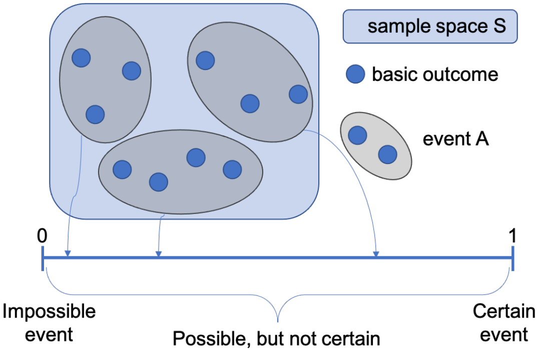
UNIVERSITÀ DEGLI STUDI DI TRIESTE

Statistics

Discrete Random Variables

Luca Pennella

September 12th, 2024



Probability model

A **probability model** is a mathematical description of a random experiment consisting of a sample space and a way of assigning probabilities to events

Random variable

A **random variable** (r.v.) X is a variable whose value represents a numerical outcome of a random phenomenon; that is, it is a well-defined but unknown number

- the number of tails on three coin tosses
- the number of defective items in a sample of 20 items from a large shipment
- the number of students attending the statistics class on Friday
- the delay time of the airplane
- the weight of a newborn
- the duration of a phone call with your mother

Random Variable

Probability distribution

The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values

$$P(x) = P(X = x), \forall x$$

- the number of tails on tree coin tosses: $X : \{0, 1, 2, 3\}$ and each value x has probability $P(X = x)$

There are two main types of random variables: discrete if it has a finite list of possible outcomes, and continuous if it can take any value in an interval.

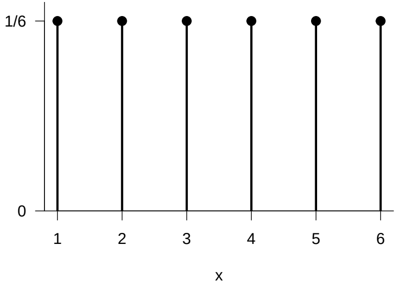
- D the number of tails on three coin tosses
- D the number of defective items in a sample of 20 items from a large shipment
- D the number of students attending the statistics class on Friday
- C the delay time of the airplane
- C the weight of a newborn
- C the duration of a phone call with your mother

For continuous random variables we can assign probabilities only to a range of values, using a mathematical function. This allows us to calculate the probability of events such as "today's high temperature will be between 25° and 26°."

Discrete Random Variables

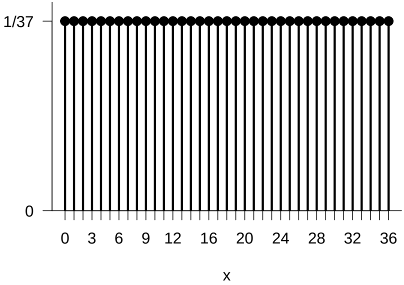
X = rolling a dice

x	P
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$



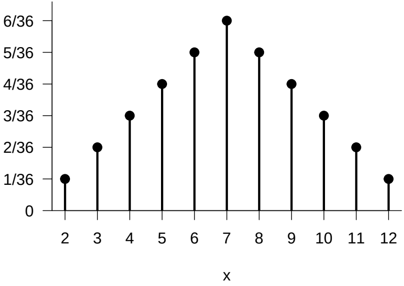
Y = roulette result

y	P
0	$1/37$
1	$1/37$
2	$1/37$
...	...
35	$1/37$
36	$1/37$



$Z =$ sum the results of rolling two dice

z	P
2	$1/36$
3	$2/36$
4	$3/36$
5	$4/36$
6	$5/36$
7	$6/36$
8	$5/36$
9	$4/36$
10	$3/36$
11	$2/36$
12	$1/36$



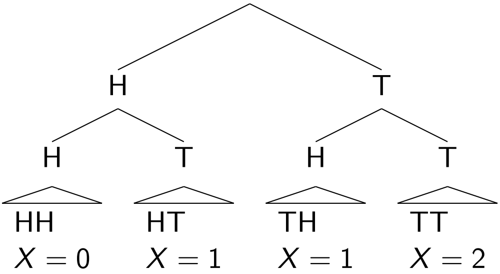
Number of tails on two flips of a coin

We toss a coin two times, then we sum the number of tails T

$X =$ number of tails in flipping a coin two times

X is a discrete random variable that can assume values: $\{0, 1, 2\}$

The random experiment is represented in the tree diagram:



4 possible outcomes $= 2^2$

Number of tails on two flips of a coin

We toss a coin two times, then we sum the number of tails T

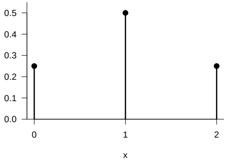
$X =$ number of tails in tossing a coin two times

X is a discrete random variable that can assume values: $\{0, 1, 2\}$

Given a fair coin, the probability distribution of X is

outcomes	P	X
HH	$1/4$	0
HT	$1/4$	1
TH	$1/4$	1
TT	$1/4$	2

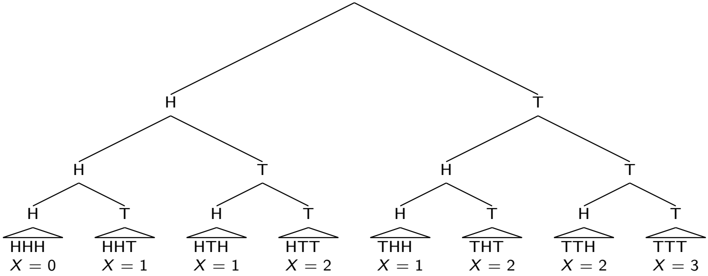
X	P
0	$P(HH) = 1/4$
1	$P(HT \cup TH) = 1/2$
2	$P(TT) = 1/4$



Number of tails on three flips of a coin

X = number of tails in tossing a coin three times

X is a discrete random variable that can assume values: $\{0, 1, 2, 3\}$



8 possible outcomes $= 2^3$

Number of tails on three flips of a coin

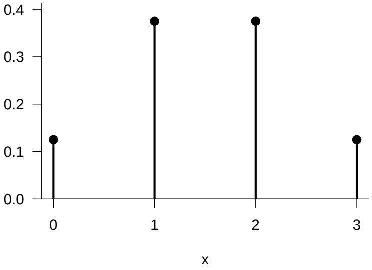
X = number of tails in tossing a coin three times

X is a discrete random variable that can assume values: $\{0, 1, 2, 3\}$

Assuming a fair coin, the probability distribution of X is

outcomes	P	X
HHH	$1/8$	0
HHT	$1/8$	1
HTH	$1/8$	1
THH	$1/8$	1
HTT	$1/8$	2
TTH	$1/8$	2
THT	$1/8$	2
TTT	$1/8$	3

X	P
0	$1/8$
1	$3/8$
2	$3/8$
3	$1/8$



$$P(X = 2) = P(HTT \cup TTH \cup THT) = P(HTT) + P(TTH) + P(THT)$$

Number of tails on n flips of a coin

$X =$ number of tails in tossing a coin n times

X is a discrete random variable that can assume values: $\{0, 1, 2, \ldots, n\}$

There are 2^n possible outcomes

a generic outcome	T	H	T	T	H	T	$\ldots$	T	H
flip	1	2	3	4	5	6	...	$n-1$	n

Given a fair coin, each outcome (sequence of n trials) has probability $\left(\frac{1}{2}\right)^n$

To compute $P(X = x)$ we have to count how many outcomes with x tails we can obtain in the random experiment:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Then, the probability distribution is:

$$P(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^n$$

Number of tails on n flips of a coin

$X =$ number of tails in tossing a coin n times

X is a discrete random variable that can assume values: $\{0, 1, 2, \dots, n\}$

There are 2^n possible outcomes

a generic outcome	T	H	T	T	H	T	$\dots$	T	H
flip	1	2	3	4	5	6	...	$n-1$	n
prob	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	...	$\frac{1}{2}$	$\frac{1}{2}$

Given a fair coin, each outcome (sequence of n trials) has probability $\left(\frac{1}{2}\right)^n$

To compute $P(X = x)$ we have to count how many outcomes with x tails we can obtain in the random experiment:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Then, the probability distribution is:

$$P(X = x) = \binom{n}{x} \left(\frac{1}{2}\right)^n$$

Binomial distribution

Binomial distribution

A random variable X follows the binomial distribution with dimension $n \in \mathbb{N}$ and parameter $p \in [0, 1]$

$$X \sim \text{Binom}(n, p)$$

if $X \in \{0, 1, \dots, n\}$ and

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$X \sim \text{Binom}(n, p)$ is the number of successes in n independent trials with success probability p

- the number of observations/trials n is fixed
- the n observations are independent
- each observation can be a success or a failure

Blood Types

Genetic says that children receive genes from their parents independently

Each child of a particular pair of parents has a probability 0.25 of having type "0" blood

If these parents have 5 children, the number who have type "0" blood is the count X of successes in 5 independent observations with probability 0.25 of success in each observation

So X has the Binomial distribution with $n = 5$ and $p = 0.25$

$$X \sim \text{Binom}(5, 0.25)$$

$$P(X = x) = \binom{5}{x} 0.25^x (1 - 0.25)^{5-x}$$

Blood Types

X has the Binomial distribution with $n = 5$ and $p = 0.25$

$$X \sim \text{Binom}(5, 0.25)$$

$$P(X = x) = \binom{5}{x} 0.25^x (1 - 0.25)^{5-x}$$

What is the probability that two children have type "0" blood?

$$P(X = 2) = \binom{5}{2} 0.25^2 (1 - 0.25)^{5-2}$$

What is the probability that more than 4 children have type "0" blood?

$$P(X > 4) = P(X = 5) = 0.25^5$$

Probability mass function with countably finite support

Probability mass function with countably finite support

Given a random variable X with finite support $\{x_1, x_2, \ldots, x_n\}$, we define the probability mass function of the rv X

$$P(X = x_i) = p(x_i), \forall i$$

such that

i. $p(x_i) \geq 0$

ii. $\sum_{i=1}^{n} p(x_i) = 1$

Probability mass function with countably infinite support

Probability mass function with countably infinite support

Given a random variable X that assumes a countably infinite set of values $\{x_1, x_2, \dots, x_n, \dots\}$, we define its probability mass function as

$$P(X = x_i) = p(x_i), \forall i$$

such that

i. $p(x_i) \geq 0$
ii. $\sum_{i=1}^{\infty} p(x_i) = 1$ (that is, the series must converge to 1)

Cumulative distribution function - discrete rv

Cumulative distribution function - discrete rv

Given a random variable X that assumes a countably infinite set of values $x_1, \ldots, x_n, \ldots$ and with probability mass function $p(x)$, we define the cumulative distribution function of X as

$$F(x) = P(X \leq x) = \sum_{i : x_i \leq x} p(x_i)$$

The cumulative distribution function represents the probability that X does not exceed the value x

i. $F(x) \geq 0, \quad \forall x \in \mathbb{R}$;

ii. $F(x)$ is non decreasing;

iii. $\lim\limits_{x \to -\infty} F(x) = 0$;

iv. $\lim\limits_{x \to +\infty} F(x) = 1$.

Assume that X is a discrete random variable that follows a Binomial distribution with $n = 4$ and $p = 0.4$, then

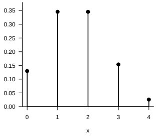
$$X \in \{0, 1, 2, 3, 4\}$$

and the probability mass function of X is

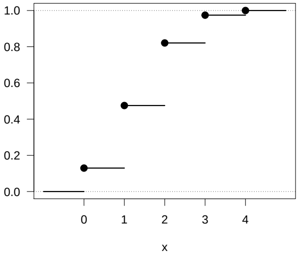
$$P(X = x_i) = \binom{4}{x_i} p^{x_i} (1 - p)^{4 - x_i}$$

x_i	p_i	F_i
0	0.12960	0.1296
1	0.34560	0.4752
2	0.34560	0.8208
3	0.15360	0.9744
4	0.02560	1.0000

Probability mass function



Cumulative distribution function



In order to obtain a measure of the center of a probability distribution, we introduce the notion of the expectation of a random variable

You know the sample mean as a measure of central location for sample data

The expected value is the corresponding measure of central location for a random variable

In order to obtain a measure of the center of a probability distribution, we introduce the notion of the expectation of a random variable

You know the sample mean as a measure of central location for sample data

The expected value is the corresponding measure of central location for a random variable

Let X be the number of errors on a page chosen at random from business area textbooks, from a review we found that 81% of all pages were error-free ($X = 0$), 17% of all pages contained one error ($X = 1$), and the remaining 2% contained two errors ($X = 2$).
Thus, the probability mass function of the variable X is

$$p(0) = 0.81, \quad p(1) = 0.17, \quad p(2) = 0.02$$

Let X be the number of errors on a page chosen at random from business area textbooks, from a review we found that 81% of all pages were error-free ($X = 0$), 17% of all pages contained one error ($X = 1$), and the remaining 2% contained two errors ($X = 2$).
Thus, the probability mass function of the variable X is

$$p(0) = 0.81, \quad p(1) = 0.17, \quad p(2) = 0.02$$

What is the expected value of X?
In computing the average number of possible values, $E(X) = (0 + 1 + 2)/3 = 1$, we are ignoring how each value is likely to occur (assuming the same probability on each value)

$$E(X) = 0 \cdot 0.81 + 1 \cdot 0.17 + 2 \cdot 0.02 = \sum_x xp(x) = 0.21$$

Expectation

The expected value E(X), of a discrete random variable X is defined as

$$E(X) = \mu = \sum_{i=1}^{\infty} x_i p(x_i)$$

Using the definition of relative frequency probability, we can view the expected value of a rv as the long-run weighted average value that it takes over a large number of trials

Variance

The variance V(X), of a discrete random variable X is defined as the expectation of the squared deviations about the mean, $(X - E(X))^2$

$$V(X) = \sigma^2 = E[(X - E(X))^2] = \sum_{i=1}^{\infty} (x_i - E(x))^2 p(x_i)$$

$$V(X) = E(X^2) - [E(X)]^2$$

The standard deviation σ is the positive square root of the variance

Binomial: expected value and variance

It can be shown that for a Binomial rv X with dimension n and probability p, that is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

then

$$E(X) = np$$
$$V(X) = np(1-p)$$

Binomial: expected value and variance

It can be shown that for a Binomial rv X with dimension n and probability p, that is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

then

$$E(X) = np$$
$$V(X) = np(1-p)$$

Overbooking example:

A small airline accepts reservations for a flight with 20 seats and knows that of the people who book a trip 10% do not show up

What is the expected number of people that show up at the airport?

Assuming $X \sim Binom(20, 0.9)$

$$E(X) = np = 20 \cdot 0.9 = 18$$

Linear transformations

We defined random variables as numbers, arithmetical operations are allowed

e.g. given a random variable X we can define a new rv Y applying a linear transformation

$$Y = aX + b$$

The values that the rv Y can assume and its probability distribution are derived from the ones of X

If X assumes values $\{x_i\}$, then $Y = aX + b$ assumes values $\{ax_i + b\}$, and the probability distribution of Y is

$$P(Y = ax_i + b) = P(X = x_i)$$

Also,

$$E(Y) = E(aX + b) = aE(X) + b$$

$$V(Y) = V(aX + b) = a^2 V(X)$$

Standardization

Given a rv X with mean $\mu = E(X)$ and variance $\sigma^2 = V(X)$, the standardization is the linear transformation

$$Z = \frac{X - \mu}{\sigma}$$

such that, the rv Z has a mean equal to 0 and variance (and standard deviation) equal to 1

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0$$

$$V(Z) = V\left(\frac{X - \mu}{\sigma}\right) = \frac{V(X)}{\sigma^2} = 1$$

the linear transformation $Z = a + Xb$, a and b are defined as: $a = -\frac{\mu}{\sigma}$ and $b = 1/\sigma$

- The number of failures in a large computer system during a given day
- The number of replacement orders for a part received by a firm in a given month
- The number of ships arriving at a loading facility during a 6-hour loading period
- The number of delivery trucks to arrive at a central warehouse in an hour
- The number of customers to arrive at a checkout aisle in your local grocery store during a particular time interval

- The number of failures in a large computer system during a given day
- The number of replacement orders for a part received by a firm in a given month
- The number of ships arriving at a loading facility during a 6-hour loading period
- The number of delivery trucks to arrive at a central warehouse in an hour
- The number of customers to arrive at a checkout aisle in your local grocery store during a particular time interval

All the random phenomena above describe the number of independent occurrences (successes) on a given interval of time

Poisson distribution

Poisson distribution

A random variable $X \in \{0, 1, 2, \ldots, n, \ldots\}$ follows a Poisson distribution with parameter λ if and only if

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$X \sim Poisson(\lambda)$ is the number of occurrences/successes of a certain event in a given continuous interval (such as time, surface area, or length)

- assume that the interval is divided into a large number of equal subintervals each with a very small probability of occurrence of an event
- the probability of the occurrence of an event is constant for all subintervals
- there can be no more than one occurrence in each subinterval
- occurrences are independent; that is, an occurrence in one interval does not influence the probability of an occurrence in another interval

Poisson: expected value and variance

It can be shown that for a Poisson rv X with parameter λ, that is

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

then

$$E(X) = \lambda, \qquad V(X) = \lambda$$

Thus λ represents the expected number of successes per space unit and it can assume only positive values

$\lambda = 1$

x_i	p_i
0	0.36788
1	0.36788
2	0.18394
3	0.06131
4	0.01533
5	0.00307
6	0.00051
7	0.00007
8	0.00001
9	0.00000
10	0.00000
> 10	0.00000

$$E(X) = 1$$

$$V(X) = 1$$

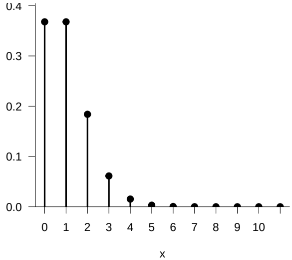
$\lambda = 0.2$

x_i	p_i
0	0.81873
1	0.16375
2	0.01637
3	0.00109
4	0.00005
5	0.00000
6	0.00000
7	0.00000
8	0.00000
9	0.00000
10	0.00000
> 10	0.00000

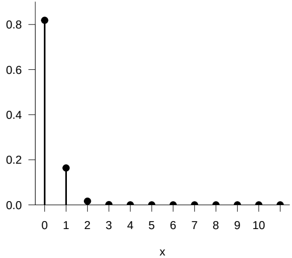
$$E(X) = 0.2$$

$$V(X) = 0.2$$

$\lambda = 1$



$\lambda = 0.2$



Football

A football team scores a number of goals per game that is assumed to be distributed as a Poisson distribution and on average, the team scores 1.5 goals per game

1. Compute the probability that in the next game, the number of goals by the football team is 0
2. Compute the probability that in the next game, the number of goals by the football team is greater than 4

Football

A football team scores a number of goals per game that is assumed to be distributed as a Poisson distribution and on average, the team scores 1.5 goals per game

1. Compute the probability that in the next game, the number of goals by the football team is 0
2. Compute the probability that in the next game, the number of goals by the football team is greater than 4

The number of goals per game follows a Poisson distribution with parameter $\lambda = 1.5$, thus

$$P(X = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda} = 0.2231$$

Football

A football team scores a number of goals per game that is assumed to be distributed as a Poisson distribution and on average, the team scores 1.5 goals per game

1. Compute the probability that in the next game, the number of goals by the football team is 0
2. Compute the probability that in the next game, the number of goals by the football team is greater than 4

$$
\begin{aligned}
P(X > 4) &= P\left(\bigcup_{i=5}^{+\infty}(X = i)\right) = \sum_{i=5}^{+\infty} \frac{\lambda^i}{i!} e^{-\lambda} \\
&= 1 - \sum_{i=0}^{4} \frac{\lambda^i}{i!} e^{-\lambda} = 1 - 0.9814 = 0.01858
\end{aligned}
$$

Uniform discrete distribution

Uniform discrete distribution

A random variable $X \in \{a, a+1, a+2, \dots, b-2, b-1, b\}$ follows a discrete Uniform distribution on the interval $[a, b]$ if and only if its probability mass function is

$$P(X = x) = \frac{1}{n}$$

where a and b are integer numbers such that $a \leq b$ and $n = b - a + 1$

If $X \sim Unif\{a, b\}$ all the values of the support are equally likely to be observed

The expected value and variance of a rv X following a discrete Uniform distribution are respectively

$$E(X) = \frac{a+b}{2} \qquad V(X) = \frac{n^2 - 1}{12}$$

Toss a coin

If we flip a coin three times we can define several random variables such as
S = number of tails
M = number of tails before the first head

Outcome	P	S	M
HHH	1/8	0	0
HHT	1/8	1	0
HTH	1/8	1	0
THH	1/8	1	1
HTT	1/8	2	0
TTH	1/8	2	2
THT	1/8	2	1
TTT	1/8	3	3

S	P
0	1/8
1	3/8
2	3/8
3	1/8

M	P
0	4/8
1	2/8
2	1/8
3	1/8

We now want to look at them jointly, so we consider events as

$$(S = s) \cap (M = m)$$

Toss a coin

If we flip a coin three times we can define several random variables such as
$S =$ number of tails
$M =$ number of tails before the first head

Outcome	P	S	M
HHH	1/8	0	0
HHT	1/8	1	0
HTH	1/8	1	0
THH	1/8	1	1
HTT	1/8	2	0
TTH	1/8	2	2
THT	1/8	2	1
TTT	1/8	3	3

		S			
		0	1	2	3
M	0	HHH	HHT HTH	HTT	-
	1	-	THH	THT	-
	2	-	-	TTH	-
	3	-	-	-	TTT

We now want to look at them jointly, so we consider events as

$$(S = s) \cap (M = m)$$

Toss a coin

If we flip a coin three times we can define several random variables such as
S = number of tails
M = number of tails before the first head

Outcome	P	S	M
HHH	1/8	0	0
HHT	1/8	1	0
HTH	1/8	1	0
THH	1/8	1	1
HTT	1/8	2	0
TTH	1/8	2	2
THT	1/8	2	1
TTT	1/8	3	3

		S			
		0	1	2	3
M	0	1/8	2/8	1/8	0
	1	0	1/8	1/8	0
	2	0	0	1/8	0
	3	0	0	0	1/8

We now want to look at them jointly, so we consider events as

$$(S = s) \cap (M = m)$$