ST202 PROBABILITY, DISTRIBUTION THEORY, AND INFERENCE
LECTURE NOTES

Matteo Barigozzi*

Michaelmas Term 2015-2016

This version April 16, 2016

Introductory Material

Course Aims and Objectives

The first part of the course aims to convey a thorough understanding of probability and distribution theory. A range of methods, related to distributions of practical importance, are taught. The course builds on material in ST102 Elementary Statistical Theory and provides the foundation for further courses in statistics and actuarial science.

The following list gives you an idea of the sort of things you will be able to do by the end of this term – it does not come close to covering everything. By the end of the first part of the course you should:

- be able to work out probabilities associated with simple (and not so simple) experiments,
- know the distinction between a random variable and an instance of a random variable,
- for any given distribution, be able to select suitable methods and use them to work out moments,
- be familiar with a large number of distributions,
- understand relationships between variables, conditioning, independence and correlation,
- feel at ease with joint distributions and conditional distributions,
- know the law of large numbers and the central limit theorem and their implications,
- be able to put together all the theory and techniques you have learned to solve practical problems.

*Office: COL7.11

Pre-requisites

Most of you will have done MA100 Mathematical Methods and ST102 Elementary Statistical Theory. If you have not taken both of these courses (or if the relevant memory has been deleted) you should (re-)familiarize yourself with their content. Below is a, far from exhaustive, list of mathematical tools that we will be using.

- **Sets**: union, intersection, complement.
- **Series**: arithmetic, geometric, Taylor.
- **Differentiation**: standard differentials, product rule, function of a function rule.
- **Integration**: standard integrals, integration by parts.

Structure of the course

During Michaelmas term each student should attend **three hours per week**, divided in lectures and seminars.

- **Lectures**: two hours per week during which I will teach you the theory of probability. The schedule for Michaelmas term is:
  - Thursday 9:00 - 10:00: in room TW1.G.01;
  - Thursday 16:00 - 17:00: in room TW1.G.01.
- **Seminars**: Dr. Miltiadis Mavrakakis will go through the solution of homework with you, attendance is compulsory and will be recorded on LSE for You, failing to attend seminars may bar you from the final exam. You are divided in three groups and the schedule is

Group 1 Friday 9:00 - 10:00 in room NAB.1.07;

Group 2 Tuesday 9:00 - 10:00 in room KSW.G.01;

Gropu 3 Tuesday 10:00 - 11:00: group 3 in room KSW.G.01;

no change of group is possible. There will be no seminars in week 1 and also in week 2 for groups 2 and 3 which will start in week 3 but with an additional session in week 11 or week 1 in LT, so that all groups have 9 sessions on MT material.

The **coursework** for Michaelmas term is made of homework, class tests, and a final exam.

- **Homework**: you will be assigned 8 weekly problem sets on
  - Thursday of weeks 1 to 4 and 7 to 9 to be returned on the next Wednesday by 12:00;

collection boxes are per GROUP NUMBER and are located in the entrance hall on ground floor of Columbia House. After being marked, problem sets are handed back to in the seminar the following week where we will go through some of the solutions. You are encouraged to solve and discuss the exercises with your colleagues and cooperate in finding solutions. Marks will be registered on LSE for You so that your academic advisors can keep track of your efforts. Failing to submit homework may bar you from exams.

- **Help sessions**: will be held on
  - Thursday 17:00-18:00 in room TW1.G.01 in weeks 1 to 5 and 7 to 10 (teachers Mr. Cheng Li and Mr. Baojun Dou);

  you are encouraged to attend them as they provide the opportunity to work on the exercises with one-to-one help of two assistant teachers.

- **Class tests**: two tests will take place in class on
  - Thursday 17:00-18:00 in room TW1.G.01 in weeks 6 and 11;

  you are strongly encouraged to participate in order to verify the degree of your preparation. As for homework, tests marks will be registered on LSE for You in order to allow advisors to monitor your attendance and progress.

- **Exam**: the course is assessed by a three hour written exam in the Summer term which covers the material taught during both terms; previous years exams with solutions are available from LSE library website. Homework and mid-term test do not count for the final exam mark, but the more effort you put in solving exercises and studying during the year the more likely you are to pass the exam.

A Guide to Content

The following is a guide to the content of the course rather than a definitive syllabus. Throughout the course examples, with varying degrees of realism, will be used to illustrate the theory. Here is an approximate list of the topics I plan to teach, however the material that goes into the exam will be determined by what is actually covered during the lectures (I will provide you with an updated list of topics at the end of the term).

1. **Events and their Probabilities**:
   - sample space;
   - elementary set theory;
   - events;
   - probability;
   - counting;
   - conditional probability;

- independence.

2. **Random Variables and their Distributions**:
    - random variables;
    - distributions;
    - discrete random variables and probability mass function;
    - continuous random variables and probability density function;
    - support and indicator functions;
    - expectations (mean and variance);
    - moments;
    - inequalities (Chebyshev, Markov, Jensen);
    - moment generating functions;
    - survival and hazard.
3. **The Distribution Zoo**:
    - discrete distributions;
        - degenerate;
        - Bernoulli;
        - binomial;
        - negative binomial;
        - geometric;
        - hypergeometric;
        - uniform;
        - Poisson and its approximation of binomial;
    - continuous distributions;
        - uniform;
        - normal;
        - gamma;
        - chi-squared;
        - exponential;
        - beta;
        - log-normal;
4. **Multivariate Distributions**:
    - joint and marginal distributions;
    - dependence;
    - joint moments;
    - inequalities (Hölder, Cauchy-Schwarz, Minkowski);

- conditional distributions.

5. **Multivariate Applications**:
    - sums of random variables;
    - mixtures and random sums;
    - random vectors;
    - multivariate normal distribution;
    - modes of convergence;
    - limit theorems for Bernoulli sums;
    - law of large numbers;
    - central limit theorem.

Books

There are a large number of books that cover at least part of the material in the course. Finding a useful book is partly a question of personal taste. I suggest you look at what is available in the library and find a text that covers the material in a way that you find appealing and intelligible. Reproduced below is the reading list along with some additional texts that may be worth looking at.

- Casella, G. and R. L. Berger. *Statistical inference.* [QA276 C33]
  (Nearly all material covered in the course can be found in this book.)
- Larson, H. J. *Introduction to probability theory and statistical inference.* [QA273.A5 L33]
- Hogg, R. V. and A. T. Craig. *Introduction to mathematical statistics.* [QA276.A2 H71]
- Freund, J. E. *Mathematical statistics.* [QA276 F88]
- Hogg, R. V. and E. A. Tanis. *Probability and statistical inference.* [QA273 H71]
- Meyer, P. L. *Introductory probability and statistical applications.* [QA273.A5 M61]
- Mood, A. M., F. A. Graybill and D. C. Boes. *Introduction to the theory of statistics.* [QA276.A2 M81]
- Bartoszyski, R. and M. Niewiadomska-Bugaj. *Probability and statistical inference.* [QA273 B29]
- Cox, D. R. and D. V. Hinkley. *Theoretical statistics.* [QA276.A2 C87]
  (Not great to learn from but a good reference source.)
- Stuart, A. and J. K. Ord. *Kendall's advanced theory of statistics* **1**, *Distribution theory.* [QA276.A2 K31]
  (A bit arcane but covers just about everything.)

- Grimmett, G. R. and D. R. Stirzaker. *Probability and random processes.* [QA273 G86]
  (Very succinct and very precise. One for those who like it mathematical.)
- Johnson, N. L. and S. Kotz (some volumes also with N. Balakrishnan) *Discrete distributions.* and *Continuous univariate distributions.* [QA273.6 J61]
  (The place to go if you have a particular question about a distribution.)
- Larsen R. J. and M. L. Marx. *An introduction to mathematical statistics and its applications.* [QA276 L33]
  (Good introduction to probability. Lots of examples.)

Practical informations

I will post all material related to the course on moodle. Lecture notes and lecture recordings will also be made available in due course. Each week I will upload the problem sets on Thursday and after you return them on Wednesday I will also upload solutions.

If you have questions related to the course you have different ways to ask me:

1. ask questions in class, feel free to interrupt me at any time;
2. come and see me in my office COL.7.11 in Columbia House during my office hours on Thursday 13:30 - 15:00. Please make an appointment through LSE for You in advance to avoid queues;
3. take advantage of the help sessions;
4. use the forum on moodle where all of you can post both questions and answers to your colleagues; if necessary I will also post my answers there; this is a way to stimulate the discussion and will avoid me repeating the same answer many times;

(5.) if you think you need to speak with me personally, but you cannot come to my office hours, send me an email and we will fix an appointment; please try to avoid emails for questions that might be of interest for all the class and use the moodle forum instead.

1 Probability Space

The main concept of probability is a random experiment, i.e. an experiment with an uncertain outcome. Associated with random experiments is the probability space which is made of three ingredients:

1. the collection of all possible outcomes: sample space Ω;
2. the collection of all possible events: σ-algebra $\mathcal{F}$;
3. the probability measure P;

and we write it as

$$(\Omega, \mathcal{F}, P).$$

1.1 Sample space

Experiment: a procedure which can be repeated any number of times and has a well-defined set of possible outcomes.

Sample outcome: a potential eventuality of the experiment. The notation ω is used for an outcome.

Sample space: the set of all possible outcomes. The notation Ω is used for the sample space of an experiment. An outcome ω is a member of the sample space Ω, that is, $\omega \in \Omega$.

Example: a fair six-sided die is thrown once. The outcomes are numbers between 1 and 6, i.e. the sample space is given by $\Omega = \{1, \ldots, 6\}$.

Example: a fair six-sided die is thrown twice. The outcomes are pairs of numbers between 1 and 6. For example, $(3, 5)$ denotes a 3 on the first throw and 5 on the second. The sample space is given by $\Omega = \{(i, j) \ : \ i = 1, \ldots, 6, \ j = 1, \ldots, 6\}$. In this example the sample space is finite so can be written out in full:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Example: the measurement of people's height has the positive real numbers as sample space, if we allow for infinite precision in the measurement.

Example: assume to have an experiment with a given sample space Ω, then the experiment corresponding to n replications of the underlying experiment has sample space Ω^n. Notice that, in principle we can repeat an experiment infinitely many times.

1.2 Elementary set theory

Notation: given a sample space Ω, we define the following objects:

	Set terminology	Probability terminology
A	Subset of Ω	Event some outcome in A occurs
A^c	Complement	Event no outcome in A occurs
$A \cap B$	Intersection	Event outcome in both A and B occur
$A \cup B$	Union	Event outcome in A and/or B occur
$A \backslash B$	Difference	Event outcome in A but not in B occur
$A \subseteq B$	Inclusion	If outcome is in A it is also in B occur
$\emptyset$	Empty set	Impossible event
Ω	Whole space	Certain event

Properties of Intersections and Unions

1. Commutative: $A \cap B = B \cap A$,
   $A \cup B = B \cup A$.
2. Associative: $A \cap (B \cap C) = (A \cap B) \cap C$,
   $A \cup (B \cup C) = (A \cup B) \cup C$.
3. Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
4. With whole space: $A \cap \Omega = A$,
   $A \cup \Omega = \Omega$.
5. With empty set: $A \cap \emptyset = \emptyset$,
   $A \cup \emptyset = A$.

Properties of the complement set: $A^c = \Omega \backslash A$, that is, $\omega \in A^c \iff \omega \notin A$.

1. $(A^c)^c = A$.
2. $A \cap A^c = \emptyset$.
3. $A \cup A^c = \Omega$.
4. $(A \cup B)^c = A^c \cap B^c$.
5. De Morgan's theorem (a generalization of 4 above): $(\bigcup_{i=1}^n A_i)^c = \bigcap_{i=1}^n A_i^c$.

Partition of Ω: $\{A_1, \ldots, A_n\}$ is a partition if:

1. mutually exclusive: $A_i \cap A_j = \emptyset$ for any $i \neq j$, so $A_1, \ldots, A_n$ are disjoint sets;
2. exhaustive: $\bigcup_{i=1}^n A_i = \Omega$;
3. not-empty: $A_i \neq \emptyset$ for any i.

Notice that n can be infinite.

1.3 Events

For any experiment, the events form a collection of all the possible subsets of Ω which we denote $\mathcal{F}$ and has the following properties:

1. $\emptyset \in \mathcal{F}$,
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$,
3. if $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$. The union has to be infinite.

Any collection of subsets with these properties is known as a σ-algebra.

If Ω has n elements, then $\mathcal{F}$ has 2^n elements. Indeed, the number of elements of $\mathcal{F}$ is made of the sum of all possible combinations of n elements, i.e., for any $0 \leq k \leq n$, we need to compute all the possible k-elements subsets of an n-elements set:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

The binomial coefficient is also used to find the coefficients of binomial powers, the general formula is

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

and by setting $x = y = 1$ we have the result above. Another useful formula for the binomial coefficient is

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

1.4 Probability

In an experiment the intuitive definition of probability is the ratio between the number of favorable outcomes over the total number of possible outcomes or with the above notation, the probability of an event $A \subset \Omega$, such that $A \in \mathcal{F}$, is:

$$P(A) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega}.$$

Slightly more sophisticated is the "frequentist" definition of probability which is based on the frequency f_n with which a given event A is realized, given a total number n of repetitions of an experiment:

$$P(A) = \lim_{n\to\infty} f_n.$$

Example: if we toss a fair coin the sample space is $\Omega = \{H, T\}$, then the event $A = \{H\}$ has probability

$$P(\{H\}) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega} = \frac{1}{2}.$$

Alternatively, we could compute this probability by tossing the coin n times, where n is large, and compute the number of times we get head say k_n. If the coin is fair, we should get

$$P(\{H\}) = \lim_{n\to\infty} \frac{k_n}{n} = \frac{1}{2}.$$

We here adopt a more mathematical definition of probability, based on the Kolmogorov axioms.

Probability measure: is a function $P : \mathcal{F} \longrightarrow [0, 1]$, such that

1. $P(A) \geq 0$,
2. $P(\Omega) = 1$,
3. if $A_1, A_2, \ldots,$ is an infinite collection of mutually exclusive members of $\mathcal{F}$ then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i),$$

This in turn implies that for any finite collection $A_1, A_2, \ldots, A_n$ of mutually exclusive members of $\mathcal{F}$ then

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

We can associate a probability space $(\Omega, \mathcal{F}, P)$ with any experiment.

Properties of probability measures

1. $P(A^c) = 1 - P(A)$.
2. $P(A) \leq 1$.
3. $P(\emptyset) = 0$.
4. $P(B \cap A^c) = P(B) - P(A \cap B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. If $A \subseteq B$ then $P(B) = P(A) + P(B \backslash A) \geq P(A)$.
7. More generally if $A_1, \ldots, A_n$ are events then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n).$$

8. For any partition $A_1, \ldots, A_n$ of Ω

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

   Notice that n can be infinite.

9. Boole's inequality:

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

2 Counting or occupancy models

Multiplication rule for counting ordered sequences: an operation A_i can be performed in n_i different ways for $i = 1, \ldots, k$. The ordered sequence (operation A_1, operation A_2, $\ldots$, operation A_k) can be performed in $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ ways. We write this product as $\prod_{i=1}^{k} n_i$.

When the sample space Ω is finite and all the outcomes in Ω are equally likely, we calculate the probability of an event A by counting the number of outcomes in the event:

$$P(A) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega} = \frac{|A|}{|\Omega|}$$

Consider the following problem: k balls are distributed among n distinguishable boxes in such a manner that all configurations are equally likely or analogously (from the modeling point of view) we extract k balls out on n. We need to define the sample space and its cardinality, i.e. the number of its elements. The balls can be distinguishable or undistinguishable which is analogous to saying that the order in the extraction matters or not. Moreover, the extraction can be with or without replacement, i.e. the choice of a ball

is independent or not from the ball previously chosen. In terms of balls and boxes this means that we can put as many balls as we want in each box (with replacement) or only one ball can fit in each box (without replacement).

There are four possible cases (three of which are named after famous physicists).

Ordered (distinct), without replacement (dependent): in this case we must have $k \leq n$ and the sample space is

$$\Omega = \{(\omega_1 \dots \omega_k) \ : \ 1 \leq \omega_i \leq n \ \forall i \ \omega_i \neq \omega_j \text{ for } i \neq j\},$$

where ω_i is the box where ball i is located. All the possible permutations of k balls that can be formed from n distinct elements, i.e. not allowing for repetition, are

$$|\Omega| = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Ordered (distinct), with replacement (independent) - Maxwell-Boltzmann: the sample space is

$$\Omega = \{(\omega_1 \dots \omega_k) \ : \ 1 \leq \omega_i \leq n \ \forall i\},$$

where ω_i is the box where ball i is located. Each ball can be selected in n ways, so the total number of outcomes is

$$|\Omega| = \underbrace{n \cdot n \cdot \ldots \cdot n}_{k \text{times}} = n^k.$$

Unordered (not distinct), without replacement (dependent) - Fermi-Dirac: again we need $k \leq n$ and the sample space is

$$\Omega = \left\{ (\omega_1 \dots \omega_n) \ : \ \omega_i = \{0,1\} \ \forall i \text{ and } \sum_{i=1}^{n} \omega_i = k \right\},$$

with box i occupied if and only if $\omega_i = 1$. Starting from the case of distinct balls, we have to divide out the redundant outcomes and we obtain the total number of outcomes:

$$|\Omega| = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

Unordered (not distinct), with replacement (independent) - Bose-Einstein: the sample space is

$$\Omega = \left\{ (\omega_1 \dots \omega_n) \ : \ 0 \leq \omega_i \leq k \ \forall i \text{ and } \sum_{i=1}^{n} \omega_i = k \right\},$$

with ω_i the number of balls in box i. This is the most difficult case to count. The easiest way is to think in terms of k balls and n boxes. We can put as many balls as we want in each box and balls are identical. To find all the possible outcomes it is enough to keep

track of the balls and of the walls separating the boxes. Excluding the 2 external walls, we have $n+1-2 = n-1$ walls and k balls, hence we have $n-1+k$ objects that can be arranged in $(n-1+k)!$ ways. However, since the balls and the walls are identical we need to divide out the redundant orderings which are $k!(n-1)!$, so

$$|\Omega| = \frac{(n-1+k)!}{k!(n-1)!} = \binom{n-1+k}{k}.$$

Example: in a lottery 5 numbers are extracted without replacement out of $\{1,\dots,90\}$. Which is the probability of extracting the exact sequence of numbers $(1,2,3,4,5)$?
The possible outcomes of this lottery are all the 5-tuples $\boldsymbol{\omega} = (\omega_1,\dots,\omega_5)$ such that $\omega_i \in \{1,\dots,90\}$. We can extract the first number in 90 ways, the second in 89 ways and so on, so

$$|\Omega| = 90 \cdot 89 \cdot 88 \cdot 87 \cdot 86 = \frac{90!}{85!}.$$

Since all the outcomes are equally likely, the probability we are looking for is $85!/90! \simeq 1/510^9$.

Example: if in the previous example the order of extraction does not matter, i.e. we look for the probability of extracting the first 5 numbers independently of their ordering, then Ω contains all the combinations of 5 numbers extracted from 90 numbers:

$$|\Omega| = \binom{90}{5}.$$

Since all the outcomes are equally likely, the probability we are looking for is $1/\binom{90}{5} \simeq 1/410^7$ so as expected it is greater than before, although still very small!

Example: which is the probability that, out of n people randomly chosen, at least two were born in the same day of the year? We can define a generic event of the sample space as $\boldsymbol{\omega} = (\omega_1,\dots,\omega_n)$ such that $\omega_i \in \{1,\dots,365\}$. Each birth date can be selected n times so

$$|\Omega| = \underbrace{365 \cdot 365 \cdot \ldots \cdot 365}_{n \text{ times}} = 365^n.$$

Now we have to compute the number of elements contained in the event $A = \{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega} \text{ has at least two identical entries}\}$. It is easier to compute the number of elements of the complement set $A^c = \{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega} \text{ has all entries distinct}\}$. Indeed A^c is made of all n-tuples of numbers that are extracted out of 365 numbers without replacement, so the first entry can be selected in 365 ways, the second in 364 ways and so on, then

$$|A^c| = \frac{365!}{(365-n)!}.$$

If we assume that the outcomes of Ω are all equally likely (which is not completely correct as we now that birth rates are not equally distributed throughout the year), then

$$P(A) = 1 - \frac{365!}{365^n(365-n)!},$$

which for $n = 23$ is 0.507, for $n = 50$ is 0.974, and for $n = 100$ is 0.9999997.

Example: an urn contains b black balls and r red balls, we extract without replacement $n \leq (b + r)$ balls, what is the probability of extracting k red balls? We first compute all the possible ways of extracting without replacement n balls out of $(b + r)$, then $|\Omega| = \binom{b+r}{n}$. Let us assume that the all the balls are numbered and that the red ones have index $\{1, \ldots, r\}$ while the black ones have index $\{r + 1, \ldots, b + r\}$ so we are interested in the event

$$A = \{\omega \ : \ \omega \text{ contains exactly } k \text{ elements with index } \leq r\},$$

then is like asking for the all possible ways of extracting k balls out of r and $n - k$ balls out of b, therefore

$$P(A) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{b+r}{n}}.$$

Reading

Casella and Berger, Sections 1.1, 1.2.

3 Conditional probability

Let A and B be events with $P(B) > 0$. The conditional probability of A given B is the probability that A will occur given that B has occurred;

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

It is as if we were updating the sample space to B, indeed $P(B|B) = 1$. Moreover, if A and B are disjoint, then $P(A|B) = P(B|A) = 0$, once one of the two events took place the other becomes impossible.

By noticing that

$$P(A \cap B) = P(A|B)P(B) \quad \text{and} \quad P(A \cap B) = P(B|A)P(A),$$

we have the useful formula

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

Law of total probability: if $A_1, \ldots, A_n$ is a partition of Ω and B is any other event defined on Ω, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Notice that n can be infinite.

Bayes' rule: if $A_1, \ldots, A_n$ is a partition of Ω and B is any other event defined on Ω, then for any $j = 1, \ldots, n$

$$P(A_j|B) = P(B|A_j)\frac{P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

Notice that n can be infinite.

Multiplication rule for intersections: let $A_1, \ldots, A_n$ be a set of events defined on Ω,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{j=1}^{n} P\left(A_j|\bigcap_{i=0}^{j-1} A_i\right),$$

where we define $A_0 = \Omega$.

4 Independence

If the occurrence of an event B has no influence on the event A then

$$P(A|B) = P(A),$$

then from previous section

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} = P(B),$$

so A has no influence on B, moreover from Bayes' rule

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B),$$

and this is the definition of statistical independence. **Two** events A and B are said to be **independent** ($A \perp B$) if and only if

$$P(A \cap B) = P(A)P(B).$$

1. If $P(A) > 0$ then $P(B|A) = P(B) \iff A \perp B$.
   If $P(B) > 0$ then $P(A|B) = P(A) \iff A \perp B$.
2. If $A \perp B$ then $A^c \perp B^c$, $A^c \perp B$ and $A \perp B^c$.

A **collection** of events $A_1, \ldots, A_n$ is said to be **mutually independent** if for every subset $A_{i_1}, \ldots, A_{i_k}$ we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Example A common misconception is that an event A is independent of its complement A^c. In fact, this is only the case when $P(A) \in \{0, 1\}$ (check this!). Otherwise, the events A and A^c since the never occur at the same time and hence the probability of their intersection is zero.

Example: another common misconception is that an event is independent of itself. If A is an event that is independent of itself, then

$$P(A) = P(A \cap A) = P(A)P(A) = (P(A))^2.$$

The only finite solutions to the equation $x = x^2$ are $x = 0$ and $x = 1$, so an event is independent of itself only if it has probability 0 or 1.

Example: consider tossing a coin 3 times, then we have $2^3 = 8$ possible outcomes and if the coin is fair each outcome has probability $\frac{1}{8}$. If we define H_i the event of having head at the the i-th toss for $i = 1, 2, 3$ we have only four possible outcomes contained in each event H_i, therefore

$$P(H_i) = \frac{4}{8} = \frac{1}{2} \text{ for } i = 1, 2, 3.$$

To verify that H_is are independent we need to compute

$$P(H_1 \cap H_2 \cap H_3) = P(\{HHH\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2)P(H_3),$$

but we also have to compute for any $i \neq j$

$$P(H_i \cap H_j) = P(H_i)P(H_j),$$

so for example when $i = 1$ and $j = 3$

$$P(H_1 \cap H_3) = P(\{HTH, HHH\}) = \frac{2}{8} = \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_3).$$

Example: consider tossing a tetrahedron (i.e. a die with just four faces) with a red, a blue, a yellow face, and a face with all three colours. Each face has equal probability $\frac{1}{4}$ to be selected.[1] We want to see if the events: red (R), green (G), blue (B) are independent. The probability of selecting any colour is then $P(R) = P(G) = P(B) = \frac{1}{2}$ since all colours appear twice on the tetrahedron. Consider the conditional probability

$$P(R|G) = \frac{P(RG)}{P(G)} = \frac{1/4}{1/2} = \frac{1}{2} = P(R),$$

so the event R is independent of the event G, by repeating the same reasoning with all couples of colours we see that colours are pairwise independent. However, we do not have mutual independence indeed, for example,

$$P(R|GB) = \frac{P(RGB)}{P(GB)} = \frac{1/4}{1/4} = 1 \neq P(R) = \frac{1}{2}.$$

[1] Due to its geometry in this case the selected face is the bottom one once the tetrahedron is tossed.

Example. Consider the following game: your ST202 lecturer shows you three cups and tells you that under one of these there is a squashball while under the other two there is nothing. The aim of the Monty Squashball problem[2] is to win the squashball by picking the right cup. Assume you choose one of the three cups, without lifting it. At this point one of the remaining cups for sure does not contain the ball and the your lecturer lifts it showing emptiness (selecting one at random if there is a choice). With two cups still candidates to hide the squashball, you are given a second chance of choosing a cup: will you stick to the original choice or will you switch to the other cup?

We can model and solve the problem by using conditional probability and Bayes' rule.

The probability of getting the ball is identical for any cup, so

$$P(\text{ball is in } k) = \frac{1}{3}, \quad k = 1, 2, 3.$$

Once you choose a cup (say i), your ST202 lecturer can lift only a cup with no ball and not chosen by yourself, he will lift cup j (different from i and k) with probability

$$P(\text{ST202 lecturer lifts } j | \text{you choose } i \text{ and ball is in } k) = \begin{cases} \frac{1}{2} & \text{if } i = k, \\ 1 & \text{if } i \neq k. \end{cases}$$

Let us call the cup you pick number 1 (we can always relabel the cups). Using Bayes' rule we compute (for $j \neq k$ and $j \neq 1$)

$$P(\text{ball is in } k | \text{ ST202 lecturer lifts } j) = \frac{P(\text{ST202 lecturer lifts } j | \text{ ball is in } k) P(\text{ball is in } k)}{P(\text{ST202 lecturer lifts } j)}.$$

Since $P(\text{ball is in } k) = 1/3$, we are left to compute (for $j \neq 1$)

$$\begin{aligned} P(\text{lecturer lifts} j) &= \sum_{k=1}^{3} P(\text{lecturer lifts } j | \text{ ball is in } k) P(\text{ball is in } k) \\ &= \frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

This can also be seen by symmetry and law of total probability.

So if you choose cup 1 and the ST202 lecturer lifts cup 2, the probability that the ball is in cup 3 is

$$P(\text{ball is in } 3 | \text{ST202 lecturer lifts } 2) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

while the probability that the ball is in cup 1, i.e. the cup you chose at the beginning

$$P(\text{ball is in } 1 | \text{ST202 lecturer lifts } 2) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Hence, switching gives a higher probability of winning the squashball.

[2] this is an eco-friendly version of the famous Monty Hall problem which has "doors" for "cups", "goats" for "nothing" and a "car" for "squashball"; no animals are harmed in the Monty Squashball problem. It is also closely related to Bertrand's box paradox and the Prisoners' paradox (not to be confused with the Prisoners' dilemma)

Reading

Casella and Berger, Sections 1.3.

5 Random variables

We use random variables to summarize in a more convenient way the structure of experiments.

Borel σ-algebra: is the σ-algebra $\mathcal{B}(\mathbb{R})$ (called the Borel σ-algebra) on $\Omega = \mathbb{R}$, i.e. the σ-algebra generated by (i.e. the smalles sigma-algebra containing) the intervals $(a, b]$ where we allow for $a = -\infty$ and $b = +\infty$.
We could have equally have taken intervals $[a, b]$ (think about this for a while!).

Random variable: a real-valued function is defined on the sample space

$$X : \Omega \longrightarrow \mathbb{R}$$

with the property that, for every $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$.

Define, for all $x \in \mathbb{R}$, the set of outcomes

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

then $A_x \in \mathcal{F}$. Thus, A_x is an event, for every real-valued x.

The function X defines a new sample space (its range) and creates a bijective correspondence between events in the probability space $(\Omega, \mathcal{F}, P)$ with events in the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ which allows for easier mathematical computations. We need to define the probability measure on the Borel σ-algebra.

Example: consider the experiment of tossing a coin n times, the sample space is made of all the n-tuples $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ such that $\omega_i = 1$ if we get head and $\omega_i = 0$ if we get tail. An example of random variable is the function: number of heads in n tosses which we can define as

$$X(\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i.$$

Consider the case in which we get m times head with $m < n$. Then, for every number m we can define the event $A_m = \{\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n) \in \Omega : X(\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i = m\}$.

Notice that in this example the random variables have only integer values which are a subset of the real line. Notice also that the original sample space is made of 2^n elements,

while the new sample space is made of the integer numbers $\{0 \ldots, n\}$ which is a smaller space.

Example: consider the random walk, i.e. a sequence of n steps $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ such that the i-th step can be to the left or to the right. We can introduce a random variable that represents the i-th step by $X_i(\boldsymbol{\omega}) = \pm 1$ where it takes the value 1 if the step is to the left and -1 if the step is to the right. We can also introduce the random variable that represents the position of the random walk after k steps: $Y_k(\boldsymbol{\omega}) = \sum_{i=1}^{k} X_i(\boldsymbol{\omega})$.

6 Distributions

We must check that the probability measure P defined on the original sample space Ω is still valid as a probability measure defined on $\mathbb{R}$. If the sample space is $\Omega = \{\omega_1, \ldots, \omega_n\}$ and the range of X is $\{x_1, \ldots, x_m\}$, we say that we observe $X = x_i$ if and only if the outcome of the experiment is ω_j such that $X(\omega_j) = x_i$.

Induced probability: we have two cases

1. finite or countable sample spaces: given a random variable X, the associated probability measure P_X is such that, for any $x_i \in \mathbb{R}$,

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\}).$$

2. uncountable sample spaces given a random variable X, the associated probability measure P_X is such that, for any $B \in \mathcal{B}(\mathbb{R})$,

$$P_X(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}).$$

Hereafter, given the above equivalences, we denote P_X simply as P.

Cumulative distribution function (cdf): given a random variable X, it is the function

$$F : \mathbb{R} \longrightarrow [0, 1], \quad \text{s.t.} \quad F(x) = P(X \leq x), \ \forall\, x \in \mathbb{R}.$$

Properties of cdfs: F is a cdf if and only if

1. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
2. Non-decreasing: if $x < y$ then $F(x) \leq F(y)$.
3. Right-continuous: $\lim_{h \to 0^+} F(x + h) = F(x)$.

Probabilities from distribution functions

1. $P(X > x) = 1 - F(x)$;
2. $P(x < X \leq y) = F(y) - F(x)$;
3. $P(X < x) = \lim_{h \to 0^-} F(x + h) = F(x^-)$;
4. $P(X = x) = F(x) - F(x^-)$.

Identically distributed random variables: the random variables X and Y are identically distributed if, for any set $A \in \mathcal{B}(\mathbb{R})$, $P(X \in A) = P(Y \in A)$. This is equivalent to saying that $F_X(x) = F_Y(x)$, for every $x \in \mathbb{R}$.

Example: in the random walk the step size random variable X_i is distributed as:

$$P(X_i = 1) = \frac{1}{2}, \qquad P(X_i = -1) = \frac{1}{2}.$$

while

$$F_X(-1) = \frac{1}{2}, \qquad F_X(1) = 1.$$

The random variables X_i are identically distributed. Moreover, they are also independent so

$$P(\boldsymbol{\omega}) = P(X_1 = \omega_1, \ldots, X_n = \omega_n) = \prod_{i=1}^{n} P(X_i = \omega_i),$$

for any choice of $\omega_1, \ldots, \omega_n = \pm 1$. Therefore, all n-tuples $\boldsymbol{\omega}$ are equally probable with probability

$$P(\boldsymbol{\omega}) = P(X_1 = \omega_1, \ldots, X_n = \omega_n) = \prod_{i=1}^{n} \frac{1}{2} = \frac{1}{2^n}.$$

Consider the random variable Z the counts the steps to the right, then the probability of having k steps to the right and $n - k$ steps to the left is

$$\begin{aligned} P(Z = k) = F_Z(k) &= (\text{\# of ways of extracting } k \text{ 1s out of } n) \, (\text{ Prob. of a generic } \boldsymbol{\omega}) \\ &= \binom{n}{k} \frac{1}{2^n}. \end{aligned}$$

We say that X_i follows a Bernoulli distribution and Z follows a Binomial distribution. The previous example of a fair coin can be modeled exactly in the same way but this time by defining $X_i(\boldsymbol{\omega}) = 0$ or 1.

7 Discrete random variables

A random variable X is *discrete* if it only takes values in some countable subset $\{x_1, x_2, \ldots\}$ of $\mathbb{R}$, then $F(x)$ is a step-function of x, but still right-continuous.

Probability mass function (pmf): given a discrete random variable X, it is the function

$$f : \mathbb{R} \longrightarrow [0,1] \text{ s.t. } f(x) = P(X = x) \quad \forall\, x \in \mathbb{R}.$$

Properties of pmfs

1. $f(x) = F(x) - F(x^-)$;
2. $F(x) = \sum_{i:x_i \leq x} f(x_i)$;
3. $\sum_i f(x_i) = 1$;
4. $f(x) = 0$ if $x \notin \{x_1, x_2, \ldots\}$.

8 Continuous random variables

A random variable X is *continuous* if it takes values in $\mathbb{R}$ and its distribution function $F(x)$ is an absolutely continuous function of x (F is differentiable "almost everywhere") .

Probability density function (pdf): given a continuous random variable X, (a version of) its density is an integrable function $f : \mathbb{R} \longrightarrow [0, +\infty)$ such that the cdf of X can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u \qquad \forall\, x \in \mathbb{R}.$$

Properties of continuous random variables

1. $P(X = x) = 0$ for any $x \in \mathbb{R}$;
2. $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$;
3. $f(x) \geq 0$ for any $x \in \mathbb{R}$;
4. $\int_a^b f(u)\mathrm{d}u = P(a < X \leq b)$.

Notice that, in principle, any nonnegative function with a finite integral over its support can be turned into a pdf. So if

$$\int_{A \subset \mathbb{R}} h(x)\mathrm{d}x = K < \infty$$

for some constant $K > 0$, then $h(x)/K$ is a pdf of a random variable with values in A.

Unified notation: given a random variable X;

$$P(a < X \leq b) = \int_a^b \mathrm{d}F(x) = \begin{cases} \sum_{i:a<x_i\leq b} f(x_i), & \text{if } X \text{ discrete,} \\ \int_a^b f(u)du, & \text{if } X \text{ continuous.} \end{cases}$$

Reading

Casella and Berger, Sections 1.4 - 1.5 - 1.6.

9 Expectations

Mean: given a random variable X, its mean is defined as

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x \mathrm{d}F(x) = \begin{cases} \sum_i x_i f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} x f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

where f is either the pmf or the pdf. The definition holds provided that $\int_{-\infty}^{+\infty} |x| \mathrm{d}F(x) < \infty$.

If we interpret μ as a good guess of X we may also be interested to have a measure of the uncertainty with which X assumes the value μ, this is known as variance of X.

Variance: given a random variable X, its variance is defined as

$$\sigma^2 = \mathrm{Var}[X] = \int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x) = \begin{cases} \sum_i (x_i - \mu)^2 f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) \mathrm{d}x, & \text{if } X \text{ continuous,} \end{cases}$$

where f is either the pmf or the pdf. The standard deviation is defined as $\sigma = \sqrt{\mathrm{Var}[X]}$. Notice that $\sigma^2 = \mathrm{E}[(X-\mu)^2]$. The definition holds provided that $\int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x) < \infty$.

Expectations: for an integrable function $g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\int_{-\infty}^{\infty} |g(x)| \mathrm{d}F(x) < \infty$, the expectation of the random variable $g(X)$ as

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \mathrm{d}F(x) = \begin{cases} \sum_i g(x_i) f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} g(x) f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

Note that we have cheated a bit here, since we need to show in fact that $g(X)$ is a random variable and also that the given expression corresponds to the one given above for the random variable $g(X)$. This can be done but is beyond the scope of ST202. Feel free to ask me if you would like to hear more about this.

Properties of expectations: for any constant a, integrable functions g_1 and g_2, and random variables X and Y:

1. $\mathrm{E}[a] = a$;

2. $\mathrm{E}[ag_1(X) + bg_2(Y)] = a\mathrm{E}[g_1(X)] + b\mathrm{E}[g_2(Y)]$;
3. if $X \geq Y$ then $\mathrm{E}[X] \geq \mathrm{E}[Y]$;
4. $\mathrm{Var}[ag_1(X) + b] = a^2\mathrm{Var}[g_1(X)]$.

The variance of X can be written in a more convenient form

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

10 Moments

Moments are expectations of powers of a random variable. They characterise the distribution of a random variable. Said differently (and somewhat informally), the more moments of X we can compute, the more precise is our knowledge of the distribution of X.

Moment: given a random variable X, for r a positive integer then the r^{th} moment, μ_r, of X is

$$\mu_r = \mathrm{E}[X^r] = \int_{-\infty}^{+\infty} x^r \mathrm{d}F(x) \begin{cases} \sum_i x_i^r f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} x^r f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

where f is either the pmf or the pdf. The definition holds provided that $\int_{-\infty}^{+\infty} |x|^r \mathrm{d}F(x) < \infty$.

Central moment: given a random variable X, the r^{th} central moment, m_r is

$$m_r = \mathrm{E}[(X - \mu_1)^r].$$

The definition holds provided that $\int_{-\infty}^{+\infty} |x|^r \mathrm{d}F(x) < \infty$. so if the r-th moment exists, then also the r-th central moment exists.

Properties of moments:

1. mean: $\mu_1 = \mathrm{E}[X] = \mu$ and $m_1 = 0$;
2. variance: $m_2 = \mathrm{E}[(X - \mu_1)^2] = \mathrm{Var}[X] = \sigma^2$;
3. coefficient of skewness: $\gamma = \mathrm{E}[(X - \mu_1)^3]/\sigma^3 = m_3/m_2^{\frac{3}{2}}$;
4. coefficient of kurtosis: $\kappa = (\mathrm{E}[(X - \mu_1)^4]/\sigma^4) = (m_4/m_2^2)$.

What would a distribution with postive skew and large kurtosis look like?

11 Inequalities involving expectations

A general inequality: let X be a random variable with $X \geq 0$ and let g be a positive increasing function on $\mathbb{R}^+$, then, for any $a > 0$,

$$P(g(X) \geq a) \leq \frac{\mathrm{E}[g(X)]}{a}.$$

There are two special cases.

1. **Markov's inequality**: let X be a random variable with $X \geq 0$ and $\mathrm{E}[X]$ defined, then, for any $a > 0$,
$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}.$$
2. **Chebyshev's inequality**: let X be a random variable with $\mathrm{E}[X^2] < \infty$, then, for any $a > 0$,
$$P((X - \mathrm{E}[X])^2 \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

Jensen's inequality: If X is a random variable with $\mathrm{E}[X]$ defined, and g is a convex function with $\mathrm{E}[g(X)]$ defined, then

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]).$$

12 Moment generating functions

These are functions that help to compute moments of a distribution and are also useful to characterie the distribution. However, it can be shown that the moments do not characterise the distribution uniquely (if you would like to know more about this, check the log-normal distribution).

Moment generating function (mgf): given a random variable X, it is a function

$$M : \mathbb{R} \longrightarrow [0, \infty) \text{ s.t. } M(t) = \mathrm{E}[e^{tX}],$$

where it is assumed $M(t) < \infty$ for $|t| < h$ and some $h > 0$, i.e. the expectation exists in a neighborhood of 0. Therefore,

$$M(t) = \int_{-\infty}^{+\infty} e^{tx} \mathrm{d}F(x) = \begin{cases} \sum_i e^{tx_i} f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) \mathrm{d}x, & \text{if } X \text{ continuous.} \end{cases}$$

Properties of mgfs: if X has mgf $M(t)$ then

1. Taylor expansion:

$$M(t) = 1 + t\mathrm{E}[X] + \frac{t^2}{2!}\mathrm{E}[X^2] + \ldots + \frac{t^r}{r!}\mathrm{E}[X^r] + \ldots = \sum_{j=0}^{\infty} \frac{\mathrm{E}[X^j]}{j!} t^j;$$

2. the r^{th} moment is the coefficient of $t^r/r!$ in the Taylor expansion;

3. derivatives at zero:

$$\mu_r = \mathrm{E}[X^r] = M^{(r)}(0) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} M(t)\right|_{t=0}.$$

Proof: by differentiating $M(t)$ (in a neighbourhood of 0 assuming existence)

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} M(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{+\infty} e^{tx} \mathrm{d}F(x) = \\
&= \int_{-\infty}^{+\infty} \frac{\mathrm{d}}{\mathrm{d}t} e^{tx} \mathrm{d}F(x) = \\
&= \int_{-\infty}^{+\infty} x e^{tx} \mathrm{d}F(x) = \\
&= \mathrm{E}[X e^{tX}],
\end{aligned}$$

and in general

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r} M(t) = \mathrm{E}[X^r e^{tX}],$$

by imposing $t = 0$ we get the desired result.

Uniqueness: let F_X and F_Y be two cdfs with all moments defined, then:

1. if X and Y have bounded support, then $F_X(x) = F_Y(x)$ for any $x \in \mathbb{R}$ if and only if $\mathrm{E}[X^r] = \mathrm{E}[Y^r]$ for any $r \in \mathbb{N}$;

2. if the mgfs exist and $M_X(t) = M_Y(t)$ for all $|t| < h$ and some $h > 0$, then $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

Cumulant generating function (cgf): given a random variable X with moment generating function $M(t)$, it is defined as

$$K(t) = \log M(t).$$

Cumulant: the r^{th} cumulant, c_r, is the coefficient of $t^r/r!$ in the Taylor expansion of the cumulant generating function $K(t)$:

$$c_r = K^{(r)}(0) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} K(t)\right|_{t=0}.$$

Properties of cgfs:

1. $c_1 = \mu_1 = \mu$ (mean, first moment);
2. $c_2 = m_2 = \sigma^2$ (variance, second central moment);
3. $c_3 = m_3$ (third central moment);
4. $c_4 + 3c_2^2 = m_4$ (fourth central moment).

Reading

Casella and Berger, Sections 2.2 - 2.3.

13 Discrete distributions

Degenerate: all probability concentrated in a single point a.

- $f(x) = 1 \quad$ for $x = a$.
- $M(t) = e^{at}, \quad K(t) = at$.
- $\mu = a, \quad \sigma^2 = 0$.

Bernoulli: trials with two, and only two, possible outcomes, here labeled $X = 0$ (failure) and $X = 1$ (success).

- $f(x) = p^x(1-p)^{1-x} \quad$ for $x = 0,\ 1$.
- $M(t) = 1 - p + pe^t, \quad K(t) = \log(1 - p + pe^t)$.
- $\mu = p, \quad \sigma^2 = p(1-p)$.

Binomial: we want to count the number of successes in n independent Bernoulli trials, each with probability of success p. Consider n random variables Y_i with just two possible outcomes $Y_i = 0,\ 1$, their sum $X = \sum_{i=1}^{n} Y_i$ is the total number of successes in n trials, so $0 \leq X \leq n$. Notation $X \sim \text{Bin}(n, p)$. We need to count all the possible ways to extract x numbers out of n and multiply this number for the probability of success given by the Bernoulli distribution.

- $f(x) = \frac{n!}{x!(n-x)!} p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x} \quad$ for $x = 0,\ 1, \ldots,\ n$.
- $M(t) = (1 - p + pe^t)^n, \quad K(t) = n\log(1 - p + pe^t)$.
- $\mu = np, \quad \sigma^2 = np(1-p)$.

The Bernoulli distribution is equivalent to a binomial distribution with $n = 1$.

Examples: tossing a coin n times and counting the number of times we get head (or tail); n steps in the random walk and counting the steps to the right (or to the left).
Suppose to roll a die k times and we want the probability of obtaining at least one 3. So we have k Bernoulli trials with success probability $p = 1/6$. Define the random variable X that counts the total number of 3 in k rolls, then $X \sim \text{Bin}(k, 1/6)$ and

$$P(\text{at least one 3}) = P(X > 0) = 1 - P(X = 0) = 1 - \binom{k}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^k = 1 - \left(\frac{5}{6}\right)^k$$

If, by throwing two dice, we were interested in the probability of at least double 3 we would get

$$P(\text{at least one double 3}) = 1 - \left(\frac{35}{36}\right)^k < P(\text{at least one 3}),$$

since $35/36 > 5/6$.

Computing the moments and mgf of the binomial distribution: just notice that a binomial random variable X is the sum of n Bernoulli independent random variables Y_i, each with mean $\text{E}[Y_i] = p$ and variance $\text{Var}[X] = p(1-p)$ hence

$$\text{E}[X] = \text{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \text{E}[Y_i] = np,$$

and (independence is crucial here)

$$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[Y_i] = np(1-p).$$

The mgf is computed as

$$M(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{k}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{k}(pe^t)^x(1-p)^{n-x}$$

use the binomial expansion

$$(u+v)^n = \sum_{x=0}^{n}\binom{n}{x}u^x v^{n-x}$$

and by substituting $u = pe^t$ and $v = 1-p$ we get

$$M(t) = (pe^t + 1 - p)^n.$$

Negative Binomial: we want to count the number of Bernoulli trials necessary to get a fixed number of successes (i.e. a waiting time). Consider a random variable X denoting the trial at which the r^{th} success occurs. We want the distribution of the event $\{X = x\}$ for $x = r, r+1, \ldots$. This event occurs only if we had $r-1$ successes in $x-1$ trials and a success at the x^{th} trial. By multiplying these probabilities we have

- $f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$ for $x = r,\ r+1, \ldots$.
- $M(t) = \left(\frac{pe^t}{1-(1-p)e^t} \right)^r, \quad K(t) = -r \log\{(1-\frac{1}{p}) + \frac{1}{p} e^{-t}\}$ for $|t| < -\log(1-p)$.
- $\mu = \frac{r}{p}, \quad \sigma^2 = \frac{r(1-p)}{p^2}$.

It is also the defined in terms of the number of failures before the r^{th} success.

Geometric: to count the number of Bernoulli trials before the first success occurs. Equivalent to a negative binomial with $r = 1$.

- $f(x) = (1-p)^{x-1} p$ for $x = 1,\ 2, \ldots$.
- $M(t) = \frac{pe^t}{1-(1-p)e^t}, \quad K(t) = -\log\{(1-\frac{1}{p}) + \frac{1}{p} e^{-t}\}$ for $|t| < -\log(1-p)$.
- $\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}$.

This distribution is memoryless, indeed, if X follows a geometric distribution, then, for integers $s > t$,

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \cap X > t)}{P(X > t)} = \frac{P(X > s)}{P(X > t)} \\ &= (1-p)^{s-t} = P(X > s - t), \end{aligned}$$

Given that we observed t failures we observe an additional $s - t$ failures with the same probability as we observed $s - t$ failures at the beginning of the experiment. The only thing that counts is the length of the sequence of failures not its position.

Hypergeometric: it is usually explained with the example of the urn model. Assume to have an urn containing a total of N balls made up of N_1 balls of type 1 and $N_2 = N - N_1$ balls of type 2, we want to count the number of type 1 balls chosen when selecting $n < N$ balls without replacement from the urn.

- $f(x) = \binom{N_1}{x} \binom{N-N_1}{n-x} / \binom{N}{n}$ for $x \in \{0, \ldots, n\} \cap \{n - (N - N_1), \ldots, N_1\}$.
- $\mu = n\frac{N_1}{N}, \quad \sigma^2 = n\frac{N_1}{N}\frac{N-N_1}{N}\frac{N-n}{N-1}$.

Uniform: for experiments with N equally probable outcomes

- $f(x) = \frac{1}{N}$ for $x = 1,\ 2, \ldots, N$.
- $\mu = \frac{N+1}{2}, \quad \sigma^2 = \frac{N^2-1}{12}$.

Poisson: to count the number of events which occur in an interval of time. The assumption is that for small time intervals the probability of an occurrence is proportional to the length of the waiting time between two occurrences. We consider the random variable X which counts the number of occurrences of a given event in a given unit time interval, it depends on a parameter λ which is the intensity of the process considered. Notation $\mathrm{Pois}(\lambda)$.

- $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,\ 1, \ldots$.
- $M(t) = e^{\lambda(e^t-1)}, \quad K(t) = \lambda(e^t - 1)$.
- $\mu = \lambda, \quad \sigma^2 = \lambda$.

The intensity is the average number of occurrences in a given unit time interval. Notice that the Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

Example: think of crossing a busy street with an average of 300 cars per hour passing. In order to cross we need to know the probability that in the next minute no car passes. In a given minute we have an average of $\lambda = 300/60 = 5$ cars passing through. If X is the number of cars passing in one minute we have

$$P(X = 0) = \frac{e^{-5}5^0}{0!} = 6.7379 \cdot 10^{-3},$$

maybe is better to cross the street somewhere else. Notice that λ has to be the intensity per unit of time. If we are interested in no cars passing in one hour then $\lambda = 300$ and clearly the probability would be even smaller. If we want to know the average number of cars passing in 5 minutes time then just define a new random variable X which counts the cars passing in 5 minutes, which is distributed as Poisson with $\lambda = 300/12 = 25$ and this is also the expected value.

The Poisson approximation: if $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Pois}(\lambda)$ with $\lambda = np$, then for large n and small p we have $P(X = x) \simeq P(Y = x)$. More rigorously we have to prove that for finite $\lambda = np$

$$\lim_{n\to\infty} F_X(x; n, p) = F_Y(x; \lambda)$$

we can use mgfs and prove equivalently that

$$\lim_{n\to\infty} M_X(t; n, p) = \lim_{n\to\infty} (1 - p + pe^t)^n = e^{\lambda(e^t-1)} = M_Y(t; \lambda).$$

Proof: we can use the following result: given a sequence of real numbers s.t. $a_n \to a$ for $n \to \infty$, then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Now

$$\begin{aligned} \lim_{n\to\infty} M_X(t; n, p) &= \lim_{n\to\infty} (1 - p + pe^t)^n = \lim_{n\to\infty} \left(1 + \frac{1}{n}(e^t - 1)np\right)^n = \\ &= \lim_{n\to\infty} \left(1 + \frac{1}{n}(e^t - 1)\lambda\right)^n = e^{\lambda(e^t-1)}. \end{aligned}$$

14 Continuous distributions

Uniform: a random number chosen from a given closed interval $[a,\ b]$. Notation $U(a, b)$.

- $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.
- $M(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$ for $t \neq 0$ and $M(0) = 1$.
- $\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}$.

Normal or Gaussian: this is the most important distribution. Notation $N(\mu, \sigma^2)$.

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ for $-\infty < x < \infty$.
- $M(t) = e^{\mu t+\sigma^2 t^2/2}$.
- $\mathrm{E}[X] = \mu, \quad \mathrm{Var}[X] = \sigma^2$.

If $X \sim N(\mu, \sigma^2)$ then $(X - \mu)/\sigma = Z \sim N(0, 1)$, is a standard normal distribution, i.e. with zero mean and unit variance. We can use the moments of Z to compute the moments of X, indeed

$$\mathrm{E}[X] = \mathrm{E}[\mu + \sigma Z] = \mu, \qquad \mathrm{Var}[X] = \mathrm{Var}[\mu + \sigma Z] = \sigma^2.$$

The shape of $f(x)$ is symmetric around μ with inflection points at $\mu \pm \sigma$. A statistical table (in the past) or a computer programme (nowadays) can be used to calculate the distribution function. The following values will be useful later on:

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = .6826, \\ P(|X - \mu| \leq 2\sigma) &= P(|Z| \leq 2) = .9544, \\ P(|X - \mu| \leq 3\sigma) &= P(|Z| \leq 3) = .9974. \end{aligned}$$

In particular, the so-called two-sigma rule states that (roughly) 95% (in a repeated sample) of the data from a normal distribution falls within two standard deviations of its mean.

The normal distribution is characterized by just its first two moments. We can compute higher order moments by using the following relation (holding for any differentiable function $g(X)$) for $X \sim N(\mu, \sigma^2)$:

$$\mathrm{E}[g(X)(X - \mu)] = \sigma^2 \mathrm{E}[g'(X)].$$

Check this (hint: use integration by parts).

From the above relation we have that all moments of a normal distribution are computable starting from the second central moment. Moreover, for a standard normal random variable Z all moments of odd order are zero, in particular

$$\begin{aligned} \text{skewness} \quad \gamma &= \frac{\mathrm{E}[Z^3]}{\mathrm{E}[Z^2]^{3/2}} = \mathrm{E}[Z^2 Z] = \mathrm{E}[2Z] = 0, \\ \text{kurtosis} \quad \kappa &= \frac{\mathrm{E}[Z^4]}{\mathrm{E}[Z^2]^2} = \mathrm{E}[Z^3 Z] = \mathrm{E}[3Z^2] = 3. \end{aligned}$$

The skewness coefficient measures the asymmetry and indeed is zero for the normal, and the kurtosis coefficient measures flatness of the tails, usually we are interested in the coefficient of excess kurtosis (with respect to the normal case), i.e. $\kappa - 3$.

Computing moments and mgf of the standard normal distribution: the mgf is computed as

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}+tz} \mathrm{d}z = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2-2tz+t^2}{2}} e^{t^2/2} \mathrm{d}z = \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(z-t)^2}{2}} \mathrm{d}z = e^{\frac{t^2}{2}}. \end{aligned}$$

The Taylor expansion of $M(t)$ is

$$\begin{aligned} M(t) &= 1 + 0 + \frac{t^2}{2} + 0 + \frac{t^4}{2^2 2!} + \ldots = \sum_{j=0}^{+\infty} \frac{t^{2j}}{2^j j!} = \\ &= 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \ldots = \sum_{r=0}^{+\infty} \mu_r \frac{t^r}{r!}, \end{aligned}$$

hence the moments of Z (which in this case are equal to the central moments) are for $r = 0, 1, 2, \ldots$

$$\begin{aligned} \mu_{2r+1} &= \mathrm{E}[Z^{2r+1}] = 0, \\ \mu_{2r} &= \mathrm{E}[Z^{2r}] = \frac{(2r)!}{2^r r!}. \end{aligned}$$

Gamma: is a family of distributions characterized by parameters $\alpha > 0$ and θ. We need the gamma function defined by

$$\Gamma(t) = \int_0^{\infty} y^{t-1} e^{-y} \mathrm{d}y, \quad \text{for } t > 0.$$

Properties of the gamma function $\Gamma(t) = (t-1)\Gamma(t-1)$ for $t > 1$ and $\Gamma(n) = (n-1)!$ for positive integer n. Notation for the gamma distribution; $\mathrm{Gamma}(\alpha, \theta)$ or $G(\alpha, \theta)$.

- $f(x) = \frac{1}{\Gamma(\alpha)} \theta^{\alpha} x^{\alpha-1} e^{-\theta x}$ for $0 \leq x < \infty$.
- $M(t) = \frac{1}{(1-t/\theta)^{\alpha}}$ for $t < \theta$.
- $\mu = \alpha/\theta, \quad \sigma^2 = \alpha/\theta^2$.

α is the shape parameter determining if the distribution has a peak or it is monotonically decreasing, while θ is the scale parameter influencing the spread of the distribution hence its peak location.

Chi-square: if Z_j are independent standard normal, then $X = \sum_{j=1}^{r} Z_j^2$ has a chi-square distribution with r degrees of freedom. Notation χ_r^2 or $\chi^2(r)$. Equivalent to a gamma distribution with $\alpha = r/2$ and $\theta = 1/2$.

- $f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}$ for $0 \leq x < \infty$.
- $M(t) = \frac{1}{(1-2t)^{r/2}}$ for $t < 1/2$.
- $\mu = r, \quad \sigma^2 = 2r$.

Exponential: waiting time between events distributed as Poisson with intensity θ. Notation $\mathrm{Exp}(\theta)$ (somewhat ambiguous). Equivalent to a gamma distribution with $\alpha = 1$.

- $f(x) = \theta e^{-\theta x}$ for $0 \leq x < \infty$.
- $M(t) = \frac{\theta}{\theta - t}$ for $t < \theta$.
- $\mu = 1/\theta, \quad \sigma^2 = 1/\theta^2$.

It is a memoryless distribution, indeed if $X \sim \mathrm{Exp}(\theta)$, then for integers $s > t$,

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \cap X > t)}{P(X > t)} = \frac{P(X > s)}{P(X > t)} \\ &= \frac{\int_s^{+\infty} \theta e^{-x\theta} \mathrm{d}x}{\int_t^{+\infty} \theta e^{-x\theta} \mathrm{d}x} = e^{-(s-t)\theta} = P(X > s - t). \end{aligned}$$

Example: it is used in modeling survival rates (see below).

Log-normal: it is the distribution of a random variable X such that $\log X \sim N(\mu, \sigma^2)$. It is used for random variables with positive support, and it is very similar to, although less flexible, and more analytically tractable than the gamma distribution.

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-(\log x - \mu)^2/(2\sigma^2)}$ for $0 < x < \infty$.
- $\mathrm{E}[X] = e^{\mu + \sigma^2/2}, \quad \mathrm{Var}[X] = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$.

Notice that in this case $M(t)$ is not defined (see ex. 2.36 Casella & Berger). Examples are the distributions of income or consumption. This choice allows to model the logs of income and consumption by means of the normal distribution which is the distribution predicted by economic theory.

15 Survival and hazard

Survival function: given a continuous non-negative random variable X, it is the function

$$\bar{F}(x) = 1 - F(x) = P(X > x).$$

where x is interpreted as a threshold and we are interested in the probability of having realizations of X beyond x. We usually assume that $\bar{F}(0) = 1$.

In the context of survival analysis the cdf and the pdf are called lifetime distribution function and event density, respectively.

Hazard function or hazard rate: it is the probability of having a realization of X in a small interval beyond the threshold x, i.e. conditional on survival of X beyond x:

$$h(x) = \lim_{\varepsilon \to 0^+} \frac{\bar{F}(x+\varepsilon) - \bar{F}(x)}{\varepsilon \bar{F}(x)} = \lim_{\varepsilon \to 0^+} \frac{P(X \leq x + \varepsilon | X > x)}{\varepsilon},$$

it is then defined as

$$h(x) = \frac{f(x)}{\bar{F}(x)} = -\frac{\bar{F}'(x)}{\bar{F}(x)}.$$

Its relationship with cdf is:

$$h(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \log(1 - F(x)), \qquad F(x) = 1 - \exp\left(-\int_0^x h(u)\mathrm{d}u\right).$$

Reading

Casella and Berger, Sections 3.1 - 3.2 - 3.3.

16 Bivariate joint and marginal distributions

For simplicity we first give the definitions for the bivariate case and then we generalise to the n-dimensional setting.

Joint cumulative distribution function: for two random variables X and Y the joint cdf is a function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$F(x, y) = P(X \leq x, Y \leq y).$$

Properties of joint cdf:

1. $F_{X,Y}(-\infty, y) = \lim_{x\to-\infty} F_{X,Y}(x, y) = 0$,
   $F_{X,Y}(x, -\infty) = \lim_{y\to-\infty} F_{X,Y}(x, y) = 0$,
   $F_{X,Y}(+\infty, +\infty) = \lim_{x\to+\infty, y\to+\infty} F_{X,Y}(x, y) = 1$;

2. Right continuous in x: $\lim_{h\to0^+} F_{X,Y}(x + h, y) = F_{X,Y}(x, y)$,
   Right continuous in y: $\lim_{h\to0^+} F_{X,Y}(x, y + h) = F_{X,Y}(x, y)$.

3. For any y, the function $F(x, y)$ is non-decreasing in x.
   For any x, the function $F(x, y)$ is non-decreasing in y.

We are interested in the probability that X and Y take values in a given (Borel !) subset of the plane $\mathbb{R} \times \mathbb{R} \equiv \mathbb{R}^2$. The simplest is a rectangular region $A = \{(x, y) \in \mathbb{R}^2 \text{ s.t. } x_1 < x \leq x_2 \text{ and } y_1 < x \leq y_2\}$. Then

$$\begin{aligned} P(A) &= P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \\ &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - [F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_1)]. \end{aligned}$$

Marginal cumulative distribution functions: if $F_{X,Y}$ is the joint distribution function of X and Y then the marginal cdfs are the usual cdfs of the single random variables and are given by

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y) = F_{X,Y}(x, \infty),$$
$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x, y) = F_{X,Y}(\infty, y).$$

Marginal cdfs are generated from the joint cdf, but the reverse is not true. The joint cdf contains information that is not captured in the marginals. In particular it tells us about the dependence structure among the random variables, i.e. how they are associated.

17 Bivariate joint and marginal pmf and pdf

Joint probability mass function: for two discrete random variables X and Y it is a function $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$f_{X,Y}(x, y) = P(X = x, Y = y) \quad \forall\, x, y \in \mathbb{R}.$$

In general

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \sum_{x_1<x\leq x_2} \sum_{y_1<y\leq y_2} f_{X,Y}(x, y).$$

Marginal probability mass functions: for two discrete random variables X and Y, with range $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$ respectively, the marginal pmfs are

$$\begin{aligned} f_X(x) &= \sum_{y\in\{y_1,y_2,\ldots\}} f_{X,Y}(x, y) \\ f_Y(y) &= \sum_{x\in\{x_1,x_2,\ldots\}} f_{X,Y}(x, y). \end{aligned}$$

Joint probability density function: for two jointly continuous random variables X and Y, it is an integrable function $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) \mathrm{d}u \mathrm{d}v \quad \forall\, x, y \in \mathbb{R},$$

notice that this implies

$$f_{X,Y}(x, y) = \left. \frac{\partial^2 F_{X,Y}(u, v)}{\partial u \partial v} \right|_{u=x, v=y},$$

Properties of joint pdf:

1. $f_{X,Y}(x, y) \geq 0$ for any $x, y \in \mathbb{R}$;
2. normalisation:
$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y = 1;$$
3. probability of a rectangular region:
$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y;$$
4. for any (Borel) set $B \subseteq R^2$ the probability that (X, Y) takes values in B is
$$P(B) = \int \int_{B} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y.$$

In the one-dimensional case events are usually intervals of $\mathbb{R}$ and their probability is proportional to their length, in two-dimensions events are regions of the plane $\mathbb{R}^2$ and their probability is proportional to their area, in three-dimensions events are regions of the space $\mathbb{R}^3$ and their probability is proportional to their volume. Lengths, areas and volumes are weighted by the frequencies of the outcomes which are part of the considered events hence they are areas, volumes and 4-d volumes under the pdfs. Probability is the measure of events with respect to the measure of the whole sample space which is 1 by definition.

Marginal probability density functions: for two jointly continuous random variables X and Y, they are integrable functions $f_X : \mathbb{R} \to [0, +\infty)$ and $f_Y : \mathbb{R} \to [0, +\infty)$ such that

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}y, \quad \forall\, x \in \mathbb{R}, \\ f_Y(y) &= \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}x, \quad \forall\, y \in \mathbb{R}. \end{aligned}$$

Therefore, the marginal cdfs are

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f_{X,Y}(u, y) \mathrm{d}y \mathrm{d}u, \quad \forall\, x \in \mathbb{R}, \\ F_Y(y) &= \int_{-\infty}^{y} \int_{-\infty}^{+\infty} f_{X,Y}(x, v) \mathrm{d}x \mathrm{d}v, \quad \forall\, y \in \mathbb{R}. \end{aligned}$$

18 Cdf, pmf, and pdf of n random variables

Multivariate generalization: for n random variables $X_1, \ldots, X_n$ we have analogous definitions:

1. the joint cdf is a function $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$ such that
$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots X_n \leq x_n);$$
2. the marginal cdfs are, for any $j = 1, \ldots, n$, the functions
$$F_{X_j}(x_j) = F_{X_1,\ldots,X_n}(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty);$$
3. the marginal pmf or pdf are, for any $j = 1, \ldots, n$, the functions
$$f_{X_j}(x_j) = \begin{cases} \sum_{x_1} \cdots \sum_{x_{j-1}} \sum_{x_{j+1}} \cdots \sum_{x_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), & \text{discrete case,} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_n, & \text{continuous case;} \end{cases}$$
4. if g is a well-behaved function $g : \mathbb{R}^n \to \mathbb{R}$, then
$$\mathrm{E}[g(X_1, \ldots, X_n)] = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), & \text{discrete,} \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n, & \text{continuous.} \end{cases}$$

Reading

Casella and Berger, Section 4.1.

19 Independence of two random variables

Besides the usual univariate measures of location (mean) and scale (variance), in the multivariate case we are interested in measuring the dependence among random variables.

Joint cdf of independent random variables: two random variables X and Y are independent if and only if the events $\{X \leq x\}$, $\{Y \leq y\}$ are independent for all choices of x and y, i.e., for all $x, y \in \mathbb{R}$,

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

Joint pmf or pdf of independent random variables: two random variables X and Y are independent if and only if, for all $x, y \in \mathbb{R}$,

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

The two above are necessary and sufficient conditions, while the following is just necessary conditions but not sufficient (see also below the distinction between independence and uncorrelation).

Expectation and independence: if X and Y are independent then

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

Moreover, if g_1 and g_2 are well-behaved functions then also $g_1(X)$ and $g_2(Y)$ are independent random variables, hence

$$\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)].$$

20 Independence of n random variables

Multivariate generalisation: in the n-dimensional case we have analogous definitions:

1. the random variables $X_1,\ X_2, \ldots,\ X_n$ are mutually independent if and only if the events $\{X_1 \leq x_1\},\ \{X_2 \leq x_2\}, \ldots,\ \{X_n \leq x_n\}$ are independent for all choices of $x_1, x_2, \ldots, x_n \in \mathbb{R}$:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_n}(x_n) = \prod_{j=1}^{n} F_{X_j}(x_j);$$

2. $X_1, X_2, \ldots, X_n$ are mutually independent if and only if $x_1, x_2, \ldots, x_n$:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n) = \prod_{j=1}^{n} f_{X_j}(x_j).$$

3. if $X_1, X_2, \ldots, X_n$ are mutually independent then

$$\mathrm{E}[X_1, X_2, \ldots, X_n] = \mathrm{E}[X_1]\mathrm{E}[X_2]\ldots\mathrm{E}[X_n] = \prod_{j=1}^{n} \mathrm{E}[X_j],$$

and if $g_1, g_2, \ldots, g_n$ are well-behaved functions then also $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are mutually independent random variables, hence

$$\mathrm{E}[g_1(X_1)g_2(X_2)\ldots g_n(X_n)] = \mathrm{E}[g_1(X_1)]\mathrm{E}[g_2(X_2)]\ldots\mathrm{E}[g_n(X_n)] = \prod_{j=1}^{n} \mathrm{E}[g_j(X_j)].$$

21 Measures of pairwise dependence

Covariance function: for two random variables X and Y we define

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])],$$

which is equivalent to

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y].$$

Properties of covariance:

1. symmetry: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;
2. bilinearity

   $$\mathrm{Cov}(X_1+X_2, Y_1+Y_2) = \mathrm{Cov}(X_1,Y_1)+\mathrm{Cov}(X_1,Y_2)+\mathrm{Cov}(X_2,Y_1)+\mathrm{Cov}(X_2,Y_2),$$

   and for any $a, b \in \mathbb{R}$

   $$\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y);$$

3. relationship with variance: $\mathrm{Var}[X] = \mathrm{Cov}(X,X)$,
   $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y)$,
   $\mathrm{Var}[X-Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}(X,Y)$;
4. if X and Y are independent, $\mathrm{Cov}(X,Y) = 0$.

Correlation coefficient: for random variables X and Y,

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

Correlation is the degree of linear association between two variables. It is a scaled covariance, $|\mathrm{Corr}(X,Y)| \leq 1$. Moreover, $|\mathrm{Corr}(X,Y)| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$ (a linear relation between variables). If $\mathrm{Corr}(X,Y) = 1$ then $a > 0$, if $\mathrm{Corr}(X,Y) = -1$ then $a < 0$.

Uncorrelation and independence: $\mathrm{Corr}(X,Y) = 0$, i.e. X and Y are uncorrelated, if and only if

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

This result implies that

$$X, Y \text{ independent } \Rightarrow X, Y \text{ uncorrelated.}$$

but not the viceversa. Indeed correlation means only linear dependence.

Example: we know that independence implies

$$\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)].$$

for any g_1, g_2 well-behaved functions. Consider the discrete random variables X and Y such that the joint pmf is

$$f_{X,Y}(x,y) = \begin{cases} 1/4 & \text{if} \quad x=0 \quad \text{and} \quad y=1 \\ 1/4 & \text{if} \quad x=0 \quad \text{and} \quad y=-1 \\ 1/4 & \text{if} \quad x=1 \quad \text{and} \quad y=0 \\ 1/4 & \text{if} \quad x=-1 \quad \text{and} \quad y=0 \\ 0 & \text{otherwise} \end{cases}$$

Now, $\mathrm{E}[XY] = 0$ and $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, thus $\mathrm{Cov}(X,Y) = 0$, the variables are uncorrelated. If we now choose $g_1(X) = X^2$ and $g_2(Y) = Y^2$ we have $\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[X^2Y^2] = 0$, but

$$\mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)] = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \neq 0.$$

So X and Y are not independent.

Example: suppose X is a standard normal random variable, i.e. with $\mathrm{E}[X^k] = 0$ for k odd, and let $Y = X^2$. Clearly X and Y are not independent: if you know X, you also know Y. And if you know Y, you know the absolute value of X. The covariance of X and Y is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X^3] - 0 \cdot \mathrm{E}[Y] = \mathrm{E}[X^3] = 0.$$

Thus $\mathrm{Corr}(X,Y) = 0$, and we have a situation where the variables are not independent, yet they have no linear dependence. A linear correlation coefficient does not encapsulate anything about the quadratic dependence of Y upon X.

22 Joint moments and mgfs for two random variables

Expectation of a function of two random variables: if g is a well-behaved function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and X and Y are random variables with joint pmf or pdf function $f_{X,Y}$ then

$$\mathrm{E}[g(X,Y)] = \begin{cases} \sum_y \sum_x g(x,y) f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y, & \text{continuous case.} \end{cases}$$

Joint moments: if X and Y are random variables with joint pmf or pdf $f_{X,Y}$ then the $(r,s)^{\text{th}}$ joint moment is

$$\mu_{r,s} = \mathrm{E}[X^r Y^s] = \begin{cases} \sum_y \sum_x x^r y^s f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^r y^s f_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y, & \text{continuous case.} \end{cases}$$

Joint central moments: the $(r,s)^{\text{th}}$ joint central moment is

$$\begin{aligned} m_{r,s} &= \mathrm{E}[(X-\mathrm{E}[X])^r(Y-\mathrm{E}[Y])^s] = \\ &= \begin{cases} \sum_y \sum_x [(x-\mu_X)^r(y-\mu_Y)^s] f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty}[(x-\mu_X)^r(y-\mu_Y)^s] f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases} \end{aligned}$$

Properties of joint moments:

1. mean for X: $\mu_{1,0} = E[X]$;
2. r^{th} moment for X: $\mu_{r,0} = \mathrm{E}[X^r]$;
3. variance for X: $m_{2,0} = E[(X-\mathrm{E}[X])^2]$;
4. r^{th} central moment for X: $m_{r,0} = \mathrm{E}[(X-\mathrm{E}[X])^r]$;
5. covariance: $m_{1,1} = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] = \mathrm{Cov}(X,Y)$;
6. correlation: $m_{1,1}/\sqrt{m_{2,0}m_{0,2}} = \mathrm{Corr}(X,Y)$.

Joint moment generating function: given two random variables X and Y is a function $M_{X,Y} : \mathbb{R}\times\mathbb{R} \to [0,+\infty)$ such that

$$M_{X,Y}(t,u) = \mathrm{E}[e^{tX+uY}] = \begin{cases} \sum_y \sum_x e^{tx+uy} f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty} e^{tx+uy} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases}$$

Properties of joint mgfs:

1. Taylor expansion:
$$M_{X,Y}(t,u) = \mathrm{E}\left[\sum_{i=0}^{+\infty}\frac{(tX)^i}{i!}\sum_{j=0}^{+\infty}\frac{(uY)^j}{j!}\right] = \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty}\mathrm{E}[X^iY^j]\frac{t^iu^j}{i!j!};$$
2. the $(r,s)^{\text{th}}$ joint moment is the coefficient of $t^ru^s/(r!s!)$ in the Taylor expansion;
3. derivatives at zero:
$$\mu_{r,s} = \mathrm{E}[X^rY^s] = M_{X,Y}^{(r,s)}(0,0) = \left.\frac{\mathrm{d}^{r+s}}{\mathrm{d}t^r\mathrm{d}u^s}M_{X,Y}(t,u)\right|_{t=0,u=0};$$
4. moment generating function for marginals: $M_X(t) = \mathrm{E}[e^{tX}] = M_{X,Y}(t,0)$, $M_Y(u) = \mathrm{E}[e^{uY}] = M_{X,Y}(0,u)$;
5. if X and Y independent:
$$M_{X,Y}(t,u) = M_X(t)M_Y(u).$$

Joint cumulants: let $K_{X,Y}(t,u) = \log M_{X,Y}(t,u)$ be the joint cumulant generating function, then we define the $(r,s)^{\text{th}}$ joint cumulant $c_{r,s}$ as the coefficient of $(t^r u^s)/(r!s!)$ in the Taylor expansion of $K_{X,Y}$. Thus,

$$\mathrm{Cov}(X,Y) = c_{1,1} \text{ and } \mathrm{Corr}(X,Y) = \frac{c_{1,1}}{\sqrt{c_{2,0}c_{0,2}}}.$$

23 Joint moments and mgfs of n random variables

Multivariate generalisation: for random variables $X_1, \ldots, X_n$ with joint pmf or pdf $f_{X_1,\ldots,X_n}$:

1. joint moments:

$$\mu_{r_1,\ldots,r_n} = \mathrm{E}[X_1^{r_1} \ldots X_n^{r_n}]$$
$$= \begin{cases} \sum_{x_1} \ldots \sum_{x_n} x_1^{r_1} \ldots x_n^{r_n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_1^{r_1} \ldots x_n^{r_n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_n, & \text{continuous case;} \end{cases}$$

2. joint central moments:

$$m_{r_1,\ldots,r_n} = \mathrm{E}[(X_1 - \mathrm{E}[X_1])^{r_1} \ldots (X_n - \mathrm{E}[X_n])^{r_n}].$$

3. joint moment generating function:

$$M_{X_1,\ldots,X_n}(t_1,\ldots,t_n) = \mathrm{E}[e^{t_1X_1+\ldots+t_nX_n}],$$

   and the coefficient of $t_1^{r_1} \ldots t_n^{r_n}/(r_1! \ldots r_n!)$ in the Taylor expansion of $M_{X_1,\ldots,X_n}$ is $\mathrm{E}[X_1^{r_1} \ldots X_n^{r_n}]$;

4. independence: if $X_1, \ldots, X_n$ are independent then

$$M_{X_1,\ldots,X_n}(t_1,\ldots,t_n) = M_{X_1}(t_1) \ldots M_{X_n}(t_n) = \prod_{j=1}^{n} M_{X_j}(t_j);$$

5. joint cumulant generating function:

$$K_{X_1,\ldots,X_n}(t_1,\ldots,t_n) = \log(M_{X_1,\ldots,X_n}(t_1,\ldots,t_n)),$$

   and the $(r_1,\ldots,r_n)^{\text{th}}$ joint cumulant is defined as the coefficient of $(t_1^{r_1} \ldots t_n^{r_n})/(r_1! \ldots r_n!)$ in the Taylor expansion of $K_{X_1,\ldots,X_n}$.

24 Inequalities

Hölder's inequality: let p and q be two integers such that $\frac{1}{p} + \frac{1}{q} = 1$, if X belongs to L^p and Y belongs to L^q, then XY belong to L^1 and

$$\mathrm{E}[|XY|] \leq \mathrm{E}[|X|^p]^{1/p}\mathrm{E}[|Y|^q]^{1/q}.$$

Cauchy-Schwarz's inequality: this is Hölder's inequality when $p = q = 2$; if X and Y belong to L^2, then XY belongs to L^1 and

$$\mathrm{E}[|XY|] \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[Y^2]}.$$

As a consequence, if X and Y have variances σ_X^2 and σ_Y^2, then

$$|\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y,$$

which means $|\mathrm{Corr}(X, Y)| \leq 1$.

Minkowski's inequality: let $p \geq 1$, if X and Y belong to L^p, then $X + Y$ belongs to L^p and

$$\mathrm{E}[|X + Y|^p]^{1/p} \leq \mathrm{E}[|X|^p]^{1/p} + \mathrm{E}[|Y|^p]^{1/p}.$$

Reading

Casella and Berger, Sections 4.2 - 4.5 - 4.7.

25 Conditional distributions

When we observe more than one random variable their values may be related. By considering conditional probabilities we can improve our knowledge of a given random variable by exploiting the information we have about the other.

Conditional cumulative distribution function: given X and Y random variables with $P(X = x) > 0$, the distribution of Y conditional (given) to $X = x$ is defined as

$$F_{Y|X}(y|x) = P(Y \leq y|X = x).$$

It is a possibly different distribution for every value of X, we have a family of distributions.

Conditional probability mass function: given X and Y discrete random variables with $P(X = x) > 0$, the conditional pmf of Y given $X = x$ is

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

such that the conditional cdf is

$$F_{Y|X}(y|x) = \sum_{y_i \leq y} f_{Y|X}(y_i|x).$$

Conditional probability density function: given X and Y jointly continuous random variables with $f_X(x) > 0$, the conditional pdf of Y given $X = x$ is

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

such that the conditional cdf is

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} \mathrm{d}v.$$

Conditional, joint and marginal densities: given $f_X(x) > 0$ we have:

1. conditional pmf or pdf:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}, & \text{discrete case,} \\ \\ \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy}, & \text{continuous case;} \end{cases}$$

2. joint pmf or pdf:

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

3. marginal pmf or pdf:

$$f_Y(y) = \begin{cases} \sum_x f_{Y|X}(y|x) f_X(x), & \text{discrete case,} \\ \\ \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx, & \text{continuous case;} \end{cases}$$

4. reverse conditioning (if also $f_Y(y) > 0$):

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

These are all direct implications of Bayes' theorem.

26 Conditional moments and mgfs

Conditional expectation: given X and Y random variables the expectation of Y given $X = x$ is

$$\mathrm{E}[Y|X = x] = \begin{cases} \sum_y y f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) \mathrm{d}y, & \text{continuous case.} \end{cases}$$

If we consider all possible values taken by X then we have a new random variable which is the conditional expectation of Y given X and it is written as $\mathrm{E}[Y|X]$. It is the best guess of Y given the knowledge of X. All properties of expectations still hold.

Law of iterated expectations: since $\mathrm{E}[Y|X]$ is a random variable we can take its expectation:

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y].$$

Indeed, in the continuous case

$$\begin{aligned}\mathrm{E}[\mathrm{E}[Y|X]] &= \int_{-\infty}^{+\infty} \mathrm{E}[Y|X=x] f_X(x)\mathrm{d}x = \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} f_X(x)\mathrm{d}x\mathrm{d}y = \mathrm{E}[Y].\end{aligned}$$

A useful consequence is that we can compute $\mathrm{E}[Y]$ without having to refer to the marginal pmf or pdf of Y:

$$\mathrm{E}[Y] = \begin{cases} \sum_x \mathrm{E}[Y|X=x] f_X(x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} \mathrm{E}[Y|X=x] f_X(x)\mathrm{d}x, & \text{continuous case.} \end{cases}$$

Conditional expectations of function of random variables: if g is a well-behaved, real-valued function, the expectation of $g(Y)$ given $X = x$ is defined as:

$$\mathrm{E}[g(Y)|X=x] = \begin{cases} \sum_y g(y) f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} g(y) f_{Y|X}(y|x)\mathrm{d}y, & \text{continuous case.} \end{cases}$$

The conditional expectation of $g(Y)$ given X is written as $\mathrm{E}[g(Y)|X]$ and it is also a random variable.

As a consequence any function of X can be treated as constant with respect to expectations conditional on X. In general for well-behaved functions g_1 and g_2

$$\mathrm{E}[g_1(X)g_2(Y)|X] = g_1(X)\mathrm{E}[g_2(Y)|X].$$

Notice that also $\mathrm{E}[Y|X]$ is a function of X so

$$\mathrm{E}[\mathrm{E}[Y|X]Y|X] = \mathrm{E}[Y|X]\mathrm{E}[Y|X] = (\mathrm{E}[Y|X])^2.$$

Conditional variance: for random variables X and Y, it is defined as

$$\begin{aligned}\mathrm{Var}[Y|X=x] &= \mathrm{E}[(Y - \mathrm{E}[Y|X=x])^2|X=x] = \\ &= \begin{cases} \sum_y [y - \mathrm{E}[Y|X=x]]^2 f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{\infty} [y - \mathrm{E}[Y|X=x]]^2 f_{Y|X}(y|x)dy, & \text{continuous case.} \end{cases}\end{aligned}$$

The conditional variance of Y given X is written as $\text{Var}[Y|X]$ and it is a random variable function of X. Moreover,

$$\text{Var}[Y|X] = \text{E}[Y^2|X] - (\text{E}[Y|X])^2,$$

By using the law of iterated expectations,

$$\begin{aligned}
\text{Var}[Y] &= \text{E}[Y^2] - (\text{E}[Y])^2 = \\
&= \text{E}[\text{E}[Y^2|X]] - \{\text{E}[\text{E}[Y|X]]\}^2 = \\
&= \text{E}[\text{Var}[Y|X] + (\text{E}[Y|X])^2] - \{\text{E}[\text{E}[Y|X]]\}^2 \\
&= \text{E}[\text{Var}[Y|X]] + \text{E}[\{\text{E}[Y|X]\}^2] - \{\text{E}[\text{E}[Y|X]]\}^2 \\
&= \text{E}[\text{Var}[Y|X]] + \text{Var}[\text{E}[Y|X]],
\end{aligned}$$

This result tells us that

$$\text{Var}[Y] \geq \text{E}[\text{Var}[Y|X]],$$

the expected value of the conditional variance is in general smaller than the unconditional variance. If X contains useful information for Y then conditioning on X makes uncertainty about the value of Y smaller. The case in which equality holds is when $\text{Var}[\text{E}[Y|X]] = 0$, i.e. when $\text{E}[Y|X]$ is no more random, which is when X contains no information on Y, i.e. they are independent.

Conditional distributions and independence: if X and Y are independent random variables then for cdfs we have

$$\begin{aligned}
F_{Y|X}(y|x) &= F_Y(y) \quad \forall\, x, y \in \mathbb{R}, \\
F_{X|Y}(x|y) &= F_X(x) \quad \forall\, x, y \in \mathbb{R}.
\end{aligned}$$

and for pmfs or pdfs we have

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y) \quad \forall\, x, y \in \mathbb{R}, \\
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x) \quad \forall\, x, y \in \mathbb{R}.
\end{aligned}$$

Finally,

$$\text{E}[Y|X] = \text{E}[Y].$$

Conditional moment generating function: given $X = x$, it is the function defined as

$$M_{Y|X}(u|x) = \text{E}[e^{uY}|X = x] = \begin{cases} \sum_y e^{uy} f_{Y|X}(y|x), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} e^{uy} f_{Y|X}(y|x) \mathrm{d}y, & \text{continuous case.} \end{cases}$$

This is a conditional expectation so it is a random variable. We can calculate the joint mgf and marginal mgfs from the conditional mgf,

$$\begin{aligned}
M_{X,Y}(t,u) &= \text{E}[e^{tX+uY}] = \text{E}[e^{tX} M_{Y|X}(u|X)], \\
M_Y(u) &= M_{X,Y}(0,u) = \text{E}[M_{Y|X}(u|X)].
\end{aligned}$$

Example: suppose that X is the number of hurricanes that form in the Atlantic basin in a given year and Y is the number making landfall. We assume we know that each hurricane has a probability p of making landfall independent of other hurricanes. If we know the number of hurricanes that form say x we can view Y as the number of success in x independent Bernoulli trials, i.e. $Y|X = x \sim \text{Bin}(x, p)$. If we also know that $X \sim \text{Pois}(\lambda)$, then we can compute the distribution of Y (notice that $X \geq Y$)

$$\begin{aligned}
f_Y(y) &= \sum_{x=y}^{+\infty} f_{Y|X}(y|x) f_X(x) = \\
&= \sum_{x=y}^{+\infty} \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!} = \\
&= \frac{\lambda^y p^y e^{-\lambda}}{y!} \sum_{x=y}^{+\infty} \frac{[\lambda(1-p)]^{x-y}}{(x-y)!} = \\
&= \frac{\lambda^y p^y e^{-\lambda}}{y!} \sum_{j=0}^{+\infty} \frac{[\lambda(1-p)]^{j}}{j!} = \\
&= \frac{\lambda^y p^y e^{-\lambda}}{y!} e^{\lambda(1-p)} = \\
&= \frac{(\lambda p)^y e^{-\lambda p}}{y!},
\end{aligned}$$

thus $Y \sim \text{Pois}\lambda p$. So $\text{E}[Y] = \lambda p$ and $\text{Var}[Y] = \lambda p$, but we could find these results without the need of the marginal pdf. Since $Y|X = x \sim \text{Bin}(x, p)$, then

$$\text{E}[Y|X = x] = Xp \qquad \text{Var}[Y|X = x] = Xp(1-p)$$

Since $X \sim \text{Pois}(\lambda)$, by using the law of iterated expectations, we have

$$\text{E}[Y] = \text{E}[\text{E}[Y|X = x]] = \text{E}[X]p = \lambda p$$

and

$$\text{Var}[Y] = \text{E}[\text{Var}[Y|X = x]] + \text{Var}[\text{E}[Y|X = x]] = \text{E}[X]p(1-p) + \text{Var}[Xp] = \lambda p(1-p) + \lambda p^2 = \lambda p.$$

Alternatively we can use the mgfs, we have

$$M_X(t) = \exp\{\lambda(e^t - 1)\} \qquad M_{Y|X}(u|X) = (1 - p + pe^u)^X,$$

therefore

$$\begin{aligned}
M_Y(u) &= \text{E}[M_{Y|X}(u|X)] = \text{E}[(1 - p + pe^u)^X] = \\
&= \text{E}[\exp\{X \log(1 - p + pe^u)\}] = \\
&= M_X(\log(1 - p + pe^u)) = \\
&= \exp\{\lambda(1 - p + pe^u - 1)\} = \\
&= \exp\{\lambda p(e^u - 1)\},
\end{aligned}$$

which is the mgf of a Poisson distribution.

27 An example of bivariate distribution

Consider the function

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

- It is a valid density, indeed it is a positive real valued function and it is normalized

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y &= \int_0^1\int_0^1 (x+y)\mathrm{d}x\mathrm{d}y = \\ &= \int_0^1 \left[\frac{x^2}{2}+xy\right]_0^1 \mathrm{d}y = \int_0^1 \left[\frac{1}{2}+y\right]\mathrm{d}y = \\ &= \left[\frac{y}{2}+\frac{y^2}{2}\right]_0^1 = 1. \end{aligned}$$

- The joint cdf is

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\mathrm{d}u\mathrm{d}v = \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} (u+v)\mathrm{d}u\mathrm{d}v = \\ &= \int_0^y \left[\frac{x^2}{2}+xv\right]\mathrm{d}v = \left[\frac{x^2v}{2}+\frac{xv^2}{2}\right]_0^1 = \\ &= \frac{1}{2}xy(x+y) \quad \text{for } 0<x<1, 0<y<1. \end{aligned}$$

  More precisely we have

$$F_{X,Y}(x,y) = \begin{cases} \frac{1}{2}xy(x+y) & \text{if } 0<x<1 \text{ and } 0<y<1, \\ \frac{1}{2}x(x+1) & \text{if } 0<x<1 \text{ and } y\geq 1, \\ \frac{1}{2}y(y+1) & \text{if } x\geq 1 \text{ and } 0<y<1, \\ 1 & \text{if } x\geq 1 \text{ and } y\geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- The marginal pdf of X is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}y = \\ &= \int_0^1 (x+y)\mathrm{d}y = x+\frac{1}{2}. \end{aligned}$$

- We can compute probabilities as $P(2X < Y)$, we first define the event $B =$

$\{(x,y)$ s.t. $0<x<\frac{y}{2}, 0<y<1\}$ then

$$\begin{aligned} P(2X<Y)=P(B) &= \int\int_B f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \\ &= \int_0^1\int_0^{y/2}(x+y)\mathrm{d}x\mathrm{d}y = \int_0^1\left[\frac{y^2}{8}+\frac{y^2}{2}\right]\mathrm{d}y = \\ &= \left[\frac{y^3}{24}+\frac{y^3}{6}\right]_0^1 = \frac{5}{24}. \end{aligned}$$

Analogously we could define $C=\{(x,y)$ s.t. $0<x<\frac{1}{2}, 2x<y<1\}$ and compute $P(C)$.

- the $(r,s)^{\text{th}}$ joint moment is

$$\begin{aligned} \mathrm{E}[X^rY^s] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^ry^sf_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_0^1\int_0^1 x^ry^s(x+y)\mathrm{d}x\mathrm{d}y = \int_0^1\left[\frac{1}{r+2}y^s+\frac{1}{r+1}y^{s+1}\right]\mathrm{d}y = \\ &= \left[\frac{1}{(r+2)(s+1)}y^{s+1}+\frac{1}{(r+1)(s+2)}y^{s+2}\right]_0^1 = \frac{1}{(r+2)(s+1)}+\frac{1}{(r+1)(s+2)}. \end{aligned}$$

Thus, $\mathrm{E}[XY]=\frac{1}{3}$, $\mathrm{E}[X]=\mathrm{E}[Y]=\frac{7}{12}$, $\mathrm{E}[X^2]=\frac{5}{12}$ so $\mathrm{Var}[X]=\frac{11}{144}$ and finally

$$\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]=\frac{1}{3}-\frac{49}{144}=-\frac{1}{144},$$

and $\mathrm{Corr}(X,Y)=-\frac{1}{11}$, so X and Y are not independent.
We find this result also by noticing that given the marginals and the joint pdfs we have

$$f_X(x)f_Y(y)=xy+\frac{x+y}{2}+\frac{1}{4},$$

therefore $f_X(x)f_Y(y)\neq f_{X,Y}(x,y)$ so X and Y are not independent.

- The conditional pdf of Y given $X=x$ is

$$f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_X(x)}=\begin{cases}\frac{x+y}{x+\frac{1}{2}} & \text{if } 0<y<1\\ 0 & \text{otherwise.}\end{cases}$$

- The conditional expectation of Y given $X=x$ is

$$\begin{aligned} \mathrm{E}[Y|X=x] &= \int_{-\infty}^{+\infty} yf_{Y|X}(y|x)\mathrm{d}y = \\ &= \int_0^1 y\frac{x+y}{x+\frac{1}{2}}\mathrm{d}y = \\ &= \frac{1}{x+\frac{1}{2}}\left[\frac{xy^2}{2}+\frac{y^3}{3}\right]_0^1 = \\ &= \frac{3x+2}{6x+3}. \end{aligned}$$

- we can use the law of iterated expectations

$$\begin{aligned}\mathrm{E}[\mathrm{E}[Y|X=x]] &= \int_0^1 \frac{3x+2}{6x+3}\left(x+\frac{1}{2}\right)\mathrm{d}x = \\ &= \frac{1}{6}\int_0^1 3x+2\mathrm{d}x \\ &= \frac{1}{6}\left(\frac{3}{2}+2\right) \\ &= \frac{7}{12} = \mathrm{E}[Y].\end{aligned}$$

Reading

Casella and Berger, Sections 4.2 - 4.4 - 4.5.

28 Sums of random variables

We start with the bivariate case and then we generalise it to n variables.

Moments of a sum: if X and Y are random variables then:

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y], \qquad \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y),$$

and, by using the linearity of expectations and the binomial expansion, we have for $r \in \mathbb{N}$

$$\mathrm{E}[(X+Y)^r] = \sum_{j=0}^{r}\binom{r}{j}\mathrm{E}[X^j Y^{r-j}].$$

Probability mass/density function of a sum: if X and Y are random variables with joint density $f_{X,Y}(x,y)$ and we define $Z = X+Y$ then the pmf/pdf of Z is

$$f_Z(z) = \begin{cases} \sum_u f_{X,Y}(u, z-u), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} f_{X,Y}(u, z-u)\mathrm{d}u, & \text{continuous case.}\end{cases}$$

In the continuous case just change variables $X = U$ and $Y = Z - U$. In the discrete case notice that

$$\{X+Y=z\} = \bigcup_u \{X = u \cap Y = z-u\}$$

and, since this is a sum of disjoint events, for any u, we have

$$P(X+Y=z) = \sum_u P(X = u \cap Y = z-u).$$

Probability mass/density function of a sum of independent random variables: if X and Y are independent random variables and we define $Z = X + Y$ then the pmf/pdf of Z is

$$f_Z(z) = \begin{cases} \sum_u f_X(u) f_Y(z-u), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u, & \text{continuous case.} \end{cases}$$

This operation is known as convolution

$$f_Z = f_X * f_Y \Leftrightarrow \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u.$$

Convolution is commutative so $f_X * f_Y = f_Y * f_X$.

Moment generating function of the sum of independent random variables: if X and Y are independent random variables and we define $Z = X + Y$ then the mgf of Z is

$$M_Z(t) = M_X(t) M_Y(t),$$

and the cumulant generating function is

$$K_Z(t) = K_X(t) + K_Y(t).$$

Example: suppose the X and Y are independent r.v. exponentially distributed, $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\theta)$, with $\lambda \neq \theta$, then the pdf of $Z = X + Y$ is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u = \\ &= \int_0^z \lambda e^{-\lambda u} \theta e^{-\theta(z-u)} \mathrm{d}u = \\ &= \lambda\theta e^{-\theta z} \left[\frac{-1}{\lambda - \theta} e^{-(\lambda-\theta)u} \right]_0^z = \\ &= \frac{\lambda\theta}{\lambda - \theta} (e^{-\theta z} - e^{-\lambda z}) \qquad 0 \leq z < +\infty. \end{aligned}$$

Note the domain of integration $[0, z]$. Indeed, since both X and Y are positive r.v., also U and $Z - U$ have to be positive, thus we need $0 < U \leq Z$.
In theory, we could also use mgfs, but in this case we get a function of t that does not have an expression that resembles one of a known distribution.

Example: suppose the X and Y are independent r.v. normally distributed, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then then to compute the pdf of $Z = X + Y$ we use the cumulant generating functions

$$K_X(t) = \mu_X t + \frac{\sigma_X^2 t^2}{2}, \quad K_Y(t) = \mu_Y t + \frac{\sigma_Y^2 t^2}{2},$$

and

$$K_Z(t) = (\mu_X + \mu_Y) t + \frac{(\sigma_X^2 + \sigma_Y^2) t^2}{2}$$

by uniqueness of cumulant generating functions $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Multivariate generalization: for n independent random variables $X_1, \ldots, X_n$ let $S = \sum_{j=1}^n X_j$ then

1. the pmf/pdf of S is
$$f_S = f_{X_1} * \ldots * f_{X_n};$$
2. the mgf of $S$is
$$M_S(t) = M_{X_1}(t) \ldots M_{X_n}(t).$$
3. if $X_1, \ldots, X_n$ are also identically distributed they have a common mgf $M_X(t)$ thus
$$f_S = \underbrace{f * f * \ldots * f}_{n-\text{times}}, \quad M_S(t) = [M_X(t)]^n, \quad K_S(t) = nK_X(t).$$

To indicate independent and identically distributed random variables we use the notation i.i.d.

Example: given n i.i.d. Bernoulli r.v. $X_1 \ldots X_n$ with probability p and mgf
$$M_X(t) = 1 - p + pe^t,$$
the sum $S = \sum_{j=1}^n X_j$ has mgf
$$M_S(t) = (1 - p + pe^t)^n,$$
thus, by uniqueness of mgf, $S \sim \text{Bin}(n, p)$.

Example: given $X_1, \ldots, X_n$ independent r.v. normally distributed $X_j \sim N(\mu_j, \sigma_j^2)$ then, for fixed constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, we have
$$S = \sum_{j=1}^n (a_j X_j + b_j) \sim N\left(\sum_{j=1}^n (a_j \mu_j + b_j), \sum_{j=1}^n a_j^2 \sigma_j^2\right).$$
If $X_j \sim iidN(\mu, \sigma^2)$, then
$$S = \sum_{j=1}^n X_j \sim N(n\mu, n\sigma^2).$$

Other examples of sums of independent random variables

1. Poisson:
$$X \sim \text{Pois}(\lambda_1),\ Y \sim \text{Pois}(\lambda_2) \Rightarrow Z \sim \text{Pois}(\lambda_1 + \lambda_2)$$
$$X_j \sim iid\text{Pois}(\lambda) \Rightarrow S \sim \text{Pois}(n\lambda) \quad j = 1, \ldots n;$$

2. Gamma:

$$X \sim \text{Gamma}(r_1, \theta),\ Y \sim \text{Gamma}(r_2, \theta) \Rightarrow Z \sim \text{Gamma}(r_1 + r_2, \theta)$$

$$X_j \sim iid\text{Exp}(\lambda) \Rightarrow S \sim \text{Gamma}(n, \lambda) \quad j = 1, \dots n;$$

3. Binomial:

$$X \sim \text{Bin}(n_1, p),\ Y \sim \text{Bin}(n_2, p) \Rightarrow Z \sim \text{Bin}(n_1 + n_2, p)$$

$$X_j \sim iid\text{Bin}(k, p) \Rightarrow S \sim \text{Bin}(nk, p) \quad j = 1, \dots n.$$

29 Limit theorems for Bernoulli sums

Assume to observe n independent Bernoulli trials X_i with an unknown probability of success p. We study the behaviour of the process $S_n = \sum_{i=1}^{n} X_i$ which counts the number of successes in n trials. If $X_i \sim iid\text{Bernoulli}(p)$, then $S_n \sim \text{Bin}(n, p)$. For any i we have that $\text{E}[X_i] = p$ and $\text{Var}[X_i] = p(1-p)$ so that $\text{E}[S_n] = np$ and $\text{Var}[S_n] = np(1-p)$.

Law of Large Numbers: there are two forms of this law:

1. Weak Law of Large Numbers: as $n \to +\infty$, $S_n/n \overset{m.s.}{\to} p$, i.e.

$$\lim_{n \to +\infty} \text{E}\left[\left(\frac{S_n}{n} - p\right)^2\right] = 0,$$

   which implies $S_n/n \overset{P}{\to} p$, i.e.

$$\lim_{n \to \infty} P\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) = 1, \quad \forall \epsilon > 0;$$

2. Strong Law of Large Numbers: as $n \to +\infty$, $S_n/n \overset{a.s.}{\to} p$, i.e.

$$P\left(\lim_{n \to \infty} \left|\frac{S_n}{n} - p\right| = 0\right) = 1.$$

The law establishes the convergence of the empirical average (or sample mean) S_n/n to the expected value of X_i, i.e. to p (or population mean). It is uesful if we observe many Bernoulli trials and we want to determine p: it is a first example of inference.

Proof of the weak law: for each n we have

$$\text{E}\left[\left(\frac{S_n}{n} - p\right)^2\right] = \frac{\text{E}\left[(S_n - np)^2\right]}{n^2} = \frac{\text{Var}[S_n]}{n^2} = \frac{p(1-p)}{n} \to 0, \quad \text{as } n \to +\infty.$$

Example: when tossing a coin X_i is 1 if we get head or 0 if we get tail (a Bernoulli trial), S_n is the number of heads we get in n independent tosses. The frequency of heads will converge to $1/2$ which is the value of p in this particular case.

The following result is a special case of the Central Limit Theorem which we shall see in due course.

De Moivre-Laplace Limit Theorem: as $n \to +\infty$, and for $Z \sim N(0,1)$,

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{S_n/n - p}{\sqrt{p(1-p)}} \leq \alpha\right) = P(Z \leq \alpha) = \int_{-\infty}^{\alpha} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z, \quad \forall\, \alpha \in \mathbb{R},$$

which implies

$$\sqrt{n}\frac{S_n/n - p}{\sqrt{p(1-p)}} \xrightarrow{d} Z.$$

We are saying that the sample mean (which is a random variable) of the Bernoulli trials converges in distribution or is asymptotically distributed as a normal random variable with mean p (this we know already from the law of large numbers) and variance $p(1-p)/n$, thus the more trials we observe the smaller the uncertainty about the expected value of the sample mean, the rate of convergence being $\sqrt{n}$. This result contains useful informations not only on the point-wise estimate of the population mean but also on the uncertainty and the speed with which we have convergence.

Finally, remember that $S_n \sim \mathrm{Bin}(np, np(1-p))$, then, by rearranging the terms, we have

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z.$$

i.e. the Binomial distribution can be approximated by a normal distribution with mean np and variance $np(1-p)$.

Reading

Casella and Berger, Sections 5.2

30 Mixtures and random sums

Hierarchies and mixtures: suppose we are interested in a random variable Y which has a distribution that depends on another random variables, say X. This is called a hierarchical model and Y has a mixture distribution. In the first instance we do not know the marginal distribution of Y directly, but we know the conditional distribution of Y given

$X = x$ and the marginal distribution of X (see the example on hurricanes of Section 26).

The key results which are necessary for characterising Y, are

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}[Y|X]] \\ \mathrm{Var} &= \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathrm{E}[Y|X]] \\ f_Y(y) &= \mathrm{E}[f_{Y|X}(y|X)] \quad \text{and } M_Y(t) = \mathrm{E}[M_{Y|X}(t|X)] \end{aligned}$$

Example: Poisson mixing. If $Y|\Lambda = \lambda \sim \mathrm{Pois}(\lambda)$, for some positive r.v. Λ, then

$$\mathrm{E}[Y|\Lambda] = \mathrm{Var}[Y|\Lambda] = \Lambda.$$

Therefore,

$$\mathrm{E}[Y] = \mathrm{E}[\Lambda], \qquad \mathrm{Var}[Y] = \mathrm{E}[\Lambda] + \mathrm{Var}[\Lambda].$$

Random sums: We consider the case in which $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables and $Y = \sum_{j=1}^{N} X_j$, where N is also a random variable which is independent of each X_i. Y is called random sum and can be viewed as a mixture such that $Y|N = n$ is a sum of random variables, so all results of previous section still hold.

Conditional results for random sums: suppose that $\{X_j\}$ is a sequence of i.i.d. random variables with mean $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$, for any j, and suppose that N is a random variable taking only positive integer values and define $Y = \sum_{j=1}^{N} X_j$, then

$$\begin{aligned} \mathrm{E}[Y|N] &= N\mathrm{E}[X], \\ \mathrm{Var}[Y|N] &= N\mathrm{Var}[X], \\ M_{Y|N}(t|N) &= [M_X(t)]^N \quad \text{and } K_{Y|N}(t|N) = NK_X(t). \end{aligned}$$

Marginal results for random sums: suppose that $\{X_j\}$ is a sequence of i.i.d. random variables with mean $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$, for any j, and suppose that N is a random variable taking only positive integer values and define $Y = \sum_{j=1}^{N} X_j$, then

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[N]\mathrm{E}[X], \\ \mathrm{Var}[Y] &= \mathrm{E}[N]\mathrm{Var}[X] + \mathrm{Var}[N]\{\mathrm{E}[X]\}^2, \\ M_Y(t) &= M_N(\log M_X(t)) \quad \text{and } K_Y(t) = K_N(K_X(t)). \end{aligned}$$

Example: each year the value of claims made by an owner of a health insurance policy is distributed exponentially with mean α independent of previous years. At the end of each year with probability p the individual will cancel her policy. We want the distribution of the total cost of the health insurance policy for the insurer. The value of claims in year j is X_j and the number of years in which the policy is held is N, thus

$$X_j \sim iid\mathrm{Exp}\left(\frac{1}{\alpha}\right), \qquad N \sim \text{Geometric}(p).$$

The total cost for the insurer is $Y = \sum_{j=1}^{N} X_j$. Therefore, $\mathrm{E}[Y] = \alpha\frac{1}{p}$. To get the distribution we use the cumulant generating function

$$K_X(t) = -\log(1-\alpha t), \qquad K_N(t) = -\log\left(1 - \frac{1}{p} + \frac{1}{p}e^{-t}\right),$$

and

$$K_Y(t) = K_N(K_X(t)) = -\log\left(1 - \frac{1}{p} + \frac{1}{p}(1-\alpha t)\right) = -\log\left(1 - \frac{\alpha}{p}t\right),$$

by uniqueness we have that $Y \sim \mathrm{Exp}\left(\frac{p}{\alpha}\right)$.

The Poisson approximation: assume to have $X_j \sim iid\mathrm{Bernoulli}(p)$, and $N \sim \mathrm{Pois}(\lambda)$. Consider $Y = \sum_{j=1}^{N} X_j$, then $Y|N = n \sim \mathrm{Bin}(n,p)$ and

$$\begin{aligned}
\mathrm{E}[Y] &= \lambda\mathrm{E}[X], \\
\mathrm{Var}[Y] &= \lambda\mathrm{E}[X^2], \\
M_Y(t) &= M_N(\log M_X(t)) = e^{\lambda(M_X(t)-1)}, \\
K_S(t) &= \lambda(M_X(t)-1).
\end{aligned}$$

By using the mgf of a Bernoulli $M_X(t) = 1 - p + pe^t$ we get

$$M_Y(t) = e^{\lambda(M_X(t)-1)} = e^{\lambda p(e^t-1)},$$

by uniqueness of mgf, $Y \sim \mathrm{Pois}(\lambda p)$ (see the example on hurricanes of Section 26).

Reading

Casella and Berger, Section 4.4

31 Random vectors

This is just a way to simplify notation when we consider n random variables. Expectations are element wise and we have to remember that the variance of a vector is a matrix.

Random vector: an n-dimensional vector of random variables, i.e. a function

$$\mathbf{X} = (X_1, \ldots, X_n)^T : \Omega \to \mathbb{R}^n.$$

The cdf, pmf or pdf, and mgf of a random vector are the joint cdf, pmf or pdf, and mgf of $X_1, \ldots, X_n$ so, for any $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$,

$$\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &= F_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \\
f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \\
M_{\mathbf{X}}(\mathbf{t}) &= M_{X_1,\ldots,X_n}(t_1, \ldots, t_n).
\end{aligned}$$

Expectation of a random vector: the expectation of a random vector is a vector of the expectations, i.e. it is taken element by element

$$\mathrm{E}[\mathbf{X}] = \begin{pmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_n] \end{pmatrix}.$$

For jointly continuous random variables we have

$$\begin{aligned} \mathrm{E}[\mathbf{X}] &= \int_{\mathbb{R}^n} \mathbf{x} f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x} = \\ &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} x_1 \dots x_n f_{X_1,\dots,X_n}(x_1,\dots,x_n) \mathrm{d}x_1 \dots \mathrm{d}x_n. \end{aligned}$$

Variance-covariance matrix: given n random variables $X_1, \dots, X_n$ we know what is the variance of each of them and we know the covariance of each couple. All these informations can be summarized in just one object, defined as

$$\mathbf{\Sigma} = \mathrm{Var}[\mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T],$$

Where $\mathbf{X}$ is $n \times 1$ (a column vector), then $\mathbf{X}^T$ is $1 \times n$ (a row vector), and $\mathbf{\Sigma}$ is a $n \times n$ matrix. Taking element by element expectation of this matrix we get

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \dots & \dots & \mathrm{Var}[X_n] \end{pmatrix}.$$

The matrix is symmetric and if the variables are uncorrelated then it is a diagonal matrix. If the variables are also identically distributed then $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$ where σ^2 is the variance of each random variable and $\mathbf{I}_n$ is the n-dimensional identity matrix. Finally, as the univaraite variance is always positive, in this case we have that $\mathbf{\Sigma}$ is a non-negative definite matrix, i.e.

$$\mathbf{b}^T \mathbf{\Sigma} \mathbf{b} \geq 0 \qquad \forall\, \mathbf{b} \in \mathbb{R}^n.$$

Example: if $N = 2$ and assume $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ then

$$\Sigma = \mathrm{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix} (X\ Y) \right] = \mathrm{E}\begin{bmatrix} X^2 & XY \\ YX & Y^2 \end{bmatrix} = \begin{pmatrix} \mathrm{E}[X^2] & \mathrm{E}[XY] \\ \mathrm{E}[YX] & \mathrm{E}[Y^2] \end{pmatrix} = \begin{pmatrix} \mathrm{Var}[X] & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}[Y] \end{pmatrix}.$$

Conditioning for random vectors: if $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, and if $f_{\mathbf{X}}(\mathbf{x}) > 0$, we can define the conditional pdf/pmf as

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

or

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

Decomposition of probability mass/density function: given an n-dimensional random vector $\mathbf{X}$ and given $\mathbf{x} \in \mathbb{R}^n$, then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_n|X_{n-1}\ldots X_1}(x_n|x_{n-1}\ldots x_1) f_{X_{n-1}|X_{n-2}\ldots X_1}(x_{n-1}|x_{n-2}\ldots x_1)\ldots f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1) = \\ &= \prod_{j=1}^{n} f_{X_j|\mathbf{X}_{j-1}}(x_j|\mathbf{x}_{j-1}), \end{aligned}$$

where the random vector $\mathbf{X}_{j-1}$ is the random vector $\mathbf{X}$ without its j-th element.

Example: consider 3 r.v. X_1, X_2 and X_3, we can group them in different ways and we get for example

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3|x_1,x_2) f_{X_1,X_2}(x_1,x_2),$$

and applying again the definition above to the joint pdf/pmf of X_1 and X_2 we have

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3|x_1,x_2) f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1).$$

32 Multivariate normal distribution

We start with the bivariate case. We want a bivariate version of the normal distribution. Given two standard normal random variables, we can build a bivariate normal that depends only on their correlation.

Standard bivariate normal: given U and V i.i.d. standard normal random variables, and for some number $|\rho| < 1$, define $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$, then we can prove that

1. $X \sim N(0,1)$ and $Y \sim N(0,1)$;
2. $\mathrm{Corr}(X,Y) = \rho$;
3. the joint pdf is that of a standard bivariate normal random variable and depends only on the parameter ρ:

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

The random vector $\mathbf{X} = (X,Y)^T$ is normally distributed and we write

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

or $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}_{X,Y})$ where $\mathbf{\Sigma}_{X,Y}$ is the 2×2 variance covariance matrix;

4. the joint mgf is

$$M_{X,Y}(s,t) = \exp\left[\frac{1}{2}(s^2 + 2\rho st + t^2)\right].$$

Bivariate normal for independent random variables: if the random variables U and V are independent and standard normal, the joint pdf and mgf are

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-(u^2+v^2)/2}, \\ M_{U,V}(s,t) &= e^{(s^2+t^2)/2}. \end{aligned}$$

The random vector (U, V) is normally distributed with variance covariance matrix

$$\mathbf{\Sigma}_{U,V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Computing the joint pdf: given $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$, we have to compute $f_{X,Y}(x,y)$ given $f_{U,V}(u,v)$. Given the function $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\mathbf{h}(X,Y) = (U,V)$ and the domain of $\mathbf{h}$ is $C \subseteq \mathbb{R}^2$ and it is in one-to-one correspondence with the support of (U, V), we have the rule

$$f_{X,Y}(x,y) = \begin{cases} f_{U,V}(\mathbf{h}(x,y))|J_{\mathbf{h}}(x,y)| & \text{for } (x,y) \in C \\ 0 & \text{otherwise} \end{cases}$$

where

$$J_{\mathbf{h}}(x,y) = \det \begin{pmatrix} \frac{\partial}{\partial x}h_1(x,y) & \frac{\partial}{\partial x}h_2(x,y) \\ \\ \frac{\partial}{\partial y}h_1(x,y) & \frac{\partial}{\partial y}h_2(x,y) \end{pmatrix}.$$

In this case, $C = \mathbb{R}^2$,

$$u = h_1(x,y) = x, \qquad v = h_2(x,y) = \frac{y - \rho x}{\sqrt{1-\rho^2}},$$

and $|J_{\mathbf{h}}(x,y)| = \frac{1}{\sqrt{1-\rho^2}}$, thus

$$f_{X,Y}(x,y) = f_{U,V}\left(x, \frac{y-\rho x}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}}.$$

Generic bivariate normal: if $X^* = \mu_X + \sigma_X X$ and $Y^* = \mu_Y + \sigma_Y Y$ then $X^* \sim N(\mu_X, \sigma_X^2)$ and $Y^* \sim N(\mu_Y, \sigma_Y^2)$ with $\mathrm{Corr}(X^*, Y^*) = \rho$ and the joint pdf is

$$f_{X^*,Y^*}(x,y) = \frac{1}{\sigma_X \sigma_Y} f_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}, \frac{y-\mu_Y}{\sigma_Y}\right).$$

A generic jointly normal random vector is distributed as

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \sigma_Y^2 \end{pmatrix}\right).$$

Conditional distribution: of Y^* given X^* is

$$Y^*|X^* = x \ \sim \ N\left(\mu_y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\right).$$

It is obtained by using the joint and the marginal pdfs.

Multivariate case

1. Multivariate normal density: let $X_1, \ldots, X_n$ be random variables and define the $n \times 1$ random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$. If $X_1, \ldots, X_n$ are jointly normal then $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}]$ is an $n \times 1$ vector and the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}]$ is an $n \times n$ matrix whose $(i,j)^{\text{th}}$ entry is $\mathrm{Cov}(X_i, X_j)$. The joint density functions is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}|\det \boldsymbol{\Sigma}|^{-1/2} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

2. Conditional expectation for multivariate normal: suppose that $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$, for some integers n and m, and $\mathbf{X} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\mathbf{Y} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. If, $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}'_{YX}$, then

$$\begin{aligned}\mathrm{E}[\mathbf{Y}|\mathbf{X}] =& \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}(\mathbf{X} - \boldsymbol{\mu}_X),\\ \mathrm{Var}[\mathbf{Y}|\mathbf{X}] =& \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}.\end{aligned}$$

Joint normality and independence:

- normally distributed and independent random variables are jointly normally distributed, however, a pair of jointly normally distributed variables need not be independent;
- while it is true that the marginals of a multivariate normal are normal too, it is not true in general that given two normal random variables their joint distribution is normal;
- in general, random variables may be uncorrelated but highly dependent, but if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent, this implies that any two or more of its components that are pairwise independent are independent;
- it is not true however that two random variables that are marginally normally distributed and uncorrelated are independent: it is possible for two random variables to be distributed jointly in such a way that each one alone is marginally normally distributed, and they are uncorrelated, but they are not independent.

Example: consider X a standard normal random variable and define

$$Y = \begin{cases} X & \text{if } |X| > c \\ -X & \text{if } |X| < c \end{cases}$$

where c is a positive number to be specified. If c is very small, then$\mathrm{Corr}(X, Y) \simeq 1$; if c is very large, then $\mathrm{Corr}(X, Y) \simeq -1$. Since the correlation is a continuous function of c, there is some particular value of c that makes the correlation 0. That value is approximately 1.54. In that case, X and Y are uncorrelated, but they are clearly not independent, since X completely determines Y. Moreover, Y is normally distributed. Indeed, its distribution is the same as that of X. We use cdfs:

$$\begin{aligned} P(Y \leq x) &= P((|X| < c \cap -X < x) \cup (|X| > c \cap X < x)) = \\ &= P((|X| < c \cap X > -x)) + P((|X| > c \cap X < x)) = \\ &= P((|X| < c \cap X < x)) + P((|X| > c \cap X < x)) \end{aligned}$$

where the last row depends on the fact that for a symmetric distribution $P(X < x) = P(X > -x)$. Thus, since the events $\{|X| < c\}$ and $\{|X| > c\}$ are a partition of the sample space which is $\mathbb{R}$, then

$$P(Y \leq x) = P(X \leq x),$$

hence Y is a standard normal random variable too. Finally, notice that the sum $X + Y$ for $c = 1.54$ has a substantial probability (about 0.88) of it being equal to 0, whereas the normal distribution, being a continuous distribution, has no discrete part, i.e., does not concentrate more than zero probability at any single point. Consequently X and Y are not jointly normally distributed, even though they are marginally normally distributed.

Reading

Casella and Berger, Definition 4.5.10

33 Bernoulli motivation for the Law of Large Numbers

This section starts off somewhat more abstract but concludes with the most important and widely-used theorem in probability, the Central Limit Theorem. Along the way we also state and prove two laws of large numbers.

To get started, as an example, consider a sequence of independent Bernoulli random variables $X_i \sim X$ with $p = 1/2$ and let $Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (2X_i - 1)$. Note that we have normalised the X_i so that $E[Y_n] = 0$ and $\mathrm{Var}(Y_n) = 1$. In particular, the mean and variance of Y_n does not depend on n. A gambler could think of Y_n as their (rescaled) earnings in case they win £1 each time a fair coin ends up head and lose £1 each time the coin leads to tail. Astonishingly, even though Y_n is constructed from a humble Bernoulli distribution, as n gets large, the distribution of Y_n approaches that of the normal distribution. Indeed,

using moment generating functions (and $M_{aX+b}(t) = e^{bt}M_X(at)$ for $a, b \in \mathbb{R}$), we get

$$\begin{aligned}
M_{Y_n}(t) &= \left(e^{-t/\sqrt{n}}M_X(2t/\sqrt{n})\right)^n \\
&= \left(e^{-t/\sqrt{n}}(1 - \frac{1}{2} + \frac{1}{2}e^{2t/\sqrt{n}})\right)^n \\
&= \left(\frac{1}{2}e^{-t/\sqrt{n}} + \frac{1}{2}e^{t/\sqrt{n}}\right)^n \\
&\to \left((\frac{1}{2} - \frac{1}{2}t/\sqrt{n} + \frac{1}{4}t^2/n) + (\frac{1}{2} + \frac{1}{2}t/\sqrt{n} + \frac{1}{4}t^2/n)\right)^n \quad \text{(Taylor)} \\
&= \left((1 + \frac{1}{2}t^2/n)\right)^n \\
&\to e^{t^2/2},
\end{aligned}$$

which we recognise as the moment generating function of a standard normal distribution. Since moment generating functions (usually, more on this below) uniquely determine the distribution, it follows that Y_n "converges" to a normally distributed random variable. We shall see below that there is nothing special here about the Bernoulli distribution as hardly any distribution (though there are some) can resist the attraction of the normal distribution. But before we get to that, we first have a closer look at the various kinds of convergence of random variables and how these notions are related.

34 Modes of convergence

In what follows we consider a sequence of random variables $X_1, X_2, \ldots$ and we consider four (and there are more!) types of convergence.

The first notion is that of almost sure convergence. Perhaps you find the terminology surprising since in mathematical statements we are used to certainty and almost sure sounds rather vague (in fact, there is even a notion of vague convergence), but almost sure in the setting here means that convergence happens on a set with probability 1.

Almost sure convergence: the sequence $\{X_n\}$ converges to X almost surely if

$$P\left(\omega \in \Omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right) = 1,$$

and we use the notation $X_n \overset{a.s.}{\to} X$.
It means that $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$ except perhaps for some $\omega \in N$ where $P(N) = 0$.
Note that in the Casella Berger book this is stated in the equivalent form

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1, \quad \forall\, \epsilon > 0.$$

Note that whenever we write $P(A)$ we should check that A is in our sigma-algebra. Indeed, with $A := \{\omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\}$ we have that $\omega \in A$ if and only if

$$\forall k \in \mathbb{N}\ \exists N \in \mathbb{N} \text{ s.t. } \forall n \geq N \quad |X_n(\omega) - X(\omega)| < \frac{1}{k}$$

and hence

$$A = \bigcap_{k\in\mathbb{N}} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \frac{1}{k}\right\}$$

is a measurable set (being the countable intersection of a countable union of a countable intersection of measurable sets!). Useful equivalent definitions are

$$P\left(|X_n - X| > \epsilon \quad \text{for infinitely many } n\right) = 0 \quad \text{for any } \epsilon > 0$$

and

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| > \epsilon\right) = 0 \quad \text{for any } \epsilon > 0.$$

To see that the latter two definitions are equivalent, first consider an increasing sequence of events B_n, meaning that $B_i \subset B_{i+1}$ for each i. Using countable additivity it follows that (with $B_0 = \emptyset$)

$$P\left(\bigcup_n B_n\right) = P\left(\bigcup_n (B_n \backslash B_{n-1})\right) = \sum_n P(B_n \backslash B_{n-1}) = \lim_{n\to\infty} P(B_n).$$

A diagram might help here to see why the above is true and the final equality is an example of a so-called telescoping series. This is called the continuity property of probability. Next, note that $\bigcup_{m=n}^{\infty}\{|X_m - X| > \epsilon\}$ is a decreasing sequence of sets and by taking complements equivalence now follows (try filling in the details).

The remaining three modes of convergence are somewhat more straightforward.

Convergence in probability: the sequence $\{X_n\}$ converges to X in probability if

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1, \quad \forall\, \epsilon > 0,$$

and we use the notation $X_n \xrightarrow{P} X$.
An obviously equivalent definition is

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0, \quad \forall\, \epsilon > 0.$$

Mean-square convergence: the sequence $\{X_n\}$ converges to X in mean-square if

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0,$$

and we use the notation $X_n \xrightarrow{m.s.} X$.

Convergence in distribution: the sequence $\{X_n\}$ converges to X in distribution if

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for any t at which F_X is continuous. We use the notation $X_n \xrightarrow{d} X$.

Relations among the modes of convergence:

1. if $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{P} X$;
2. if $X_n \xrightarrow{m.s.} X$ then $X_n \xrightarrow{P} X$;
3. if $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$.

Proof:

1. If X_n converges to X almost surely, this means that for any $\epsilon > 0$

   $$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| > \epsilon\right) = 0.$$

   Since $\{|X_n - X| > \epsilon\} \subset \bigcup_{m=n}^{\infty}\{|X_m - X| > \epsilon\}$ it follows that

   $$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0,$$

   so X_n converges to X in probability.

2. From Chebyshev's inequality we know that for any $\epsilon > 0$

   $$P\left(|X_n - X| > \epsilon\right) \leq \frac{E[(X_n - X)^2]}{\epsilon^2} \to 0$$

   as $n \to \infty$ and hence mean-square convergence indeed implies convergence in probability.

3. Suppose for simplicity that X_n and X are continuous random variables and assume that $X_n \xrightarrow{P} X$. From the bounds

   $$P\left(X \leq t - \epsilon\right) \leq P\left(X_n \leq t\right) + P\left(|X_n - X| \geq \epsilon\right)$$

   and

   $$P\left(X_n \leq t\right) \leq P\left(X \leq t + \epsilon\right) + P\left(|X_n - X| \geq \epsilon\right)$$

   it follows by letting $\epsilon > 0$ arbitrarily small that

   $$P\left(X_n \leq t\right) \to P\left(X \leq t\right) \quad \text{as } n \to \infty.$$

   This argument can be adapted to the case when X_n or X are not continuous random variables as long as t is a point of continuity of F_X.

Note that it follows that convergence in distribution is implied by any of the other modes of convergence. None of the other implications hold in general. For some of the examples and also for the proof of (a special case of) the Strong Law of Large Numbers the so-called Borel Cantelli Lemmas are incredibly useful.

35 Borel Cantelli Lemmas

The Borel Cantelli Lemmas are two fundamental lemmas in probability theory. Let A_n be a sequence of events and denote by $A := \cap_n \cup_{m=n}^{\infty} A_m$ the event that infinitely many of the A_n occur. The Borel Cantelli Lemmas give sufficient conditions on the A_n under which either $P(A) = 0$ or $P(A) = 1$.

Borel Cantelli 1: Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then $P(A) = 0$.

Proof: Note that since by definition $A \subset \cup_{m=n}^{\infty} A_m$ for each n, it follows that

$$P(A) \leq P(\bigcup_{m=n}^{\infty} A_m) \leq \sum_{m=n}^{\infty} P(A_m) \to 0 \quad \text{as } n \to \infty$$

since $\sum_{n=1}^{\infty} P(A_n) < \infty$.

Borel Cantelli 2 Suppose that $A_1, A_2, \ldots$ are independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$. Then $P(A) = 1$.

Proof: It suffices to show that $P(A^c) = 0$. Note that

$$\begin{aligned}
P(A^c) &= P\left(\bigcup_n \bigcap_{m=n}^{\infty} A_m^c\right) \\
&= \lim_{n\to\infty} P\left(\bigcap_{m=n}^{\infty} A_m^c\right) \quad (\text{as } \textstyle\bigcap_{m=n}^{\infty} A_m^c \text{ is increasing in } n) \\
&= \lim_{n\to\infty} \prod_{m=n}^{\infty} (1 - P(A_m)) \quad (\text{independence}) \\
&\leq \lim_{n\to\infty} \prod_{m=n}^{\infty} e^{-P(A_m)} \quad (\text{since } 1 - x \leq e^{-x}) \\
&= \lim_{n\to\infty} e^{-\sum_{m=n}^{\infty} P(A_m)} \\
&= 0
\end{aligned}$$

whenever $\sum_{n=1}^{\infty} P(A_n) = \infty$.

36 Examples of various modes of convergence

Example "in probability" does not imply "almost surely"

Let X_n be independent Bernoulli random variables with parameter $p = 1/n$. Then it obviously holds that $X_n \xrightarrow{P} 0$ since $P(|X_n - 0| > \epsilon) = P(X_n = 1) = 1/n \to 0$ as $n \to \infty$. You may find it surprising (at least upon first reading) that X_n **does not** converge to 0 almost surely. Indeed, considering $A_n := \{X_n = 1\}$ it holds that

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

as the harmonic series diverges[3]. Now, from the second Borel Cantelli Lemma it follows that $P(X_n = 1 \text{ for infinitely many n}) = 1$, so X_n does not converge to 0 almost surely.

Example "square mean" does not imply "almost surely"

Let X_n be defined as in the previous example. Then

$$E[(X_n - 0)^2] = \frac{1}{n} \to 0$$

as $n \to \infty$, so X_n converges in mean square to 0 but not almost surely.

Example "in probability" does not imply "square mean"

Convergence in probability only means that the probability that X_n and X differ by at most $\epsilon > 0$ goes to zero as $n \to \infty$, and, in particular, it does not lead to any restriction on the values of X_n when it is not close to X. Take for example $X_n = 0$ with $p = 1 - 1/n$ and $X_n = n$ with $p = 1/n$. Again, it holds that have that $X_n \xrightarrow{P} 0$. However, since

$$E[(X_n - 0)^2] = \frac{n^2}{n} = n$$

does not converge to zero, the random variables X_n do not converge to 0 in square mean.

Example "almost surely" does not imply "square mean"

If we tweak X_n and define the sequence now with $P(X_n = 0) = 1 - 1/n^2$ and $P(X_n = n) = 1/n^2$ we have that for any $\epsilon > 0$

$$P(|X_n - 0| > \epsilon) = \frac{1}{n^2}.$$

Since $\sum_{n=1}^{\infty} n^{-2} < \infty$, (in fact[4], it is $\pi^2/6$), it now follows from the first Borel Cantelli Lemma that

$$P\left(|X_n - 0| > \epsilon \quad \text{for infinitely many } n\right) = 0 \quad \text{for any } \epsilon > 0,$$

or equivalently, $X_n \xrightarrow{a.s.} 0$. On the other hand,

$$E[(X_n - 0)^2] = n^2/n^2 = 1$$

and so X_n does not converge to 0 in mean square.

Example "in distribution" does not imply anything

Let Z be a standard normal random variable and let $X_n = (-1)^n Z$. Than X_n converges in distribution to Z but does not converge in any of the other three modes.

Example "almost surely" implies "in probability"

[3]for example, this follows form the fact that the harmonic series $1 + 1/2 + (1/3 + 1/4) + (1/5 + 1/6 + 1/7 + 1/8) + \ldots$ has a lower bound $1 + 1/2 + (1/4 + 1/4) + (1/8 + 1/8 + 1/8 + 1/8) + \ldots = 1 + 1/2 + 1/2 + 1/2 + \ldots = \infty$

[4]for various proofs of this surprising result see `http://empslocal.ex.ac.uk/people/staff/rjchapma/etc/zeta2.pdf`

Consider X_n and $X \sim U[0,1]$ such that $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$ for any $\omega \in [0,1]$. Then if $\omega \in [0,1)$ we have $\omega^n \to 0$ and so $X_n(\omega) \to X(\omega) = \omega$. When $\omega = 1$ we have $X_n(1) = 2$ but $X(1) = 1$. However, the set in which we have problems is $A = \{\omega \text{ s.t. } \omega = 1\}$ and we have

$$P(A) = 1 - P(A^c) = 1 - P(\{\omega \text{ s.t. } \omega \in [0,1)\}) = 1 - [F_X(1) - F_X(0)] = 1 - [1-0] = 0.$$

We have also convergence in probability. We can write $X_n = X + X^n$, then

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P(|X^n| > \epsilon) = \\ &= P(X^n < -\epsilon \cup X^n > \epsilon) = \\ &= P(X < -\epsilon^{1/n} \cup X > \epsilon^{1/n}) = \\ &\to P(X < -1 \cup X > 1) = 0. \end{aligned}$$

Example "in probability" does not imply "almost surely"

Consider X_n and $X \sim U[0,1]$ such that

$$\begin{aligned} X_1(\omega) &= \omega + I_{[0,1]}(\omega), \\ X_2(\omega) &= \omega + I_{[0,1/2]}(\omega), \\ X_3(\omega) &= \omega + I_{[1/2,1]}(\omega), \\ X_4(\omega) &= \omega + I_{[0,1/3]}(\omega), \\ X_5(\omega) &= \omega + I_{[1/3,2/3]}(\omega), \\ X_6(\omega) &= \omega + I_{[2/3,1]}(\omega). \end{aligned}$$

Define also $X(\omega) = \omega$. Let's compute the probability limit

$$P(|X_n - X| > \epsilon) = P(X + I_{\delta_n} - X > \epsilon) \to 0,$$

since δ_n is an interval that becomes smaller and smaller as $n \to \infty$. Then $X_n \to X$ in probability. However, for any ω we have an n such that $X_n(\omega) = \omega$, $X_{n+1}(\omega) = \omega + 1$, and $X_{n+2}(\omega) = \omega$. Therefore the set of outcomes such that X_n does not converge to X is the whole sample space $[0,1]$ which implies that no almost sure convergence exists.

Example "in probability" implies "in distribution"

Convergence in probability implies convergence in distribution. Assume that $X_n \sim U[0,1]$ and are i.i.d. such that $X_n = \max_{1 \le i \le n} X_i$. We prove that X_n converges in probability to the random variable $X = 1$.

$$\begin{aligned} P(|X_n - 1| > \epsilon) &= P(X_n - 1 > \epsilon \cup X_n - 1 < -\epsilon) = \\ &= P(X_n > \epsilon + 1) + P(X_n < 1 - \epsilon) = \\ &= 0 + P(\cap_{i=1}^n X_i < 1 - \epsilon) = \\ &= \prod_{i=1}^n P(X_i < 1 - \epsilon) = \\ &= (1 - \epsilon)^n \to 0. \end{aligned}$$

Then consider $\epsilon = t/n$

$$P(X_n \leq 1 - t/n) = (1 - t/n)^n \to e^{-t},$$

therefore

$$P(X_n \geq 1 - t/n) = P(n(1 - X_n) \leq t) \to 1 - e^{-t},$$

which is a cdf of an Exponential r.v. thus, $n(1 - X_n) \sim \text{Exp}(1)$.

Example "in distribution" and continuity of cdf

Define $X_n \sim U[1/2 - 1/n, 1/2 + 1/n]$ then as $n \to \infty$, $X_n \to X = 1/2$ in distribution, where the limiting r.v. is a degenerate r.v. with all its mass in $1/2$. We have

$$F_{X_n}(t) = \begin{cases} 0 & t \leq 1/2 - 1/n \text{ or } t \geq 1/2 + 1/n \\ \frac{t-(1/2-1/n)}{2/n} & t \in [1/2 - 1/n, 1/2 + 1/n]. \end{cases}$$

As $n \to \infty$ the cdf converges to $F_{X_n}(1/2) = 1/2$, however the limiting r.v. has cdf $F_X(1/2) = 1$ as all the mass of X is in $t = 1/2$. So in $t = 1/2$ the cdf of X_n does not converge to the cdf of X. However, $t = 1/2$ is a point where F_X is not continuous, thus we still have convergence in distribution.

37 Two Laws of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with moments $\text{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$, such that $\sigma^2 < \infty$, for all i. We define $S_n = \sum_{i=1}^n X_i$ and S_n/n is the sample mean (a random variable). Then we have two results.

Weak Law of Large Numbers:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1, \qquad \forall \epsilon > 0$$

that is $S_n/n \xrightarrow{P} \mu$.

Proof: for every $\epsilon > 0$, we use Chebychev's inequality

$$P(|S_n/n - \mu| > \epsilon) \leq \frac{\text{E}[(S_n/n - \mu)^2]}{\epsilon^2} = \frac{\text{Var}[S_n/n]}{\epsilon^2} = \frac{\text{Var}[S_n]}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which converges to 0 as n goes to ∞.

Whereas the weak law of large numbers numbers is straightforward to prove, perhaps not surprisingly the strong law of large numbers requires some more effort.

Strong Law of Large Numbers:

$$P\left(\lim_{n\to\infty}\left|\frac{S_n}{n} - \mu\right| = 0\right) = 1,$$

that is $S_n/n \stackrel{a.s.}{\to} \mu$.

Proof: Here we give the proof in the case that we have the additional assumption that $E[X_i^4] < \infty$. In that case, note that

$$E[(S_n/n - \mu)^4] = \frac{1}{n^4} E\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^4\right].$$

Note that this is a rather humongous sum. Justify (exercise) that it is equal to

$$\frac{1}{n^4}\left\{nE[(X-\mu)^4] + 3n(n-1)\left(E\left[(X-\mu)^2\right]\right)^2\right\}.$$

Note that this expression can be bounded by Cn^{-2} for some $C > 0$ which does not depend on n. Using Chebyshev's inequality with $g(x) = x^4$ we have that for $\epsilon > 0$

$$P(|S_n/n - \mu| \geq \epsilon) \leq \frac{E[(S_n/n-\mu)^4]}{\epsilon^4} \leq \frac{C}{\epsilon^4 n^2}.$$

Since $1/n^2$ is summable we deduce from Borel Cantelli 1 that $S_n/n \stackrel{a.s.}{\to} \mu$. (to see why, reconsider the example above of almost sure convergence but not convergence in mean square).

On the assumptions: for the proof of weak and strong law above we have used the assumption of finite second and fourth moment, respectively. This is in fact stronger than what is needed. A sufficient condition is the weaker assumption $\mathrm{E}[|X|] < \infty$; the proof is much more demanding though.

The Strong Law of Large Numbers implies the Weak Law of Large Numbers and also convergence in distribution $S_n/n \stackrel{d}{\to} \mu$ which can be interpreted as convergence to the degenerate distribution with all of the mass concentrated at the single value μ. We shall soon see that, just as in the case of the sum of Bernoulli random variables at the beginning, we can say a lot more about the limiting distribution of S_n by proper rescaling. To be more specific, since $S_n/n - \mu$ converges to zero and since $\mathrm{Var}(S_n/n - \mu) = 1/n$, a scaling with factor $\sqrt{n}$, i.e. $\sqrt{n}(S_n/n - \mu)$ seems promising. This is the subject of the next section.

38 Central Limit Theorem

In this section we state and prove the fundamental result in probability and statistics, namely that the normalised sample mean from an i.i.d. sample (with finite variance) converges to a standard normal distribution. We shall make use of moment generating functions and the following result from the theory of so-called Laplace transforms.

Convergence of mgfs (Theorem 2.3.12 in Casella Berger) If X_n is a sequence of random varaibles with a moment generating functions satisfying

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

for all t in a neighbourhood of 0 and if $M_X(t)$ is a moment generating function of a random variable X, then $X_n \xrightarrow{d} X$.

Assumptions: given an i.i.d. sequence of random variables $X_1, X_2, \ldots$ with finite variance $\sigma^2 > 0$, define $S_n = \sum_{i=1}^n X_i$.

Central Limit Theorem: if the mgf $M_X(t)$ of X_i exists in some neighborhood of 0, then, as $n \to +\infty$, and for $Z \sim N(0,1)$,

$$\sqrt{n}\frac{S_n/n - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z.$$

We can state the convergence in distribution as

$$P\left(\sqrt{n}\frac{S_n/n - \mu}{\sigma} \leq \alpha\right) \to \int_{-\infty}^{\alpha} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z, \quad \forall\, \alpha \in \mathbb{R}.$$

Notice that both μ and σ^2 exist and are finite since the mgf exists in a neighbourhood of 0.

Proof: Define $Y_i = (X_i - \mu)/\sigma$, then

$$\sqrt{n}\frac{S_n/n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$$

therefore the mgf of Y_i exists for t in some neighbourhood of 0 (and we shall take t sufficiently small from now on), given that Y_i are i.i.d.

$$\begin{aligned} M_{\sqrt{n}\sigma^{-1}(S_n/n-\mu)}(t) &= M_{n^{-1/2}\sum_{i=1}^n Y_i}(t) = \\ &= \left(\mathrm{E}\left[\exp\left(tY_i/\sqrt{n}\right)\right]\right)^n = \\ &= \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n. \end{aligned}$$

By expanding in Taylor series around $t = 0$, we have

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} \mathrm{E}[Y_i^k]\frac{(t/\sqrt{n})^k}{k!}.$$

Now notice that $\mathrm{E}[Y_i] = 0$ and $\mathrm{Var}[Y_i] = 1$ for any i, thus

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2} + o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right],$$

where the last term is the remainder term in the Taylor expansion such that

$$\lim_{n\to\infty} \frac{o[(t/\sqrt{n})^2]}{(t/\sqrt{n})^2} = 0.$$

Since t is fixed we also have

$$\lim_{n\to\infty} \frac{o[(t/\sqrt{n})^2]}{(1/\sqrt{n})^2} = \lim_{n\to\infty} n\, o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right] = 0,$$

thus

$$\begin{aligned}\lim_{n\to\infty} M_{\sqrt{n}\sigma^{-1}(S_n/n-\mu)}(t) &= \lim_{n\to\infty}\left[M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \\ &= \lim_{n\to\infty}\left\{1+\frac{1}{n}\left(\frac{t^2}{2}+n\, o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right]\right)\right\}^n = e^{t^2/2},\end{aligned}$$

which is the mgf of a standard normal random variable. Therefore, by uniqueness of the moment generating function, $\sqrt{n}(S_n/n-\mu)/\sigma$ converges in distribution to a standard normal random variable.

On the assumptions:

1. we can relax the assumption of finite variances, it is enough to have X_i that are small with respect to S_n; this can be assured by imposing two conditions by Lyapunov and Lindeberg of asymptotic negligibility;

2. Independence can also be relaxed by asking for asymptotic independence.

3. The assumption on the existence on moment generating functions can be dropped and a similar proof can be given in terms of the so-called characteristic function. This is defined similarly to the moment generating function by

   $$\Phi(t) := E[e^{\mathrm{i}tX}] \quad \text{for } t \in \mathbb{R}.$$

   Here $\mathrm{i} = \sqrt{-1}$ and $e^{ix} = \cos(x) + \mathrm{i}\sin(x)$ for $x \in \mathbb{R}$. The advantage of the characteristic function over the moment generating function is that the former always exists. This is due to the property that $|e^{ix}| = \cos^2(x) + \sin^2(x) = 1$ and hence $\Phi(t) \leq 1$. Characteristic functions also uniquely determine distributions and there is a convergence result equivalent to the one above for moment generating functions. Once you have calculated the moment generating functions, it is usually straightforward to find the characteristic function. For example, if X is standard normal, then

   $$\Phi_X(t) = E[e^{\mathrm{i}tX}] = e^{(\mathrm{i}t)^2/2} = e^{-t^2/2}.$$

Reading

Casella and Berger, Sections 5.5