.

Random variable - definition

Obviously $A_x \cap A_y = \emptyset$ iff $x \neq y$ and

$$\bigcup_{x \in \mathbb{R}} A_x = S.$$

Therefore the set of events $\{A_x\}_{x \in \mathbb{R}}$ defines an event space and we will often prefer to work in this event space rather than in the original sample space. We usually abbreviate A_x as $[X = x]$.

The image of a discrete random variable (this is the case in this course) is at most countable.

Following the definition of $A_x = [X = x]$ we calculate its probability as

$$P([X = x]) = P(\{s | X(s) = x\}) = \sum_{X(s)=x} P(s).$$

Random variable - probability distribution

Definition

Probability distribution of a random variable X is a function $p_X : \mathbb{R} \rightarrow [0,1]$ satisfying the properties:

(p1) $0 \leq p_X(x) \leq 1$ for all $x \in \mathbb{R}$

(p2) For a discrete random variable X, the set $\{x | p_X(x) > 0\}$ is a finite or countable infinite subset of real numbers. Let us denote it by $\{x_1, x_2, \dots\}$. We require that

$$\sum_i p_X(x_i) = 1.$$

A real valued function $p_X(x)$ defined on $\mathbb{R}$ is a probability distribution of some random variable if it satisfies properties (p1)-(p3).

Random variable - probability distribution

When the random variable is clear from the context, we denote the probability distribution as $p(x)$.
Do not mistake the probability distribution with the **distribution function**, which is a non-decreasing function which tends to 0 as $x \to -\infty$ and to 1 as $x \to \infty$.

Distribution functions

We often are interested in computing the probability of the set $\{s|X(s) \in A\}$ for some subset $A \subseteq \mathbb{R}$. We know that

$$\{s|X(s) \in A\} = \bigcup_{x_i \in A} \{s|X(s) = x_i\} \stackrel{def}{=} [X \in A].$$

If $-\infty < a < b < \infty$ and A is an interval $A = (a, b)$, we usually write $P(a < X < b)$ instead of $P(X \in (a, b))$. If $A = (a, b]$, then $P(X \in A)$ will be written as $P(a < X \leq b)$. Of special interest is the infinite interval $A = (-\infty, x]$ and we denote it by $[X \leq x]$. We calculate the probability of A as

$$P(X \in A) = \sum_{x_i \in A} p_X(x_i).$$

Probability distribution function

Definition

The **probability distribution function** (or simply **distribution function**) of a random variable X is

$$F_X(t) = P(-\infty < X \leq t) = P(X \leq t) = \sum_{x \leq t} p_X(x), \ -\infty < t < \infty.$$

It follows that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

If X is an integer-valued random variable, then

$$F(t) = \sum_{-\infty < x \leq \lfloor t \rfloor} p_X(x).$$

Probability distribution function - properties

(F1) $0 \leq F(x) \leq 1$ for $-\infty < x < \infty$

(F2) $F(x)$ is a monotone increasing function of x, that is if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$. It is easy to see that $(-\infty, x_1] \subseteq (-\infty, x_2]$ if $x_1 \leq x_2$ and we have

$$P(-\infty < X \leq x_1) \leq P(-\infty < X \leq x_2)$$

giving $F(x_1) \leq F(x_2)$.

(F3) $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to\infty} F(x) = 1$. If the random variable X has a finite image, then there exist $u, v \in \mathbb{R}$ such that $F(x) = 0$ for all $x < u$ and $F(x) = 1$ for all $x \geq v$.

(F4) $F(x)$ has a positive increase equal to $p_X(x_i)$ at $i = 1, 2, \ldots$ and in the interval $[x_i, x_{i+1})$ it has a constant value. Thus

$$F(x) = F(x_i) \text{ for } x_i \leq x < x_{i+1}$$

and

$$F(x_{i+1}) = F(x_i) + p_X(x_{i+1}).$$

Probability distribution function

- Any function satisfying properties (F1)-(F4) is the distribution function of some discrete random variable.
- In most cases we simply forget the theoretical background (random experiment, sample space, events,...) and examine random variables, probability distributions and probability distribution functions.
- Often the initial information is *we have a random variable X with the probability distribution $p_X(x)$*. We can construct probability space consistent with the random variable as follows. Let $S = \mathbb{R}$, $X(s) = s$ for $s \in S$, $\mathcal{F}$ is a union of inverse images A_x of all subsets $\{x\}$ and

$$P(A) = \sum_{x \in A} p_X(x).$$

Part I

Examples of probability distributions

Examples of probability distributions

In this part of the lecture we introduce the most common probability distributions occurring in practical situations. In fact, we can always derive the distributions and all related results ourselves, however, it is anyway useful to remember these distributions and situations they describe both as examples and to speed up our calculations. These probability distributions are so important that they have specific names and sometimes also notation.

Constant random variable

- For $c \in \mathbb{R}$ the function defined for all $s \in S$ by $X(s) = c$ is a discrete random variable with $P(X = c) = 1$.
- The probability distribution of this variable is

$$p_X(x) = \begin{cases} 1 & \text{if } x = c \\ 0 & \text{otherwise.} \end{cases}$$

- Such a random variable is called the **constant random variable**.
- The corresponding distribution function is

$$F_X(x) = \begin{cases} 0 & \text{for } x < c \\ 1 & \text{for } x \geq c. \end{cases}$$

Indicator random variable

- Let us suppose that the event A partitions the sample space S, i.e. $A \cup \overline{A} = S$.
- The indicator of an event A is the random variable I_A defined by

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A \end{cases}$$

- The event A occurs if and only if $I_A = 1$.
- The probability distribution is

$$\begin{aligned} p_{I_A}(0) &= P(\overline{A}) = 1 - P(A) \\ p_{I_A}(1) &= P(A). \end{aligned}$$

- The corresponding distribution function reads

$$F_{I_A}(x) = \begin{cases} 0 & \text{for } x < 0 \\ P(\overline{A}) & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

Discrete uniform probability distribution

- Let X be a discrete random variable with a finite image $\{x_1, x_2 \dots x_n\}$ and let us assign to all elements of the image the same probability $p_X(x_i) = p$.
- From the requirement that the probabilities must sum to 1 we have

$$1 = \sum_{i=1}^{n} p_X(x_i) = \sum_{i=1}^{n} p = np$$

and the probability is

$$p_X(x_i) = \begin{cases} 1/n & x_i \in \mathbf{Im}(X) \\ 0 & \text{otherwise.} \end{cases}$$

- Such a random variable is said to have the **uniform probability distribution**.
- This concept cannot be extended to random variable with countably infinite image.

Discrete uniform probability distribution

- If $\mathbf{Im}(X) = \{1, 2, \ldots n\}$ with $p_X(i) = 1/n$, $1 \leq i \leq n$, the probability distribution function is

$$F_X(x) = \sum_{i=1}^{\lfloor x \rfloor} p_X(i) = \frac{\lfloor x \rfloor}{n}, \quad 1 \leq x \leq n.$$

Bernoulli probability distribution

- The Bernoulli probability distribution of a random variable X origins from the random experiment consisting of a single bernoulli trial (e.g. a coin toss).
- The only possible values of the random variable X are 0 and 1 (often denoted as *failure* and *success*, respectively).
- The distribution is given by

$$\begin{aligned} p_X(0) = & p_0 = P(X = 0) = q \\ p_X(1) = & p_1 = P(X = 1) = p = 1 - q \end{aligned}$$

- The corresponding probability distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ q & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

Bernoulli probability distribution

Example

Let X be a Bernoulli random variable with parameter p and image $\{0, 1\}$. X is the indicator of the event

$$A = \{s|X(s) = 1\}$$

and its probability distribution is $p_X(0) = 1 - p$ and $P_X(1) = p$.

Binomial probability distribution

- The Binomial probability distribution of a random variable Y_n is the number of successes (outcomes 1) in n consecutive Bernoulli trial with the same fixed probability p of success in each trial.
- The domain of the random variable Y_n are all n–tuples of 0s and 1s. The image is $\{0, 1, 2, \ldots n\}$.
- As already demonstrated in the previous lecture, the probability distribution of Y_n is

$$\begin{aligned} p_k =& P(Y_n = k) = p_{Y_n}(k) \\ =& \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } 0 \leq k \leq n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- The binomial distribution is often denoted as $b(k; n, p) = p(k)$ and represents the probability that there are k successes in a sequence of n bernoulli trials with probability of success p.
- In example, $b(3; 5, 0.5) = \binom{5}{3}(1/2)^3(1/2)^2 = 0.3125$

Binomial probability distribution

After specifying the distribution of a random variable we should verify that this function is a valid probability distribution, i.e. to verify properties (p1) and (p2). While (p1) is usually clear (it is easy to see that the function is nonnegative), the property (p2) may be not so straightforward and should be verified explicitly.

Binomial probability distribution

- The name 'binomial' comes from the equation verifying that the probabilities sum to 1

$$\sum_{i=0}^{n} p_i = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$
$$= [p + (1-p)]^n = 1.$$

- The corresponding distribution function, denoted by $B(t; n, p)$ is given by

$$B(t; n, p) = F_{Y_n}(t) = \sum_{i=0}^{\lfloor t \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

Binomial probability distribution

We can apply the binomial model when the following conditions hold:

- Each trial has exactly two mutually exclusive outcomes.
- The probability of 'success' is constant on each trial.
- The outcomes of successive trials are mutually independent.

When n becomes very large, it is difficult to calculate the binomial probability distribution, however, we can approximate it using the Laplace (or normal) approximation, which states that when n approaches infinity

$$b(k;n,p) \simeq \frac{1}{\sqrt{2\pi np(1-p)}} e^{\frac{-(k-np)^2}{2np(1-p)}}.$$

We omit proof of this statement.

The difference between the value of the normal approximation and the binomial distribution depend both on values of n and p. For fixed n the approximation is best when $p = 0.5$.

Geometric probability distribution

- The geometric probability distribution origins from Bernoulli trials as well, but this time we count the number of trials until the first 'success' occurs.
- The sample space consists of binary strings of the form $S = \{0^{i-1}1 | i \in \mathbb{N}\}$.
- We define the random variable $Z : \{0^i 1 | i \in \mathbb{N}_0\}$ as $Z(0^{i-1}1) \overset{def}{=} i$.
- Z is the number of trials up to and including the first success.
- The outcome $0^{i-1}1$ arises from a sequence of independent Bernoulli trials, thus we have

$$p_Z(i) = (1-p)^{i-1} p. \tag{1}$$

We use the formula for the sum of geometric series to obtain (verify property (p2))

$$\sum_{i=1}^{\infty} p_Z(i) = \sum_{i=1}^{\infty} p(1-p)^{i-1} = \frac{p}{1-(1-p)} = \frac{p}{p} = 1.$$

Geometric probability distribution

- A random variable with the image $\{1, 2 \dots\}$ and the probability distribution (1) is said to have the **geometric distribution**.
- The corresponding probability distribution function is defined by (for $t \geq 0$)

$$F_Z(t) = \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} = 1 - (1-p)^{\lfloor t \rfloor}.$$

- We are often interested in the number of failures before the first success, let us denote this by the random variable X with the image $\{0, 1, 2 \dots\}$.
- The random variable X is said to have the **modified geometric distribution** if the corresponding probability distribution is

$$p_X(i) = p(1-p)^i.$$

- The distribution function is (for $t \geq 0$)

$$F_X(t) = \sum_{i=0}^{\lfloor t \rfloor} p(1-p)^i = 1 - (1-p)^{\lfloor t+1 \rfloor}$$

Negative binomial probability distribution

- Let us observe the number of trials before the r–th success occurs (in contrast to 1st success in geometric distribution).
- Let T_r be the random variable representing this number.
- Let us define the following events
  - $A =' T_r = n'$.
  - $B =$'Exactly $(r-1)$ successes occur in $n-1$ trials.'
  - $C =$'the nth trial results in a success.'
- We have that $A = B \cap C$, and B and C are independent giving $P(A) = P(B)P(C)$.
- Consider a particular sequence of $n-1$ trials with $r-1$ successes and $n-1-(r-1) = n-r$ failures. The probability associated with each such sequence is $p^{r-1}(1-p)^{n-r}$ and there are $\binom{n-1}{r-1}$ such sequences. Therefore

$$P(B) = \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}.$$

Negative binomial probability distribution

- Since $P(C) = p$ we have

$$\begin{aligned} p_{T_r}(n) =& P(T_r = n) = P(A) \\ =& \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad n = r, r+1, r+2, \ldots \end{aligned}$$

- This can be rewritten (see e.g. Knuth: The art of computer programming, vol. 1, p. 57) to

$$p_{T_r}(n) = p^r \binom{-r}{n-r} (-1)^{n-r} (1-p)^{n-r}.$$

- This formula holds for any $r \in \mathbb{R}$, but the interpretation with the number of successes is no longer valid.
- The special case $r = 1$ gives the geometric distribution.

Negative binomial probability distribution

- To verify that the probabilities sum to 1 we use the Taylor expansion of $(1-t)^{-r}$ for $-1 < t < 1$

$$(1-t)^{-r} = \sum_{n=r}^{\infty} \binom{-r}{n-r} (-t)^{n-r}.$$

The substitution $t = (1-p)$ gives

$$p^{-r} = \sum_{n=r}^{\infty} \binom{-r}{n-r} (-1)^{n-r} (1-p)^{n-r}$$

and the desired result is

$$1 = \sum_{n=r}^{\infty} p^r \binom{-r}{n-r} (-1)^{n-r} (1-p)^{n-r}.$$

Note that the summation from r is correct since clearly $p_{T_r}(n) = 0$ for $n < r$.

Negative binomial probability distribution

There is also the **modified negative binomial distribution** describing the number of failures until the rth success occurs. The probability distribution

$$p_Z(n) = \binom{n+r-1}{r-1} p^r (1-p)^n, \quad n \geq 0.$$

For $r = 1$ we obtain the modified geometric distribution.

Poisson probability distribution

- Let us consider a time interval $\Delta t = (0, t]$ and observe the number of messages received by a particular server. Depending on the usual traffic we can expect that during the time interval Δt there arrive $\lambda \Delta t$ messages, where λ is a parameter derived from the average traffic on the server. It is clear that if Δt is sufficiently small, there is only a negligible probability that two messages arrive within the interval Δt.
- Let us divide the interval into n parts of length t/n and suppose that an arrival of a message in one time interval is independent of message arrival in any other (non-overlapping) time interval. For sufficiently large n we can consider n interval as constituting n Bernoulli trials with the probability of success $p = \lambda t/n$.

Poisson probability distribution

Using this model the probability of k messages arriving in n intervals of a duration t/n is approximately given by

$$b\left(k; n, \frac{\lambda t}{n}\right) = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k}, \quad k = 0, 1, \dots n.$$

Since the assumption that more than one message arrives in a given time interval is reasonable only for large n we calculate the limit of probability distribution as n approaches ∞. We start with

$$\begin{aligned} b\left(k; n, \frac{\lambda t}{n}\right) =& \frac{n(n-1)(n-2)\dots(n-k+1)}{k!n^k}(\lambda t)^k\left(1-\frac{\lambda t}{n}\right)^{n-k} \\ =& \frac{n}{n}\cdot\frac{n-1}{n}\dots\frac{n-k+1}{n}\cdot\frac{(\lambda t)^k}{k!}\left(1-\frac{\lambda t}{n}\right)^{-k}\left(1-\frac{\lambda t}{n}\right)^{n}. \end{aligned}$$

Poisson probability distribution

Let us now observe what happens when n goes to infinity. The first k factors approach 1, the next factor is fixed, the next approaches 1 and the last factor we rewrite as

$$\lim_{n\to\infty}\left[\left(1-\frac{\lambda t}{n}\right)^{-n/(\lambda t)}\right]^{-\lambda t}.$$

Substituting $-\lambda t/n = h$ we have

$$\left[\lim_{h\to 0}(1+h)^{1/h}\right]^{-\lambda t} = e^{-\lambda t}$$

since the expression in the parentheses is the definition of euler constant e. Thus the binomial probability distribution approaches

$$\lim_{n\to\infty} b\left(k; n, \frac{\lambda t}{n}\right) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, 2 \ldots$$

Poisson probability distribution

By replacing λt by α we get the well known **Poisson distribution**

$$f(k;\alpha) = e^{-\alpha}\frac{\alpha^k}{k!}.$$

Following our derivations we can use the Poisson distribution to approximate the binomial distribution when n is large and p is small (compare to normal approximation):

$$\binom{n}{k}p^k(1-p)^{n-k} \simeq e^{-\alpha}\frac{\alpha^k}{k!}, \text{ where } \alpha = np.$$

Poisson probability distribution

It remains to verify that the probabilities sum to 1 (they are obviously positive)

$$\sum_{k=0}^{\infty} f(k;\alpha) = \sum_{k=0}^{\infty} e^{-\alpha}\frac{\alpha^k}{k!} = e^{-\alpha}\sum_{k=0}^{\infty}\frac{\alpha^k}{k!} = e^{-\alpha}e^{\alpha} = 1,$$

since $\sum_{k=0}^{\infty}\frac{\alpha^k}{k!}$ is the Maclaurin expansion of e^{α}.
The probabilities $f(k;\alpha)$ are easily calculable using

$$f(0;\alpha) = e^{-\alpha}$$

and the recurrent relation

$$f(k+1;\alpha) = \frac{\alpha f(k;\alpha)}{k+1}.$$

Hypergeometric probability distribution

- The Hypergeometric probability distribution is close in its interpretation to the binomial probability distribution except that we consider sampling without replacement.
- Let us suppose we have two kinds of objects - e.g. r red and $n-r$ black socks in a basket. We have the probability r/n to select a red sock in the first trial. However, the probability of selecting red sock in the second trial is $(r-1)/(n-1)$ if red sock was selected in the first trial, or $r/(n-1)$ if black sock was selected in the first trial. It follows that the assumption of constant probability of every outcome in all trials, as required by the binomial distribution, does not hold. Also, the trials are not independent.
- In this case we are facing the **hypergeometric distribution** $h(k; m, r, n)$ defined as the probability that k red objects in a set of m object chosen randomly without replacement from n objects containing r red objects.

Hypergeometric probability distribution

- There are $\binom{n}{m}$ sample points.
- The k red socks can be selected from r red socks in $\binom{r}{k}$ ways and $m-k$ black socks can be selected from $n-r$ in $\binom{n-r}{m-k}$ ways.
- The sample of m socks with k red ones can be selected in $\binom{r}{k}\binom{n-r}{m-k}$ ways.
- Assuming uniform probability distribution on the sample space the required probability is

$$h(k;m,r,n) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}, \quad k = 1, 2, \dots \min\{r, m\}$$

- Good approximation of the hypergeometric distribution for large n is the binomial distribution

$$h(k;m,r,n) \simeq b(k;m,r/n).$$

Part II

Discrete random vectors

Discrete random vectors

- Suppose we want to study relationship between two or more random variables defined on a given sample space.
- Let $X_1, X_2, \ldots X_r$ be r discrete random variables defined on a sample space S.
- For each sample point $s \in S$, each of the random variables $X_1, X_2, \ldots X_r$ takes on one of its possible values

$$X_1(s) = x_1, X_2(s) = x_2, \ldots X_r(s) = x_r.$$

- The random vector $\mathbf{X} = (X_1, X_2, \ldots X_r)$ is an r-dimensional vector-valued function $\mathbf{X} : S \to \mathbb{R}^r$ with $\mathbf{X}(s) = \mathbf{x} = (x_1, x_2, \ldots x_r)$.

Discrete random vectors

Definition

The **joint** (or **compound**) **probability distribution** of a random vector $\mathbf{X}$ is defined to be

$$p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots X_r = x_r).$$

The properties of random vectors are

(j1) $p_{\mathbf{X}}(\mathbf{x}) \geq 0,\ \mathbf{x} \in \mathbb{R}^r$.

(j2) $\{\mathbf{x} | p_{\mathbf{x}}(\mathbf{x}) \neq 0\}$ is a finite or countably infinite subset of $\mathbb{R}^r$, which will be denoted as $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$.

(j3) $P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} p_{\mathbf{X}}(\mathbf{x})$.

(j4) $\sum_i p_{\mathbf{X}}(\mathbf{x}_i) = 1$.

Marginal probability distributions

- In situation when we are examining more that one random variable, the probability distribution of a single variable, e.g. $p_X(x)$, is referred to as **marginal probability distribution** (in contrast to joint probability distribution).
- Considering joint probability distribution $p_{X,Y}(x,y)$ of random variables X and Y we can calculate the marginal probability distribution of X as

$$\begin{aligned} p_X(x) =& P(X = x) = P\left(\bigcup_j \{X = x, Y = y_j\}\right) \\ =& \sum_j P(X = x, Y = y_j) = \sum_j p_{X,Y}(x, y_j). \end{aligned}$$

Similarly we obtain the marginal probability distribution of Y

$$p_Y(y) = \sum_i p_{X,Y}(x_i, y).$$

Marginal probability distributions

- While it is relatively easy to calculate the marginal probability distributions from the joint distribution, in general there is no way how to determine the joint distribution from corresponding marginal distributions.
- The only exception are independent random variables (see below), when the joint probability distribution is the product of marginal distributions.

Multinomial probability distribution

- Interesting example of joint probability distribution is the multinomial distribution.
- Consider a sequence of n generalized bernoulli trials, where each of them has a finite number r of outcomes having probabilities $p_1, p_2, \dots p_r$.
- Let us define the random vector $\mathbf{X} = (X_1, X_2, \dots X_r)$ such that X_i is the number of trials that resulted in ith outcome.
- Then the compound probability distribution of $\mathbf{X}$ is

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{n}) =& P(X_1 = n_1, X_2 = n_2, \dots X_r = n_r) \\ =& \binom{n}{n_1, n_2, \dots n_r} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}, \end{aligned}$$

where $\mathbf{n} = (n_1, n_2 \dots n_r)$ and $\sum_{i=1}^{r} n_i = n$.

Multinomial probability distribution

The marginal probability distribution of X_i may be computed by

$$
\begin{aligned}
p_{X_i}(n_i) &= \sum_{\mathbf{n}:\left[\left(\sum_{j\neq i} n_j\right)=n-n_i\right]} \binom{n}{n_1, n_2 \ldots n_r} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r} \\
&= \frac{n! p_i^{n_i}}{(n-n_i)! n_i!} \sum_{\mathbf{n}:\left[\left(\sum_{j\neq i} n_j\right)=n-n_i\right]} \frac{(n-n_i)! p_1^{n_1} \ldots p_{i-1}^{n_{i-1}} p_{i+1}^{n_{i+1}} \ldots p_r^{n_r}}{n_1! n_2! \ldots n_{i-1}! n_{i+1}! \ldots n_r!} \\
&= \binom{n}{n_i} p_i^{n_i} (p_1 + \cdots + p_{i-1} + p_{i+1} + \cdots + p_r)^{n-n_i} \\
&= \binom{n}{n_i} p_i^{n_i} (1-p_i)^{n-n_i}.
\end{aligned}
$$

Part III

Independent random variables

Independent random variables

Definition

Two discrete random variables are **independent** provided their joint probability distribution is a product of the marginal probability distributions, i.e.

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ for all } x \text{ and } y.$$

- If X and Y are two independent random variables, then for any two subsets $A, B \subseteq \mathbb{R}$ the events $X \in A$ and $Y \in B$ are independent:

$$P(X \in A \cap Y \in B) = P(X \in A)P(Y \in B)$$

To see this

$$\begin{aligned} P(X \in A \cap Y \in B) &= \sum_{x \in A} \sum_{y \in B} p_{X,Y}(x,y) \\ &= \sum_{x \in A} \sum_{y \in B} p_X(x)p_Y(y) \end{aligned}$$

Independent random variables

Definition

Let $X_1, X_2, \dots X_r$ be discrete random variables with probability distributions $p_{X_1}, p_{X_2}, \dots p_{X_r}$. These random variables are **pairwise independent** if

$$\forall 1 \leq i < j \leq r,\ \forall x_i \in \mathbf{Im}(X_i), x_j \in \mathbf{Im}(X_j),\ p_{X_i,X_j}(x_i, x_j) = p_{X_i}(x_i)p_{X_j}(x_j).$$

Definition

Let $X_1, X_2, \dots X_r$ be discrete random variables with probability distributions $p_{X_1}, p_{X_2}, \dots p_{X_r}$. These random variables are **mutually independent** if for all $x_1 \in \mathbf{Im}(X_1), x_2 \in \mathbf{Im}(X_2), \dots, x_r \in \mathbf{Im}(X_r)$

$$p_{X_1,X_2,\dots X_r}(x_1, x_2, \dots x_r) = p_{X_1}(x_1)p_{X_2}(x_2)\dots p_{X_r}(x_r).$$

Note that pairwise independence of a set of random variables does not imply their mutual independence.

Independent random variables

- Let X and Y be non-negative independent random variables. Then the probability distribution of the random variable $Z = X + Y$ is

$$p_Z(t) = p_{X+Y}(t) = \sum_{x=0}^{t} p_X(x) p_Y(t - x).$$

In case X and Y can also take negative values, the sum should go from $-\infty$ instead of 0.

Independent random variables

Theorem

Let $X_1, X_2, \ldots X_r$ be mutually independent

1. *If X_i has the binomial distribution with parameters n_i and p, then $\sum_{i=1}^{r} X_i$ has the binomial distribution with parameters $n_1 + n_2 + \cdots + n_r$ and p.*
2. *If X_i has the (modified) negative binomial distribution with parameters α_i and p, then $\sum_{i=1}^{r} X_i$ has the (modified) negative binomial probability distribution with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_r$ and p.*
3. *If X_i has the Poisson distribution with parameter α_i, then $\sum_{i=1}^{r} X_i$ has the Poisson distribution with parameter $\sum_{i=1}^{r} \alpha_i$.*