MAST20004 Probability

Lecturers: Mark Fackrell and Aihua Xia

Administration

- Web site: accessible through the Learning Management System, contains lecture slides, assignments and solutions, ...
- Lectures: attendance
- Consultation hours
- Tutorial/Computer Lab classes
- Homework
- Assessment

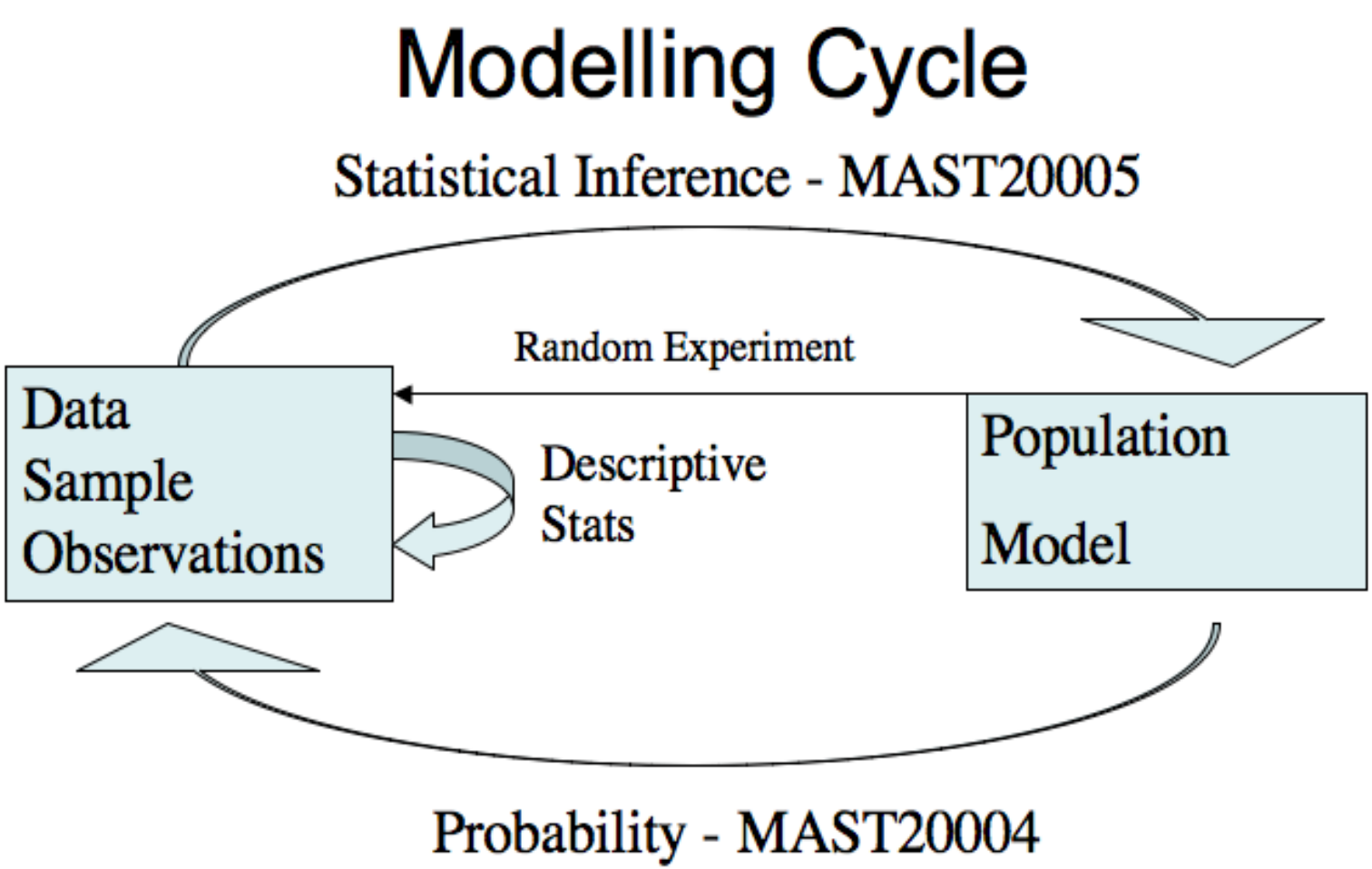
Three levels of learning

- I see, I forget
- I read, I know
- I do, I understand

Together MAST20004 and MAST20005

- Introduce mathematical statistics
  - A branch of mathematics applied to the real world
  - About 90 years
- Useful whenever measurements are subject to random variation

Modelling Cycle



Probability is in our daily lives

- The Monty Hall Problem
- The Birthday Problem
- The Bus-Stop Paradox
- Voting in US Presidential Elections

The Monty Hall Problem

- A prize lies behind one of three doors.
- The contestant chooses a door.
- Monty Hall (who knows which door the prize is behind) opens a door not chosen by the contestant that does not have the prize behind. There must be at least one such door.
- Monty Hall then offers the contestant the option of changing his/her original selection to the other unopened door.
- Should the contestant change?

The Birthday Problem

Twenty three people are on a soccer pitch. What is the probability that there are two people present with the same birthday?

The Bus-Stop Paradox

- Buses on a particular route arrive at randomly-spaced intervals throughout the day.
- On average a bus arrives every hour.
- A passenger comes to the bus-stop at a random instant.
- What is the expected length of time that the passenger will have to wait for a bus?

Probability

There are many applications of probability in society. For example, we need to use probability theory to

- design and analyse experiments in almost any field of science and social science,
- assign values to financial derivatives,
- design, dimension and control telecommunications systems, and
- understand the process of evolution of gene sequences.

Random Experiments (Ghahramani 1.2)

- *Random experiment:* a process leading to a number (which may be infinite) of possible outcomes and the actual outcome that occurs depends on influences that cannot be predicted beforehand.

- The *outcome space* or *sample space* Ω is the set of *all* possible outcomes of an experiment, survey or other observation

  NB: Note that Ghahramani uses S to denote the sample space but we will use the more common Ω.

Examples

- Toss of a coin or die.
- Spin of a roulette wheel.
- A horse race.
- Measurement of the number of phone calls passing through a telephone exchange in a fixed time period.
- A record of the proportion of people in a survey who approve of the prime minister.
- An observation of whether the greater bilby (macrotis lagotis) is extinct in 100 years time.

Examples

Toss of a coin.

$$\Omega = \{H, T\} \qquad \text{where} \qquad H = \text{“head up”}$$
$$T = \text{“tail up”}$$

Spin of a roulette wheel.

$$\Omega = \{0, 1, \ldots, 36\}$$

(There are 37 numbers on an Australian roulette wheel.)

A horse race.

Here the actual experiment needs to be defined more precisely. If we observe only the winner we might take

since the winner has to be one of the horses. If we observe the placings we could take

More generally, if we observe the whole race we might take

or, even

This example illustrates that a given physical situation can lead to different sample space depending on what we choose to observe.

Some Further Examples

Some further examples are

- A coin is tossed until a head occurs and the number of tosses required is observed
- A machine automatically fills a one litre bottle with fluid, and the actual quantity of fluid in the bottle is measured in litres
- A car is filled up with petrol and then driven until it runs out, the distance it travels is measured in kilometres

Exercise

Write down appropriate sample spaces for the experiments:

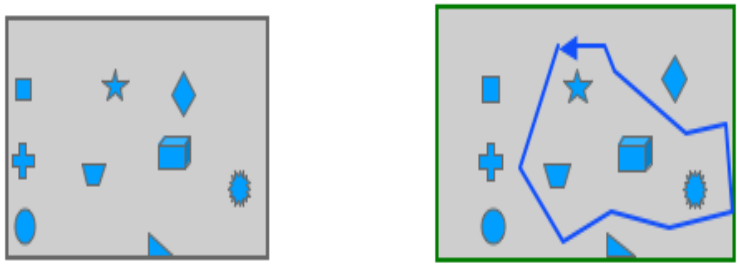
- Measurement of the number of phone calls passing through a telephone exchange in a fixed time period.

- A record of the proportion of people in a survey who approve of the prime minister.

- An observation of whether the greater bilby (*macrotis lagotis*) is extinct in 100 years time.

Events

- We are often interested in a group of outcomes.
- An *event* is a set of possible outcomes, that is a subset of Ω.



- We say that the event A occurs if the observed outcome of the random experiment is one of the outcomes ω that is in the set A.

Examples

Toss of a die, the event that "the number on the die is even" is

Spin of a roulette wheel, the event that "one of the first three numbers occurs" is

and the event that "the number 0 comes up" is

Since Ω is a set of outcomes, Ω itself is an event. This is known as the *certain event*. One of the outcomes in Ω must occur.

The empty set $\emptyset$ is also an event, known as the *impossible event*.

Events are sets and so they are subject to the normal set operations. Thus

- The event $A \cup B$ is the event that A *or* B *or* both occur.
- The event $A \cap B$ is the event that A *and* B both occur.
- The event A^c is the event that A does not occur.
- We write $\omega \in A$ to say that the outcome ω is in the event A.
- We write $A \subseteq B$ to say that A is a subset of B. This includes the possibility that $A = B$.
- If A is finite (which will often not be the case), we write $\#A$ for the number of elements of A.

For illustrative purposes, and to gain intuition, the relationship between events is often depicted using a Venn diagram.

Two events A_1, A_2 which have no outcomes in common ($A_1 \cap A_2 = \emptyset$) are called *mutually exclusive* or *disjoint* events.

Similarly, events $A_1, A_2, \ldots$ are *disjoint* if no two have outcomes in common, that is

$$A_i \cap A_j = \emptyset \quad \forall\, i \neq j.$$

Two events are *exhaustive* if they contain all possible outcomes between them,

$$A_1 \cup A_2 = \Omega.$$

Similarly, events $A_1, A_2, \ldots$ are exhaustive if their union is the whole sample space,

$$\bigcup_i A_i = \Omega.$$

Examples

1. Since $A \cap A^c = \emptyset$, A and A^c are disjoint.
2. Since $A \cup A^c = \Omega$, A and A^c are exhaustive.
3. Throw of a die. Let

   $$A = \{1, 3, 5\}, \quad B = \{2, 4, 6\}, \quad C = \{1, 2, 4, 6\}, \quad D = \{2, 4\}$$

   Then A and B are disjoint and exhaustive, A and C are exhaustive but not disjoint and A and D are disjoint but not exhaustive.

Set operations satisfy the distributive laws

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned}$$

and De Morgan's laws

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c \end{aligned}$$

Simulation

Simulation of random experiments is a tool which probabilists often use. It consists of performing the experiment on a computer, instead of in real life. This has many advantages.

- It enables us to try out multiple possibilities before going to the expense of building a system.
- It is possible to perform multiple repetitions of an experiment in a short time, so that precise estimates of the behaviour can be derived.
- It is possible to study behaviour of random experiments which are so complicated that it is hard for us to study them analytically.

In our computer lab classes, we shall be using simulation.

Defining Probability (Ghahramani 1.1)

Up to now we have talked about ways of describing results of random experiments – an event A happens if the outcome of the experiment is in the set A. We haven't yet talked about ways of assigning a measure to the "likelihood" of an event happening.

That is, we are yet to define what we mean by a probability.

First let us think about some intuitive notions.

What do we mean when we say "The probability that a toss of a coin will result in 'heads' is 1/2"?

An interpretation that is accepted by most people for practical purposes, that such statements are made based upon some information about *relative frequencies*.

People	#trials	#heads	frequency of heads
Buffon	4040	2048	0.5069
DeMorgan	4092	2048	0.5005
Feller	10000	4979	0.4979
Pearson	12000	6019	0.5016
Pearson	24000	12012	0.5005

Similar statements can be made about tossing dice, spinning roulette wheels, arrivals of phone calls in a given time period, etc.

Hence it seems that we can think of a probability as a long term relative frequency. However there are problems with this interpretation. Consider the statement

"The probability that horse X will win the Melbourne Cup this year is 1/21".

A similar statement is

"The probability that *macrotis lagotis* will be extinct in 100 years is 1/100".

Both of the above-mentioned experiments will be performed only once under unique conditions, so a repetitive relative frequency definition makes no sense.

One way to think of probability in these experiments is that it reflects the odds at which a person is willing to bet on an event.

Thus probability takes on a "personal" definition: my evaluation of a probability may not be the same as yours. This interpretation of probability is known as the *Bayesian interpretation*.

The way that mathematicians approach such issues is to define precisely the system they are studying via sets of *axioms*, derive results in that system, and then real world interpretations can be made in individual situations whenever the axioms correspond well to the real world situation.

The study of mathematical probability is like this. It is based on axioms, under which probabilities behave "sensibly".

The definition of what we mean by "sensibly" is remarkably simple.

- We arbitrarily assign the value 1 to be the probability of the certain event and require that the probability of any event be nonnegative.
- If A and B are disjoint events, then if A occurs then B can't, and vice versa. Thus we would expect that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

  where we have written $\mathbb{P}(A)$ to denote the probability of event A.

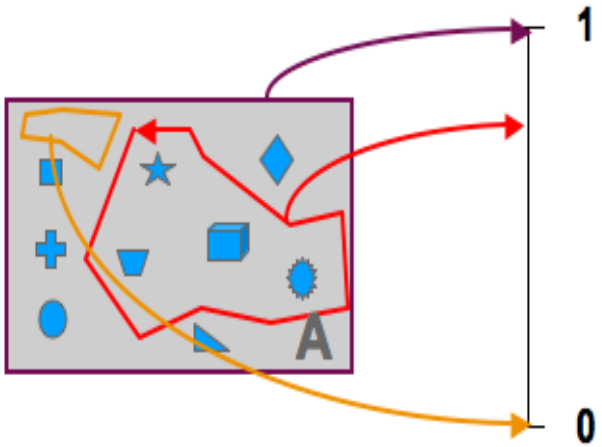
Probability axioms (Ghahramani 1.3, 1.4)

These considerations lead to the following *axioms*:

1. $\mathbb{P}(A) \geq 0$, for all events A
2. $\mathbb{P}(\Omega) = 1$

3∗. (Finite additivity) For a set of mutually exclusive events $\{A_1, A_2, A_3, \ldots, A_n\}$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$



In fact, it turns out that we need a slightly stronger version of Axiom 3 (otherwise, some weird things may happen: see slide 104). Specifically, it has to hold for infinite sequences of mutually exclusive events. Thus, we use

3. (Countable additivity)

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

where $\{A_1, A_2, A_3, \ldots\}$ is any sequence of mutually exclusive events.

Andrey Kolmogorov [25/04/1903 - 20/10/1987]

We use countable, rather than finite, additivity because we sometimes need to calculate probabilities for countable unions.

For example, the event that a 6 eventually occurs when tossing a die can be expressed as $\bigcup_{i=1}^{\infty} A_i$, where A_i is the event that the 6 occurs for the first time on the ith toss.

From the axioms, we can deduce the following properties of the probability function:

(4) $\mathbb{P}(\emptyset) = 0$, since $\emptyset \cup \emptyset \cup \cdots = \emptyset$

(5) Finite additivity

(6) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since $A \cup A^c = \Omega$

(7) $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$, since $A \cup (A^c \cap B) = B$

(8) $\mathbb{P}(A) \leq 1$, since $A \subset \Omega$

(9) Addition theorem:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Notes

- $\mathbb{P}(.)$ is a set function. It maps $\mathcal{A} \to [0, 1]$, where $\mathcal{A}$ denotes the class of events, that is the set of subsets of the outcome space.
- For a discrete outcome space, we can write

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

- In general, "possible" outcomes are allowed to have zero probability, thus $\mathbb{P}(E) = 0 \not\Rightarrow E = \emptyset$. Similarly, there can be sets other than Ω that can have probability 1.

Evaluating Probabilities

So far we have said nothing about how numerical values are assigned to the probability function, just that if we assign values in such a way that the Axioms (1) – (3) hold, then the properties (4) – (9) will also hold.

Assigning probabilities to events is a large part of what the subject is about.

- There may be no 1 "right" answer!
  - Simple problems may have a single reasonable solution
  - Real life problems often have many possible solutions
    - each OK, if they obey the rules
    - selection uses art and science

The simplest case

When the outcome space is finite $\#(\Omega) = N$ and there is no difference between outcomes so make them equally likely, then it follows easily from the axioms that

$$\mathbb{P}(\{\omega\}) = 1/N$$

for all $\omega \in \Omega$.

Further,

$$\mathbb{P}(A) = \#(A)/N.$$

Examples

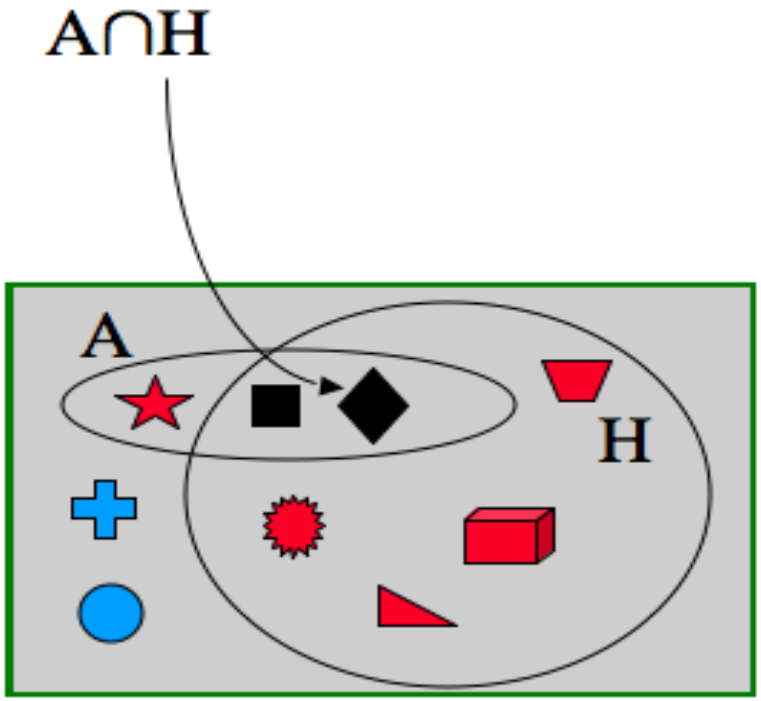
- Toss of one fair coin.
- Toss of two fair coins, find the probability of one tail and one head (D'Alembert (1717-1783): 1/3)
- Toss of two fair dice.
- Three of twenty tyres in a store are defective. Four tyres are randomly-selected for inspection. What is the probability that a defective tyre will be included?
- The birthday problem.
- Randomly draw a chord in a circle with radius 1, find the probability that the distance between the chord and the centre is less than 1/2.

Example

- Fewer dying on R.I. roads, Providence Journal, Dec 24, 1999, Jonathan Saltzman.
- The article claims: Forty-two percent of all fatalities occurred on Friday, Saturday, and Sunday, apparently because of increased drinking on the weekends.
- Of course 42% is remarkably close to $3/7 = 43\%$.

Conditional Probability (Ghahramani 3.1)

If A and H are two events, and it is known that event H has occurred, what effect does this information have on the probability of occurrence of A?



Example

Toss two fair dice. If we know that the first die is a 3, what is the probability that the sum of the two dice is 8?

The original sample space Ω has 36 outcomes $\{(1,1),(1,2),\ldots,(6,6)\}$. Given the first die is a 3 there are six outcomes of interest, $\{(3,1),(3,2),\ldots,(3,6)\}$. Since the dice are fair, each of these outcomes has the same probability of occurring. Hence, given that the first die is a 3, the probability of the sum being 8 is

If A denotes "sum of the dice is 8" and H denotes "first die is a 3" the probability we have calculated is called the *conditional probability* of A given H and is denoted $\mathbb{P}(A|H)$.

In general, imagine we conduct the experiment n times, and observe H n_H times and $A \cap H$ $n_{A\cap H}$ times. The proportion of times A occurs in the n_H experiments when H occurs is

$$\frac{n_{A\cap H}}{n_H} = \frac{n_{A\cap H}/n}{n_H/n} \to \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)}.$$

Hence the probability of A given H should be defined as the probability of $A \cap H$ relative to the probability of H:

$$\mathbb{P}(A|H) = \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)} \quad \text{if} \quad \mathbb{P}(H) > 0.$$

Multiplication Theorem (Ghahramani 3.2)

Sometimes we know $\mathbb{P}(H)$ and $\mathbb{P}(A|H)$ but not $\mathbb{P}(A \cap H)$. If this is the case we can use the definition of conditional probability to express the probability of $A \cap H$, that is

$$\mathbb{P}(A \cap H) = \mathbb{P}(H)\mathbb{P}(A|H).$$

Examples

- Toss a fair die. Let $A = \{2\}$ and $H = \{x : x \text{ is even}\}$, then

- Toss two fair dice. Let $A = \{(i, j) : |i - j| \leq 1\}$ and $H = \{(i, j) : i + j = 7\}$, then

Let us assume that $\mathbb{P}(A|H) > \mathbb{P}(A)$. That is, it is more likely that A will occur if we know that H has occurred than if we know nothing about H.

We observe that

$$\begin{aligned}
&\mathbb{P}(A|H) > \mathbb{P}(A) \\
\Leftrightarrow\ &\frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)} > \mathbb{P}(A) \\
\Leftrightarrow\ &\frac{\mathbb{P}(A \cap H)}{\mathbb{P}(A)} > \mathbb{P}(H) \\
\Leftrightarrow\ &\mathbb{P}(H|A) > \mathbb{P}(H).
\end{aligned}$$

We say that there exists a *positive relationship* between A and H if $\mathbb{P}(A|H) > \mathbb{P}(A)$ and a *negative relationship* between A and H if $\mathbb{P}(A|H) < \mathbb{P}(A)$.

If there is a positive relationship between A and H, then the occurrence of H will increase the chance of A occuring. If there is a negative relationship between A and H, then the occurrence of H will decrease the chance of A occuring.

Question: What about the situation when $\mathbb{P}(A|H) = \mathbb{P}(A)$?

Example

You are one of seven applicants for three jobs.

Consider the random experiment that occurs when the decision is made to offer the jobs to three applicants. The outcome space Ω consists of all combinations of three appointees from seven applicants. There are $\binom{7}{3} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$ of these.

Another one of the applicants (Ms X) and you are the only applicants with a particular skill. You think that it is likely that the employer will want the skill and employ one of you, but unlikely that both of you will be employed. Specifically you make the assessments listed on the following slide about the likelihood of the various possible combinations.

- the five combinations where both you and Ms X get the job have equal probability 1/60,
- the ten combinations where you get the job and Ms X doesn't have equal probability 1/24,
- the ten combinations where Ms X gets the job and you don't have equal probability 1/24, and
- the ten combinations where neither of you get the job have equal probability 1/120.

You find out on the grapevine that Ms X has not got the job. How does your assessment of the probability that you will get the job change?

Solution

Let the A be the event "you are selected" and the event H "Ms X is not selected".

More rigorously A is the subset of Ω that consists of those outcomes in which you are selected and H is the subset of Ω that consists of those outcomes in which Ms X is not selected.

Then

$$\mathbb{P}(A) =$$

$$\mathbb{P}(H) =$$

and

$$\mathbb{P}(A \cap H) =$$

Therefore

$$\mathbb{P}(A|H) =$$

and so if you know that Ms X did not get the job, your probability of getting the job is quite high.

There is a positive relationship between the events A and H.

Independence of Events (Ghahramani 3.5)

- If $\mathbb{P}(A|B) > \mathbb{P}(A)$ then $\mathbb{P}(B|A) > \mathbb{P}(B)$. A and B tend to occur together and we call them positively related.
- If $\mathbb{P}(A|B) < \mathbb{P}(A)$ then $\mathbb{P}(B|A) < \mathbb{P}(B)$. A and B tend to occur separately and we call them negatively related.
- If $\mathbb{P}(A|B) = \mathbb{P}(A)$ then $\mathbb{P}(B|A) = \mathbb{P}(B)$. A and B don't appear to influence each other and we call them *independent*.

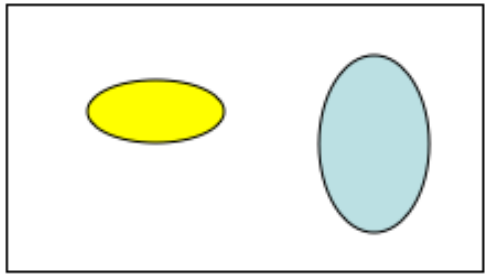
As we have just seen $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$ are algebraically equivalent to

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

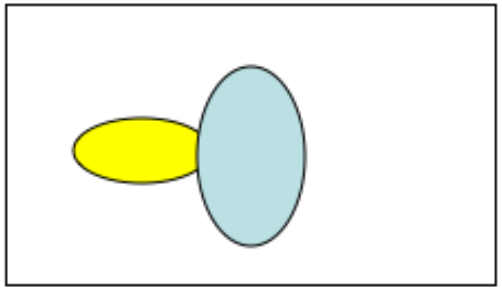
This equation is taken as the mathematical definition of the *independence* of two events. It is a special case of the general multiplication theorem

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A).$$

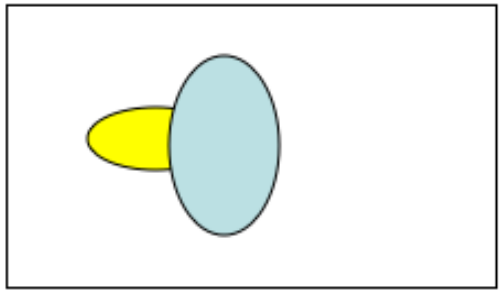
Two events that are not independent are said to be *dependent*.



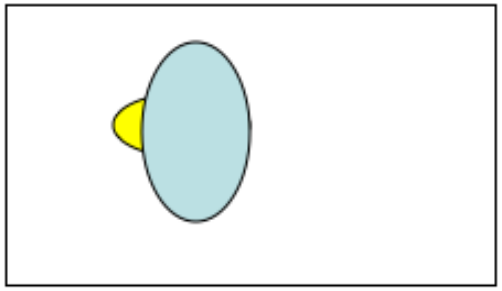
Disjoint-negative relation



P(A|H)<P(A) -negative relation



independent



P(A|H)>P(A) -positive relation

If A and B are independent events then so are

- A^c and B
- A and B^c
- A^c and B^c

Remarks about independence

- Physical indept and math indept
- Consider tossing a fair six-sided dice once and define events $A = \{2, 4, 6\}$, $B = \{1, 2, 3\}$ and $C = \{1, 2, 3, 4\}$, then A and C are independent while A and B are not independent.

Independence of $n > 2$ events (Ghahramani 3.5)

Now let's think about extending the idea of independence to more than two events. We talk of the "mutual" independence of $n > 2$ events. But can we cover all dependencies by checking for pairwise independence of all possible pairs? Consider the random experiment of tossing two fair coins and the following three events:

- A: First coin is H
- B: Second coin is H
- C: Exactly one coin is H

Events $A_1, A_2, \ldots, A_n$ are said to be *mutually independent* if for any subcollection $\{j_1, j_2, \ldots, j_m\} \subset \{1, 2, \ldots, n\}$

$$\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_m}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}) \ldots \mathbb{P}(A_{j_m})$$

- How many equations?

If the events $A_1, A_2, \ldots, A_n$ are mutually independent then the following derived collections of events are also mutually independent

- $A_1^c, A_2, \ldots, A_n$
- $A_1^c, A_2^c, A_3, \ldots, A_n$
- $A_1 \cap A_2, A_3, \ldots, A_n$
- $A_1 \cup A_2, A_3, \ldots, A_n$

An important and frequently-applied consequence of mutual independence:

If the events $A_1, A_2, \ldots, A_n$ are mutually independent then

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\ldots\mathbb{P}(A_n)$$

NB: The converse is not true. Why?

This result is particularly useful for analysing n mutually independent repetitions of a random experiment.

Independence vs Exclusion

Independence of events A and B is a different concept from A and B being *mutually exclusive* or *disjoint*. You can test for mutual exclusion simply by inspecting the outcomes in A and B to see if there are any in common, even before any probability function is defined (say by inspecting the Venn diagram).

But you cannot test independence without knowing the probabilities.

Unless one or both have probability zero, disjoint events A and B *cannot* be independent, since

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0 < \mathbb{P}(A)\mathbb{P}(B)$$

In fact, the events A and B are negatively related: the occurence of A excludes the occurence of B.

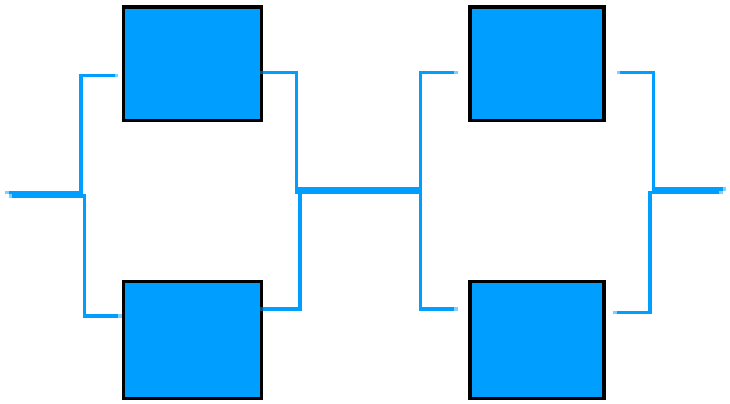
Network reliability

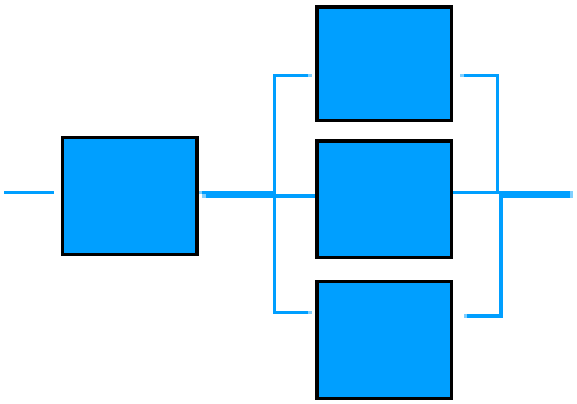
We can calculate the reliability of a network of mutually independent components connected in series and parallel.

For example imagine a network of interconnected switches which should close in an emergency to set off an alarm. Individual switches fail at random and independently. For the alarm to sound there must be at least one path for current to flow from left to right.

Example

Two systems in which components operate independently, each with a probability of failure equal to 0.01. Which system is more reliable?

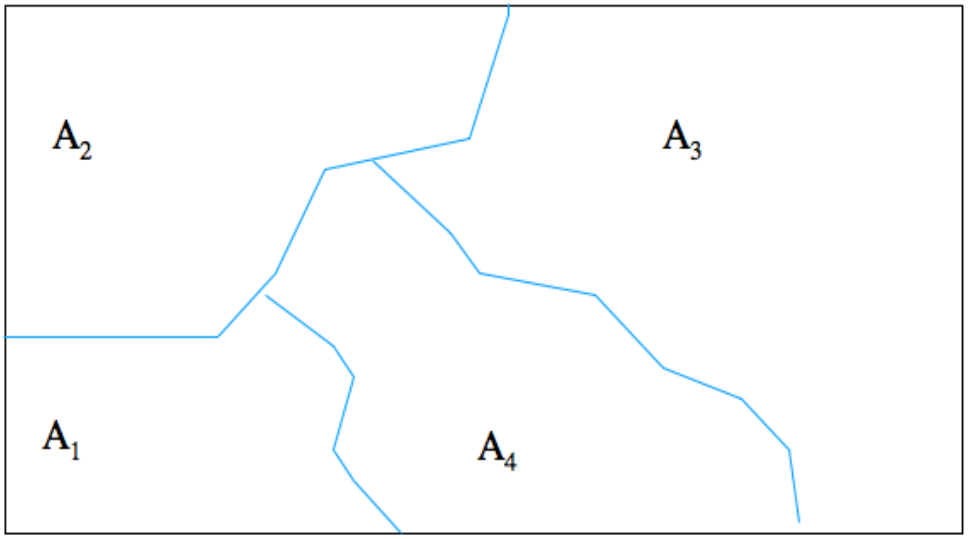




Law of Total Probability (Ghahramani 3.3)

A *partition* of the outcome space Ω is a collection of disjoint and exhaustive events $(A_1, A_2, \ldots)$. That is, for all i and j, $A_i \cap A_j = \emptyset$ and

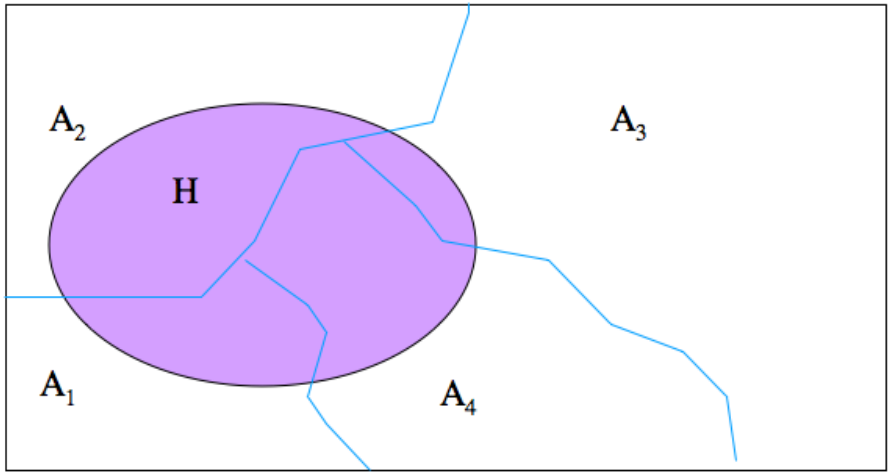
$$\bigcup_i A_i = \Omega$$



Now, for any event H,

$$\begin{aligned} H &= H \cap \Omega \\ &= H \cap \left(\bigcup_i A_i \right) \\ &= \bigcup_i (H \cap A_i) \end{aligned}$$

where the last equation follows from the distributive law.



Using probability axiom 3 and the multiplication formula, we have

$$\begin{aligned}\mathbb{P}(H) &= \sum_i \mathbb{P}\left(H \cap A_i\right) \\ &= \sum_i \mathbb{P}(H|A_i)\mathbb{P}(A_i).\end{aligned}$$

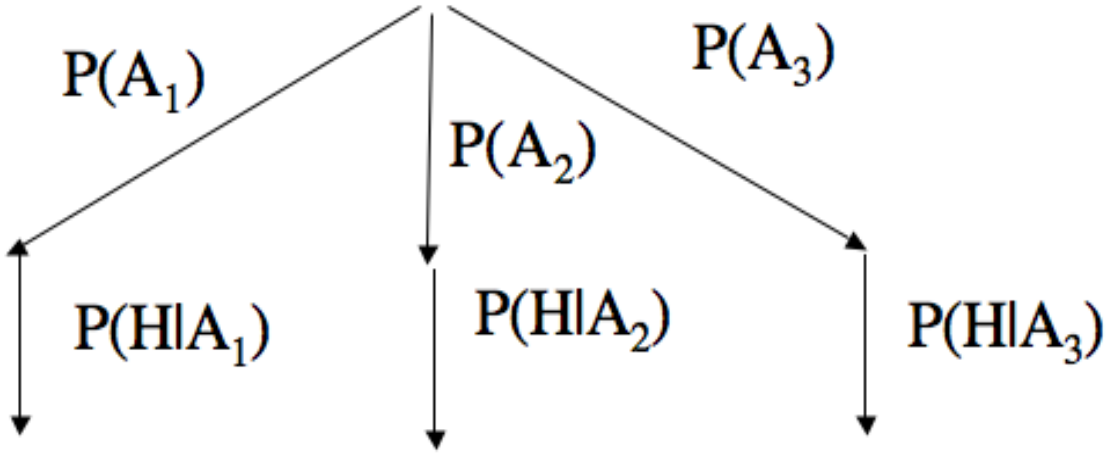
This brings us to the Law of Total Probability.

If $A_1, A_2, \ldots$ are disjoint and exhaustive events then, for any event H,

$$\mathbb{P}(H) = \sum_i \mathbb{P}(H|A_i)\mathbb{P}(A_i).$$

Law of Total Prob – tree diagrams

Multiply along the branches and add the results to get $\mathbb{P}(H)$:



The Law of Total Probability is one of the most important equations in probability theory. By choosing the right partition, it can be used (in different forms, of course) to set up equations concerning financial markets, queues, teletraffic networks, epidemics, genetics and computer systems, to name just a few applications.

Example

Suppose a test for HIV is 90% effective in the sense that if a person is HIV positive then the test has a 90% chance of saying that they are HIV positive. If they are not positive assume there is still a 5% chance that the test says that they are. Assume that in a large population, there are 0.01% of people who are HIV positive. What is the probability that a randomly selected person from the population will have a positive HIV test result? If a person receives a positive HIV test result, what is the probability that the person is actually HIV positive?

Solution:

Bayes' Formula (Ghahramani 3.4)

Let $A_1, A_2, \ldots$ be a set of disjoint and exhaustive events.

Then for an event H,

$$\mathbb{P}(A_i|H) = \frac{\mathbb{P}(H|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(H|A_j)\mathbb{P}(A_j)}.$$

Proof

From Bayes' Formula

$$\mathbb{P}(A_i|H) = \frac{\mathbb{P}(H|A_i)\mathbb{P}(A_i)}{\mathbb{P}(H)}$$

Substitution of

$$\mathbb{P}(H) = \sum_j \mathbb{P}(H|A_j)\mathbb{P}(A_j)$$

from the Law of Total Probability gives the result.

HIV example (conclusion)

- Under these numbers a person is unlikely to be HIV positive even if the test says that they are. Such phenomena are well known in the epidemiological literature. Tests for rare diseases have to be very accurate.
- NB: Reversing conditional probs can be worse than reversing implications in logic

Example (Multiple Choice Exams)

Consider a multiple choice exam that has m choices of answer for each question. Assume that the probability that a student knows the correct answer to a question is p. A student that doesn't know the correct answer marks an answer at random. Suppose that the answer marked to a particular question was correct. What is the probability that the student was guessing?

Solution

Here we have the disjoint and exhaustive events

- B_1 – the student knew the correct answer
- B_2 – the student was guessing

and the observed event A – the correct answer was marked.

The conditional probabilities are

$$\mathbb{P}(A|B_1) = 1 \quad \mathbb{P}(A|B_2) = 1/m$$

We want to find $\mathbb{P}(B_2|A)$:

Random Variables (Ghahramani 4.1)

In many random experiments we are interested in some function of the outcome rather than the actual outcome itself.

For instance, in tossing two dice (as in Monopoly) we may be interested in the sum of the two dice (e.g. 7) and not in the actual outcome (e.g. (1,6), (2,5), (3,4), (4,3), (5,2) or (6,1)).

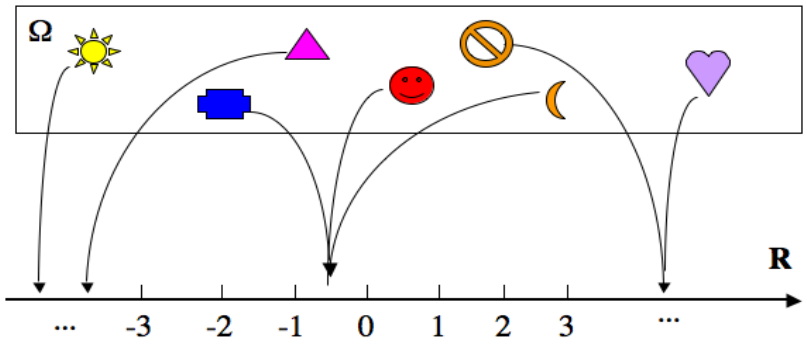
In these cases we wish to assign a real number x to each outcome ω in the sample space Ω. That is

$$x = X(\omega)$$

is the value of a *function* X from Ω to the real numbers $\mathbb{R}$.

Definition

Consider a random experiment with sample space Ω. A *function* X which assigns to every outcome $\omega \in \Omega$ a real number $X(\omega)$ is called a *random variable*.



NB. In more advanced courses, there are some restrictions on the function X, but we won't worry about them here.

- The terminology "random variable" is unfortunate because X is neither random nor a variable. However it is universally accepted.
- It is standard to denote random variables by capital letters X, Y etc. and the values they take by lower case letters x, y etc.
- We shall denote the *set of possible values* (or *state space*) of X by $S_X \subseteq \mathbb{R}$ (this differs from the notation in Ghahramani)

Example

Suppose we toss two coins. The sample space is

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Let $X(\omega),\ \omega \in \Omega$ be the number of heads in ω. Then

$$\begin{aligned} X((H,H)) &= 2 \\ X((H,T)) &= X((T,H)) = 1 \\ X((T,T)) &= 0 \end{aligned}$$

Here $S_X = \{0,1,2\}$.

- X is not necessarily a 1-1 function. Different values of ω may lead to the same value of $X(\omega)$, e.g. $X((H,T)) = X((T,H))$.
- The sets

$$\begin{aligned} A_2 &= \{\omega : X(\omega) = 2\} = \{(H,H)\} \\ A_1 &= \{\omega : X(\omega) = 1\} = \{(T,H),(H,T)\} \\ A_0 &= \{\omega : X(\omega) = 0\} = \{(T,T)\} \end{aligned}$$

are subsets of Ω and hence are *events* of the random experiment.

So we can see that a probability function defined on the events of the experiment indirectly leads to a distribution of probabilities across the possible values of the random variable. We formalise this as follows.

Definition: Consider a random experiment with sample space Ω. Let X be a random variable defined on Ω. Then, for $x \in S_X$ the probability that X is equal to x, denoted $\mathbb{P}(X = x)$, is the probability of the event $A_x \equiv \{\omega : X(\omega) = x\}$. Thus

$$\mathbb{P}(X = x) = \mathbb{P}(A_x).$$

Consequently we can think of statements involving random variables as a form of shorthand eg

- $X = x$ for $\{\omega : X(\omega) = x\}$.
- $X \leq x$ for $\{\omega : X(\omega) \leq x\}$.
- $x < X \leq y$ for $\{\omega : x < X(\omega) \leq y\}$.

This shorthand reflects a shift in our interest from the random experiment as a whole $(\Omega, \mathbb{P})$ towards the distribution of the random variable of interest $(X, \mathbb{P}(X = x))$.

Example

Toss two dice. The sample space is

$$\Omega = \{(1,1), \ldots, (6,6)\}$$

Let X denote the random variable whose value is the sum of the two faces. Assuming each outcome in Ω is equally likely,

$$\mathbb{P}(X = 2) = \mathbb{P}(\{\omega : X(\omega) = 2\}) = \mathbb{P}(\{(1,1)\}) = 1/36$$

$$\mathbb{P}(X = 3) =$$

$$\mathbb{P}(X = 4) =$$

etc. Of course other random variables may be of interest eg the minimum number showing or the maximum.

Definition

A set is said to be *countable* if it is either finite or can be put into a 1-1 correspondence with the set of natural numbers $\{1, 2, 3, \ldots\}$. That is, a set is countable if it is possible to list its elements in the form $x_1, x_2, \ldots$. Otherwise a set is *uncountable*.

$$
\begin{aligned}
N &= \{1, 2, \ldots\} \\
Z &= \{0, 1, -1, 2, -2, \ldots\} \\
Z \times Z &= \{(0,0), (0,1), (1,0), (0,-1), (-1,0), \ldots\}
\end{aligned}
$$

are all countable sets.

It is known, via a very elegant proof, that $[0, 1]$ and $\mathbb{R}$ are uncountable.

Discrete Random Variables (Ghahramani 4.3, 4.2)

A *discrete random variable* is one for which the set of possible values S_X is countable. That is, X can take only a countable number of values.

Definition

Let X be a discrete random variable. The *probability mass function (pmf)* $p_X(x)$ of X is the function from S_X to $[0, 1]$ defined by

$$p_X(x) = \mathbb{P}(X = x).$$

You can think of the $p_X(x)$ as discrete *masses* of probability assigned to each possible $x \in S_X$.

In the above, x is a dummy variable: we could use t or ζ or ξ or anything else. However, it is common to use x as a reminder that X is the random variable, and if it is clear that the pmf of X is intended, the subscript X may then be omitted.

We talk about the *probability mass function (pmf)* determining the *probability distribution* (or just *distribution* for short) of the discrete random variable X.

Note in Ghahramani, the *pmf* is defined on the domain $\mathbb{R}$, but as $p_X(x) = 0$ for all $x \notin S_X$ we prefer to restrict the domain to S_X.

Example

Let X be the sum of the numbers shown on the toss of two fair dice. Then $S_X = \{2, \ldots, 12\}$ and $p(x)$ is given by

$$\begin{aligned} p(2) &= \mathbb{P}(X = 2) = 1/36, \quad p(8) = 5/36, \\ p(3) &= \mathbb{P}(X = 3) = 2/36, \end{aligned}$$

Theorem : The *probability mass function* $p_X(x)$ of a discrete random variable X satisfies the following

1. $p_X(x) \geq 0, \quad \forall\, x.$

2. $$\sum_{x \in S_X} p_X(x) = 1.$$

Indeed any function satisfying (1) and (2) can be thought of as the pmf for some random variable.

Proof

Part (1) is obvious as:

$$\begin{aligned} p(x) &= \mathbb{P}(X = x) \\ &= \mathbb{P}(\{\omega : X(\omega) = x\}) \end{aligned}$$

and $0 \leq \mathbb{P}(A) \leq 1$ for all events A.

For (2) first note that for $x_1 \neq x_2$, the events

$$\{\omega : X(\omega) = x_1\}$$
$$\text{and} \quad \{\omega : X(\omega) = x_2\}$$

are disjoint. So

$$\begin{aligned}
\mathbb{P}(X = x_1 \text{ or } x_2) &= \mathbb{P}(\{\omega : X(\omega) = x_1 \text{ or } x_2\}) \\
&= \mathbb{P}(\{\omega : X(\omega) = x_1\} \cup \{\omega : X(\omega) = x_2\}) \\
&= \mathbb{P}(\{\omega : X(\omega) = x_1\}) + \mathbb{P}(\{\omega : X(\omega) = x_2\}) \\
&= \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2).
\end{aligned}$$

As S_X is the set of possible values of $X(\omega)$ for $\omega \in \Omega$, it follows that

$$\mathbb{P}(\{\omega : X(\omega) \in S_X\}) = \mathbb{P}(\Omega) = 1$$

but also

$$\begin{aligned} \mathbb{P}(\{\omega : X(\omega) \in S_X\}) &= \sum_{x \in S_X} \mathbb{P}(\{\omega : X(\omega) = x\}) \\ &= \sum_{x \in S_X} \mathbb{P}(X = x) \end{aligned}$$

Hence

$$\sum_{x \in S_X} \mathbb{P}(X = x) = 1. \quad \blacksquare$$

From the proof we can see that, for any set $B \subseteq \mathbb{R}$, given the *pmf*, we can compute the probability that $X \in B$ via

$$\mathbb{P}(X \in B) = \sum_{x \in B \cap S_X} p_X(x).$$

In particular

$$\mathbb{P}(X \leq x) = \sum_{y \leq x} p_X(y).$$

Example

Suppose that the discrete random variable X has pmf given by:

x	1	2	3	4	5
$p_X(x)$	α	2α	3α	4α	5α

Calculate α and $\mathbb{P}(2 \leq X \leq 4)$.

Distribution function (Ghahramani 4.2)

Definition: Let X be a discrete random variable. The *distribution function* $F_X(x)$ of X is the function from $\mathbb{R}$ to $[0, 1]$ defined by

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

Particularly in the statistical literature, the distribution function is sometimes referred to as the *cumulative distribution function (Cdf)*.

Properties of the distribution function (Ghahramani 4.2)

1. $0 \leq F_X(x) \leq 1$, since it is a probability.

2. $F_X(-\infty) = 0, F_X(\infty) = 1$, from the definition of the distribution function.

3. $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$, if $a < b$ since:

$$\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\}$$

and the events on the LHS are mutually exclusive.

4. $F_X(x)$ is non-decreasing. This follows from Property 3, since if $b > a$, then $F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b) \geq 0$.

5. $F_X(\cdot)$ is continuous on the right, that is $\lim_{h \downarrow 0} F_X(x+h) = F_X(x)$. Using Property 3 (**NB:** finite additivity is not enough), if $h > 0$, then

$$[F_X(x+h) - F_X(x)] = \mathbb{P}(x < X \leq x+h)$$

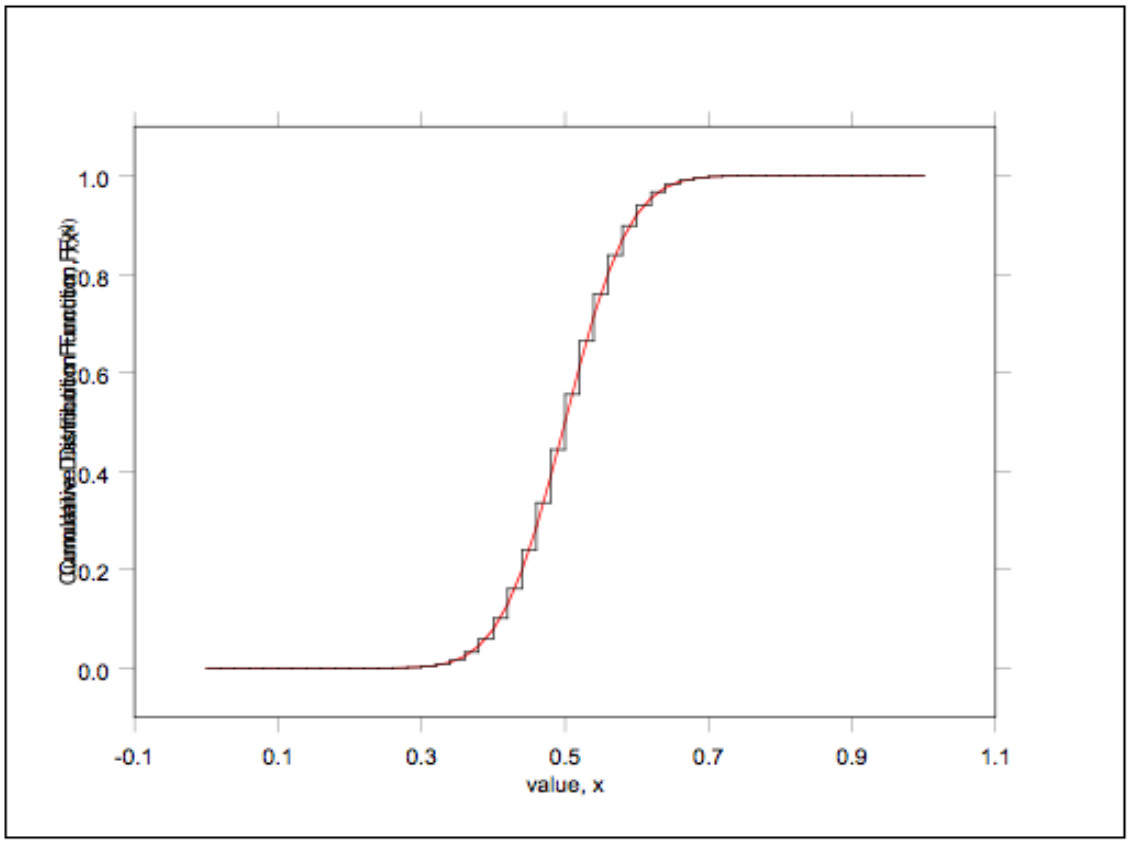
As $h \downarrow 0$, $\{x < X \leq x+h\} \to \emptyset$ and so the probability on the right hand side approaches zero.

6. $\mathbb{P}(X = x)$ is the jump in F_X at x. That is $\mathbb{P}(X = x) = F_X(x) - \lim_{h \downarrow 0} F_X(x-h)$. Again if $h > 0$, then

$$F_X(x) - F_X(x-h) = \mathbb{P}(x-h < X \leq x)$$

As $h \downarrow 0$, $\{x-h < X \leq x\} \to \{X = x\}$.

Large number of values with small prob



When there are many many values with small prob, the CDF can be approximated by a continuous curve, here shown in red

Continuous random variables (Ghahramani 6.1, 4.2)

- If S_X is *uncountable* we call X a *continuous random variable*.
- It is not possible to assign probability masses directly to every possible value of a continuous random variable.
- We deal with this by assigning probabilities to intervals.

Distribution function

Definition

Let X be a continuous random variable. The *distribution function* $F_X(x)$ of X is the function from $\mathbb{R}$ to $[0, 1]$ defined by

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

Probability density function

Let X be a continuous random variable. A function $f_X(x)$ which is such that, for all $x \in \mathbb{R}$,

$$\int_{-\infty}^{x} f_X(y)dy = F_X(x)$$

is called a *probability density function (pdf)* of X.

If such a function exists, then it is unique.

If a *pdf* exists then for "almost all" values of x we also have

$$\frac{dF_X(x)}{dx} = f_X(x).$$

Properties of the pdf

1. $f_X(x) \geq 0$ since $F_X(x)$ is non-decreasing.
2. $\int_a^b f_X(t)dt = F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b)$, i.e., probability is represented by the area under the graph of $f_X(x)$. For a random variable that has a density function

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(a \leq X < b) \\ = \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a < X < b) \end{aligned}$$

   since the end points have zero probability.

3. $\int_{-\infty}^{\infty} f_X(t)dt = 1$ since $F_X(\infty) = 1$ and $F_X(-\infty) = 0$.

Note that properties 1 and 3 are sufficient for f_X to be a *pdf*.

Discrete vs Continuous

Discrete	Continuous
pmf $\quad p_X(x)$	*pdf* $\quad f_X(x)$
prob. masses $p_X(x)$ at x	no masses $p_X(x) = 0 \quad \forall\, x$
$\sum_{x \in S_X} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(t)dt = 1$
$\mathbb{P}(X \in I) = \sum_{x \in I} p_X(x)$	$\mathbb{P}(X \in I) = \int_a^b f_X(t)dt$
$0 \leq p_X(x) \leq 1$	$0 \leq f_X(x)$

where $I = [a, b]$

There is no need to have $f_X(x) \leq 1$ since areas, not the value of the density function, represent probabilities. Thus, for example

$$f_X(x) = \begin{cases} 10^6 & 0 \leq x \leq 10^{-6} \\ 0 & \text{otherwise} \end{cases}$$

is a pdf since $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Pdf interpretation

Whilst the value of the *pdf* $f_X(x)$ is not the probability at x it can be interpreted as a probability density "around" x.

Define "$\{X \approx x\}$" to mean "$\{x - \frac{1}{2}\delta x < X \leq x + \frac{1}{2}\delta x\}$".
Then

$$\begin{aligned}
\mathbb{P}(X \approx x) &= \mathbb{P}\left(x - \frac{1}{2}\delta x < X \leq x + \frac{1}{2}\delta x\right) \\
&= \int_{x-\frac{1}{2}\delta x}^{x+\frac{1}{2}\delta x} f_X(u)du \\
&\approx f_X(x)\delta x.
\end{aligned}$$

So "almost everywhere" we have $\mathbb{P}(X \approx x) \approx f_X(x)\delta x$.

It follows that $f_X(x)$ gives a measure of the "relative probability" of an observed value near x, in the sense that

$$\frac{\mathbb{P}(X \approx x_1)}{\mathbb{P}(X \approx x_2)} \approx \frac{f_X(x_1)}{f_X(x_2)}.$$

The Story So Far

It is important to make sure that you are fully aware of the subtle distinctions between probability measures, random variables, probability mass and density functions and cumulative distribution functions. Now that we have seen them all, we will take a moment to review them and point out the differences. Consider a random experiment with sample space Ω.

Then

1. The probability measure $\mathbb{P}$ maps the set of events (i.e. subsets of Ω) to $[0, 1]$.
2. A random variable X maps Ω to $\mathbb{R}$.
3. For a discrete random variable, the probability mass function maps the set of possible values S_X to $[0, 1]$.
4. The distribution function maps $\mathbb{R}$ to $[0, 1]$.
5. For a continuous random variable with no point masses, the probability density function maps $\mathbb{R}$ to $[0, \infty)$.

We often talk about random variables and their probability mass and distribution functions without explicit reference to the underlying sample space. For example, in talking about an experiment in which a coin is tossed n times we may define the random variable X to be the number of heads that turns up, and then go on to talk about the probability mass and distribution functions of X (which are?).

This is an example of shorthand expression which mathematicians often use. However they only use it when they fully understand the situation. In a case such as that described above, it is understood that the underlying sample space is the set of sequences of H and T of length n without this fact having to be mentioned explicitly.

Expectation (Ghahramani 4.4)

The distribution function contains all the information about the distribution of a random variable. However, this information can be difficult to digest. Because of this, we often summarise the information by reducing it in some way.

The most common such measure is the expected value.

The concept of expectation first arose in gambling problems: Is a particular game a good investment? Consider the game where the winnings $\$W$ has pmf

w	-1	1	10
$\mathbb{P}(W = w)$	0.75	0.20	0.05

Is it worthwhile? If you played the game 1000 times, you would expect to lose $\$1$ about 750 times, to win $\$1$ about 200 times and to win $\$10$ about 50 times. Thus you will win about

$$\$\left(\frac{-1 \times 750 + 1 \times 200 + 10 \times 50}{1000}\right)$$

per game.

Your "expected winnings" are -5 cents per game. We say that the "expected value of W" is -0.05.

This gives an indication of the worth of the game: in the long run, you can expect to lose an average of about 5 cents per game.

Expectations of Discrete RV's (Ghahramani 4.4)

Let X be a discrete random variable with possible values in the set S_X, and probability mass function $p_X(x)$.

The *expected value* or *mean* of X, denoted by $\mathbb{E}[X]$, is defined by

$$\mathbb{E}[X] = \sum_{x \in S_X} x p_X(x)$$

provided the sum on the right hand side converges absolutely.

Note: Ghahramani uses A rather than S_X. We prefer the latter as A is often used to denote events.

Why absolutely convergence?

We toss a fair coin repeatedly until we get a head and let Y be the number of tosses needed to get the first head. The "reward" is $X = (-2)^Y$, find the pmf of X and its mean if exists.

Example

Find $\mathbb{E}[X]$ if X is the value of the upturned face after a toss of a fair die.

Solution

$\mathbb{E}[X]$ is not necessarily a possible value of X. It isn't in the example. We can never get $7/2$ to show on the face of a die. However, if we toss a die n times and let X_i denote the result of the ith toss, then we would expect that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

Thus, after a large number of tosses we expect the average of all the values of X to be close to $\mathbb{E}[X]$.

More generally, suppose any random experiment is repeated a large number of times, and the random variable X observed each time, then the average of the observed values should approximately equal $\mathbb{E}[X]$.

Another way to think of the expected value of a random variable is as the location of the "centre of mass" of its probability distribution.

We often denote the expected value by μ or μ_X to be clear which random variable is involved.

Example

A manufacturer produces items of which 10% are defective and 90% are non-defective. If a defective item is produced the manufacturer loses \$1, while a non-defective item yields a profit of \$5. If X is the profit on a single item, find $\mathbb{E}[X]$.

For any given item the manufacturer will either lose \$1 or make \$5. The interpretation of $\mathbb{E}[X]$ is that if the manufacturer makes a lot of items he or she can expect to make an average \$4.40 per item.

Expectations of Continuous RV's (Ghahramani 6.3)

The definition of the expected value of a continuous random variable is analogous to that for a discrete random variable.

Let X be a continuous random variable with possible values in the set S_X, and probability density function $f_X(x)$.

The *expected value* or *mean* of X, denoted by $\mathbb{E}[X]$ is defined by

$$\mathbb{E}[X] = \int_{x \in S_X} x f_X(x) dx$$

provided the integral on the right hand side converges absolutely.

The connection with the definition of the expected value of a discrete random variable can be seen by approximating the integral with a Riemann sum. Assume we can divide S_X up into n intervals of length δx. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_{x \in S_x} x f_X(x) dx \\
&\approx \sum_{i=1}^{n} x_i f_X(x_i) \delta x \\
&\approx \sum_{i=1}^{n} x_i \mathbb{P}(x_i \leq X < x_i + \delta x)
\end{aligned}$$

Example

If X has pdf $f_X(x) = cx^2(1-x) \quad (0 < x < 1)$, find c and $\mathbb{E}[X]$.

Expectation of functions (Ghahramani 4.4, 6.3)

In many situations we are interested in calculating the expected value of a function $\psi(X)$ of a random variable X. We need an accounting trick to do this.

Example

A couple plans to have three children. Assume that a child is equally likely to be a boy or a girl, find the expected number of girls in the three children.

Method 1 Let X be the number of girls in the three children, then $p_X(0) = 1/8$, $p_X(1) = 3/8$, $p_X(2) = 3/8$, $p_X(3) = 1/8$ with

$$\mathbb{E}(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5.$$

Method 2 We write

$\Omega = \{bbb, bbg, bgb, gbb, ggb, gbg, bgg, ggg\}$, then

$$\begin{aligned}
\mathbb{E}(X) &= X(bbb)\mathbb{P}(bbb) + X(bbg)\mathbb{P}(bbg) + X(bgb)\mathbb{P}(bgb) + X(gbb)\mathbb{P}(gbb) \\
&\quad + X(ggb)\mathbb{P}(ggb) + X(gbg)\mathbb{P}(gbg) + X(bgg)\mathbb{P}(bgg) \\
&\quad + X(ggg)\mathbb{P}(ggg) \\
&= 1.5
\end{aligned}$$

An accounting trick

$$\mathbb{E}(X) = \sum_{\text{all } \omega} X(\omega)\mathbb{P}(\omega).$$

Theorem

If X is a discrete random variable with set of possible values S_X and probability mass function $p_X(x)$, then, for any real-valued function ψ,

$$\mathbb{E}[\psi(X)] = \sum_{x \in S_X} \psi(x) p_X(x)$$

provided the sum converges absolutely.

Proof

$\psi(X)$ is a discrete rv, so by the accounting trick,

$$\begin{aligned}
\mathbb{E}[\psi(X)] &= \sum_{\text{all } \omega} \psi(X(\omega))\mathbb{P}(\omega) \\
&= \sum_{x \in S_X} \psi(x)\mathbb{P}(X = x) \\
&= \sum_{x \in S_X} \psi(x) p_X(x).
\end{aligned}$$

Example

Let's return to the toss of a fair die with X the number on the upturned face, find $\mathbb{E}[X^2]$ and $\mathbb{E}[X]^2$.

Theorem

If X is a continuous random variable with set of possible values S_X and probability density function $f_X(x)$, then, for any real-valued function ψ,

$$\mathbb{E}[\psi(X)] = \int_{x \in S_X} \psi(x) f_X(x) dx$$

provided the integral converges absolutely.

Example

If X has pdf $f_X(x) = 2x \quad (0 < x < 1)$, find $\mathbb{E}(X)$ and $\mathbb{E}\left[\frac{1}{X}\right]$.

Sol:

Note that

$$\mathbb{E}\left[\frac{1}{X}\right] \neq \frac{1}{\mathbb{E}[X]} = \frac{3}{2}.$$

Generally, $\mathbb{E}[\psi(X)] \neq \psi(\mathbb{E}[X])$, with one important exception, when ψ is a linear function.

Theorem If X is a random variable and a and b are constants, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof We do the discrete case. The continuous is similar.

$$\begin{aligned}
\mathbb{E}[aX + b] &= \sum_{x \in S_X} (ax + b) p_X(x) \\
&= \sum_{x \in S_X} axp_X(x) + \sum_{x \in S_X} bp_X(x) \\
&= a \textstyle\sum_{x \in S_X} xp_X(x) + b \sum_{x \in S_X} p_X(x) \\
&= a\mathbb{E}[X] + b. \quad \blacksquare
\end{aligned}$$

Variance (Ghahramani 4.5, 6.3)

One particular function of a random variable gives rise to a measure of the "spread" of the random variable.

Definition The *variance* $V(X)$ or $\mathrm{Var}(X)$ of a random variable X is defined by

$$V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

$V(X)$ measures the *consistency* of outcome – a small value of $V(X)$ implies that X is more often near $\mathbb{E}[X]$, whereas a large value of $V(X)$ means that X varies around $\mathbb{E}[X]$ quite a lot.

Example

Consider the batting performance of two cricketers, one of whom hits a century (exactly 100) with probability 1/2 or gets a duck (0) with probability 1/2. The other scores 50 every time.

Let X_1 be the random variable giving number of runs scored by the first batsman and X_2 the number of runs scored by the second batsman. Then

$$\begin{aligned} \mathbb{E}[X_1] &= \tfrac{1}{2} \times 0 + \tfrac{1}{2} \times 100 = 50 \\ \mathbb{E}[X_2] &= 1 \times 50 = 50. \end{aligned}$$

However

$$\begin{aligned} V(X_1) &= \tfrac{1}{2}(0-50)^2 + \tfrac{1}{2}(100-50)^2 \\ &= \tfrac{1}{2}\,(2500) + \tfrac{1}{2}\,(2500) \\ &= 2500 \\ V(X_2) &= 1 \cdot (50-50)^2 = 0 \end{aligned}$$

which reflects the fact that the second batsman is more consistent.

From the definition of the variance we can see that the more widespread the likely values of X, the larger the likely values of $(X - \mu)^2$ and hence the larger the value of $V(X)$.

This is why the variance is a measure of spread. We often denote the variance by σ^2, or σ_X^2 to be clear which random variable is involved.

The square root of $V(X)$ is called the *standard deviation* and is denoted by σ_X, $sd(X)$ or just σ if the random variable involved is clear. As the units of the standard deviation and the random variable are the same, spread is often measured in standard deviation units.

There are alternative measures of spread.

For example the *mean deviation*, $d = \mathbb{E}(|X - \mu|)$. However for various mathematical reasons the variance (and standard deviation) are by far the most frequently used.

Notes on variance

1. $V(X) \geq 0$ since $(X - \mu)^2 \geq 0$.
2. $V(X) = 0 \Leftrightarrow \mathbb{P}(X = \mu) = 1$.
3. $V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. This form is almost always the best to use when evaluating $V(X)$.
4. If $Y = aX + b$, then $V(Y) = a^2 V(X)$ and $sd(Y) = |a|\ sd(X)$.
5. If X has mean μ and variance σ^2, $X_s = \frac{X-\mu}{\sigma}$ has mean 0 and variance 1. X_s is called a standardised random variable.
6. The mean and variance do not determine the distribution - they just give some idea of the centre and spread.

Theorem

$V(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Proof: Let $\mu = \mathbb{E}[X]$, then

$$\begin{aligned} V(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mu + \mu^2 \\ &= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2. \quad \blacksquare \end{aligned}$$

Example

Calculate $V(X)$ where X represents the roll of a fair die.

Solution We saw before that

$$\begin{aligned} \mathbb{E}[X] &= \frac{7}{2} \\ \mathbb{E}[X^2] &= \frac{91}{6} \\ \text{and so } V(X) &= \frac{91}{6} - \frac{49}{4} = \frac{35}{12}. \end{aligned}$$

Higher moments of a random variable (Ghahramani 4.5, 11.1)

We will consider these in more detail later in the course but simply note these definitions at this point:

The k^{th} moment (about the origin) of a random variable X is given by $\mu_k = \mathbb{E}(X^k)$.

The k^{th} central moment (about the mean) of a random variable X is given by $\nu_k = \mathbb{E}\big((X - \mu)^k\big)$.

So the mean $\mathbb{E}(X)$ is the first moment of X and the variance $V(X)$ is the second central moment of X.

Special Probability Distributions (Ghahramani, Ch. 5 & 7)

Certain classes of random experiment and random variables defined upon them turn up so often that we define and name standard functions as their distribution functions, pmfs or pdfs as appropriate.

We shall now discuss the most common examples of such random variables.

Discrete random variables

Some discrete distributions which arise frequently in modelling real world phenomena are:

- Bernoulli
- Binomial
- Geometric
- Negative Binomial
- Hypergeometric
- Poisson
- Uniform

Continuous random variables

Some continuous distributions which arise frequently in modelling real world phenomena are:

- Uniform
- Normal
- Exponential
- Gamma

Bernoulli Random Variables (Ghahramani 5.1)

If the random experiment has two possible outcomes (or two categories of outcomes)

- success and failure
- up and down
- defective and non-defective
- right and wrong
- true and false
- ...

A random experiment in which such a dichotomy is observed is called *a Bernoulli trial*.

We write $\Omega = \{S, F\}$ and let $X(S) = 1$ and $X(F) = 0$.

The random variable X is known as a *Bernoulli random variable*. If p is the probability of a success, then the probability mass function is

$$\begin{aligned} p(0) &= 1 - p \\ p(1) &= p \end{aligned}$$

The value $p \in [0, 1]$ is a *parameter*. By varying it, we get different members of the family of Bernoulli random variables.

We say X has a *Bernoulli distribution* with parameter p.

Bernoulli mean and variance

Applying our formulae for expectations we have

$$\begin{aligned}
\mathbb{E}(X) &= 0 \times (1-p) + 1 \times p = p. \\
\mathbb{E}(X^2) &= 0^2 \times (1-p) + 1^2 \times p = p. \\
V(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1-p).
\end{aligned}$$

Many random experiments take the form of sequences of Bernoulli trials. Moreover the physical processes that govern the experiments are such that we judge that it is reasonable to assume that the outcomes of the Bernoulli trials are mutually independent.

Thus, for example, the Bernoulli trials might be tosses of separate coins, or the same coin at different time points.

The Bernoulli, Binomial, Geometric and Negative Binomial random variables all arise in the context of a *sequence of independent Bernoulli trials*. They each summarise different aspects of the observed sequence of "successes" and "failures".

Binomial random variables (Ghahramani 5.1)

Consider a sequence of n independent Bernoulli trials with $p = \mathbb{P}(\text{success})$. The sample space Ω for such an experiment could be taken to be the set of all sequences of the form

$$\omega = \underbrace{S\ S\ F\ F\ S\ S\ F \ldots S}_{n \text{ letters}}$$

and the probability of any given sequence occurring is

$$\mathbb{P}(\{\omega\}) = p^{\text{no. of successes}} \times (1-p)^{\text{no. of failures}} \qquad (*)$$

However we usually aren't interested in the precise sequence ω that comes up. More interesting is the total number of successes.

Define the random variable $N(\omega) = \text{No. of successes in } \omega$.

To find the *pmf* p_N of N, we need $\mathbb{P}(N = k)$, which is the probability of the event:

$$A_k = \{\omega : N(\omega) = k\} \quad k = 0, 1, \ldots, n$$

From $(*)$ above we know that the probability of any given $\omega \in A_k$ is

$$p^k (1-p)^{n-k}$$

Thus to get $\mathbb{P}(A_k)$ all we need to do is count how many ω there are in A_k.

This is known to be

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}.$$

So we have

$$p_N(k) = \mathbb{P}(N = k) = \mathbb{P}(A_k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Binomial distribution

If $X =$ no. of successes in n independent Bernoulli trials with $p = \mathbb{P}(\text{success})$ then

$$p_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

and we say X has a *Binomial distribution* with parameters n and p and write $X \stackrel{d}{=} \text{Bi}(n, p)$.

Note that we can then write $X \stackrel{d}{=} \text{Bi}(1, p)$ for a Bernoulli distribution.

Using the Binomial Theorem we can verify that

$$\begin{aligned}\sum_{x=0}^{n} p_X(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p+1-p)^n \\ &= 1^n = 1\end{aligned}$$

Note: The *Binomial Theorem* is very important in many areas of mathematics and is certainly something you should know. It states that for integer n

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} .$$

Binomial mean and variance

Applying our formulae for expectations we can deduce that

$$\begin{aligned} \mathbb{E}(X) &= np. \\ \mathbb{E}(X(X-1)) &= n(n-1)p^2. \\ V(X) &= \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 \\ &= np(1-p). \end{aligned}$$

Example - Tay-Sachs disease

This is a hereditary metabolic disorder caused by a recessive genetic trait. When both parents are carriers their child has probability $\frac{1}{4}$ of being born with the disease (Mendel's first law).

If such parents have four children, what is the probability distribution for X = no. of children born with the disease? State assumptions.

Binomial distribution shape

We can show that the ratio of successive binomial probabilities $r(x)$ satisfies

$$r(x) = \frac{p_X(x)}{p_X(x-1)} = \frac{\frac{n+1}{x} - 1}{\frac{1}{p} - 1} \quad x = 1, 2, \ldots, n$$

which decreases as x increases.

The formula

$$\frac{p_X(x)}{p_X(x-1)} = \frac{\frac{n+1}{x} - 1}{\frac{1}{p} - 1} \quad x = 1, 2, \ldots, n$$

is useful for computing the binomial probabilities and is called the recursive formula.

If $x < p(n+1)$ then $r(x) > 1$ and the pmf is increasing.

If $x > p(n+1)$ then $r(x) < 1$ and the pmf is decreasing.

So the Binomial distribution only has a single "peak".

Exercise: Find values of n and p so that two successive binomial probabilities are the same.

Sampling with replacement

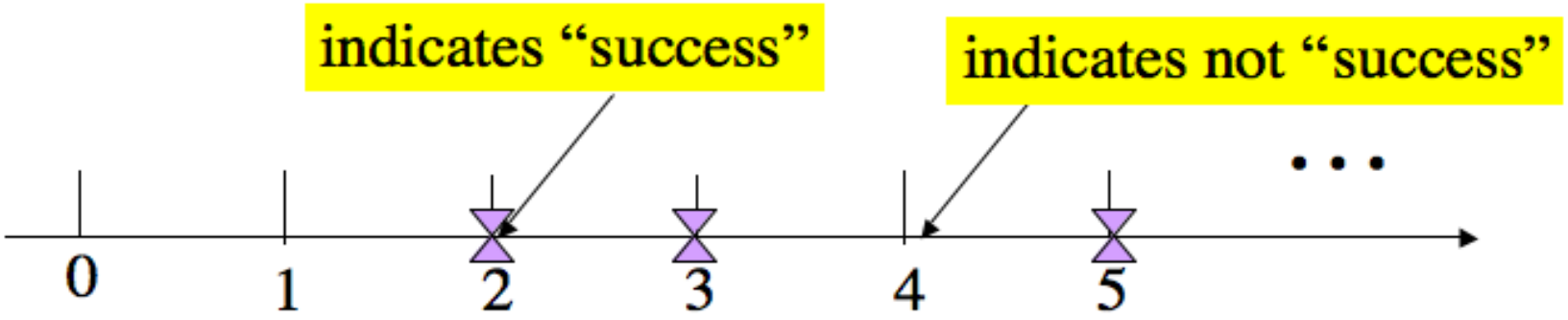
Suppose that a population consists of N objects, a proportion p of which are defective. A sample of n is obtained by selecting one object at random from the population, replacing it, selecting at random again, and so on.

We therefore have a sequence of independent Bernoulli trials with $\mathbb{P}(\text{success}) = p$.

Thus if X = number of defectives obtained in the sample, then $X \stackrel{d}{=} \text{Bi}(n, p)$.

Geometric random variables (Ghahramani 5.3)

- Why, in principle, we can't have infinitely many independent events all with the same probability ie infinitely many Bernoulli trials



- Let N be the random variable which gives the number of "failures" before there is a "success".

- The values for N are $\{0, 1, 2, \dots\}$ and N is therefore a discrete random variable
- N is a time if the trials occur in time
- The pmf of N:

$$\begin{aligned}
\mathbb{P}(N=0) &= \mathbb{P}(\{\omega : N(\omega) = 0\}) = \mathbb{P}(S) = p \\
\mathbb{P}(N=1) &= \mathbb{P}(\{\omega : N(\omega) = 1\}) = \mathbb{P}(FS) = (1-p)p \\
\mathbb{P}(N=2) &= \mathbb{P}(\{\omega : N(\omega) = 2\}) = \mathbb{P}(FFS) = (1-p)(1-p)p \\
&\dots \\
\mathbb{P}(N=n) &= (1-p)(1-p)\dots(1-p)p.
\end{aligned}$$

So the *pmf* for N is

$$p_N(n) = \mathbb{P}(N=n) = (1-p)^n p \quad n = 0, 1, 2, \dots$$

- We say N has a *Geometric distribution* with parameter p and write $N \stackrel{d}{=} \mathrm{G}(p)$.

As a check look at

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbb{P}(N=n) &= \sum_{n=0}^{\infty} (1-p)^n p \\ &= \frac{p}{1-(1-p)} \\ &= 1.\end{aligned}$$

Note: Remember that if $|x| < 1$ then

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

This is the formula for the sum of the geometric series.

Important Note: The Geometric distribution is defined slightly differently in some texts, including Ghahramani, by counting the number of *trials* until the first success rather than the number of failures before it.

If we define $M =$ to be the number of trials until the first success then clearly $M = N + 1$ where $N \stackrel{d}{=} G(p)$ under our definition. So instead of taking the values $0, 1, 2, \ldots$ the random variable M takes values $1, 2, 3, \ldots$. The distribution of M is simply shifted in location by one unit to the right.

Example

A computer communication channel transmits bits correctly with probability 0.95 independently of all other bits. What is the probability that there are at least 2 bits transmitted correctly before there is one transmitted incorrectly?

Find the probability of at least 10 correct before the first incorrect transmission in a communication channel given that there are at least 8 correct.

Lack of memory property

A curious property of the geometric distribution is the so called "lack of memory" property. If $T \stackrel{d}{=} G(p)$ then, for $t = 0, 1, 2, \ldots$

$$\mathbb{P}(T \geq t) = p(1-p)^t + p(1-p)^{t+1} + \ldots = \frac{p(1-p)^t}{p} = (1-p)^t$$

So for given $a, t = 0, 1, 2, \ldots$, we have:

$$\mathbb{P}(T - a \geq t \,|\, T \geq a) \quad = \quad (1-p)^t$$

Hence given that the first a trials were all failures, the *"residual"* time $T - a$ till the first success will have the same $\mathrm{G}(p)$ distribution as the original T.

The information that there has been no successes in the past a trials has no effect on the future waiting time to a success: the process "forgets" — the past has no effect on the future.

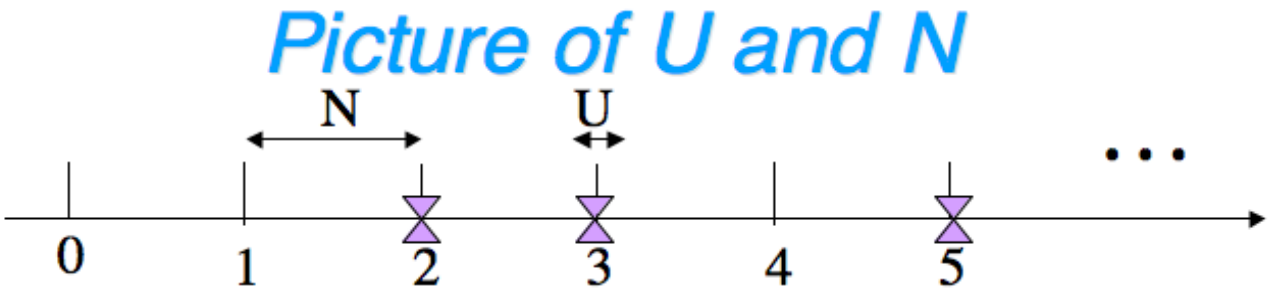
Geometric mean and variance

Applying our formulae for expectations we can deduce that

$$\begin{aligned} \mathbb{E}(X) &= \frac{(1-p)}{p}. \\ V(X) &= \frac{(1-p)}{p^2}. \end{aligned}$$

Waiting time to the second success

- Let U be the random variable defined as the number of "failures" after the first "success" before the second "success".



- Also U follows $G(p)$.
- Clearly, U and N will often be different in value and we say they equal in distribution.

Total number of trials till rth

- Let $r = 1, 2, \ldots$
- Consider now the total number of failures, Z, till the rth success
- For $r = 1$, Z is always N (ie $Z = N$, not just in distribution)
- For $r = 2$, $Z = N + U$
- In the picture, for $r = 2$, $N = 1$, $U = 0$, $Z = 1$
- Z is the sum of r independent Geometric random variables. We want to find the distribution of Z.

Negative Binomial random variables (Ghahramani 5.3)

We first observe that one way in which the event "$Z = z$" can occur is

$$\begin{array}{c|c} F\ F\ \ldots\ F \quad S\ S\ \ldots\ S & S \\ \leftarrow \quad z \quad \rightarrow \quad \leftarrow r-1 \rightarrow & \end{array}$$

The probability of this sequence is $(1-p)^z p^r$. If the first $z+r-1$ results are arranged amongst themselves, leaving the final rth S, the event $Z = z$ still occurs. This can be done in

$$\binom{z+r-1}{r-1}$$

ways; and for each arrangement, the probability is $(1-p)^z p^r$.

The pmf of Z is therefore given by

$$\begin{aligned} p_Z(z) &= \mathbb{P}(Z = z) = \binom{z+r-1}{r-1} p^r (1-p)^z \\ &= \frac{r(r+1)\dots(r+z-1)}{z!} p^r (1-p)^z \\ &= \frac{(-1)^z(-r)(-r-1)\dots(-r-z+1)}{z!} p^r (1-p)^z \\ &= \binom{-r}{z} p^r (-(1-p))^z, \quad z = 0, 1, 2, \dots \end{aligned}$$

where we extended

$$\binom{x}{k} = \frac{x(x-1)\dots(x-k+1)}{k!}$$

for nonnegative integers x and k to all real x and nonnegative integer k

Note: We define $\binom{x}{0} = 1$ for all real x.

If r happens to be an integer we have shown that

$$\binom{-r}{z} = (-1)^z \binom{z+r-1}{r-1}.$$

Using

$$p_Z(z) = \binom{-r}{z} p^r (p-1)^z, \quad z = 0, 1, 2, \ldots$$

we can show that this is a well defined pmf for all real $r > 0$.
We will need to use an extended version of the Binomial Theorem, corresponding to our extended definition of $\binom{x}{k}$.

Extended Binomial Theorem

For any real r (as opposed to just integer r) we have

$$(1+b)^r = \sum_{k=0}^{\infty} \binom{r}{k} b^k$$

which converges provided $|b| < 1$.

Provided $p \neq 0$, it follows that

$$p^r \sum_{z=0}^{\infty} \binom{-r}{z} (p-1)^z = p^r (1+(p-1))^{-r} = p^r p^{-r} = 1.$$

However to ensure that all the individual terms are non-negative we need to have $r > 0$.

Negative Binomial Distribution

If the random variable Z has pmf

$$p_Z(z) = \binom{-r}{z} p^r (p-1)^z, \quad z = 0, 1, 2, \ldots$$

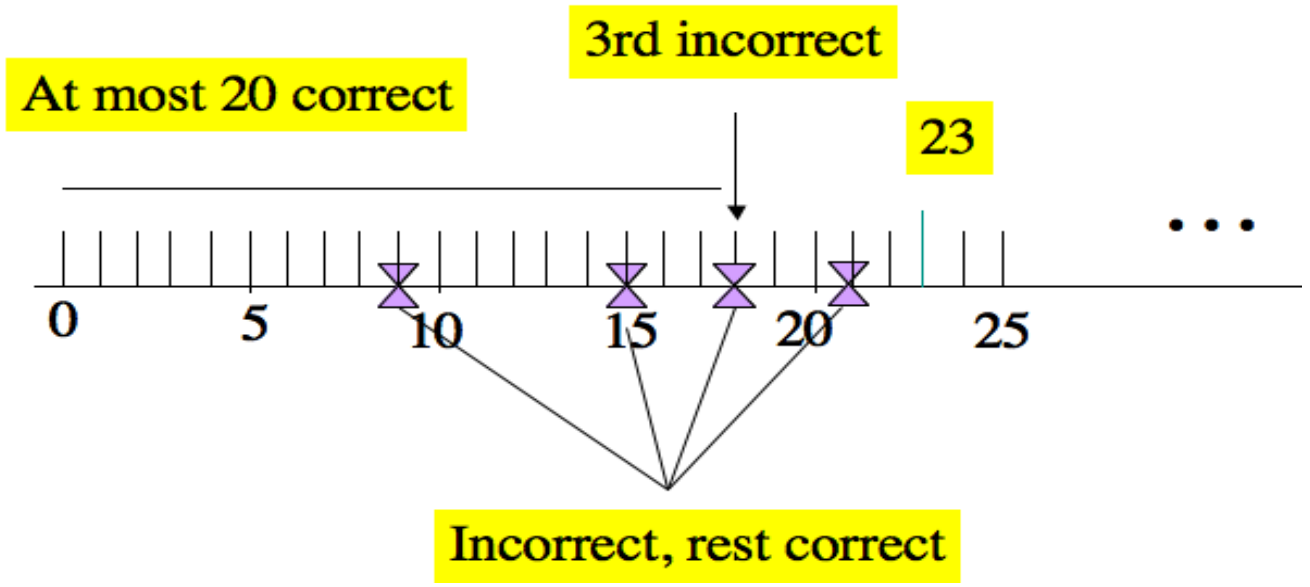
where $r > 0$ and $0 < p \leq 1$, then we say Z has a *Negative Binomial distribution* with parameters r and p and write $Z \stackrel{d}{=} \text{Nb}(r, p)$.

In the special case where r is an integer then Z can be interpreted as the number of failures before the rth success in a sequence of independent Bernoulli trials with $p = \mathbb{P}(\text{success})$.

Example (cont)

Consider the previous example of bits being transmitted correctly with probability 0.95. What is the probability that there are at most 20 correct transmissions till the third incorrect transmission?

- Here $r = 3$ and we require $\mathbb{P}(Z \leq 20)$
- Let X be the number of incorrect transmissions in the first 23 transmissions


3rd incorrect
At most 20 correct
23
0
5
10
15
20
25
Incorrect, rest correct

Example

To complete his degree a part time student needs to do three more subjects. Assuming he can only take one subject per semester, and that he passes a subject with probability 0.85 independently of his past results, find the probability that he will need more than 2 but not more than 3 years to graduate?

Important Note: Ghahramani defines the Negative binomial distribution as the sum of r independent Geometric random variables using its definition for the Geometric (which differs from ours). So in Ghahramani the Negative binomial takes values $(r, r+1, r+2, \ldots)$ and is simply shifted in location r units to the right.

Negative binomial mean and variance

We can deduce that

$$\begin{aligned} \mathbb{E}(X) &= \frac{r(1-p)}{p}. \\ V(X) &= \frac{r(1-p)}{p^2}. \end{aligned}$$

These results are most easily proved using techniques which we will examine later in the course.

Negative Binomial shape

We can show that the ratio of successive negative binomial probabilities $r(x)$ satisfies

$$r(z) = \frac{p_Z(z)}{p_Z(z-1)} = \left(\frac{r-1}{z} + 1\right)(1-p) \quad z = 1, 2, \ldots$$

which decreases as z increases.

The formula

$$\frac{p_Z(z)}{p_Z(z-1)} = \left(\frac{r-1}{z} + 1\right)(1-p) \quad z = 1, 2, \ldots$$

is useful for computing negative binomial probabilities and is called the recursive formula.

- If $z < \frac{1-p}{p}(r-1)$ then $r(z) > 1$ and the pmf is increasing.
- If $z > \frac{1-p}{p}(r-1)$ then $r(z) < 1$ and the pmf is decreasing.

So the Negative Binomial distribution only has a single "peak".

There is a relationship between the distribution functions of the Negative Binomial and Binomial distributions. Indeed we have $\{Z \leq n - r\}$ is the same as {at most $n - r$ failures before the rth success} which is the same as {at most n trials to get r successes} which is the same as {no of successes in first n trials $\geq r$}.

Hypergeometric random variables (Ghahramani 5.3)

When we looked at binomial random variables, we considered the experiment of sampling with replacement. In many, if not most, experiments of this type it is more natural to sample without replacement. If we perform such an experiment, then the number of successes is no longer binomially distributed.

We need a different distribution to describe the number of successes - the hypergeometric distribution.

- The population consists of N objects, a proportion p of which are defective.
- The number of defective items in the population is therefore $D = Np$.
- A sample of n is obtained by selecting n objects at random either all at once or sequentially without replacement. The two procedures are equivalent.
- Easiest to assume that the items are labelled: say, r_1, $\ldots$, r_D for the defetives and b_1, $\ldots$, b_{N-D} for the non-defectives.

The pmf

Let X be number of defectives obtained in the sample of n.

Then the pmf of X is given by

$$p_X(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \quad (x = 0, 1, 2, \ldots, n).$$

If $X = x$, then the sample contains x defectives and $n - x$ non-defectives. The defectives can be chosen in $\binom{D}{x}$ ways, and for each way of choosing the defectives, the non-defectives can be chosen in $\binom{N-D}{n-x}$ ways. So the number of ways of choosing a sample which contains x defectives is $\binom{D}{x}\binom{N-D}{n-x}$.

There are $\binom{N}{n}$ ways of choosing a sample of n from a population of N, and each is equally likely since the selection is made at random.

The expression for the pmf follows.

Note that $p_X(x) > 0$ only if $0 \leq x \leq D$ and $0 \leq n - x \leq N - D$, since otherwise one or other of $\binom{D}{x}$ or $\binom{N-D}{n-x}$ is zero.

Therefore $p_X(x) > 0$ only if $A \leq x \leq B$, where $A = \max(0, n + D - N)$, and $B = \min(n, D)$.

Nevertheless, we usually denote the set of possible values S_X as $\{0 \leq x \leq n\}$ allowing that some of these values may actually have zero probability.

Clearly, $p_X(x) \geq 0$. It can be shown that $\sum p_X(x) = 1$, by equating coefficients of s^n on both sides of the identity $(1 + s)^D (1 + s)^{N-D} = (1 + s)^N$.

Hypergeometric Distribution

If X has pmf $p_X(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$, $(x = A, \ldots, B)$, where A and B are defined above, then we say that X has a *hypergeometric distribution* with parameters n, D and N and we write $X \stackrel{d}{=} \mathrm{Hg}(n, D, N)$.

Hypergeometric mean and variance

It can be shown that

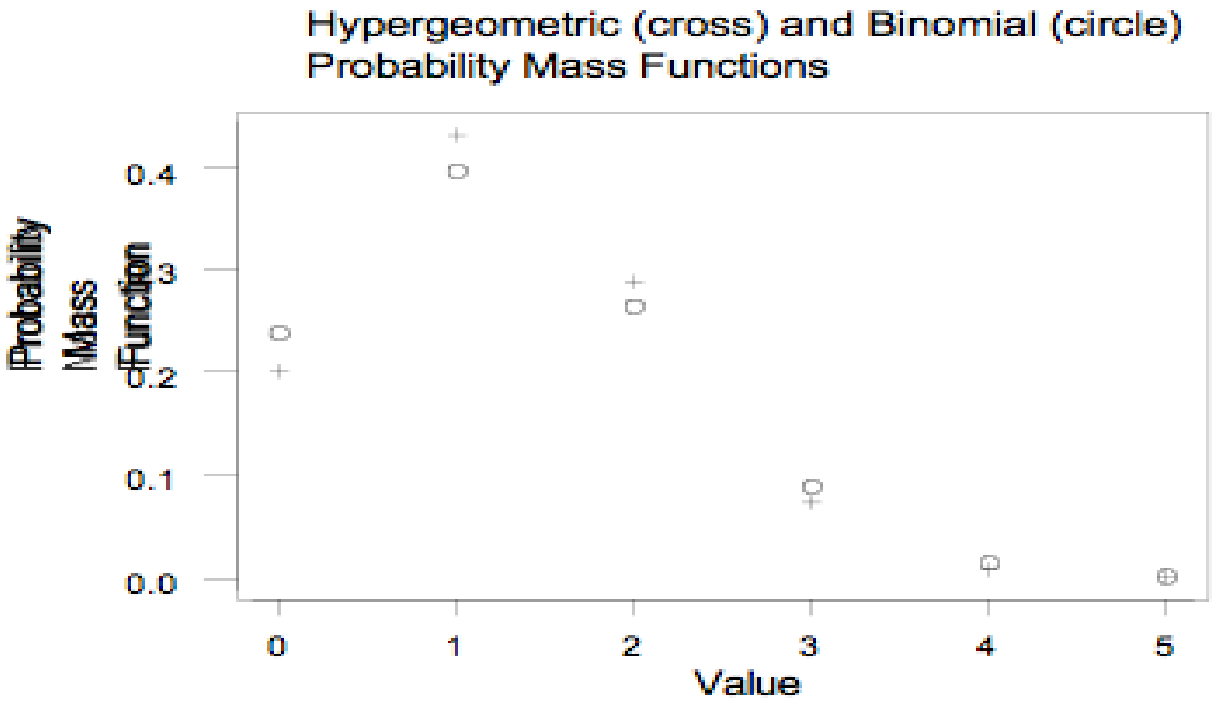
$$\begin{aligned} \mathbb{E}(X) &= \frac{nD}{N}. \\ V(X) &= \frac{nD(N-D)}{N^2} \times \left(1 - \frac{n-1}{N-1}\right). \end{aligned}$$

It is interesting to compare these formulae to those for the Binomial distribution with p replaced by the percentage of defectives $\frac{D}{N}$.

In particular, a hypergeometric distribution is approximated by the binomial distribution with parameters n and p if n is small compared to N, because there is not much difference between sampling with and without replacement

Example

Compute the hyper-geometric probabilities for $N = 24$, $n = 5$ and $p = 0.25$ and the approximating binomial probabilities and plot them



Example

If a hand of five cards is dealt from a well-shuffled pack of fifty-two cards, the number of spades in the hand,
$X \stackrel{d}{=} \mathrm{Hg}(n = 5, D = 13, N = 52)$.

Poisson (1837) Random Variables (Ghahramani 5.2)

Recall that a Binomial random variable counts the total number of successes in a sequence of n independent Bernoulli trials. If we think of a "success" as an "event" then a Binomial random variable effectively counts events occurring in discrete time.

A Poisson random variable is an analogue of the Binomial random variable which effectively counts "events" occurring in continuous time. However both types count "events" so both are discrete random variables.

We derive the Poisson distribution via a limiting process involving sequences of Bernoulli trials as follows.

Assume that each Bernoulli trial takes a time $1/n$ to complete and that the probability of success in a Bernoulli trial is proportional to this time, say $\mathbb{P}(success) = \alpha/n$. Then, by time 1, we can complete n trials. Let N = number of "events" which occur by time 1. Then $N \stackrel{d}{=} \text{Bi}(n, \frac{\alpha}{n})$ and hence

$$\mathbb{P}\big(N = k\big) = \tbinom{n}{k}(\tfrac{\alpha}{n})^k(1 - \tfrac{\alpha}{n})^{n-k}$$

Now we shrink the length of time for each trial and the success probability at the same rate, by letting $n \to \infty$.

It is a basic mathematical fact that

$$\lim_{n\to\infty}(1+\frac{x}{n})^n = e^x$$

and so

$$\lim_{n\to\infty}\left(1-\frac{\alpha}{n}\right)^n = e^{-\alpha}.$$

So we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\big(N = k\big) &= \lim_{n\to\infty} \tbinom{n}{k} (\tfrac{\alpha}{n})^k (1 - \tfrac{\alpha}{n})^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{n^k (n-k)!} \, \frac{\alpha^k}{k!} \, (1 - \tfrac{\alpha}{n})^n (1 - \tfrac{\alpha}{n})^{-k} \\
&= 1 \times \frac{\alpha^k}{k!} \times e^{-\alpha} \times 1 \\
&= \frac{e^{-\alpha}\alpha^k}{k!} \quad (k = 0, 1, 2, \ldots).
\end{aligned}$$

If N has pmf $p_N(k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ $(k = 0, 1, 2, \ldots)$, we say that N has a *Poisson distribution* with parameter λ, and we write $N \stackrel{d}{=} \mathrm{Pn}(\lambda)$.

So in this case $N \stackrel{d}{=} \mathrm{Pn}(\alpha)$.

Note that $p_N(k) \geq 0$, and that $\sum_{k=0}^{\infty} p_N(k) = 1$, since

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

This is the Taylor series expansion of e^{λ} and is another basic mathematical fact that is worth remembering.

Siméon Denis Poisson [21 June 1781 – 25 April 1840]

Poisson in action

von Bortkiewicz (1898) collected the number of fatalities that resulted from being kicked by a horse for 10 corps of Prussian cavalry over a period of 20 years, giving 200 corps-years worth of data.

number of deaths per year	Observed	Relative frequency	Poisson fitting
0	109	.545	.543
1	65	.325	.331
2	22	.110	.101
3	3	.015	.021
4	1	.005	.003

Example

Assume cars pass an isolated petrol station on a country road at a constant mean rate of 5 per hour, or equivalently 2.5 per half hour. Let N denote the number of cars which pass the petrol station whilst it is temporarily closed for half an hour one Friday afternoon. What is the probability that the station missed out on three or more potential customers?

The average number of cars ('events') in half an hour is 2.5 so $N \stackrel{d}{=} \mathrm{Pn}(2.5)$.

Poisson mean and variance

Applying our formulae for expectations we can deduce that

$$\begin{aligned} \mathbb{E}(X) &= \lambda. \\ \mathbb{E}(X(X-1)) &= \lambda^2. \\ V(X) &= \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 \\ &= \lambda. \end{aligned}$$

Note the interesting fact that the mean and variance of a Poisson random variable are equal.

Our original derivation of the Poisson distribution shows that it approximates the Binomial distribution under the right circumstances. As we shrunk the length of time for each trial and the success probability at the same rate, in the Binomial distribution for $N \stackrel{d}{=} \text{Bi}(n, \frac{\alpha}{n})$ we had

Number of trials:	$n \to \infty$
Probability of success:	$\frac{\alpha}{n} \to 0$
Average number of "events":	$(n)\frac{\alpha}{n} = \alpha$

Poisson approximation to Binomial

If p is small, the Poisson distribution with $\lambda = np$ can be used as a convenient approximation to the Binomial distribution.

$\mathrm{Bi}(n, p) \approx \mathrm{Pn}(np)$ for p small.

A rough rule is that the approximation is satisfactory if $p \leq 0.05$.

Poisson as an approximate model

Toss n coins with probabilities of heads $p_1, \ldots, p_n$, respectively, let X be the number of heads in the n tosses.

- If $p_1 = p_2 = \cdots = p_n = p$, then X has $\mathrm{Bi}(n, p)$
- If $p_1, \ldots, p_n$ are different, what is the distribution of X?

Poisson as an approximate model (2)

The Poisson process might be a reasonable model for

- Radioactive emissions
- Arrival of calls at a telephone exchange
- Times that people are diagnosed with a rare disease
- #earthquakes

Example

One in $10,000$ items from a production line is defective. The occurence of defects in successive items are independent. What is the probability that in a batch of $20,000$ items there will be at least 4 defective items?

If there is another production line which produces one defective in $5,000$ items. In a batch of $20,000$ items with half from each of the two production lines, what is the probability of at least 4 defective items? State assumptions.

Example

$100,000,000$ games of Super 66 Lotto are played. The probability of winning the Division 1 prize is $1/(45,360,620)$. How many Division 1 winners can be expected?

Discrete uniform random variables

Consider the discrete random variable X having *pmf*

$$p_X(x) = \frac{1}{n-m+1} \qquad (x = m, m+1, \ldots, n)$$

where m and n are integers such that $m \leq n$. We say that X has a *discrete uniform distribution* on $m \leq x \leq n$, and we write $X \stackrel{d}{=} \mathrm{U}(m, n)$.

If X denotes the result of throwing a fair die, then $X \stackrel{d}{=} \mathrm{U}(1, 6)$.

If Y denotes the result of spinning a roulette wheel (with one zero), then $Y \stackrel{d}{=} \mathrm{U}(0, 36)$.

Discrete uniform mean and variance

Applying our formulae for expectations we can deduce that for $X \stackrel{d}{=} U(0, n)$

$$\begin{aligned} \mathbb{E}(X) &= \frac{n}{2}. \\ \mathbb{E}(X^2) &= \frac{1}{6}n(2n+1). \\ V(X) &= \frac{1}{12}n(n+2) \end{aligned}$$

Example

Consider a sequence of independent Bernoulli trials with probability of success p. We are given the additional information that in the first n trials there is exactly one success. Let X denote the number of the (random) trial at which this single success occurred. What is the pmf of X?

Continuous uniform random variables (Ghahramani 7.1)

We have seen that a discrete uniform random variable X has a pmf that is constant over the possible values in the set S_X, which must be finite. In contrast a *continuous uniform random variable* X has a pdf that is constant over the possible values in the set S_X, which must be bounded.

Let $a < b$ be real numbers. Then a continuous random variable X, having pdf given by

$$f_X(x) = \frac{1}{b-a} \quad (a < x < b)$$

has a *continuous uniform* or *rectangular* distribution on the interval (a, b) and we write $X \stackrel{d}{=} \mathrm{R}(a, b)$.

Examples

1. If a set of real numbers are rounded down to the integer below, then for a number randomly selected from the set, it is often reasonable to assume that the error is distributed as $\mathrm{R}(0, 1)$.

2. If a needle is thrown randomly onto a horizontal surface, the acute angle that it makes with a specified direction, $\Theta \stackrel{d}{=} \mathrm{R}(0, \frac{\pi}{2})$.

Mean and variance of a Continuous Uniform Random Variable

Exercise: Show that if $X \stackrel{d}{=} R(a,b)$ then

$$\begin{aligned} \mathbb{E}(X) &= \frac{a+b}{2}. \\ V(X) &= \frac{1}{12}(b-a)^2. \end{aligned}$$

Exponential Random Variables (Ghahramani 7.3)

Recall that a geometric random variable modelled the number of failures before the first success in a sequence of Bernoulli trials.

Exponential random variables can be thought of as a continuous version of geometric random variables. They model the waiting time until a randomly-generated event occurs in continuous time. We derive them via a limiting process involving sequences of Bernoulli trials as follows.

Assume that each Bernoulli trial takes a time $1/n$ to complete and that the probability of success in a Bernoulli trial is proportional to this time, say $\mathbb{P}(success) = \alpha/n$. Then, in time t, we can complete nt trials and the probability that there is no success in a time period t is

$$\left(1 - \frac{\alpha}{n}\right)^{nt}.$$

Now we shrink the length of time for each trial and the success probability at the same rate, by letting $n \to \infty$.

Thus the probability that there is no event in time t is $e^{-\alpha t}$.

Let T be the waiting time to the first event. Then this is equivalent to saying that

$$\mathbb{P}(T > t) = e^{-\alpha t}.$$

The distribution function of T is therefore given by

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - \mathbb{P}(T > t) = 1 - e^{-\alpha t}, \quad \text{for} \quad t \geq 0$$

and as T can't be negative, $F_T(t) = 0$ for $t < 0$.

A continuous random variable T with this distribution is known as an *exponential random variable* with parameter α and we write $T \stackrel{d}{=} \exp(\alpha)$.

Differentiating tells us that the pdf of T is

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Note that $f_T(t) \geq 0$ and that $\int_{-\infty}^{\infty} f_T(t)dt = 1$.

Example

The waiting times between successive cars on a country road are exponentially distributed with parameter $\alpha = 10$. Find the probability that a gap between successive cars exceeds 0.2 time units.

Solution

$$T = \text{time gap between cars} \stackrel{d}{=} \exp(\alpha = 10)$$

Therefore $\mathbb{P}(T > t) = e^{-10 \times 0.2} = e^{-2} = 0.1353$.

Exponential Mean and Variance

Using the definition of expectation, for $X \stackrel{d}{=} exp(\alpha)$ we can write

$$\mathbb{E}(X) = \int_0^\infty x f_X(x) dx = \int_0^\infty x \times \alpha e^{-\alpha x} dx.$$

Exercise: Use integration by parts to find $\mathbb{E}(X)$.

However there is another approach which uses a generally useful alternative formula for calculating $\mathbb{E}(X)$.

We can derive a formula for the expected value in terms of the distribution function, rather than the pmf or pdf. This is most easily seen in the case of a continuous random variable. Suppose $X \geq 0$ with pdf f_X, using

$$\frac{d}{dx}\{-(1 - F_X(x))\} = f_X(x)$$

and the integration by parts, we have

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \int_{0}^{\infty} x f_X(x) dx \\
&= \int_{0}^{\infty} (1 - F_X(x))\, dx.
\end{aligned}$$

Example

If $X \stackrel{d}{=} \exp(\alpha)$, then, for $x > 0$, $F_X(x) = 1 - e^{-\alpha x}$. Hence

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\, dx = \int_0^\infty e^{-\alpha x} dx = \frac{1}{\alpha}.$$

To calculate the variance for $X \stackrel{d}{=} exp(\alpha)$ we first use repeated integration by parts to show

$$\mathbb{E}(X^2) = \int_0^\infty x^2 \alpha e^{-\alpha x} dx = \frac{2}{\alpha^2}.$$

Hence we have

$$V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}.$$

We note the interesting fact that the mean and standard deviation of the exponential distribution are equal.

Lack of memory property

Like the geometric distribution, the exponential distribution has the "lack of memory" property. If $X \stackrel{d}{=} exp(\alpha)$ then, for $x \in [0, \infty)$,

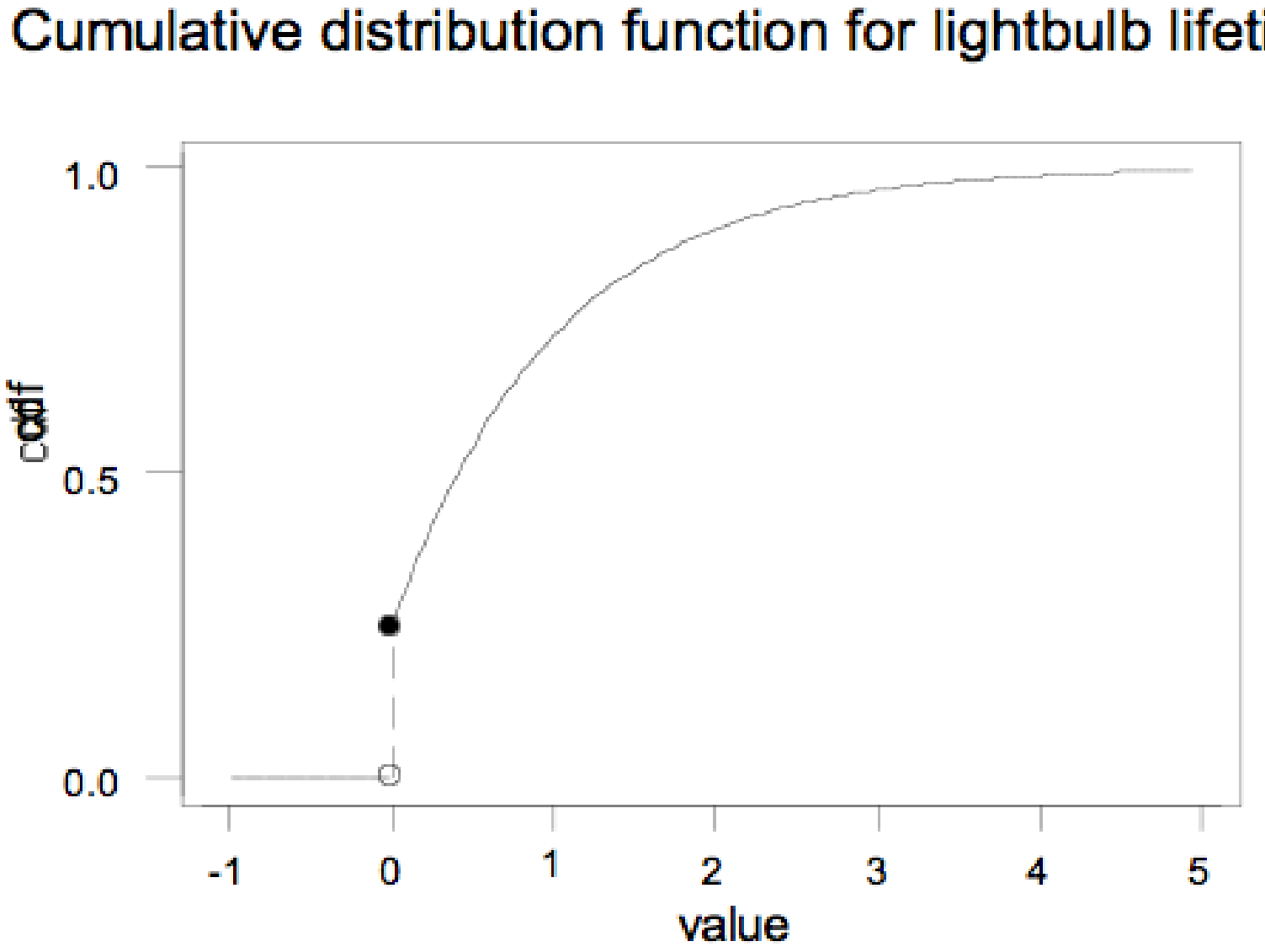
$$\mathbb{P}(X \geq x) = e^{-\alpha x}.$$

So for given $x, y \in [0, \infty)$, we have:

$$\begin{aligned}
\mathbb{P}(X - y \geq x \,|\, X \geq y) &= \mathbb{P}(X \geq x + y \,|\, X \geq y) \\
&= \frac{\mathbb{P}(X \geq x + y)}{\mathbb{P}(X \geq y)} \\
&= \frac{e^{-\alpha(x+y)}}{e^{-\alpha y}} = e^{-\alpha x} \\
&= \mathbb{P}(X \geq x).
\end{aligned}$$

Example

A lightbulb fails immediately with probability 0.25, and if it does not fail, then it has an exponential distribution with rate 0.01 days. Plot the cdf for the lifetime of the light bulb.


Cumulative distribution function for lightbulb lifetime
1.0
0.5
0.0
cdf
-1
0
1
2
3
4
5
value

[Slide 232]

Gamma random variables (Ghahramani 7.4)

An exponential random variable can be thought of as a continuous analogue of a geometric random variable. Both model the waiting time until the first occurrence of an event, one in continuous time and one in discrete time.

Similarly, a gamma random variable can be thought of as a continuous analogue of a negative binomial random variable. Both model the waiting time until the rth occurrence of an event, one in continuous time and one in discrete time.

Consider the same sequence of Bernoulli trials that we used to define the Poisson and exponential random variables. That is, each trial takes a time $1/n$ to complete and that the probability of success is $\mathbb{P}(success) = \alpha/n$.

Now, the probability that we have to wait for more than nz trials for the rth success is the same as the probability that we have less than r successes in the first nz trials.

The number of successes that we have in nz trials is binomially distributed, so the probability that we have less than r successes in the first nz trials is

$$\sum_{k=0}^{r-1} \binom{nz}{k} \left(\frac{\alpha}{n}\right)^k \left(1 - \frac{\alpha}{n}\right)^{nz-k}.$$

As $n \to \infty$, this approaches

$$\sum_{k=0}^{r-1} \frac{(\alpha z)^k}{k!} e^{-\alpha z}.$$

So, if we let Z denote the waiting time until the rth event, then

$$\mathbb{P}(Z > z) = \sum_{k=0}^{r-1} \frac{(\alpha z)^k}{k!} e^{-\alpha z}.$$

Therefore the distribution function of Z is

$$F_Z(z) = \mathbb{P}(Z \leq z) = 1 - \sum_{k=0}^{r-1} \frac{(\alpha z)^k}{k!} e^{-\alpha z}.$$

To get the pdf of Z, we differentiate this with respect to z. Thus

$$\begin{aligned} f_Z(z) &= \frac{d}{dz}\left[1 - \sum_{k=0}^{r-1} \frac{(\alpha z)^k}{k!} e^{-\alpha z}\right] \\ &= \alpha e^{-\alpha z} + \sum_{k=1}^{r-1} \left[-\frac{z^{k-1}\alpha^k}{(k-1)!} e^{-\alpha z} + \frac{z^k \alpha^{k+1}}{(k)!} e^{-\alpha z}\right] \\ &= \frac{\alpha^r z^{r-1}}{(r-1)!} e^{-\alpha z}. \end{aligned}$$

A continuous random variable Z with pdf given by

$$f_Z(z) = \frac{\alpha^r z^{r-1}}{(r-1)!} e^{-\alpha z}.$$

is known as an *gamma* or *Erlang* random variable with parameters r and α. We write $Z \stackrel{d}{=} \gamma(r, \alpha)$.

As we saw above, for $z > 0$, the distribution function of a Gamma random variable is

$$F_Z(z) = 1 - \sum_{k=0}^{r-1} \frac{(\alpha z)^k}{k!} e^{-\alpha z}.$$

Note that $\exp(\alpha) = \gamma(1, \alpha)$.

The scope of the gamma distribution can be extended to all real r by using the continuous analogue of the factorial, which is the gamma function

$$\Gamma(r) = \int_0^\infty e^{-x} x^{r-1} dx.$$

Using integration by parts, it is readily shown that $\Gamma(r) = (r-1)\Gamma(r-1)$ for all $r > 0$. Then, since $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$, it follows that if k is a positive integer then $\Gamma(k) = (k-1)!$

For $z > 0$, the function

$$f(z) = \frac{\alpha^r e^{-\alpha z} z^{r-1}}{\Gamma(r)}$$

is a pdf.

This follows because $f(z) \geq 0$ and also $\int_{-\infty}^{\infty} f(z)dz = 1$. This last statement follows from

$$\begin{aligned} \int_0^{\infty} e^{-\alpha z} \alpha^r z^{r-1} dz &= \int_0^{\infty} e^{-u} u^{r-1} du \\ &= \Gamma(r), \end{aligned}$$

using the substitution $u = \alpha z$.

Example

A machine has a component which fails once every 100 hours on average. The failures are randomly distributed in time. There are only three replacement components available. The time (in hours) for which the machine will remain operative Z is the waiting time until the fourth "event" when events occur at 0.01 (failures/hour).

Therefore $Z \stackrel{d}{=} \gamma(r = 4, \alpha = 0.01)$.

Gamma mean and variance

We start by deriving a formula for the kth moment of $X \stackrel{d}{=} \gamma(r, \alpha)$ as follows:

$$\begin{aligned}
\mathbb{E}(X^k) &= \int_0^\infty \frac{\alpha^r e^{-\alpha x} x^{r+k-1}}{\Gamma(r)} dx \\
&= \frac{1}{\alpha^k \Gamma(r)} \int_0^\infty e^{-u} u^{r+k-1} du \qquad [u = \alpha x] \\
&= \frac{\Gamma(r+k)}{\Gamma(r)\alpha^k}.
\end{aligned}$$

Note that:

1. $\mathbb{E}[X^k] \neq \mathbb{E}[X]^k = \frac{r^k}{\alpha^k}$ except if $k = 1$ or $k = 0$
2. if k is a positive integer, then $\mathbb{E}[X^k] = \frac{r(r+1)\cdots(r+k-1)}{\alpha^k}$
3. this result applies for all values of k for which the integral converges: for example $\mathbb{E}[X^{-1}] = \frac{\alpha}{r-1}$, provided $r > 1$.

Hence for $X \stackrel{d}{=} \gamma(r, \alpha)$:

$$\begin{aligned} \mathbb{E}(X) &= \frac{r}{\alpha} \\ V(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{(r+1)r}{\alpha^2} - \frac{r^2}{\alpha^2} = \frac{r}{\alpha^2}. \end{aligned}$$

Normal Random Variables (Ghahramani 7.2)

We turn now to one of the most important distributions in probability and statistics - the *Normal distribution*. It turns out, as we shall see later (Chapter 7), that the normal distribution occurs frequently as a limit. In particular the "Central Limit Theorem" says that the sum of a large number of independent and identically distributed random variables is approximately normally distributed *irrespective* of the specific underlying distribution, provided that it has finite mean and variance.

This is the primary reason for the importance of the normal distribution, which is endemic in nature.

If X has pdf

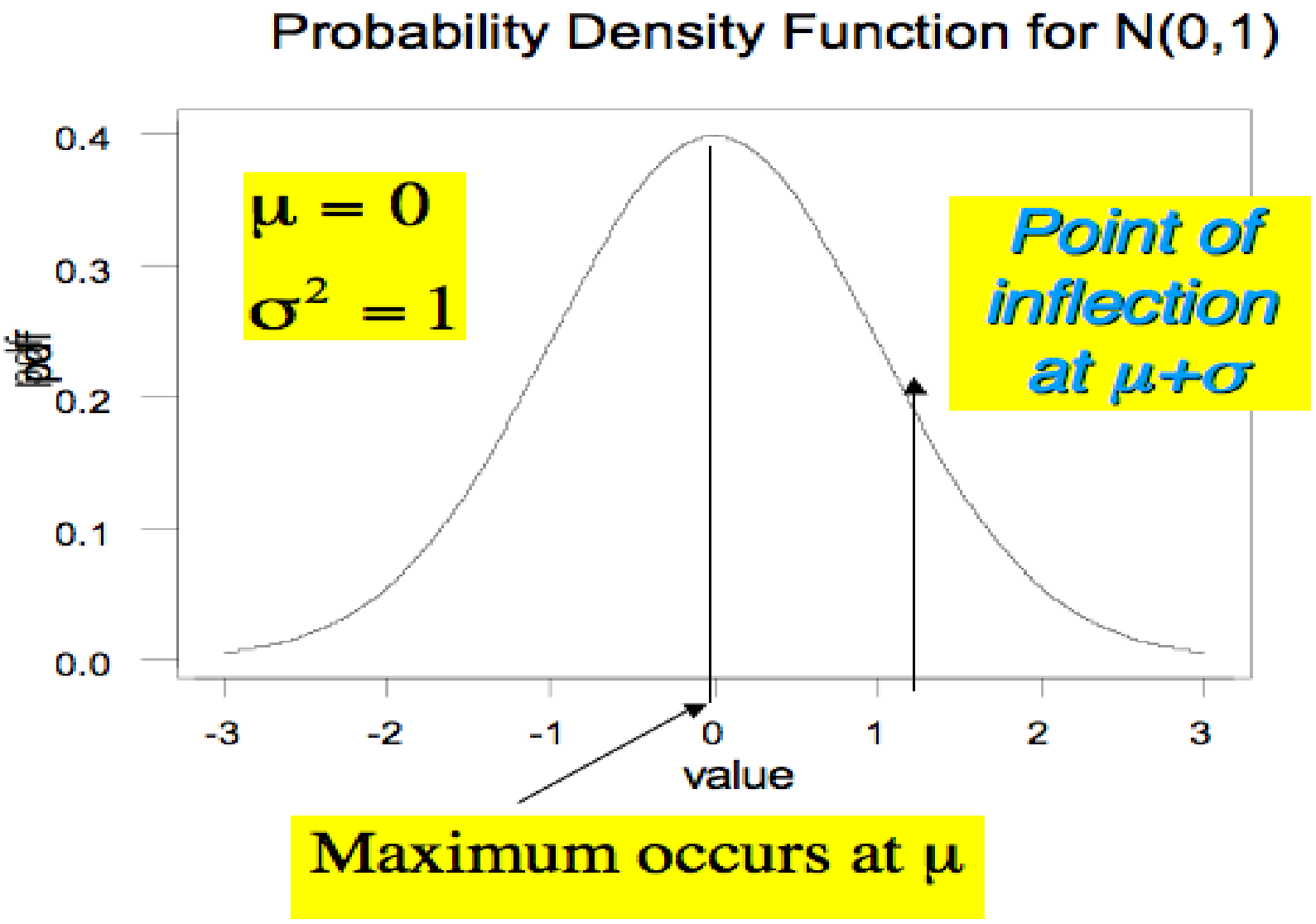
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (-\infty < x < \infty)$$

then we say that X has a *Normal distribution* with parameters μ and σ^2, and we write $X \stackrel{d}{=} \mathbb{N}(\mu, \sigma^2)$.

If $Z \stackrel{d}{=} \mathrm{N}(0,1)$, then we say that Z has a *Standard normal distribution*. The pdf of Z is given by

$$f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \quad (-\infty < z < \infty)$$

It is clear that $\varphi(z) \geq 0$. It is not so clear that $\int_{-\infty}^{\infty} \varphi(z)dz = 1$. We prove this later.


Probability Density Function for N(0,1)
μ = 0
σ² = 1
Point of inflection at μ+σ
Maximum occurs at μ
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
value

[Slide 246]

Standardisation

If $Z = \frac{X-\mu}{\sigma}$, with $\sigma > 0$, then the distribution function of Z is given by

$$\begin{aligned} F_Z(z) = \mathbb{P}(Z \leq z) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= \mathbb{P}(X \leq \mu + \sigma z) \\ &= F_X(\mu + \sigma z). \end{aligned}$$

Differentiating with respect to z gives

$$f_Z(z) = f_X(\mu + \sigma z)\sigma = \varphi(z).$$

Hence, we have shown that:

$$\text{If } X \stackrel{d}{=} \mathbb{N}(\mu, \sigma^2), \text{ then } Z = \frac{X-\mu}{\sigma} \stackrel{d}{=} \mathbb{N}(0, 1).$$

This means we can calculate probabilities for any normal distribution using the standard normal distribution.

The distribution function of $Z \stackrel{d}{=} \mathbb{N}(0,1)$ is denoted by

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

There is no "explicit formula" for this integral, so tables of $\Phi(z)$ have been compiled using numerical integration.

Tables 1 and 2 in Ghahramani give values of $\Phi(z)$. The standard normal distribution function is also available on many calculators. As $\varphi(z)$ is an even function, $\Phi(-z) = 1 - \Phi(z)$.

Example

If $X \stackrel{d}{=} \mathbb{N}(10, 25)$ and $Z \stackrel{d}{=} \mathbb{N}(0, 1)$, find $\mathbb{P}(X < 8)$.

The converse of our standardisation result is also easy to prove, namely:

$$\text{If } Z \stackrel{d}{=} \mathbb{N}(0,1), \text{ then } X = \mu + \sigma Z \stackrel{d}{=} \mathbb{N}(\mu, \sigma^2)$$

This is used in simulations to generate observations on any normal random variable by first generating observations on a standard normal and then transforming them as indicated.

It also implies that

$$\mathbb{E}(X) = \mu + \sigma \mathbb{E}(Z) \text{ and } V(X) = \sigma^2 V(Z)$$

so we can find the mean and variance of $X \stackrel{d}{=} \mathbb{N}(\mu, \sigma^2)$ from the corresponding quantities for Z. As an exercise we will actually find all of the moments of Z.

Moments of standard normal rv

For $Z \stackrel{d}{=} N(0,1)$, we have

$$\mathbb{E}[Z^n] = \int_{-\infty}^{\infty} z^n \, \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

Integrating by parts with $u = z^{n-1}$ and $dv = \left(1/\sqrt{2\pi}\right) z e^{-\frac{1}{2}z^2} dz$, we obtain

$$\mathbb{E}[Z^n] = \left[-\frac{1}{\sqrt{2\pi}} \, z^{n-1} e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} (n-1) z^{n-2} \, \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}z^2} dz.$$

Therefore

$$\mathbb{E}[Z^n] = (n-1)\mathbb{E}[Z^{n-2}].$$

We saw that

$$\mathbb{E}[Z^n] = (n-1)\mathbb{E}[Z^{n-2}].$$

We also know that $\mathbb{E}[Z^0] = 1$ (as $\phi(z)$ is a pdf) and $\mathbb{E}[Z^1] = 0$ (as the pdf is an even function). It follows that

$$\mathbb{E}[Z^{2k+1}] = 0$$

and

$$\mathbb{E}[Z^{2k}] = (2k-1)(2k-3)\ldots.1 = (2k)!/(2^k k!).$$

For $k = 1$ we have $\mathbb{E}(Z^2) = 1$ and hence $V(Z) = 1 - 0 = 1$.

Consequently for $X \stackrel{d}{=} \mathbb{N}(\mu, \sigma^2)$

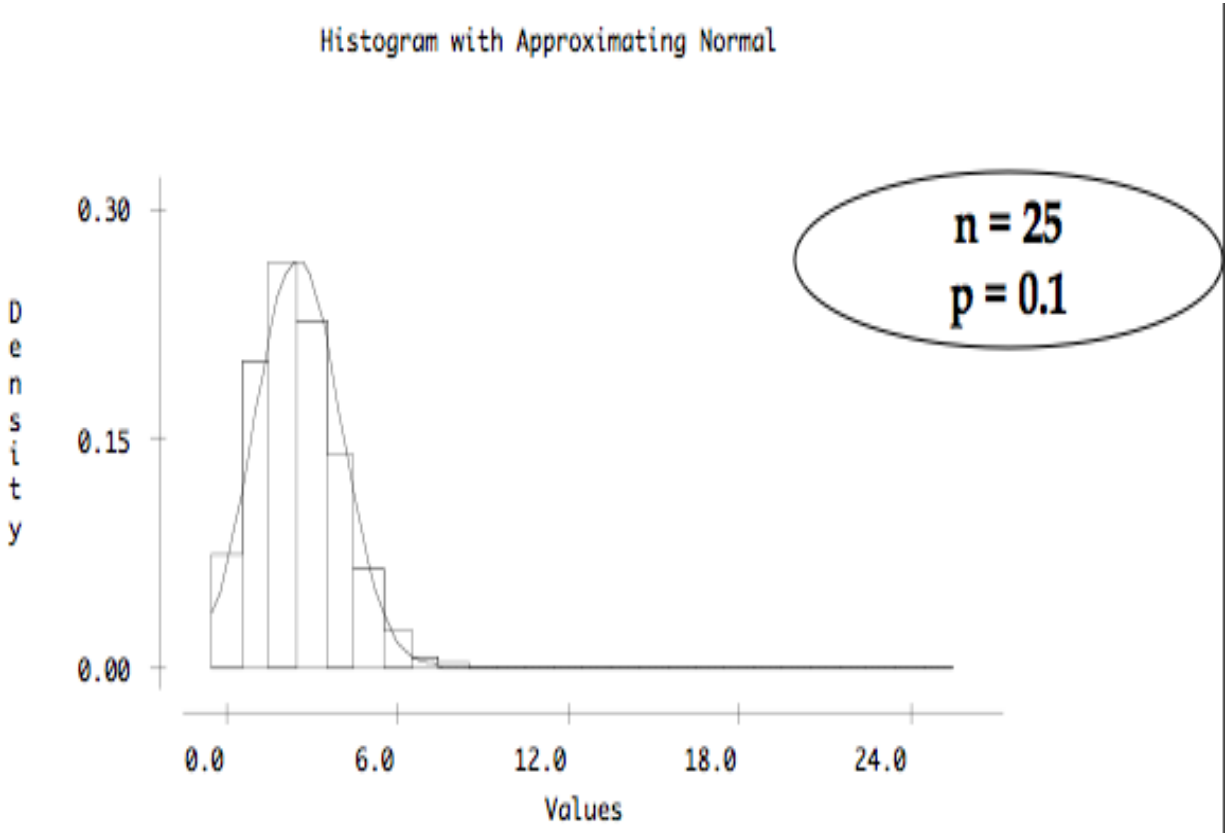
$$\mathbb{E}(X) = \mu + \sigma \times 0 = \mu \text{ and } V(X) = \sigma^2 \times 1 = \sigma^2.$$

So the parameters for the normal distribution are in fact equal to its mean μ and variance σ^2 (justifying our choice of notation). We say that the normal distribution is *parameterised* by its mean and variance.

Binomial with large n

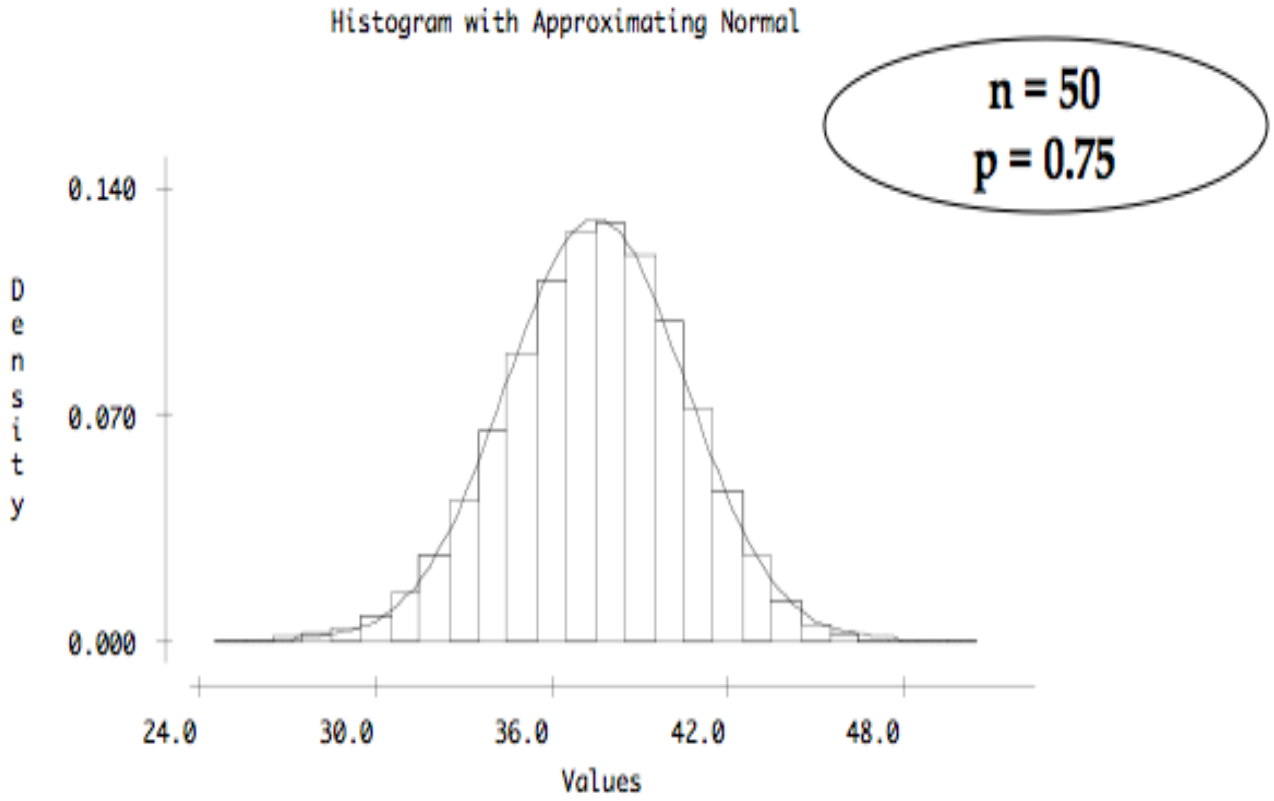
- We can't find probabilities because they are individually too small, and large "n choose k", for $n > 100$.
- If p is small, we can use Poisson to approximate
- If p is close to 1?
- If p is away from 0 and 1, we can use normal to approximate.

Normal approximation to Binomial



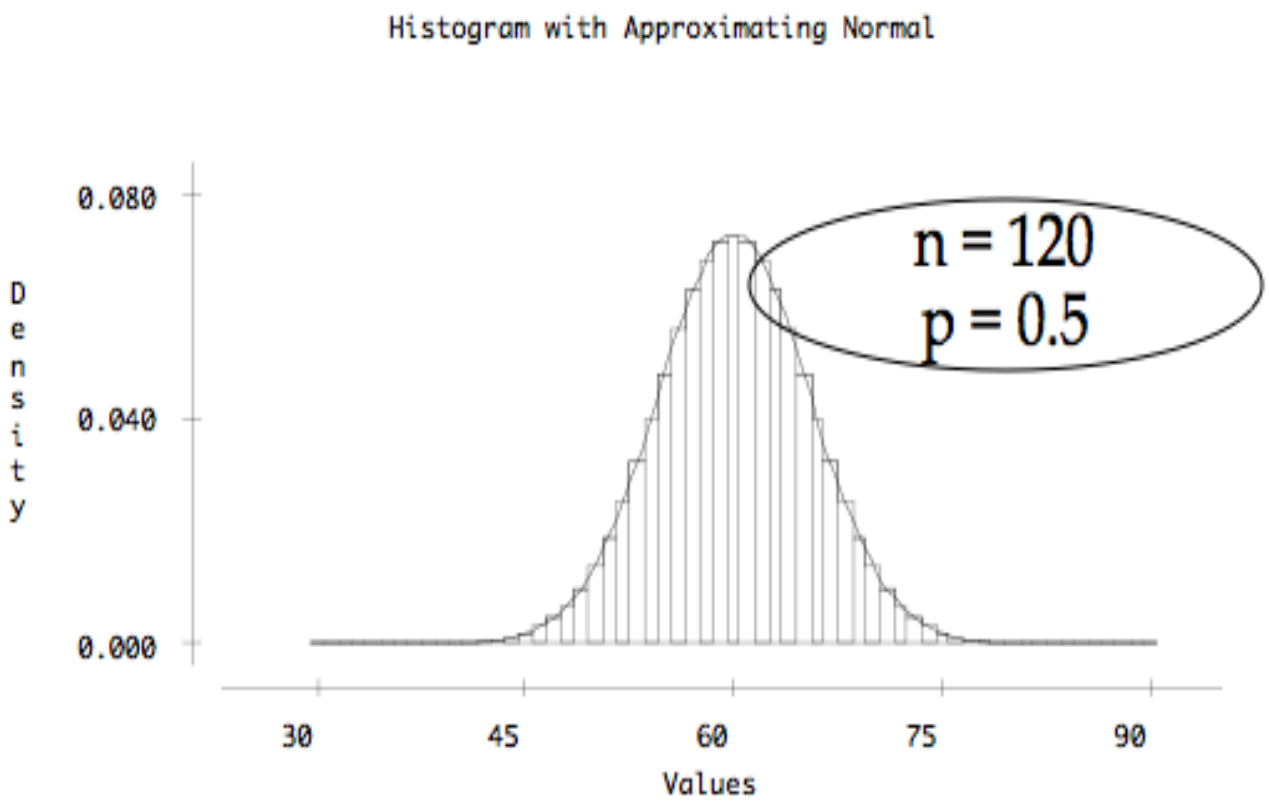
For moderate n, and p close to 0 or 1, the distribution of X is skew

Normal approximation to Binomial



For larger n, and p close to 0.5, dist'bn close to normal.

Normal approximation to Binomial



Probs can be approx. computed as areas under normal curve with $\mu = np$ and $\sigma^2 = np(1-p)$

The magic of normal

1. If $X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n,p)$ and n is large, p is not close to 0 or 1, then $X \stackrel{\mathrm{d}}{\approx} N(np, np(1-p))$
2. If $X \stackrel{\mathrm{d}}{=} \mathrm{Pn}(\lambda)$ and λ is large, then $X \stackrel{\mathrm{d}}{\approx} N(\lambda, \lambda)$.
3. if $X \stackrel{\mathrm{d}}{=} \gamma(r,\alpha)$ and r is large, then $X \stackrel{\mathrm{d}}{\approx} N\left(\frac{r}{\alpha}, \frac{r}{\alpha^2}\right)$

Transformations of Random Variables (Ghahramani 6.2)

We are often interested in random variables Y which are some function ψ of another random variable X. In particular we want to write the distribution function of Y in terms of the distribution function of X.

Thus if X has distribution function F_X and $Y = \psi(X)$, we want to write F_Y in terms of F_X and ψ.

Linear functions

In the case that $Y = aX + b$, where $a > 0$, the distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(aX + b \leq y) \\ &= \mathbb{P}\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

Thus if $Y = aX + b$ with $a > 0$ then $F_Y(y) = F_X\left(\frac{y-b}{a}\right)$.

Similarly, if $a < 0$, then $F_Y(y) = 1 - F_X(\frac{y-b}{a})$. It follows that if X is a continuous random variable with pdf f_X, then, for $a \neq 0$, Y has pdf given by

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Monotonic functions

Suppose that $Y = \psi(X)$, where ψ is a continuous function and increasing on S_X. Define the inverse function $\psi^{-1}(y) = \sup\{x : \ \psi(x) \leq y\}$. Then

$$\psi(x) \leq y \Longleftrightarrow x \leq \psi^{-1}(y).$$

Now, for $y \in \mathbb{R}$, we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(\psi(X) \leq y) \\ &= \mathbb{P}(X \leq \psi^{-1}(y)) \\ &= F_X(\psi^{-1}(y)). \end{aligned}$$

Example

Let X be a continuous random variable. Then F_X is an increasing function which maps S_X to $S_Y \equiv [0,1]$. Let $\psi = F_X$ and so $Y = F_X(X)$. Then, the distribution function of Y is given by

$$F_Y(y) = \begin{cases} F_X(F_X^{-1}(y)) = y & \text{if } 0 < y < 1, \\ 0 & \text{if } y \leq 0, \\ 1 & \text{if } y \geq 1. \end{cases}$$

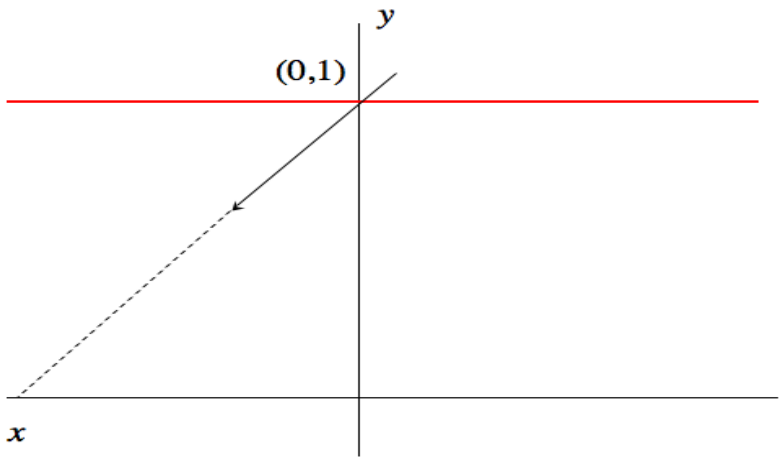
We end up with $Y = F_X(X) \stackrel{d}{=} \mathrm{R}(0,1)$.

Pseudorandom numbers

Generate pseudorandom numbers from $U(0,1)$ according to some algorithm convert to pseudorandom numbers of F as long as F^{-1} can be calculated easily.

Example

A spinner is mounted at the point $(0, 1)$. Let Θ be the smallest angle between the y-axis and the spinner:



Assume that Θ has a uniform distribution on the interval $(-\pi/2, \pi/2)$

1. Find the cdf of Θ.
2. $X = \tan\Theta$, find the pdf of X.

Sol:

The distribution function of X is given by

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad (-\infty < x < \infty)$$

This distribution is endowed with many remarkable properties. It is a special case of the Cauchy distribution.

Definition

If X has pdf

$$f_X(x) = \frac{1}{\pi} \frac{a}{a^2 + (x-m)^2} \ (-\infty < x < \infty),$$

then we say X has a *Cauchy distribution* with parameters m and a, and we write $X \stackrel{d}{=} \mathrm{C}(m, a)$.

Square functions

If $Y = X^2$ then we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X < -\sqrt{y}) \\ F_Y(y) &= F_X(\sqrt{y}) - F_X(-\sqrt{y} - 0) \end{aligned}$$

If X is a continuous random variable, then we have:

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

and therefore, using the chain rule for differentiation:

$$f_Y(y) = \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})]$$

If $X \stackrel{d}{=} N(0,1)$ and $Y = X^2$, then from above

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\sqrt{y})^2} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(-\sqrt{y})^2}\right] \\ &= \left(\frac{1}{2}\right)^{1/2}\frac{1}{\sqrt{\pi}}y^{\frac{1}{2}-1}e^{-\frac{y}{2}}, \qquad y > 0 \\ &= \frac{\alpha^r e^{-\alpha y}y^{r-1}}{\Gamma(r)} \end{aligned}$$

with $\alpha = r = 1/2$. We see that $Y \stackrel{d}{=} \gamma(1/2, 1/2)$.

We have used the fact that:

$$\begin{aligned}
\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx \\
&= \sqrt{2}\int_0^\infty e^{-\frac{z^2}{2}} dz \quad (\text{put } z = \sqrt{2x}) \\
&= \sqrt{2}\ \ \frac{1}{2}\ \ \int_{-\infty}^\infty e^{-\frac{z^2}{2}} dz \\
&= \sqrt{2}\ \ \frac{1}{2}\ \ \sqrt{2\pi}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \sqrt{\pi}.
\end{aligned}$$

Example

X is a rv with pmf

x	-1	0	1	2
$p(x)$	$1/8$	$1/4$	$1/2$	$1/8$

Compute the pmf of $Y = X^2$.

Bivariate Random Variables (Ghahramani Ch 8)

- Ecological studies: counts of several species are made (as r.v.s). One species is often the prey of another; the number of predators will be related to the number of prey
- Atmospheric turbulence: wind velocity is measured in three components x, y, z
- Classification of transmitted and received signals: numbers of high, medium and low quality (height,weight), (age,blood pressure), ...

A *bivariate random variable* is a *function* which maps Ω into $\mathbb{R}^2$, the real plane. It can be thought of as a real-valued number pair $(X(\omega), Y(\omega))$, the value of which depends on the outcome ω of a random experiment.

The *set of possible values* of (X, Y) is given by:

$$S_{(X,Y)} = \{(x, y) : (X(\omega), Y(\omega)) = (x, y) \text{ for some } \omega \in \Omega\} \subset \mathbb{R}^2.$$

Example

Consider the random experiment of tossing a coin three times and observing the sequence of results.

If X = number of heads in the first two tosses, and Y = number of heads in the last two tosses; then Ω is mapped into $\mathbb{R}^2$.

We often think of a univariate random variable as a "random point" on the real line. Analogously we can think of a bivariate random variable as a "random point" on the plane.

There is little difficulty in extending the definitions of distribution function, pmf and pdf to the case of bivariate variables as we shall see.

Distribution Function of a Bivariate Random Variable

The cumulative distribution function of (X, Y) is defined by:

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

This *bivariate distribution function* has properties analogous to those of the univariate distribution function. The most important of these is the following:

$$\begin{aligned}
&\mathbb{P}(a < X \leq b, c < Y \leq d) \\
&\quad = \mathbb{P}(\{a < X \leq b\} \cap \{Y \leq d\}) - \mathbb{P}(\{a < X \leq b\} \cap \{Y \leq c\}) \\
&\quad = \mathbb{P}(\{X \leq b\} \cap \{Y \leq d\}) - \mathbb{P}(\{X \leq a\} \cap \{Y \leq d\}) \\
&\qquad\quad -\mathbb{P}(\{X \leq b\} \cap \{Y \leq c\}) + \mathbb{P}(\{X \leq a\} \cap \{Y \leq c\}) \\
&\quad = F_{(X,Y)}(b, d) - F_{(X,Y)}(a, d) - F_{(X,Y)}(b, c) + F_{(X,Y)}(a, c).
\end{aligned}$$

From the definition of the distribution function:

$$F_{(X,Y)}(x,\infty) = F_X(x) \quad \text{and} \quad F_{(X,Y)}(\infty,y) = F_Y(y)$$

so that the distributions of X and Y can be obtained from the distribution of (X,Y).

In general the converse is not true, because the distributions of X and Y are not sufficient to specify the distribution of (X,Y). We also need to know something about the relationship between X and Y.

However in principle we can imagine choosing a random point by choosing its X and Y coordinates completely "independently". Then the individual distributions of X and Y *should* determine the joint distribution.

If X and Y are "independent" then the events $\{X \leq x\}$ and $\{Y \leq y\}$ should also be "independent" for all x and y and so

$$\begin{aligned} F_{(X,Y)}(x,y) &= \mathbb{P}(X \leq x, Y \leq y) \\ &= \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y) \\ &= F_X(x)F_Y(y). \end{aligned}$$

In this special case the two univariate distributions do determine the joint distribution. Factorisation of the bivariate distribution function in this way can be taken as one definition of the independence of random variables, which we will define more carefully later.

Joint and marginal pmf's (Ghahramani 8.1)

If $S_{(X,Y)}$ is countable then we refer to (X, Y) as *bivariate discrete random variable* and define the *joint pmf* in the obvious way:

$$\begin{aligned} p_{(X,Y)}(x, y) &= \mathbb{P}(\{X = x\} \cap \{Y = y\}) \\ &= \mathbb{P}(X = x, Y = y). \end{aligned}$$

The pmf is such that

$$p_{(X,Y)}(x,y) \geq 0 \quad \text{and} \quad \sum_x \sum_y p_{(X,Y)}(x,y) = 1.$$

Just as we did for "univariate" or one dimensional discrete random variables, we can again interpret this *pmf* as assigning discrete probability masses to particular points in the plane.

Marginal pmf's

The formula for the marginal distributions follows from the observation that the events $\{Y = y\}$ are disjoint and exhaustive as y ranges over the set $S_Y = \{y : (x, y) \in S_{(X,Y)}\}$. Then,

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

that is

$$p_X(x) = \sum_{y \in S_Y} p_{(X,Y)}(x, y)$$

and similarly

$$p_Y(y) = \sum_{x \in S_X} p_{(X,Y)}(x, y).$$

Coin example

If in the coin tossing example the coin is fair then the pmf is given by

$p_{(X,Y)}(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	$\frac{1}{8}$	$\frac{1}{8}$	0
$y=1$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
$y=2$	0	$\frac{1}{8}$	$\frac{1}{8}$

We can also find the marginal pmf's for X and Y from here.

CDF for discrete rv's

For discrete (X, Y),

$$\begin{aligned} F_{(X,Y)}(x,y) &= \mathbb{P}(X \le x, Y \le y) \\ &= \sum_{u \le x \text{ and } v \le y} \mathbb{P}(X = u, Y = v) \\ &= \sum_{u \le x \text{ and } v \le y} p_{(X,Y)}(u,v). \end{aligned}$$

However, obtaining the pmf from cdf for a pair of rv's requires more care!

Coin example

If in the coin tossing example the coin is fair then the pmf is given by

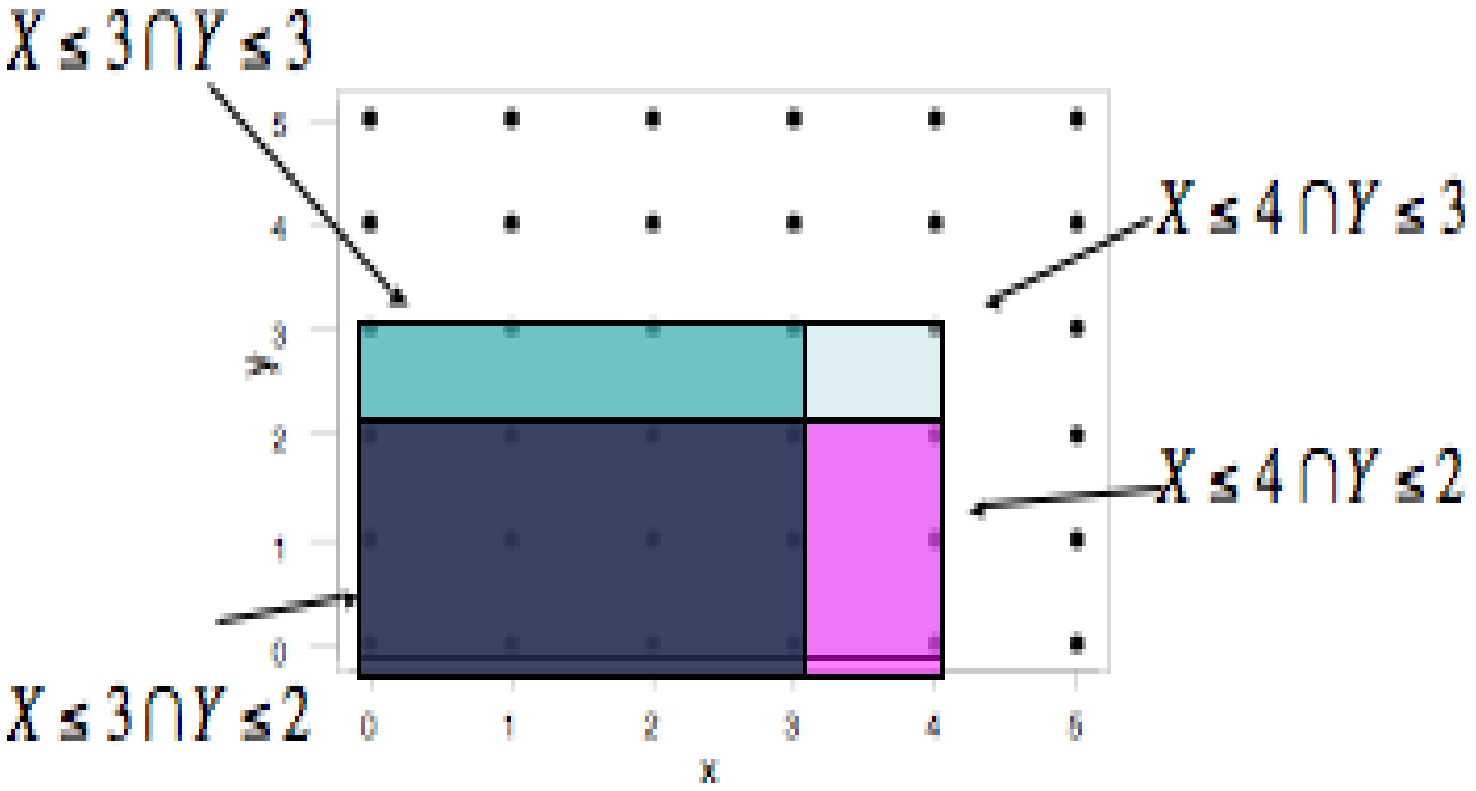
$p_{(X,Y)}(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	$\frac{1}{8}$	$\frac{1}{8}$	0
$y=1$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
$y=2$	0	$\frac{1}{8}$	$\frac{1}{8}$

Example

Suppose the cumulative distribution function of the pair of rv's (X, Y) (each with values $0, 1, 2, \dots$) is given by the table:

y \ x	1	2	3	4
1	0	0.1	0.2	0.3
2	0.1	0.2	0.35	0.6
3	0.2	0.3	0.6	0.95
4	0.25	0.35	0.65	1

1. Find $\mathbb{P}(X = 4, Y = 3)$.
2. Find the complete probability mass function.


$X \leq 3 \cap Y \leq 3$
$X \leq 4 \cap Y \leq 3$
$X \leq 4 \cap Y \leq 2$
$X \leq 3 \cap Y \leq 2$
y
x
0
1
2
3
4
5
6

[Slide 286]

Pmf from cdf

For a pair of rv's (X, Y) with values in $0, 1, 2, \ldots,$

$$\begin{aligned}
&\mathbb{P}(X = x, Y = y) \\
&= \mathbb{P}(X = x, Y \le y) - \mathbb{P}(X = x, Y \le y - 1) \\
&= \mathbb{P}(X \le x, Y \le y) - \mathbb{P}(X \le x - 1, Y \le y) \\
&\quad - \mathbb{P}(X \le x, Y \le y - 1) + \mathbb{P}(X \le x - 1, Y \le y - 1) \\
&= F_{(X,Y)}(x, y) - F_{(X,Y)}(x - 1, y) \\
&\quad - F_{(X,Y)}(x, y - 1) + F_{(X,Y)}(x - 1, y - 1).
\end{aligned}$$

pmf

y \ x	1	2	3	4
1	0	0.1	0.1	0.1
2	0.1	0	0.05	0.15
3	0.1	0	0.15	0.1
4	0.05	0	0	0

Probabilities from cdf - in general

For a pair of rv's (X, Y) and numbers a, b, c, d:

$$\begin{aligned}
&\mathbb{P}(\{a < X \leq b\} \cap \{c < Y \leq d\}) \\
&= F_{(X,Y)}(b, d) - F_{(X,Y)}(a, d) \\
&\quad -F_{(X,Y)}(b, c) + F_{(X,Y)}(a, c),
\end{aligned}$$

as before.

Example

Suppose there are 4 factories which are ranked 1 to 4 in quality of output. If a random sample of size 2 is taken, find the pmf and cdf for the pair of random variables (X, Y) if X is the sum of ranks in the sample and Y is the maximum of ranks in the sample.

Joint and marginal pdfs (Ghahramani 8.1)

What is the analogy for cts rv's?

A pair of rv's (X, Y) is said to have a continuous distribution if the cdf $F_{(X,Y)}$ can be written as an integral of a two variable, non-negative function $f_{(X,Y)}$, called the probability density function (pdf):

$$\begin{aligned} F_{(X,Y)}(x,y) &= \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) \\ &= \iint_{\{(u,v):\ u \leq x \text{ and } v \leq y\}} f_{(X,Y)}(u,v)\,du\,dv. \end{aligned}$$

Double integrals

Let $f(u,v) \geq 0$ and

$$V = \int_a^b \int_c^d f(u,v)dvdu = \lim \sum_i f(u_i, v_i)\Delta_i$$

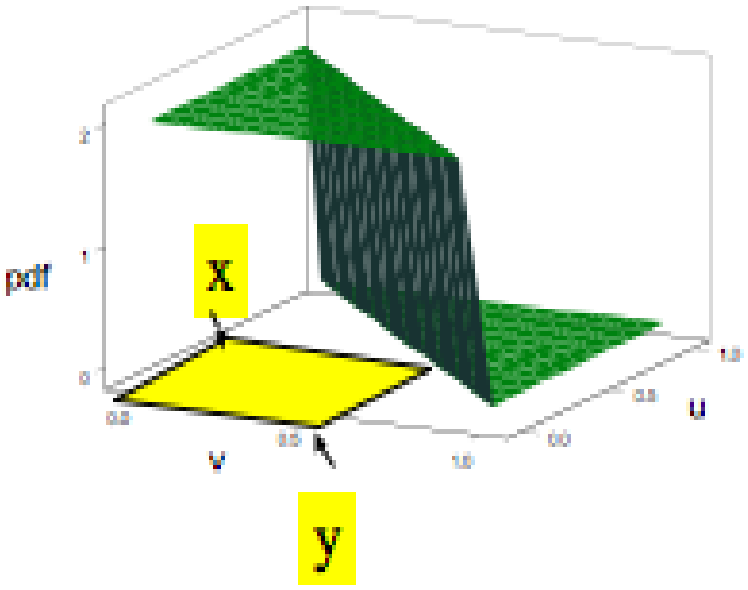
where the limit is over all regular partitions of the rectangle $(a,b) \times (c,d)$ into the small areas Δ_i. If we interpret $f(u,v)$ as a surface above the xy plane then V can be interpreted as the volume below that surface.

Double integrals (2)

Conceptually, the double integral is the volume under the surface given by $f_{(X,Y)}$ in the region

$\{(u, v) : \ u \leq x \text{ and } v \leq y\}$

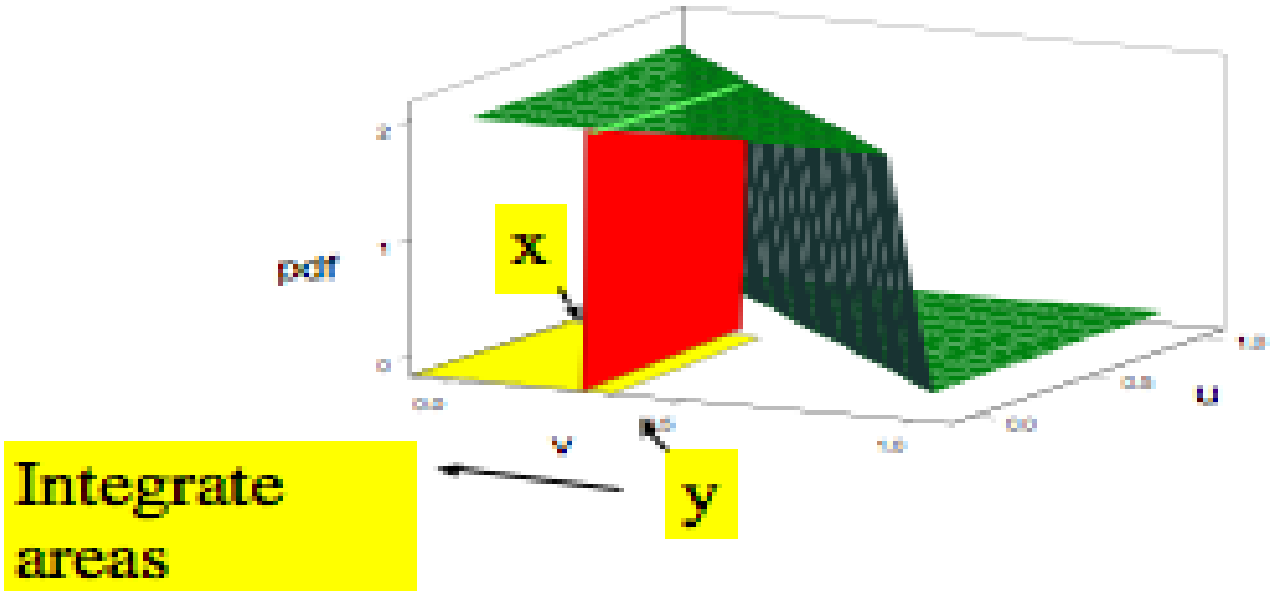
Total volume under green surface equal to 1



Cdf is volume under green surface down to yellow

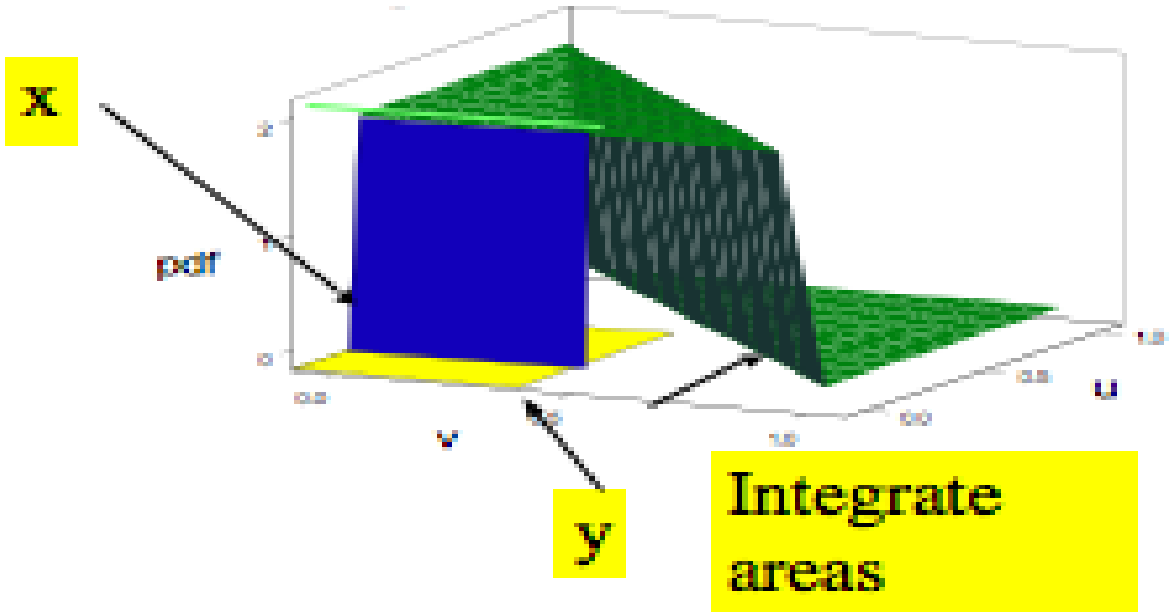
Conceptually, either

Find the area under the green curve illustrated in red, and then integrate the areas over $\{v \leq y\}$:



or

Find the area of the blue rectangle under the green surface, and then integrate this over different $\{u \leq x\}$:



Double integrals (3)

The double integrals can be evaluated by treating them as two iterated single integrals:

$$\int_a^b \int_c^d f(u,v)dvdu = \int_a^b \Big(\int_c^d f(u,v)dv \Big) du.$$

In the inner integral here we think of u as fixed. Consider intersecting the hill by the plane through this fixed u - the resulting cross-sectional area is given by the inner integral. So the whole iterated integral just amounts to saying that the total volume is the sum of cross-sectional "slices" parallel to the x axis of volume $\Big(\int_c^d f(u,v)dv \Big) du$. You can reverse the order to get the sum of cross-sectional slices parallel to the y axis.

pdf's

- Just as the pdf for a single rv must be non-negative and have 1 as total area under the curve it defines, the pdf for a pair of rv's must be non-negative and have 1 as total volume under the surface that it defines.
- If such a function exists, then it is unique.
- If a *pdf* exists then for "almost all" (x, y) values we also have

$$\frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x, y) = f_{(X,Y)}(x, y).$$

To understand the connection between the distribution function and pdf consider:

$$
\begin{aligned}
f_{(X,Y)}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x,y) \\
&\approx \frac{\partial}{\partial x}\Big(\frac{1}{h}\Big(F(x, y+\frac{h}{2}) - F(x, y-\frac{h}{2})\Big)\Big) \\
&\approx \frac{1}{h}\Big(\frac{F(x+\frac{h}{2}, y+\frac{h}{2}) - F(x+\frac{h}{2}, y-\frac{h}{2})}{h} \\
&\qquad - \frac{F(x-\frac{h}{2}, y+\frac{h}{2}) - F(x-\frac{h}{2}, y-\frac{h}{2})}{h}\Big) \\
&\approx \frac{1}{h^2}\mathbb{P}(x-\frac{h}{2} < X \leq x+\frac{h}{2}, y-\frac{h}{2} < Y \leq y+\frac{h}{2}).
\end{aligned}
$$

So

$$\mathbb{P}((X,Y) \in \text{small rectangle area } h^2 \text{ near}(x,y)) \approx h^2 f_{(X,Y)}(x,y)$$

justifying our interpretation of the pdf as measuring the probability density around the point (x, y).

For continuous univariate random variables, probabilities are assigned to intervals using the area under the pdf curve $f_X(x)$. For continuous bivariate random variables probabilities are assigned to regions in the plane using volumes under the pdf "hill" or surface $f_{(X,Y)}(x,y)$.

The probability that the random point (X, Y) lies in any rectangle in the plane is given by the volume below the two dimensional surface represented by $f_{(X,Y)}(x, y)$ over that rectangle:

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{(X,Y)}(x, y) dy dx$$

As we did with discrete bivariate random variables, we have

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

and similarly

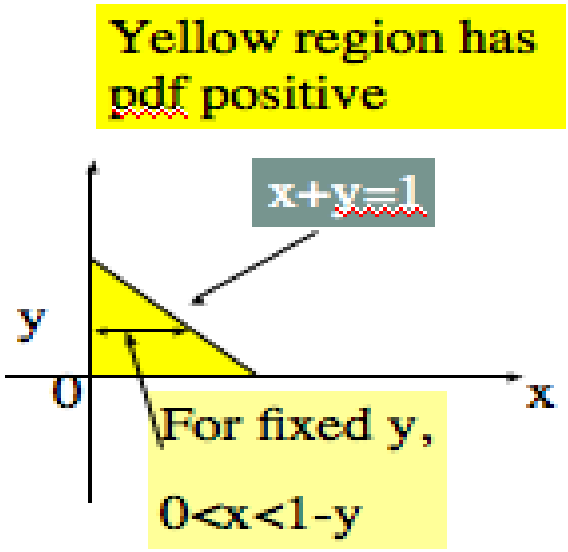
$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx.$$

pdf example

Suppose the pdf of the bivariate rv (X, Y) is

$$f_{(X,Y)}(x,y) = \begin{cases} cxy, & \text{if } 0 < x < 1,\ 0 < y < 1,\ 0 < x + y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find c and $\mathbb{P}(\{X \leq 0.5\} \cap \{Y \leq 0.7\})$.



A surprise?

1. $f_{(X,Y)}(x,y) = 2x + 2y - 4xy$ for $0 < x < 1$, $0 < y < 1$, find the marginal pdf's.

2. $f_{(X,Y)}(x,y) = 2 - 2x - 2y + 4xy$ for $0 < x < 1$, $0 < y < 1$, find the marginal pdf's.

3. What can we see from here?

Conditional pmfs (Ghahramani 8.3)

It seems natural to ask about the distribution of one component of our bivariate distribution (say X) when we have information about the value of the other component (Y).

By analogy with the way that we defined conditional probability for events, this is called a *conditional distribution* and for discrete random variables the expression for the conditional distribution follows directly from the conditional probability formula.

We define the conditional distribution of X given Y as:

$$\begin{aligned} p_{X|Y}(x|y) &= \mathbb{P}(X = x|Y = y) \\ &= \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \\ &= \frac{p_{(X,Y)}(x, y)}{p_Y(y)} \end{aligned}$$

Note in this formula we think of x as a variable and of y as fixed with $y \in S_Y$ (as previously defined) and $p_Y(y) \neq 0$.

Conditional pdfs (Ghahramani 8.3)

The pdf of X *conditional on* $Y = y$ is given by

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

This can derived by letting $\delta y \to 0$ in the equation

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

Example

A particular type of rock is analysed. Let X and Y denote the proportions of minerals A and B respectively found in a sample of the rock. Assume that the pdf of (X, Y) is given by

$$f_{(X,Y)}(x, y) = 2 \quad (x + y \leq 1, x \geq 0, y \geq 0).$$

The pdf of X is given by

Similarly, $f_Y(y) = 2 - 2y$ $(0 < y < 1)$. Thus, $X \stackrel{d}{=} Y$. This is a consequence of the fact that $f_{(X,Y)}(x, y)$ is symmetric in x and y.

The pdf of X conditional on $Y = y$ is given by

That is, $(X|Y = y) \stackrel{d}{=} \mathrm{R}(0, 1 - y)$.

The distribution of X given $Y = y$ is different from the unconditional distribution of X, hence X and Y are *not* independent.

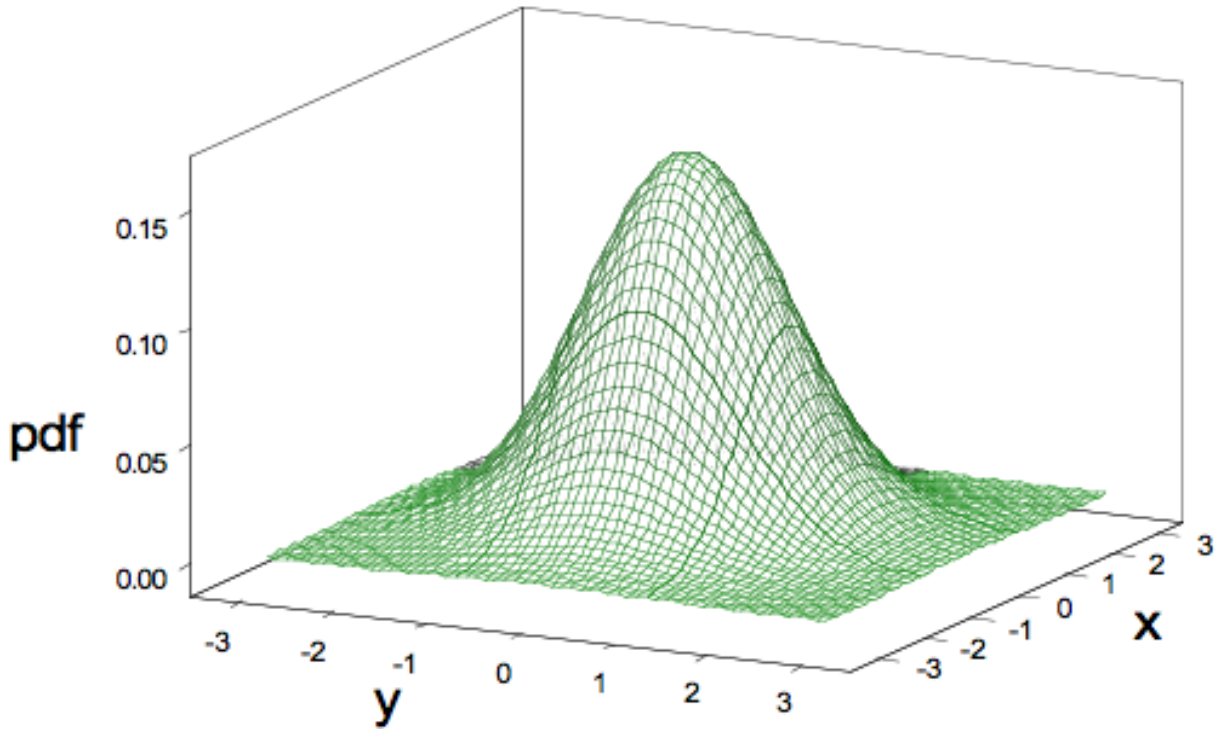
Bivariate normal distribution (Ghahramani 10.5)

If the pdf of (X, Y) is given by

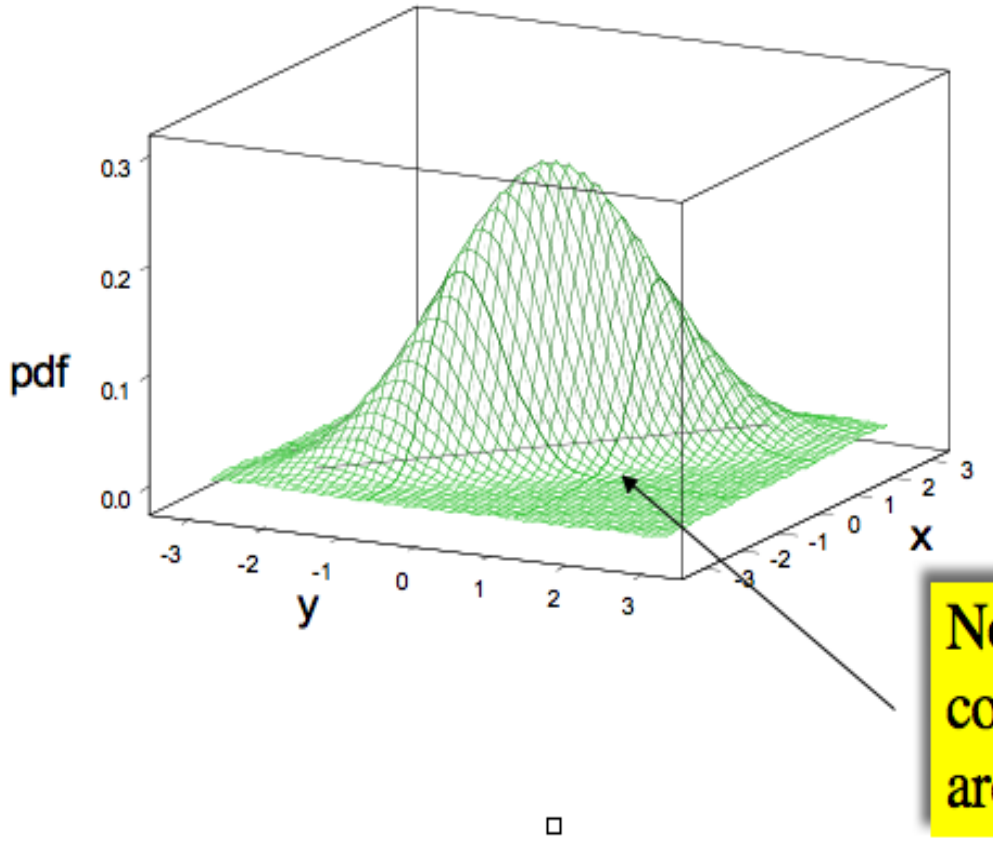
$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \right)$$

where $\rho \in [-1, 1]$, then we say that (X, Y) has the *standard bivariate normal distribution with parameter* ρ, and we write $(X, Y) \stackrel{d}{=} N_2(\rho)$.

$$\rho = 0$$

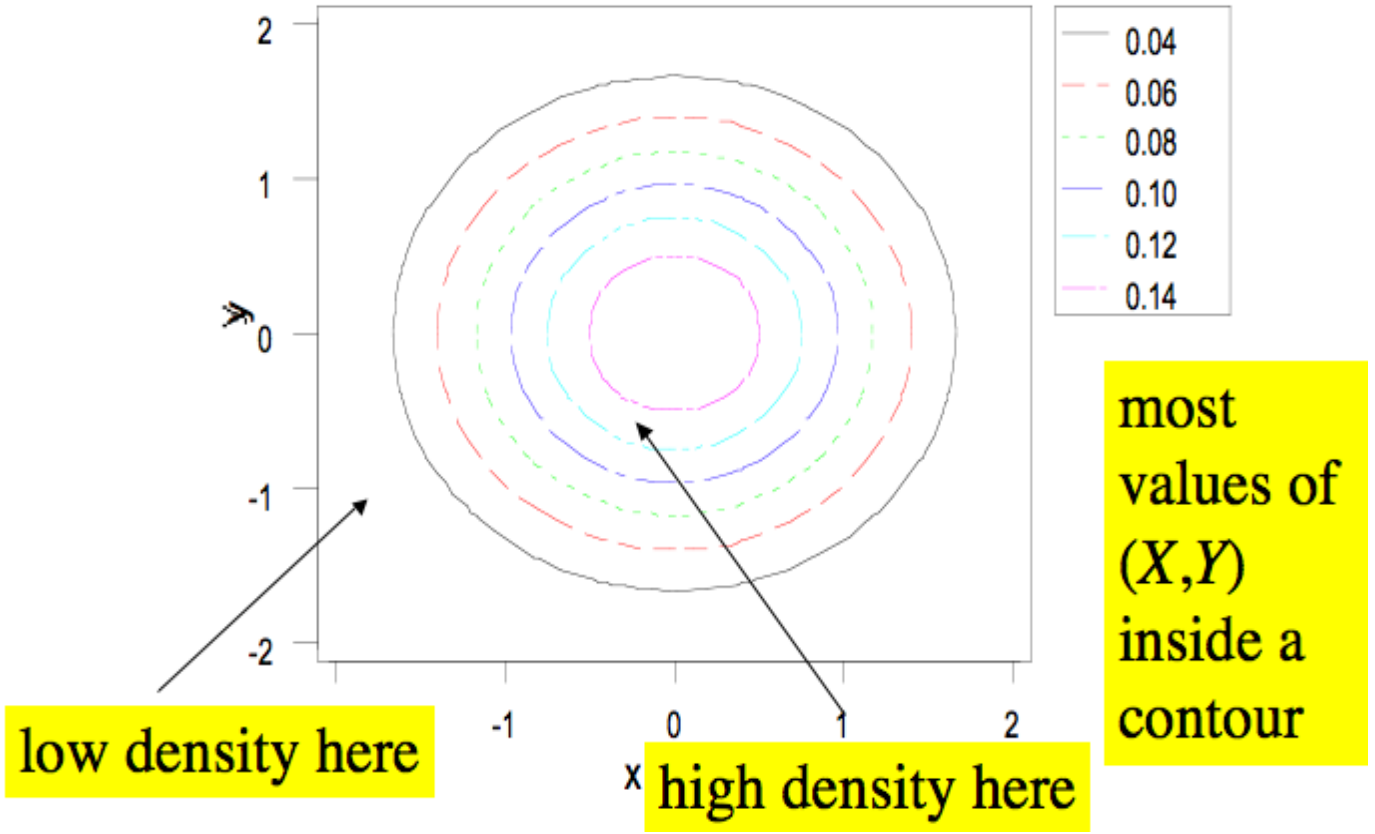


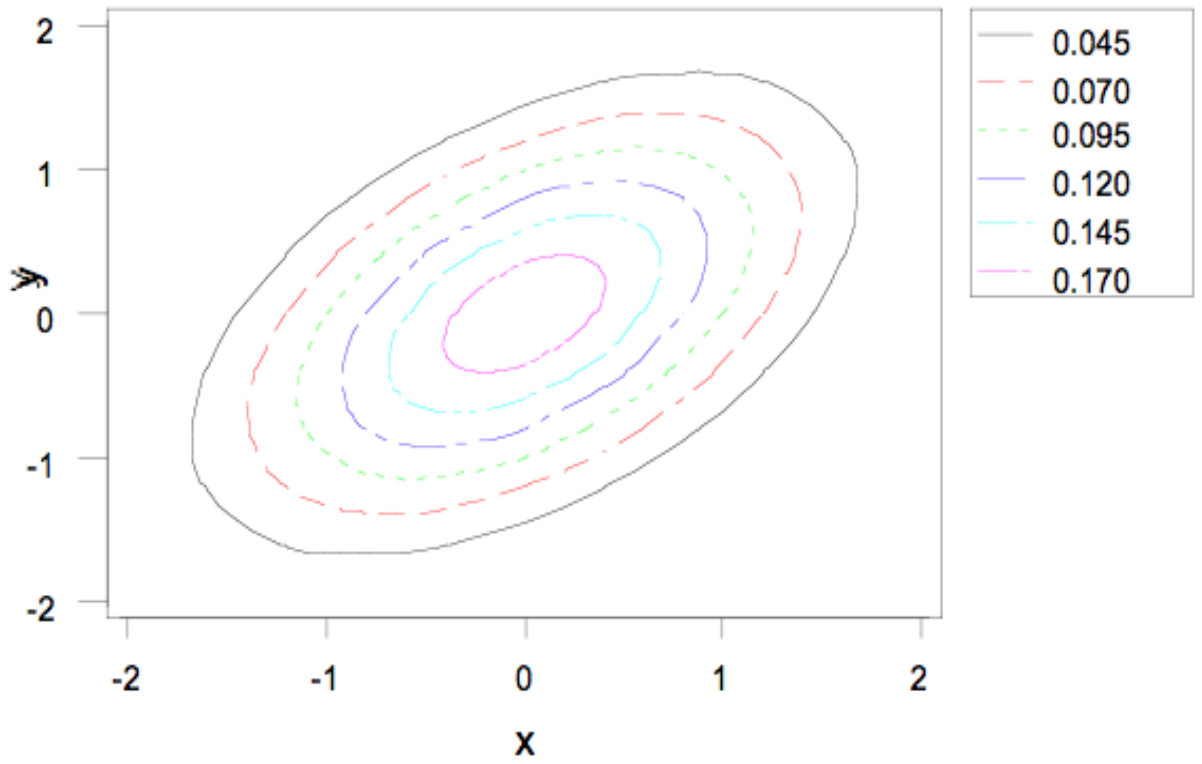
$$\rho = 0.8$$



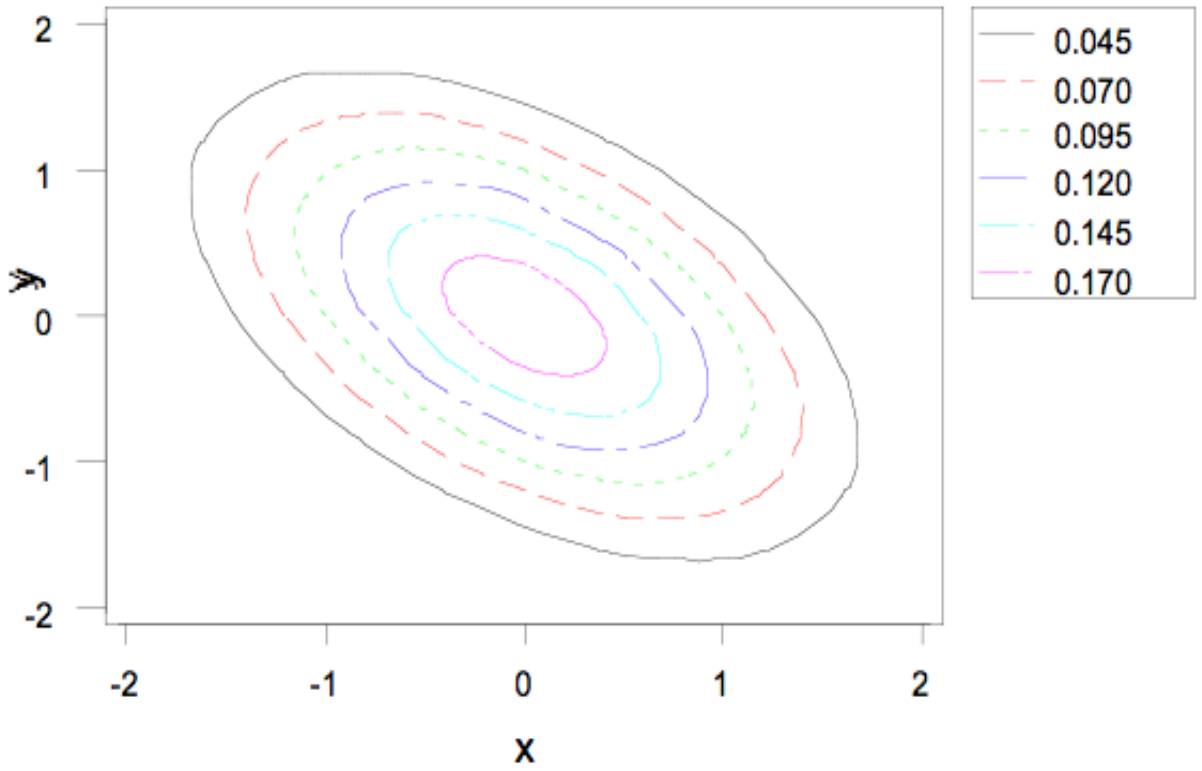
Note: pdf concentrated around line

$$\rho = 0$$

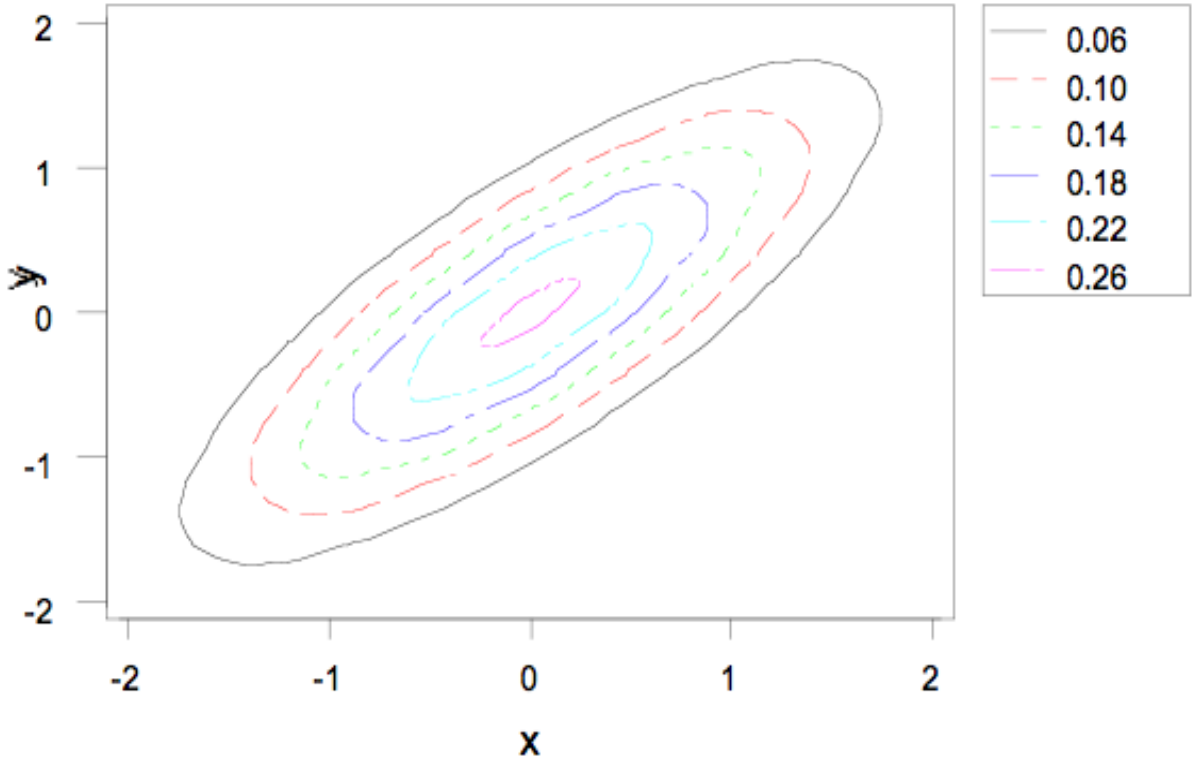

2
1
0
-1
-2
y
-1
0
1
2
x
0.04
0.06
0.08
0.10
0.12
0.14
low density here
high density here
most values of (X,Y) inside a contour



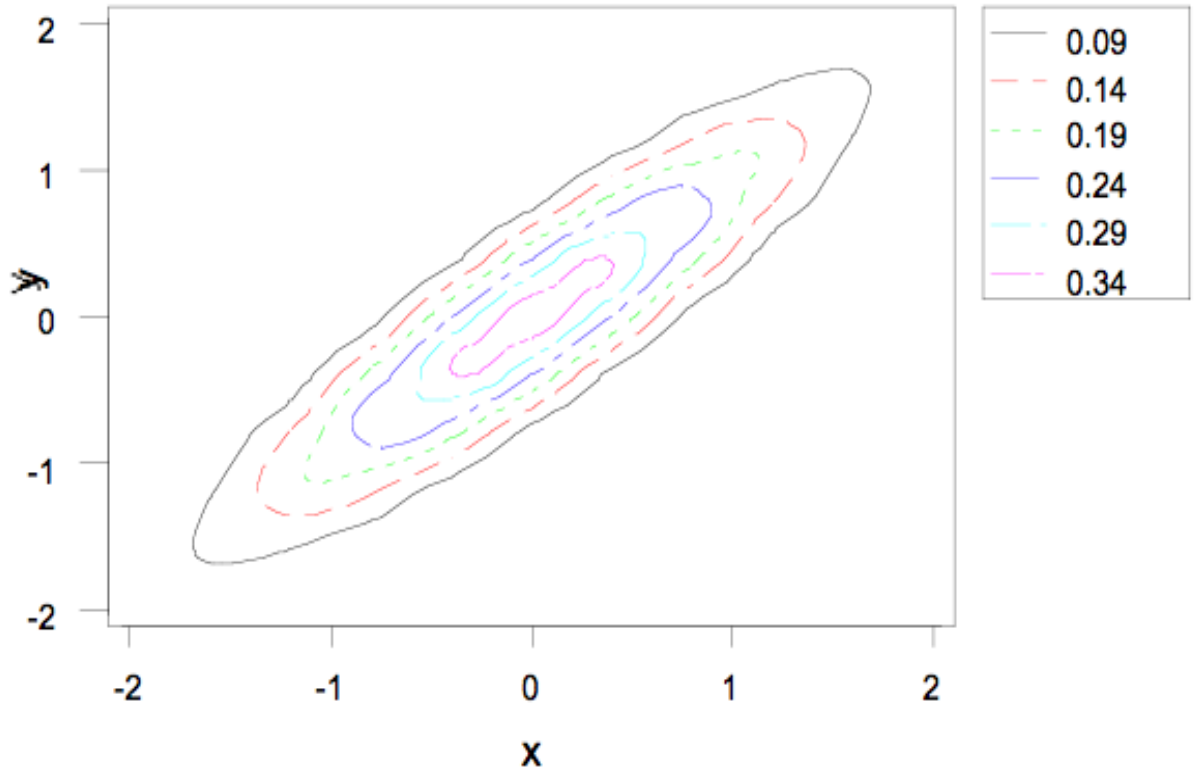
$\rho = 0.5$



$\rho = -0.5$



$\rho = 0.8$



$\rho = 0.9$

The contours of $f_{(X,Y)}(x, y)$ are ellipses with axes inclined at an angle of $\pi/4$ to the x- and y-axes.

If $\rho > 0$ then the major axis lies along $y = x$ and the minor axis lies along $y = -x$. This means that X and Y tend to be large together, and small together. In this case we say that X and Y are *positively-related*.

On the other hand, if $\rho < 0$ then the axes are reversed and X tends to be large when Y is small and vice versa. In this case X and Y are said to be *negatively-related*.

As $|\rho|$ gets close to one, the ellipses become "skinnier" in whatever orientation that they have and the relationship between X and Y becomes stronger.

We see that the parameter ρ tells us a good deal about the relationship between X and Y.

The pdf of Y is given by

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{(x^2-2\rho xy+y^2)}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{2\pi} e^{-\frac{1}{2}y^2} \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x^2-2\rho xy+\rho^2y^2)}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du \qquad \left[\text{where } u = \frac{x-\rho y}{\sqrt{1-\rho^2}}\right] \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}
\end{aligned}
$$

Therefore $Y \stackrel{d}{=} N(0,1)$. Similarly, we can show that $X \stackrel{d}{=} N(0,1)$.

Note that $X \stackrel{d}{=} N(0,1)$ and $Y \stackrel{d}{=} N(0,1)$ does not imply that $(X,Y) \stackrel{d}{=} N_2(\rho)$.

For example if $X \stackrel{d}{=} N(0,1)$ and we define

$$Y = \begin{cases} X & \text{with probability } 1/2 \\ -X & \text{with probability } 1/2 \end{cases}$$

then $Y \stackrel{d}{=} N(0,1)$ also, but (X,Y) is not bivariate normal — since the bivariate pdf is non-zero only on the lines $y = \pm x$.

The conditional pdf of X given $Y = y$ is given by

$$
\begin{aligned}
& f_{X|Y}(x|y) \\
&= \frac{f_{(X,Y)}(x,y)}{f_Y(y)} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + \rho^2 y^2)\Big) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x-\rho y)^2\Big)
\end{aligned}
$$

hence $(X|Y = y) \stackrel{d}{=} N(\rho y, 1-\rho^2)$. It follows that $\mathbb{E}[X|Y = y] = \rho y$.

Example

Suppose that $(X, Y) \stackrel{d}{=} N_2(\rho = 0.5)$. Find $\mathbb{P}(X > 1)$ and $\mathbb{P}(X > 1|Y = 1)$.

Solution

Since $X \stackrel{d}{=} N(0, 1)$, we have $\mathbb{P}(X > 1) = 0.159$ from tables.

For $\mathbb{P}(X > 1|Y = 1)$,

We see that $\mathbb{P}(X > 1|Y = 1)$ is greater than $\mathbb{P}(X > 1)$. Since $\rho > 0$, X and Y are positively related. It follows that we expect Y to be larger (greater than its mean) when X is larger.

If

$$\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) \stackrel{d}{=} N_2(\rho),$$

then (X, Y) has a bivariate normal distribution with parameters μ_X, μ_Y, σ_X, σ_Y and ρ, and we write

$$(X, Y) \stackrel{d}{=} N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).$$

An alternative notation is

$$\begin{bmatrix} X \\ Y \end{bmatrix} \stackrel{d}{=} N_2 \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right).$$

Results for the general bivariate normal distribution can be derived from those for the standard case (try this yourself)

$$\begin{aligned} X &\stackrel{d}{=} N(\mu_X, \sigma_X^2) \\ Y &\stackrel{d}{=} N(\mu_Y, \sigma_Y^2) \\ (X|Y=y) &\stackrel{d}{=} N\left(\mu_X + \rho\sigma_X \frac{(y-\mu_Y)}{\sigma_Y},\ \sigma_X^2(1-\rho^2)\right). \end{aligned}$$

Example

Suppose that $(X, Y) \stackrel{d}{=} N_2(50, 50; 10^2, 5^2; -0.7)$. Find $\mathbb{P}(X > 55)$ and $\mathbb{P}(X > 55 | Y = 60)$.

Solution

$X \stackrel{d}{=} N(50, 10^2)$, so

To find $\mathbb{P}(X > 55|Y = 60)$,

In the above example, $\mathbb{P}(X > 55|Y = 60)$ is less than $\mathbb{P}(X > 55)$ because, $\rho < 0$ and so when Y is large we expect X to be small.

In this case, $|\rho| = 0.7$ and so the relationship between X and Y is quite strong.

Example

Suppose the study scores (X, Y) of a random student in VCE English and Mathematical Methods can be considered as bivariate normal with $\rho = 0.7$, μ's $= 30$ and σ's $= 7$. What is the conditional probability that the student gets more than 30 in Mathematical Methods given the students study score in English is (a) 30 (b) 35 (c) 40?

Sol

$$
\begin{aligned}
& \mathbb{P}(Y > 30 | X = x) \\
= \quad & \mathbb{P}\left(\frac{Y - 30 - 0.7(x-30)}{7\sqrt{1-0.7^2}} > \frac{30 - 30 - 0.7(x-30)}{7\sqrt{1-0.7^2}} \middle| X = x \right) \\
= \quad & \mathbb{P}\left(Z > \frac{-0.7(x-30)}{7\sqrt{1-0.7^2}} \right) = \mathbb{P}\left(Z < \frac{0.7(x-30)}{7\sqrt{1-0.7^2}} \right)
\end{aligned}
$$

- $\mathbb{P}(Y > 30 | X = 30) = \frac{1}{2}$
- $\mathbb{P}(Y > 30 | X = 35) = 0.758080$
- $\mathbb{P}(Y > 30 | X = 40) = 0.919285$

Independence of random variables (Ghahramani 8.2)

- Independence of "events".
- We now extend the notion of independence to random variables: two random variables X and Y are *independent* if

$$\mathbb{P}(X \in M,\ Y \in N) = \mathbb{P}(X \in M)\mathbb{P}(Y \in N)$$

  for any (appropriate) subsets $M, N \in \mathbb{R}$.
- There are more equivalent but easy ways: to prove these results today rigorously, measure theory is needed (MSc study)

Independence by cdf

X, Y independent rv's if and only if

$$F_{(X,Y)}(x,y) = F_X(x)F_Y(y) \text{ for all } x \text{ and } y$$

For discrete rv's

Often more convenient to think of pmf.

Rv's are independent if and only if the joint pmf is the product of the marginals:

$$X,\, Y \text{ independent}$$

$$\Updownarrow$$

$$\left\{ \begin{array}{ll} p_{(X,Y)}(x,y) = p_X(x)p_Y(y) & \forall x, y \\ p_X(x) = p_{X\,|\,Y}(x\,|\,Y=y), p_Y(y) = p_{Y\,|\,X}(y\,|\,X=x) & \forall x, y \end{array} \right\}$$

For continuous rv's

Often more convenient to think of pdf.

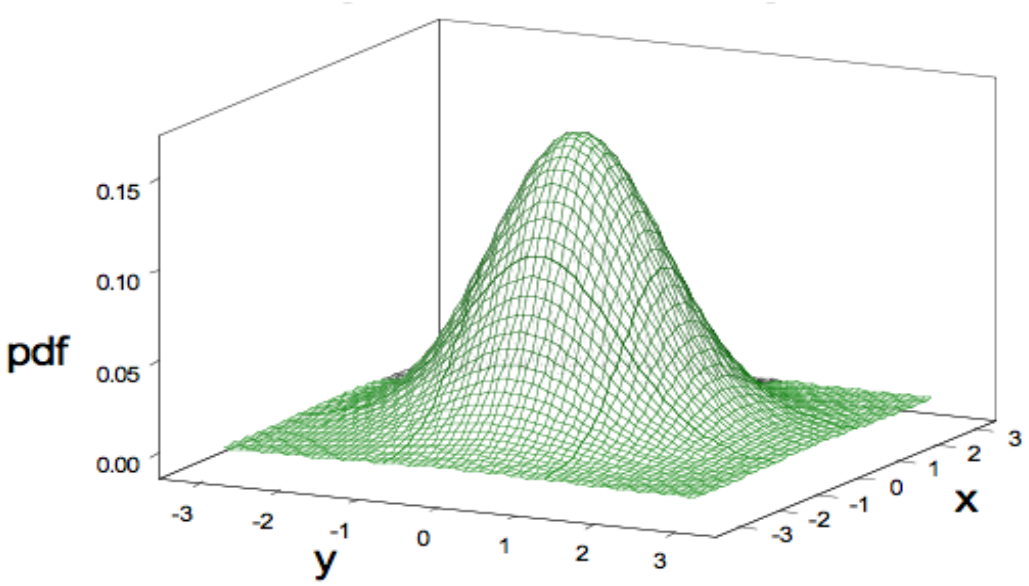
Rv's are independent if and only if the pdf is the product of the marginals:

$$X,\, Y \text{ independent}$$

$$\Updownarrow$$

$$\left\{ \begin{array}{ll} f_{(X,Y)}(x,y) = f_X(x) f_Y(y) & \forall x, y \\ f_X(x) = f_{X\,|\,Y}(x \,|\, Y = y), f_Y(y) = f_{Y\,|\,X}(y \,|\, X = x) & \forall x, y \end{array} \right\}$$

Pair of indept N(0,1)



Pdf example

Suppose the pdf for the bivariate rv (X, Y) is

$$f_{(X,Y)}(x, y) = \begin{cases} c, & \text{if } 0 < x < 1,\ 0 < y < 1,\ 0 < x + y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Symmetry

If the density is symmetric in x and y (ie the value does not change if x and y are swapped), the marginal pdf's are the same: X and Y are identically distributed

Summary: three criteria for independence

Rv's are independent if and only if

- cdf equals the product of its marginals
- pmf or pdf is the product of its marginals
- conditional pmf or pdf is same as its marginal

Example

Suppose that $(X, Y) \stackrel{d}{=} N_2(0)$, that is, X and Y are independent standard normal. Then (X, Y) defines a point in $\mathbb{R}^2$. What is the distribution of (R, Θ), the polar coordinates of this point? Are the polar coordinates independent?

For $(x, y) \in (-\infty, \infty)^2$, we have

$$\mathbb{P}(X \approx x, Y \approx y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \delta x \delta y.$$

In polar coordinates, $r = \sqrt{x^2 + y^2}$ and so, for $(r, \theta) \in [0, \infty) \times [0, 2\pi)$, we have

Hence, the pdf of (R, Θ) is given by

It follows that

$$f_R(r) = re^{-\frac{1}{2}r^2} \ (r > 0) \quad \text{and} \quad f_\Theta(\theta) = \frac{1}{2\pi} \ (0 < \theta < 2\pi).$$

The transformation described in the above example can be used to derive a proof of the result that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi},$$

which we have used to derive the constant $1/(\sqrt{2\pi})$ in the formula for the pdf of the normal distribution.

Assume that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = c.$$

Then

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,dy = c,$$

and, multiplying them together,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\,dx\,dy = c^2.$$

Now, transforming to polar coordinates,

$$\int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}r^2} r dr d\theta = c^2.$$

This is equivalent to

$$2\pi \int_0^{\infty} e^{-\frac{1}{2}r^2} r dr = c^2$$

or

$$2\pi \left[-e^{-\frac{1}{2}r^2} \right]_0^{\infty} = c^2.$$

This implies $c = \sqrt{2\pi}$.

The accounting trick

The accounting trick: $\mathbb{E}(Z) = \sum_{\text{all } \omega} Z(\omega)\mathbb{P}(\omega)$

Hence, for $\psi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(\psi(Z)) = \begin{cases} \sum_{z \in S_Z} \psi(z) p_Z(z) & Z \text{ discrete} \\ \int_{-\infty}^{\infty} \psi(z) f_Z(z) dz & Z \text{ continuous} \end{cases}$$

Expectation of function (Ghahramani 8.1)

Theorem If (X, Y) is a discrete bivariate random variable with set of possible values $S_{(X,Y)}$ and probability mass function $p_{(X,Y)}(x, y)$, then, for any real-valued function ψ,

$$\mathbb{E}[\psi(X, Y)] = \sum_{(x,y) \in S_{(X,Y)}} \psi(x, y) p_{(X,Y)}(x, y)$$

provided the sum converges absolutely.

Example

If a fair coin is tossed three times, and X denotes the number of heads in the first two tosses, and Y the number of heads in the last two tosses, then we have seen that the pmf of (X, Y) is given by:

$p_{(X,Y)}(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	$\frac{1}{8}$	$\frac{1}{8}$	0
$y=1$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
$y=2$	0	$\frac{1}{8}$	$\frac{1}{8}$

Therefore $\mathbb{E}(XY) = 1 \times \frac{1}{4} + 2 \times \frac{1}{8} + 2 \times \frac{1}{8} + 4 \times \frac{1}{8} = \frac{5}{4}$.
Note that $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y) = 1$.

Theorem If (X, Y) is a continuous bivariate random variable with set of possible values $S_{(X,Y)}$ and probability density function $f_{(X,Y)}(x, y)$, then, for any real-valued function ψ,

$$\mathbb{E}[\psi(X, Y)] = \int \int_{(x,y)\in S_{(X,Y)}} \psi(x, y) f_{(X,Y)}(x, y) dy dx$$

provided the integral converges absolutely.

Example

Consider choosing a point at "random" in the unit square $S{=}(0,1)\times(0,1)$. What we mean by this is that the probability of being in any small rectangle in S must be constant wherever the rectangle is located in S. So we must have:

$$f_{(X,Y)}(x,y)=\begin{cases}1 & (x,y)\in S\\ 0 & \text{otherwise.}\end{cases}$$

Now let $\psi(X,Y)=X+Y$. Note that $0\leq X+Y\leq 2$ and that $X+Y$ is constant on straight lines of slope -1 which cut diagonally through S. Since the line $X+Y=1$ cuts S in half, by symmetry we should have $\mathbb{E}(X+Y)=1$. Let's check this using the definition of $\mathbb{E}(\psi(X,Y))$.

$$
\begin{aligned}
\mathbb{E}[\psi(X,Y)] &= \mathbb{E}[X+Y] \\
&= \int_0^1 \int_0^1 (x+y) \times 1 \, dy dx \\
&= \int_0^1 \Big(\int_0^1 (x+y) dy \Big) dx \\
&= \int_0^1 \Big[xy + \frac{1}{2} y^2 \Big]_0^1 \Big) dx \\
&= \int_0^1 \Big(x + \frac{1}{2} \Big) dx = \Big[\frac{1}{2} x^2 + \frac{1}{2} x \Big]_0^1 = 1.
\end{aligned}
$$

Expectation of a product (Ghahramani 8.2)

If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

The converse is not true, that is, the fact that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ does not imply that X and Y are independent.

Proof

We consider the case when X and Y are discrete random variables. The continuous case is analogous. Because X and Y are independent $p_{(X,Y)}(x,y) = p_X(x)p_Y(y)$ so that, by the accounting trick,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{\text{all } \omega} X(\omega)Y(\omega)\mathbb{P}(\omega) = \sum_{x\in S_X, y\in S_Y} xy\mathbb{P}(X=x, Y=y) \\
&= \sum_{x\in S_X, y\in S_Y} xyp_{(X,Y)}(x,y) = \sum_{y\in S_Y}\sum_{x\in S_X} xyp_X(x)p_Y(y) \\
&= \sum_{x\in S_X} xp_X(x) \sum_{y\in S_Y} yp_Y(y) = \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

Again, by induction we obtain,

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mathbb{E}[X_1]\mathbb{E}[X_2] \ldots \mathbb{E}[X_n]$$

provided that $X_1, X_2, \ldots, X_n$ are independent.

Transformations of Bivariate RV's (Ghahramani 8.4)

For other types of questions, it may be necessary to find the pdf or pmf of function of two rvs.

Example: If the prob that an item is defective is 0.1 for one line and 0.15 for the other line, and a sample of 10 is taken on each of two lines, find the prob that at least two items in the total sample are defective.

In general

If X and Y are independent rv's with values $0, 1, 2, \ldots,$

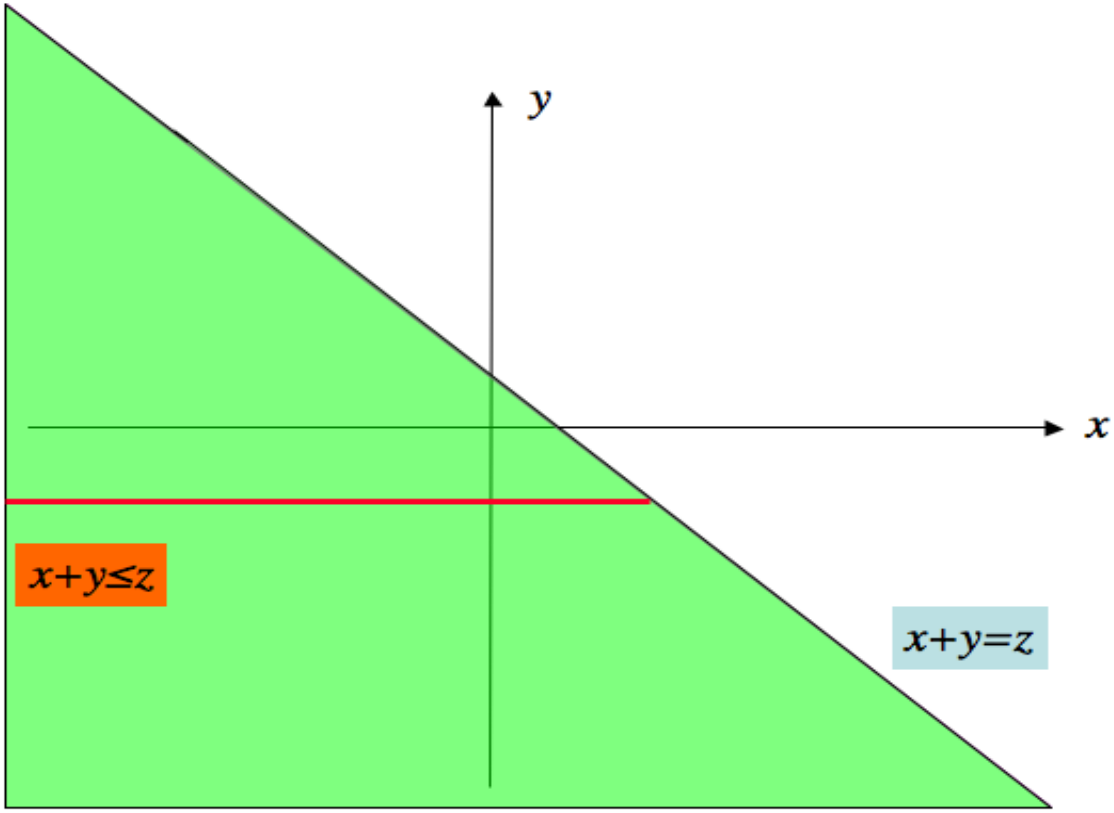
$$\begin{aligned} \mathbb{P}(X+Y=i) &= \sum_{j=0}^{i} \mathbb{P}(X=j)\mathbb{P}(Y=i-j) \\ &= \sum_{j=0}^{i} p_X(j)p_Y(i-j). \end{aligned}$$

To do this for cts. rvs, use pdf

We need to do cdf rather than pdf direct.

Suppose X and Y are independent and continuous, then

$$\begin{aligned}\mathbb{P}(X+Y\leq z) &= \iint_{\{(x,y):\ x+y\leq z\}} f_{(X,Y)}(x,y)dxdy \\ &= \iint_{\{(x,y):\ x+y\leq z\}} f_X(x)f_Y(y)dxdy.\end{aligned}$$



[Slide 360]

Hence,

$$\begin{aligned} F_{X+Y}(z) &= \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{z-y} f_X(x)dx \right) dy \\ &= \int_{-\infty}^{\infty} f_Y(y) F_X(z-y)dy. \end{aligned}$$

Differentiating with respect to z under the integral:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(y) f_X(z-y)dy.$$

Integrals of this form are called *convolution integrals*.

Summary

If X and Y are independent:

$$\begin{cases} p_{X+Y}(z) = \sum_{x \in S_X} p_X(x) p_Y(z-x) \\ \qquad\qquad = \sum_{y \in S_Y} p_X(z-y) p_Y(y), & (X,Y) \text{ discrete}, \\ f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \\ \qquad\qquad = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x) dx, & (X,Y) \text{ cont.} \end{cases}$$

Example

If $Z = X + Y$ with $X \stackrel{d}{=} \mathrm{Pn}(\lambda)$ and $Y \stackrel{d}{=} \mathrm{Pn}(\mu)$ and X and Y independent, show that $Z \stackrel{d}{=} \mathrm{Pn}(\lambda + \mu)$.

Example

If X and Y are independent $\exp(\lambda)$ random variables, then the pdf of $Z = X + Y$ is given by

Probability distribution of product

Let $U = XY$, where X and Y are independent random variables. Then the distribution function of U can be derived as follows:

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) = \mathbb{P}(XY \leq u) \\ &= \int_{S_X} \mathbb{P}(XY \leq u | X = x) f_X(x) dx \\ &= \int_{S_X} \mathbb{P}(xY \leq u | X = x) f_X(x) dx \\ &= \int_{S_X} \mathbb{P}(xY \leq u) f_X(x) dx \end{aligned}$$

$$
\begin{aligned}
&= \int_{S_X \cap (-\infty,0)} \mathbb{P}(Y \geq \frac{u}{x}) f_X(x) dx \\
&\quad + \int_{S_X \cap (0,\infty)} \mathbb{P}(Y \leq \frac{u}{x}) f_X(x) dx \\
&= \int_{S_X \cap (-\infty,0)} \left(1 - F_Y(\frac{u}{x})\right) f_X(x) dx \\
&\quad + \int_{S_X \cap (0,\infty)} F_Y(\frac{u}{x}) f_X(x) dx.
\end{aligned}
$$

The pdf of U can be obtained by differentiating. Thus

$$\begin{aligned} f_U(u) &= \int_{S_X \cap (-\infty,0)} \frac{-1}{x} f_Y(\frac{u}{x}) f_X(x) dx \\ &\quad + \int_{S_X \cap (0,\infty)} \frac{1}{x} f_Y(\frac{u}{x}) f_X(x) dx. \end{aligned}$$

An alternative expression can be obtained by interchanging X and Y.

Expected values of sums of RV's (Ghahramani 10.1)

A most important property of expectation is that it is additive. In other words, for any random variables X and Y,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof

We shall do the continuous case. The discrete case is analogous.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \int\int_{(x,y)\in S_{(X,Y)}} (x+y) f_{(X,Y)}(x,y)\,dy\,dx \\
&= \int_{x\in S_X} x \left(\int_{y\in S_Y} f_{(X,Y)}(x,y)\,dy \right) dx \\
&\quad + \int_{y\in S_Y} y \left(\int_{x\in S_X} f_{(X,Y)}(x,y)\,dx \right) dy \\
&= \int_{x\in S_X} x f_X(x)\,dx + \int_{y\in S_Y} y f_Y(y)\,dy \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

By induction, this can be extended to any number of random variables:

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots \mathbb{E}[X_n].$$

Try this yourself.

Variability of sums (Covariance) (Ghahramani 10.2)

Consider $Z = X + Y$. Intuitively you would expect the variability of the Z to reflect the individual variability of X and Y, but it is also clearly affected by their relationship. If we let $Y = X$ then:

$$V(X + Y) = V(2X) = 4V(X)$$

If we let $Y = -X$ then:

$$V(X + Y) = V(0) = 0.$$

In neither case is $V(X + Y) = V(X) + V(Y)$.

To derive a general expression for $V(X+Y)$ we start with the basic definition of the variance.

$$
\begin{aligned}
& V(X+Y) \\
&= \mathbb{E}\big((X+Y-(\mu_X+\mu_Y))^2\big) \\
&= \mathbb{E}\Big(\big((X-\mu_X)+(Y-\mu_Y)\big)^2\Big) \\
&= \mathbb{E}\big((X-\mu_X)^2\big)+2\mathbb{E}\big((X-\mu_X)(Y-\mu_Y)\big)+\mathbb{E}\big((Y-\mu_Y)^2\big) \\
&= V(X)+2\mathrm{Cov}(X,Y)+V(Y).
\end{aligned}
$$

where we define $\mathrm{Cov}(X,Y)=\mathbb{E}\big((X-\mu_X)(Y-\mu_Y)\big)$. Clearly $\mathrm{Cov}(X,Y)$ contains the required information about the relationship between X and Y.

If $\mathrm{Cov}(X, Y) = 0$ then the variance of the sum is the sum of the variances and the variance function is additive.

If $\mathrm{Cov}(X, Y) = 0$ we call X and Y *uncorrelated* random variables.

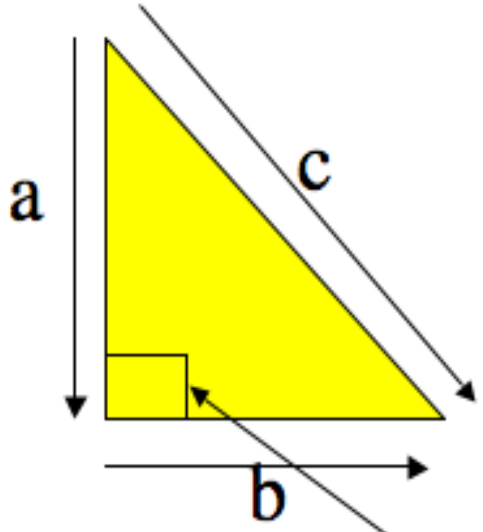
Thus if X and Y are *uncorrelated* random variables then $V(X + Y) = V(X) + V(Y)$.

Moreover, if X and Y are independent then we have seen that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ and so

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = 0.$$

Thus independent random variables are uncorrelated which gives us the special case that if X and Y are *independent* random variables then $V(X + Y) = V(X) + V(Y)$.

Pythagoras



$$c = \sqrt{a^2 + b^2}$$

If X and Y are independent

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Using induction, this can be extended to

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$$

for mutually independent $X_1, X_2, \ldots, X_n$.

Note that even when the variance is additive the standard deviation is not, $sd(X+Y) \neq sd(X) + sd(Y)$. In fact we have

$$sd(X+Y) = \sqrt{sd(X)^2 + sd(Y)^2}.$$

Example

If X denotes the number of successes in n independent Bernoulli trials with probability of success p, then

$$X = Z_1 + Z_2 + \cdots + Z_n$$

where the Z_i are independent and identically distributed Bernoulli random variables with pmf

z	0	1
$p(z)$	$1-p$	p

Now, $V(Z_i) = \mathbb{E}(Z_i^2) - \mathbb{E}(Z_i)^2 = p - p^2 = p(1-p)$ and $V(X) = V(Z_1) + V(Z_2) + \cdots + V(Z_n)$. So $V(X) = np(1-p)$.

Covariance of a linear combination (Ghahramani 10.2)

To discuss results of this type it is convenient to define $X_0 = X - \mu_X$ and $Y_0 = Y - \mu_Y$, so that

$$\begin{aligned} \mathbb{E}(X_0) &= \mathbb{E}(Y_0) = 0 \\ V(X) &= \mathbb{E}(X_0^2), V(Y) = \mathbb{E}(Y_0^2) \\ \mathrm{Cov}(X, Y) &= \mathbb{E}(X_0 Y_0). \end{aligned}$$

Further, if $Z = aX + bY$, then

$$\begin{aligned} Z_0 &= Z - \mu_Z \\ &= aX + bY - a\mu_X - b\mu_Y \\ &= a(X - \mu_X) - b(Y - \mu_Y) \\ &= aX_0 + bY_0. \end{aligned}$$

Therefore

$$\begin{aligned}
&\mathrm{Cov}(aX + bY,\ cX + dY) \\
&\quad = \quad \mathbb{E}\big((aX_0 + bY_0)(cX_0 + dY_0)\big) \\
&\quad = \quad ac\mathbb{E}(X_0^2) + (ad + bc)\mathbb{E}(X_0Y_0) + bd\mathbb{E}(Y_0^2) \\
&\quad = \quad acV(X) + (ad + bc)\mathrm{Cov}(X, Y) + bdV(Y).
\end{aligned}$$

Since $V(Z) = \mathrm{Cov}(Z, Z)$, a special case of this result is

$$V(aX + bY) = a^2V(X) + 2ab\mathrm{Cov}(X, Y) + b^2V(Y).$$

Example

If $V(X) = 2$, $V(Y) = 3$, $\mathrm{Cov}(X, Y) = -1$, and $U = X + Y$, $W = 2X - Y$, compute $\mathrm{Cov}(U, W)$.

Describing the Relationship Between X and Y (Ghahramani 10.2)

We saw that two *events* A and B are positively-related if

$$\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B)$$

and are negatively-related if

$$\mathbb{P}(A \cap B) < \mathbb{P}(A)\mathbb{P}(B).$$

This idea can be extended from events to random variables.

Definition

We say that random variables X and Y are *positively related* if $\mathrm{Cov}(X, Y) > 0$, *negatively related* if $\mathrm{Cov}(X, Y) < 0$ and *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$.

To see the connection, let X_A and X_B be the *indicator random variables* for the events A and B. This is

$$X_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$ and similarly for X_B. Then

$$\mathrm{Cov}(X_A, X_B) = \mathbb{E}(X_A X_B) - \mathbb{E}(X_A)\mathbb{E}(X_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

so the "relatedness" of the random variables X_A and X_B is the same as that for the events A and B.

Correlation (Ghahramani 10.3)

The *correlation coefficient* $\rho(X, Y)$ is defined by standardising the covariance to remove the scale effect. Thus

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

Below we shall show that $-1 \leq \rho(X, Y) \leq 1$.

Note on Covariance and Correlation

1. $\mathrm{Cov}(X, Y)$ is also denoted by σ_{XY}. Then $\rho(X, Y)$ is equal to
$$\frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

3. $\mathrm{Cov}(X, X) = V(X)$.

4. $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. This form is generally the best to use when evaluating $\mathrm{Cov}(X, Y)$. Note the similarity with $V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Proof

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

We have previously defined X and Y to be *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$, or equivalently $\rho(X, Y) = 0$. We have also seen that independent random variables are always uncorrelated:

$$X, Y \text{ independent} \quad \Longrightarrow \quad X, Y \text{ uncorrelated.}$$

However it is important to note that two variables being uncorrelated is not the same as them being independent, that is

$$X, Y \text{ uncorrelated} \quad \not\Longrightarrow \quad X, Y \text{ independent.}$$

Example

Suppose that X has a symmetric probability distribution with mean zero, and $Y = X^2$. Then

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) \\ &= 0 \end{aligned}$$

since $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^3) = 0$. Thus X and Y are uncorrelated. However, they are not independent, since they are functionally related.

The statement that X and Y are independent is equivalent to requiring that

$$\mathbb{E}\big(\phi(X)\psi(Y)\big) = \mathbb{E}\big(\phi(X)\big)\mathbb{E}\big(\psi(Y)\big)$$

for all functions ϕ and ψ.

This is a much stronger statement than

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

which is all that is required for zero correlation.

Proof that $|\rho| \leq 1$ (Ghahramani 10.3)

For any real number z, we have

$$V(zX + Y) = z^2 V(X) + 2z\text{Cov}(X, Y) + V(Y).$$

We know that $V(zX + Y) \geq 0$, since any variance is non-negative. So, viewing the right side of the above equation as a function in z, we have a quadratic function which is never negative.

Hence, the discriminant of the quadratic function is non-positive, and so

$$\Delta = b^2 - 4ac = \big(2\mathrm{Cov}(X,Y)\big)^2 - 4V(X)V(Y) \leq 0.$$

It follows that

$$\frac{\mathrm{Cov}(X,Y)^2}{V(X)V(Y)} \leq 1$$

and so $\rho(X,Y)^2 \leq 1$.

Example

Let X, Y and Z be independent Poisson-distributed random variables with parameter $\lambda = 10$. Find $\mathrm{Cov}(X + Y, Y + Z)$.

Now, by a previous example, $X + Y \stackrel{d}{=} \mathrm{Pn}(20)$ and $Y + Z \stackrel{d}{=} \mathrm{Pn}(20)$ and so $V(X + Y) = V(Y + Z) = 20$. Therefore

$$\rho(X + Y, Y + Z) = \frac{10}{\sqrt{20 \times 20}} = \frac{1}{2}.$$

A revisit of bivariate normal

If the pdf of (X, Y) is given by

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\Big)$$

then $\rho(X, Y) = \rho$.

Example: Total claims

Suppose the number of insurance claims is Poisson with mean 20 per year. Suppose the claim sizes are independent rv's each with mean 200 and standard deviation 200 and suppose the number and claim sizes are independent. Determine the mean and standard deviation of the total value of claims in a year.

Solution

Let N be the rv for the number of claims and X_1, X_2, ... be the claim size rv's.

The total value of claims is

$$T = \sum_{i=1}^{N} X_i.$$

By the law of total probability, for any $t \geq 0$,

$$\mathbb{P}(T = t) = \sum_{n=0}^{\infty} \mathbb{P}(T = t | N = n)\mathbb{P}(N = n).$$

Hence,

$$\begin{aligned}
\mathbb{E}(T) &= \sum_{t=1}^{\infty} t\mathbb{P}(T=t) \\
&= \sum_{t=1}^{\infty} t \sum_{n=0}^{\infty} \mathbb{P}(T=t|N=n)\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n) \sum_{t=1}^{\infty} t\mathbb{P}(T=t|N=n) \text{ swapping sum order} \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)\mathbb{E}(T|N=n).
\end{aligned}$$

This leads to conditioning on rv's.

Conditioning on RV's (Ghahramani 10.4)

Consider the function

$$\eta(y) = \mathbb{E}[X|Y = y] = \sum_{x \in S_X} x p_{X|Y}(x|y).$$

This is the expected value of X given that $Y = y$.

We can apply this function to the random variable Y to derive a random variable $Z = \eta(Y)$.

Z is known as the *conditional expectation of X given Y*. We write it as $\mathbb{E}[X|Y]$.

About $\mathbb{E}[X|Y]$

If Y takes values $0, 1, 2, \ldots$, then $\mathbb{E}(X|Y)$ is a random variable with the following distribution

Values	$\eta(0)$	$\eta(1)$	$\eta(2)$	$\ldots$
Prob	$p_Y(0)$	$p_Y(1)$	$p_Y(2)$	$\ldots$

In general, if Y is discrete, $\mathbb{E}(X|Y)$ takes value $\eta(y)$ with probability $p_Y(y)$; if Y is cont., $\mathbb{E}(X|Y)$ takes values "around" $\eta(y)$ with "probability" $f_Y(y)dy$.

Now let's work out the expected value of Z. This is given by

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[\eta(Y)] \\
&= \int_{-\infty}^{\infty} \eta(y) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} x f(x \mid y) dx \Big) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{(X,Y)}(x, y) dy dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \mathbb{E}[X].
\end{aligned}$$

We therefore have the useful result:

$$\mathbb{E}[X] = \mathbb{E}\big(\mathbb{E}[X|Y]\big).$$

This can be thought of as an expected value version of the Law of Total Probability.

Example

If (X, Y) has pdf $f(x, y) = 2$ $(x + y \leq 1, x \geq 0, y \geq 0)$ then $\mathbb{E}[X] = \mathbb{E}[Y] = 1/3$. We have seen that $(X|Y = y) \stackrel{d}{=} \mathrm{R}(0, 1 - y)$.

It follows that $\mathbb{E}[X|y] = (1 - y)/2$. Therefore

$$\mathbb{E}\big(\mathbb{E}(X|Y)\big) = \mathbb{E}\big((1 - Y)/2\big) = 1/2 - 1/2 \times 1/3 = 1/3 = \mathbb{E}[X].$$

The above result can be extended to functions of X and Y.
Thus, using exactly the same argument,

$$\mathbb{E}\big(\psi(X,Y)\big) = \mathbb{E}\big(\mathbb{E}(\psi(X,Y)|Y)\big).$$

Try to prove this yourself.

As an example

$$\mathbb{E}(XY) = \mathbb{E}\big(\mathbb{E}(XY|Y)\big) = \mathbb{E}\big(Y\mathbb{E}(X|Y)\big).$$

Conditional variance (Ghahramani 10.4)

We have previously seen that, when $(\eta(y) = \mathbb{E}(X|Y = y))$,

$$\mathbb{E}(X) = \mathbb{E}(\eta(Y)) = \mathbb{E}(\mathbb{E}(X|Y)),$$

which is useful for calculating the mean particularly in the case of two stage random experiments.

It is reasonable to ask whether we can we find a similar formula for $V(X)$. Consider $v(y) = V(X|Y = y)$. Analogously to $\eta(Y)$ we will write the random variable $v(Y)$ as $V(X|Y)$.

Then

$$v(y) = V(X|Y = y) = \mathbb{E}(X^2|Y = y) - \mathbb{E}(X|Y = y)^2.$$

So

$$v(Y) = V(X|Y) = \mathbb{E}(X^2|Y) - \mathbb{E}(X|Y)^2 = \mathbb{E}(X^2|Y) - \eta(Y)^2.$$

Taking expectations gives

$$\mathbb{E}(V(X|Y)) = \mathbb{E}(\mathbb{E}(X^2|Y)) - \mathbb{E}(\eta(Y)^2) = \mathbb{E}(X^2) - \mathbb{E}(\eta(Y)^2).$$

Now consider the variance of $\mathbb{E}(X|Y)$. This is

$$V(\eta(Y)) = \mathbb{E}(\eta(Y)^2) - \mathbb{E}(\eta(Y))^2 = \mathbb{E}(\eta(Y)^2) - \mathbb{E}(X)^2.$$

Adding the two equations gives

$$\begin{aligned} V(\mathbb{E}(X|Y)) + \mathbb{E}(V(X|Y)) &= \mathbb{E}(\eta(Y)^2) - \mathbb{E}(X)^2 \\ &\quad + \mathbb{E}(X^2) - \mathbb{E}(\eta(Y)^2) \\ &= V(X). \end{aligned}$$

The rather beautiful formula

$$V(X) = V(\mathbb{E}(X|Y)) + \mathbb{E}(V(X|Y))$$

can be remembered as

'The variance of X equals the variance of the conditional mean of X given Y plus the mean of the conditional variance of X given Y'.

Claims Example (cont)

$T = X_1 + X_2 + \cdots + X_N$, where the X_i's are independent with mean μ and variance σ^2, and N is a non-negative integer-valued random variable independent of the X_is.

Given $N = n$, $T = X_1 + X_2 + \cdots + X_n$.

So $\mathbb{E}(T|N = n) = n\mu$ and $V(T|N = n) = n\sigma^2$.

Therefore $\mathbb{E}(T|N) = N\mu$ and $V(T|N) = N\sigma^2$. Hence,

$$\begin{aligned} V(T) &= \mathbb{E}\big(V(T|N)\big) + V\big(\mathbb{E}(T|N)\big) \\ &= \mathbb{E}(N\sigma^2) + V(N\mu) \\ &= \sigma^2\mathbb{E}(N) + \mu^2 V(N) \\ &= 1.6 \times 10^6. \end{aligned}$$

Approximations for the mean and variance of functions

In theory, values of $\mathbb{E}\big(\psi(X)\big)$ and $V\big(\psi(X)\big)$ can be found exactly from $f_X(x)$ (or $p_X(x)$). However, often the integrals (or sums) involved are complicated or can only be evaluated by numerical integration. Thus approximations, even rough ones, for the mean and variance can be useful.

Assume that $\psi(X)$ can be expanded in a Taylor's series about $X = \mu$ so that

$$\psi(X) = \psi(\mu) + (X - \mu)\psi'(\mu) + \frac{1}{2}(X - \mu)^2\psi''(\mu) + \cdots .$$

Then,

$$\psi(X) \approx \psi(\mu) + (X - \mu)\psi'(\mu) + \frac{1}{2}(X - \mu)^2\psi''(\mu).$$

Taking expectations of both sides, we obtain the approximation

$$\mathbb{E}\big(\psi(X)\big) \approx \psi(\mu) + \frac{1}{2}\psi''(\mu)V(X).$$

To derive an approximation to the variance, we use the first-order Taylor series approximation

$$\psi(X) \approx \psi(\mu) + (X - \mu)\psi'(\mu).$$

If this is a reasonable approximation within the range of probable values of X, then

$$V\big(\psi(X)\big) \approx \psi'(\mu)^2 V(X)$$

In spite of their rough origins, these two results are remarkably useful and often surprisingly accurate. They perform best when $V(X)$ is small, since then the deviations from μ are likely to be small.

Example

If $X \stackrel{d}{=} \mathrm{R}(0,1)$, that is, $f_X(x) = 1 \quad (0 < x < 1)$; and $Z = \psi(X) = e^X$, then:

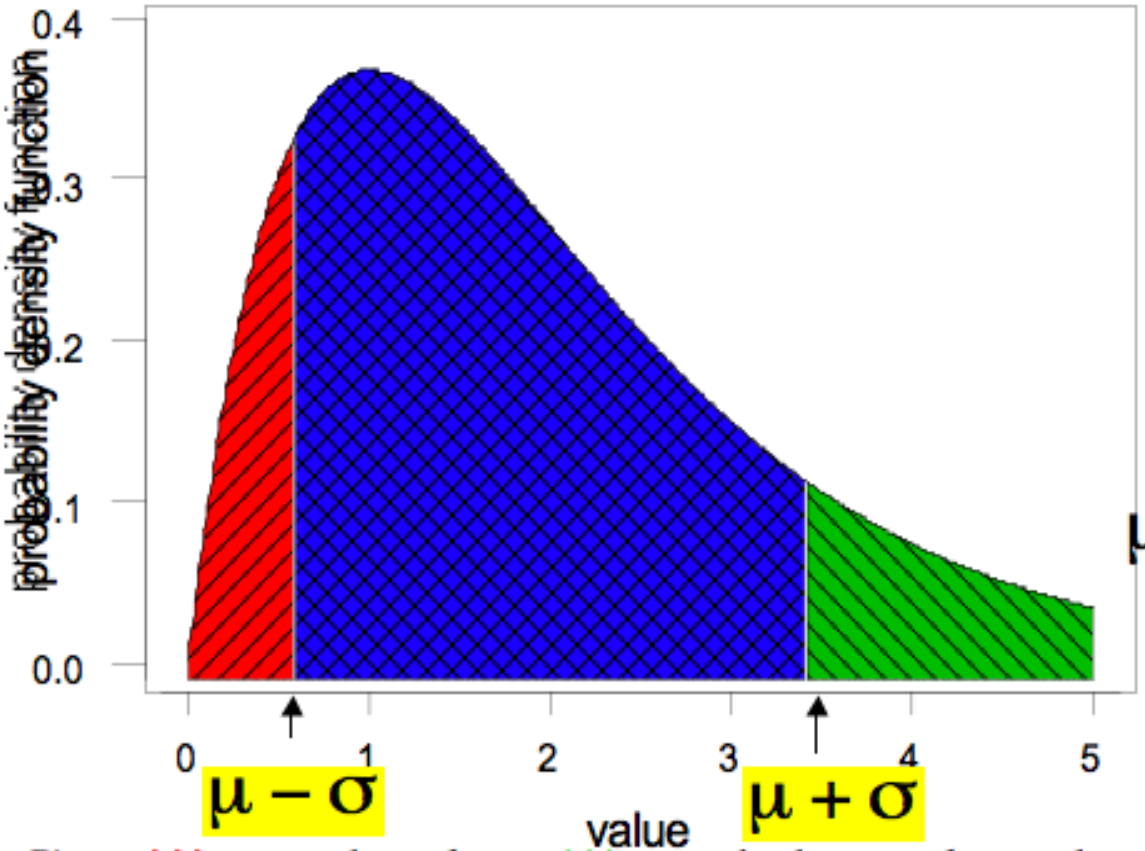
$$\begin{aligned} \mathbb{E}(Z) &= \int_0^1 e^x dx = e - 1 = 1.718 \\ \mathbb{E}(Z^2) &= \int_0^1 e^{2x} dx = \frac{1}{2}(e^2 - 1) = 3.195 \\ V(Z) &= 3.195 - 1.718^2 = 0.242. \end{aligned}$$

To compute the approximations we need

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{2} \\
\mathbb{E}(X^2) &= \int_0^1 x^2 dx = \frac{1}{3} \\
V(X) &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.
\end{aligned}$$

So the approximations give:

Is the spread related to prob?



$\mu = 2;\ \sigma = \sqrt{2}$

Can \\\ prob plus /// prob be related to sd?

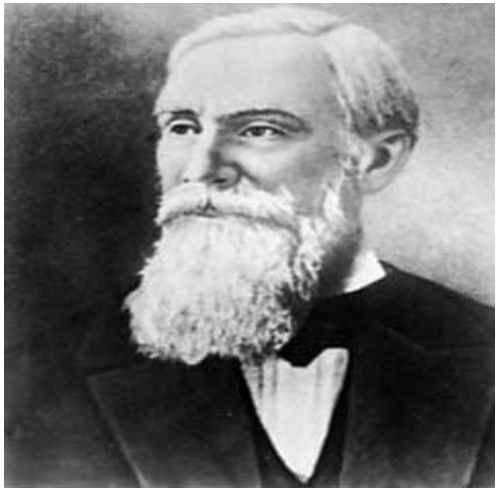
Bienaymé inequality (1853) (Ghahramani 11.3)

Bienaymé inequality (aka Chebyshev's inequality (1867)) gives some quantitative idea of standard deviation as a measure of spread. It is stated as follows:

If X has mean μ and variance σ^2, then

$$\mathbb{P}\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}.$$

Irénée-Jules Bienaymé
[28/08/1796 – 19/10/1878]



Pafnuty Lvovich Chebyshev
[16/05/1821 – 8/12/1894]

Some facts

- $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 0.75.$
- $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \geq 0.89.$
- Challenge: For fixed k, can we find a rv X such that

$$\mathbb{P}\left(\frac{|X - \mu_X|}{\sigma_X} \geq k\right) = \frac{1}{k^2}?$$

It's a crude estimate!

The estimate is relatively crude.

For example, when X is normally distributed it is easily verified using tables that:

$$\mathbb{P}(|X - \mu| \geq 2\sigma) = 0.0455.$$

$$\mathbb{P}(|X - \mu| \geq 3\sigma) = 0.0027.$$

So the probability that an observation is more than 2 (or 3) standard deviations away from the mean (for any normal distribution) are roughly 0.05 and 0.003.

Bienaymé's proof

Let $B = \{|X - \mu| \geq k\sigma\}$ and $\mathbb{1}_B$ be the random variable mapping $\omega \in B$ to 1 and $\omega \in B^c$ to 0, then we can check that $(X - \mu)^2 \mathbb{1}_B \geq \sigma^2 k^2 \mathbb{1}_B$ and $\mathbb{E}\left(\sigma^2 k^2 \mathbb{1}_B\right) = \sigma^2 k^2 \mathbb{P}(B)$, hence

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(X - \mu)^2] \\
&\geq \mathbb{E}\left((X - \mu)^2 \mathbb{1}_B\right) \\
&\geq \mathbb{E}\left(\sigma^2 k^2 \mathbb{1}_B\right) \\
&= \sigma^2 k^2 \mathbb{E}\left(\mathbb{1}_B\right) \\
&= \sigma^2 k^2 \mathbb{P}(B).
\end{aligned}$$

The result follows on dividing both sides of this inequality by $k^2\sigma^2$.

Example

Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 500 and variance 100, what can be said about the probability that this week's production will be between 400 and 600?

Generating Functions and Applications (Ghahramani Ch 11)

Let $\{a_k,\ k = 0, 1, 2, \ldots\}$ be a sequence of real numbers. Then the *generating function* of the sequence is the series defined by

$$A(z) = \sum_{k=0}^{\infty} a_k z^k.$$

This series will converge for $z \in C \subseteq \mathbb{R}$. This is the range of z for which we consider the generating function to be defined.

We shall only consider sequences for which $C \neq \{0\}$.

The function $A(z)$ is just a convenient way of storing the sequence $\{a_k\}$. Clearly $\{a_k\}$ defines $A(z)$. By the uniqueness of power series, it is also true that $A(z)$ defines $\{a_k\}$.

We can extract the sequence $\{a_k\}$ from $A(z)$ by either

1. expanding $A(z)$ as a power series and evaluating the coefficients, or
2. differentiating:

$$k!a_k = \frac{d^k}{dz^k}\big(A(z)\big)_{z=0}.$$

Example

If $a_k = 1/k!$, then $A(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} = e^z$. Note that the series converges for all $z \in \mathbb{R}$.

Moreover, if $A(z) = e^z$ then

$$\frac{d^k}{dz^k}\big(A(z)\big)_{z=0} = e^z|_{z=0} = 1,$$

and therefore $a_k = 1/k!$.

Example

If $a_k = \binom{n}{k}$, then $A(z) = \sum_{k=0}^{\infty} \binom{n}{k} z^k = (1+z)^n$. If n is a positive integer, then the coefficients of the series are zero for $k > n$ and so the series converges for all $z \in \mathbb{R}$. Otherwise, we can use the ratio test to show that the series converges for all z with $|z| < 1$.

If $A(z) = (1+z)^n$, then, for $|z| < 1$, the binomial theorem tells us that $A(z) = \sum_{k=0}^{\infty} \binom{n}{k} z^k$ from which we can deduce that $a_k = \binom{n}{k}$.

Probability Generating Functions Definition

The *probability generating function (pgf)* of a non-negative integer valued rv X is given by

$$P_X(z) = \sum_{x=0}^{\infty} p_X(x) z^x.$$

When $|z| \leq 1$,

$$\begin{aligned} |P_X(z)| &\leq \sum_{x=0}^{\infty} |p_X(x) z^x| \\ &= \sum_{x=0}^{\infty} p_X(x) |z|^x \\ &\leq \sum_{x=0}^{\infty} p_X(x) \\ &= 1, \end{aligned}$$

so the generating function always converges for $|z| \leq 1$.

From pgf to pmf

$$p_X(k) = \mathbb{P}(X = k) = \frac{P_X^{(k)}(0)}{k!},$$

where $P_X^{(k)}$ is the kth derivative.

We can also easily compute $\mathbb{P}(X \text{ is even})$ and $\mathbb{P}(X \text{ is odd})$.

On discrete distribution on integers

So knowing the pgf will tell you the probability distribution (by differentiating and evaluating at 0).

A discrete distribution on integers can be described by:

- probability mass function
- cumulative distribution function
- OR probability generating function

Example - Binomial random variable

If $X \stackrel{d}{=} \text{Bi}(n, p)$, then

$$\begin{aligned} P_X(z) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} z^x \\ &= \sum_{x=0}^{n} \binom{n}{x} (pz)^x (1-p)^{n-x} \\ &= (1-p+pz)^n \end{aligned}$$

by applying the Binomial theorem. $P_X(z)$ converges for all $z \in \mathbb{R}$.

Example - Poisson random variable

If $X \stackrel{d}{=} \mathrm{Pn}(\lambda)$, then

$$\begin{aligned} P_X(z) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x z^x}{x!} \\ &= e^{-\lambda}e^{\lambda z} \\ &= e^{-\lambda(1-z)} \end{aligned}$$

using the Taylor series expansion for $e^{\lambda z}$.

$P_X(z)$ converges for all $z \in \mathbb{R}$.

Example - Negative Binomial random variable

If $X \stackrel{d}{=} \mathrm{Nb}(r,p)$, so that, for $x = 0, 1, 2, \ldots$,
$p_X(x) = \binom{-r}{x} p^r (p-1)^x$, then

$$\begin{aligned} P_X(z) &= \sum_{x=0}^{\infty} \binom{-r}{x} p^r (p-1)^x z^x \\ &= p^r \sum_{x=0}^{\infty} \binom{-r}{x} ((p-1)z)^x \\ &= p^r (1-(1-p)z)^{-r}. \end{aligned}$$

$P_X(z)$ converges for all for all z with $|z| < 1/(1-p)$.

Properties of the pgf

1. $P_X(1) = 1$, since $\sum_{x=0}^{\infty} p_X(x) = 1$.
2. $P'_X(1) = \mathbb{E}(X)$, since $P'_X(z) = \sum_{x=0}^{\infty} x p_X(x) z^{x-1}$.
3. $P''_X(1) = \mathbb{E}\left(X(X-1)\right)$, since
   $P''_X(z) = \sum_{x=0}^{\infty} x(x-1) p_X(x) z^{x-2}$.
4. Using 2 and 3 we have $V(X) = P''_X(1) + P'_X(1) - P'_X(1)^2$.
5. From the definition of $\mathbb{E}(z^X)$, we see that we can write $P_X(z)$ as $\mathbb{E}(z^X)$.
6. If X and Y are independent, and $W = X + Y$, then
   $P_W(z) = P_X(z) P_Y(z)$ since
   $P_W(z) = \mathbb{E}(z^{X+Y}) = \mathbb{E}(z^X)\mathbb{E}(z^Y)$.

From our work on bivariate distributions we know that the pmf of the sum W of two independent, non-negative random variables X and Y is given by

$$p_W(w) = \sum_{x=0}^{w} p_Y(w-x)p_X(x).$$

Property 6 has told us that the pgf of W is the product of the pgfs of X and Y. That is

$$P_W(z) = P_X(z)P_Y(z).$$

This is known as *the convolution theorem*.

It gives another method to find the pmf of W.

Example - Means and variances

Once we have the pgf, finding the mean and variance is a simple matter of differentiation. If $X \stackrel{d}{=} \text{Bi}(n,p)$:

$$\begin{aligned} P'_X(z) &= \frac{d}{dz}(1-p+pz)^n = n(1-p+pz)^{n-1}p \\ P''_X(z) &= n(n-1)(1-p+pz)^{n-2}p^2. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}(X) &= P'_X(1) = np \\ V(X) &= P''_X(1) + P'_X(1) - P'_X(1)^2 \\ &= n(n-1)p^2 + np - (np)^2 = np(1-p). \end{aligned}$$

If $X \stackrel{d}{=} \mathrm{Pn}(\lambda)$ then

$$\begin{aligned} P'_X(z) &= \frac{d}{dz} e^{-\lambda(1-z)} = \lambda e^{-\lambda(1-z)} \\ P''_X(z) &= \lambda^2 e^{-\lambda(1-z)}. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}(X) &= P'_X(1) = \lambda \\ V(X) &= P''_X(1) + P'_X(1) - P'_X(1)^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

If $X \stackrel{d}{=} G(p) \stackrel{d}{=} Nb(1,p)$ then

$$\begin{aligned} P'_X(z) &= \frac{d}{dz} p^1 (1-(1-p)z)^{-1} = \frac{p(1-p)}{(1-(1-p)z)^2} \\ P''_X(z) &= 2p(1-p)^2(1-(1-p)z)^{-3}. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}(X) &= P'_X(1) = p(1-p)p^{-2} = \frac{(1-p)}{p} \\ V(X) &= P''_X(1) + P'_X(1) - P'_X(1)^2 \\ &= \frac{2p(1-p)^2}{p^3} + \frac{(1-p)}{p} - \left(\frac{(1-p)}{p}\right)^2 = \frac{(1-p)}{p^2}. \end{aligned}$$

Example - Distribution of sums

Once we have their pgf's, finding the distribution of the sum of independent random variables simply requires us to identify the distribution associated with the product of the pgf's. Suppose that $X \stackrel{d}{=} \mathrm{Pn}(\lambda)$, $Y \stackrel{d}{=} \mathrm{Pn}(\mu)$ are independent. Then

$$\begin{aligned} P_{X+Y}(z) &= P_X(z)P_Y(z) \\ &= e^{-\lambda(1-z)}e^{-\mu(1-z)} \\ &= e^{-(\lambda+\mu)(1-z)}. \end{aligned}$$

Hence $X + Y \stackrel{d}{=} \mathrm{Pn}(\lambda + \mu)$.

If $X \stackrel{d}{=} \text{Bi}(n, p)$ then

$$X = X_1 + X_2 + \ldots + X_n$$

where $X_i \stackrel{d}{=} \text{Bernoulli } (p)$. So

$$P_{X_i}(z) = \mathbb{E}(z^{X_i}) = z^0(1-p) + z^1 p = 1 - p + pz$$

and hence

$$P_X(z) = P_{X_1}(z) \times \ldots \times P_{X_n}(z) = (1 - p + pz)^n$$

which we have already seen is the pgf of $\text{Bi}(n, p)$.

Moment generating functions (Ghahramani Ch 11)

Probability generating functions are a very useful tool for deriving properties of discrete random variables. However, they are not usually used for continuous random variables. It is much more common to use the moment generating function in this case.

Before defining the moment generating function we will look in more detail at the meaning of some of the so called higher moments of a random variable.

Moments of a random variable (Ghahramani 4.5, 11.1)

The k^{th} moment (about the origin) of a random variable X is given by $\mu_k = \mathbb{E}(X^k)$.

The k^{th} central moment (about the mean) of a random variable X is given by $\nu_k = \mathbb{E}\big((X - \mu)^k\big)$.

We can express ν_k in terms of the μ_js as follows:

$$\begin{aligned}
\nu_k &= \mathbb{E}\big((X-\mu)^k\big) \\
&= \mathbb{E}\left(\sum_{j=0}^{k}\binom{k}{j}X^j(-\mu)^{k-j}\right) \\
\nu_k &= \sum_{j=0}^{k}\binom{k}{j}\mu_j(-\mu)^{k-j}.
\end{aligned}$$

For $k=2$, this gives us

$$\begin{aligned}
\nu_2 &= (-\mu)^2 + 2\mu^2 + \mu_2 \\
&= \mu_2 - \mu^2,
\end{aligned}$$

noting that $\mu_0 = 1$ and $\mu_1 = \mu$. This is our familiar formula for the variance.

Skewness and Kurtosis (Ghahramani 11.1)

We have seen that $\mu_1 = \mu$ and $\nu_2 = \sigma^2$ are used as measures of location and spread of a probability distribution.

The third and fourth moments about the mean are used as measures of *skewness* and *kurtosis* respectively.

Measure of Skewness

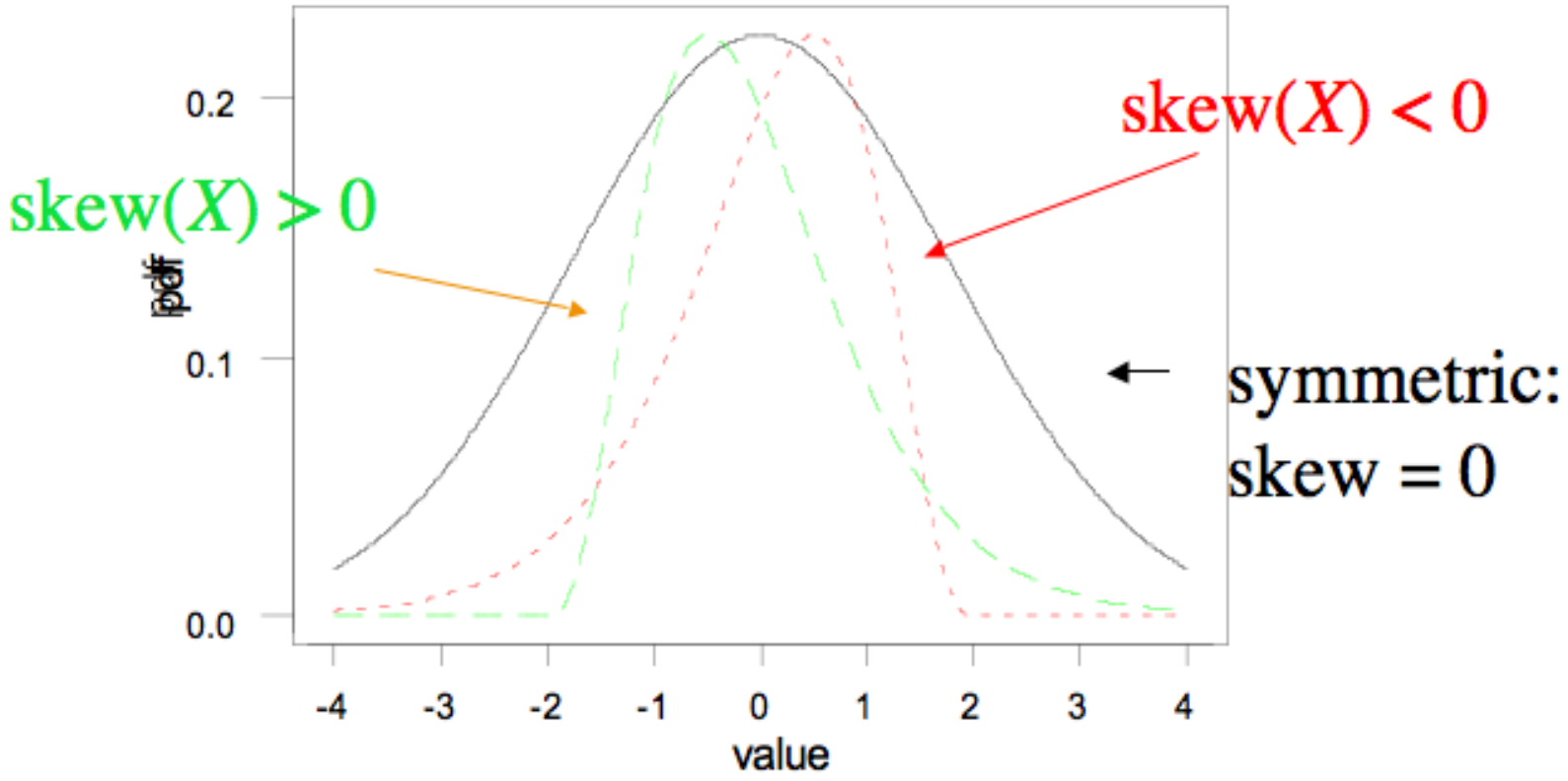
The third central moment: $\nu_3 = \mathbb{E}\big((X-\mu)^3\big)$ is an indicator of skewness. If the distribution has a long positive tail then $(X-\mu)^3$ has large positive values but does not have large negative values. Thus $\nu_3 > 0$ and the distribution is *positively skew*. Similarly, if the distribution has a long negative tail, $\nu_3 < 0$ and the distribution is *negatively skew*.

The *coefficient of skewness* is obtained by standardising to remove the scale effect

$$\text{skew}(X) = \frac{\nu_3}{\sigma^3}$$

If the pdf of X is symmetric, then $\text{skew}(X)$ is identically zero. This follows because as $(x-\mu)^3 f_X(x)$ is an odd function around μ.

rv's, X, in stat units ie mean 0, var 1



Example

If X has pdf $f_X(x) = 2x \quad (0 < x < 1)$, then:

$$\mu = \tfrac{2}{3}, \;\; \mu_2 = \tfrac{1}{2}, \;\; \mu_3 = \int_0^1 2x^4 dx = \tfrac{2}{5}, \; \sigma^2 = \tfrac{1}{2} - \tfrac{4}{9} = \tfrac{1}{18}.$$

For $k = 3$,

$$\begin{aligned} \nu_3 &= (-\mu)^3 + 3\mu(-\mu)^2 + 3\mu_2(-\mu) + \mu_3 \\ &= \mu_3 - 3\mu_2(\mu) + 2\mu^3. \end{aligned}$$

So

$$\nu_3 = \tfrac{2}{5} - 3 \times \tfrac{2}{3} \times \tfrac{1}{2} + 2 \times \tfrac{8}{27} = -\tfrac{1}{135}$$

Hence:

$$\text{skew}(X) = -\frac{18^{3/2}}{135} = -0.5657.$$

Measure of kurtosis

The fourth central moment $\nu_4 = \mathbb{E}\big((X-\mu)^4\big)$ is an indicator of the peakedness and the length of the tails of the distribution — often called *kurtosis*. The *coefficient of kurtosis* is then obtained by standardising to remove scale effect:

$$\text{kurt}(X) = \frac{\nu_4}{\sigma^4} - 3.$$

We subtract 3 so that $\text{kurt}(X)$ is zero for normal distributions. Then, if the distribution is flatter and shorter tailed than a normal distribution, the coefficient of kurtosis is negative; while if the distribution is more peaked and has longer tails than a normal distribution, then the coefficient of kurtosis is positive. (NB Ghahramani does not subtract 3 in its definition of kurtosis)

Example

We compare the pdfs

$$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \quad (-\infty < t < \infty)$$

$$f_Y(t) = \frac{1}{2\sqrt{3}} \quad (-\sqrt{3} < t < \sqrt{3}).$$

We can easily show that both have mean 0 and variance 1. Also since both are symmetric, both have skewness coefficient 0. Lets compare the kurtosis of the two distributions.

$$
\begin{aligned}
\mathbb{E}(X^4) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 \, x e^{-\frac{1}{2}x^2} dx \\
&= \frac{1}{\sqrt{2\pi}} \left[-x^3 e^{-\frac{1}{2}x^2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} 3x^2 e^{-\frac{1}{2}x^2} dx \\
&= 0 + 3\mathbb{E}(X^2) \\
&= 3 \\
\mathrm{kurt}(X) &= 3 - 3 \;=\; 0 \\
\mathbb{E}(Y^4) &= \int_{-\sqrt{3}}^{\sqrt{3}} x^4 \frac{1}{2\sqrt{3}} dx = \frac{1}{\sqrt{3}} \left[\frac{x^5}{5} \right]_0^{\sqrt{3}} = \frac{9}{5} = 1.8 \\
\mathrm{kurt}(Y) &= 1.8 - 3 \;=\; -1.2
\end{aligned}
$$

f_Y is flatter and has shorter tails than f_X.

Mgf definition (Ghahramani 11.1)

Definition The *moment generating function (mgf)* of a random variable X is defined by

$$M_X(t) = \mathbb{E}(e^{tX})$$

for all t in the set T for which $\mathbb{E}(e^{tX}) < \infty$.

So, when X is continuous,

$$M_X(t) = \int_{S_X} e^{tx} f_X(x) dx$$

whenever the integral converges.

Expanding the exponential as a power series, we obtain:

$$\begin{aligned} M_X(t) &= \mathbb{E}\left(\sum_{k=0}^{\infty} X^k \frac{t^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \mathbb{E}(X^k)\frac{t^k}{k!} \\ &= \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!} \quad \text{for } t \in T \end{aligned}$$

Thus $M_X(t) \leftrightarrow \{\mu_k\}$. That is, the moments of X are uniquely determined by the mgf. Is this sufficient to uniquely determine the distribution of X? The answer is "no" unless we impose more conditions.

Exercise

Can you find a Rectangular Distribution which has the same mean and variance as a standard normal distribution?

Note that both the range of possible values and the shapes are very different for these two distributions with the same mean (location) and variance ("spread").

Properties of the mgf (Ghahramani 11.1, 11.2)

1. $M_X(0) = 1$.
2. $M'_X(0) = \mathbb{E}(X)$, since $M'_X(t) = \mathbb{E}(Xe^{tX})$.
   $M''_X(0) = \mathbb{E}(X^2)$, since $M''_X(t) = \mathbb{E}(X^2e^{tX})$
3. Using 2 and 3 we have $V(X) = M''_X(0) - M'_X(0)^2$.
4. In fact, for any k, $\mu_k = M_X^{(k)}(0) = \frac{d^k}{dz^k} M_X(t)\Big|_{t=0}$. We can find all $\{\mu_k\}$ directly (without differentiation) if we can write out $M_X(t)$ as a power series in t.
5. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$ since $\mathbb{E}(e^{Yt}) = e^{bt}\mathbb{E}(e^{Xat})$.

6. If X and Y are independent, and $Z = X + Y$, then $M_Z(t) = M_X(t)M_Y(t)$ since $M_Z(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY})$.

7. If X is a discrete random variable defined on the non-negative integers, having pgf $P_X(z)$, then

$$M_X(t) = P_X(e^t), \quad P_X(z) = M_X(\log z).$$

8. For some random variables, the mgf does not exist, that is, the set T for which $M_X(t) < \infty$ contains only zero, that is $T = \{0\}$.

9. The central moment generating function is given by

$$N_X(t) = \mathbb{E}\left(e^{(X-\mu)t}\right) = \mathbb{E}\left(\sum_{k=0}^{\infty}(X-\mu)^k\frac{t^k}{k!}\right) = \sum_{k=0}^{\infty}\nu_k\frac{t^k}{k!}$$

and we see that $N_X(t) = e^{-\mu t}M_X(t)$.

10. The mgf $M_X(t)$ uniquely determines the distribution of X, that is its distribution function $F_X(x)$.

Examples - Deriving moments

If $X \stackrel{d}{=} exp(\alpha)$, that is, $f(x) = \alpha e^{-\alpha x} \quad (x > 0)$, then

$$\begin{aligned} M(t) &= \int_0^\infty e^{xt} \alpha e^{-\alpha x} dx \\ &= \alpha \int_0^\infty e^{-(\alpha - t)x} dx \\ &= \frac{\alpha}{\alpha - t}, \end{aligned}$$

provided that $t < \alpha$. Therefore $T = \{t : t < \alpha\}$ and

$$M(t) = \frac{1}{1 - \frac{t}{\alpha}} = \sum_{k=0}^{\infty} (\frac{t}{\alpha})^k = \sum_{k=0}^{\infty} \left(\frac{k!}{\alpha^k} \right) \frac{t^k}{k!} \quad \Longrightarrow \quad \mu_k = \frac{k!}{\alpha^k}.$$

If $X \stackrel{d}{=} \mathrm{Pn}(\lambda)$, that is $p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad (x = 0, 1, 2, \ldots)$; then

$$\begin{aligned}
M_X(t) &= \sum_{x=0}^{\infty} e^{xt} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda+\lambda e^t} \\
&= e^{\lambda(e^t-1)} \\
&= \sum_{k=0}^{\infty} \frac{(\lambda(e^t-1))^k}{k!} \\
&= 1 + \lambda(t + \frac{t^2}{2} + \cdots) + \frac{\lambda^2}{2}(t + \cdots)^2 + \cdots \\
&= 1 + \lambda t + (\lambda + \lambda^2)\frac{t^2}{2} + \cdots
\end{aligned}$$

so that $\mu_1 = \lambda$, $\mu_2 = \lambda + \lambda^2$ and $\sigma^2 = \mu_2 - \mu_1^2 = \lambda$.

If $X \stackrel{d}{=} \gamma(r, \alpha)$ so that $f_X(x) = \alpha^r e^{-\alpha x} x^{r-1}/\Gamma(r)$ $(x > 0)$, then

$$\begin{aligned} M_X(t) &= \int_0^\infty \frac{e^{xt}\alpha^r e^{-\alpha x} x^{r-1}}{\Gamma(r)} dx \\ &= \alpha^r \int_0^\infty \frac{e^{-(\alpha-t)x} x^{r-1}}{\Gamma(r)} dx \\ &= \frac{\alpha^r}{(\alpha-t)^r} \int_0^\infty \frac{e^{-u} u^{r-1}}{\Gamma(r)} du \qquad \text{with } u = (\alpha - t)x \end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha^r}{(\alpha - t)^r} \qquad \text{provided } t < \alpha \\
&= (1 - \frac{t}{\alpha})^{-r} \\
&= \sum_{k=0}^{\infty} \binom{-r}{k} (-\frac{t}{\alpha})^k \\
&= 1 + \frac{rt}{\alpha} + \frac{r(r+1)}{\alpha^2} \frac{t^2}{2} + \cdots
\end{aligned}
$$

Therefore for $X \stackrel{d}{=} \gamma(r, \alpha)$ we have

$$\begin{aligned} \mathbb{E}(X) &= \frac{r}{\alpha} \\ \mathbb{E}(X^2) &= \frac{r(r+1)}{\alpha^2} \\ V(X) &= \frac{r}{\alpha^2}. \end{aligned}$$

If $Z \stackrel{d}{=} N(0,1)$ then the mgf of Z is given by

$$\begin{aligned} M_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{zt - \frac{1}{2}z^2} dz \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz \\ &= e^{\frac{1}{2}t^2}. \end{aligned}$$

It follows that the mgf of $X \stackrel{d}{=} N(\mu, \sigma^2)$ is given by

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

since $X \stackrel{d}{=} \sigma Z + \mu$, so that $M_X(t) = e^{\mu t} M_Z(\sigma t)$.

Hence the central moment generating function of X is given by

$$N_X(t) = e^{\frac{1}{2}\sigma^2 t^2} = \sum_{k=0}^{\infty} \frac{(2k)!}{2^k k!}\, \sigma^{2k}\, \frac{t^{2k}}{(2k)!}.$$

It follows that the odd central moments of X are zero, and that the even central moments of X are given by

$$\nu_{2k} = \frac{(2k)!}{2^k k!}\sigma^{2k} = (1 \times 3 \times 5 \times \cdots \times (2k-1))\, \sigma^{2k}$$

In particular, $\nu_2 = \sigma^2$ and $\nu_4 = 3\sigma^4$.

Laplace transforms

The *Laplace transform* of a random variable X is defined as

$$L_X(t) = M_X(-t) = \mathbb{E}(e^{-tX}).$$

The Laplace transform exists for all $t > 0$ whenever $\mathbb{P}(X \geq 0) = 1$, or more generally $\mathbb{P}(X \geq c) = 1$ for some $c > -\infty$. This transform is very popular and important in engineering and financial applications. It has very similar properties to the mgf with the obvious changes, for example $L'_X(0) = -\mu_X$.

Recovering the distribution from the mgf

Remember the mgf of X uniquely determines its distribution. But how do we determine the distribution, given the mgf?

The easiest case is where we can recognise the form of the mgf. For example $M_X(t) = e^{t^2}$, which tells you $X \stackrel{d}{=} N(0, 2)$.

If X is integer valued we can set $P_X(z) = M_X(\log z)$ and then recover $\mathbb{P}(X = k)$ from P_X by differentiation. If X is not discrete there is no simple universal "inversion" formula for the mgf, but there is one for the related *characteristic function*.

The characteristic function is defined by

$$g_X(t) = \mathbb{E}(e^{itX}), \quad i = \sqrt{-1}$$

Its advantage is that it is defined for all $t \in \mathbb{R}$ for all random variables as $|e^{itX}| = 1$. If X is a continuous random variable with pdf f_X then the characteristic function of X:

$$g_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) dx$$

is the *Fourier transform* of the pdf f_X.

Jean Baptiste Joseph Fourier [21/03/1768 – 16/05/1830]

The Fourier Transform is one of the most useful and famous tools in mathematics. In a sense, it is inverse to itself:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} g_X(t) dt.$$

This "inversion formula" allows us to recover the distribution from the characteristic function.

Since $e^{-itX} = \cos(tX) - i\sin(tX)$, the characteristic function allows us to interpret f_X as a mixture of "harmonic oscillations" at different frequencies where $g_X(t)$ gives the coefficient of the harmonic at frequency t. Again the properties of the characteristic function are very similar to the mgf. For example

$$g_X(0) = 1, \quad g_{aX+b}(t) = e^{ibt} g_X(at).$$

ch.f. vs mgf

- No assumptions of moments needed.
- We can recover the pdf through an "inversion formula".
- We can prove the central limit theorem and other limit theorems without having to worry about the existence of moments.

Limiting distributions (Ghahramani 11.5)

One of the main applications for generating and characteristic functions is deriving approximations to the distributions of complicated random variables - for example sums

$$S_n = X_1 + X_2 + \ldots + X_n$$

and averages

$$\overline{X} = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

This application follows from the fact that not only does the mgf uniquely determine the distribution, but also when the mgf $M_X(t)$ is "close" to the mgf $M_Y(t)$ then the distribution of X will be "close" to that of Y. We make this notion more precise by defining convergence in distribution.

Convergence in distribution

Definition

We say a sequence of random variables $\{Y_n\}$ converges to the random variable Y in distribution, and write $Y_n \xrightarrow{d} Y$ if

$$F_{Y_n}(x) \to F_Y(x) \quad \text{as } n \to \infty$$

for all x such that $F_Y(x) = F_Y(x-0)$ (that is for all x where there is no jump in F_Y at x).

We do not give a proof, but to establish that $Y_n \xrightarrow{d} Y$ it suffices to show that

$$M_{Y_n}(t) \to M_Y(t)$$

for all $t \in T$.

Law of Large Numbers (Ghahramani 11.4)

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}(X_i) = \mu$, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then

$$\frac{S_n}{n} \xrightarrow{d} \mu.$$

Note that, this makes sense only if we interpret the right hand side as a random variable which is constant with probability one.

We prove the Law of Large Numbers as follows:

$$\begin{aligned}
M_{\frac{S_n}{n}}(t) &= M_{S_n}\left(\frac{t}{n}\right) \\
&= \left[M_X\left(\frac{t}{n}\right)\right]^n \\
&= \left(M_X(0) + \frac{t}{n}M'_X(0) + \ldots\right)^n \\
&\approx \left(1 + \frac{\mu t}{n}\right)^n \\
&\to e^{\mu t} \\
&= M_\mu(t)
\end{aligned}$$

where $M_\mu(t)$ is the mgf of the constant μ.

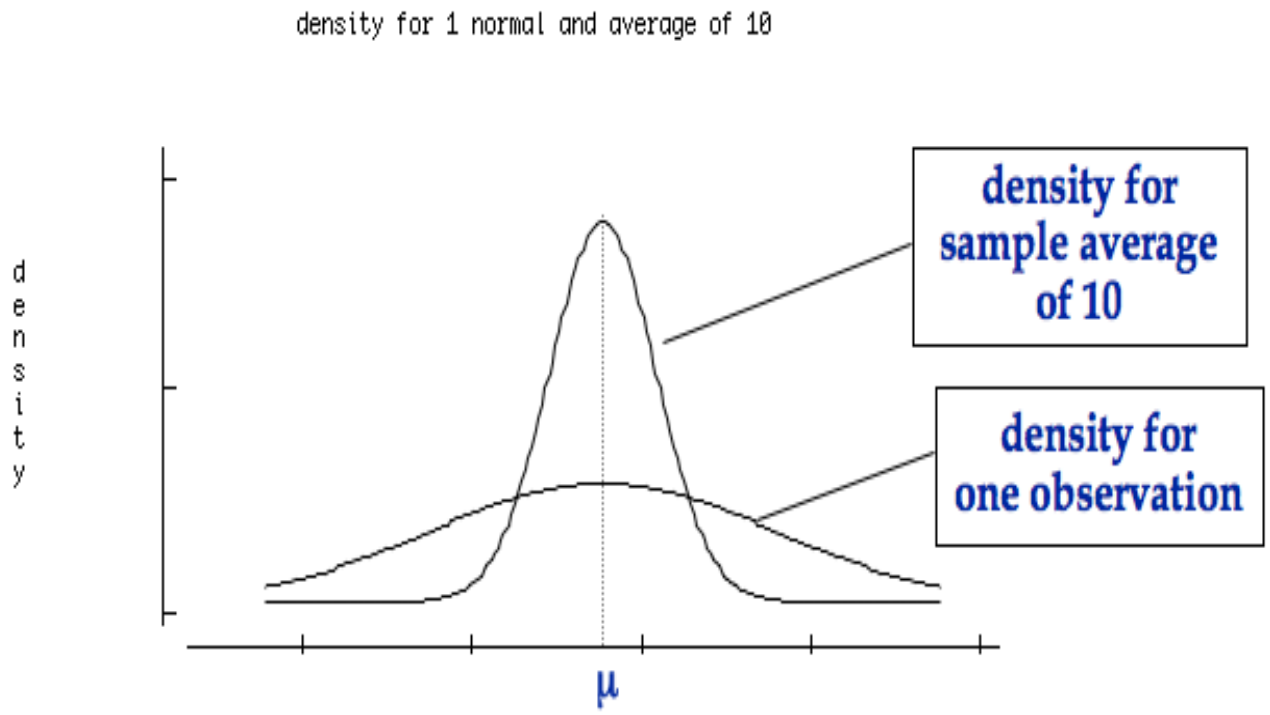
The Law of Large Numbers shows us that the random variable $\bar{X} = \frac{S_n}{n}$ is a good "estimator" of μ as it lies near to μ with high probability as $n \to \infty$. In fact

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

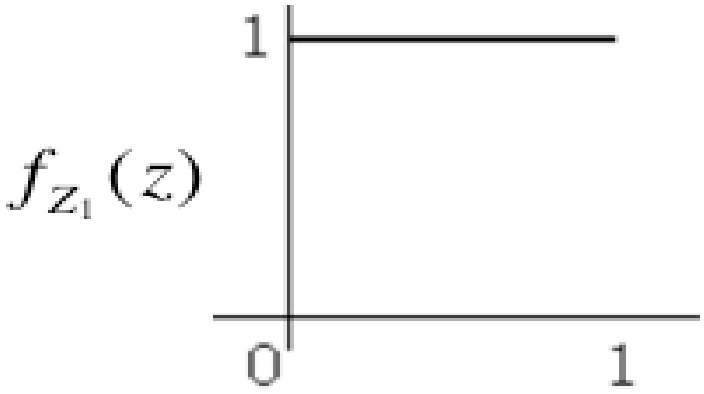
We can be much more precise, however, about the way this convergence occurs.

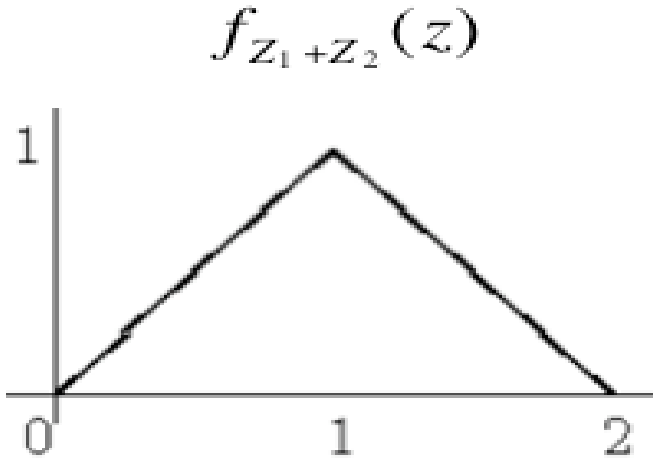
Central Limit Theorem (Ghahramani 11.5)

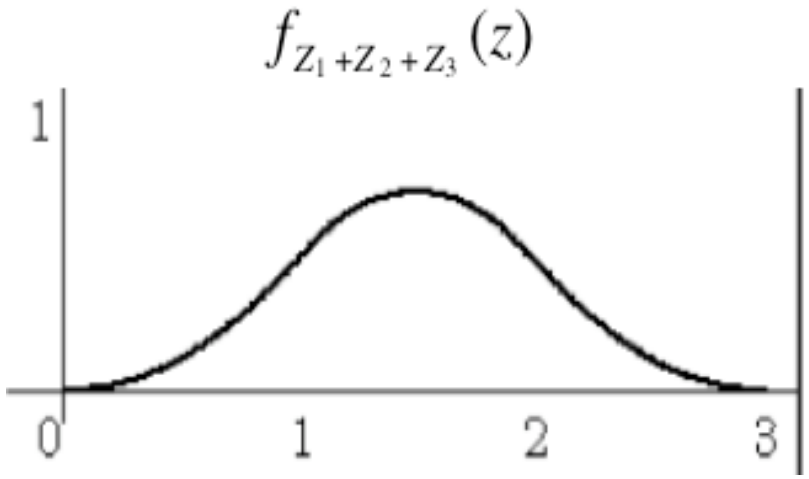
If the observations are normally distributed, then so is the sample mean:



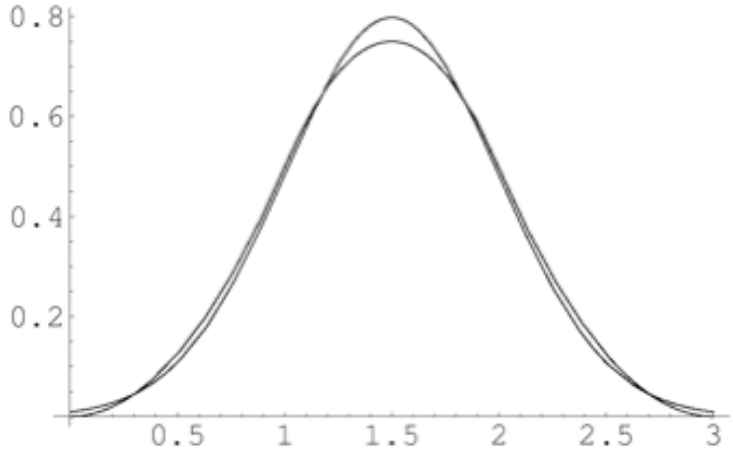
Random sample from R(0,1)







$f_{Z_1+Z_2+Z_3}$ with the normal pdf of same mean and var



[Slide 478]

Central Limit Theorem:

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $V(X_i) = \sigma^2$, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty.$$

In other words for large n, $S_n \overset{d}{\approx} N(n\mu, n\sigma^2)$ or $\overline{X} \overset{d}{\approx} N(\mu, \frac{\sigma^2}{n})$.

We prove the Central Limit Theorem as follows:

$$
\begin{aligned}
M_{Z_n}(t) &= M_{S_n - n\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\
&= \left[M_{X-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\
&= \left(M_{X-\mu}(0) + \frac{t}{\sigma\sqrt{n}} M'_{X-\mu}(0)\right. \\
&\qquad \left. + \left(\frac{t}{\sigma\sqrt{n}}\right)^2 \frac{M''_{X-\mu}(0)}{2} + \ldots\right)^n \\
&\approx \left(1 + \frac{\sigma^2}{2} \times \frac{t^2}{\sigma^2 n}\right)^n \\
&\to e^{\frac{t^2}{2}}
\end{aligned}
$$

where $e^{\frac{t^2}{2}}$ is the mgf of $N(0,1)$.

Example - Sum of exponential random variables

If $X_1, X_2, \ldots, X_{100}$ are independent random variables distributed as $\exp(1)$, so that $\mu = 1$, $\sigma^2 = 1$, and $T = X_1 + X_2 + \cdots + X_{100}$ then $T \overset{d}{\approx} N(100, 100)$ so that

$$\mathbb{P}(90 < T < 110) \simeq \mathbb{P}(-1 < T_s < 1) = 0.6828.$$

Sum of Bernoulli random variables

If $T = X_1 + X_2 + \cdots + X_n$ is the sum of n independent Bernoulli random variables each with parameter p then $\mu_X = p$, $\sigma_X^2 = p(1-p)$ and by the Central Limit Theorem

$$T \overset{d}{\approx} N(n\mu_X, n\sigma_X^2) = N(np, np(1-p)).$$

Of course we have the exact result $T \overset{d}{=} \text{Bi}(n, p)$, so we have the approximation:

$$\text{Bi}(n, p) \overset{d}{\approx} N(np, np(1-p)).$$

Example

The Australian Bureau of Statistics shows that the distribution of household size is as follows:

size	1	2	3	4	5	6	7
prob	0.237	0.317	0.178	0.157	0.07	0.026	0.015

Find the approximate probability that the sample average of 100 households exceeds 3.

($\mu_X = 2.644$, $\sigma_X^2 = 2.050$)

Stochastic Processes (Ghahramani Ch 12)

A *stochastic process* is a sequence of random variables, $X(t), t \in T$.

For example,

1. $X(t)$ could be the number of individuals in a population at time t.
2. $X(t)$ could be the number of t^{th} generation individuals in a population.
3. $X(t)$ could be the air pressure at a particular location at time t.

4. $X(t)$ could be the stock market closing price of BHP shares on day t.

5. $X(t)$ number of flaws in a length t of twine.

6. $X(t)$ number of organisms in a volume t of fluid.

For each $t \in T$, $X(t)$ takes values in a set S, called the *state space* of the stochastic process. The set T is called the *index set*, with t usually denoting time though, as indicated in the above examples, it might denote a spatial variable.

The word *stochastic* is of Greek origin and is concerned with time and chance.

Exercise

Specify the state space S and the index set T for each of the stochastic processes described above.

Classification of Stochastic Processes

The state space may be either discrete or continuous, depending on whether $X(t)$ is a discrete or a continuous random variable. For example, if $S = \{0, 1, 2, \ldots\}$, then it is discrete and if $S = \{x : 0 \leq x < \infty\}$, then it is continuous.

In this course, we shall consider only the case where S is discrete.

The index set may also be either discrete or continuous. If $T = \{0, 1, 2, \ldots\}$ then we say that $X(t)$ is a discrete-time stochastic process. If $T = \{t : 0 \leq x < \infty\}$ or $T = [a, b]$ then we say that $X(t)$ is a continuous-time stochastic process.

It is common practice to denote a discrete-time stochastic process by $\{X_n, n = 0, 1, 2, \ldots\}$, while using $\{X(t), t \geq 0\}$ or $\{X(t), t \in [a, b]\}$ for a continuous-time stochastic process.

Sequences of Bernoulli Trials

A sequence of Bernoulli trials is a discrete-state, discrete-time stochastic process.

Here X_n is the number of successes after n trials, the state space for the stochastic process $\{X_n\}$ is given by $S = \mathbb{Z}_+$ and the index set is given by $T = \mathbb{Z}_+$.

Poisson processes (Ghahramani 5.2, 12.2)

The discrete-state, continuous-time analogue of a sequence of Bernoulli trials is called the Poisson process, which we have already touched on in Chapter 2. The Poisson process is a process in which points occur "randomly" in continuous time.

The Poisson process is one of the most important concepts in all of applied probability - models for a myriad of applications are built upon it.

For the Poisson process, $X(t)$ is the number of "points" that occur in time $[0, t]$, the state space for the stochastic process $\{X(t)\}$ is given by $S = \mathbb{Z}_+$ and the index set is given by $T = \mathbb{R}_+$.

Using the same arguments as in Chapter 2, we see that in a Poisson process

1. The number of events in $[0, t]$ has a Poisson distribution with parameter αt,
2. The waiting time until the first event has an exponential distribution with parameter α, and
3. The waiting time until the rth event has a gamma distribution with parameters r and α.

Exercise - go back to Chapter 2 and work through these arguments. Hint: For (1), take a given time t, split the interval $[0, t]$ into n subintervals of length t/n. In each interval, the probability of success is $\alpha t/n$. Then shrink the intervals by letting $n \to \infty$.

Poisson processes can be used to model

1. the arrival of radioactive particles at a Geiger counter.
2. the passing of cars on a minor road.
3. the occurrence of accidents in a factory.
4. the arrival of calls at a telephone exchange.
5. the occurrence of flaws in a manufactured substance.

Discrete-time Markov Chains (Ghahramani 12.3)

Suppose that $X_1, \ldots, X_n$ are discrete rv's and $x_1, \ldots, x_n$ are numbers. Then, the joint pmf is

$$\mathbb{P}(X_1 = x_1 \cap \cdots \cap X_n = x_n)$$

To compute this prob, we may use

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1 \cap \cdots \cap X_n = x_n) \\
&= \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1} \cap \cdots \cap X_1 = x_1) \\
&\quad \times \mathbb{P}(X_{n-1} = x_{n-1} \cap \cdots \cap X_1 = x_1) \\
&= \cdots \\
&= \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1} \cap \cdots \cap X_1 = x_1) \\
&\quad \times \mathbb{P}(X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2} \cap \cdots \cap X_1 = x_1) \\
&\quad \times \cdots \times \mathbb{P}(X_1 = x_1)
\end{aligned}$$

Markov chain

- When $X_1, \ldots, X_n$ are iid
- The next simplest case is that of a Markov chain, where the conditional probabilities only depend on the value of the previous rv and are the same for all time.

Andrey (Andrei) Andreyevich
[14/06/1856 - 20/07/1922]

Andrei Andreevich Markov

- 1860s schooling at St Petersburg Gymnasium No 5: outstanding talents for math but poor in everything else
- 1874: University of Petersburg (under Chebyshev)
- 1878: gold medal for his scientific work
- Research areas: number theory, the approximation of functions, the problem of moments, the calculus of finite differences, etc.
- Theory of chains of stochastic processes (Markov chains)

A discrete-time Markov chain is a stochastic process with index set $T = \mathbb{Z}_+$ and a countable state space. Since S is countable, for notational ease, we can consider it to be a subset of $\mathbb{Z}_+$.

Let $i \in S$. Then $\mathbb{P}(X_n = i)$ is the probability that the state of the process at time point n is i. The process is said to be a *Markov chain* if

$$\mathbb{P}\left(X_{n+1} = j | X_0 = i_0 \cap \ldots \cap X_n = i_n\right) = \mathbb{P}\left(X_{n+1} = j | X_n = i_n\right).$$

That is, the future behaviour of the process depends on its history only through its present state.

- $\mathbb{P}\left(X_{n+1} = j | X_n = i_n\right)$ is called the transition probability

A discrete-time Markov chain is *homogeneous* or *time-homogeneous* if it also has the property that, for all i, j, m and n,

$$\mathbb{P}\left(X_{n+m} = j \middle| X_n = i\right) = \mathbb{P}\left(X_m = j \middle| X_0 = i\right).$$

That is, the changes in the Markov chain depend on the number m of time steps and not on the starting time n.

Transition Matrix

For a homogeneous discrete-time Markov chain, we can define

$$\begin{aligned} P_{ij} &= \mathbb{P}\left(X_{n+1} = j \middle| X_n = i\right) \\ P &= \left[P_{ij}\right]_{i,j \in S}. \end{aligned}$$

In particular, when $S = \{0, 1, 2, \ldots, m\}$, we have

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \ldots & P_{0m} \\ P_{10} & P_{11} & P_{12} & \ldots & P_{1m} \\ . & . & . & . & . \\ P_{m0} & P_{m1} & P_{m2} & \ldots & P_{mm} \end{bmatrix}$$

In general

We define

$$\begin{aligned} P_{ij}^{(m)} &= \mathbb{P}\left(X_{n+m} = j \middle| X_n = i\right) \\ P_{ij} &= P_{ij}^{(1)} \\ P^{(m)} &= \left[P_{ij}^{(m)}\right]_{i,j \in S} \\ P &= P^{(1)}. \end{aligned}$$

Usually we know P, and our task is to find $P^{(m)}$.

Properties of transition matrices

- Every entry is non-negative
- Every row sums to 1

  Proof

- A square matrix with these properties is called a *stochastic matrix*.

Theorem

$$P^{(m)} = P^m.$$

Proof

We use induction on m. The result is obviously true for $m = 1$. Assume it is true for $m - 1$, that is $P^{(m-1)} = P^{m-1}$.

Condition $P_{ij}^{(m)}$ on the state after $m-1$ steps. Thus

$$\begin{aligned}
P_{ij}^{(m)} &= \mathbb{P}\left(X_m = j | X_0 = i\right) \\
&= \sum_k \mathbb{P}\left(X_m = j | X_{m-1} = k, X_0 = i\right) \mathbb{P}\left(X_{m-1} = k | X_0 = i\right) \\
&= \sum_k \mathbb{P}\left(X_m = j | X_{m-1} = k\right) \mathbb{P}\left(X_{m-1} = k | X_0 = i\right) \\
&= \sum_k P_{kj} P_{ik}^{(m-1)} \\
&= \sum_k P_{kj} \left[P^{m-1}\right]_{ik} \\
&= \left[P^m\right]_{ij} .
\end{aligned}$$

Example - spread of disease

- Population size 3.
- During any single period of time, 2 people are selected at random to interact
- Transmission occurs with prob 0.1 if one is diseased and the other is healthy
- X_n is the number of diseased people at the end of time n

Example - number of patients

Let $Y_1, Y_2, \ldots$ be the numbers of patients arriving at a hospital on days $1, 2, \ldots$ and suppose the Y_k's are iid with $\mathbb{P}(Y_k = j) = 2^{-j-1},\ j = 0,\ 1,\ 2,\ \ldots$ If we let X_n be the total number of arrivals after n days, then $X_n = \sum_{0 \le i \le n} Y_i$ and $\{X_n;\ n = 0,\ 1,\ 2,\ \ldots\}$ $(X_0 = 0)$ is a Markov chain.

Example

If

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix},$$

what is $P_{12}^{(3)}$?

We know that $P^{(3)} = P^3$ and so

$$
\begin{aligned}
P_{12}^{(3)} &= \left[P^3\right]_{12} \\
&= \left[\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}^3\right]_{12} \\
&= \begin{bmatrix} \frac{7}{32} & \frac{13}{32} & \frac{3}{8} \\ \frac{15}{64} & \frac{25}{64} & \frac{3}{8} \\ \frac{11}{32} & \frac{13}{32} & \frac{1}{4} \end{bmatrix}_{12} \\
&= \frac{13}{32}.
\end{aligned}
$$

Example

If

$$P = \begin{pmatrix} 0.28 & 0.72 \\ 0.08 & 0.92 \end{pmatrix},$$

then we find

$$P^2 = \begin{pmatrix} 0.136 & 0.864 \\ 0.096 & 0.904 \end{pmatrix}$$

Furthermore

$$P^3 = \begin{pmatrix} 0.1072 & 0.8928 \\ 0.0992 & 0.9008 \end{pmatrix},$$

$$P^4 = \begin{pmatrix} 0.1014 & 0.8986 \\ 0.0998 & 0.9002 \end{pmatrix},$$

$$P^5 = \begin{pmatrix} 0.1003 & 0.8997 \\ 0.1000 & 0.9000 \end{pmatrix}.$$

We see that

$$P^n \to \begin{pmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{pmatrix}.$$

as $n \to \infty$.

Distribution of X_1

The law of total probability gives

$$\mathbb{P}(X_1 = y) = \sum_{\text{values of } X_0} \mathbb{P}(X_1 = y | X_0 = x)\mathbb{P}(X_0 = x).$$

- The distribution of X_0 is often called the initial distribution

Notation

- To simplify our computation, we introduce $\vec{\pi}_n = (\mathbb{P}(X_n = 0), \mathbb{P}(X_n = 1), \ldots)$ as the vector containing the probability mass function of X_n.
- The initial distribution is $\vec{\pi}_0$
- The distribution of X_1 can be computed by

$$\vec{\pi}_1 = \vec{\pi}_0 \, P.$$

- More generally,

$$\vec{\pi}_n = \vec{\pi}_0 \, P^n.$$

Proof

By the law of total probability,

$$\mathbb{P}(X_n = i) = \sum_k \mathbb{P}(X_0 = k)\mathbb{P}(X_n = i | X_0 = k).$$

Using our notation defined above, this can be rewritten as

$$\mathbb{P}(X_n = i) = \vec{\pi}_0 \left[P^{(n)}\right]_i,$$

where $\left[P^{(n)}\right]_i$ is the ith column of the matrix $P^{(n)}$.

Now, since $P^{(n)} = P^n$, we can rewrite this as

$$\vec{\pi}_n = \vec{\pi}_0 P^n.$$

Example

In our previous example assuming $\vec{\pi}_0 = (1, 0)$ and

$$P = \begin{pmatrix} 0.28 & 0.72 \\ 0.08 & 0.92 \end{pmatrix},$$

$$\begin{aligned}
\vec{\pi}_1 &= (0.28, 0.72) \\
\vec{\pi}_2 &= (0.136, 0.864) \\
\vec{\pi}_3 &= (0.1072, 0.8928) \\
\vec{\pi}_4 &= (0.1014, 0.8986) \\
\vec{\pi}_5 &= (0.1003, 0.8997).
\end{aligned}$$

Whatever probability vector we start with, we see that $\overrightarrow{\pi}_5 \to (0.1, 0.9)$ as $n \to \infty$. You can check this yourself.

Thus if n is large, the effect of the initial condition wears off: no matter where the process started, for large n the probability that the state will be 1 at time point n is close to 0.1.

Markov chains forget

- Markov chains forget the distant past, in general
- Regardless of the initial distribution, the probabilities for the states generally settle down

Long-run distribution

- We shall assume that the state space S is finite.
- We shall also assume that

$$P_{ij}^{(m)} \to \pi_j$$

as $m \to \infty$. That is, the probability of being in state j after many steps is independent of the initial state.

- The vector $\vec{\pi} = (\pi_1, \pi_2, \ldots)$ is known as the equilibrium distribution or long-run distribution or stationary distribution.
- Note that this property does not always occur. More details in MAST30001 Stochastic Modelling

Letting $m \to \infty$ in the equation

$$P_{ij}^{(m)} = \sum_k P_{ik}^{(m-1)} P_{kj},$$

we see that

$$\pi_j = \sum_k \pi_k P_{kj}, \quad \text{for all } j.$$

These equations are known as the equilibrium equations for a Markov Chain. They can be written as

$$\vec{\pi} = \vec{\pi}\, P.$$

We see that $\vec{\pi}$ is a *left-eigenvector* of P associated with the eigenvalue 1.

What does the equilibrium distribution mean?

The limiting distribution has more than one interpretation.

It can be seen as

1. limiting,
2. stationary, or
3. ergodic.

The limiting interpretation

By definition

$$\pi_j = \lim_{n \to \infty} P_{ij}^{(n)}$$

and so π_j is the *limiting probability* of the process being in state j. This means that so long as the process has been going for quite a long time, the probability that the process is in state j will be approximately π_j.

The stationary interpretation

We showed that

$$\vec{\pi}\, P = \vec{\pi}$$

and so $\vec{\pi}$ has a *stationary interpretation*. If the process starts with probability distribution $\vec{\pi}$ it will persist with that distribution forever.

The ergodic interpretation

It can be shown that there is also an *ergodic interpretation*. This means that for sample paths of the process with probability one, the proportion of time that the process spends in state j is π_j.

Summary

- Interpretation 1: $\overrightarrow{\pi}$ gives limiting distribution of the process
- Interpretation 2: Start with $\overrightarrow{\pi}$ and stay there
- Interpretation 3: $\overrightarrow{\pi}$ gives long term relative frequencies for values of rv's

Example

Consider a three-state Markov chain with transition probability matrix

$$P = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

The equilibrium distribution for this Markov chain is given by solving

$$\vec{\pi}\ P = \vec{\pi}$$

which is the same as

$$\begin{aligned} 0.5\pi_1 + 0.2\pi_2 + 0.1\pi_3 &= \pi_1 \\ 0.4\pi_1 + 0.4\pi_2 + 0.3\pi_3 &= \pi_2 \\ 0.1\pi_1 + 0.4\pi_2 + 0.6\pi_3 &= \pi_3. \end{aligned}$$

One of the three equations is redundant.

To see this, add the equations. The result is $\pi_1 + \pi_2 + \pi_3 = \pi_1 + \pi_2 + \pi_3$, which shows that the columns are linearly dependent.

This is always the case. For an m-state Markov chain, $\vec{\pi}\, P = \vec{\pi}$ provides at most $m - 1$ linearly independent equations.

The mth equation, which we need in order to solve for $\overrightarrow{\pi}$ is provided by

$$\pi_1 + \pi_2 + \pi_3 = 1$$

and this must hold since $\overrightarrow{\pi}$ defines a distribution.

This gives us three equations in three unknowns which we can solve. The result in this case is

$$\pi_1 = 12/53 \approx 0.226, \quad \pi_2 = 19/53 \approx 0.359, \quad \pi_3 = 22/53 \approx 0.415.$$

Solution of the equilibrium equations

In general, when there are N states, the equation

$$\vec{\pi}\, P = \vec{\pi}$$

is a system of N linear equations in N unknowns. In the cases that we shall consider, there will be exactly one redundant equation. The normalising equation will supply the extra equation that we need and so there are N linearly independent equations in N unknowns.

The unique solution can be found by any of the usual methods, such as Gauss-Jordan reduction. We need to be careful to transpose the matrix P before we do this.

Example

Consider a manufacturing process which produces items classified as 1 = defective, or 2 = non-defective. Suppose that the probability of a defective item following a defective item is 0.81, while if a non-defective item is produced the probability that the next item is also non-defective is 0.99.

This process can be modelled as a discrete-time Markov chain with transition matrix

$$P = \begin{pmatrix} 0.81 & 0.19 \\ 0.01 & 0.99 \end{pmatrix}.$$

What is the long-term proportion of defective items?

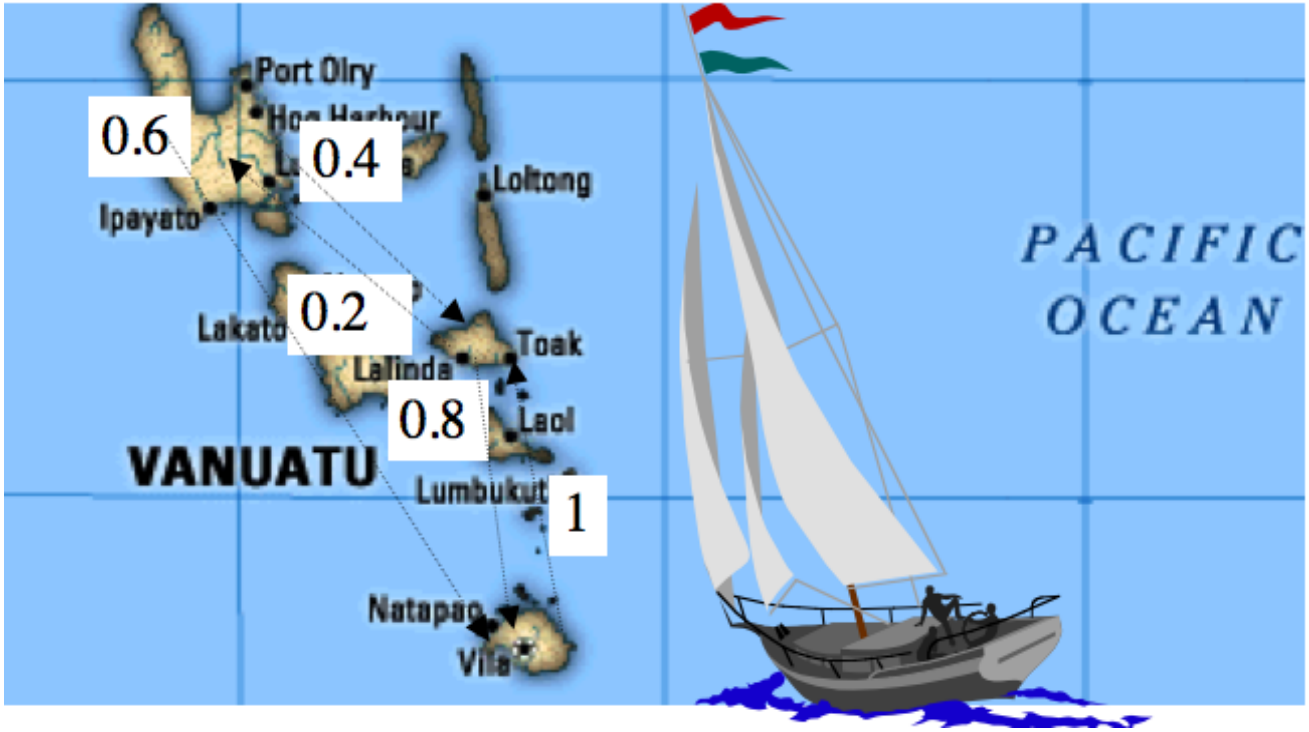
State Transition Graphs

The behaviour of the Markov chain can often be usefully visualised using a state transition graph.

- Change of state is represented by a circle
- arrows for non-zero probability of a transition
- The transition probabilities are added to the arrows.

To remember state space diagrams

Imagine a yachtsman hopping from one Pacific island to another represented by the circles on the state diagram. The yachtsman chooses the next island to travel to according to the probabilities on the arrows.



Example

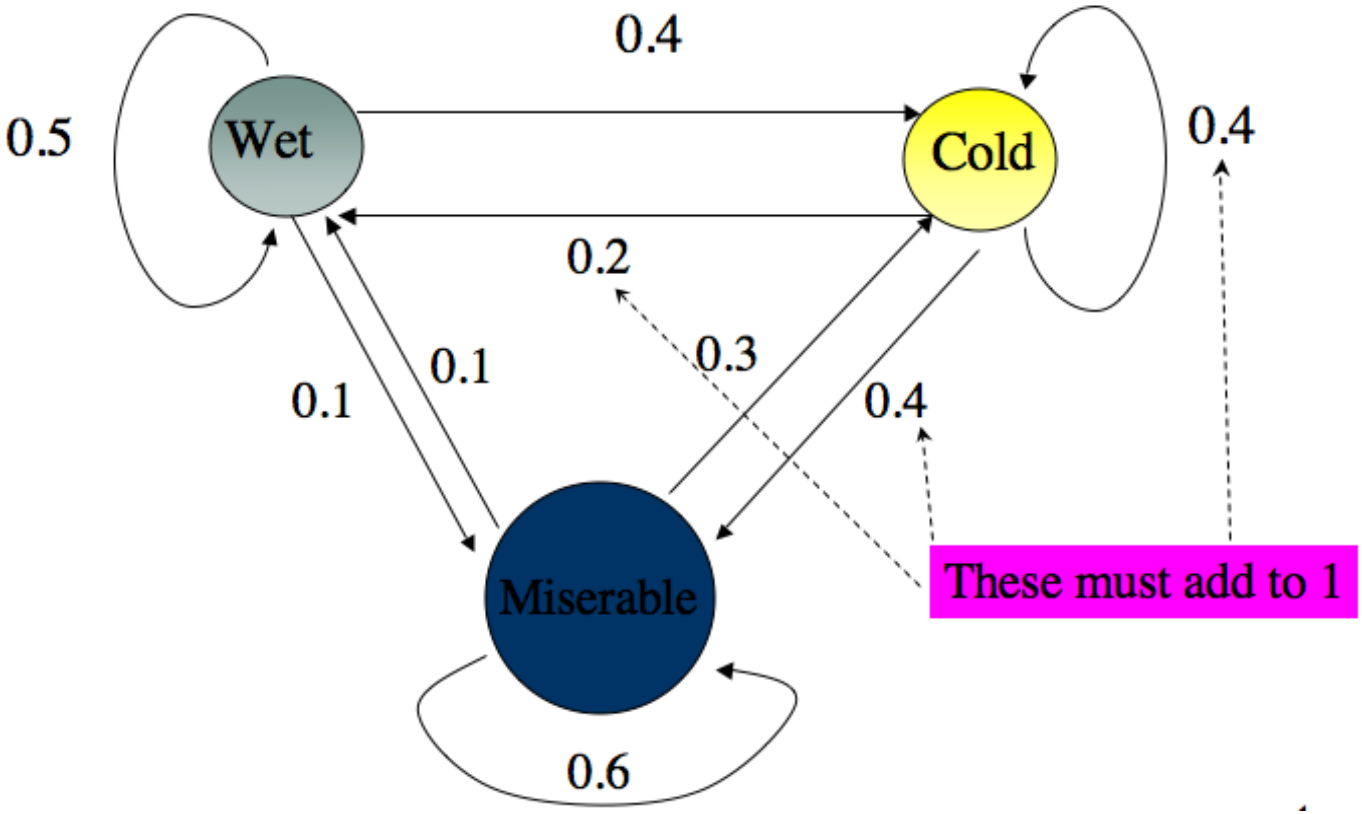
The weather in Markovograd is in one of three states: 1 = "wet", 2 ="cold" or 3 = "miserable". The state of the weather is well described by a Markov chain model with transition probability matrix

$$P = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

If it is wet on Monday, what is the probability that it is wet again on Tuesday and Thursday?

What is the long-term proportion of wet days?.

State space diagram - Weather



Example

Consider a model for road surface deterioration: the state of thenroad surface is either 1 = "good", 2 = "moderate" or 3 = "needs replacing". Suppose that the state of the road after n years is described by a Markov chain model with

$$P = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{pmatrix}.$$

If the road starts out good in year 0, find the probability that it needs replacement after 5, 10, 15 years?

The state transition diagram for this Markov chain is shown in Figure 4.2 of the course notes.

Calculation — preferably using a computer — gives

$$P^5 = \begin{pmatrix} 0.59049 & 0.26281 & 0.14670 \\ 0 & 0.32768 & 0.67232 \\ 0 & 0 & 1 \end{pmatrix},$$

$$P^{10} = \begin{pmatrix} 0.34868 & 0.24130 & 0.41002 \\ 0 & 0.10737 & 0.89263 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$P^{15} = \begin{pmatrix} 0.20589 & 0.17071 & 0.62340 \\ 0 & 0.03518 & 0.96482 \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore the probability that the road needs replacement after five years is given by $[P^5]_{13} = 0.14670$ and similarly the probability that the road needs replacement after ten years is 0.41002, and after fifteen years 0.62340.

We can show that

$$P^n \to \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

and so the road will eventually need replacement. In this case the equilibrium probabilities are not very interesting. It is the length of time taken to get there is of interest.

Non-Markov Stochastic Process

Insurance claims for household theft arise in each year with a probability which depends on the suburb: in Carlton there is a claim with probability 0.05 and in Collingwood the probability is 0.1. We assume that for a given house, the claims for successive years are independent.

If a house is chosen at random (with probability 0.5 for Carlton and 0.5 for Collingwood), show that the successive claims for that house are not Markovian.

- Intuitively, learning that there were claims in both of two years makes it more likely that the randomly chosen house is in Collingwood, the high burglary suburb, than learning that there was a claim in one year.
- The Markov property does not permit this kind of learning from the values of the stochastic process.
- Let $X_0, X_1, \ldots$ be rv's which are 1/0 if the 1st, 2nd, ... years in the chosen household do/do not have claims. Let S be 1/0 if the suburb chosen is Carlton/Collingwood.

Example

- A population consisting of individuals able to produce offspring of the same kind.
- Each individual will, by the end of its lifetime, have produced j new offspring with probability

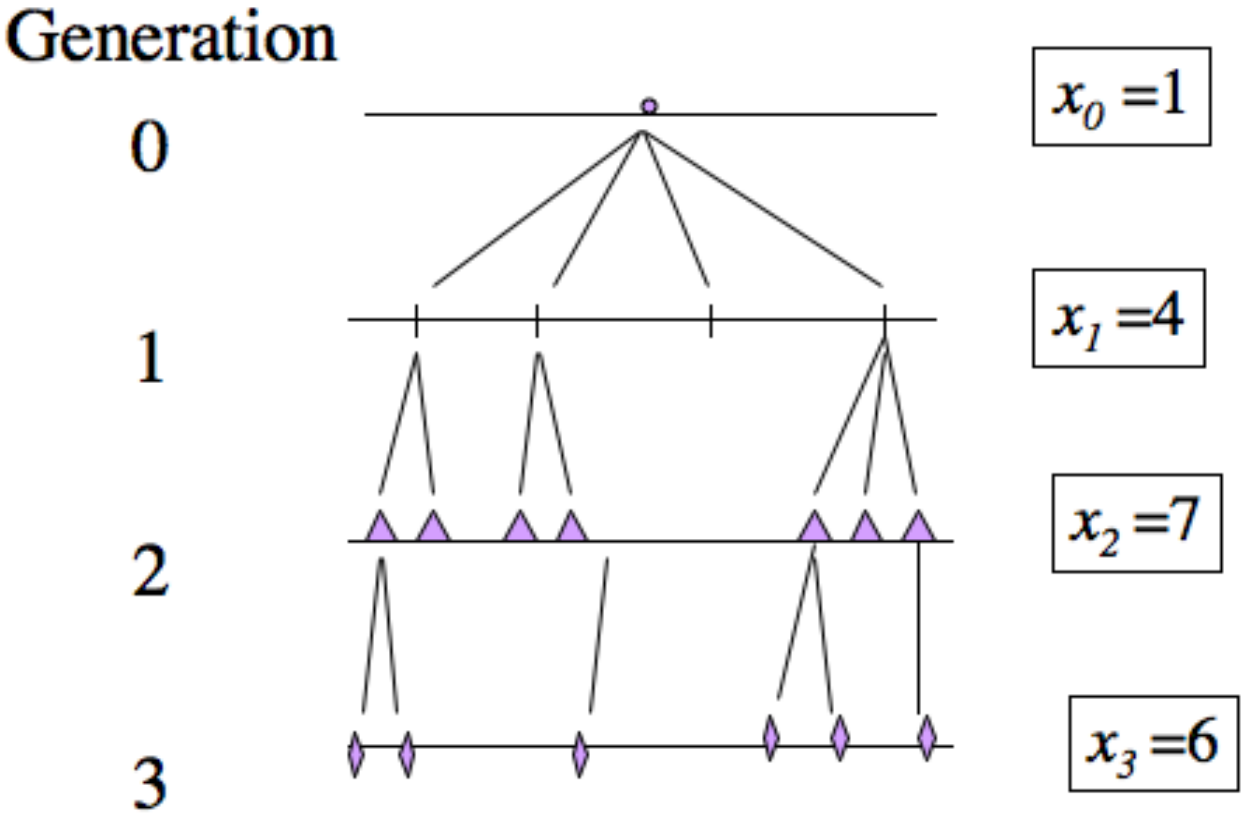
$$a_j, \quad j = 0, 1, 2, \ldots$$

indeptly of the number produced by any other individual

- Assume finite mean μ and variance σ^2 for the offspring distribution

- The number of individuals initially present, denoted by X_0, is called the size of the zeroth generation
  - We generally start with a single individual in the 0th generation so $\mathbb{P}(X_0 = 1) = 1$.
- All offspring of the zeroth generation constitute the first generation and their number is denoted by X_1
- In general, let X_n denote the (random) size of the n^{th} generation

A realisation



- Let $N_i(n)$ equal the number of offspring produced by the ith individual in the nth generation. We assume all the $N_i(n)$s are independent and identically distributed. The common offspring distribution N has pmf

$$\mathbb{P}(N_i(n) = j) = a_j, \quad j = 0, 1, 2, \ldots$$

- The population sizes of successive generations are related by

$$\begin{aligned} X_{n+1} &= N_1(n) + N_2(n) + \ldots + N_{X_n}(n) \\ &= \sum_{k=1}^{X_n} N_k(n), \quad n = 0, 1, 2, \ldots \end{aligned}$$

  - These equations merely reflect the fact that the number of individuals in the $(n+1)$st generation is equal to the sum of the offspring of all the individuals in the nth generation.

Branching Process

- **Definition** $\{X_n,\ n = 0,\ 1,\ 2,\ \ldots\}$ is a Markov chain and it is called a branching process.
- Galton, F. and Watson, H.W. (1874). On the probability of extinction of families, J.R. Anthropol. Inst. 4, 138-144.
- Branching processes have been applied to
  - growth of animal, plant and insect populations
  - neutron chain reactions
  - survival of mutant genes
  - extinction of family surnames
  - phylogenetic trees

Mean and variance of population size

From the fundamental equations

$$\mathbb{E}(X_{n+1}|X_n = x_n) = x_n\mu \quad \text{hence} \quad \mathbb{E}(X_{n+1}|X_n) = \mu X_n.$$
$$V(X_{n+1}|X_n = x_n) = x_n\sigma^2 \quad \text{hence} \quad V(X_{n+1}|X_n) = \sigma^2 X_n.$$

It follows that

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(\mathbb{E}(X_{n+1}|X_n)) = \mathbb{E}(\mu X_n) = \mu\mathbb{E}(X_n).$$

As $\mathbb{E}(X_0)$=1 we obtain $\mathbb{E}(X_n) = \mu^n, \quad n = 0, 1, \ldots.$

For the variance we have

$$\begin{aligned} V(X_{n+1}) &= \mathbb{E}(V(X_{n+1}|X_n)) + V(\mathbb{E}(X_{n+1}|X_n)) \\ &= \mathbb{E}(\sigma^2 X_n) + V(\mu X_n) \\ &= \sigma^2 \mu^n + \mu^2 V(X_n) \quad n = 0, 1, 2, \ldots \end{aligned}$$

Since $V(X_0) = 0$ we obtain

$$\begin{aligned} V(X_n) &= \sigma^2 \mu^{n-1} \left(1 + \mu + \ldots + \mu^{n-1}\right) \\ &= \begin{cases} \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1} & \mu \neq 1 \\ n\sigma^2 & \mu = 1. \end{cases} \end{aligned}$$

Pgf for population size

- We denote the offspring pgf by $A(z) = \sum_{k=0}^{\infty} a_k z^k$.
- We use the abbreviated notation $P_n(z)$ to denote $P_{X_n}(z)$ the pgf of X_n. Since $\mathbb{E}(z^{X_{n+1}} | X_n = x_n) = A(z)^{x_n}$ we have

$$P_{n+1}(z) = \mathbb{E}(\mathbb{E}(z^{X_{n+1}} | X_n)) = \mathbb{E}(A(z)^{X_n}) = P_n(A(z)).$$

- Iterating this recurrence relation we obtain

$$P_n(z) = P_{n-1}(A(z)) = P_{n-2}(A(A(z))) = \ldots = P_0(A^{[n]}(z)).$$

where $A^{[n]}(z)$ is the nth iterate of the pgf A.

As $P_0(z) = z$

$$P_n(z) = A^{[n]}(z).$$

Hence also $P_{n+1}(z) = A^{[n+1]}(z) = A(A^{[n]}(z)) = A(P_n(z))$.

Probability of extinction

- Let D be the generation at which extinction occurs
- $q_n = \mathbb{P}(D \leq n) = \mathbb{P}(X_n = 0) = P_n(0)$.
  - The sequence $\{q_n\}$ is non decreasing (since $\{X_n = 0\} \subset \{X_{n+1} = 0\}$) and is bounded above by 1, so $\lim_{n\to\infty} q_n$ must exist.
  - $q := \lim_{n\to\infty} q_n$: the prob of eventual extinction.

- Since $P_{n+1}(0) = A(P_n(0))$, we know that $q_{n+1} = A(q_n)$. Letting $n \to \infty$ we obtain

$$q = A(q).$$

  - Any solution of an equation of this type is called a *fixed point* of the function A.

Thus, to find the extinction probability q, we solve the equation $z = A(z)$. However, this equation often has multiple solutions and we need to know which is the one that we are looking for.

Theorem The extinction probability q is the *minimal nonnegative solution* of the equation $z = A(z)$.

Proof

Let r be any nonnegative solution to $z = A(z)$. We show by induction that $q_n \leq r$ for all n. Clearly $q_0 = 0 \leq r$. Assume that $q_m \leq r$. Then

$$\begin{aligned} q_{m+1} &= A(q_m) \\ &= \sum_{k=0}^{\infty} a_k (q_m)^k \\ &\leq \sum_{k=0}^{\infty} a_k r^k = A(r) = r. \end{aligned}$$

Since $q_n \leq r$ for all n, then $\lim_{n\to\infty} q_n \leq r$, which implies that $q \leq r$ and so q is the minimal nonnegative solution to $z = A(z)$.

Conclusions

- If $a_0 = 0$, then $q = 0$
- If $a_0 > 0$ and $a_0 + a_1 = 1$, then $q = 1$
- If $a_0 > 0$ and $a_0 + a_1 < 1$, then
  - $q = 1$ if and only if $\mu \leq 1$
  - if $\mu > 1$, there is a solution q in $[0, 1)$, which is the minimal nonnegative solution and hence the extinction probability.

We can summarise our results in the following table

Parameters	Prob of extinction
$\mu \leq 1$ and $a_0 > 0$	$q = 1$
$\mu > 1$ and $a_0 > 0$	$0 < q < 1.$
$a_0 = 0$	$q = 0$

Example

Assume that, for $k \geq 0$, an individual has k offspring with probability $p(1-p)^k$, where $p < 1$. What is the eventual probability of extinction? Here

$$\begin{aligned} A(z) &= \sum_{k=0}^{\infty} p(1-p)^k z^k \\ &= \frac{p}{1-(1-p)z} \end{aligned}$$

for $z \leq 1/(1-p)$.

The equation $A(z) = z$ becomes

$$z = \frac{p}{1 - (1-p)z}$$

which leads to a quadratic equation with solutions $p/(1-p)$ and 1. If $p \geq 1/2$ then the minimal nonnegative solution is 1 and extinction is certain. If $p < 1/2$ then the minimal nonnegative solution is $p/(1-p)$, which is the extinction probability.

Surname extinction

Lotka (1931) found that for white males in the United States in 1920 the pgf of the number of male offspring per male was given approximately by

$$A(z) = \frac{0.482 - 0.041z}{1 - 0.559z}.$$

Consequently the equation $q = A(q)$ leads to the quadratic

$$0.559q^2 - (1 + 0.041)q + 0.482 = 0.$$

This quadratic can be factorised easily using the fact that $q = 1$ is a solution yielding the smaller positive root

$$q = \frac{0.482}{0.559} = 0.86.$$

Hence the probability of extinction of the surname for white males in the United States in 1920 was 0.86. Note from the pgf that 48.2% of males had no offspring at all (for example they may have died in infancy). This analysis also assumed that male descendants would automatically take the surname of the father.