Probability

Computer Science Tripos, Part IA

R.J. Gibbens

Computer Laboratory
University of Cambridge

Lent Term 2012/13

Last revision: 2013-01-07/r-56

Outline

- Elementary probability theory (2 lectures)
  - Preamble on set theory, probability spaces, random variables, discrete/continuous distributions, means and variances, independence, conditional probabilities, Bayes's theorem.
- Probability generating functions (1 lecture)
  - Definitions and properties; use in calculating moments of random variables and for finding the distribution of sums of independent random variables.
- Multivariate distributions and independence (1 lecture)
  - Random vectors and independence; joint and marginal density functions; variance, covariance and correlation; conditional density functions.
- Elementary stochastic processes (2 lectures)
  - Simple random walks; recurrence and transience; the Gambler's Ruin Problem and solution using difference equations.
- Case studies (2 lectures)
  - A selection of illustrative applications in Computer Science.

Reference books

- (*) Grimmett, G. & Welsh, D.
  Probability: an introduction.
  Oxford University Press, 1986

- Ross, Sheldon M.
  Probability Models for Computer Science.
  Harcourt/Academic Press, 2002

Elementary probability theory

Preamble on set theory

Our approach to Probability will neccessitate formulating concepts based on notions of sets of random outcomes and so we will begin with a brief look at some basic ideas of sets.

Mathematically, a set is just a collection of items. The items may be of any type and there need be no notion of ordering between the items. However, what is essential is the notion of membership. Any particular item, a, say, may be a member of some set A and this is written $a \in A$.

Sets can be specified either by a property that all members need to satisfy or by explicitly listing all the items. For example, $\{$even numbers$\}$, $\{$odd numbers$\}$, $\{1,2,3\}$, $\{2,4,6,\ldots\}$ and $\{1,3,5,\ldots\}$ are all examples of sets.

The empty set $\{\}$ contains no members and is often written as $\emptyset$.

Preamble on set theory, ctd

Given two sets A and B it may be that all the items in A are also in B. In this case, we say that A is a subset of B and that B is a superset of A. We use the notation $A \subset B$ to denote the subset relation between sets. If $A \subset B$ and $B \subset A$ then the two sets are equal, written $A = B$. We can write $A \subseteq B$ if we want to emphasize that A and B may be equal as well as $A \subset B$.

We may define various operations to combine sets. The intersection of two sets A and B, written $A \cap B$, is set of items which are members of both A and B. The union of A and B, written $A \cup B$, is the set of items which are members of either A or B (or both).

Preamble on set theory, ctd

The complement of a set A is the collection of items which are not members of A and we write $a \notin A$ when item a is not a member of A. The complement of A can be written as

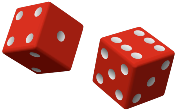
$$A^c = \{a | a \notin A\} .$$

The complement is taken relative to some universal set Ω, say, of possible items of interest. The difference between two sets A and B, written $A \setminus B$ is defined as

$$A \setminus B = \{a \in A | a \notin B\} .$$

So then $A^c = \Omega \setminus A$.
Finally, note that the set $A = \{a\}$ is a set containing one item a and so is different from the item a itself. If we consider the set $B = \{a, b\}$ consisting of the two distinct items a and b then we can say that $a \in B$ and that $A \subset B$. We can not say that $A \in B$ or that $a \subset B$.

Random experiments



We will describe randomness by conducting experiments (or trials) with uncertain outcomes. The set of all possible outcomes of an experiment is called the sample space and is denoted by Ω.

Identify random events with particular subsets of Ω and write

$$\mathscr{F} = \{E \mid E \subseteq \Omega \text{ is a random event}\}$$

for the collection of possible events.
For each such random event, $E \in \mathscr{F}$, we will associate a number called its probability, written $\mathbb{P}(E) \in [0, 1]$.
Before introducing probabilities we need to look closely at our notion of collections of random events.

Event spaces

We formalize the notion of an event space, $\mathscr{F}$, by requiring the following to hold.

Definition (Event space)

1. $\mathscr{F}$ is non-empty
2. $E \in \mathscr{F} \Rightarrow \Omega \setminus E \in \mathscr{F}$
3. $(\forall i \in I \,.\, E_i \in \mathscr{F}) \Rightarrow \cup_{i \in I} E_i \in \mathscr{F}$

Example

Ω any set and $\mathscr{F} = \mathscr{P}(\Omega)$, the power set of Ω.

Example

Ω any set with some event $E' \subset \Omega$ and $\mathscr{F} = \{\emptyset, E', \Omega \setminus E', \Omega\}$.

Note that $\Omega \setminus E$ is often written using the shorthand E^c for the complement of E with respect to Ω.

Probability spaces

Given an experiment with outcomes in a sample space Ω with an event space $\mathscr{F}$ we associate probabilities to events by defining a probability function $\mathbb{P} : \mathscr{F} \to \mathbb{R}$ as follows.

Definition (Probability function)

1. $\forall E \in \mathscr{F} \,.\, \mathbb{P}(E) \geq 0$
2. $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$
3. $E_i \in \mathscr{F}$ for $i \in I$ disjoint (that is, $E_i \cap E_j = \emptyset$ for $i \neq j$) then

$$\mathbb{P}\left(\cup_{i \in I} E_i\right) = \sum_{i \in I} \mathbb{P}(E_i)\,.$$

We call the triple $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space.

Examples of probability spaces

- Ω any set with some event $E' \subset \Omega$ ($E' \neq \emptyset$, $E' \neq \Omega$). Take $\mathscr{F} = \{\emptyset, E', \Omega \setminus E', \Omega\}$ as before and define the probability function $\mathbb{P}(E)$by

$$\mathbb{P}(E) = \begin{cases} 0 & E = \emptyset \\ p & E = E' \\ 1 - p & E = \Omega \setminus E' \\ 1 & E = \Omega \end{cases}$$

  for any $0 \leq p \leq 1$.

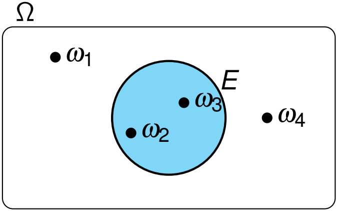
- $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ with $\mathscr{F} = \mathscr{P}(\Omega)$ and probabilities given for all $E \in \mathscr{F}$ by

$$\mathbb{P}(E) = \frac{|E|}{n}.$$

- For a six-sided fair die $\Omega = \{1, 2, 3, 4, 5, 6\}$ we take $\mathscr{F} = \mathscr{P}(\Omega)$ and

$$\mathbb{P}(\{i\}) = \frac{1}{6} \qquad i = 1, 2, \dots, 6.$$

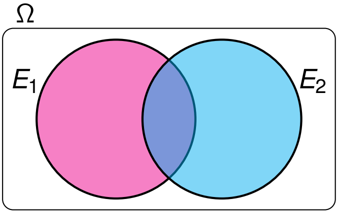
Examples of probability spaces, ctd



- More generally, for each outcome $\omega_i \in \Omega$ ($i = 1, \ldots, n$) assign a value p_i where $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. If $\mathscr{F} = \mathscr{P}(\Omega)$ then take

$$\mathbb{P}(E) = \sum_{i:\omega_i \in E} p_i \qquad \forall\, E \in \mathscr{F}\,.$$

Conditional probabilities



Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and two events $E_1, E_2 \in \mathscr{F}$ how does knowledge that the random event E_2, say, has occurred influence the probability that E_1 has also occurred?
This question leads to the notion of conditional probability.

Definition (Conditional probability)

If $\mathbb{P}(E_2) > 0$, define the conditional probability, $\mathbb{P}(E_1|E_2)$, of E_1 given E_2 by

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \,.$$

Note that $\mathbb{P}(E_2|E_2) = 1$.
Exercise: check that for any $E' \in \mathscr{F}$ such that $\mathbb{P}(E') > 0$ then $(\Omega, \mathscr{F}, \mathbb{Q})$ is a probability space where $\mathbb{Q} : \mathscr{F} \to \mathbb{R}$ is defined by

$$\mathbb{Q}(E) = \mathbb{P}(E|E') \qquad \forall\, E \in \mathscr{F} \,.$$

Independent events

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we can define independence between random events as follows.

Definition (Independent events)

Two events, $E_1, E_2 \in \mathscr{F}$ are independent if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$

Otherwise, the events are dependent. Note that if E_1 and E_2 are independent events then

$$\begin{aligned}\mathbb{P}(E_1|E_2) &= \mathbb{P}(E_1)\\ \mathbb{P}(E_2|E_1) &= \mathbb{P}(E_2).\end{aligned}$$

Independence of multiple events

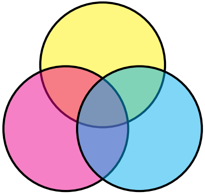
More generally, a collection of events $\{E_i \,|\, i \in I\}$ are independent events if for all subsets J of I

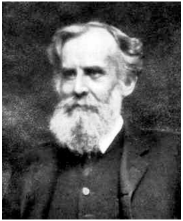
$$\mathbb{P}(\cap_{j\in J} E_j) = \prod_{j\in J} \mathbb{P}(E_j)\,.$$

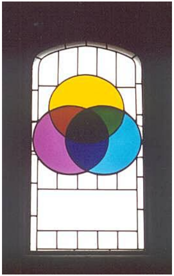
When this holds just for all those subsets J such that $|J| = 2$ we have pairwise independence.
Note that pairwise independence does not imply independence (unless $|I| = 2$).

Venn diagrams

John Venn 1834–1923







Example ($|I| = 3$ events)

E_1, E_2, E_3 are independent events if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$
$$\mathbb{P}(E_1 \cap E_3) = \mathbb{P}(E_1)\mathbb{P}(E_3)$$
$$\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(E_2)\mathbb{P}(E_3)$$
$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_1)\mathbb{P}(E_2)\mathbb{P}(E_3)$$

Bayes' theorem

Thomas Bayes (1702–1761)

Theorem (Bayes' theorem)

If E_1 and E_2 are two events with $\mathbb{P}(E_1) > 0$ and $\mathbb{P}(E_2) > 0$ then

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)}\,.$$

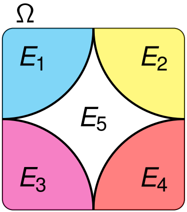
Proof.
We have that

$$\mathbb{P}(E_1|E_2)\mathbb{P}(E_2) = \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_2 \cap E_1) = \mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\,.$$

□

Thus Bayes' theorem provides a way to reverse the order of conditioning.

Partitions



Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ define a partition of Ω as follows.

Definition (Partition)

A partition of Ω is a collection of disjoint events $\{E_i \in \mathscr{F} \mid i \in I\}$ with

$$\cup_{i \in I} E_i = \Omega\,.$$

We then have the following theorem (a.k.a. the law of total probability).

Theorem (Partition theorem)

If $\{E_i \in \mathscr{F} \mid i \in I\}$ is a partition of Ω and $\mathbb{P}(E_i) > 0$ for all $i \in I$ then

$$\mathbb{P}(E) = \sum_{i \in I} \mathbb{P}(E|E_i)\mathbb{P}(E_i) \qquad \forall\, E \in \mathscr{F}\,.$$

Proof of partition theorem

We prove the partition theorem as follows.

Proof.

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}\left(E \cap (\cup_{i\in I} E_i)\right)\\ &= \mathbb{P}\left(\cup_{i\in I}(E \cap E_i)\right)\\ &= \sum_{i\in I} \mathbb{P}(E \cap E_i)\\ &= \sum_{i\in I} \mathbb{P}(E|E_i)\mathbb{P}(E_i)\end{aligned}$$

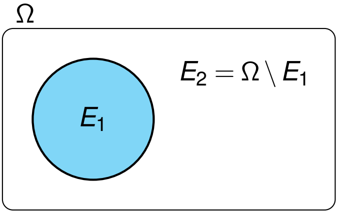
□

Bayes' theorem and partitions

A (slight) generalization of Bayes' theorem can be stated as follows combining Bayes' theorem with the partition theorem.

$$\mathbb{P}(E_i|E) = \frac{\mathbb{P}(E|E_i)\mathbb{P}(E_i)}{\sum_{j\in I}\mathbb{P}(E|E_j)\mathbb{P}(E_j)} \qquad \forall\, i \in I$$

where $\{E_i \in \mathscr{F} \mid i \in I\}$ forms a partition of Ω.



As a special case consider the partition $\{E_1, E_2 = \Omega \setminus E_1\}$.

Then we have

$$\mathbb{P}(E_1|E) = \frac{\mathbb{P}(E|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E|E_1)\mathbb{P}(E_1) + \mathbb{P}(E|\Omega \setminus E_1)\mathbb{P}(\Omega \setminus E_1)}\,.$$

Bayes' theorem example

Suppose that you have a good game of table football two times in three, otherwise a poor game.
Your chance of scoring a goal is 3/4 in a good game and 1/4 in a poor game.

What is your chance of scoring a goal in any given game?
Conditional on having scored in a game, what is the chance that you had a good game?
So we know that

- $\mathbb{P}(\text{Good}) = 2/3$,
- $\mathbb{P}(\text{Poor}) = 1/3$,
- $\mathbb{P}(\text{Score}|\text{Good}) = 3/4$,
- $\mathbb{P}(\text{Score}|\text{Poor}) = 1/4$.

Bayes' theorem example, ctd

Thus, noting that $\{\text{Good}, \text{Poor}\}$ forms a partition of the sample space of outcomes,

$$\begin{aligned}\mathbb{P}(\text{Score}) &= \mathbb{P}(\text{Score}|\text{Good})\mathbb{P}(\text{Good}) + \mathbb{P}(\text{Score}|\text{Poor})\mathbb{P}(\text{Poor}) \\ &= (3/4)\times(2/3) + (1/4)\times(1/3) = 7/12\,.\end{aligned}$$

Then by Bayes' theorem we have that

$$\mathbb{P}(\text{Good}|\text{Score}) = \frac{\mathbb{P}(\text{Score}|\text{Good})\mathbb{P}(\text{Good})}{\mathbb{P}(\text{Score})} = \frac{(3/4)\times(2/3)}{(7/12)} = 6/7.$$

Random variables

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we may wish to work not with the outcomes $\omega \in \Omega$ directly but with some real-valued function of them, say using the function $X : \Omega \to \mathbb{R}$. (Alternative to writing $X \subseteq \Omega \times \mathbb{R}$.)
This gives us the notion of a random variable (RV) measuring, for example, temperatures, profits, goals scored or minutes late.
We shall first consider the case of discrete random variables.

Definition (Discrete random variable)

A function $X : \Omega \to \mathbb{R}$ is a discrete random variable on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ if

1. the image set, $\mathrm{Im}(X) = \{x \in \mathbb{R} \,|\, \exists \omega \in \Omega \,.\, X(\omega) = x\}$, is a countable subset of $\mathbb{R}$
2. $\{\omega \in \Omega \,|\, X(\omega) = x\} \in \mathscr{F} \qquad \forall\, x \in \mathbb{R}$

The first condition ensures discreteness of the values obtained. The second condition says that the set of outcomes $\omega \in \Omega$ mapped to a common value, x, say, by the function X must be an event E, say, that is in the event space $\mathscr{F}$ (so that we can actually associate a probability $\mathbb{P}(E)$ to it).

Probability mass functions

Suppose that X is a discrete RV. We shall write

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}) \qquad \forall x \in \mathbb{R}\,.$$

So that

$$\sum_{x \in \mathrm{Im}(X)} \mathbb{P}(X = x) = \mathbb{P}(\cup_{x \in \mathrm{Im}(X)} \{\omega \in \Omega \,|\, X(\omega) = x\}) = \mathbb{P}(\Omega) = 1$$

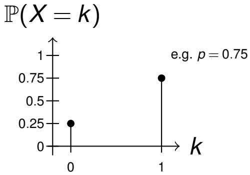
and $\mathbb{P}(X = x) = 0$ if $x \notin \mathrm{Im}(X)$. It is usual to abbreviate all this by writing

$$\sum_{x \in \mathbb{R}} \mathbb{P}(X = x) = 1\,.$$

The RV X is then said to have probability mass function $\mathbb{P}(X = x)$ thought of as a function $x \in \mathbb{R} \to [0, 1]$. The probability mass function describes the distribution of probabilities over the collection of outcomes for the RV X.

Examples of discrete distributions

Example (Bernoulli distribution)



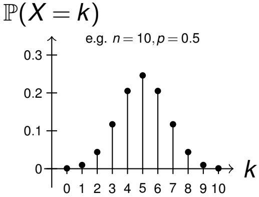
Here $\mathrm{Im}(X)=\{0,1\}$ and for given $p\in[0,1]$

$$\mathbb{P}(X=k)=\begin{cases} p & k=1 \\ 1-p & k=0\,. \end{cases}$$

RV, X	Parameters	$\mathrm{Im}(X)$	Mean	Variance
Bernoulli	$p\in[0,1]$	$\{0,1\}$	p	$p(1-p)$

Examples of discrete distributions, ctd

Example (Binomial distribution, Bin(n,p))



Here $\mathrm{Im}(X)=\{0,1,\ldots,n\}$ for some positive integer n and given $p \in [0,1]$

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

RV, X	Parameters	$\mathrm{Im}(X)$	Mean	Variance
Bin(n,p)	$n \in \{1,2,\ldots\}$ $p \in [0,1]$	$\{0,1,\ldots,n\}$	np	$np(1-p)$

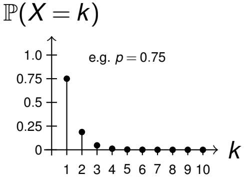
We use the notation

$$X \sim \mathrm{Bin}(n,p)$$

as a shorthand for the statement that the RV X is distributed according to stated Binomial distribution. We shall use this shorthand notation for our other named distributions.

Examples of discrete distributions, ctd

Example (Geometric distribution, Geo(p))



Here $\text{Im}(X) = \{1,2,\ldots\}$ and $0 < p \leq 1$

$$\mathbb{P}(X = k) = p(1-p)^{k-1} \qquad \forall\, k \in \{1,2,\ldots\}.$$

RV, X	Parameters	Im(X)	Mean	Variance
Geo(p)	$0 < p \leq 1$	$\{1,2,\ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

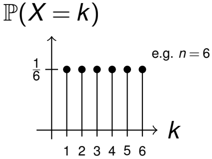
Notationally we write

$$X \sim \text{Geo}(p).$$

Beware possible confusion: some authors prefer to define our $X-1$ as a 'Geometric' RV!

Examples of discrete distributions, ctd

Example (Uniform distribution, $U(1,n)$)



Here n is some positive integer and

$$\mathbb{P}(X=k)=\frac{1}{n} \qquad \forall\, k \in \{1,2,\ldots,n\}\,.$$

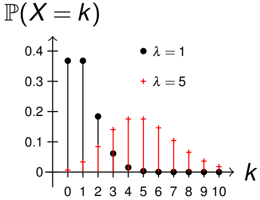
RV, X	Parameters	Im(X)	Mean	Variance
$U(1,n)$	$n \in \{1,2,\ldots\}$	$\{1,2,\ldots,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

Notationally we write

$$X \sim U(1,n)\,.$$

Examples of discrete distributions, ctd

Example (Poisson distribution, Pois(λ))



Here $\text{Im}(X) = \{0, 1, \ldots\}$ and $\lambda > 0$

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \forall k \in \{0, 1, \ldots\}.$$

RV, X	Parameters	Im(X)	Mean	Variance
Pois(λ)	$\lambda > 0$	$\{0, 1, \ldots\}$	λ	λ

Notationally we write

$$X \sim \text{Pois}(\lambda).$$

Expectation

One way to summarize the distribution of some RV, X, would be to construct a weighted average of the observed values, weighted by the probabilities of actually observing these values. This is the idea of expectation defined as follows.

Definition (Expectation)

The expectation, $\mathbb{E}(X)$, of a discrete RV X is defined as

$$\mathbb{E}(X) = \sum_{x \in \mathrm{Im}(X)} x \mathbb{P}(X = x)$$

so long as this sum is (absolutely) convergent (that is, $\sum_{x \in \mathrm{Im}(X)} |x \mathbb{P}(X = x)| < \infty$).

The expectation of a RV X is also known as the expected value, the mean, the first moment or simply the average.

Expectations and transformations

Suppose that X is a discrete RV and $g : \mathbb{R} \to \mathbb{R}$ is some transformation. We can check that $Y = g(X)$ is again a RV defined by $Y(\omega) = g(X)(\omega) = g(X(\omega))$.

Theorem

We have that

$$\mathbb{E}(g(X)) = \sum_x g(x)\mathbb{P}(X = x)$$

whenever the sum is absolutely convergent.

Proof.

$$\begin{aligned}
\mathbb{E}(g(X)) = \mathbb{E}(Y) &= \sum_{y \in g(\mathrm{Im}(X))} y\mathbb{P}(Y = y) \\
&= \sum_{y \in g(\mathrm{Im}(X))} y \sum_{x \in \mathrm{Im}(X) : g(x) = y} \mathbb{P}(X = x) \\
&= \sum_{x \in \mathrm{Im}(X)} g(x)\mathbb{P}(X = x)
\end{aligned}$$

□

Expectation is linear

Suppose that X is a discrete RV and consider the special case where $g : \mathbb{R} \to \mathbb{R}$ is given by the transformation: $g(x) = ax + b$ with a and b any real numbers.
We have that

$$\begin{aligned}
\mathbb{E}(aX + b)) &= \sum_x (ax + b)\mathbb{P}(X = x) \\
&= \sum_x ax\mathbb{P}(X = x) + \sum_x b\mathbb{P}(X = x) \\
&= a\sum_x x\mathbb{P}(X = x) + b\sum_x \mathbb{P}(X = x) \\
&= a\mathbb{E}(X) + b\,.
\end{aligned}$$

Variance

For a discrete RV X with expected value $\mathbb{E}(X)$ we define the variance, written $\mathrm{Var}(X)$, as follows.

Definition (Variance)

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

Thus, writing $\mu = \mathbb{E}(X)$ and taking $g(x) = (x - \mu)^2$

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(g(X)) = \sum_x (x - \mu)^2 \mathbb{P}(X = x)\,.$$

Just as the expected value summarizes the location of outcomes taken by the RV X, the variance measures the dispersion of X about its expected value.
The standard deviation of a RV X is defined as $+\sqrt{\mathrm{Var}(X)}$.
Note that $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ are real numbers not RVs.

First and second moments of random variables

Just as the expectation or mean, $\mathbb{E}(X)$, is called the first moment of the RV X, $\mathbb{E}(X^2)$ is called the second moment of X.

The variance $\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$ is called the second central moment of X since it measures the dispersion in the values of X centred about their mean value.

Note that we have the following property where $a, b \in \mathbb{R}$.

$$\begin{aligned}
\text{Var}(aX+b) &= \mathbb{E}\left((aX+b-\mathbb{E}(aX+b))^2\right) \\
&= \mathbb{E}\left((aX+b-a\mathbb{E}(X)-b)^2\right) \\
&= \mathbb{E}\left(a^2(X-\mathbb{E}(X))^2\right) \\
&= a^2\text{Var}(X)\,.
\end{aligned}$$

Calculating variances

Note that we can expand our expression for the variance where again we use $\mu = \mathbb{E}(X)$ as follows

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_x (x-\mu)^2 \mathbb{P}(X=x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2) \mathbb{P}(X=x) \\
&= \sum_x x^2 \mathbb{P}(X=x) - 2\mu \sum_x x \mathbb{P}(X=x) + \mu^2 \sum_x \mathbb{P}(X=x) \\
&= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\
&= \mathbb{E}(X^2) - \mu^2 \\
&= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \,.
\end{aligned}$$

This useful result determines the second central moment of a RV X in terms of the first and second moments of X. This usually is the best method to calculate the variance.

An example of calculating means and variances

Example (Bernoulli)

The expected value is given by

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= 0 \times \mathbb{P}(X = 0) + 1 \times \mathbb{P}(X = 1) \\ &= 0 \times (1 - p) + 1 \times p = p\,.\end{aligned}$$

In order to calculate the variance first calculate the second moment, $\mathbb{E}(X^2)$

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_x x^2\mathbb{P}(X = x) \\ &= 0^2 \times \mathbb{P}(X = 0) + 1^2 \times \mathbb{P}(X = 1) = p\,.\end{aligned}$$

Then the variance is given by

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p - p^2 = p(1 - p)\,.$$

Bivariate random variables

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, we may have two RVs, X and Y, say. We can then use a joint probability mass function

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\} \cap \{\omega \in \Omega \,|\, Y(\omega) = y\})$$

for all $x, y \in \mathbb{R}$.

We can recover the individual probability mass functions for X and Y as follows

$$\begin{aligned}
\mathbb{P}(X = x) &= \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}) \\
&= \mathbb{P}\left(\cup_{y \in \mathrm{Im}(Y)} (\{\omega \in \Omega \,|\, X(\omega) = x\} \cap \{\omega \in \Omega \,|\, Y(\omega) = y\})\right) \\
&= \sum_{y \in \mathrm{Im}(Y)} \mathbb{P}(X = x, Y = y) \,.
\end{aligned}$$

Similarly,

$$\mathbb{P}(Y = y) = \sum_{x \in \mathrm{Im}(X)} \mathbb{P}(X = x, Y = y) \,.$$

Transformations of random variables

If $g : \mathbb{R}^2 \to \mathbb{R}$ then we get a similar result to that obtained in the univariate case

$$\mathbb{E}(g(X,Y)) = \sum_{x \in \mathrm{Im}(X)} \sum_{y \in \mathrm{Im}(Y)} g(x,y)\mathbb{P}(X = x, Y = y)\,.$$

This idea can be extended to probability mass functions in the multivariate case with three or more RVs.
The linear transformation occurs frequently and is given by $g(x,y) = ax + by + c$ where $a,b,c \in \mathbb{R}$. In this case we find that

$$\begin{aligned}\mathbb{E}(aX + bY + c) &= \sum_x \sum_y (ax + by + c)\mathbb{P}(X = x, Y = y)\\ &= a\sum_x x\mathbb{P}(X = x) + b\sum_y y\mathbb{P}(Y = y) + c\\ &= a\mathbb{E}(X) + b\mathbb{E}(Y) + c\,.\end{aligned}$$

Independence of random variables

We have defined independence for events and can use the same idea for pairs of RVs X and Y.

Definition

Two RVs X and Y are independent if $\{\omega \in \Omega \,|\, X(\omega) = x\}$ and $\{\omega \in \Omega \,|\, Y(\omega) = y\}$ are independent for all $x, y \in \mathbb{R}$.

Thus, if X and Y are independent

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) .$$

If X and Y are independent discrete RV with expected values $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ respectively then

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y xy\mathbb{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \sum_x x\mathbb{P}(X = x) \sum_y y\mathbb{P}(Y = y) \\
&= \mathbb{E}(X)\mathbb{E}(Y) .
\end{aligned}$$

Variance of sums of RVs and Covariance

Given a pair of RVs X and Y consider the variance of their sum $X + Y$

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}\left(((X+Y)-\mathbb{E}(X+Y))^2\right)\\ &= \mathbb{E}\left(((X-\mathbb{E}(X))+(Y-\mathbb{E}(Y)))^2\right)\\ &= \mathbb{E}((X-\mathbb{E}(X))^2)+2\mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))+\\ &\qquad\qquad \mathbb{E}((Y-\mathbb{E}(Y))^2)\\ &= \mathrm{Var}(X)+2\mathrm{Cov}(X,Y)+\mathrm{Var}(Y)\end{aligned}$$

where the covariance of X and Y is given by

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))\\ &= \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)\,.\end{aligned}$$

So, if X and Y are independent RV then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ and so $\mathrm{Cov}(X, Y) = 0$ and we have that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)\,.$$

Notice also that if $Y = X$ then $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Covariance and correlation

The covariance of two RVs can be used as a measure of dependence but it is not invariant to a change of units. For this reason we define the correlation coefficient of two RVs as follows.

Definition (Correlation coefficient)

The correlation coefficient, $\rho(X,Y)$, of two RVs X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

whenever the variances exist and the product $\mathrm{Var}(X)\mathrm{Var}(Y) \neq 0$.

It may further be shown that we always have

$$-1 \leq \rho(X,Y) \leq 1\,.$$

We have seen that when X and Y are independent then $\mathrm{Cov}(X,Y) = 0$ and so $\rho(X,Y) = 0$. When $\rho(X,Y) = 0$ the two RVs X and Y are said to be uncorrelated. In fact, if $\rho(X,Y) = 1(\text{or} -1)$ then Y is a linearly increasing (or decreasing) function of X.

Random samples

An important situation is where we have a collection of n RVs, $X_1, X_2, \ldots, X_n$ which are independent and identically distributed (IID). Such a collection of RVs represents a random sample of size n taken from some common probability distribution. For example, the sample could be of repeated measurements of given random quantity. Consider the RV given by

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is known as the sample mean.
We have that

$$\begin{aligned} \mathbb{E}(\overline{X}_n) &= \mathbb{E}(\frac{1}{n} \sum_{i=1}^{n} X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu \end{aligned}$$

where $\mu = \mathbb{E}(X_i)$ is the common mean value of X_i.

Distribution functions

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we have so far considered discrete RVs that can take a countable number of values. More generally, we define $X : \Omega \to \mathbb{R}$ as a random variable if

$$\{\omega \in \Omega \,|\, X(\omega) \leq x\} \in \mathscr{F} \qquad \forall x \in \mathbb{R}.$$

Note that a discrete random variable, X, is a random variable since

$$\{\omega \in \Omega \,|\, X(\omega) \leq x\} = \cup_{x' \in \mathrm{Im}(X) : x' \leq x} \{\omega \in \Omega \,|\, X(\omega) = x'\} \in \mathscr{F}.$$

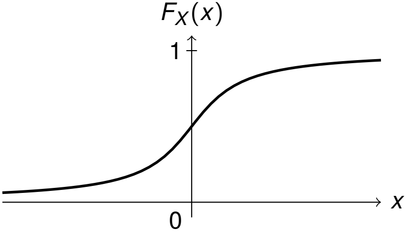
Definition (Distribution function)

If X is a RV then the distribution function of X, written $F_X(x)$, is defined by

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \leq x\}) = \mathbb{P}(X \leq x).$$

Properties of the distribution function

$F_X(x) = \mathbb{P}(X \le x)$



1. If $x \le y$ then $F_X(x) \le F_X(y)$.
2. If $x \to -\infty$ then $F_X(x) \to 0$.
3. If $x \to \infty$ then $F_X(x) \to 1$.
4. If $a < b$ then $\mathbb{P}(a < X \le b) = F_X(b) - F_X(a)$.

Continuous random variables

Random variables that take just a countable number of values are called discrete. More generally, we have that a RV can be defined by its distribution function, $F_X(x)$. A RV is said to be a continuous random variable when the distribution function has sufficient *smoothness* that

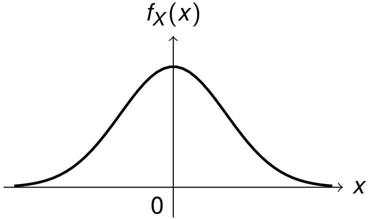
$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

for some function $f_X(x)$. We can then take

$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx} & \text{if the derivative exists at } x \\ 0 & \text{otherwise}\,. \end{cases}$$

The function $f_X(x)$ is called the probability density function of the continuous RV X or often just the density of X.
The density function for continuous RVs plays the analogous rôle to the probability mass function for discrete RVs.

Properties of the density function

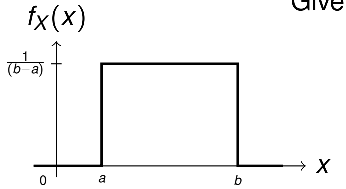


1. $\forall x \in \mathbb{R} \,.\, f_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
3. If $a \leq b$ then $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$.

Examples of continuous random variables

We define some common continuous RVs, X, by their density functions, $f_X(x)$.

Example (Uniform distribution, $U(a,b)$)



Given $a \in \mathbb{R}$ and $b \in \mathbb{R}$ with $a < b$ then

$$f_X(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } a < x < b \\ 0 & \text{otherwise}. \end{cases}$$

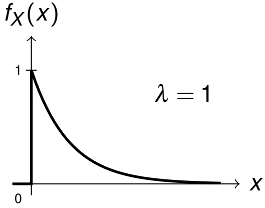
RV, X	Parameters	Im(X)	Mean	Variance
$U(a,b)$	$a,b \in \mathbb{R}$ $a < b$	(a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Notationally we write

$$X \sim U(a,b).$$

Examples of continuous random variables, ctd

Example (Exponential distribution, Exp(λ))



Given $\lambda > 0$ then

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise}. \end{cases}$$

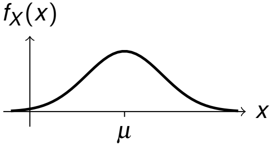
RV, X	Parameters	Im(X)	Mean	Variance
Exp(λ)	$\lambda > 0$	$\mathbb{R}_+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Notationally we write

$$X \sim \text{Exp}(\lambda).$$

Examples of continuous random variables, ctd

Example (Normal distribution, $N(\mu, \sigma^2)$)



Given $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

RV, X	Parameters	Im(X)	Mean	Variance
$N(\mu, \sigma^2)$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$\mathbb{R}$	μ	σ^2

Notationally we write

$$X \sim N(\mu, \sigma^2).$$

Expectations of continuous random variables

Just as for discrete RVs we can define the expectation of a continuous RV with density function $f_X(x)$ by a weighted averaging.

Definition (Expectation)

The expectation of X is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

whenever the integral exists.

In a similar way to the discrete case we have that if $g : \mathbb{R} \to \mathbb{R}$ then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

whenever the integral exists.

Variances of continuous random variables

Similarly, we can define the variance of a continuous RV X.

Definition (Variance)

The variance, $\mathrm{Var}(X)$, of a continuous RV X with density function $f_X(x)$ is defined as

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx$$

whenever the integral exists and where $\mu = \mathbb{E}(X)$.

Exercise: check that we again find the useful result connecting the second central moment to the first and second moments.

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \,.$$

Example: exponential distribution, Exp(λ)

Suppose that the RV X has an exponential distribution with parameter $\lambda > 0$ then using integration by parts

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\ &= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\ &= 0 + \frac{1}{\lambda}\left(\int_0^\infty \lambda e^{-\lambda x}\,dx\right) = \frac{1}{\lambda}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \\ &= \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,dx \\ &= 0 + \frac{2}{\lambda}\left(\int_0^\infty x\lambda e^{-\lambda x}\,dx\right) = \frac{2}{\lambda^2}\,.\end{aligned}$$

Hence, $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$.

Bivariate continuous random variables

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, we may have multiple continuous RVs, X and Y, say.

Definition (joint probability distribution function)

The joint probability distribution function is given by

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \leq x\} \cap \{\omega \in \Omega \,|\, Y(\omega) \leq y\}) \\ &= \mathbb{P}(X \leq x, Y \leq y) \end{aligned}$$

for all $x, y \in \mathbb{R}$.

Independence follows in a similar way to the discrete case and we say that two continuous RVs X and Y are independent if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all $x, y \in \mathbb{R}$.

Bivariate density functions

The bivariate density of two continuous RVs X and Y satisfies

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v) du dv$$

and is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) & \text{if the derivative exists at } (x,y) \\ 0 & \text{otherwise} . \end{cases}$$

We have that

$$f_{X,Y}(x,y) \geq 0 \qquad \forall x, y \in \mathbb{R}$$

and that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1 .$$

Marginal densities and independence

If X and Y have a joint density function $f_{X,Y}(x,y)$ then we have marginal densities

$$f_X(x) = \int_{v=-\infty}^{\infty} f_{X,Y}(x,v)dv$$

and

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u,y)du\,.$$

In the case that X and Y are also independent then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $x, y \in \mathbb{R}$.

Conditional density functions

The marginal density $f_Y(y)$ tells us about the variation of the RV Y when we have no information about the RV X. Consider the opposite extreme when we have full information about X, namely, that $X = x$, say. We can not evaluate an expression like

$$\mathbb{P}(Y \le y \,|\, X = x)$$

directly since for a continuous RV $\mathbb{P}(X = x) = 0$ and our definition of conditional probability does not apply.

Instead, we first evaluate $\mathbb{P}(Y \le y \,|\, x \le X \le x + \delta x)$ for any $\delta x > 0$. We find that

$$\begin{aligned}\mathbb{P}(Y \le y \,|\, x \le X \le x + \delta x) &= \frac{\mathbb{P}(Y \le y, x \le X \le x + \delta x)}{\mathbb{P}(x \le X \le x + \delta x)} \\ &= \frac{\int_{u=x}^{x+\delta x} \int_{v=-\infty}^{y} f_{X,Y}(u,v) du dv}{\int_{u=x}^{x+\delta x} f_X(u) du}\,.\end{aligned}$$

Conditional density functions, ctd

Now divide the numerator and denominator by δx and take the limit as $\delta x \to 0$ to give

$$\begin{aligned}\mathbb{P}(Y \le y \,|\, x \le X \le x + \delta x) &\to \int_{v=-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} dv \\ &= G(y), \text{say}\end{aligned}$$

where $G(y)$ is a distribution function with corresponding density

$$g(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \,.$$

Accordingly, we define the notion of a conditional density function as follows.

Definition

The conditional density function of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

defined for all $y \in \mathbb{R}$ and $x \in \mathbb{R}$ such that $f_X(x) > 0$.

Probability generating functions

Probability generating functions

A very common situation is when a RV, X, can take only non-negative integer values, that is $\mathrm{Im}(X) \subset \{0, 1, 2, \ldots\}$. The probability mass function, $\mathbb{P}(X = k)$, is given by a sequence of values $p_0, p_1, p_2, \ldots$ where

$$p_k = \mathbb{P}(X = k) \qquad \forall\, k \in \{0, 1, 2, \ldots\}$$

and we have that

$$p_k \geq 0 \qquad \forall\, k \in \{0, 1, 2, \ldots\} \qquad \text{and} \qquad \sum_{k=0}^{\infty} p_k = 1\,.$$

The terms of this sequence can be wrapped together to define a certain function called the probability generating function (PGF).

Definition (Probability generating function)

The probability generating function, $G_X(z)$, of a (non-negative integer-valued) RV X is defined as

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k$$

for all values of z such that the sum converges appropriately.

Elementary properties of the PGF

1. $G_X(z) = \sum_{k=0}^{\infty} p_k z^k$ so

$$G_X(0) = p_0 \qquad \text{and} \qquad G_X(1) = 1\,.$$

2. If $g(t) = z^t$ then

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} g(k)\mathbb{P}(X = k) = \mathbb{E}(g(X)) = \mathbb{E}(z^X)\,.$$

3. The PGF is defined for all $|z| \leq 1$ since

$$\sum_{k=0}^{\infty} |p_k z^k| \leq \sum_{k=0}^{\infty} p_k = 1\,.$$

4. Importantly, the PGF characterizes the distribution of a RV in the sense that

$$G_X(z) = G_Y(z) \qquad \forall\, z$$

if and only if

$$\mathbb{P}(X = k) = \mathbb{P}(Y = k) \qquad \forall\, k \in \{0, 1, 2, \ldots\}\,.$$

Examples of PGFs

Example (Bernoulli distribution)

$$G_X(z) = q + pz \qquad \text{where } q = 1 - p.$$

Example (Binomial distribution, Bin(n, p))

$$G_X(z) = \sum_{k=0}^{n} \binom{n}{k} p^k (q)^{n-k} z^k = (q + pz)^n \qquad \text{where } q = 1 - p.$$

Example (Geometric distribution, Geo(p))

$$G_X(z) = \sum_{k=1}^{\infty} pq^{k-1} z^k = pz \sum_{k=0}^{\infty} (qz)^k = \frac{pz}{1 - qz} \text{ if } |z| < q^{-1} \text{ and } q = 1 - p.$$

Examples of PGFs, ctd

Example (Uniform distribution, $U(1,n)$)

$$G_X(z) = \sum_{k=1}^{n} z^k \frac{1}{n} = \frac{z}{n} \sum_{k=0}^{n-1} z^k = \frac{z}{n} \frac{(1-z^n)}{(1-z)} .$$

Example (Poisson distribution, Pois(λ))

$$G_X(z) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} z^k = e^{\lambda z} e^{-\lambda} = e^{\lambda(z-1)} .$$

Derivatives of the PGF

We can derive a very useful property of the PGF by considering the derivative, $G'_X(z)$, with respect to z of the PGF $G_X(z)$. Assume we can interchange the order of differentiation and summation, so that

$$\begin{aligned} G'_X(z) &= \frac{d}{dz}\left(\sum_{k=0}^{\infty} z^k \mathbb{P}(X=k)\right) \\ &= \sum_{k=0}^{\infty} \frac{d}{dz}\left(z^k\right) \mathbb{P}(X=k) \\ &= \sum_{k=0}^{\infty} k z^{k-1} \mathbb{P}(X=k) \end{aligned}$$

then putting $z = 1$ we have that

$$G'_X(1) = \sum_{k=0}^{\infty} k\mathbb{P}(X=k) = \mathbb{E}(X)$$

the expectation of the RV X.

Further derivatives of the PGF

Taking the second derivative gives

$$G''_X(z) = \sum_{k=0}^{\infty} k(k-1)z^{k-2}\mathbb{P}(X=k)\,.$$

So that,

$$G''_X(1) = \sum_{k=0}^{\infty} k(k-1)\mathbb{P}(X=k) = \mathbb{E}(X(X-1))$$

Generally, we have the following result.

Theorem

If the RV X has PGF $G_X(z)$ then the r-th derivative of the PGF, written $G_X^{(r)}(z)$, evaluated at $z=1$ is such that

$$G_X^{(r)}(1) = \mathbb{E}\left(X(X-1)\cdots(X-r+1)\right)\,.$$

Using the PGF to calculate $\mathbb{E}(X)$ and $\mathrm{Var}(X)$

We have that

$$\mathbb{E}(X) = G'_X(1)$$

and

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= [\mathbb{E}(X(X-1)) + \mathbb{E}(X)] - (\mathbb{E}(X))^2 \\ &= G''_X(1) + G'_X(1) - G'_X(1)^2 .\end{aligned}$$

For example, if X is a RV with the Pois(λ) distribution then $G_X(z) = e^{\lambda(z-1)}$.
Thus, $G'_X(z) = \lambda e^{\lambda(z-1)}$ and $G''_X(z) = \lambda^2 e^{\lambda(z-1)}$.
So, $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$.
Finally,

$$\mathbb{E}(X) = \lambda \qquad \text{and} \qquad \mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda .$$

Sums of independent random variables

The following theorem shows how PGFs can be used to find the PGF of the sum of independent RVs.

Theorem

If X and Y are independent RVs with PGFs $G_X(z)$ and $G_Y(z)$ respectively then

$$G_{X+Y}(z) = G_X(z)G_Y(z)\,.$$

Proof.

Using the independence of X and Y we have that

$$\begin{aligned} G_{X+Y}(z) &= \mathbb{E}(z^{X+Y}) \\ &= \mathbb{E}(z^X z^Y) \\ &= \mathbb{E}(z^X)\mathbb{E}(z^Y) \\ &= G_X(z)G_Y(z) \end{aligned}$$

□

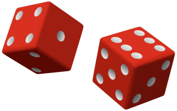
PGF example: Poisson RVs

For example, suppose that X and Y are independent RVs with $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, respectively.
Then

$$\begin{aligned} G_{X+Y}(z) &= G_X(z) G_Y(z) \\ &= e^{\lambda_1(z-1)} e^{\lambda_2(z-1)} \\ &= e^{(\lambda_1+\lambda_2)(z-1)} . \end{aligned}$$

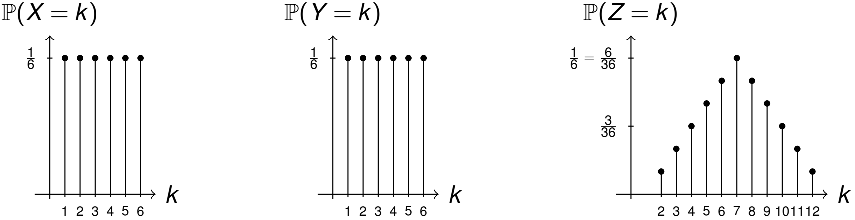
Hence $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$ is again a Poisson RV but with the parameter $\lambda_1 + \lambda_2$.

PGF example: Uniform RVs



Consider the case of two fair dice with IID outcomes X and Y, respectively, so that $X \sim U(1,6)$ and $Y \sim U(1,6)$. Let the total score be $Z = X + Y$ and consider the probability generating function of Z given by $G_Z(z) = G_X(z)G_Y(z)$. Then

$$\begin{aligned} G_Z(z) &= \frac{1}{6}(z + z^2 + \cdots + z^6)\frac{1}{6}(z + z^2 + \cdots + z^6) \\ &= \frac{1}{36}[z^2 + 2z^3 + 3z^4 + 4z^5 + 5z^6 + 6z^7 + \\ &\qquad 5z^8 + 4z^9 + 3z^{10} + 2z^{11} + z^{12}]. \end{aligned}$$



Elementary stochastic processes

Random walks

Consider a sequence $Y_1, Y_2, \ldots$ of independent and identically distributed (IID) RVs with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$ with $p \in [0, 1]$.

Definition (Simple random walk)

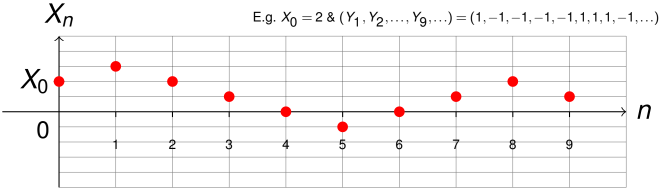
The simple random walk is a sequence of RVs $\{X_n \mid n \in \{1, 2, \ldots\}\}$ defined by

$$X_n = X_0 + Y_1 + Y_2 + \cdots + Y_n$$

where $X_0 \in \mathbb{R}$ is the starting value.

Definition (Simple symmetric random walk)

A simple symmetric random walk is a simple random walk with the choice $p = 1/2$.



Examples

Practical examples of random walks abound across the physical sciences (motion of atomic particles) and the non-physical sciences (epidemics, gambling, asset prices).

The following is a simple model for the operation of a casino. Suppose that a gambler enters with a capital of £X_0. At each stage the gambler places a stake of £1 and with probability p wins the gamble otherwise the stake is lost. If the gambler wins the stake is returned together with an additional sum of £1.
Thus at each stage the gambler's capital increases by £1 with probability p or decreases by £1 with probability $1 - p$.
The gambler's capital X_n at stage n thus follows a simple random walk except that the gambler is bankrupt if X_n reaches £0 and then can not continue to any further stages.

Returning to the starting state for a simple random walk

Let X_n be a simple random walk and

$$r_n = \mathbb{P}(X_n = X_0) \qquad \text{for } n = 1, 2, \ldots$$

the probability of returning to the starting state at time n.
We will show the following theorem.

Theorem

If n is odd then $r_n = 0$ else if $n = 2m$ is even then

$$r_{2m} = \binom{2m}{m} p^m (1-p)^m .$$

Proof.

The position of the random walk will change by an amount

$$X_n - X_0 = Y_1 + Y_2 + \cdots + Y_n$$

between times 0 and n. Hence, for this change $X_n - X_0$ to be 0 there must be an equal number of up steps as down steps. This can never happen if n is odd and so $r_n = 0$ in this case. If $n = 2m$ is even then note that the number of up steps in a total of n steps is a binomial RV with parameters $2m$ and p. Thus,

$$r_{2m} = \mathbb{P}(X_n - X_0 = 0) = \binom{2m}{m} p^m (1-p)^m \,.$$

□

This result tells us about the probability of returning to the starting state at a given time n.

We will now look at the probability that we ever return to our starting state. For convenience, and without loss of generality, we shall take our starting value as $X_0 = 0$ from now on.

Recurrence and transience of simple random walks

Note first that $\mathbb{E}(Y_i) = p - (1-p) = 2p - 1$ for each $i \in \{1, 2, \ldots\}$. Thus there is a net drift upwards if $p > 1/2$ and a net drift downwards if $p < 1/2$. Only in the case $p = 1/2$ is there no net drift upwards nor downwards.

We say that the simple random walk is recurrent if it is certain to revisit its starting state at some time in the future and transient otherwise.

We shall prove the following theorem.

Theorem

For a simple random walk with starting state $X_0 = 0$ the probability of revisiting the starting state is

$$\mathbb{P}(X_n = 0 \textit{ for some } n \in \{1, 2, \ldots\}) = 1 - |2p - 1|\,.$$

Thus a simple random walk is recurrent only when $p = 1/2$.

Proof

We have that $X_0 = 0$ and that the event $R_n = \{X_n = 0\}$ indicates that the simple random walk returns to its starting state at time n. Consider the event

$$F_n = \{X_n = 0, X_m \neq 0 \text{ for } m \in \{1, 2, \ldots, (n-1)\}\}$$

that the random walk first revisits its starting state at time n. If R_n occurs then exactly one of $F_1, F_2, \ldots, F_n$ occurs. So,

$$\mathbb{P}(R_n) = \sum_{m=1}^{n} \mathbb{P}(R_n \cap F_m)$$

but

$$\mathbb{P}(R_n \cap F_m) = \mathbb{P}(F_m)\mathbb{P}(R_{n-m}) \qquad \text{for } m \in \{1, 2, \ldots, n\}$$

since we must first return at time m and then return a time $n - m$ later which are independent events. So if we write $f_n = \mathbb{P}(F_n)$ and $r_n = \mathbb{P}(R_n)$ then

$$r_n = \sum_{m=1}^{n} f_m r_{n-m} .$$

Given the expression for r_n we now wish to solve these equations for f_m.

Proof, ctd

Define generating functions for the sequences r_n and f_n by

$$R(z) = \sum_{n=0}^{\infty} r_n z^n \qquad \text{and} \qquad F(z) = \sum_{n=0}^{\infty} f_n z^n$$

where $r_0 = 1$ and $f_0 = 0$ and take $|z| < 1$. We have that

$$\begin{aligned}
\sum_{n=1}^{\infty} r_n z^n &= \sum_{n=1}^{\infty} \sum_{m=1}^{n} f_m r_{n-m} z^n \\
&= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} f_m z^m r_{n-m} z^{n-m} \\
&= \sum_{m=1}^{\infty} f_m z^m \sum_{k=0}^{\infty} r_k z^k \\
&= F(z) R(z) .
\end{aligned}$$

The left hand side is $R(z) - r_0 z^0 = R(z) - 1$ thus we have that

$$R(z) = R(z)F(z) + 1 \qquad \text{if } |z| < 1.$$

Proof, ctd

Now,

$$\begin{aligned} R(z) &= \sum_{n=0}^{\infty} r_n z^n \\ &= \sum_{m=0}^{\infty} r_{2m} z^{2m} \quad \text{as } r_n = 0 \text{ if } n \text{ is odd} \\ &= \sum_{m=0}^{\infty} \binom{2m}{m} (p(1-p)z^2)^m \\ &= (1-4p(1-p)z^2)^{-\frac{1}{2}} \,. \end{aligned}$$

The last step follows from the binomial series expansion of $(1-4\theta)^{-\frac{1}{2}}$ and the choice $\theta = p(1-p)z^2$.
Hence,

$$F(z) = 1 - (1-4p(1-p)z^2)^{\frac{1}{2}} \quad \text{for } |z| < 1 \,.$$

Proof, ctd

But now

$$\begin{aligned}
\mathbb{P}(X_n = 0 \,\text{for some } n = 1, 2, \ldots) &= \mathbb{P}(F_1 \cup F_2 \cup \cdots) \\
&= f_1 + f_2 + \cdots \\
&= \lim_{z \uparrow 1} \sum_{n=1}^{\infty} f_n z^n \\
&= F(1) \\
&= 1 - (1 - 4p(1-p))^{\frac{1}{2}} \\
&= 1 - ((p + (1-p))^2 - 4p(1-p))^{\frac{1}{2}} \\
&= 1 - ((2p-1)^2)^{\frac{1}{2}} \\
&= 1 - |2p - 1|\,.
\end{aligned}$$

So, finally, the simple random walk is certain to revisit its starting state just when $p = 1/2$.

Mean return time

Consider the recurrent case when $p = 1/2$ and set

$$T = \min\{n \geq 1 \,|\, X_n = 0\} \qquad \text{so that} \qquad \mathbb{P}(T = n) = f_n$$

where T is the time of the first return to the starting state. Then

$$\begin{aligned} \mathbb{E}(T) &= \sum_{n=1}^{\infty} n f_n \\ &= G'_T(1) \end{aligned}$$

where $G_T(z)$ is the PGF of the RV T and for $p = 1/2$ we have that $4p(1-p) = 1$ so

$$G_T(z) = 1 - (1 - z^2)^{\frac{1}{2}}$$

so that

$$G'_T(z) = z(1 - z^2)^{-\frac{1}{2}} \to \infty \quad \text{as } z \uparrow 1\,.$$

Thus, the simple symmetric random walk ($p = 1/2$) is recurrent but the expected time to first return to the starting state is infinite.

The Gambler's ruin problem

We now consider a variant of the simple random walk. Consider two players A and B with a joint capital between them of £N. Suppose that initially A has $X_0 = £a$ $(0 \leq a \leq N)$.

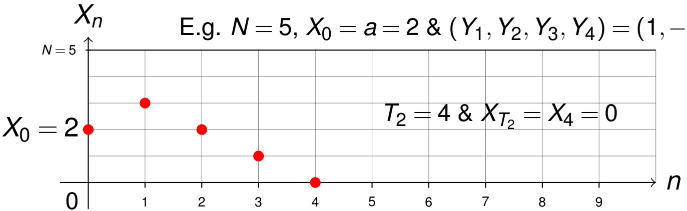
At each time step player B gives A £1 with probability p and with probability $q = (1 - p)$ player A gives £1 to B instead. The outcomes at each time step are independent.

The game ends at the first time T_a if either $X_{T_a} = £0$ or $X_{T_a} = £N$ for some $T_a \in \{0, 1, \ldots\}$.

We can think of A's wealth, X_n, at time n as a simple random walk on the states $\{0, 1, \ldots, N\}$ with absorbing barriers at 0 and N.

Define the probability of ruin for gambler A as

$$\rho_a = \mathbb{P}(\text{A is ruined}) = \mathbb{P}(\text{B wins}) \quad \text{for } 0 \leq a \leq N\,.$$



E.g. $N = 5$, $X_0 = a = 2$ & $(Y_1, Y_2, Y_3, Y_4) = (1, -1, -1, -1)$

$T_2 = 4$ & $X_{T_2} = X_4 = 0$

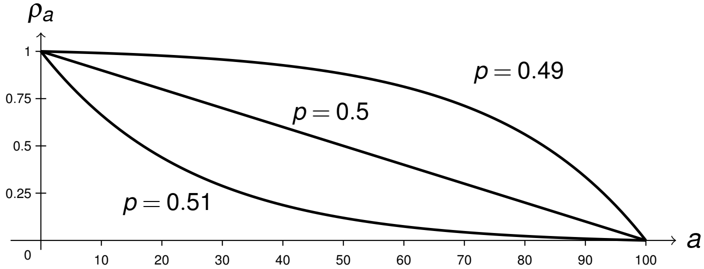
Solution of the Gambler's ruin problem

Theorem
The probability of ruin when A starts with an initial capital of a is given by

$$\rho_a = \begin{cases} \frac{\theta^a - \theta^N}{1-\theta^N} & \text{if } p \neq q \\ 1 - \frac{a}{N} & \text{if } p = q = 1/2 \end{cases}$$

where $\theta = q/p$.

For illustration here is a set of graphs of ρ_a for $N = 100$ and three possible choices of p.



Proof

Consider what happens at the first time step

$$\begin{aligned}\rho_a &= \mathbb{P}(\text{ruin} \cap Y_1 = +1 | X_0 = a) + \mathbb{P}(\text{ruin} \cap Y_1 = -1 | X_0 = a) \\ &= p\mathbb{P}(\text{ruin} | X_0 = a+1) + q\mathbb{P}(\text{ruin} | X_0 = a-1) \\ &= p\rho_{a+1} + q\rho_{a-1}\end{aligned}$$

Now look for a solution to this difference equation of the form λ^a with boundary conditions $\rho_0 = 1$ and $\rho_N = 0$.
Try a solution of the form $\rho_a = \lambda^a$ to give

$$\lambda^a = p\lambda^{a+1} + q\lambda^{a-1}$$

Hence,

$$p\lambda^2 - \lambda + q = 0$$

with solutions $\lambda = 1$ and $\lambda = q/p$.

Proof, ctd

If $p \neq q$ there are two distinct solutions and the general solution of the difference equation is of the form $A + B(q/p)^a$.
Applying the boundary conditions

$$1 = \rho_0 = A + B \qquad \text{and} \qquad 0 = \rho_N = A + B(q/p)^N$$

we get

$$A = -B(q/p)^N$$

and

$$1 = B - B(q/p)^N$$

so

$$B = \frac{1}{1-(q/p)^N} \qquad \text{and} \qquad A = \frac{-(q/p)^N}{1-(q/p)^N}\,.$$

Hence,

$$\rho_a = \frac{(q/p)^a - (q/p)^N}{1-(q/p)^N}\,.$$

Proof, ctd

If $p = q = 1/2$ then the general solution is $C + Da$.
So with the boundary conditions

$$1 = \rho_0 = C + D(0) \qquad \text{and} \qquad 0 = \rho_N = C + D(N)\,.$$

Therefore,

$$C = 1 \qquad \text{and} \qquad 0 = 1 + D(N)$$

so

$$D = -1/N$$

and

$$\rho_a = 1 - a/N\,.$$

Mean duration time

Set T_a as the time to be absorbed at either 0 or N starting from the initial state a and write $\mu_a = \mathbb{E}(T_a)$.
Then, conditioning on the first step as before

$$\mu_a = 1 + p\mu_{a+1} + q\mu_{a-1} \quad \text{for } 1 \le a \le N-1$$

and $\mu_0 = \mu_N = 0$.
It can be shown that μ_a is given by

$$\mu_a = \begin{cases} \frac{1}{p-q}\left(N\frac{(q/p)^a-1}{(q/p)^N-1} - a\right) & \text{if } p \neq q \\ a(N-a) & \text{if } p = q = 1/2\,. \end{cases}$$

We skip the proof here but note the following cases can be used to establish the result.
Case $p \neq q$: trying a particular solution of the form $\mu_a = ca$ shows that $c = 1/(q-p)$ and the general solution is then of the form $\mu_a = A + B(q/p)^a + a/(q-p)$. Fixing the boundary conditions gives the result.
Case $p = q = 1/2$: now the particular solution is $-a^2$ so the general solution is of the form $\mu_a = A + Ba - a^2$ and fixing the boundary conditions gives the result.

Case studies

Two short cases studies where probability has played a pivitol role:

1. Birthday problem (“birthday attack”)
   - cryptographic attacks
2. Probabilistic classification (“naive Bayes classifier”)
   - email spam filtering

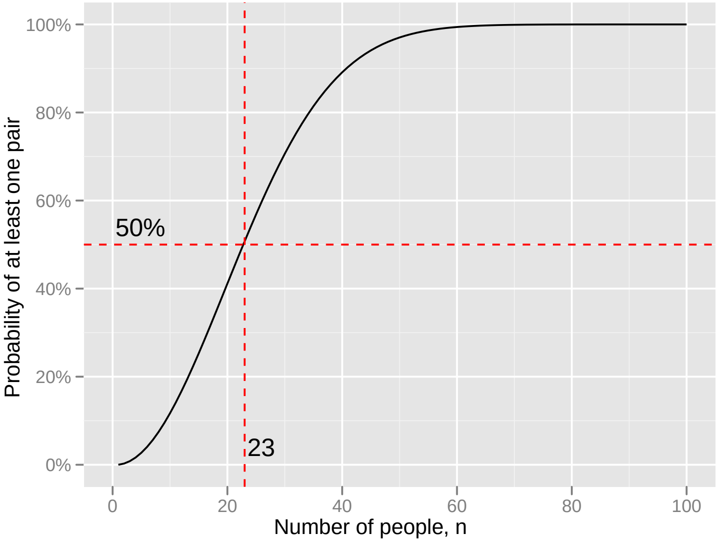
The birthday problem

Consider the problem of computing the probability, $p(n)$, that in a party of n people at least two people share a birthday (that is, the same day and month but not necessarily same year).
It is easiest to first work out $1 - p(n) = q(n)$, say,
where $q(n) = \mathbb{P}(\text{none of the } n \text{ people share a birthday})$ then

$$\begin{aligned} q(n) &= \left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365-n+1}{365}\right) \\ &= \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\cdots\left(1-\frac{n-1}{365}\right) \\ &= \prod_{k=1}^{n-1}\left(1-\frac{k}{365}\right) . \end{aligned}$$

Surprisingly, $n = 23$ people suffice to make $p(n)$ greater than 50%.

Graph of $p(n)$



100%
80%
60%
40%
20%
0%
50%
23
0
20
40
60
80
100
Probability of at least one pair
Number of people, n

Assumptions

We should record some of our assumptions behind the calculation of $p(n)$.

1. Ignore leap days (29 Feb)
2. Each birthday is equally likely
3. People are selected independently and without regard to their birthday to attend the party (ignore twins, etc)

Examples: coincidences on the football field

Ian Stewart writing in Scientific American illustrates the birthday problem with an interesting example. In a football match there are 23 people (two teams of 11 plus the referee) and on 19 April 1997 out of 10 UK Premier Division games there were 6 games with birthday coincidences and 4 games without.

Examples: cryptographic hash functions

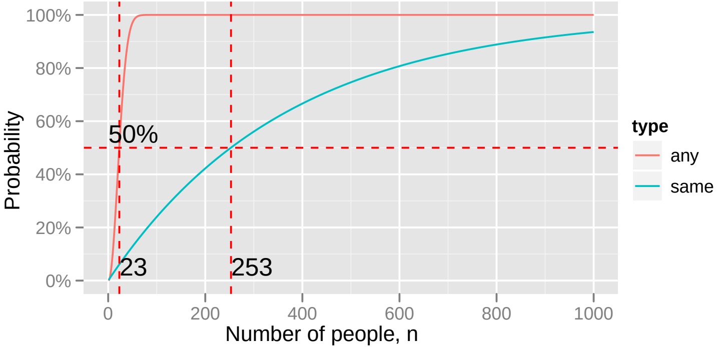
A hash function $y = f(x)$ used in cryptographic applications is usually required to have the following two properties (amongst others):

1. one-way function: computationally intractible to find an x given y.
2. collision-resistant: computationally intractible to find distinct x_1 and x_2 such that $f(x_1) = f(x_2)$.

Probability of same birthday as you

Note that in calculating $p(n)$ we are not specifying which birthday (for example, your own) matches. For the case of finding a match to your own birthday amongst a party of n other people we would calculate

$$1 - \left(\frac{364}{365}\right)^n .$$



General birthday problem

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ of size n where X_i are IID with $X_i \sim U(1, d)$ and let $p(n, d)$ be the probability that there are at least two outcomes that coincide.
Then

$$p(n,d) = \begin{cases} 1 - \prod_{k=1}^{n-1} \left(1 - \frac{k}{d}\right) & n \leq d \\ 1 & n > d\,. \end{cases}$$

The usual birthday problem is the special case when $d = 365$.

Approximations

One useful approximation is to note that for $x \ll 1$ then $1 - x \approx e^{-x}$. Hence for $n \leq d$

$$\begin{aligned} p(n,d) &= 1 - \prod_{k=1}^{n-1}\left(1 - \frac{k}{d}\right) \\ &\approx 1 - \prod_{k=1}^{n-1} e^{-\frac{k}{d}} \\ &= 1 - e^{-(\Sigma_{k=1}^{n-1} k)/d} \\ &= 1 - e^{-n(n-1)/(2d)} . \end{aligned}$$

We can further approximate the last expression as

$$p(n,d) \approx 1 - e^{-n^2/(2d)} .$$

Inverse birthday problem

Using the last approximation

$$p(n,d) \approx 1 - e^{-n^2/(2d)}$$

we can invert the birthday problem to find $n = n(p,d)$, say, such that $p(n,d) \approx p$ so then

$$\begin{aligned} e^{-n(p,d)^2/(2d)} &\approx 1 - p \\ -\frac{n(p,d)^2}{2d} &\approx \log(1-p) \\ n(p,d)^2 &\approx 2d \log\left(\frac{1}{1-p}\right) \\ n(p,d) &\approx \sqrt{2d \log\left(\frac{1}{1-p}\right)}. \end{aligned}$$

In the special case of $d = 365$ and $p = 1/2$ this gives the approximation $n(0.5, 365) \approx \sqrt{2 \times 365 \times \log(2)} \approx 22.49$.

Expected waiting times for a collision/match

Let W_d be the random variable specifiying the number of iterations when you choose one of d values independently and uniformly at random (with replacement) and stop when any value is selected a second time (that is, a "collision" or "match" occurs).
It is possible to show that

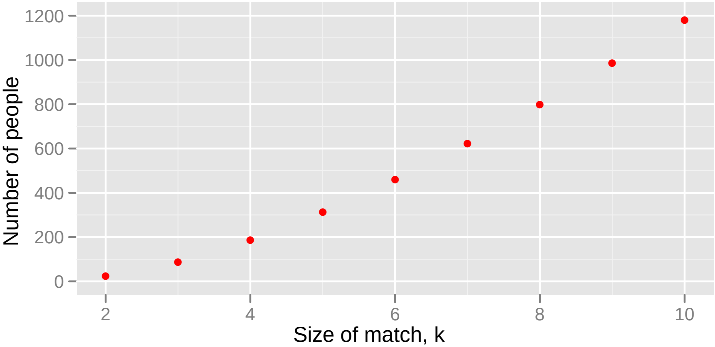
$$\mathbb{E}(W_d) \approx \sqrt{\frac{\pi d}{2}}\,.$$

Thus in the special case of the birthday problem where $d = 365$ we have that $\mathbb{E}(W_{365}) \approx \sqrt{\frac{\pi \times 365}{2}} \approx 23.94$.
In the case that we have a cryptographic hash function with 160-bit outputs ($d = 2^{160}$) then $\mathbb{E}(W_{2^{160}}) \approx 1.25 \times 2^{80}$. This level of reduction leads to so-called "birthday attacks". (See the IB course Security I for further details.)

Further results

Persi Diaconis and Frederick Mosteller give results on the minimum number n_k required to give a probability greater than $1/2$ of k or more matches with $d = 365$ possible choices.

k	2	3	4	5	6	7	8	9	10
n_k	23	88	187	313	460	623	798	985	1181



Email spam filtering

Suppose that an email falls into exactly one of two classes (spam or ham) and that various features $F_1, F_2, \ldots, F_n$ of an email message can be measured. Such features could be the presence or absence of particular words or groups of words, etc, etc.

We would like to determine $\mathbb{P}(C \mid F_1, F_2, \ldots, F_n)$ the probability that an email message falls into a class C given the measured features $F_1, F_2, \ldots, F_n$. We can use Bayes' theorem to help us.

Bayes' theorem for emails

We have that

$$\mathbb{P}(C \mid F_1, F_2, \ldots, F_n) = \frac{\mathbb{P}(C)\mathbb{P}(F_1, F_2, \ldots, F_n \mid C)}{\mathbb{P}(F_1, F_2, \ldots, F_n)}$$

which can be expressed in words as

$$\text{posterior probability} = \frac{\text{prior probability} \times \text{likelihood}}{\text{evidence}} .$$

Naive Bayes classifier

In the naive Bayes classifier we make the assumption of independence across features. So that

$$\mathbb{P}(F_1, F_2, \ldots, F_n \mid C) = \prod_{i=1}^{n} \mathbb{P}(F_i \mid C)$$

and then

$$\mathbb{P}(C \mid F_1, F_2, \ldots, F_n) \propto \mathbb{P}(C) \prod_{i=1}^{n} \mathbb{P}(F_i \mid C) \,.$$

Decision rule for naive Bayes classifier

We then use the decision rule to classify an email with observed features $F_1, F_2, \ldots, F_n$ as spam if

$$\mathbb{P}(C = \text{spam}) \prod_{i=1}^{n} \mathbb{P}(F_i \mid C = \text{spam}) > \mathbb{P}(C = \text{ham}) \prod_{i=1}^{n} \mathbb{P}(F_i \mid C = \text{ham}) \,.$$

This decision rule is known as the maximum a posteriori (MAP) rule.

Surveys and a training set of manually classified emails are needed to estimate the values of $\mathbb{P}(C)$ and $\mathbb{P}(F_i \mid C)$.

References

Ian Stewart
What a coincidence!
Mathematical Recreations, Scientific American, Jun 1998, 95–96.

Persi Diaconis and Frederick Mosteller
Methods for studying coincidences.
Journal of American Statistical Association, Vol 84, No 408, Dec 1989, 853–861.

Properties of discrete RVs

RV, X	Parameters	$\mathrm{Im}(X)$	$\mathbb{P}(X=k)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$	$G_X(z)$
Bernoulli	$p \in [0,1]$	$\{0,1\}$	$(1-p)$ if $k=0$ or p if $k=1$	p	$p(1-p)$	$(1-p+pz)$
$\mathrm{Bin}(n,p)$	$n \in \{1,2,\ldots\}$ $p \in [0,1]$	$\{0,1,\ldots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(1-p+pz)^n$
$\mathrm{Geo}(p)$	$0 < p \leq 1$	$\{1,2,\ldots\}$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-(1-p)z}$
$U(1,n)$	$n \in \{1,2,\ldots\}$	$\{1,2,\ldots,n\}$	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{z(1-z^n)}{n(1-z)}$
$\mathrm{Pois}(\lambda)$	$\lambda > 0$	$\{0,1,\ldots\}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$e^{\lambda(z-1)}$

Properties of continuous RVs

RV, X	Parameters	$\mathrm{Im}(X)$	$f_X(x)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$
$U(a,b)$	$a,b \in \mathbb{R}$ $a < b$	(a,b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\lambda > 0$	$\mathbb{R}_+$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$N(\mu,\sigma^2)$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2

Notation

Ω	sample space of possible outcomes $\omega \in \Omega$
$\mathscr{F}$	event space: set of random events $E \subset \Omega$
$\mathbb{I}(E)$	indicator function of the event $E \in \mathscr{F}$
$\mathbb{P}(E)$	probability that event E occurs, e.g. $E = \{X = k\}$
RV	random variable
$X : \Omega \to \mathbb{R}$	Alternative to writing function as a relation $X \subseteq \Omega \times \mathbb{R}$.
Im(X)	image set under RV X, i.e. $\{x \in \mathbb{R} \mid \exists \omega \in \Omega \,.\, X(\omega) = x\}$
$X \sim U(0,1)$	RV X has the distribution $U(0,1)$
$\mathbb{P}(X = k)$	probability mass function of RV X
$F_X(x)$	distribution function, $F_X(x) = \mathbb{P}(X \leq x)$
$f_X(x)$	density of RV X given, when it exists, by $F'_X(x)$
PGF	probability generating function $G_X(z)$ for RV X
$\mathbb{E}(X)$	expected value of RV X
$\mathbb{E}(X^n)$	n^{th} moment of RV X, for $n = 1, 2, \ldots$
Var(X)	variance of RV X
IID	independent, identically distributed
$\overline{X}_n$	sample mean of random sample $X_1, X_2, \ldots, X_n$