36-217 Probability Theory and Random Processes Lecture Notes

Riccardo Fogliato

June 28, 2019

Lecture 1

Recommended Readings: WMS 2.1-2.4

Populations, Statistics and Random Processes

What is Statistics?
Quoting John Tukey,[1]

> "Statistics is a science, not a branch of mathematics, but uses mathematical models as an essential tool."

However, on the Merriam-Webster dictionary[2] we read that (Statstics is)

> "a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

Although one may argue that Statistics is, or is not, a branch of mathematics, both quotes seem to suggest that statisticians need mathematical tools. The second quote further suggests that such tools are used to collect and analyse numerical data. What does this mean?

Broadly speaking, Statistics is about *inference* and *estimation* of *features* or *parameters* for a 'population' using observed data from that population. What do we mean by 'population'? We can think of a population as an entity which encompasses all possible data that can be observed in an experiment. This can be a tangible collection of objects or a more abstract entity.

Examples of populations are:

[1] https://www.stat.berkeley.edu/~brill/Papers/boas.pdf
[2] https://www.merriam-webster.com/dictionary/statistics

- the deers in Frick Park;
- the runners of all the marathons in the USA;
- the subscribers of HBO;
- the number of burritos sold daily at Chipotle.

Most often, we perform an experiment because we are interested in learning about a particular feature or parameter of a population. Examples of parameters for the above populations are respectively:

- the total number;
- the average time of the finishers;
- the proportion of subscribers that watched the final episode of GoT;
- the variability of the sales.

In order to learn such features we usually proceed as follows:

1. collect data
2. specify a statistical model for the unknown true population (e.g. specify a class of functions that approximate well the unobserved "true" distribution of the preferences of the subscribers of HBO)
3. collect data $X_1, \ldots, X_n$, possibly by performing a particular experiment (e.g. go to Frick park and count the deers... make sure that you do not double count them)
4. compute a *statistic*, i.e. a function of the data (and of the data exclusively!), to estimate the parameter of interest (e.g. compute the average time of arrival of the runners)
5. further statistical analyses (in the next future).

How do we perform such such a statistical analysis?
Based on Tukey's quote we use "mathematical models as an essential tools'. Probability Theory provides the foundations of these tools. From a purely mathematical point of view, Probability Theory is an application of Measure Theory to real-world problems. At a more introductory and conceptual level, Probability Theory is a set of mathematical tools that allows us to carry out and analyze the process described above in a scientific way.

A statistic is always a function of the data (it cannot be a function of some parameter!). As an example, some frequently used statistics or estimators are

- the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

In particular, they estimate the mean μ of the distribution of interest (which is a measure of central tendency) and its *variance* σ^2 (which is a measure of variability/dispersion).

With respect to the examples above, we can use the sample mean to estimate the mean weight μ of all the burritos sold in Chipotles by visiting the store on Craig street every day for a month and asking the cashier. We have now measured the sales $X_1, \ldots, X_n$, and we can compute $\bar{X}$. We expect that, as n gets larger, $\bar{X} \approx \mu$. If we are also interested in estimating the variance of the sales of burritos σ^2 or its standard deviation $\sigma = \sqrt{\sigma^2}$, we might compute the sample variance S^2 or the sample standard deviation $S = \sqrt{S^2}$ (note that, as opposed to the sample variance, the sample standard deviation has the same unit of measure of the data $X_1, \ldots, X_n$). Again, as n gets larger, we expect that $S^2 \approx \sigma^2$.

For the marathon's example, we could scrape some data from the marathons website. Unfortunately we are not expert in the topic, and we choose the Boston marathon. The sample mean is 3 hours and 54 minutes, so we conclude the population mean will be the same. However, from a quick search on Google we find out that some website claims that the population mean should be 4 hours and 22 minutes. Then, who is correct? Likely, nor you nor the website. However, who is more correct? A friend suggests that in order to run the Boston marathon one needs to qualify for it; from this insight we desume that these runners probably are faster than average. We have incurred in a problem of *sample bias*: our sample does not reflect the entire population; statistically speaking, we say that the *distribution* of our sample does not correspond to the distribution of the population of interest.

How do we solve this problem? We could imagine that the runners population is a gamma distribution with some unknown mean, and all the individuals in our sample have times lower than the mean. Therefore we have made a *modeling assumption*, that is a statement that you believe it is true, in order to estimate the parameter of interest, the mean.

This class will provide you with the tools to approach these problems in a statistical manner. Statistics is a result of the following process:

TRUTH → MODELS → STATISTICS

There are situations in which the quantity that we want to study evolves with time (or with space, or with respect to some other dimension). For example, suppose that you are interested in the number of people walking into a store on a given day. We can model that quantity as a random quantity X_t varying with respect to time t. Most often there exists some kind of 'dependence' between the number of people in the store at time t and the number of people in the store at time t' (especially if $|t-t'|$ is small). There is a branch of Probability Theory that is devoted to study and model this type of problems. We usually refer to the collection $\{X_t : t \in \mathcal{T}\}$ as a 'random process' or as a 'stochastic process'. The set $\mathcal{T}$ can represent a time interval, a spatial region, or some more elaborate domain.

Another example of a random process is observing the location of robberies occurring in a given city. (Question: what is T)?

Another typical example of a random process is the evolution of a stock price in Wall Street as a (random) function of time. (Question: what is T)?

We will devote part of this course to study (at an introductory level) some of the theory related to random processes.

We have mentioned probability several times, but what do we mean by that? There are two major interpretations:

- objective probability: the long-run frequency of the occurrence of a given event (e.g. seeing heads in a coin flip)
- subjective probability: a subjective *belief* in the rate of occurrence of a given event.

Methods based on objective probability are usually called *frequentist* or *classical* whereas those based on subjective probability are usually called *Bayesian* as they are associated to the Bayesian paradigm, although Larry argues that frequentist and Bayesian should be characterized in terms of their goals, and not in terms of their methods.[3] In this class, we focus on frequentist methods (but we will discuss later in the course the basic idea at the basis of Bayesian Statistics).

In Probability Theory, the notion of 'event' is usually described in terms of a set (of outcomes). Therefore, before beginning our discussion of Probability Theory, it is worthwhile to review some basic facts about set theory.

[3] https://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

Set Theory

Throughout the course we will adopt the convention that Ω denotes the universal set (the superset of all sets) and $\emptyset$ denotes the empty set (i.e. a set that does not contain any element). Recall that a set A is a subset of another set B (or A is contained in B) if for any element x of A we have $x \in A \implies x \in B$. We denote the relation A is a subset of B by $A \subset B$ (similarly, $A \supset B$ means A is a superset of B or A contains B). Clearly, if $A \subset B$ and $B \subset A$, then $A = B$. Let $\wedge$ and $\vee$ stand for 'and' and 'or' respectively. Given three subsets A, B, C of Ω, recall the following set properties:

- commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$;
- associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$;
- distributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
- De Morgan's laws: $(A \cup B)^c = A^c \cap B^c$ and $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$.

and the following basic set operations:

- union of sets: $A \cup B = \{x \in \Omega : x \in A \vee x \in B\}$
- intersection of sets: $A \cap B = \{x \in \Omega : x \in A \wedge x \in B\}$
- complement of a set: $A^c = \{x \in \Omega : x \notin A\}$
- set difference $A \setminus B = \{x \in \Omega : x \in A \wedge x \notin B\}$
- symmetric set difference $A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$.

These can be extended to the union and the intersection of any $n > 0$ sets:

- $(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c$
- $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$.

The same strategy for the proof can be used for the second statement.

We typically use *Venn's diagrams* to represent logical relations between sets.

Notice that, for any set $A \subset \Omega$, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. Two sets $A, B \subset \Omega$ for which $A \cap B = \emptyset$ are said to be disjoint or mutually exclusive.

A **partition** (disjoint union) of Ω is a collection of subsets $A_1, \cdots, A_n$ that satisfy

1. $A_i \cap A_j = \emptyset$, $\forall i \neq j$
2. $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$

We used before the word 'experiment' to describe the process of collecting data or observations. An experiment is, broadly speaking, any process by which an observation is made. This can be done actively, if you have control on the apparatus that collects the data, or passively, if you only get to see the data, but you have no control on the apparatus that collects them. An experiment generates observations or outcomes. Consider the following example: you toss a fair coin twice (your experiment). The possible outcomes (simple events) of the experiment are HH, HT, TH, TT (H: heads, T: tails). The collection of all possible outcomes of an experiment forms the 'sample space' of that experiment. In this case, the sample space is the set $\Omega = \{HH, HT, TH, TT\}$. Any subset of the sample space of an experiment is called an event. For instance, the event 'you observe at least one tails in your experiment' is the set $A = \{HT, TH, TT\} \subset \Omega$.

Note: These are *discrete* events, i.e. they are a collection of sample points from a *discrete sample space*. A discrete sample space Ω is one that contains either a finite or countable number of distinct sample points. An event in a discrete sample space is simply a collection of sample points – any subset of Ω.

In this part of the course, just assume we deal with discrete events from a discrete sample spaces. Later, we will deal with *continuous* sample spaces.

Exercises in class

1. A fair coin is tossed twice. You have decided to play a game whose outcome is

   - If the first flip is **H**, you win \$1,000,
   - If the first flip is **T**, you lose \$1,000.

   (a) Writing a particular outcome as the concatenation of single outcomes such as **H** or **T**. What is the sample space Ω?

(b) Write as A, a subset of Ω, the set corresponding to the *event* that the first flip is **H**. What are elements in A? What is the probability of A?

(c) Denote as B the same as above, of the event that the second toss is **T**. What are the elements in B? What is the probability of B?

(d) What is the probability of the first is head *and* second being tail?

(e) Draw A, B, and Ω in a Venn diagram.

(f) What is your *expected return*?

Lecture 2

Recommended readings: WMS 2.1→ 2.6

Unconditional Probability

Consider again the example from the previous lecture. We toss a fair coin twice. The sample space for the experiment is $\Omega = \{HH, HT, TH, TT\}$.

If we were to repeat the experiment over and over again, what is the frequency of the occurrence of the sequence HT? What about the other sequences? What about the frequency of $\{HT\} \cup \{TT\}$?

Modern Probability Theory is built on a set of axioms[4] formulated by the Russian mathematician Andrey Kolmogorov in his masterpiece *Foundations of the Theory of Probability.*
A *probability measure* P on Ω is a set function[5] satisfying the following axioms:

1. for any $A \subset \Omega$, $P(A) \in [0, 1]$;
2. (norming) $P(\Omega) = 1$;
3. (countable additivity) for any countable collection of disjoint events $\{A_i\}_{i=1}^{\infty} \subset \Omega$, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

From the second and third axiom, it follows that $P(\emptyset) = 0$.

Now, what do these axioms imply about $P(A \cup B)$? We can show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. First notice that $A = (A \cap B) \cup (A \cap B^c)$, and these events are disjoint. By the third axiom of probability we have

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ P(B) &= P(A \cap B) + P(A^c \cap B) \end{aligned} \tag{1}$$

[4] An axiom is a statement believed to be true.
[5] A set function is a function whose domain is a collection of sets.

Moreover, $A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$ and again these sets are all disjoint. Putting all together we get

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned} \tag{2}$$

□

In case of $A \cap B = \emptyset$ we clearly obtain

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(\emptyset) = P(A) + P(B). \end{aligned} \tag{3}$$

This can be extended to more than 2 events. For instance, given $A, B, C \subset \Omega$, we have

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) \\ - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned} \tag{4}$$

and, in general, for $A_1, \ldots, A_n \subset \Omega$ we have the so called inclusion-exclusion formula (note the alternating signs for the summands):

$$\begin{aligned} P(\cup_{i=1}^n A_i) &= \sum_{i=1}^n P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) \\ &+ \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n). \end{aligned} \tag{5}$$

Notice that in general we have the following *union bound* for any $A_1, \ldots, A_n \subset \Omega$:

$$P(\cup_{i=1}^n A_i) \le \sum_{i=1}^n P(A_i). \tag{6}$$

For a given experiment with m possible outcomes, how can we compute the probability of an event of interest? We can always do the following:

1. define the experiment and describe its simple events E_i, $i \in \{1, \ldots, m\}$;
2. define reasonable probabilities on the simple events, $P(E_i)$, $i \in \{1, \ldots, m\}$;
3. define the event of interest A in terms of the simple events E_i;
4. compute $P(A) = \sum_{i: E_i \in A} P(E_i)$.

Here is an example (and we will describe it in terms of the above scheme). There are 5 candidates for two identical job positions: 3 females and 2 males. What is the probability that a completely random selection process will appear discriminatory? (i.e. exactly 2 males or exactly 2 females are chosen for these job positions) We can approach this problem in the following way.

1. we introduce 5 labels for each of the 5 candidates: M_1, M_2, F_1, F_2, F_3. The sample space is then

$$\begin{aligned}\Omega =\{&M_1M_2, M_1F_1, M_1F_2, M_1F_3, M_2M_1, M_2F_1, M_2F_2, M_2F_3, F_1M_1, F_1M_2,\\ &F_1F_2, F_1F_3, F_2M_1, F_2M_2, F_2F_1, F_2F_3, F_3M_1, F_3M_2, F_3F_1, F_3F_2\}\end{aligned}$$

2. because the selection process is completely random, each of the simple events of the sample space is equally likely. Therefore the probability of any simple event is just $1/|\Omega|$

3. the event of interest is $A = \{M_1M_2, M_2M_1, F_1F_2, F_1F_3, F_2F_1, F_2F_3, F_3F_1, F_3F_2\}$

4. $P(A) = P(M_1M_2) + P(M_2M_1) + \cdots + P(F_3F_2) = |A|/|\Omega| = 8/20 = 2/5 = 0.4$.

If the simple events are equally likely to occur, then the probability of a composite event A is just $P(A) = |A|/|\Omega|$.

Questions to ask when you define the sample space and the probabilities on the simple events:

- is the sampling done with or without replacement?
- does the order of the labels matter?
- can we efficiently compute the size of the sample space, $|\Omega|$?

This leads to our next topic, which will equip us with *tools* to conveniently calculate probability.

Tools for counting sample points

Basic techniques from combinatorial analysis come handy for this type of questions when the simple events forming the sample space are equally likely to occur. Let's take a closer look to some relevant cases.

Suppose you have a group of m elements $a_1, \ldots, a_m$ and another group of n elements $b_1, \ldots b_n$. You can form mn pairs containing one element from each group. That is, if $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, then we are interested in the elements of the cartesian product $A \times B$. Of course you can easily extend this reasoning to more than just two groups. This is a useful fact when we are sampling with replacement and the order matters.

Consider the following example. You toss a six-sided die twice. You have $m = 6$ simple outcomes on the first roll and $n = 6$ possible outcomes in the second roll. The sample space is

$$\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,5), (6,6)\}.$$

Therefore the size of Ω is $|\Omega| = mn = 6^2 = 36$.
Is the experiment performed with replacement? Yes, if the first roll is a 2, nothing precludes the fact that the second roll is a 2 again. Does the order matter? Yes, the pair (2,5) corresponding to a 2 on the first roll and a 5 on the second roll is not equal to the pair (5,2) corresponding to a 5 on the first roll and a 2 on the second roll.

Sampling without replacement when order matters
An ordered arrangement of r distinct elements is called a *permutation*. The number of ways of ordering n distinct elements taken r at a time is denoted P_r^n where

$$P_r^n = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}. \tag{7}$$

Consider the following example. There are 30 members in a student organization and they need to choose a president, a vice-president, and a treasurer. In how many ways can this be done? The size of the n distinct elements (persons) is $n = 30$, the number of persons to be appointed is $r = 3$. The sampling is done without replacement (a president is chosen out of 30 people, then among the 29 people left a vice-president is chosen, and finally among the 28 people left a treasurer is chosen). Does the order matter? Yes, the president is elected first, then the vice-president, and finally the treasurer (think about 'ranking' the 30 candidates). The number of ways in which the three positions can be assigned is therefore $P_3^{30} = 30!/(30-3)! = 30 * 29 * 28 = 24360$.

Sampling without replacement when order does not matter
The number of *combinations* of n elements taken r at a time corresponds to

the number of subsets of size r that can be formed from the n objects. This is denoted C_r^n where

$$C_r^n = \binom{n}{r} = \frac{P_r^n}{r!} = \frac{n!}{(n-r)!r!}. \tag{8}$$

How did we get such a formula? To get some intuition, first think about all the ordered sets that contain the same r objects: this number is given by P_r^n. Fixed this r elements, let the objects be $x_1, \ldots, x_r$. Then we will have $r!$ different possible permutations of these r objects. Why? It is clear that x_1 will appear in any of r positions in the set, hence we have r choices. For any choice, x_2 will appear in any of the $r-1$ positions left. And so on .... This is exactly equal to the permutation of r objects. Therefore we divide P_r^n by $r!$ to obtain the number of combinations.

Here is an example. How many different subsets of 3 people can become officers of the organization formed by 30 people, if chosen randomly? Order doesn't matter here (we are not interested in the exact appointments for a given candidate). The answer is therefore $C_3^{30} = 30!/(27! * 3!) = 30 * 29 * 28/6 = 4060$.

The *binomial coefficient* $\binom{n}{r}$ can be extended to the multinomial case in a straightforward way. Suppose that we want to partition n distinct elements in k distinct groups of size $n_1, \ldots, n_k$ in such a way that each of the n elements appears in exactly one of the groups. This can be done in

$$\binom{n}{n_1 \ldots n_k} = \frac{n!}{n_1! n_2! \ldots n_k!} \tag{9}$$

possible ways.

The connection to the combinations seen above is more clear through the following decomposition:

$$\binom{n}{n_1 \ldots n_k} = C_{n_1}^n C_{n_2}^{n-n_1} C_{n_3}^{n-n_1-n_2} \ldots C_{n_k}^{n-\sum_{i=1}^{k-1} n_i}.$$

Exercises in class:

1. The final exam of 36-217 has 5 multiple choice questions. Questions 1 to 3 and question 5 each have 3 possible choices. Questions 4 has 4

possible choices. Suppose that you haven't studed at all for the exam and your answers to the final exam are given completely at random. What is the probability that you will get a full score at the exam?

2. You have a string of 10 letters. How many substrings of length 7 can you form based on the original string with replacement? And what about without replacement?

3. Google is looking to hire 3 software engineers and asked CMU for potential candidates. CMU advised that Google hires 3 students from the summer 36-217 class. Because the positions must be filled as quickly as possible, Google decided to skip the interview process and hire 3 of you completely at random. What is the probability that *you* will become a data scientist at Google?

4. In a game of 5 card poker (played out of a standard 52-card deck ? 13 denominations, 4 suits):

    - What is the probability of getting "4 of a kind" (4 cards of one denomination, 1 card of a different denomination)?
    - What is the probability of getting a "full house" (3 cards of one denomination, 2 cards of another denomination)?

Lecture 3

Recommended readings: WMS, sections 2.7 → 2.13

Conditional Probability

The probability of an event A may vary as a function of the occurrence of other events. Then, it becomes interesting to compute *conditional* probabilities, e.g. the probability that the event A occurs *given* the knowledge that another event B occurred.

As an example of unconditional probability, think about the event $A =$'my final grade for the 36-217 summer class is higher than 90/100' (this event is not conditional on the occurrence of another event) and its probability $P(A)$. As an example of conditional probability, consider now the event $B =$ 'my midterm grade for the 36-217 class was higher than 80/100': how are $P(A)$ and $P(A$ given that B occurred) related? Intuitively, there is quite a lot of uncertainty: grade of the final exam, homeworks, etc.... but we might expect that $P(A) < P(A$ given that B occurred)!

The conditional probability of an event A given another event B is usually denoted $P(A|B)$.

By definition, the conditional probability of the event A given the event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{10}$$

Observe that the quantity $P(A|B)$ is well-defined only if $P(B) > 0$.

Consider the following table:

	B	B^c
A	0.3	0.1
A^c	0.4	0.2

The unconditional probabilities of the events A and B are respectively $P(A) = 0.3 + 0.1 = 0.4$ and $P(B) = 0.3 + 0.4 = 0.7$. The conditional probability of the event A given the event B is $P(A|B) = P(A \cap B)/P(B) = 0.3/0.7 = 3/7$ while the probability of the event B given the event A is $P(B|A) = P(B \cap A)/P(A) = 0.3/0.4 = 3/4$. The conditional probability of the event A given that B^c occurs is $P(A|B^c) = P(A \cap B^c)/P(B^c) = 0.1/0.3 = 1/3$.
Does the probability of A vary as a function of the occurrence of the event B?

The conditional probability $P(\cdot|B)$ with respect to an event B with $P(B) > 0$ is a proper probability measure, therefore the three axioms of probability that we discussed in Lecture hold for $P(\cdot|B)$ as well. When it comes to standard computations, $P(\cdot|B)$ has the same properties as $P(\cdot)$: for instance, for disjoint events $A_1, A_2 \subset \Omega$ we have $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$.

Exercise in class
There is 20% chance that you go to Craig Street to have lunch at Sushi Fuku, a 30% chance that you get a coffee at Starbucks, and a 10% chance that you both have lunch at Sushi Fuku and get a coffee at Starbucks. What's the probability that you get a coffee at Starbucks if you have been to Sushi Fuku? What about the probability that you get lunch at Sushi Fuku given that you have been to Starbucks?

Independence

The event $A \subset \Omega$ is said to be independent of the event $B \subset \Omega$ if $P(A \cap B) = P(A)P(B)$. This means that the occurrence of the event B does not alter the chance that the event A happens. In fact, assuming that $P(B) > 0$, we easily see that this is equivalent to $P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A)$.

Furthermore, assuming that also $P(A) > 0$, we have that $P(A|B) = P(A)$ is equivalent to $P(B|A) = P(B)$ (independence is a symmetric relation!).

Exercise in class
Let $A, B \subset \Omega$ and $P(B) > 0$.

- What is $P(A|\Omega)$?
- What is $P(A|A)$?
- Let $B \subset A$. What is $P(A|B)$?
- Let $A \cap B = \emptyset$. What is $P(A|B)$?

Notice that from the definition of conditional probability, we have the following:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A). \tag{11}$$

This can be generalized to more than two events. For instance, for three events $A, B, C \subset \Omega$, we have

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) \tag{12}$$

and for n events $A_1, \ldots, A_n$

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}). \tag{13}$$

Warning: Independence and Disjoint are not the same.
Two events being disjoint simply means that they do not share any jointly occurring elements. For instance, in a single coin flip example, $A = \{\mathbf{H}\}$ and $B = \{\mathbf{T}\}$ are disjoint events, but A gives *perfect* knowledge of B, as we know that $P(A|B) = 0$ and $P(A|B^C) = 1$, so that $P(A|B) \neq P(A) = 1/2$.
Exercise in class
Consider the following events and the corresponding table of probabilities:

	B	B^c
A	0	0.2
A^c	0.4	0.4

Are the events A and B disjoint?
Are the events A and B independent?

Exercise in class
Consider the following events and the corresponding table of probabilities:

	B	B^c
A	1/4	1/12
A^c	1/2	1/6

Are the events A and B disjoint?
Are the events A and B independent?

Law of Total Probability and Bayes' Rule

Assume that $\{B_i\}_{i=1}^{\infty}$ is a partition of Ω, i.e. for any $i \neq j$ we have $B_i \cap B_j = \emptyset$ and $\cup_{i=1}^{\infty} B_i = \Omega$. Assume also that $P(B_i) > 0 \, \forall i$. Then, for any $A \subset \Omega$, we have the so-called *law of total probability*

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i). \tag{14}$$

Indeed, $A = \cup_{i=1}^{\infty}(A \cap B_i)$ when $\cup_{i=1}^{n} B_i = \Omega$. But this time the sets $(A \cap B_i)$ are also disjoint since $\{B_i\}_{i=1}^{\infty}$ is a partition of Ω. Furthermore, by equation (11) we have $P(A \cap B_i) = P(A|B_i)P(B_i)$. Thus, $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) =$

$\sum_{i=1}^{\infty} P(A|B_i)P(B_i)$.

We can now use the law of total probability to derive the so-called Bayes' rule. We have

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)}. \tag{15}$$

For two events A, B this reduces to

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \tag{16}$$

Exercise in class
You are diagnosed with a disease that has two types, A and B. In the population, the probability of having type A is 10% and the probability of having type B is 90%. You undergo a test that is 80% accurate, i.e. if you have type A disease, the test will diagnose type A with probability 80% and type B with probability 20% (and vice versa). The test indicates that you have type A. What is the probability that you really have the type A disease?

Let A ='you have type A', B='you have type B', T_A ='the test diagnoses type A', and T_B ='the test diagnoses type B'. We know that $P(A) = 0.1$ and $P(B) = 0.9$. The test is 80% accurate, meaning that

$$P(T_A|A) = P(T_B|B) = 0.8$$
$$P(T_B|A) = P(T_A|B) = 0.2$$

We want to compute $P(A|T_A)$. We have

$$P(A|T_A) = \frac{P(A \cap T_A)}{P(T_A)} = \frac{P(T_A|A)P(A)}{P(T_A|A)P(A) + P(T_A|B)P(B)}$$
$$= \frac{0.8 * 0.1}{0.8 * 0.1 + 0.2 * 0.9} = \frac{8}{26} = \frac{4}{13}.$$

Conditional Independence

We are now equipped to talk about another 'parallel' concept to independence. Let C be an event with $P(C) > 0$. We say that the events A and B are conditional independent of C if

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

This implies that

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

To see this notice that conditional independence of A,B given C implies that

$$\begin{aligned} P(A|C)P(B|C) &= P(A \cap B|C) \\ &= \frac{P(A \cap B \cap C)}{P(C)} \\ &= \frac{P(C)P(B|C)P(A|B \cap C)}{P(C))} \\ &= P(B|C)P(A|B \cap C). \end{aligned}$$

Conditional independence does not imply nor is it implied by independence.

Independence does not imply conditional independence Toss two coins. Let H_1 be the event that the first toss is H and H_2 the event that the second toss is H. Let D the even that the two tosses are different. H_1 and H_2 are clearly independent but $P(H_1|D) = P(H_2|D) = 1/2$ and $P(H_1 \cap H_2|D) = 0$.

Conditional independence does not imply independence We have two coins, a blue one and a red one. For the blue coin the probability of H is 0.99 and for the red coin it is 0.01. A coin is chosen at random and then tossed twice. Let B the event that the blue coin is selected, R the event that the red coin is selected. Let H_1 and H_2 the events that the first and second toss is H. By the law of total probability

$$P(H_1) = P(H_1|B)P(B) + P(H_1|R)P(R) = 1/2.$$

Similarly $P(H_2) = 1/2$. But, using again the law of total probability,

$$P(H_1 \cap H_2) = P(H_1 \cap H_2|B)P(B) + P(H_1 \cap H_2|R)P(R) = 1/2 \cdot 0.99^2 + 1/2 \cdot 0.01^2 \approx 1/2$$

which is different than $P(H_1)P(H_2) = 1/4$.

Pairwise independence does not imply independence.

Consider $A_1, \ldots, A_n \subset \Omega$. If for every $1 \leq i < j \leq n$ we have $P(A_i \cap A_j) = P(A_i)P(A_j)$, the events are said to be pairwise independent. Mutual independence, instead, is defined as $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$. Notice that mutual independence implies pairwise independence, but the converse is not true.

Lecture 4

Recommended readings: WMS, sections 3.1, 3.2, 4.1 $\rightarrow$ 4.3

Random Variables and Probability Distributions

Let's start with an example. Suppose that you flip a fair coin three times. The sample space for this experiment is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Since the coin is fair, we have that

$$P(\{HHH\}) = P(\{HHT\}) = \cdots = P(\{TTT\}) = \frac{1}{|\Omega|} = \frac{1}{8}.$$

Suppose that we are interested in a particular quantity associated to the experiment, say the number of tails that we observe. This is of course a random quantity. In particular, we can conveniently define a function $X : \Omega \rightarrow \mathbb{R}$ that counts the number of tails. Precisely,

$$\begin{aligned} &X(HHH) = 0 \\ &X(HHT) = 1 \\ &X(HTH) = 1 \\ &X(HTT) = 2 \\ &\vdots \\ &X(TTT) = 3. \end{aligned}$$

We say that P induces through X a *probability distribution* on $\mathbb{R}$. In particular, we can easily see that the probability distribution of the random variable X is

$$\begin{aligned} &P(X = 0) = P(\{HHH\}) = 1/8 \\ &P(X = 1) = P(\{HHT, HTH, THH\}) = 3/8 \\ &P(X = 2) = P(\{HTT, THT, TTH\}) = 3/8 \\ &P(X = 3) = P(\{TTT\}) = 1/8 \\ &P(X = \text{any other number}) = 0. \end{aligned}$$

Remark: although the probability measures are denoted by P on both sides, they do refer to probability measure on different spaces. The measure

on the LHS is on $\mathbb{R}$, while the one on the RHS is on Ω. This is more clear if we rewrite it, with some abuse of notation, as $P_{\mathbb{R}}(X = 0) = P_{\Omega}(X^{-1}(X = 0)) = P_{\Omega}(\{HHH\}) = 1/8$. Ω is typically referred to as the *underlying probability space*. Therefore $P_{\mathbb{R}}$ is the probability distribution induced by P_{Ω} through X.

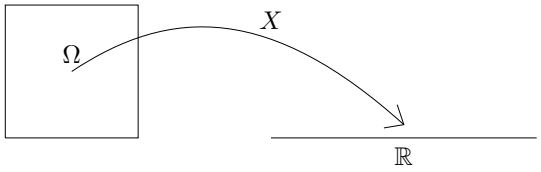
Why is X said to be random if it is just a function? First, the outcome of the experiment is random, so the value that the function X takes is also random. Second, we could modify the underlying probability space Ω without modifying X; therefore "random" aims at highlighting the fact that we are not truly interested in Ω, but in the distribution of X.

Formal Definition of a random variable

Suppose we have a sample space Ω. A **random variable** X is a function from Ω into the real line. In other words, $X : \Omega \to \mathbb{R}$ or

$$\omega \in \Omega \mapsto X(\omega) \in \mathbb{R}$$

Random variables will be denoted with capital letters, such as X. This is a consistent, standard notation for a random variable.



Why we care about random variables? After all, based on the example, it seems that we still have to clearly define and specify the sample space in order to determine $P(X = x)$ for $x \in \mathbb{R}$. Actually, it turns out when we model a random quantity of interest (such as the number of tails in the coin tossing example) most of the times one assumes (or knows) a distribution to use. It's just easier and more natural. Of course, this assumption must to be reasonable and needs to be rigorously checked by the modeller. By modelling the randomness of a phenomenon as a random variable whose distribution is known, we can bypass the trouble of defining/specifying a sample space. In the example above, if we simply assume that X has a Binomial distribution, then the sample space is automatically the discrete space of 3-length binary outcomes (HHH, HHT, etc.), and the probability of events of interest (e.g.

X=1, which corresponds to one head out of three throws, which is the subset of the sample space $\{HTT, THT, TTH\}$) is precisely calculable.

To summarise, one should specify Ω and a probability measure on this space, map Ω to $\mathbb{R}$ through X, and analyse the induced probability measure on $\mathbb{R}$. However, one can skip the first step, that is leave Ω undefined, and just specify the induced probability measure on $\mathbb{R}$, which is the object of interest. This is enough to guarantee the existence of some Ω for which this probability measure exists.

Exercise in class

Let X and Y be random variables with the following distributions:

$$P(X = x) = \begin{cases} 0.4 \text{ if } x = 0 \\ 0.6 \text{ if } x = 1 \\ 0 \text{ if } x \notin \{0, 1\} \end{cases}$$

and

$$P(Y = y) = \begin{cases} 0.7 \text{ if } y = -1 \\ 0.3 \text{ if } y = 1 \\ 0 \text{ if } y \notin \{0, 1\} \end{cases}$$

Suppose that for any $x, y \in \mathbb{R}$ the events $\{X = x\}$ and $\{Y = y\}$ are independent. What is the probability distribution of the random variable $Z = X + Y$? We have

$$Z = \begin{cases} -1 \text{ if } \{X = 0\} \cap \{Y = -1\} \text{ occurs} \\ 0 \text{ if } \{X = 1\} \cap \{Y = -1\} \text{ occurs} \\ 1 \text{ if } \{X = 0\} \cap \{Y = 1\} \text{ occurs} \\ 2 \text{ if } \{X = 1\} \cap \{Y = 1\} \text{ occurs.} \end{cases}$$

Thus,

$$P(Z = z) = \begin{cases} 0.4 * 0.7 \text{ if } z = -1 \\ 0.6 * 0.7 \text{ if } z = 0 \\ 0.4 * 0.3 \text{ if } z = 1 \\ 0.6 * 0.3 \text{ if } z = 2 \\ 0 \text{ if } z \notin \{-1, 0, 1, 2\} \end{cases} = \begin{cases} 0.28 \text{ if } z = -1 \\ 0.42 \text{ if } z = 0 \\ 0.12 \text{ if } z = 1 \\ 0.18 \text{ if } z = 2 \\ 0 \text{ if } z \notin \{-1, 0, 1, 2\}. \end{cases}$$

At this point it is worthwhile making an important distinction between two types of random variables, namely *discrete* random variables and *continuous* random variables.

We say that a random variable is discrete if the set of values that it can take is at most countable. On the other hand, a random variable taking values in an uncountably infinite set is called continuous.

Question
Consider the following:

- you draw a circle on a piece of paper and one diameter of the circle; at the center of the circle, you keep your pencil standing orthogonal to the plane where the circle lies. At some point you let go the pencil. X is the random variable corresponding to the angle that the pencil forms with the diameter of the circle that you drew after the pencil fell on the piece of paper.

- you roll a die. Y is the random variable corresponding to the number of rolls needed until you observe tail for the first time.

What are the possible values that X and Y can take? Are X and Y discrete or continuous random variables?

Depending on whether a given random variable X is discrete or continuous, we use two special types of functions in order to describe its distribution.

- if X is discrete, let's define its support as the set $\text{supp}(X) = \{x \in \mathbb{R} : P(X = x) > 0\}$ (if X is discrete, $\text{supp}(X)$ is either a finite or countable set). We can describe the probability distribution of X in terms of its *probability mass function* (p.m.f.), i.e. the function

$$p(x) = P(X = x) \tag{17}$$

  mapping $\mathbb{R}$ into $[0, 1]$. The function p satisfies the following properties:

  1. $p(x) \in [0, 1] \quad \forall x \in \mathbb{R}$
  2. $\sum_{x \in \text{supp}(X)} p(x) = 1.$

- if X is continuous, we can describe the probability distribution of X by means of the *probability density function* (p.d.f.) $f : \mathbb{R} \to \mathbb{R}_+$. Define in this case $\text{supp}(X) = \{x \in \mathbb{R} : f(x) > 0\}$. The function f satisfies the following properties:

  1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
  2. $\int_{\mathbb{R}} f(x)\, dx = \int_{\text{supp}(X)} f(x)\, dx = 1.$

We use f to compute the probability of events of the type $\{X \in (a, b]\}$ for $a < b$. In particular, for $a < b$, we have

$$P(X \in (a, b]) = \int_a^b f(x)\, dx \tag{18}$$

Notice that this implies that, for any $x \in \mathbb{R}$, $P(X = x) = 0$ if X is a continuous random variable! Also, if X is a continuous random variable, it is clear from above that

$$P(X \in (a, b]) = P(X \in [a, b]) = P(X \in [a, b)) = P(X \in (a, b)). \tag{19}$$

In general, for any set $A \subset \mathbb{R}$, we have

$$P(X \in A) = \int_A f(x)\, dx.$$

Exercise in class
A group of 4 components is known to contain 2 defectives. An inspector randomly tests the components one at a time until the two defectives are found. Let X denote the number of tests on which the second defective is found. What is the p.m.f. of X? What is the support of X? Graph the p.m.f. of X.

Let D stand for defective and N stand for non-defective. The sample space for this experiment is $\Omega = \{DD, NDD, DND, DNND, NDND, NNDD\}$. The simple events in Ω are equally likely because the inspector samples the components completely at random. Thus, the probability of each simple event in Ω is just $1/|\Omega| = 1/6$. We have

$$\begin{aligned}
&P(X = 2) = P(\{DD\}) = 1/6 \\
&P(X = 3) = P(\{NDD, DND\}) = 2/6 = 1/3 \\
&P(X = 4) = P(\{DNND, NDND, NNDD\}) = 3/6 = 1/2 \\
&P(X = \text{any other number}) = 0.
\end{aligned}$$

Thus, the p.m.f. of X is

$$p(x) = \begin{cases} 1/6 \text{ if } x = 2 \\ 1/3 \text{ if } x = 3 \\ 1/2 \text{ if } x = 4 \\ 0 \text{ if } x \notin \{2, 3, 4\} \end{cases}$$

and $\text{supp}(X) = \{2, 3, 4\}$.

Exercise in class
Consider the following p.d.f. for the random variable X:

$$f(x) = e^{-x}\mathbb{1}_{[0,\infty)}(x) = \begin{cases} e^{-x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0. \end{cases}$$

What is the support of X? Compute $P(2 < X < 3)$. Graph the p.d.f. of X.

The support of X is clearly the set $[0, \infty)$. We have

$$P(2 < X < 3) = P(X \in (2, 3)) = \int_2^3 f(x)\, dx = -e^{-x}\big|_2^3 = e^{-2} - e^{-3}.$$

Another way to describe the distribution of a random variable is by means of its cumulative distribution function (c.d.f). Again we will separate the discrete case and the continuous case.

- If X is a discrete random variable, its c.d.f is defined as the function

$$F(x) = \sum_{\substack{y \leq x \\ y \in \text{supp}(X)}} p(y).$$

  Notice that for a discrete random variable, F is not a continuous function.

- If X is a continuous random variable, its c.d.f is defined as the function

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

  Notice that for a continuous random variable, F is a continuous function.

In both cases, the c.d.f. of X satisfies the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$
2. $\lim_{x \to +\infty} F(x) = 1$
3. $x \leq y \implies F(x) \leq F(y)$
4. F is a right-continuous function, i.e. for any $x \in \mathbb{R}$ we have $\lim_{y \to x^+} F(y) = F(x)$.

Exercise in class
Compute the c.d.f. of the random variable X in the two examples above and draw its graph.

For the example in the discrete case, we have

$$F(x) = \begin{cases} 0 \text{ if } x < 2 \\ 1/6 \text{ if } 2 \leq x < 3 \\ 1/6 + 1/3 = 1/2 \text{ if } 3 \leq x < 4 \\ 1/2 + 1/2 = 1 \text{ if } x \geq 4. \end{cases}$$

For the example in the continuous case, we have

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ 0 + \int_0^x e^{-y}\, dy \text{ if } x \geq 0 \end{cases} = \begin{cases} 0 \text{ if } x < 0 \\ -e^{-y}|_0^x \text{ if } x \geq 0 \end{cases} = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-x} \text{ if } x \geq 0. \end{cases}$$

What is the relationship between the c.d.f. of a random variable and its p.m.f./p.d.f.?

- For a discrete random variable X, let $x_{(i)}$ denote the i-th largest element in $\text{supp}(X)$. Then,

$$p(x_{(i)}) = \begin{cases} F(x_{(i)}) - F(x_{(i-1)}) \text{ if } i \geq 2 \\ F(x_{(i)}) \text{ if } i = 1 \end{cases}$$

- For a continuous random variable X,

$$f(x) = \frac{d}{dy} F(y)\Big|_{y=x}$$

  for any x at which F is differentiable.

Furthermore, notice that for a continuous random variable X with c.d.f. F and for $a < b$ one has

$$\begin{aligned} P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b) \\ = \int_a^b f(x)\, dx = \int_{-\infty}^b f(x)\, dx - \int_{-\infty}^a f(x)\, dx = F(b) - F(a). \end{aligned}$$

This is easy. However, for a discrete random variable one has to be careful with the bounds. Using the same notation as before, for $i < j$ one has

$$\begin{aligned}
P(x_{(i)} < X \leq x_{(j)}) &= F(x_{(j)}) - F(x_{(i)}) \\
P(x_{(i)} \leq X \leq x_{(j)}) &= F(x_{(j)}) - F(x_{(i)}) + p(x_{(i)}) \\
P(x_{(i)} \leq X < x_{(j)}) &= F(x_{(j-1)}) - F(x_{(i)}) + p(x_{(i)}) \\
P(x_{(i)} < X < x_{(j)}) &= F(x_{(j-1)}) - F(x_{(i)}).
\end{aligned}$$

Suppose that the c.d.f. F of a continuous random variable X is strictly increasing. Then F is invertible, meaning that there exists a function $F^{-1} : (0,1) \to \mathbb{R} \cup \{-\infty, +\infty\}$ such that for any $x \in \mathbb{R}$ we have $F^{-1}(F(x)) = x$. Then, for any $\alpha \in (0,1)$ we can define the *α-quantile* of X as the number

$$x_\alpha = F^{-1}(\alpha) \tag{20}$$

with the property that $P(X \leq x_\alpha) = \alpha$. This can be extended to c.d.f. that are not strictly increasing and to p.m.f.'s, but for this class we will not use that extension.

Exercise in class
Consider again a random variable X with p.d.f.

$$f(x) = e^{-x}\mathbb{1}_{[0,\infty)}(x) = \begin{cases} e^{-x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0. \end{cases}$$

Compute the α-quantile of X.

We know from above that

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-x} \text{ if } x \geq 0. \end{cases}$$

For $\alpha \in (0,1)$, set $\alpha = F(x_\alpha) = 1 - e^{-x_\alpha}$. We then have that the α-quantile is

$$x_\alpha = -\log(1-\alpha).$$

Lecture 5

Recommended readings: WMS, sections 3.3, 4.3

Expectation and Variance

The *expectation* (or expected value or mean) is an important operator associated to a probability distribution. Given a random variable X with p.m.f. p (if X is discrete) or p.d.f. f (if X is continuous), its expectation $E(X)$ is defined as

- $E(X) = \sum_{x \in \text{supp}(X)} xp(x)$, if X is discrete
- $E(X) = \int_{\mathbb{R}} xf(x)\,dx = \int_{x \in \text{supp}(X)} xf(x)\,dx$, if X is continuous.

Roughly speaking, $E(X)$ is the 'center' of the distribution of X.

Exercise in class
Consider the random variable X and its p.m.f.

$$p(x) = \begin{cases} 0.2 \text{ if } x = 0 \\ 0.3 \text{ if } x = 1 \\ 0.1 \text{ if } x = 2 \\ 0.4 \text{ if } x = 3 \\ 0 \text{ if } x \notin \{0, 1, 2, 3\}. \end{cases}$$

What is $E(X)$?

By definition we have

$$E(X) = \sum_{x \in \text{supp}(X)} xp(x) = 0{*}0.2{+}1{*}0.3{+}2{*}0.1{+}3{*}0.4 = 0.3{+}0.2{+}1.2 = 1.7.$$

Exercise in class
Consider the random variable X and its p.d.f

$$f(x) = 3x^2 \mathbb{1}_{[0,1]}(x).$$

What is $E(X)$?

Again, by definition

$$E(X) = \int_{\mathbb{R}} xf(x)\,dx = \int_{\text{supp}(X)} xf(x)\,dx = \\ = \int_0^1 x * 3x^2\,dx = 3\int_0^1 x^3\,dx = \frac{3}{4}\,x^4\Big|_0^1 = \frac{3}{4}.$$

Consider a function $g : \mathbb{R} \to \mathbb{R}$ of the random variable X and the new random variable $g(X)$. The expectation of $g(X)$ is simply

- $E(g(X)) = \sum_{x\in\text{supp}(X)} g(x)p(x)$, if X is discrete
- $E(g(X)) = \int_{\mathbb{R}} g(x)f(x)\,dx = \int_{x\in\text{supp}(X)} g(x)f(x)\,dx$, if X is continuous.

Exercise in class
Consider once again the two random variables above and the function $g(x) = x^2$. What is $E(X^2)$?

- In the discrete example, we have

$$E(X^2) = \sum_{x\in\text{supp}(X)} x^2p(x) = 0^2{*}0.2{+}1^2{*}0.3{+}2^2{*}0.1{+}3^2{*}0.4 = 0.3{+}0.4{+}3.6 = 4.3.$$

- In the continous example, we have

$$E(X^2) = \int_{\mathbb{R}} x^2f(x)\,dx = \int_{\text{supp}(X)} x^2f(x)\,dx = \\ = \int_0^1 x^2 * 3x^2\,dx = 3\int_0^1 x^4\,dx = \frac{3}{5}\,x^5\Big|_0^1 = \frac{3}{5}.$$

A technical note: for a random variable X, the expected value $E(X)$ is a well-defined quantity whenever $E(|X|) < +\infty$. However, the opposite is not true. Why is it the case? Let's first prove that $E[|X|] < +\infty$ implies $E[X] < +\infty$. This simply follows from the following inequality

$$E[X] = E[X\mathbb{1}(X \geq 0)] + E[X\mathbb{1}(X < 0)] \\ \leq E[X\mathbb{1}(X \geq 0)] + E[-X\mathbb{1}(X < 0)] = E[|X|] < +\infty.$$

What about the other direction? If we assume that $E[X] < +\infty$, then at most one between $E[X\mathbb{1}(X \geq 0)]$ and $E[X\mathbb{1}(X < 0)]$ needs to be finite. In

particular, we can take $E[X\mathbb{1}(X<0)] = -\infty$, which implies $E[X] = -\infty$, hence $E[|X|] = +\infty$.

The expected value operator E is a linear operator: given two random variables X, Y and scalars $a, b \in \mathbb{R}$ we have

$$\begin{aligned} E(aX) &= aE(x) \\ E(X+Y) &= E(X) + E(Y). \end{aligned} \tag{21}$$

For any scalar $a \in \mathbb{R}$, we have $E(a) = a$. To see this, consider the random variable Y with p.m.f.

$$p(y) = \mathbb{1}_{\{a\}}(x) = \begin{cases} 1 \text{ if } x = a \\ 0 \text{ if } x \neq a. \end{cases}$$

From the definition of $E(Y)$ it is clear that $E(Y) = a$.

Exercise in class
Consider again the continuous random variable X of the previous example. What is $E(X + X^2)$?

We have $E(X + X^2) = E(X) + E(X^2) = \frac{3}{4} + \frac{3}{5} = \frac{27}{20}$.

Beware that in general, for two random variables X, Y, it is not true that $E(XY) = E(X)E(Y)$. However, we shall see later in the class that this is true when X and Y are independent.

Remark: If X and Y are independent, then $f(x)$ and $g(Y)$ are independent. However, the opposite is not true!

The letter μ is often used to denote the expected value $E(X)$ of a random variable X, i.e. $\mu = E(X)$.

Another very important operator associated to a probability distribution is the *variance*. The variance measures how 'spread' the distribution of a random variable is. The variance of a random variable X is defined as

$$V(X) = E[(X-\mu)^2] = E(X^2) - \mu^2. \tag{22}$$

Equivalently, one can write

- if X is discrete:

$$V(X) = \sum_{x \in \text{supp}(X)} (x-\mu)^2 p(x) = \sum_{x \in \text{supp}(X)} x^2 p(x) - \mu^2$$

- if X is continuous:

$$V(X) = \int_{x \in \text{supp}(X)} x^2 f(x)\, dx - \mu^2.$$

Exercise in class
Consider the random variables of the previous examples. What is $V(X)$?

- In the discrete example, $V(X) = E(X^2) - [E(X)]^2 = 4.3 - (1.7)^2 = 4.3 - 2.89 = 1.41$.
- In the continuous example $V(X) = E(X^2) - [E(X)]^2 = 3/5 - (3/4)^2 = 3/5 - 9/16 = 94/80 = 47/40$.

σ^2 is often used to denote the variance $V(X)$ of a random variable X, i.e. $\sigma^2 = V(X)$. The variance of a random variable X is finite as soon as $E(X^2) < \infty$.

Here are two important properties of the variance operator. For any $a \in \mathbb{R}$ we have

- $V(aX) = a^2 V(X)$
- $V(a + X) = V(X)$.

Exercise in class
Consider the continuous random variable of the previous examples. What is $V\left(\sqrt{40/47}X\right)$?

We have

$$V\left(\sqrt{40/47}X\right) = \frac{40}{47}V(X) = \frac{40}{47}\frac{47}{40} = 1.$$

Beware that in general, for two random variables X, Y, it is not true that $V(X + Y) = V(X) + V(Y)$. However, we shall see later in the course that this is true when X and Y are independent.

Markov's Inequality and Tchebysheff's Inequality

Sometimes we may want to compute or approximate the probability of certain events involving a random variable X even if we don't know its distribution (but given that we know its expectation or its variance). Let $a > 0$. The following inequalities are useful in these cases:

$$\begin{aligned} P(|X| \geq a) &\leq \frac{E(|X|)}{a} \text{ (Markov's inequality)} \\ P(|X - \mu| \geq a) &\leq \frac{V(X)}{a^2} \text{ (Tchebysheff's inequality)} \end{aligned} \tag{23}$$

There are some conditions! The Markov's inequality can be used if X is a positive random variable. The Tchebysheff's Inequality can be used for any random variable with finite (literally, not infinite) expected value. Let $\sigma^2 = V(X)$ as usual so that σ is the standard deviation of X. Note that we can conveniently take $a = k\sigma$ and the second inequality then reads

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k\sigma^2} = \frac{1}{k^2} \tag{24}$$

where $\mu = E(X)$. Thus, we can easily bound the probability that X deviates from its expectation by more than k times its standard deviation.

Let's first prove Markov's inequality.

$$aE[\mathbb{1}(|X| \geq a)] = E[a\mathbb{1}(|X| \geq a)] = \begin{cases} 0 \text{ if } |X| < a \\ a \text{ if } |X| \geq a \end{cases}$$

therefore we can upper bound it by $E[|X|]$ and the proof is completed.
Tchebysheff's inequality can be proved via Markov's. Indeed,

$$P(|X - \mu| \geq a) = P\left(\frac{|X - \mu|}{a} \geq 1\right) = P\left(\frac{|X - \mu|^2}{a^2} \geq 1\right) \leq \frac{E[|X - \mu|^2]}{a^2}$$

where the last inequality is thanks to Markov's.
Exercise in class
A call center receives an average of 10,000 phone calls a day, with a standard deviation of $\sqrt{(}2000)$. What is the probability that that there will be more than 15,000 calls ?

Call X the number of phone calls (a random variable) with

1. Using Markov's inequality, we knot that

$$P(X \geq 15,000) \leq \frac{E(X)}{15,000} = 2/3$$

This is quick and easy, but we can do better

2. Using Tchebysheff's, we get

$$P(X \geq 15,000) = P(X - 10,000 \geq 5,000) \leq P(|X - 10,000| \geq 5,000) \leq \frac{2,000}{5,000^2} = 0.000008.$$

This is *much* better than the previous result.

Lecture 6

Recommended readings: WMS, sections 3.4 → 3.8

Independent Random Variables

Consider a collection of n random variables $X_1, \ldots, X_n$ and their *joint probability distribution*. Their joint probability distribution can be described in terms of the joint c.d.f.

$$F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \cdots \cap X_n \leq x_n), \quad (25)$$

in terms of the joint p.m.f. (if the random variables are all discrete)

$$p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n), \quad (26)$$

or in terms of the joint p.d.f. $f_{X_1,\ldots,X_n}$ (if the random variables are all continuous).

The random variables $X_1, \ldots, X_n$ are said to be independent if either of the following holds:

- $F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$
- (discrete case) $p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$
- (continuous case) $f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

Then, if we consider an arbitrary collection of events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$, we have that

$$P(X_1 \in A_1 \cap X_2 \in A_2 \cap \cdots \cap X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

If the random variables also share the same *marginal distribution*, i.e. we have

- $F_{X_i} = F \quad \forall i \in \{1, \ldots, n\}$
- $p_{X_i} = p \quad \forall i \in \{1, \ldots, n\}$ (if the random variables are all discrete)
- $f_{X_i} = f \quad \forall i \in \{1, \ldots, n\}$ (if the random variables are all continuous)

then the random variables $X_1, \ldots, X_n$ are said to be *independent and identically distributed*, usually shortened in *i.i.d.*.

Probability calculations using sums and products

We can make good use of some basic probability concepts to invoke some tools for probability calculations! We know that if random events A and B are independent, then $P(A \cap B) = P(A)P(B)$. Also, if A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$. Take an example of 5 coin flips of an uneven coin that lands on heads with probability p. Then, the probability of 3 heads happening, $P(X = 3)$ for the random variable X which counts the number of heads in this scenario, can be calculated using these two tricks (instead of counting, which we have been doing so far!). Let's do this in two steps:

1. What is the probability of a particular outcome $HHHTT$? This is simply obtained by multiplying p three times and $(1 - p)$ two times. Why is this? because *within* a particular draw, those five coin flips are independent, so each single outcome's probability should be multiplied. (They are not disjoint, as one event does not preclude the possibility of another event!)

2. What about $HHTHT$? This outcome is *disjoint* from the first outcome $HHHTT$, or for that matter, any other outcome who is a combination of 3 H's and 2 T's. They can never happen together, so they are *disjoint* events! (They are not independent, because they will never happen together – in fact, if you know that $HHTHT$ happened, then you definitely know that $HHHTT$ didn't happen!) Also, we know how to count how many such events can happen – it is $\binom{5}{3}$! So, we can calculate the aggregate probability of 3 heads and 2 tails happening by adding the single event's probability $p^3(1-p)^2$ up exactly $\binom{5}{3}$ times! i.e.
$$P(3successesin5trials) = \binom{5}{3} p^3(1-p)^2 \tag{27}$$

In general, we are interested in the probability distribution of the r number of heads out of n trials. We now proceed to learn some common useful discrete probability distributions.

Frequently Used Discrete Distributions

Notation

The symbol $\sim$ is an identifier for a random variable, and specifies its pmf. In statistical jargon, we will say that a random variable has a certain distribution to signify that it has a a certain pmf/pdf.

Preview

Some 'named' discrete distributions that are frequently used are the following:

1. Binomial distribution (outcome of n coin flips)
2. Bernoulli distribution (special case of Binomial, $n = 1$)
3. Multinomial distribution (outcome of n (unfair) die rolls)
4. Geometric distribution (# times to 1st success)
5. Negative binomial distribution (# times to r'th success)
6. Hypergeometric distribution (# times x successes of one class will happen in n samples taken from population N that has two classes of size r and $N - r$)
7. Poisson distribution (number of successes expected in i.i.d., memoryless process; linked to exponential distribution)

The Binomial Distribution

Consider again tossing a coin (not necessarily fair) n times in such a way that each coin flip is independent of the other coin flips. By this, we mean that if H_i denotes the event 'observing heads on the i-th toss', then $P(H_i \cap H_j) = P(H_i)P(H_j)$ for all $i \neq j$. Suppose that the probability of seeing heads on each flip is $p \in [0, 1]$ (and let's call the event 'seeing heads' a *success*). Introduce the random variables

$$Y_i = \begin{cases} 1 \text{ if the i}-th \text{ flip is a success} \\ 0 \text{ otherwise} \end{cases}$$

fo $i \in \{0, 1, 2, \ldots, n\}$. The number of heads that we observe (or the number of successes in the experiment) is

$$X = \sum_{i=1}^{n} Y_i.$$

Under the conditions described above, the random variable X is distributed according to the Binomial distribution with parameters $n \in \mathcal{Z}_+$ (number of trials) and $p \in [0, 1]$ (probability of success in each trial). We denote this by $X \sim \text{Binomial}(n, p)$. The p.m.f. of X is

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ if } x \in \{0, 1, 2, \ldots, n\} \\ 0 \text{ if } x \notin \{0, 1, 2, \ldots, n\} \end{cases} \tag{28}$$

Its expectation is $E(X) = np$ and its variance is $V(X) = np(1-p)$.

In particular, when $n = 1$, we usually say that X is distributed according to the Bernoulli distribution of parameter p, denoted $X \sim \text{Bernoulli}(p)$. In this case $E(X) = p$ and $V(X) = p(1-p)$. Every X_i above is a Bernoulli distributed random variable.

The sum of binomial random variables is also binomial; i.e. it is closed to summation. If X_1, X_2 are independent binomial random variables each with distribution $X_1, X_2 \sim \text{Binom}(n_i, p)$, Then, the sum $Y = X_1 + X_2$ also follows a binomial distribution, with parameters $(n_1 + n_2, p)$.

Exercise in class:

Define $Z = X_1 + X_2$, then calculate $P(Z = z) = P(X_1 + X_2 = z)$ by:

$$\begin{aligned} P(X_1 + X_2 = z) &= \sum_{x=0}^{z} p_{X_1}(x) p_{X_2}(z-x) \\ &= \sum_{x=0}^{z} \binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{z-x} p^{z-x} (1-p)^{n_2 - z + x} \\ &= p^z (1-p)^{n_1 + n_2 - z} \sum_{x=0}^{z} \binom{n_1}{x} \binom{n_2}{z-x} \\ &= \binom{n_1 + n_2}{z} p^z (1-p)^{n_1 + n_2 - z} \end{aligned}$$

where the last equality follows from Vandermonde's identity.

Exercise in class:

Let X denote the number of 6 observed after rolling 4 times a fair die. Then $X \sim \text{Binomial}(n, p)$ with $n = 4$ and $p = 1/6$. What is the probability that we observe at least 3 times the number 6 in the 4 rolls? What is the expected number of times that the number 6 is observed in 4 rolls?

We have

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) = p(3) + p(4) \\ &= \binom{4}{3} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{4-3} + \binom{4}{4} \left(\frac{1}{6}\right)^4 \left(1 - \frac{1}{6}\right)^{4-4} \\ &= 4 * \frac{1}{6^3} \frac{5}{6} + \frac{1}{6^4} = \frac{21}{6^4} \approx 0.016 \end{aligned}$$

and $E(X) = np = 4/6 = 2/3$.

The Geometric Distribution

Consider now counting the number of coin flips needed before the first success is observed in the setting described above and let X denote the corresponding random variable. Then X has the Geometric distribution with parameter $p \in [0,1]$, denoted $X \sim \text{Geometric}(p)$. Its p.m.f. is given by

$$p(x) = \begin{cases} (1-p)^{x-1}p \text{ if } x \in \{1,2,3,\dots\} \\ 0 \text{ if } x \notin \{1,2,3,\dots\} \end{cases} \tag{29}$$

The expected value of X is $E(X) = 1/p$, while its variance is $V(X) = (1-p)/p^2$.

The Geometric distribution is one of the distribution that have the so-called 'memoryless property'. This means that if $X \sim \text{Geometric}(p)$, then

$$P(X > x + y | X > x) = P(X > y)$$

for any $0 < x \leq y$. To see this, let's first compute the c.d.f. of X. We have

$$F(x) = P(X \leq x) = \begin{cases} 0 \text{ if } x < 1 \\ p\sum_{y=1}^{\lfloor x \rfloor}(1-p)^{y-1} \text{ if } x \geq 1 \end{cases} = \begin{cases} 0 \text{ if } x < 1 \\ p\sum_{y=0}^{\lfloor x \rfloor - 1}(1-p)^{y} \text{ if } x \geq 1 \end{cases}$$
$$= \begin{cases} 0 \text{ if } x < 1 \\ p\frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} \text{ if } x \geq 1 \end{cases} = \begin{cases} 0 \text{ if } x < 1 \\ 1-(1-p)^{\lfloor x \rfloor} \text{ if } x \geq 1. \end{cases}$$

Let's look for example to the case $1 < x < y$. We have

$$P(X > x + y | X > x) = \frac{P(X > x+y)}{P(X > x)} = \frac{1 - F(x+y)}{1 - F(x)} = \frac{(1-p)^{\lfloor x+y \rfloor}}{(1-p)^{\lfloor x \rfloor}}$$
$$= (1-p)^{\lfloor y \rfloor} = P(X > y).$$

Exercise in class

Consider again rolling a fair die. What is the probability that the first 6 is rolled on the 4-th roll?

Let X denote the random variable counting the number of rolls needed to observe the first 6. We have $X \sim \text{Geometric}(p)$ with $p = 1/6$. Then,

$$P(X = 4) = p(4) = p(1-p)^{4-1} = \frac{1}{6}\frac{5^3}{6^3} = \frac{5^3}{6^4} \approx 0.096.$$

The Negative Binomial Distribution

Suppose that we are interested in counting the number of coin flips needed in order to observe the r-th success. If X denotes the corresponding random variable, then X has the Negative Binomial distribution with parameters $r \in \mathbb{Z}_+$ (number of successes) and $p \in [0,1]$ (probability of success on each trial). We use the notation $X \sim \text{NBinomial}(r,p)$. Its p.m.f. is then

$$p(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} \text{ if } x \in \{r, r+1, \dots\} \\ 0 \text{ if } x \notin \{r, r+1, \dots\}. \end{cases} \tag{30}$$

The expected value of X is $E(X) = r/p$ and its variance is $V(X) = r(1-p)/p^2$.

Exercise in class:
Consider again rolling a fair die. What is the probability that the 3rd 6 is observed on the 5-th roll?

We have $X \sim \text{NBinomial}(r,p)$ with $r = 3$ and $p = 1/6$ and

$$P(X = 5) = \binom{5-1}{3-1} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{5-3} = 6 * \frac{5^2}{6^5} \approx 0.019.$$

Question:
How do the Geometric distribution and the Negative Binomial ditribution relate?

The Hypergeometric Distribution

Suppose that you have a population of N elements that can be divided into 2 subpopulations of size r and $N - r$ (say the population of 'successes' and 'failures', respectively). Imagine that you sample $n \leq N$ elements from this population without replacement. What is the probability that your sample contains x successes? To answer this question we can introduce the random variable X with the Hypergeometric distribution with parameters $r \in \mathbb{Z}_+$ (number of successes in the population), $N \in \mathbb{Z}_+$ population size, and $n \in \mathbb{Z}_+$ (number of trials). We write $X \sim \text{HGeometric}(r, N, n)$. Its p.m.f. is

$$p(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \text{ if } x \in \{0, 1, \dots, r\} \\ 0 \text{ if } x \notin \{0, 1, \dots, r\}. \end{cases} \tag{31}$$

The expectation and the variance of X are $E(X) = nr/N$ and

$$V(X) = \frac{nr}{N}\frac{N-r}{N}\frac{N-n}{N-1}.$$

Exercise in class:
There are 10 candidates for 4 software engineering positions at Google. 7 of them are female candidates and the remaining 3 are male candidates. If the selection process at Google is completely random, what is the probability that only 1 male candidate will be eventually hired?

The number of male candidates hired by Google can be described by means of the random variable $X \sim \text{Hypergeometric}(r, N, n)$ with $r = 3$, $N = 10$, and $n = 4$. Then,

$$P(X = 1) = \frac{\binom{3}{1}\binom{7}{3}}{\binom{10}{4}} = \frac{3\frac{7!}{4!3!}}{\frac{10!}{6!4!}} = \frac{7!6!}{2 * 10!} = \frac{6 * 5 * 4 * 3}{10 * 9 * 8} = \frac{1}{2}.$$

The Poisson Distribution

The Poisson distribution can be thought of as a limiting case of the Binomial distribution. Consider the following example: you own a store and you want to model the number of people who enter in your store on a given day. In any time interval of that day, the number of people walking in the store is certainly discrete, but in principle that number can be any non-negative integer. We could try and divide the day into n smaller subperiods in such a way that, as $n \to \infty$, only one person can walk into the store in any given subperiod. If we let $n \to \infty$, it is clear however that the probability p that a person will walk in the store in an infinitesimally small subperiod of time is such that $p \to 0$. The Poisson distribution arises as a limiting case of the Binomial distribution when $n \to \infty$, $p \to 0$, and $np \to \lambda \in (0, \infty)$.

The p.m.f. of a random variable X that has the Poisson distribution with parameter $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$ is

$$p(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} \text{ if } x \in \{0, 1, 2, 3, \dots\} \\ 0 \text{ if } x \notin \{0, 1, 2, 3, \dots\} \end{cases} \tag{32}$$

Both the expected value and the variance of X are equal to λ, $E(X) = V(X) = \lambda$.

Poisson random variables have the convenient property of being closed to summation. if $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ and they are

independent, then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. **Exercise in class:** Prove this, using pmfs. Define Z, then calculate $P(Z = z) = P(X + Y = z)$ by:

$$P(X + Y = z) = \sum_{x=0}^{z} f_X(x) f_Y(z - x)$$

Exercise in class:
At the CMU USPS office, the expected number of students waiting in line between 1PM and 2PM is 3. What is the probability that you will see more than 2 students already in line in front of you, if you go to the USPS office in that period of time?

Let $X \sim \text{Poisson}(\lambda)$ with $\lambda = 3$ be the number of students in line when you enter into the store. We want to compute $P(X > 2)$. We have

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \sum_{x=0}^{2} p(x) = 1 - e^{-3} \sum_{x=0}^{2} \frac{3^x}{x!} =$$
$$= 1 - e^{-3} \left(1 + 3 + \frac{9}{2}\right) = 1 - e^{-3} \frac{17}{2} \approx 0.577.$$

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. Suppose n i.i.d. experiments are performed. Each experiment can lead to r possible outcomes with probability $p_1, p_2, \cdots, p_r$ such that

$$p_i > 0, \sum_{i=0}^{r} p_i = 1$$

Now, let X_i be the number of experiments resulting in outcome $i \in [1, r]$. Then, the multinomial distribution characterizes the joint probability of $(X_1, X_2, \cdots, X_n)$; in other words, it fully describes

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_r = x_r) \tag{33}$$

The probability mass function is

$$p_{X_1,\cdots,X_r}(x_1, \cdots, x_r) = \begin{cases} \binom{n}{x_1\ x_2 \cdots x_r} p_1^{x_1} \cdots p_r^{x_r} & \text{if } \sum_{i=1}^{r} x_i = n \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

What is the mean? The mean should be defined on each of the X_i, and is $E(X_i) = np_i$. The variance is $V(X_i) = np_i(1 - p_i)$. We will later prove

Age	Proportion
18-24	.18
25-34	.23
35-44	.16
45-64	.27
65+	.16

(after we have learned about the concept of covariance) that the variance of each X_i is $V(X_i) = Cov(X_i, X_i) = np_i(1 - p_i)$. The covariance of X_i and X_j is $Cov(X_i, X_j) = -np_ip_j$.

How should we understand the expectation and variance? We can see that the *marginal* probability distribution of X_i as simply Binom(n, p_i) – if we only focus on one variable at a time, each is simply the number of successes (each success having probability p_i) out of n trials!

One might argue that it somehow doesn't seem like the other outcomes are 'irrelevant', so that each X_i can be treated as a separate binomial random variable. He would be partially correct! (about his former point) The outcomes of X_i and X_j are 'pitted' against each other; if X_i is high, then $X_j (j \neq i)$ should be low, since there are only n draws in total! Indeed, we will find that these are negatively correlated (and have negative covariance).

Exercise in class

A fair die is rolled 9 times. What is the probability of 1 appearing 3 times, 2 appearing 2 times, 3 appearing 2 times, 4 appearing 1 times, 5 appearing 1 time, and 6 appearing 0 times?

Exercise in class

According to recent census figures, the proportion of adults (18 years or older of age) in the U.S. associated with 5 age categories are as given in the following table If the figures are accurate and five adults are randomly sampled, find the probability that the sample contains one person between the ages of 18 and 24, two between the ages of 25 and 34, and two between 45 and 64. (Hint: see WMS)

Lecture 7

Recommended readings: WMS, sections 4.4 → 4.8

Frequently Used Continuous Distributions

In this section, we will describe some of the most commonly used continuous distribution. In particular, we will focus on three important families which are often used in the statistical modeling of physical phenomena:

1. the Normal family: this is a class of distributions that is commonly used to describe physical phenomena where the quantity of interest takes values (at least in principle) in the range $(-\infty, \infty)$

2. the Gamma family: this is a class of distributions which are frequently used to describe physical phenomena where the quantity of interest takes non-negative values, i.e. it takes values in the interval $[0, \infty)$

3. the Beta family: this is a class of distributions that is commonly used to describe physical phenomena where the quantity of interest takes values in some interval $[a, b]$ of the real line.

We will also focus on some relevant subfamilies of the families mentioned above.

The Uniform Distribution

The Uniform distribution is the simplest among the continuous distributions. We say that a random variable X has the Uniform distribution with parameters $a, b \in \mathbb{R}$, $a < b$, if its p.d.f. is

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} \text{ if } x \in [a, b] \\ 0 \text{ if } x \notin [a, b]. \end{cases} \tag{35}$$

In this case, we use the notation $X \sim \text{Uniform}(a, b)$. We have $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$.

The Normal Distribution

The Normal distribution is of the most relevant distributions for applications in Statistics. We say that a random variable X has a Normal distribution

with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$ if its p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In this case, we write $X \sim \mathcal{N}(\mu, \sigma)$. The parameters of X correspond to its expectation $\mu = E(X)$ and its variance $\sigma^2 = V(X)$. The c.d.f. of $X \sim \mathcal{N}(\mu, \sigma^2)$ can be expressed in terms of the *error function* $\text{erf}(\cdot)$. However, for our purposes, we will not need to investigate this further.

The Standardized Normal Distribution

Given $X \sim \mathcal{N}(\mu, \sigma^2)$, we can always obtain a 'standardized' version Z of X such that Z still has a Normal distribution, $E(Z) = 0$, and $V(Z) = 1$ (i.e. $Z \sim \mathcal{N}(0, 1)$). This can be done by means of the transformation

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

The random variable Z is said to be *standardized*. Of course, one can also standardize random variables that have other distributions (as long as they have finite variance), but unlike the Normal case the resulting standardized variable may not belong anymore to the same family to which the original random variable X belonged to.

The Normal family is *closed* with respect to translation and scaling: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ for $a \neq 0$ and $b \in \mathbb{R}$.

Furthermore, if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$ and X and Y are independent, then $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$.

We finally mention that it is common notation to indicate the c.d.f. of $Z \sim \mathcal{N}(0, 1)$ by $\Phi(\cdot)$. Notice that $\Phi(-x) = 1 - \Phi(x)$ for any $x \in \mathbb{R}$.

The Gamma Distribution

We say that the random variable X has a Gamma distribution with parameters $\alpha, \beta > 0$ if its p.d.f. is

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \mathbb{1}_{[0,\infty)}(x). \tag{36}$$

Notice that the p.d.f. of the Gamma distribution includes the Gamma *function*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \tag{37}$$

for $\alpha > 0$. Useful properties of the Gamma function:

- For $\alpha \in \mathbb{Z}_+$, we have that $\Gamma(\alpha) = (\alpha - 1)!$.
- For any $\alpha > 0$, we have $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. [6]

Make sure not to confuse the Gamma distribution, described by the p.d.f. of equation (36), and the Gamma function of equation (37)!

The expectation and the variance of X are respectively $E(X) = \alpha\beta$ and $V(X) = \alpha\beta^2$.

The c.d.f. of a Gamma-distributed random variable can be expressed explicitly in terms of the *incomplete Gamma function*. Once again, for our purposes, we don't need to investigate this further.

We saw that the Normal family is closed with respect to translation and scaling. The Gamma family is closed with respect to positive scaling only. If $X \sim \text{Gamma}(\alpha, \beta)$, then $cX \sim \text{Gamma}(\alpha, c\beta)$, provided that $c > 0$. Furthermore, if $X_1 \sim \text{Gamma}(\alpha_1, \beta)$, $X_2 \sim \text{Gamma}(\alpha_2, \beta)$, ..., $X_n \sim \text{Gamma}(\alpha_n, \beta)$ are independent random variables, then $\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$

The Exponential Distribution

The Exponential distribution constitutes a subfamily of the Gamma distribution. In particular, X is said to have an Exponential distribution with parameter $\beta > 0$ if $X \sim \text{Gamma}(1, \beta)$. In that case, we write $X \sim \text{Exponential}(\beta)$.

The p.d.f. of X is therefore

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \mathbb{1}_{[0,\infty)}(x).$$

Because the Exponenetial distribution is a subfamily of the Gamma distribution, we have $E(X) = \alpha\beta = \beta$ and $V(X) = \alpha\beta^2 = \beta^2$.

Exercise in class

Compute the c.d.f. of $X \sim \text{Exponential}(\beta)$.

We have,

$$F(x) = P(X \leq x) = \begin{cases} 0 \text{ if } x < 0 \\ \int_0^x \frac{1}{\beta} e^{-\frac{y}{\beta}}\, dy \text{ if } x \geq 0 \end{cases} = \begin{cases} 0 \text{ if } x < 0 \\ -\, e^{-\frac{y}{\beta}} \Big|_0^x \text{ if } x \geq 0. \end{cases}$$

$$= \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-\frac{x}{\beta}} \text{ if } x \geq 0. \end{cases}$$

[6] (This recursive property often comes in handy for computations that involve Gamma-distributed random variables.)

We will learn a bit more about this when we revisit Poisson processes and exponential wait times of events, but we note here that the exponential distribution also has the property of being 'memoryless'. For an exponential random variable X with parameter λ, the following hold

$$P(X > t + s | X > t) = P(X > s) \tag{38}$$

or, equivalently

$$P(X > t + s) = P(X > s)P(X > t) \tag{39}$$

Exercise in class
Prove this. Hint: consider using 1-CDF of exponential distributions.

The Chi-Square Distribution

The Chi-Square distribution is another relevent subfamily of the Gamma family of distributions. It frequently arises in statistical model-fitting procedures. We say that X has a Chi-Square distribution with $\nu > 0, \nu \in \mathbb{N}^+$ 'degrees of freedom', if $X \sim \text{Gamma}(\nu/2, 2)$. In this case, we write $X \sim \chi^2(\nu)$. We have $E(X) = \nu$ and $V(X) = 2\nu$.

The Chi-square distribution can be formed by the sum of ν independent standard normal distributions:

$$\text{If } X = \sum_{i=1}^{\nu} X_i^2, \quad \text{then} \quad X \sim \chi^2(\nu)$$

This distribution is useful when we want to verify (i.e. conduct hypothesis tests) if two categories are independent (see Pearson's Chi-square test of independence) or if a prior belief about proportions of the population in some category (say, male/female, or income group) is plausible (see Goodness of fit test for further study).

The Beta distribution

Suppose that you are interested in studying a quantity Y that can only take values in the interval $[a, b] \subset \mathbb{R}$. We can easily transform Y in such a way that it can only take values in the standard unit interval $[0, 1]$: it is enough to consider the normalized version of Y

$$X = \frac{Y - a}{b - a}.$$

Then, a flexible family of distributions that can be used to model Y is the Beta family. We say that a random variable X has a Beta distribution with parameters $\alpha, \beta > 0$, denoted $X \sim \text{Beta}(\alpha, \beta)$, if its p.d.f. is of the form

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \mathbb{1}_{[0,1]}(x). \tag{40}$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The c.d.f. of X can be expressed explicitly in terms of the *incomplete Beta function*

$$B(x;\alpha,\beta) = \int_0^x y^{\alpha-1}(1-y)^{\beta-1}\,dx,$$

but we don't need to investigate this further for our purposes.

The expected value of X is $E(X) = \frac{\alpha}{\alpha+\beta}$ and its variance is $V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

A Note on the Normalizing Constant of a Probability Density Function

Most frequently, a given p.d.f. f takes the form

$$f(x) = cg(x)$$

where c is a positive constant and g is a function. The part of f depending on x, i.e. the function g, is usually called the *kernel* of the p.d.f. f. Very often one can guess whether f belongs to a certain family by simply inspecting g. Then, if f is indeed a density, $c > 0$ is exactly the 'right' constant which makes f integrate to 1. Therefore, if for any reason c is unknown, but you can guess that $X \sim f$ belongs to a certain family of distributions for a particular value of its parameters, in order to figure out c one does not necessarily have to compute

$$c = \left(\int_{\text{supp}(X)} g(x)\,dx \right)^{-1}.$$

Let us illustrate this by means of two examples:

- Let $f(x) = ce^{-\frac{x}{2}} \mathbb{1}_{[0,\infty)}(x)$ and we want to figure out what c is. Here $g(x) = e^{-\frac{x}{2}} \mathbb{1}_{[0,\infty)}(x)$ is the kernel of an Exponential p.d.f. with parameter $\beta = 2$. We know therefore that c must be equal to $c = 1/\beta$.

- Let $f(x) = cx^4(1-x)^5\mathbb{1}_{[0,1]}(x)$ and again suppose that we want to figure out the value of c. Here $g(x) = x^4(1-x)^5\mathbb{1}_{[0,1]}(x)$ which is the kernel of a Beta p.d.f. with parameters $\alpha = 5$ and $\beta = 6$. We know therefore that c must be the number

$$c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\Gamma(5+6)}{\Gamma(5)\Gamma(6)} = \frac{10!}{4!5!} = 1260.$$

Extra: Simulations of random variables

Fun stuff: If we have time, we will try some simulations of some random variables (Normal, Poisson, Exponential, Gamma with different parameters, etc.), overlay pmf/pdfs on them to show dependence on n. Also, as a preview, we will see simulated examples of the central limit theorem phenomenon. Also, we will see 2d scatter plots and 3d density plots of a independent/positively correlated normal distributions.

Lecture 8

Recommended readings: WMS, sections 5.1 → 5.4

Multivariate Probability Distributions

So far we have focused on *univariate* probability distributions, i.e. probability distributions for a single random variable. However, when we discussed independence of random variables in Lecture 6, we introduced the notion of joint c.d.f., joint p.m.f., and joint p.d.f. for a collection of n random variables $X_1, \ldots, X_n$. In this lecture we will elaborate more on these objects.

Let $X_1, \ldots, X_n$ be a collection of n random variables.

- Regardless of whether they are discrete or continuous, we denote by $F_{X_1,\ldots,X_n}$ their joint c.d.f., i.e. the function

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1 \cap \cdots \cap X_n \leq x_n).$$

- If they are all discrete, we denote by $p_{X_1,\ldots,X_n}$ their joint p.m.f., i.e. the function

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1 \cap \cdots \cap X_n = x_n).$$

- If they are all continuous, we denote by $f_{X_1,\ldots,X_n}$ their joint p.d.f..

The above functions satisfy properties that are similar to those satisfied by their univariate counterparts.

- The joint c.d.f. $F_{X_1,\ldots,X_n}$ satisfies:
  - $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \in [0, 1]$ for any $x_1, \ldots, x_n \in \mathbb{R}$.
  - $F_{X_1,\ldots,X_n}$ is monotonically non-decreasing in each of its variables
  - $F_{X_1,\ldots,X_n}$ is càdlàg (right-continuous with left limits with respect to every variable)
  - $\lim_{x_i \to -\infty} F_{X_1,\ldots,X_n}(x_1, \ldots, x_i, \ldots, x_n) = 0$ for any $i \in \{1, \ldots, n\}$
  - $\lim_{x_1 \to +\infty, \ldots, x_n \to +\infty} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$
- The joint p.m.f. satisfies:
  - $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \in [0, 1]$ for any $x_1, \ldots, x_n \in \mathbb{R}$.
  - $\sum_{x_1 \in \text{supp}(X_1)} \cdots \sum_{x_n \in \text{supp}(X_n)} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$.

- The joint p.d.f. satisfies:
  - $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \geq 0$ for any $x_1,\ldots,x_n \in \mathbb{R}$.
  - $\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$ $= \int_{\text{supp}(X_1)} \cdots \int_{\text{supp}(X_n)} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_n \ldots dx_1 = 1.$

Furthermore we have

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \sum_{\substack{y_1 \leq x_1 \\ y_1 \in \text{supp}(X_1)}} \cdots \sum_{\substack{y_n \leq x_n \\ y_n \in \text{supp}(X_n)}} p(y_1,\ldots,y_n)$$

and

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(y_1,\ldots,y_n)\, dy_n \ldots dy_1$$

Exercise in class:
You are given the following bivariate p.m.f.:

$$p_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1,x_2) = (0,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (0,0) \\ \frac{1}{8} \text{ if } (x_1,x_2) = (0,1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,0) \\ 0 \text{ otherwise.} \end{cases}$$

What is $P(X_1 \leq 1 \cap X_2 \leq 0) = F_{X_1,X_2}(1,0)$? We have

$$F_{X_1,X_2}(1,0) = p_{X_1,X_2}(0,-1) + p_{X_1,X_2}(0,0) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}.$$

Draw figure of bivariate pmf here.

Exercise in class:
You are given the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1,x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1,x_2)$$

- What is $P(X_1 \leq 1 \cap X_2 > 5)$?
- What is $P(X_1 + X_2 \leq 3)$?

We have

$$P(X_1 \leq 1 \cap X_2 > 5) = \int_{\infty}^{1} \int_{5}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 = \int_{0}^{1} \int_{5}^{\infty} e^{-(x_1+x_2)}\, dx_1 dx_2$$
$$= \int_{0}^{1} e^{-x_1}\, dx_1 \int_{5}^{\infty} e^{-x_2}\, dx_2 = \left(-\, e^{-x_1}\big|_0^1\right)\left(-\, e^{-x_2}\big|_5^\infty\right)$$
$$= \left(1 - e^{-1}\right) e^{-5}.$$

and

Draw figure of bivariate pdf here.

$$P(X_1 + X_2 \leq 3) = \int_{0}^{3} \int_{0}^{3-x_1} e^{-(x_1+x_2)}\, dx_2 dx_1 = \int_{0}^{3} e^{-x_1} \int_{0}^{3-x_1} e^{-x_2}\, dx_2 dx_1$$
$$= \int_{0}^{3} e^{-x_1} \left(-e^{-x_2}\big|_0^{3-x_1}\right) dx_1 = \int_{0}^{3} e^{-x_1} \left(1 - e^{x_1-3}\right) dx_1$$
$$= \int_{0}^{3} \left(e^{-x_1} - e^{-3}\right) dx_1 = \left(-\, e^{-x_1}\big|_0^3\right) - 3e^{-3} = 1 - 4e^{-3}.$$

Marginal Distributions

Given a collection of random variables $X_1, \ldots, X_n$ and their joint distribution, how can we derive the *marginal* distribution of only one of them, say X_i? The idea is summing or integrating the joint distribution over all the variables except for X_i.

Draw figure of marginalizing a bivariate pmf here.

Thus, given $p_{X_1,\ldots,X_n}$ we have that

$$p_{X_i}(x_i) =$$
$$\sum_{y_1 \in \text{supp}(X_1)} \cdots \sum_{y_{i-1} \in \text{supp}(X_{i-1})} \sum_{y_{i+1} \in \text{supp}(X_{i+1})} \cdots \sum_{y_n \in \text{supp}(X_n)} p_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n).$$

and given $f_{X_1,\ldots,X_n}$ we have

$$f_{X_i}(x_i) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)\, dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1$$
$$= \int_{\text{supp}(X_1)} \cdots \int_{\text{supp}(X_{i-1})} \int_{\text{supp}(X_{i+1})} \cdots \int_{\text{supp}(X_n)} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$$
$$dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1.$$

Exercise in class:
Consider again the bivariate p.m.f.

$$p_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1,x_2) = (0,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (0,0) \\ \frac{1}{8} \text{ if } (x_1,x_2) = (0,1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,0) \\ 0 \text{ otherwise.} \end{cases}$$

Derive the marginal p.m.f. of X_2.

Notice first that $\text{supp}(X_2) = \{-1, 0, 1\}$. We have

$$\begin{aligned} p_{X_2}(-1) &= p_{X_1,X_2}(0,-1) + p_{X_1,X_2}(2,-1) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8} \\ p_{X_2}(0) &= p_{X_1,X_2}(0,0) + p_{X_1,X_2}(2,0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\ p_{X_2}(1) &= p_{X_1,X_2}(0,1) = \frac{1}{8}. \end{aligned}$$

Thus,

$$p_{X_2}(x_2) = \begin{cases} \frac{3}{8} \text{ if } x_2 = -1 \\ \frac{1}{2} \text{ if } x_2 = 0 \\ \frac{1}{8} \text{ if } x_2 = 1 \\ 0 \text{ if } x_2 \notin \{-1,0,1\} \end{cases}$$

Exercise in class:
Consider again the bivariate p.d.f.

$$f_{X_1,X_2}(x_1,x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1,x_2).$$

Derive the marginal p.d.f. of X_1.

Notice first that $\text{supp}(X_1) = [0,\infty)$. For $x_1 \in [0,\infty)$ we have

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{X_1,X_2}(x_1,x_2)\, dx_2 = e^{-x_1} \int_0^\infty e^{-x_2}\, dx_2 = e^{-x_1}.$$

Thus,

$$f_{X_1}(x_1) = e^{-x_1} \mathbb{1}_{[0,\infty)}(x_1).$$

Conditional Distributions

We will limit our discussion here at the bivariate case, but all that follows is easily extended to more than two random variables.

Suppose that we are given a pair of random variables X_1, X_2 and we want to compute probabilities of the type

$$P(X_1 \in A_1 | X_2 = x_2)$$

for a particular fixed value x_2 of X_2. To do this, we need either the conditional p.m.f. (if X_1 is discrete) or the conditional p.d.f. (if X_1 is continuous) of X_1 given $X_2 = x_2$. By definition, we have

$$p_{X_1|X_2=x_2}(x_1) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$
$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Notice the two following impotant facts:

1. $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ are not well-defined if x_2 is such that $p_{X_2}(x_2) = 0$ and $f_{X_2}(x_2) = 0$ respectively (i.e. $x_2 \notin \text{supp}(X_2)$)

2. given $x_2 \in \text{supp}(X_2)$, $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ are null whenever $p_{X_1,X_2}(x_1, x_2) = 0$ and $f_{X_1,X_2}(x_1, x_2) = 0$ respectively.

So whenever you are computing a conditional distribution, it is good practice to 1) determine the support of the conditioning variable X_2 and clearly state that the conditional distribution that you are about to compute is only well-defined for $x_2 \in \text{supp}(X_2)$ and 2) given $x_2 \in \text{supp}(X_2)$, clarify for which values of $x_1 \in \mathbb{R}$ the conditional distribution $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ is null.

Exercise in class:
Consider again the bivariate p.m.f.

$$p_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1, x_2) = (0, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (0, 0) \\ \frac{1}{8} \text{ if } (x_1, x_2) = (0, 1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, 0) \\ 0 \text{ otherwise.} \end{cases}$$

Derive the conditional distribution of X_1 given X_2.

First of all notice that the conditional p.m.f. of X_1 given $X_2 = x_2$ is only well-defined for $x_2 \in \text{supp}(X_2) = \{-1, 0, 1\}$. Then we have

$$p_{X_1|X_2=-1}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,-1)}{p_{X_2}(-1)} \text{ if } x_1 = 0 \\ \frac{p_{X_1,X_2}(2,-1)}{p_{X_2}(-1)} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases} = \begin{cases} \frac{1/8}{3/8} \text{ if } x_1 = 0 \\ \frac{1/4}{3/8} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$= \begin{cases} \frac{1}{3} \text{ if } x_1 = 0 \\ \frac{2}{3} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$p_{X_1|X_2=0}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,0)}{p_{X_2}(0)} \text{ if } x_1 = 0 \\ \frac{p_{X_1,X_2}(2,0)}{p_{X_2}(0)} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases} = \begin{cases} \frac{1/4}{1/2} \text{ if } x_1 = 0 \\ \frac{1/4}{1/2} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$= \begin{cases} \frac{1}{2} \text{ if } x_1 = 0 \\ \frac{1}{2} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$p_{X_1|X_2=1}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,1)}{p_{X_2}(1)} \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0 \end{cases} = \begin{cases} \frac{1/8}{1/8} \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0 \end{cases}$$

$$= \begin{cases} 1 \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0. \end{cases}$$

Exercise in class:
Consider again the bivariate p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2).$$

Derive the conditional p.d.f. of X_2 given X_1.

First of all notice that the conditional p.m.f. of X_2 given $X_1 = x_1$ is only well-defined for $x_1 \in \text{supp}(X_1) = [0, \infty)$. For $x_1 \in [0, \infty)$, we have

$$f_{X_2|X_1=x_1}(x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} = \frac{e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2)}{e^{-x_1} \mathbb{1}_{[0,\infty)}(x_1)}$$

$$= \frac{e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)}(x_2)}{e^{-x_1}} = e^{-x_2} \mathbb{1}_{[0,\infty)}(x_2).$$

Technical note: you may wonder why, if X_1, X_2 are continuous r.v's, $P(X_1 \in A_1 | X_2 = x_2)$ is even defined after all. Indeed, the event $P(X_2 = x_2)$ has probability zero. However, intuitively this probability should still make sense. In order to overcome this kind of issue, we should dive into *regular* conditional probabilities. In its simplest form, think about the following decompoition

$$F_{X_1}(x_1) = \int_{-\infty}^{+\infty} F_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) dx_2.$$

We also know that

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} f_{X_1}(y_1) dy_1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} f_{X_1,X_2}(y_1, y_2) dy_1 dy_2.$$

Therefore we would like the following equality to hold

$$F_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) = \int_{-\infty}^{x_1} f_{X_1,X_2}(y_1, y_2) dy_1.$$

We call exactly this expression the conditional density function of X_1 with respect to X_2.

A Test for the Independence of Two Random Variables

Suppose that you are given $(X_1, X_2) \sim f_{X_1,X_2}$. If

1. the support of f_{X_1,X_2} is a 'rectangular' region and
2. $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ for all $x_1, x_2 \in \mathbb{R}$

then X_1 and X_2 are independent. The same is true for the case where (X_1, X_2) is discrete, after obvious changes to 1). Notice that 1) can be conveniently captured by appropriately using indicator functions.

Exercise in class:
Consider the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1, x_2) = 6(1 - x_2) \mathbb{1}_{\{0 \leq x_1 \leq x_2 \leq 1\}}(x_1, x_2).$$

Are X_1 and X_2 independent?

A Note on Independence and Normalizing Constants

Frequently, the p.d.f. (and the same holds true for a p.m.f.) of a pair of random variables takes the form

$$f_{X_1,X_2}(x_1, x_2) = cg(x_1)h(x_2) \tag{41}$$

where $c > 0$ is a constant, g is a function depending only on x_1 and h is a function depending only on x_2. If this is the case, the two random variables X_1 and X_2 are independent and g and h are the kernels of the p.d.f. of X_1 and X_2 respectively. Thus, by simply inspecting g and h we can guess to which family of distributions X_1 and X_2 belong to. We do not need to worry about the constant c, as we know that if f_{X_1,X_2} is a bona fide p.d.f., then $c = c_1c_2$ is exactly equal to the product of the normalizing constants c_1 and c_2 that satisfy

$$c_1 \int_{\mathbb{R}} g(x)\,dx = c_2 \int_{\mathbb{R}} h(x)\,dx = 1. \tag{42}$$

This easily generalizes to a collection of $n > 2$ random variables.

Lecture 9

Recommended readings: WMS, sections 5.5 → 5.8

Expected Value in the Context of Bivariate Distributions (part 1)

In Lecture 5, we introduced the concept of expectation or expected value of a random variable. We will now introduce other operators that are based on the expected value operator and we will investigate how they act on bivariate distributions. While we focus on bivariate distributions, it is worthwile to keep in mind that all that we discuss in this Lecture can be extended to collections of random variables $X_1, \ldots, X_n$ with $n > 2$.

For a function

$$\begin{aligned} g :& \mathbb{R}^2 \to \mathbb{R} \\ & (x_1, x_2) \mapsto g(x_1, x_2) \end{aligned}$$

and a pair of random variables (X_1, X_2) with joint p.m.f. p_{X_1,X_2} (if discrete) or joint p.d.f. f_{X_1,X_2} (if continuous), the expected value of $g(X_1, X_2)$ is defined as

$$E(g(X_1, X_2)) = \sum_{x_1 \in \text{supp}(X_1)} \sum_{x_2 \in \text{supp}(X_2)} g(x_1, x_2) p_{X_1,X_2}(x_1, x_2) \text{ (discrete case)}$$

$$E(g(X_1, X_2)) = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \text{ if } X_1 \text{ and } X_2 \text{ (continuous case).}$$

Expectation, Variance and Independence

In Lecture 5, we pointed out that in general, given two random variables X_1 and X_2, it is not true that $E(X_1X_2) = E(X_1)E(X_2)$. We said, however, that it is true as soon as X_1 and X_2 are independent. Let's see why. Without loss of generality, assume that X_1 and X_2 are continous (for the discrete case, just change integration into summation as usual). If they are independent, $f_{X_1,X_2} = f_{X_1} f_{X_2}$. Take $g(x_1, x_2) = x_1 x_2$. Then, we have

$$\begin{aligned} E(X_1X_2) &= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1 dx_2 = \int_{\mathbb{R}} x_1 f_{X_1}(x_1)\, dx_1 \int_{\mathbb{R}} x_2 f_{X_2}(x_2)\, dx_2 \\ &= E(X_1)E(X_2). \end{aligned}$$

The above argument is easily extended to show that for any two functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$, if X_1 and X_2 are independent, then $E(g_1(X_1)g_2(X_2)) = E(g_1(X_1)g_2(X_2))$.

Therefore remember that independence implies $E[XY] = E[X]E[Y]$, but the opposite does not hold always hold.

Exercise in class:
Consider the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1, x_2) = \frac{3}{2}e^{-3x_1}\mathbb{1}_{[0,\infty)\times[0,2]}(x_1, x_2).$$

Are X_1 and X_2 independent? Why? What are their marginal distributions? Compute $E(3XY + 3)$.

Exercise in class:
Consider the pair of random variables X_1 and X_2 with joint p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{8}x_1e^{-(x_1+x_2)/2}\mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2)$$

and consider the function $g(x, y) = y/x$. What is $E(g(X_1, X_2))$?

First of all, notice that X_1 and X_2 are independent. Therefore, we know already that $E(g(X_1, X_2)) = E(X_2/X_1) = E(X_2)E(1/X_1)$. Moreover, the kernel of the p.d.f. of X_1 is

$$x_1e^{-x_1/2}\mathbb{1}_{[0,\infty)}(x_1)$$

while that of the p.d.f. of X_2 is

$$e^{-x_2/2}\mathbb{1}_{[0,\infty)}(x_2).$$

Thus, X_1 and X_2 are distributed according to Gamma$(2, 2)$ and Exponential(2) respectively. It follows that $E(X_2) = \beta = 2$. We only need to compute $E(1/X_1)$. This is equal to

$$\begin{aligned} E\left(\frac{1}{X_1}\right) &= \int_0^\infty \frac{1}{x_1}f_{X_1}(x_1)\,dx = \int_0^\infty \frac{1}{x_1}\frac{1}{4}x_1e^{-\frac{x_1}{2}}\,dx \\ &= \frac{1}{4}\int_0^\infty e^{-\frac{x_1}{2}}\,dx = \frac{1}{4}2 = \frac{1}{2}. \end{aligned}$$

It follows that $E(X_2/X_1) = E(X_2)E(1/X_1) = 2(1/2) = 1$.

Back to Lecture 5 again. There we mentioned that for two random variables X and Y, $V(X+Y) \neq V(X)+V(Y)$ usually, but that the result is true if X and Y are independent. Now, we can easily see that if X and Y are independent and we take the functions $g_1 = g_2 = g$ with $g(X) = X - E(X)$, by our previous results we have

$$\begin{aligned} V(X+Y) &= E[(X+Y-(E(X)+E(Y)))^2] = E[((X-E(X))+(Y-E(Y)))^2] \\ &= E[(X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y))] \\ &= E[(X-E(X))^2] + E[(Y-E(Y))^2] + 2E(X-E(X))E(Y-E(Y)) \\ &= V(X) + V(Y) + 2*0*0 = V(X) + V(Y). \end{aligned} \tag{43}$$

Notice that we exploited the independence of X and Y together with the results above to deal with the expectation of the cross-product $2(X-E(X))(Y-E(Y))$ with $g_1 = g_2 = g$.

Question: what about $V(X-Y)$?

Covariance

The *covariance* of a pair of random variables X and Y is a measure of their linear dependence. By definition, the covariance between X and Y is

$$Cov(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y). \tag{44}$$

It is easy to check that the covariance operator satisfies, for $a, b, c, d \in \mathbb{R}$

$$Cov(a+bX, c+dY) = bdCov(X,Y).$$

Also, notice that $Cov(X,Y) = Cov(Y,X)$ and $Cov(X,X) = V(X)$.

Question: suppose that X and Y are independent. What is $Cov(X,Y)$?

There exists a scaled version of the covariance which takes values in $[-1,1]$. This is the *correlation* between X and Y. The correlation between X and Y is defined as

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}. \tag{45}$$

It is easy to check that for $a, b, c, d \in \mathbb{R}$ with $b, d \neq 0$, we have $Cor(a+bX, c+dY) = Cor(X,Y)$. So, unlike the covariance operator, the correlation operator is not affected by affine transformations of X and Y.

Beware that while we saw that X and Y are independent $\implies Cov(X,Y) = 0$, the converse is not true (i.e. independence is stronger than *uncorrelation*). Consider the following example. Let X be a random variable with $E(X) = E(X^3) = 0$. Consider $Y = X^2$. Clearly, X and Y are not independent (in fact, Y is a deterministic function of X!). However,

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X * X^2) - 0 = E(X^3) = 0.$$

So X and Y are uncorrelated, because $Cor(X,Y) = 0$, but they are not independent!

Exercise in class:
Look back at equation (43). Without assuming that X and Y are independent, but rather only making the weaker assumption that X and Y are *uncorrelated* (i.e. $Cov(X,Y) = 0$), show that

$$V(X+Y) = V(X-Y) = V(X) + V(Y).$$

From the discussion above it is clear that in general

$$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$$
$$V(X-Y) = V(X) + V(Y) - 2Cov(X,Y).$$

This generalizes as follows. Given $a_1, \ldots, a_n \in \mathbb{R}$ and $X_1, \ldots, X_n$,

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 V(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_i a_j Cov(X_i, X_j)$$
$$= \sum_{i=1}^{n} a_i^2 V(X_i) + \sum_{1 \leq i < j \leq n} 2 a_i a_j Cov(X_i, X_j).$$

If the random variables $X_1, \ldots, X_n$ are all pairwise uncorrelated, then obviously the second summand above is null.

Exercise in class:
You are given three random variables X, Y, Z with $V(X) = 1$, $V(Y) = 4$, $V(Z) = 3$, $Cov(X,Y) = 0$, $Cov(X,Z) = 1$, and $Cov(Y,Z) = 1$. Compute $V(3X + Y - 2Z)$.

We have

$$\begin{aligned} V(3X+Y-2Z) &= V(3X) + V(Y) + V(-2Z) + 2Cov(3X,Y) + 2Cov(3X,-2Z) + 2Cov(Y,-2Z) \\ &= 9V(X) + V(Y) + 4V(Z) + 6Cov(X,Y) - 12Cov(X,Z) - 4Cov(Y,Z) \\ &= 9 + 4 + 12 - 12 - 4 = 9. \end{aligned}$$

Exercise in class:
Recall the multinomial distribution from the lecture about discrete distributions. Suppose that $(X_1, \ldots, X_k)$ has a Multinomial$(p_1 \ldots, p_k)$ distribution. We want to prove:

$$\text{Cov}[X_i, X_j] = -np_ip_j.$$

The trick is to treat a multinomial experiment as a sequence of n independent trials, Y_t. $1_{Y_t=i}$ is the random variable which takes the value of 1 if the t'th outcome Y_t took value i, and notice that $X_i = \sum_{t=1}^n 1_{Y_t=i}$ and $X_j = \sum_{t=1}^n 1_{Y_t=j}$.

Then, proceed as follows: Now, if $s = t$,

$$\text{Cov}[1_{Y_t=i}, 1_{Y_s=j}] = -\mathbb{E}[1_{Y_t=i}]\mathbb{E}[1_{Y_t=j}] = -p_ip_j,$$

while if $s \neq t$

$$\text{Cov}[1_{Y_t=i}, 1_{Y_s=j}] = 0$$

by independence.

$$\text{Cov}[X_i, X_j] = -np_ip_j = \sum_{t=1}^n \sum_{s=1}^n \text{Cov}[1_{Y_t=i}, 1_{Y_s=j}].$$

Hence the result.

Remark: Why is $\rho(X, Y) := Cor(X, Y) \in [-1, 1]$? In order to answer this question, we will need *Cauchy-Schwarz* (CS) inequality. An application of this inequality gives us the following bound:

$$|E[ZQ]| \leq (E[Z^2])^2(E[Q^2])^2$$

In particular, notice the absolute value. The more general version of this inequality is known as *Holder inequality*. CS inequality, although simple, is one of the most powerful tools used in Statistics. Now, if we take $Z := X - E[X]$ and $Q := Y - E[Y]$, we have

$$|cov(X, Y)| \leq \sqrt{Var(X)}\sqrt{Var(Y)}.$$

It follows that

$$-\sqrt{Var(X)}\sqrt{Var(Y)} \leq cov(X, Y) \leq \sqrt{Var(X)}\sqrt{Var(Y)}$$

hence

$$Cor(X, Y) \in [-1, 1].$$

Lecture 10

Recommended readings: WMS, section 5.11

Expected Value in the Context of Bivariate Distributions (part 2)

Using conditional distributions, we can define the notion of *conditional expectation*, *conditional variance*, and *conditional covariance*.

Introduction Conditional expectation can first be thought of as simply the expectation of a (discrete for now) random variable X given an event A:

$$E(X|A) = \sum_{x \in \text{supp}(X)} x p_{X|A}(x|A)$$

Conditional expectation $E(X|A)$ has all the properties of regular expectation. In particular:

1. $E(f(X)|A] = \sum_{x \in \text{supp}(X)} f(x) p_{X|A}(x|A)$
2. $E[aX + b|A] = aE[X|A] + b \ \forall a, b$

Set $A = \{Y = y\}$:

$$\Rightarrow E[X|Y = y] = \sum_{x \in \text{supp}(X)} x p_{X|Y}(x|y)$$

Where $E[X|Y = y]$ is the *conditional expectation* of X given $Y = y$. Now, Write $E[X|Y]$ where X and Y are r.v.'s *as the random variable*, which is a function of Y, whose value when $Y = y$ is $E[X|Y = y]$. So think of $E[X|Y] = g(Y)$ for some function g.

From here on, the distinction between how we refer to the two may be blurred i.e. we will call both $E(X|A)$ and $E(X|Y)$ conditional expectations, but it should be clear which it is referring to, from context.

Conditional Expectation

Given a conditional p.m.f. $p_{X_1|X_2=x_2}$ or a conditional p.d.f. $p_{X_1|X_2=x_2}$ the conditional expectation of X_1 given $X_2 = x_2$ is defined as

$$E(X_1|X_2 = x_2) = \sum_{x_1 \in \text{supp}(X_1)} x_1 p_{X_1|X_2=x_2}(x_1) \tag{46}$$

in the discrete case and as

$$E(X_1|X_2 = x_2) = \int_{x_1 \in \text{supp}(X_1)} x_1 f_{X_1|X_2=x_2}(x_1)\, dx_1 \tag{47}$$

in the continuous case.

It is clear from the definition that the conditional expectation is a function of the value of X_2. Since X_2 is a random variable, this means that the conditional expectation (despite its name) is itself a random variable! This is a fundamental difference with respect to the standard concept of expectation for a random variable (for instance $E(X_1)$) which is just a constant depending on the distribution of X_1. In terms of notation, we often denote the random variable corresponding to the conditional expectation of X_1 given X_2 simply as $E(X_1|X_2)$.

Note that if X_1 and X_2 are independent, then $E(X_1|X_2) = E(X_1)$.

Conditional expectation has the so-called 'tower property', or law of total expectation, which says that

$$E(E(X_1|X_2)) = E(X_1)$$

where the outer expectation is taken with respect to the probability distribution of X_2. This is easier to understand in the context of discrete random variables.

Exercise: Prove this, for discrete and continuous case. Hint: carefully write down the expression for $E(E(X_1|X_2))$, where $E(X_1|X_2)$ can be treated as a random variable and a function of X_2. You may even write it as $g(X_2)$ if it makes it clear. Then, work on the double sum (discrete) or double integral (continuous).

Conditional Variance

We can also define the conditional variance of X_1 given X_2. We have

$$V(X_1|X_2) = E[(X_1 - E(X_1|X_2))^2|X_2] = E(X_1^2|X_2) - [E(X_1|X_2)]^2 \tag{48}$$

Obtaining the unconditional variance from the conditional variance is a little 'harder' than obtaining the unconditional expectation from the conditional expectation:

$$V(X_1) = E[V(X_1|X_2)] + V[E(X_1|X_2)].$$

Note that if X_1 and X_2 are independent, then $V(X_1|X_2) = V(X_1)$.

Conditional Covariance

It is worthwhile mentioning that we can also define the conditional covariance between X_1 and X_2 given a third random variable X_3. We have

$$\begin{aligned} Cov(X_1, X_2|X_3) &= E[(X_1 - E(X_1|X_3))(X_2 - E(X_2|X_3))|X_3] \\ &= E(X_1X_2|X_3) - E(X_1|X_3)E(X_2|X_3). \end{aligned} \tag{49}$$

To get the unconditional covariance we have the following formula:

$$Cov(X_1, X_2) = E[Cov(X_1, X_2|X_3)] + Cov[E(X_1|X_3), E(X_2|X_3)]. \tag{50}$$

Note that if X_1 and X_3 are independent, and X_2 and X_3 are independent, then $Cov(X_1, X_2|X_3) = Cov(X_1, X_2)$.

In terms of computation, everything is essentially unchanged. The only difference is that we sum or integrate against the conditional p.m.f./conditional p.d.f. rather than the marginal p.m.f./marginal p.d.f..

Exercise in class:
Consider again the p.d.f. of exercise 7 in Homework 5:

$$f_{X_1,X_2}(x_1, x_2) = 6(1 - x_2)\mathbb{1}_{\{0 \le x_1 \le x_2 \le 1\}}(x_1, x_2).$$

What is $E(X_1|X_2)$? Compute $E(X_1)$.

We first need to compute $f_{X_1|X_2=x_2}$ for $x_2 \in \text{supp}(X_2)$. For $x_2 \in [0, 1]$ we have

$$f_{X_2}(x_2) = \int_0^{x_2} 6(1 - x_2)\mathbb{1}_{[0,1]}(x_1)\, dx_1 = 6x_2(1 - x_1)\mathbb{1}_{[0,1]}(x_1)$$

Notice that from the marginal p.d.f. of X_2 we can see that $X_2 \sim \text{Beta}(2, 2)$. For $x_2 \in (0, 1)$ we have

$$\begin{aligned} f_{X_1|X_2=x_2}(x_1) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{6(1 - x_2)\mathbb{1}_{[0,x_2]}(x_1)}{6x_2(1 - x_2)} \\ &= \frac{1}{x_2}\mathbb{1}_{[0,x_2]}(x_1) \end{aligned}$$

Notice that $f_{X_1|X_2=x_2}$ is the p.d.f. of a Uniform$(0, x_2)$ distribution. It follows that $E(X_1|X_2 = x_2) = x_2/2$. We can write this more concisely (and in a way that stresses more the fact that the conditional expectation is a random variable!) as $E(X_1|X_2) = X_2/2$.

We have $E(X_1) = E[E(X_1|X_2)] = E(X_2/2) = E(X_2)/2 = [(2/(2+2)]/2 = 1/4$.

Exercise in class:
Let $N \sim \text{Poisson}(\lambda)$ and let $Y_1, \ldots, Y_N \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$ with N independent of each of the Y_i's. Let $T = \sum_{i=1}^{N} Y_i$. Compute $E(T|N)$, $E(T)$, $V(T|N)$, and $V(T)$.

We have that

$$\begin{aligned} E(T|N=n) &= E\left(\sum_{i=1}^{N} Y_i | N=n\right) = \sum_{i=1}^{n} E(Y_i|N=n) \\ &= \sum_{i=1}^{n} E(Y_i) = n\alpha\beta. \end{aligned}$$

Thus, $E(T|N) = N\alpha\beta$. Now,

$$E(T) = E[E(T|N)] = E(N\alpha\beta) = \alpha\beta E(N) = \alpha\beta\lambda.$$

The conditional variance is

$$\begin{aligned} V(T|N=n) &= V\left(\sum_{i=1}^{N} Y_i | N=n\right) = V\left(\sum_{i=1}^{n} Y_i | N=n\right) \\ &= \sum_{i=1}^{n} V(Y_i|N=n) = \sum_{i=1}^{n} V(Y_i) = n\alpha\beta^2. \end{aligned}$$

Thus, $V(T|N) = N\alpha\beta^2$. The unconditional variance of T is then

$$\begin{aligned} V(T) &= V[E(T|N)] + E[V(T|N)] = V(N\alpha\beta) + E(N\alpha\beta^2) \\ &= \alpha^2\beta^2 V(N) + \alpha\beta^2 E(N) \\ &= \alpha^2\beta^2\lambda + \alpha\beta^2\lambda \\ &= \alpha\beta^2\lambda(1+\alpha). \end{aligned}$$

Exercise in class:
Let $Q \sim \text{Uniform}(0,1)$ and $Y|Q \sim \text{Binomial}(n, Q)$. Compute $E(Y)$ and $V(Y)$.

We have

$$E(Y) = E[E(Y|Q)] = E(nQ) = nE(Q) = n/2$$

and

$$\begin{aligned}
V(Y) &= V[E(Y|Q)] + E[V(Y|Q)] = V(nQ) + E(nQ(1-Q)) \\
&= n^2 V(Q) + nE(Q - Q^2) = \frac{n^2}{12} + nE(Q) - nE(Q^2) \\
&= \frac{n^2}{12} + \frac{n}{2} - n(V(Q) + [E(Q)]^2) \\
&= \frac{n^2}{12} + \frac{n}{2} - \frac{n}{12} - \frac{n}{4} \\
&= \frac{n^2}{12} + \frac{n}{6} = \frac{n}{6}(n/2 + 1).
\end{aligned}$$

Lecture 11

Recommended readings: WMS, sections 6.1 → 6.4

Functions of Random Variables and Their Distributions

Suppose that you are given a random variable $X \sim F_X$ and a function $g : \mathbb{R} \to \mathbb{R}$. Consider the new random variable $Y = g(X)$. How can we find the distribution of Y? This is the focus of this lecture.

There are two approaches that one can typically use to find the distribution of a function of a random variable $Y = g(X)$: the first approach is based on the c.d.f., the other approach is based on the change of variable technique. The former is general and works for any function g, the latter requires some assumptions on g, but it is generally faster than the general approach based on the c.d.f..

Remark: recall that thanks to the *law of the unconscious statistician* (LOTUS), you already know how to compute quantities such as $E[g(X)]$, etc... Here we only focus on computing the *distribution* of $g(X)$. The computation of quantities such as $E[g(X)]$ is typically easier with

The method of the cumulative distribution function

The idea is pretty simple. You know that $X \sim F_X$ and you want to find F_Y. The process is as follows:

1. $F_Y(y) = P(Y \leq y)$ by definition
2. $P(Y \leq y) = P(g(X) \leq y)$, since $Y = g(X)$
3. now you want to express the event $\{g(X) \leq y\}$ in terms of X, since you know the distribution of X only
4. let $A = \{x \in \mathbb{R} : g(x) \leq y\}$; then $\{g(X) \leq y\} = \{X \in A\}$
5. it follows that $P(g(X) \leq y) = P(X \in A)$ which can typically be expressed in terms of F_X
6. (EXTRA) once you have F_X, the p.m.f. or the p.d.f. of X can be easily derived from it.

Exercise in class:
Let $Z \sim \mathcal{N}(0, 1)$ and $g(x) = x^2$. Consider $Y = g(Z) = Z^2$. What is the probability distribution of Y?

Following the steps above, we have

$$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = \begin{cases} 0 \text{ if } y < 0 \\ P(|Z| \leq \sqrt{y}) \text{ if } y \geq 0. \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ P(Z \in [-\sqrt{y}, \sqrt{y}]) \text{ if } y \geq 0. \end{cases}$$

Notice that in this case $A = [-\sqrt{y}, \sqrt{y}]$. It follows that

$$F_Y(y) = \begin{cases} 0 \text{ if } y < 0 \\ P(Z \in [-\sqrt{y}, \sqrt{y}]) \text{ if } y \geq 0. \end{cases} = \begin{cases} 0 \text{ if } y < 0 \\ \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \text{ if } y \geq 0. \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ 2\Phi(\sqrt{y}) - 1 \text{ if } y \geq 0. \end{cases}$$

Let ϕ denote the standard normal p.d.f.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}.$$

The p.d.f. of $Y = Z^2$ is therefore

$$f_Y(y) = \begin{cases} 0 \text{ if } y \leq 0 \\ \frac{1}{\sqrt{y}} \phi(\sqrt{y}) \text{ if } y > 0 \end{cases} = \begin{cases} 0 \text{ if } y \leq 0 \\ \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} \text{ if } y > 0 \end{cases}$$

Notice that this is the p.d.f. of a Gamma$(\frac{1}{2}, 2) \equiv \chi^2(1)$ distribution.

Exercise in class: the probability integral transform
Consider a continuous random variable $X \sim F_X$ where F_X, the distribution of X is strictly increasing. Consider the random variable $Y = F_X(X)$. What is the probablity distribution of Y?

The transformation $Y = F_X(X)$ is usually called the *probability integral transform.* To see why, notice that

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

We have

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

In this case, therefore, $A = (-\infty, F_X^{-1}(y)]$. Then,

$$F_Y(y) = \begin{cases} 0 \text{ if } y < 0 \\ P(X \leq F_X^{-1}(y)) \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1 \end{cases} = \begin{cases} 0 \text{ if } y < 0 \\ F_X(F_X^{-1}(y)) \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1 \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ y \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1. \end{cases}$$

The p.d.f. of Y is therefore

$$f_Y(y) = \mathbb{1}_{[0,1]}(y),$$

i.e. $Y \sim \text{Uniform}(0,1)$.

The method of the change of variable

In order to apply the method of the change of variable, the function g must be strictly increasing continuously differentiable and it must also admit an inverse g^{-1} (this is not required by the method based on the c.d.f.). A sufficient condition is that g is strictly monotone (increasing or decreasing).

Suppose that you are given $X \sim F_X$ and you want to compute the probability distribution of $Y = g(X)$. First of all, determine the support of Y. Then, use this formula to obtain f_Y on the support of Y from f_X and g:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dz} g^{-1}(z) \Big|_{z=y} \right|.$$

Exercise in class:
Consider $X \sim \text{Beta}(1,2)$ and $g(x) = 2x - 1$. What is the p.d.f. of $Y = g(X)$?

First of all, notice that (with a little abuse of notation) $\text{supp}(Y) = g(\text{supp}(X))$. Since $\text{supp}(X) = [0,1]$, it follows that $\text{supp}(Y) = [-1,1]$.

We have

$$f_X(x) = 2(1-x)\mathbb{1}_{[0,1]}(x),$$

$$g^{-1}(y) = \frac{y+1}{2},$$

and

$$\frac{d}{dx} g^{-1}(x) = \frac{1}{2}.$$

Thus, for $y \in [-1, 1]$,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dx} g^{-1}(x) \Big|_{x=y} \right|$$
$$= 2\left(1 - \frac{y+1}{2}\right)\left|\frac{1}{2}\right| = 1 - \frac{y+1}{2} = \frac{1-y}{2}.$$

The complete description of the p.d.f. of Y is therefore

$$f_Y(y) = \frac{1-y}{2} \mathbb{1}_{[-1,1]}(y).$$

Other exercises in class:
• $X \sim Exp(\lambda)$ and $Y = 3X + 1$. Find the pdf of Y and $E[Y]$.
• $X \sim \text{Gamma}(\alpha, \beta)$. Prove that $cX \sim \text{Gamma}(\alpha, c\beta)$ where β is the scale parameter.

Inverse transform method

Sometimes it is useful to *simulate* $X_1, \ldots, X_n \overset{iid}{\sim} F_X$. However, this would require the knowledge and developments of methods to simulate from every distribution F_X, which would be unbelievably expensive. Instead, there is a simple way to do it. It is called the *inverse transform method*. We will only see the case of continuous distributions; for the discrete case, the sampling strategy is analogous and one needs to take into account the partitioned space.

The method is fairly simple and it is based on the results about the transformations of random variables that we have already studied. Let $U \sim \text{Uniform}(0, 1)$, and F_X any strictly increasing cumulative distribution function (cdf) admitting inverse. Let X be the transformation of U through F_X^{-1}, that is $X = F_X^{-1}(U)$. We would like to show that the equality $X \sim F_X$. Notice that

$$F_U(F_X(x))P(U \leq F_X(x)) = F_X(x) \qquad \forall x \text{ s.t. } F_X(x) \in [0, 1]s$$

as already shown for the uniform distribution. Now, the event $\{U \leq F_X(x)\}$ occurs if and only if the event $\{F_X^{-1}(U) \leq F_X^{-1}(F_X(x))\}$ occurs. This is thanks to the strict monotonicity condition on F_X. Moreover, the latter event is equivalent to $\{F_X^{-1}(U) \leq x\}$, again due to this condition. Therefore we can finally conclude that

$$P(X \leq x) = P(F_X^{-1}(U) \leq x) = F_U(F_X(x)) = F_X(x).$$

This means that $X \sim F_X$. In other words, now we know how to generate every random variable just knowing (1) its *inverse cdf* and (2) how to sample form the uniform distribution. The theory behind random number generation (RNG) is wide, and here we have only touched the surface.

The requirement for the cdf to be monotonic and have a closed-form inverse is not always satisfied. The monotonicity requirement can be easily relaxed. The proof is the same once we define

$$F^{-1}(u) = \inf\{x : F_X(x) \geq u\} \qquad \forall 0 < u < 1.$$

Regarding the sample problem, recall, for instance, that the cdf of the normal distribution does not have a closed-form form. Typically approximate methods are developed.

Exercises in class:
• Let $F_X(x) = 1 - e^{-\sqrt{x}}$ for $x \in [0, \infty)$. Find the transformation g such that $X = g(U)$ where $X \sim F_X$ and $U \sim \text{Uniform}(0, 1)$.
In other words, we want to find F_X^{-1} since we know that $P(F^{-1}(U) \leq x) = F_X(x)$ thanks to the previous result. Let's find it. Then

$$F_X(x) = 1 - e^{-\sqrt{x}} \iff x = [\log(1 - F_X(x))]^2$$

therefore, taking $F(x) = u$,

$$g(u) = F^{-1}(u) = [\log(1 - u)]^2 .$$

• Let $F_X(x) = 1 - e^{-\lambda x}$ for $x \in [0, \infty)$ and $\lambda > 0$. Find the transformation g such that $X = g(U)$ where $X \sim F_X$ and $U \sim \text{Uniform}(0, 1)$.
Here, similarly as in the example above, we need to have $F(F^{-1}(u)) = u$, therefore

$$1 - e^{-\lambda F^{-1}(u)} = u \iff F^{-1}(u) = -\frac{1}{\lambda} \log(1 - u).$$

therefore $g(U) = F^{-1}(U) = \frac{1}{\lambda} \log(1 - U)$.

Lecture 12

Recommended readings: WMS, sections 8.1-8.3, 9.7, Cosma Shalizi's 36-401 lecture notes

Estimation

What to do when you don't know $\lambda = 5$?

The concepts and framework of probability theory covered so far has proven useful in calculating theoretical ('ideal') long run frequency properties of random events. The problem settings so far have always included a known probability distribution of the random variable of interest – for instance, the number of people entering a coffeeshop was simply given as 5 people an hour, on average. This led us to use a Poisson random variable $X \sim \text{Poisson}(\lambda = 5)$ to calculate an event like $P(5 \leq X \leq 20)$, by summing the probability mass function in the appropriate region. Of course, we don't have perfect information that says the truth of this process is $\lambda = 5$, in practice.

Data as realization of random variables.

Data Instead of having the perfect information that $\lambda = 5$, as a practitioner and probabilistic modeler, you will have *data* about the arrivals in a coffee shop, over several days. If you are modeling gambling probabilities involving unfair coin flips (Binomial(n, p)) or unevenly shaped dice rolls ($X \sim$ Multinomial$(p_1, \cdots, p_6, n)$, then you will have data about the *outcomes* of coin flips or dice rolls. The goal of estimation is to recover the statistical parameters from observed data, by calculating *statistics* from the data (make *inference*). Indeed, Statistics are just functions of the data. Here is an example of the typical setting in which you will be asked to estimate a parameter. You have observed n dice rolls The data will look like this:

Draw	1	2	3	$\cdots$	n
Outcome	0	1	1	$\cdots$	0

We will call the outcome of each draw random variables (in upper case):

$$X_1, \cdots, X_n$$

Each data in the table above is a *realization* of random variables X_i; we can write this as $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ for the particular outcome $(x_1, \cdots, x_n)$.

Statistical Model The statistical model (interchangeably used are: probabilistic model, or data model) is what we assume (impose) about the nature of randomness of the data. In other words, it is what we impose as the *type* of randomness that $X_1, \cdots, X_n$ follow. A sensible model in this example is

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$$

Under this assumption, the number of heads $X = \sum_{i=1}^{n} X_i$ is distributed as $X \sim \text{Binomial}(n, p)$. We will write $\underline{\text{X}} = (X_1, \cdots, X_n)$ and $\underline{\text{x}} = (x_1, \cdots, x_n)$ sometimes.

Estimation Lastly, the *estimation* of the parameter p is done by:

$$\hat{p} = g(X_1, \cdots, X_n) = \frac{1}{n}(\sum_{i=1}^{n} X_i).$$

The $\hat{\cdot}$ notation is used to emphasize the $\hat{p}$ is aimed at learning about the value of p. The estimate $\hat{p}$ is sometimes written $\hat{p}(\underline{\text{X}})$ to emphasize that it is a function of the data! Also, $\hat{p}$ is a good estimate of the parameter p if it is close to it.

How do you find a good function g of the data whose function value provides a good estimate of p? This is a core matter of interest in statistical modeling, and further studies in statistics and machine learning attempt to answer this question using cool mathematical and algorithmic insights. We will get a flavor to one such technique in this class.

Maximum likelihood estimation One way of finding a good g is to think about the 'likelihood' of the data, given my statistical model. If your model parameter is $p = 0.8$, and your data is mostly zero – let us say that 95 percent of them were zero, then it is quite unlikely that they are realizations from the statistical model of $X_1, \cdots, X_n \sim \text{Bernoulli}(0.8)$ – this model (which is completely determined by p) is quite implausible. What about $p = 0.5$? If the data is mostly zero, then, Bernoulli($p = 0.5$) is still not quite plausible, but it is more likely than before that mostly zero draws came from it.

We can continue this guessing game of finding the most plausible estimate of p, or we can take a more principled route. Write the *likelihood*

function of the data, given the model

$$\begin{aligned}
\mathcal{L}(p|\underline{\mathrm{x}}) &= f_{\underline{\mathrm{X}}}(\underline{\mathrm{x}}|p) && (51)\\
&= P(\underline{\mathrm{X}} = \underline{\mathrm{x}}|p) && (52)\\
&= P(X_1 \cdots = x_1, \cdots, X_n = x_n|p) && (53)\\
&= \prod_{i=1}^{n} p_{X_i}(x_i) && (54)\\
&= p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} && (55)
\end{aligned}$$

where $p_{X_i}(\cdot)$ are the probability mass function of $X_i \sim$ Bernoulli(p). For our purposes, the likelihood function is the joint probability of $\underline{\mathrm{x}}$. We would like to find the value p that maximizes the joint probability (likelihood) for the data on hand, $x_1, \cdots, x_n$. The value of p is what is called the *maximum likelihood estimator* (MLE) of the parameter of interest (p):

$$\hat{p}_{\mathrm{MLE}} = \mathrm{argmax}_p \mathcal{L}(p|x)$$

How would you maximize (51)? Since this is a polynomial function in p, we can use differentiation to find the 'zeros' of this function – this is because we recall from calculus that in order to maximize a polynomial, a good strategy is to take the derivative and set it equal to zero (and some other steps such as checking concavity/convexity..).

Exercise in class

- Obtain the maximum likelihood estimator for the parameter p of a Bernoulli random variable (the above setting). (Hint: maximizing a function $g(x)$ is the same as maximizing $\log(g(x))$.)

Bias The *bias* of an estimator is the difference between the expectation of $\hat{p}$ and the actual parameter:

$$\mathrm{bias} = E(\hat{p}(\underline{\mathrm{X}})) - p.$$

One salient quality of a statistical estimator $\hat{p}$ is for the expectation (mean) of $\hat{p}$ to be equal to p! So that *on average*, $\hat{p}$ is not consistently misled in its mean – that the bias is zero. This quality of an estimator is called *unbiasedness*, and such an estimator is called an *unbiased estimator*.

Exercise in class

- In the above Bernoulli data example, what is the bias of the estimator $\hat{p} = \bar{X}$?
- Is $\hat{p} = \bar{X}$ unbiased?

What is an example of a biased estimate? Take a Normal data model, where $X_i \sim \mathcal{N}(0, \sigma^2)$. We would like to estimate the variance (noise level) σ^2. We will show that the MLE of the $\hat{\sigma}^2$.
Derivation of $\hat{\sigma}^2_{\mathbf{MLE}}$ goes here:

$$\frac{\partial}{\partial \sigma^2} \sum_{i=1}^{n} [-\frac{1}{2} \log 2\pi\sigma^2 - \frac{x_i^2}{2\sigma^2}] = \frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{x_i^2}{2\sigma^4} = 0$$

$$\implies \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

Exercise in class

- Obtain the maximum likelihood estimator for the parameter λ of a Poisson random variable. What is $E(\hat{\lambda}_{\text{MLE}} - \lambda)$?
- Obtain the maximum likelihood estimator for the parameter λ of an Exponential random variable. What is $E(\hat{\lambda}_{\text{MLE}} - \lambda)$?
  Remember that in general $E[1/X] \neq 1/E[X]$!!!

(note, the subscript MLE is sometimes written to emphasize that it is a maximum likelihood estimate)

Here is a summarizing table distinguishing parameters and statistics:

	μ	$\hat{\mu}$
Where does it come from?	Underlying truth	Data
What is it?	The goal of estimation	Function of data
What is it a function of?	Population	Sample
What is it called?	Parameter	Statistic
For example, it is called	True mean	Sample mean
Is it random?	Constant value	Random variable
Example	$p = 0.6$	$\hat{p} = g(Y_1, \cdots, Y_n) = \bar{Y}$
Example	$\mu = 3.54$	$\hat{\mu} = g(X_1, \cdots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Invariance Property of MLEs Maximum likelihood estimators have a convenient property of being *invariant* to transformation by a function g. Formally, if $\hat{\theta}_{\text{MLE}}$ is an MLE of θ, and g is some real valued function, then $g(\hat{\theta}_{\text{MLE}})$ is a maximum likelihood estimator for $\tau = g(\theta)$.

Let's start with the case of g being a one-to-one function, that is every value of τ corresponds to at most one value of θ. In this case

$$\mathcal{L}^*(\tau|\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i; g^{-1}(\tau)) = \mathcal{L}(g^{-1}(\tau)|\mathbf{x}).$$

Then

$$\sup_{\tau} \mathcal{L}^*(\tau|\mathbf{x}) = \sup_{\tau} \mathcal{L}(g^{-1}(\tau)|\mathbf{x}) = \sup_{\theta} \mathcal{L}(\theta|\mathbf{x}).$$

If g is not a one-to-one function, there may be multiple value of θ that correspond to the same value of τ. This is a problem since maximizing $\mathcal{L}^*$ does not correspond anymore to maximizing $\mathcal{L}$. However, we can overcome this problem by defining the *induced likelihood function*:

$$\mathcal{L}^*(\tau|\mathbf{x}) = \sup_{\{\theta: g(\theta)=\tau\}} \mathcal{L}(\theta|\mathbf{x}).$$

For instance, we are ruling out the cases of $\hat{\theta}$ being the MLE of $\mathcal{L}$, but both $\hat{\theta}$ and θ' corresponding to $g(\hat{\theta}) = g(\theta') = \tau$. Then we have

$$\begin{aligned}
\mathcal{L}^*(\tau|\mathbf{x}) &= \sup_{\tau} \sup_{\{\theta: g(\theta)=\tau\}} \mathcal{L}(\theta|\mathbf{x}) \\
&= \sup_{\theta} \mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}(\hat{\theta}|\mathbf{x}) \\
&= \sup_{\{\theta: g(\theta)=g(\hat{\theta})\}} \mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}^*(g(\hat{\theta})|\mathbf{x})
\end{aligned}$$

and the proof is now complete.

Lecture 13

Recommended readings: WMS, sections 7.1 → 7.4

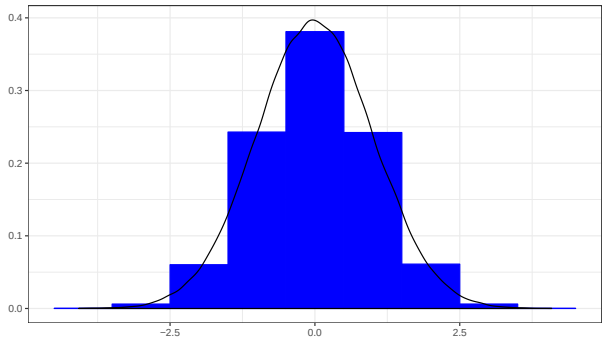
The Empirical Rule, Sampling Distributions, the Central Limit Theorem, and the Delta Method

In this lecture we will focus on some basic statistical concepts at the basis of statistical inference. Before starting, let's quickly discuss a useful rule of thumb associated to the Normal distribution which is often mentioned and used in practice.

The Empirical Rule Based on the Normal Distribution

Suppose that you collect data $X_1, \ldots, X_n \sim f$ where f is an unknown probability density function which, however, you know is 'bell-shaped' (or you expect to be 'bell-shaped' given the information you have on the particular phenomenon you are observing, or you can show to 'bell-shaped' using, for example, an histogram).

Example of histograms and density estimation goes here:



We know that we can estimate the mean μ_f and the standard deviation σ_f of f by means of the sample mean and the sample standard deviation $\bar{X}$ and S, respectively. On the basis of these two statistics, you may be interested in approximately quantifying the probability content of intervals of the form $[\mu_f - k\sigma_f, \mu_f + k\sigma_f]$ where k is a positive integer. Because for

$X \sim \mathcal{N}(\mu, \sigma^2)$ one has

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 68\% \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95\% \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99\%, \end{aligned}$$

one would expect that, based on $\bar{X}$ and S,

$$\begin{aligned} P([\bar{X} - S, \bar{X} + S]) &\approx P([\mu_f - \sigma_f, \mu_f + \sigma_f]) \approx 68\% \\ P([\bar{X} - 2S, \bar{X} + 2S]) &\approx P([\mu_f - 2\sigma_f, \mu_f + 2\sigma_f]) \approx 95\% \\ P([\bar{X} - 3S, \bar{X} + 3S]) &\approx P([\mu_f - 3\sigma_f, \mu_f + 3\sigma_f]) \approx 99\% \end{aligned} \tag{56}$$

if the probability density f is 'bell-shaped'. The approximations of equation (56) are frequently referred to as the *empirical rules* based on the Normal distribution.

Sampling Distributions

To illustrate the concept of sampling distribution, we will consider the Normal model, which assumes that the data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with a Normal distribution with parameters μ and σ^2 that are usually assumed to be unknown. In order to estimate μ and σ^2, we saw that we can use the two statistics $\bar{X}$ and S^2 (the sample mean and the sample variance). These statistics, or *estimators*, are random variables too, since they depend on the data $X_1, \ldots, X_n$. If they are random variables, then they must have their own probability distributions! The probability distribution of a statistic (or an appropriate stabilizing transformation of it) is called the *sampling distribution* of that statistic. We have the following results:

1. $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0, 1)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$.

Some comments: 1. is easy to understand and prove (it's just standardization of a Normal random variable!). 1. is useful when we are interested in making inferences about μ and we know σ^2. 2. is a little harder to prove. We use this result when we are interested in making inferences about σ^2 and μ is unknown. 3. is a classical result. It is useful when we want to make inferences on μ and σ^2 is unknown. We won't study in detail the t distribution. However, this distribution arises when we take the ratio of a standard

Normal distribution and the square root of a χ^2 distribution divided by its degrees of freedom (under the assumption that the Normal and the χ^2 distributed random variables are also independent). The t distribution looks like a Normal distribution with 'fatter' tails. As the number of degrees of freedom of the t distribution goes to ∞, the t distribution 'converges in distribution' to a standard Normal distribution.

For the purposes of this course, we say that a sequence of random variables $X_1, X_2, \ldots X_n, \ldots$ *converges in distribution* to a certain distribution F if their c.d.f.'s $F_1, F_2, \ldots, F_n, \ldots$ are such that

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for each $x \in \mathbb{R}$ at which F is continuous. We will use the notation $\xrightarrow{\text{d}}$ to denote convergence in distribution. The idea here is that the limiting c.d.f. F (often called the *limiting distribution* of the X's) can be used to approximate probability statements about the random variables in the sequence. This idea will be key to the next result, the Central Limit Theorem.

The Central Limit Theorem

The Central Limit Theorem is a remarkable result. Simply put, suppose that you have a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots \overset{\text{iid}}{\sim} F$ distributed according to some c.d.f. F, such that their expectation μ and their variance σ^2 exist and are finite. Then, the standardized version of their average converges in distribution to a standard Normal distribution.

In other words,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} Z \tag{57}$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0,1)$. Another way to express the Central Limit Theorem is to say that, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x). \tag{58}$$

This result is extremely frequently used and invoked to approximate probability statements regarding the average of a collection of i.i.d. random variables when n is 'large'. However, rarely the population variance σ^2 is known. The Central Limit Theorem can be extended to accomodate this case. We have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \xrightarrow{\text{d}} Z \tag{59}$$

where $Z \sim \mathcal{N}(0,1)$. Its typical proof involves characteristic functions, that we will study in a few classes.

Exercise in class:
The number of errors per computer program has a Poisson distribution with mean $\lambda = 5$. We receive $n = 125$ programs written by $n = 125$ different programmers. Let $X_1, \ldots, X_n$ denote the number of errors in each of the n programs. Then $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. We want to approximate the probability that the average number of errors per program is not larger than 5.5.

We have $\mu = E(X_1) = \lambda$ and $\sigma = \sqrt{V(X_1)} = \sqrt{\lambda}$. Furthermore,

$$\begin{aligned} P(\bar{X}_n \leq 5.5) &= P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq \frac{\sqrt{n}(5.5-\mu)}{\sigma}\right) \\ &\approx P\left(Z \leq \frac{\sqrt{125}(5.5-5)}{\sqrt{5}}\right) = P\left(Z \leq 2.5\right) \\ &= \Phi(2.5) \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$.

The Delta Method

If we have a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ which converges in distribution to a standard Normal and a differentiable function $g : \mathbb{R} \to \mathbb{R}$, then the Delta Method allows us to find the limiting distribution for the sequence of random variables $g(X_1), g(X_2), \ldots, g(X_n), \ldots$. Assume that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

and that $g : \mathbb{R} \to \mathbb{R}$ is differentiable with $g'(\mu) \neq 0$. Then,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Exercise in class:
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ where F is some c.d.f. such that both the expectation μ and the variance σ^2 of the X's exist and are finite. By the Central Limit Theorem,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Consider the function $g(x) = e^x$. Find the limiting distribution of $g(\bar{X}_n)$.

We have $g'(x) = g(x) = e^x > 0$ for any $x \in \mathbb{R}$; thus, $g(\mu) \neq 0$ necessarily. By applying the Delta Method, we have that

$$\frac{\sqrt{n}[g(\bar{X}_n) - g(\mu)]}{|g'(\mu)|\sigma} = \frac{\sqrt{n}\left(e^{\bar{X}_n} - e^{\mu}\right)}{e^{\mu}\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Thus, we can use the distribution

$$\mathcal{N}\left(e^{\mu}, \frac{e^{2\mu}\sigma^2}{n}\right).$$

to approximate probability statements about $g(\bar{X}_n) = e^{\bar{X}_n}$ when n is 'large'.

The Law of Large Numbers

We have so far studied the Central Limit Theorem and a particular mode of convergence, the convergence in distribution. Another important result on the convergence of the average of a sequence of random variables is the *law of large numbers*[7]. Simply put, the law of large numbers says that the sample mean of a sequence of i.i.d. random variables converges to the common expectation of the random variables in the sequence. More precisely, let $X_1, \ldots, X_n, \ldots$ be a sequence of iid random variables with common mean μ. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0. \tag{60}$$

In this case, we use the notation $\bar{X}_n \xrightarrow{\text{p}} \mu$.

Example in class:
Suppose that you repeatedly flip a coin which has probability $p \in (0,1)$ of showing head. Introduce the random variables

$$X_i = \begin{cases} 1, & \text{if the i-th flip is head} \\ 0, & \text{if the i-th flip is tails.} \end{cases}$$

Consider the random variable $\bar{X}_n$ describing the observed proportion of heads. Then, as $n \to \infty$, $\bar{X}_n \xrightarrow{\text{p}} \mu$. In English, this means that for any

[7] This is also called the (weak) law of large numbers; the stronger version exists, but we do not discuss this in this course.

arbitrarily small $\epsilon > 0$, the probability of the event 'the proportion of heads differs from p by more than ϵ' converges to 0 as $n \to \infty$. Let's prove it. By the central limit theorem, we know that

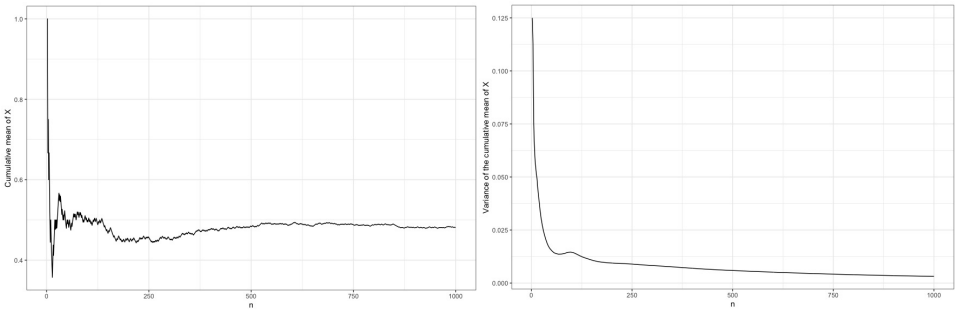
$$\bar{X} \sim \mathcal{N}(\mu, \mu(1-\mu)/n)$$

therefore, for any $\epsilon > 0$,

$$\begin{aligned} P(|\bar{X}_n - \mu| > \epsilon) &= 1 - P(-\epsilon \leq \bar{X}_n - \mu \leq \epsilon) \\ &= 1 - P\left(-\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}} \leq Z \leq \frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) \\ &= 1 - \phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) + \phi\left(-\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) \overset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

Summary: two ways in which Random Variables converge.

We introduced two notions of convergence for random variables: *convergence in distribution* (which we associated to the Central Limit Theorem) and *convergence in probability* (which we associated to the weak law of large numbers). There are other ways of convergence of random variables, all of which have the common goal: What does the distribution of $Y_n = f(X_1, \cdots, X_n)$ look like as we collect more and more samples? $(n \to \infty)$? These results allow us to understand the *precision* or *accuracy* in which our statistics (estimators, necessarily functions of the data) estimate some unknown quantity (usually a parameter). For further study, read a good probability textbook or consider taking a graduate class!



(a) Convergence of $\bar{X}_n$. (b) Convergence of $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Figure 2: Convergence of cumulative sample mean and sample variance of Bernoulli trials as a function of n.

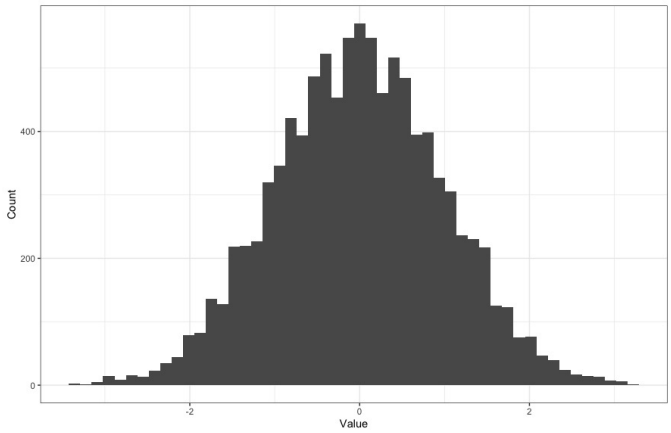


Figure 3: Histogram (counts) of the values taken by $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ for Bernoulli trials with $\mu = 0.5$, and consequently variance $\sigma^2 = \mu(1-\mu)$. Each value is computed with 10^4 Bernoulli trials. As you see, $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0, 1)$.

Lecture 14

Recommended readings: WMS, sections 3.9, 4.9, 6.5

Moments

Remember how we learned that a CDF or a PDF/PMF *completely* characterizes the distribution of a random variable – in other words, it contains all you need to know about the law of randomness of that random variable. If two random variables (from different origins, or sampling procedure) give the same CDF, then they have the same distribution.

What is a *moment* of F? In mathematics/statistics/mechanics, a moment is a specific quantitative measure of the shape of a set of points. (For simplicity, think of such points as potential realizations random variables.) Would you agree that, if you have a set of points, and you knew (1) the center of balance (2) how spread out they are, then you already have a pretty good rough idea of the distribution of variables? (1) corresponds to the first moment (and the mean). The spread (variance) is the second moment minus the square of the first moment. The story goes on – what if you knew which side of the mean the points are *more* populated/concentrated in? This is given by the third moment, plus some multiples of the first and second moment.

Would you agree that, as you go further and learn more such information about the shape of the points, that you learn more about the exact *distribution* of the random variables? Indeed, knowing all the moments is equivalent to knowing the CDF (from which we know all we need to know about the distribution)!

In this lecture, we introduce a particular function associated to a probability distribution F which *uniquely* characterizes F. This function is called the *moment-generating function* (henceforth shortened to m.g.f.) of F because, if it exists, it allows to easily compute any *moment* of F, i.e. any expectation of the type $E(X^k)$ with $k \in \{0, 1, \dots\}$ for $X \sim F$.

Moment generating functions

There are other functions that are similar to the m.g.f.s which we will not discuss in this course: these include the *probability-generating function* for discrete probability distributions (which is a compact power series representation of the p.m.f.), the *characteristic function* (which is the inverse Fourier transform of a p.m.f./p.d.f. and always exists) and the *cumulant-generating*

function (which is the logarithm of the m.g.f.).

The moments $\{E(X^k)\}_{k=1}^{\infty}$ associated to a distribution $X \sim F$ completely characterize F (if they exist). They can all be encapsulated in the m.g.f.

$$m_X(t) = E(e^{tX}) = \begin{cases} \int_{\text{supp}(X)} e^{tx} f_X(x)\, dx \text{ if } X \text{ is continuous} \\ \sum_{\text{supp}(X)} e^{tx} p_X(x) \text{ if } X \text{ is discrete.} \end{cases}$$

If two random variables X and Y are such that $m_X = m_Y$ then their c.d.f.'s F_X and F_Y are equal at almost all points (i.e. they can differ at at most countably many points). If you are interested in the proof, I suggest you to look it up in more advanced books.

We say that the moment generating function m_X of $X \sim F$ exists if there exists an open neighborhood around $t = 0$ in which $m_X(t)$ is finite. [8]Note that it is always true that, for $X \sim F$ with an arbitrary distribution F, $m_X(0) = 1$.

The name of this function comes from the following feature: suppose that m_X exists, then for any $k \in \{0, 1, \dots\}$

$$\frac{d^k}{dt^k} m_X(t)\bigg|_{t=0} = E(X^k). \tag{61}$$

This means that we can 'generate' the moments of $X \sim F$ from m_X by differentiating m_X and evaluating its derivatives at $t = 0$. This is more clear if we rewrite the m.g.f. in terms of its series expansion: $\forall t \in \mathbb{R}$

$$\begin{aligned} E[e^{tX}] &= E\left[1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots + \frac{t^nX^n}{n!} + \dots\right] \\ &= 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots + \frac{t^nE[X^n]}{n!} + \dots \end{aligned}$$

Exercise in class:
Show that the m.g.f. of $X \sim \text{Binomial}(n, p)$ is

$$m_X(t) = [pe^t + 1 - p]^n$$

and use it to compute $V(X)$. Let's see how.

$$\begin{aligned} m_X(t) &= \sum_{x=0}^{n} e^{tx} p_X(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = [pe^t + (1-p)]^n \end{aligned}$$

[8]For instance, the mgf of a Cauchy distribution does not exists.

where we used the binomial theorem

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

Notice that the function above is well-defined and finite for any $t \in \mathbb{R}$. Then,

$$E(X) = \frac{d}{dt} m_X(t)\Big|_{t=0} = npe^t[pe^t + (1-p)]^{n-1}\big|_{t=0} = np$$

and

$$\begin{aligned} E(X^2) = \frac{d^2}{dt^2} m_X(t)\Big|_{t=0} &= np^2 e^{2t}(n-1)[pe^t + (1-p)]^{n-2} + npe^t[pe^t + (1-p)]^{n-1}\big|_{t=0} \\ &= np^2(n-1) + np. \end{aligned}$$

Thus,

$$V(X) = E(X^2) - [E(X)]^2 = np^2(n-1) + np - n^2p^2 = -np^2 + np = np(1-p).$$

Exercise in class:
Show that the m.g.f. of $X \sim \text{Uniform}(0,1)$ is

$$m_X(t) = \frac{1}{t}(e^t - 1).$$

We have

$$m_X(t) = \int_0^1 e^{tx}\, dx = \frac{1}{t} e^{tx}\Big|_0^1 = \frac{1}{t}(e^t - 1)$$

which is well-defined and finite for any $t \in \mathbb{R}$ (recall that $m_X(0) = 1$ for any X).

It is easy to verify that a m.g.f. m_X satisfies

$$m_{aX+b}(t) = e^{bt} m_X(at). \tag{62}$$

M.g.f.'s also provide a tool which is often useful to identify the distribution of a linear combination of random variables. In the m.g.f. world sums becomes products! In particular, consider a collection of n independent random variables $X_1, \ldots, X_n$ with m.g.f's $m_{X_1}, \ldots, m_{X_n}$. It is easy to check from the definition of m.g.f. that, if we consider $Y = \sum_{i=1}^n (a_i X_i + b_i)$, then

$$m_Y(t) = e^{\sum_{i=1}^n b_i t} \prod_{i=1}^{n} m_{X_i}(a_i t). \tag{63}$$

Exercise in class:

1. What is the mgf of $X \sim \text{Poisson}(\lambda)$?

$$m_X(t) = e^{\lambda(e^t-1)}$$

for $t \in \mathbb{R}$.

2. Consider $Y \sim \text{Poisson}(\mu)$ with X and Y independent. What is the distribution of $X + Y$?

Since X and Y are independent, we have

$$m_{X+Y}(t) = m_X(t)m_Y(t) = e^{\lambda(e^t-1)}e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$$

and we recognize this as the m.g.f. of a Poisson$(\lambda + \mu)$ distribution.

Why are moment generating functions useful?
First, for X and Y whose moment generating functions are finite, $m_X(t) = m_Y(t)$ for all t if and only if $P(X \leq x) = P(Y \leq x)$ for all x.

We will briefly prove this, for discrete distributions X and Y. One direction is trivial; if the two have the same distribution, then of course

$$m_X(t) = E(e^{tX}) = E(e^{tY}) = m_Y(t)$$

is true. The other direction is harder. Call $A = \text{supp}(X) \cup \text{supp}(Y)$, and $a_1, \cdots, a_n$ the elements of A. Then, the mgf of X is

$$\begin{aligned} m_X(t) &= E(e^T X) \\ &= \sum_{x \in \text{supp}(X)} e^{tx} \cdot p_X(x) \\ &= \sum_{i=1\cdots,n} e^{ta_i} \cdot p_X(a_i) \end{aligned}$$

and likewise, $m_Y(t) = \sum_{i=1,\cdots,n} e^{ta_i} \cdot p_X(a_i)$. Subtract the two, to get

$$m_X(t) - m_Y(t) = \sum_{i=\cdots,n} e^{ta_i} \cdot [p_X(a_i) - p_Y(a_i)] = 0$$

must be true when t is close to zero (because it is true for all t). If t is close to zero, then no matter what a_i is, e^{ta_i} must be close to 1. So, it must be the case that

$$p_X(a_i) = p_Y(a_i)$$

for all i; hence, the pmfs are the same, and the distributions are the same.

Second, if $m_{X_n}(t) \to m_X(t)$ for all t, and $P(X \leq x)$ is continuous in x, then $P(X_n \leq x) \to P(X \leq x)$. You can prove the central limit theorem with moment generating functions, assuming the moments are finite. (This proof is in the homework.)

Lecture 15

Recommended readings: Ross, sections 5.3.1 → 5.3.3

Introduction to Random Processes

Consider a collection of random variables

$$\mathbb{X} = \{X_t\}_{t \in \mathcal{T}}$$

defined on a sample space Ω endowed with a probability measure P, where the collection is indexed by a parameter t taking values in the set $\mathcal{T}$. Such a collection is commonly referred to as a *random process* or as a *stochastic process*.

It suffices to think about t as 'time' [9], and $\mathcal{T}$ to be the domain of time (for example, $[0, \infty)$). Recall that a random variable is a function mapping the sample space Ω into a subset $\mathcal{S}$ of the real numbers. The set $\mathcal{S}$ is usually called the *state space* of the random process $\mathbb{X}$. In fact, if at the time $t \in \mathcal{T}$ the random variable X_t takes the value $X_t = x_t \in \mathcal{S}$, we say that the state of the random process $\mathbb{X}$ at time t is x_t.

Based on the features of the sets $\mathcal{T}$ and the sets $\mathcal{S}$ we say that $\mathbb{X}$ is a

- discrete-time random process, if $\mathcal{T}$ is a discrete set [10]
- continuous-time random process, if $\mathcal{T}$ is not a discrete set
- discrete-state random process, if $\mathcal{S}$ is a discrete set
- continuous-state random process, if $\mathcal{S}$ is a not a discrete set.

It is clear from above that $\mathbb{X} = \{X_t\}_{t \in \mathcal{T}}$ is a *function* or *mapping* of both the elements of $\omega \in \Omega$ and of the time parameter t.

$$\begin{aligned} \mathbb{X}: \quad & \Omega \times \mathcal{T} \to \mathcal{S} \\ & (\omega, t) \mapsto X_t(\omega) \end{aligned}$$

We can therefore think of $\mathbb{X}$ in at least two ways:

- as a collection of random variables X_t taking values in $\mathcal{S}$; i.e. for any fixed time t, the random process $\mathbb{X}$ corresponds to a random variable X_t valued in $\mathcal{S}$

[9] t can be more generally thought of as time or space (we call this a 'spatial process', or just any indexing that keeps track of the evolution of the random variable.

[10] i.e. $\mathcal{T}$ is finite or countably infinite

- as the collection of random functions of time ('trajectories'); i.e. for any fixed $\omega \in \Omega$, we can view $\mathbb{X}$ as the *sample path* or 'trajectory' $t \mapsto X_t(\omega)$ which is a (random) function of time.

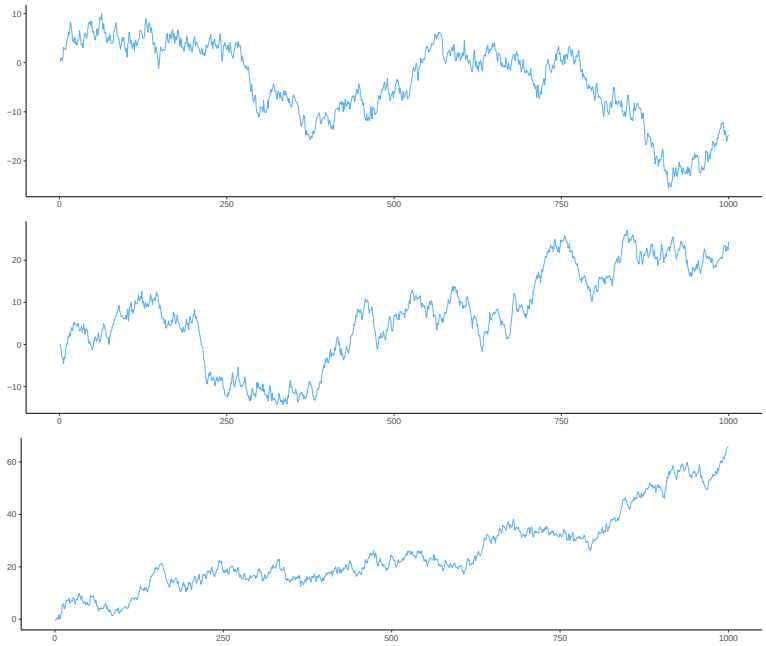


Figure 4: Sample path of X_t for three different ω's, that is for $X_t(\omega_1)$, $X_t(\omega_2)$, $X_t(\omega_3)$.

In contrast to the study of a single random variable or to the study of a collection of independent random variables, the analysis of a random process puts much more focus on the *dependence structure* between the random variables in the process at different times and on the implications of that dependence structure.

I think that by now you might be confused with high probability. For this reason, let's see a few examples of stochastic processes.

- Any single random variable is a stochastic process. For instance, $X \sim$Poisson(5) is a stochastic process.

- Let $\mathbb{X} = \{X_i\}_{i \in T}$ and let $T = \{1, 2\}$. Then we can consider a coin toss, that is the sample space will be $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$.

Call ω_i^t the element in the $t-th$ position of the $i-th$ element of Ω. For instance, $\omega_2^1 = H$. We endow Ω with a probability measure P such that $P(\omega_i) = 1/4\ \forall i\{1,2,3,4\}$. Then define the random variable $X_t(\omega) = \mathbb{1}(\{\omega^t \text{ is H}\})$ if ω_i for $t = 1, 2$. Clearly, for fixed $t \in T$, $X_t(\omega)$ is a Bernoulli random variables. For fixed ω, X_t is a function of t, and there are four ($|\Omega|$) possible trajectories.

Because there are many real-world phenomena that can be interpreted and modeled as random processes, there is a wide variety of random processes. In this class, however, we will focus only on the following processes:

1. *Bernoulli process*: discrete-time and discrete-state,
2. *Poisson process*: continuous-time and discrete-state,
3. *Brownian motion* (the Wiener process): continuous-time and continuous-state,
4. *Discrete-time Markov chains*: discrete-time and any state.

The Bernoulli Process and the Poisson Process

In lecture (and in-class examples), we will define Bernoulli and Poisson processes and their useful properties. Then, in the homework problems, you will learn further aspects and how to apply them to several settings, and calculate/identify useful quantities.

The Bernoulli Process

Let $\mathbb{X} = \{X_i\}_{i=1}^{\infty}$ be a sequence of iid Bernoulli(p) random variables. The random process $\mathbb{X}$ is called a Bernoulli process.

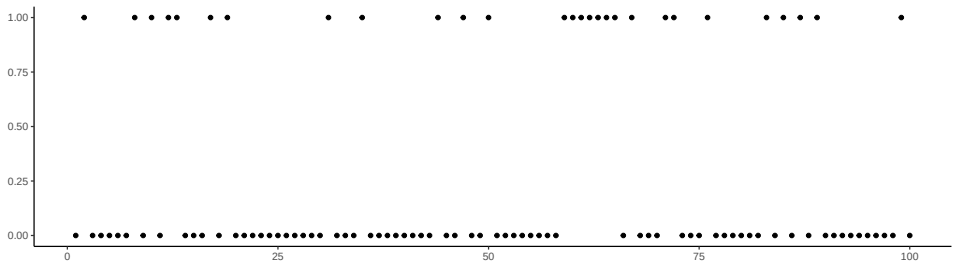


Figure 5: Sample path for a Bernoulli process with $p = 0.3$. Horizontal and vertical axes represent time and value of the process respectively.

The Bernoulli process can be thought of as an *arrival process*: at each time i either there is an arrival (i.e. $X_i = 1$, or equivalently stated, we observe a *success*) or there is not an arrival (i.e. $X_i = 0$, or equivalently stated, we observe a *failure*). Many of the properties of a Bernoulli process are already known to us from previous discussions.

Some interesting questions:

- Consider n distinct times; what is the probability distribution of the number of arrivals in those n times? This is clearly binomially distributed with parameters n and p!
- What is the probability distiribution of the first arrival? This even follows a geometric distribution with parameter p.
- Finally, consider the time needed until the r-th arrival ($r \in \{1, 2, \dots\}$); what is the probability distribution associated to the time of the r-th arrival? Here we need the negative binomial distribution.

The independence of the random variables in the Bernoulli process has an important implication about the process. Consider a Bernoulli process $\mathbb{X} = \{X_i\}_{i=1}^{\infty}$ and the process $\mathbb{X}_{-n} = \{X_i\}_{i=n+1}^{\infty}$. Because the random variables in $\mathbb{X}_{-n}$ are independent of the random variables in $\mathbb{X} \setminus \mathbb{X}_{-n}$, it follows that $\mathbb{X}_{-n}$ is itself a Bernoulli process (starting at time n) which does not depend on the initial n random variables of $\mathbb{X}$. We say that the Bernoulli process thus satisfies the *fresh-start property*.

Recall the memoryless property of the Geometric distribution. How do you interpret this property in light of the fresh-start property of the Bernoulli process?

The Poisson Process

Motivation: Poisson process is one of the most important models used in queueing theory. The arrival process of customers is well modeled by a Poisson process. In teletraffic theory the "customers" may be calls or packets. Poisson process is a viable model when the calls or packets originate from a large population of independent users. In the following it is instructive to think that the Poisson process we consider represents discrete arrivals (of e.g. calls or packets).

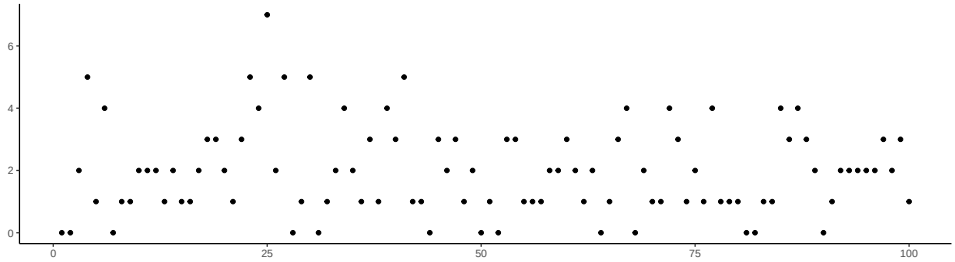


Figure 6: Sample path for a Poisson process with $\lambda = 2$.

A random process $\mathbb{X} = \{X_t\}_{t \in \mathcal{T}}$ is called a *counting process* if X_t represents the number of events that occur by time t. More properly, if $\mathbb{X}$ is a counting process, it must also satisfy the following intuitive properties:

- for any $t \in \mathcal{T}$, $X_t \in \mathbb{Z}^+$;
- for $s \leq t$, $X_t - X_s$ is the number of events occurring in the time interval $(s, t]$. Consequently, $X_s \leq X_t$.

A counting process $\mathbb{X}$ is said to have *independent increments* if the number of events occurring in disjoint time intervals are independent. This means that the random variables $X_t - X_s$ and $X_v - X_u$ are independent whenever $(s, t] \cap (u, v] = \emptyset$.

Furthermore, a counting process $\mathbb{X}$ is said to have *stationary increments* if the distribution of the number of events that occur in a time interval only depends on the length of the time interval. This means that the random variables $X_{t+s} - X_s$ have the same distribution for all $s \in \mathcal{T}$.

For instance, consider a Bernoulli process $\mathbb{X} = \{X_i\}_{i=1}^{\infty}$ and the process $\mathbb{Y} = \{Y_i\}_{i=1}^{\infty}$ with $Y_i = \sum_{j \leq i} X_j$. Then $\mathbb{Y}$ is a discrete-time counting process with independent and stationary increments.

Let's now see three equivalent definitions of random processes.

1. *homogeneous Poisson process.* A counting process $\mathbb{X} = \{X_t\}_{t \geq 0}$ is said to be a homogeneous Poisson process with rate $\lambda > 0$ if
   - $X_0 = 0$
   - $\mathbb{X}$ has independent increments
   - the number of events in any time interval of length $t > 0$ is Poisson distributed with expectation λt, i.e. for any $s, t \geq 0$ we have $X_{t+s} - X_s \sim \text{Poisson}(\lambda t)$.

Consequently, if the parameter λ does not depend on time, then the Poisson process does have stationary increments. In this class we will deal only with homogeneous process, that is processes whose rate does not depend on t, hence we will frequently refer to them simply as Poisson processes.

2. A *pure birth process* (Yule process) is a stochastic process for which in an infinitesimal dt time interval, there is only one arrival, and this happens with λdt, independent of arrivals outside of the interval.

3. A stochastic process whose *inter-arrival* times follow an exponential(λ) distribution.

Notice, in all of these definitions, there is no restriction that t must take values in some discrete set – so this is a *continuous-time* random process.

It is interesting to prove the equivalence of the three definitions. Let's see it.

1 → 2: First, definition 1 says:

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

In order not to burden the notation, let $h := dt$. Then, consider these three outcomes of X_{dt}:

1. $P(X_{dt} = 0) = e^{-\lambda \cdot dt} = e^{-\lambda \cdot dt} = 1 - \lambda \cdot dt + o(dt) = 1 - \lambda h + o(h)$ where the second equality follows from the Maclaurin series[11] of the exponential function.
2. $P(X_{dt} = 1) = \frac{\lambda dt}{1!} e^{-\lambda \cdot dt} = \lambda \cdot dt - \lambda^2 (dt)^2 + \cdots = \lambda h + o(h)$
3. This is a little bit more tricky.

$$\begin{aligned} P(X_{dt} \geq 2) &= e^{-\lambda h} \sum_{x \geq 2} \frac{(\lambda h)^x}{x!} = e^{-\lambda h}(e^{-\lambda h} - 1 - \lambda h) \\ &= 1 - e^{-\lambda h} - \lambda h e^{-\lambda h} \end{aligned}$$

Now, substituting from above,

$$= 1 - 1 + \lambda h + o(h) - \lambda h + o(h) = o(h).$$

In the proof we have used the "little o" notation. In general we say that $a_n = o(b_n)$ where a_n and b_n, where b_n is nonzero, are two sequences if $\lim_{n \to \infty} \frac{a_n}{b_n} = 0$.

[11] The Maclaurin series is just a Taylor expansion centred at 0.

1 → 3: Denote Y = time between two arrivals. Then, notice the clever fact that

$$\{Y > t\} \equiv \{X_t = 0\}$$

Now, take the probability of these two events and equate them:

$$P(\{Y > t\}) = P(\{X_t = 0\}) = e^{-\lambda t}$$

and further manipulate this to get the CDF of Y:

$$P(\{Y \leq t\}) = 1 - P(Y > t) = 1 - e^{-\lambda t}$$

which we recognize as an exponential(λ) CDF.

3 → 2: If $Y \sim exp(\lambda)$, then,

$$\begin{aligned} P(Y \ \leq dt) &= 1 - e^{-\lambda \cdot dt} \\ &= 1 - (1 - \lambda dt + \lambda^2 (dt)^2 - \cdots) \\ &= \lambda dt + o(dt) \end{aligned}$$

The last part is exactly the 2nd definition above.

2 → 1: First, we introduce the *probability generating function* (PGF for short). A PGF is a convenient representation for discrete random variables that take values in $\{0, 1, \cdot, s\}$:

$$G(z) = E[z^X] = \sum_{x=0}^{\infty} p(x) z^x$$

This is useful because it allows a succinct description of the probability distribution of $P(X = i)$ via

$$P(X = i) = \frac{G^{(k)}(0)}{k!}$$

where $G^{(k)}$ means the k'th derivative of G with respect to z.

What is the PGF of a Poisson random variable $X \sim \text{Pois}(\lambda)$?

$$G_X(z) = E(z^X) = \sum_{x=0}^{\infty} z^x \frac{1}{x!} e^{-\lambda} \lambda^x = e^{-\lambda} \sum_{x=0}^{\infty} \frac{1}{x!} (z\lambda)^x = e^{-\lambda(1-z)}$$

Now, let's consider the PGF of the *counter* $N_{0,t} := N(0,t)$ which counts how many occurrences happen between time 0 and t.

$$G_{N_{0,t}}(z) = E(z^{N_{0,t}})$$

$$G_{N_{0,t+dt}}(z) = E(z^{N_{0,t+dt}}) = E(z^{N_{0,t}+N_{t,t+dt}}) = E(z^{N_{0,t}})E(z^{N_{t,t+dt}})$$

The last expression is the same as

$$G_{N_{0,t}}(z) \times [(1-\lambda dt)z^0 + \lambda dt z^1] = G_{N_{0,t}} - \lambda(1-z)G_{N_{0,t}}dt.$$

This is an ordinary differential equation in t. Although you don't know how to solve it, I will tell you that the solution is actually

$$G_{N_{0,t}}(z) = e^{-\lambda t(1-z)}$$

that corresponds to the PGF of a Poisson random variable with parameter λt.

What is definition 2 saying? Now that we've examined it a bit more carefully, we can say that a counting process $\mathbb{X} = \{X_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$ if

1. $X_0 = 0$
2. $\mathbb{X}$ has stationary and independent increments
3. $P(X_h = 1) = \lambda h + o(h)$
4. $P(X_h \geq 2) = o(h)$

Condition 3 in definition 2 says that the probability of observing an arrival in a small amount of time is roughly proportional to h. Condition 4 says that it is very unlikely to observe more than an arrival in a short amount of time.

Although we will not prove the following claim, you might find it useful: a non-negative random variable has the memoryless property if and only if it follows the exponential distribution.

So, we know about the probability distribution of the count of arrivals up to any time in a Poisson process. What about the probability distribution of the inter-arrival times $\{T_i\}_{i=1}^{\infty}$. First, define the n'th arrival time ($n = 0, 1, \cdots$) as:

$$Y_0 = 0, \;\; Y_n = \min\{t : Y_t = n\}$$

Then, we define the *inter-arrival* times as the random variables

$$T_n = Y_n - Y_{n-1}, \;\; n = 1, 2, \cdots$$

As we saw in the proof of definition 3, notice that the event $\{T_1 > t\}$ is equivalent to the event 'there are no arrivals in the time interval $[0, t]$, i.e. $\{X_t = 0\}$. Thus,

$$P(T_1 > t) = P(X_t = 0) = e^{-\lambda t}$$

from which we see that $T_1 \sim$ Exponential$(1/\lambda)$. Now, what is the distribution of T_2? We have

$$\begin{aligned} P(T_2 > t | T_1 = s) &= P(X_{t+s} - X_s = 0 | T_1 = s) = P(X_{t+s} - X_s = 0 | X_s - X_0 = 1) \\ &= P(X_{t+s} - X_s = 0) = e^{-\lambda t} \end{aligned}$$

which implies that $T_2 \sim$ Exponential(λ) as well. By using the same argument, it is easy to see that all the inter-arrival times of a Poisson process are iid Exponentially distributed random variables with parameter λ.

What is the distribution of the *waiting time* until the n-th arrival?

$$P(T_n > t) = 1 - P(T_n \leq t) = 1 - P(X_t \geq n) = 1 - \sum_{x=0}^{n} \frac{(\lambda t)^x e^{-\lambda t}}{x!}.$$

A Poisson process is also a Markov process (i.e. it is a random process that satisfies the (weak) Markov property), as an immediate consequence of the first definition. In other words, it satisfies

$$P(X_t = x_t | X_u = x_u, X_s = x_s) = P(X_t = x_t | X_s = x_s)$$

for $u < s < t$. Why?

$$\begin{aligned} &P(X_t = x_t | X_u = x_u, X_s = x_s) \\ &= P(X_t - X_s = x_t - x_s | X_s - X_u = x_s - x_u, X_u = x_u) \\ &= P(X_t - X_s = x_t - x_s) = P(X_t = x_t | X_s = x_s). \end{aligned}$$

where we used independence of the increments.

The Poisson Process as a Continuous-time Version of the Bernoulli Process

Fix an arrival rate $\lambda > 0$, some time $t > 0$, and divide the time interval $(0, t]$ into n subintervals of length $h = t/n$. Consider now a Bernoulli process $\mathbb{X} = \{X_i\}_{i=1}^n$ defined over these time subintervals, where each X_i is a Bernoulli random variable recording whether there was an arrival in the time interval $((i-1)h, ih]$. Imposing the condition of the pure birth process, we have that the probability p of observing at least one arrival in any of these subintervals is

$$p = \lambda h + o(h) = \lambda \frac{t}{n} + o\left(\frac{1}{n}\right).$$

Thus, the number of subintervals in which we record an arrival has a Binomial(n, p) distribution with p.m.f.

$$p(x) = \binom{n}{x} \left(\lambda \frac{t}{n} + o\left(\frac{1}{n}\right)\right)^x \left(1 - \lambda \frac{t}{n} + o\left(\frac{1}{n}\right)\right)^{n-x} \mathbb{1}_{\{0,1,\ldots,n\}}(x)$$

Now, let $n \to \infty$ or equivalently $h \to 0$ so that the partition of $(0, t]$ becomes finer and finer, and we approach the continuous-time regime. Following the same limit calculations that we did in the Bernoulli approximation to the Poisson, as $n \to \infty$ we have

$$p(x) \to e^{-\lambda t} \frac{(\lambda t)^x}{x!} \mathbb{1}_{\{0,1,\ldots\}}(x),$$

which is the p.m.f. of a Poisson(λt) distribution. Thus, the Poisson process can be thought of as a continuous-time version of the Bernoulli process.

Lecture 16

Recommended readings: Ross, sections 5.3.4, 5.3.5, 5.4.1

Splitting, Merging, Further Properties of the Poisson Process, the Nonhomogeneous Poisson Process, and the Spatial Poisson Process

Splitting a Bernoulli Process

Consider a Bernoulli process of parameter $p \in [0,1]$. Suppose that we keep any arrival with probability $q \in [0,1]$ or otherwise discard it with probability $1-q$. The new process thus obtained is still a Bernoulli process with parameter pq. This is an example of *thinned* Bernoulli process. Similarly, the process obtained by the discarded arrivals is a thinned Bernoulli process as well (with parameter $p(1-q)$).

In other terms, define the Bernoulli process with parameter p $\mathbb{X} = \{X_i\}_{i=1}^{\infty}$ and let $Z_1, \ldots$ be independent Bernoulli r.v.'s with parameter q. We are claiming that $\mathbb{Z}_1 = \{Z_i X_i\}_{i=1}^{\infty}$ and $\mathbb{Z}_0 = \{(1-Z_i)X_i\}_{i=1}^{\infty}$ are two Bernoulli processes as well. This is actually true. Indeed, (1)

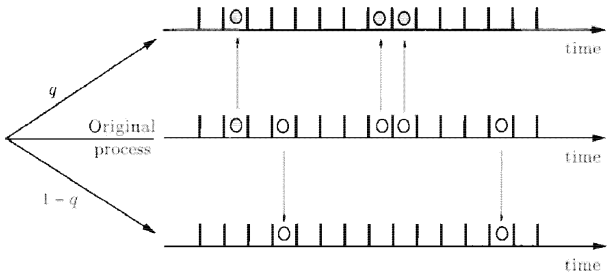
$$Z_i X_i \perp\!\!\!\perp Z_j X_j \; \forall j \neq i$$

and (2) $X_i Z_i$ is clearly a Bernoulli random variable with parameter

$$E[X_i Z_i] = E[X_i]E[Z_i] = pq.$$

Therefore $\mathbb{Z}_1$ is a Bernoulli process. The same can be said for the process $\mathbb{Z}_0$.

This can be easily generalized to the case where we split the original process in more than two subprocesses.



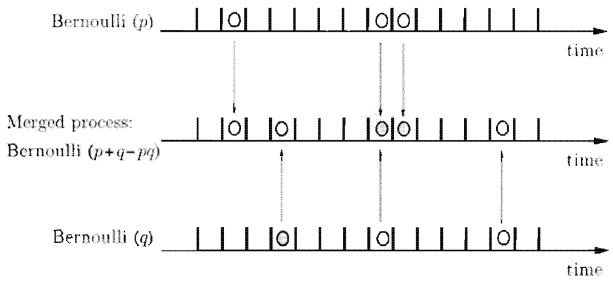
Merging a Bernoulli Process

Consider two independent Bernoulli processes, one with parameter p and the other with parameter q. Consider the new process obtained by

$$X_i = \begin{cases} 1 \text{ if there was an arrival at time } i \text{ in either process} \\ 0 \text{ otherwise.} \end{cases}$$

The process thus obtained is a Bernoulli process with parameter $p + q - pq$.

This is easily generalized to the case where we merge more than two independent Bernoulli processes.



Thinning of a Poisson Process

Consider a Poisson process $\mathbb{X}$ with rate $\lambda > 0$. Suppose that there can be two different types of arrivals in the process, say type A and type B, and that each arrival is classified as an arrival of type A with probability $p \in [0, 1]$ and as an arrival of type B with probability $1 - p$ and that the classification is fone independently from the rest of the process. Let $\mathbb{X}_A$ and $\mathbb{X}_B$ denote the counting processes associated to arrivals of type A and type B respectively. Then $\mathbb{X}_A$ and $\mathbb{X}_B$ are Poisson processes with parameters λp and $\lambda(1 - p)$ respectively. Furthermore, $\mathbb{X}_A$ and $\mathbb{X}_B$ are independent processes. The two processes $\mathbb{X}_A$ and $\mathbb{X}_B$ are examples of thinned Poisson processes.

Why is this true? To gain some intuition, in the homework you will prove that for $X \sim \text{Poisson}(\lambda)$ and $Z_1, \ldots$ independent Bernoulli r.v.'s with parameter p, then $\sum_{i=1}^{X} Z_i \sim \text{Poisson}(\lambda p)$.

This can be easily generalized to the case where we consider $n \geq 3$ different types of arrivals.

Superposition of two Poisson Processes

Consider two Poisson processes $\mathbb{X}_1$ and $\mathbb{X}_2$ with rates $\lambda > 0$ and $\mu > 0$ respectively. Consider the new Poisson process $\mathbb{X} = \mathbb{X}_1 + \mathbb{X}_2$. The *merged* Poisson process $\mathbb{X}$ is a Poisson process with rate $\lambda + \mu$.

Furthermore, any arrival in the process $\mathbb{X}$ has probability $\lambda/(\lambda + \mu)$ of being originated from $\mathbb{X}_1$ and probability $\mu/(\lambda+\mu)$ of being originated from $\mathbb{X}_2$.

This is easily generalized to the case where more than two Poisson processes are merged.

Conditional Distribution of the Arrival Times

Suppose that we know that the number of arrivals up to time $t > 0$ is given by a Poisson process N_t with rate λ. What is the conditional distribution of the n arrival times $T_1, \ldots, T_n$ given that there were exactly n arrivals? Let T_i denote the i-th arrival's time and $A \subset [0, t]$.

$$P(T_n \in A | N_t = n) = \int_{A \times (t,\infty)} \frac{n!}{t^n} \mathbb{1}(0 \leq t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_n) d\mathbf{t}.$$

Let's prove this result. First, recall that

$$P(T_n \in A | N_t = n) = P((T_1, \ldots, T_n) \in A | N_t = n) = \frac{P((T_1, \ldots, T_n) \in A, T_{n+1} > t)}{P(T_{n+1} > t)}$$

The denominator is simply

$$P(T_{n+1} \notin A) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

while for the numerator

$$\begin{aligned} P((T_1, \ldots, T_n) \in A, T_{n+1} > t) &= \int_A \int_{(t,\infty)} f_{T_1,\ldots,T_n,T_{n+1}}(t_1, \ldots, t_{n+1}) dt_{n+1} d\mathbf{t} \\ &= \int_A \int_{(t,\infty)} f_{T_1}(t_1) \prod_{i=2}^{n+1} f_{T_i|T_{i-1}}(t_i - t_{i-2}) dt_{n+1} d\mathbf{t} \end{aligned}$$

where we used the fact that the Poisson process has stationary and independent increments. Then

$$= \int_A \int_{(t,\infty)} \lambda^{n+1} e^{-\lambda t_1} \prod_{i=2}^{n+1} e^{-\lambda t_i - t_{i-1}} \mathbb{1}(0 \le t_1 \le \cdots \le t_{n+1}) dt_{n+1} d\mathbf{t}$$

$$= \int_A \int_{(t,\infty)} \lambda^{n+1} e^{-\lambda t_{n+1}} (0 \le t_1 \le \cdots \le t_{n+1}) dt_{n+1} d\mathbf{t}$$

$$= \int_A \lambda^n e^{-\lambda t} \mathbb{1}(0 \le t_1 \le \cdots \le t_n) d\mathbf{t}$$

hence

$$P((T_1, \ldots, T_n) \in A | N_t = n) = \int_A \frac{n!}{t^n} \mathbb{1}(0 \le t_1 \le \cdots \le t_n) d\mathbf{t}$$

Therefore the density is

$$f_{T_1,\ldots,T_n|N_t=n}(t_1, \ldots, t_n) = \frac{n!}{t^n}.$$

Surprisingly, this is actually the distribution of the order statistics of a uniform distribution.

Generating arrival times of a Poisson processes

What does the result above suggest? In order to simulate the arrival times of a Poisson process on the time interval $[0, t]$, one can

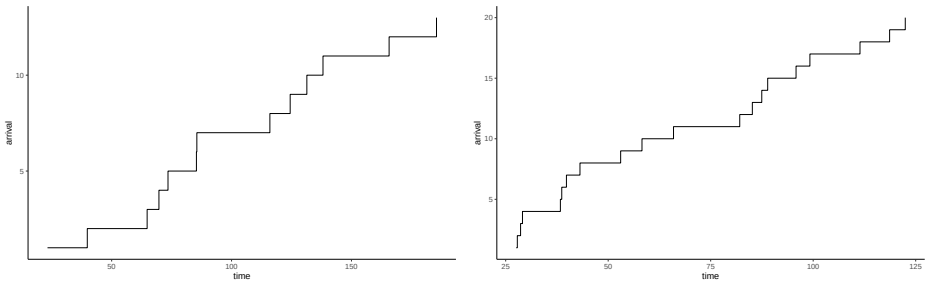
1. simulate a random variable $X_t \sim \text{Poisson}(t\lambda)$;
2. generate X independent uniform random variables $(U_1, \ldots, U_{X_t})$;
3. the arrival times of the Poisson process will be given by $(tU_{(1)}, \ldots, tU_{(X_t)})$.

Let's call this method (1).
I also remind you of another way to generate arrival times of a Poisson processes. Call this method (2). In a previous characterization of this process, we have seen that the inter-arrival times are exponentially distributed. Consequently, one can generate samples from a Poisson process sampling directly from the exponential, that is from $t = 0$, for $i \ge 1$, repeat the following loop:

1. sample $T_i \sim \text{Exponential}(\lambda)$;

2. the $i-th$ arrival time will be given by $t + T_i$;

3. set $t = t + T_i$;



(a) Arrival times with method (1). (b) Arrival times with method (2).

Figure 7: Mmethod (1): $\lambda = 0.1$, $n = 20$. Method (2), $\lambda = 0.1$, $t = 200$.

Some code if you are curious:

```
## method (1)
lambda <- 0.1
t <- 200
X <- runif(rpois(1, lambda*t),0,1)

times <- data.frame(time = sort(X)*t,
                    arrival = seq_along(X))

ggplot(times, aes(time, arrival)) +
  geom_step() +
  theme_classic() +
  theme(legend.position="none")


## method (2)
lambda <- 0.1
n <- 20
X <- rexp(n,lambda)
times <- data.frame(time = cumsum(X),
                    arrival = seq_along(X))

ggplot(times, aes(time, arrival)) +
  geom_step() +
  theme_classic() +
  theme(legend.position="none")
```

The Nonhomogeneous Poisson Process

So far we fixed the rate of a Poisson process to be a fixed scalar $\lambda > 0$. However, we can generalize the definition of Poisson process and allow the arrival rate to vary with as a function of time.

We say that a random process $\mathbb{X}$ is a nonhomogeneous Poisson process with intensity function $\lambda(t)$ for $t \geq 0$ if

- $\mathbb{X}$ is a counting process

- $X_0 = 0$
- $\mathbb{X}$ has independent increments
- $P(X_{t+h} - X_t = 1) = \lambda(t)h + o(h)$
- $P(X_{t+h} - X_t \geq 2) = o(h)$

In this case, for any $0 \leq s < t$, we have

$$N_t - N_s \sim \text{Poisson}\left(\int_s^t \lambda(u)\, du\right).$$

Clearly the nonhomogeneous Poisson process described above does not have stationary increments.

The Spatial Poisson Process

We saw that the Poisson process is a probabilistic model for random scattering of values across the positive real numbers. What if we want to randomly scatter points over a spatial domain?

We say that a counting process $\mathbb{N} = \{N(A)\}_{A \subset \mathcal{X}}$ on a set $\mathcal{X}$ is a spatial Poisson process with rate $\lambda > 0$ if

- $N(\emptyset) = 0$
- $\mathbb{N}$ has independent increments, i.e. if $A_1, \ldots, A_n$ are disjoint sets then $N(A_1), \ldots, N(A_n)$ are independent random variables
- $N(A) \sim \text{Poisson}(\lambda \,\text{vol}(A))$.

The spatial Poisson process is a basic example of a *point process*. If we denote $Y_1, Y_2, \ldots$ the 'points' of the process, the spatial Poisson random process induces a random measure (the Poisson random measure) over the subsets of $\mathcal{X}$ by means of

$$N(A) = \sum_{i=1}^{N(\mathcal{X})} \mathbb{1}_A(Y_i).$$

Lecture 17

Recommended readings: Ross, sections 4.1 → 4.3

Markov Chains

The Markov Property

Let $\mathbb{X} = \{X_n\}_{n=1}^{\infty}$ be a discrete-time random process with discrete state space $\mathcal{S}$. $\mathbb{X}$ is said to be a *Markov chain* if it satisfies the Markov property

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n),$$

for any $n \geq 0$.

The Markov property can be stated in the following equivalent way. Let h be any bounded function mapping $\mathcal{S}^{\infty} \to \mathbb{R}$. Then $\mathbb{X}$ satisfies the Markov property if

$$\begin{aligned} &E(h(X_{n+1}, X_{n+2}, \ldots) | X_n, X_{n-1}, \ldots, X_0) \\ &= E(h(X_{n+1}, X_{n+2}, \ldots) | X_n). \end{aligned}$$

How can we interpret the above condition in an algorithmic fashion? Think about h being any algorithm you may use on the future states of the process (i.e. h is any feature of the future states of the process that you may be interested to consider). Then your 'best' prediction of the algorithm's output on the process's future states given the entire history of the process up to the current state is a function only of the current state.

Example in class:
Let $\mathbb{X} = \{X_n\}_{n=0}^{\infty}$ be a random process where $\{X_n\}_{n=0}^{\infty}$ is a collection of independent random variables. Then, for all $n \geq 0$,

$$\begin{aligned} &E(h(X_{n+1}, X_{n+2}, \ldots) | X_n, X_{n-1}, \ldots, X_0) \\ &= E(h(X_{n+1}, X_{n+2}, \ldots)) = E(h(X_{n+1}, X_{n+2}, \ldots) | X_n). \end{aligned}$$

Thus $\mathbb{X}$ satisfies the Markov property.

We are usually interested in two kinds of questions about Markov chains:

- how will the system evolve in the short term? In this case, we use conditional probabilities to compute the likelihood of various events and evolutionary paths.

- how will the system evolve in the long term? In this case, we use conditional probabilities to identify the 'limiting behavior' over vast stretches of time, or the 'equilibrium' of the system.

Transition Probabilities

From now on, to simplify the notation, we will assume that the state space $\mathcal{S}$ is the set (or a subset) of the positive integers $\{0, 1, 2, \dots\}$. The evolution of a Markov chain with respect to time is described in terms of the transition probabilities

$$P(X_{n+1} = j | X_n = i) = P_{n;i,j}$$

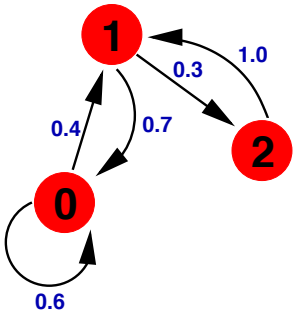
Notice that in general the transition probability from state i to state j can vary with time. It is very convenient, especially for computational reasons, to arrange the transition probabilities into a *transition probability matrix* (t.p.m.) $\boldsymbol{P}_n$.

When the transition probabilities $P_{n;i,j}$ do not change over time, we say that the Markov chain is *time-homogeneous*. In this case, we drop the subscript $_n$ and we simply write

$$P(X_{n+1} = j | X_n = i) = P_{i,j}.$$

Similarly, the t.p.m. of a time-homogeneous Markov chain is denoted $\boldsymbol{P}$. In this class we will mainly focus on time-homogeneous Markov chain.

Exercise in class:
Write the t.p.m. of the time-homogeneous Markov chain in the figure below.



Exercise in class:
Consider a simple random walk starting from 0. At each time n, the pro-

cess increments by 1 with probability $p \in [0,1]$ or decrements by 1 with probability $1-p$. Describe the t.p.m. of this process.

Stochastic Matrices

Consider a matrix A. The matrix A is said to be a *stochastic matrix* if it satisfies the two following properties:

1. all the entries of A are non-negative, i.e. $A_{i,j} \geq 0$ for all i, j
2. each row of A sums to 1: $\sum_j A_{i,j} = 1$ for all i.

Notice that we call the matrix 'stochastic', but such matrix is not random! If the matrix A is such that 1. and 2. are satisfied and, furthermore, also each column of A sums to 1 ($\sum_i A_{i,j} = 1$ for all j), then A is said to be a *doubly stochastic matrix*.

Question: is a t.p.m. a stochastic matrix? Why?

Question: is the t.p.m. of the 3-state Markov chain above a doubly stochastic matrix?

Question: is the t.p.m. of the simple random walk above a doubly stochastic matrix?

Using Transition Probabilities

Let's start with an example. Consider again the 3-state Markov chain with its t.p.m.

$$\boldsymbol{P} = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.7 & 0 & 0.3 \\ 0 & 1 & 0 \end{pmatrix}$$

Suppose that the initial distribution of the Markov chain at time 0 is $p_{X_0} = (P(X_0 = 0), P(X_0 = 1), P(X_0 = 2)) = (0.2, 0, 0.8)$. What is the probability

distribution of X_1? What is the probability distribution of X_n, for any $n \geq 1$?

Let $\boldsymbol{P}^{(n)}$ denote the n-step t.p.m.. Based on the example above, we see that $\boldsymbol{P}^{(n)} = \boldsymbol{P}\boldsymbol{P}\boldsymbol{P}\ldots\boldsymbol{P}$, i.e. $\boldsymbol{P}^{(n)}$ is the n-th power of $\boldsymbol{P}$, $\boldsymbol{P}^n$, with respect to the usual matrix multiplication.

This is made precise in terms of the Chapman-Kolmogorov equations, which state that for all positive integers n, m we have

$$\boldsymbol{P}^{(n+m)} = \boldsymbol{P}^{(n)}\boldsymbol{P}^{(m)}$$

with $\boldsymbol{P}^{(0)} = I$.

Therefore, we conclude that a discrete-state time-homogeneous Markov chain is completely described by the initial proabability distribution p_{X_0} and the t.p.m. $\boldsymbol{P}$.

Probability of Sample Paths

Consider a sample path $(i_0, i_1, \ldots, i_n)$. Note that the Markov property makes it very easy to compute the probability of observing this sample path. In fact we have

$$\begin{aligned}
&P(\text{path } (i_o \cdots, i_n) \text{ is observed}) \\
&P(X_0 = i_0 \cap X_1 = i_1 \cap \cdots \cap X_n = i_n) \\
&= P(X_0 = i_0)P(X_1 = i_1 | X_0 = i_0)P(X_2 = i_2 | X_1 = i_0 \cap X_2 = i_1) \times \ldots \\
&\times P(X_n = i_n | X_{n-1} = i_{n-1} \cap \cdots \cap X_1 = i_1 \cap X_0 = i_0) \\
&= P(X_0 = i_0)P(X_1 = i_1 | X_0 = i_0)P(X_2 = i_2 | X_1 = i_1) \ldots P(X_n = i_n | X_{n-1} = i_{n-1}).
\end{aligned}$$

Communicating Classes

We now turn to the questions of which states of the process can be reached from which. This type of analysis will help us understand both the short term and the long term behavior of a Markov chain.

Let $i, j \in \mathcal{S}$ be two states (possibly with $i = j$). We say that j is *accessible* from i if, for some $n \geq 0$, $\boldsymbol{P}^{(n)}_{i,j} > 0$. This is often denoted $i \to j$.

Let again $i, j \in \mathcal{S}$ be two states (possibly with $i = j$). We say that i and j *communicate* if $i \to j$ and $j \to i$. This is often denoted $i \leftrightarrow j$.

The relation of communication satisfies the following properties for each $i, j, k \in \mathcal{S}$:

1. $i \leftrightarrow i$

2. $i \leftrightarrow j \implies j \leftrightarrow i$

3. $i \leftrightarrow j \wedge j \leftrightarrow k \implies i \leftrightarrow k$.

Furthermore, communication is an *equivalence relation*. This means that the state space $\mathcal{S}$ can be partitioned into disjoint subsets, where all the states that communicate with each other belong to the same subset. These subsets are called *communicating classes*. Each state of the state space $\mathcal{S}$ lies in exactly one communicating class.

Exercise in class:
Consider the following t.p.m. of a Markov chain:

$$\boldsymbol{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

What are the communicating classes of this Markov chain?

Exercise in class:
Consider the following t.p.m. of a Markov chain:

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 0.4 & 0 & 0.4 & 0 & 0.1 \\ 0 & 0.3 & 0 & 0.7 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 0 & 0.5 & 0 & 0.2 \\ 0 & 0.3 & 0 & 0.3 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & 0.6 & 0.4 \end{pmatrix}$$

What are the communicating classes of this Markov chain?

Absorbing Classes

A communicating class is called *absorbing* if the Markov chain never leaves that class once it enters.

Question: what communicating classes are absorbing in the above examples?

Irreducibility

A Markov chain is said to be *irreducible* if it has only one communicating class (trivial partition of the state space). In other words, it is possible to get from any state to any other state in finite time, so the entire state space *is* a communicating class.

If a Markov chain is not irreducible, then we can partition its state space $\mathcal{S}$ into disjoint sets

$$\mathcal{S} = \mathcal{D} \cup (\cup_i \mathcal{C}_i) \tag{64}$$

where each C_i is an absorbing communicating class and $\mathcal{D}$ is a union of non-absorbing communicating classes.

Note the following key implication: eventually, the Markov chain will leave $\mathcal{D}$ and enter exactly one of the C_i's. At that point the Markov chain will act like an irreducible Markov chain on the reduced state space C_i.

Recurrent and Transient States

Informally, we say that a state $i \in \mathcal{S}$ is *recurrent* if the probability that starting from state i the Markov chain will ever visit again state i is 1. On the other hand, if the probability that starting from state i the Markov chain will ever visit again state i is strictly less than 1, then we say that i is a *transient* state.

Note that, based on the above definition,

- if $i \in \mathcal{S}$ is recurrent then, if the Markov chain starts from state i, state i will be visited infinitely often
- if $i \in \mathcal{S}$ is transient then, each time the Markov chain visits state i, there is a positive probability $p \in (0, 1)$ that state i will never be visited again. Therefore, the probability that starting from state i the Markov chain will visit state i exactly n times is $(1-p)^{n-1}p$; this means that the amount of time that the Markov chain spends in state i (starting from state i) is a Geometric random variable with parameter p. This implies that, if the Markov chain starts from state i, the expected amount of time spent in state i is $1/p$.

From the above, it follows that a state $i \in \mathcal{S}$ is recurrent if and only if, starting from state i, the expected number of times that the Markov chain visits state i is infinite. This allows us to characterize recurrent and transient states in terms of transition probabilities. Let $I_n = \mathbb{1}_{\{X_n=i\}}(X_n)$

be the random variable indicating whether the Markov chain is visiting state i at time n. Then $\sum_{n=0}^{\infty} I_n$ is the amount of time spent by the Markov chain in state i. We have

$$E\left(\sum_{n=0}^{\infty} I_n | X_0 = i\right) = \sum_{n=0}^{\infty} E(I_n | X_0 = i)$$
$$= \sum_{n=0}^{\infty} P(X_n = i | X_0 = i) = \sum_{n=0}^{\infty} \boldsymbol{P}_{i,i}^{(n)}$$

and therefore

$$\sum_{n=0}^{\infty} \boldsymbol{P}_{i,i}^{(n)} = \infty \iff i \text{ is recurrent}$$
$$\sum_{n=0}^{\infty} \boldsymbol{P}_{i,i}^{(n)} < \infty \iff i \text{ is transient.}$$

This is the main way of checking if a state is recurrent/transient.

Notice also that if $i \in \mathcal{S}$ is recurrent then every state in the same communicating class is also recurrent. The property of being recurrent or transient is a class property.

Example in class:
Consider the random walk on the integers. Since this Markov chain is irreducible (all the states communicates), either all the states are recurrent or they are all transient. It turns out that the random walk on the integers is recurrent if and only if the probability of a positive/negative increment is exactly 1/2. Otherwise, the random walk on the positive integers is transient (see Ross, pages 208-209 for a proof).

Exercise in class: Consider again the previous two Markov chains. Which classes are recurrent/transient?

Lecture 18

Recommended readings: Ross, sections 4.4

Markov Chains (part 2)

Periodicity

Let's start with a simple example. Consider the t.p.m.

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

We have that

$$\boldsymbol{P}^{(2)} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

and

$$\boldsymbol{P}^{(3)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Starting from any of the states $0, 1, 2$, how often does the Markov chain return to the initial state?
The above example suggests that this simple Markov chain has a built-in periodicity of order 3. How can we make this idea more formal and generalize it to more complicated Markov chains?

For any state $s \in \mathcal{S}$ of a Markov chain, we define the *period* of s as

$$d(s) = \gcd\{n \geq 1 : P^n_{s,s} > 0\}$$

where gcd stands for 'greatest common divisor'. This implies that $\boldsymbol{P}^{(n)}_{s,s} = 0$ unless $n = md(s)$ for some $m \in \mathbb{Z}$.

An irreducible Markov chain is said to be *aperiodic* if $d(s) = 1$ for all states $s \in \mathcal{S}$.

Exercise in class:
Consider

$$\boldsymbol{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{8} & \frac{7}{8} & 0 \end{pmatrix}.$$

Is this Markov chain periodic or aperiodic?

Exercise in class:
Consider

$$\boldsymbol{P} = \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{8} & \frac{7}{8} & 0 \end{pmatrix}.$$

Is this Markov chain periodic or aperiodic?

Periodicity is another class property. This means that the period is constant over any fixed communicating class of a Markov chain.

Cyclic Class Decomposition[12]

Let $\mathbb{X}$ be an irreducible Markov chain with period d. There exists a partition of S

$$\mathcal{S} = \cup_{k=1}^{d} \mathcal{U}_k$$

such that

$$P_{s,\mathcal{U}_{k+1}} = 1$$

for $s \in \mathcal{U}_k$ and $k = 0, 1, \ldots, d-1$.

The sets $\mathcal{U}_1, \ldots, \mathcal{U}_s$ are called *cyclic classes* of $\mathbb{X}$ because $\mathbb{X}$ cycles through them successively. It thus follows that the 'accelerated' Markov chain $\mathbb{X}^d = \{X_{id}\}_{i=1}^{\infty}$ has transition probability matrix $\boldsymbol{P}^d$ and each $\mathcal{U}_k$ is an absorbing, irreducible, aperiodic class.

In light of this result, we can rewrite the state space decomposition of equation (64) as

$$\mathcal{S} = \mathcal{D} \cup (\cup_i \mathcal{C}_i) = \mathcal{D} \cup \left(\cup_i \cup_{j=1}^{d_i} \mathcal{U}_{i,j} \right)$$

where each absorbing class C_i is an irreducible Markov chain with period d_i whose state space can be partitioned into the d_i cyclic classes $\mathcal{U}_{i,1}, \ldots, \mathcal{U}_{i,d_i}$.

Positive and Null Recurrence

We already defined the notion of recurrence in the previous lecture. Here, we introduce a somewhat technical (but important) point. Suppose that

[12] We have not covered this topic in class.

state $s \in \mathcal{S}$ is recurrent. Then, based on the previous lecture, this means that if the Markov chain starts from state s, the chain will visit state s in the future infinitely often. We now make the following distinction for a recurrent state s (or, equivalently, for the communicating class of s, since recurrence is a class property):

- s is *positive recurrent* if, starting from s, the expected amount of time until the Markov chain visits again state s is finite
- s is *null recurrent* if, starting from s, the expected amount of time until the Markov chain visits again state s is infinite.

While there is indeed a difference between positive and null recurrent states, it can be shown that in a Markov chain with finite state space all recurrent states are positive recurrent.

Ergodicity, Equilibrium and Limiting Probabilities

Consider a Markov chain with t.p.m. $\boldsymbol{P}$. Suppose there exists a probability distribution π over the state space such that $\pi \boldsymbol{P} = \pi$. This implies that $\pi = \pi \boldsymbol{P}^{(n)}$ for all integers $n \geq 0$. The probability distribution π is called the *equilibrium distribution* or the *stationary distribution* of the Markov chain. In equilibrium π_j is the propotion of time spent by the Markov chain in state j.

A communicating class is said to be *ergodic* if it is positive recurrent and aperiodic. We have the following important result: for an irreducible and ergodic Markov chain the limiting probabilities

$$\pi_j = \lim_{n \to \infty} P_{ij}^n \quad j \in \mathcal{S}$$

exist and the limit does not depend on i. Furthermore, the vector π of the above limiting probabilities is the unique solution to $\pi = \boldsymbol{P}\pi$ and $\sum_j \pi_j = 1$ (hence it corresponds to the stationary distribution of the Markov chain).

The limiting probabilities can also be shown to correspond to the long-term proportion of time that the Markov chain spends in each state. More precisely, we have that

$$\pi_j = \lim_{n \to \infty} \frac{v_{ij}(n)}{n}$$

where $v_{ij}(n)$ denotes the expected number of visits to state j, starting from state i, within the first n steps.

Equilibrium When the State Space is Finite

Let U be a $|\mathcal{S}| \times |\mathcal{S}|$ matrix with entries all equal to 1, let u be a $|\mathcal{S}|$-dimensional vector with entries all equal to 1 and let o be a $|\mathcal{S}|$-dimensional vector with entries all equal to 0. Let $\mathbb{X}$ be an irreducible, aperiodic, and finite state (and thus ergodic) Markov chain. If we want to find the stationary distribution of $\mathbb{X}$ we must solve

$$\begin{cases} \pi(I - \boldsymbol{P}) = o \\ \pi U = u. \end{cases}$$

The solution to the above system is $\pi = u(I - P + U)^{-1}$.

Exercise in class:
Consider the t.p.m.

$$\boldsymbol{P} = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.7 & 0 & 0.3 \\ 0 & 1 & 0 \end{pmatrix}.$$

Apply the above formula to verify that $\pi \approx (0.5738, 0.3279, 0.0984)$.

It can be shown that if the matrix $\boldsymbol{P}$ is doubly stochastic, we have that $\pi_j = 1/|\mathcal{S}|$ for all $j \in \mathcal{S}$.

Exercise in class:
Consider again the random walk on the pentagon. What is the stationary distribution of this Markov chain?

Lecture 19

Recommended readings: Ross, section 4.6

Markov Chains (part 4)

More on Equilibrium Distribution and Limiting Probabilities

In the previous lecture, we studied that if an irreducible Markov chain is ergodic (i.e. positive recurrent and aperiodic), the limiting probabilities

$$\lim_{n\to\infty} P_{ij}^n \quad j \in \mathcal{S} \tag{65}$$

exist and these limits do not depend on i or on the initial distribution p_{X_0}. Furthermore, we said that, whenever a stationary distribution π exists, the limiting probabilities above correspond exactly to the entries of π.

Let's clarify a slightly technical point. While it is true that in order for the Markov chain to admit the above limiting probabilities we need the Markov chain to be irreducible and ergodic, the existence of a stationary distribution does not require the aperiodicity assumption.

Hitting Probabilities (part 1)

Recall the decomposition of the state space of equation (64). Suppose that the Markov chain starts at a state $s \in \mathcal{D}$ which is not absorbing. At some point in time, the Markov chain will eventually enter one of the absorbing classes C_i of the decomposition (64) and will be 'stuck' in that class from that time on. What is the probability that, starting from $s \in \mathcal{D}$, the chain will hit the absorbing class C_i?

Define $h_i(s)$ as the probability that, starting from a state $s \in \mathcal{S}$, the chain will hit the absorbing class C_i. We have

$$h_i(s) = \sum_{s' \in C_i} P_{ss'} + \sum_{u \in \mathcal{D}} P_{su} h_i(u)$$

and the vector h_i is the smallest non-negative solution to the above equation.

Define $\boldsymbol{P}_{\mathcal{D}}$ to be the submatrix of $\boldsymbol{P}$ associated to $\mathcal{D}$ and $b_i(s) = \sum_{s' \in C_i} P_{ss'}$ for $s \in \mathcal{D}$. When $\mathcal{S}$ is finite, we can rewrite the above equation in vector/matrix form as

$$(I - \boldsymbol{P}_{\mathcal{D}})h_i = b_i \implies h_i = (I - \boldsymbol{P}_{\mathcal{D}})^{-1} b_i.$$

Exercise in class:
Let $\mathbb{X}$ be a Markov chain with t.p.m.

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

defined on $\mathcal{S} = \{0, 1, 2, 3, 4\}$. We have that $\mathcal{C}_1 = \{0\}$, $\mathcal{C}_2 = \{4\}$, and $\mathcal{D} = \{1, 2, 3\}$. Then,

$$\boldsymbol{P}_{\mathcal{D}} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix},$$

$$b_1 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix},$$

$$b_2 = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{3} \end{pmatrix},$$

and

$$(I - \boldsymbol{P}_{\mathcal{D}})^{-1} = \frac{1}{19} \begin{pmatrix} 30 & 14 & 12 \\ 24 & 34 & 21 \\ 18 & 16 & 30 \end{pmatrix}.$$

Thus,

$$h_1 = \frac{1}{19} \begin{pmatrix} 15 \\ 12 \\ 9 \end{pmatrix}$$

and

$$h_2 = \frac{1}{19} \begin{pmatrix} 4 \\ 7 \\ 10 \end{pmatrix}.$$

Block Decomposition of the T.P.M.

From now on, we consider a finite state Markov chain with a set $\mathcal{A}$ of $|\mathcal{A}| = a$ absorbing states and a set $\mathcal{T}$ of $|\mathcal{T}| = t$ of transient states. We can rearrange the t.p.m. of the Markov chain as

$$\boldsymbol{P} = \begin{pmatrix} I & 0 \\ S & T \end{pmatrix}$$

where I is a $a \times a$ matrix, T is a $t \times t$ matrix, S is a $t \times a$ matrix, and 0 is a $a \times t$ matrix of zeroes. Notice that the values of S and T are not unique: they depend on how you order the states.

Hitting Probabilities (part 2), Expected Hitting Times, Expected Number of Visits

In terms of the block decoposition of the t.p.m., it can be shown that we can obtain the hitting probabilities matrix by simply computing

$$(I - T)^{-1}S.$$

On the other hand, the matrix $(I - T)^{-1}$ gives us the expected number of visits to state s' starting from state s, with $s, s' \in \mathcal{T}$.

Finally, if we let u denote a vector of ones, it can be shown that $m = (I - T)^{-1}u$ returns the vector of the expected time until, starting from a transient state $s \in \mathcal{T}$, the Markov chain hits an absorbing state.

Example in class (the rat maze):
A rat is placed in the maze depicted below. If the rat is in the rooms 2, 3, 4, or 5, it chooses one of the doors at random with equal probability. The rat stays put once it reaches either the food or the shock.

The t.p.m. for the chain is

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Using the formulae above, we have that the matrix of the expected number of visits is

$$\begin{pmatrix} \frac{26}{21} & \frac{2}{7} & \frac{5}{7} & \frac{2}{21} \\ \frac{4}{21} & \frac{10}{7} & \frac{4}{7} & \frac{10}{21} \\ \frac{10}{21} & \frac{4}{7} & \frac{10}{7} & \frac{4}{21} \\ \frac{2}{21} & \frac{5}{7} & \frac{2}{7} & \frac{26}{21} \end{pmatrix},$$

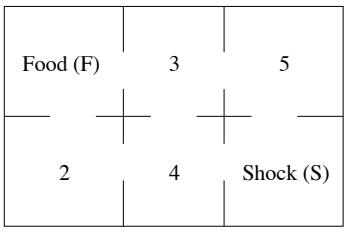
while the matrix of hitting probabilities is

$$\begin{pmatrix} \frac{5}{7} & \frac{2}{7} \\ \frac{4}{7} & \frac{3}{7} \\ \frac{3}{7} & \frac{4}{7} \\ \frac{2}{7} & \frac{5}{7} \end{pmatrix}$$

and the expected times until absorption are

$$\begin{pmatrix} \frac{49}{21} \\ \frac{56}{21} \\ \frac{56}{21} \\ \frac{49}{21} \end{pmatrix}.$$

How do we interpret these numbers?



An introduction to Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms that provide approximate sampling techniques by means of Markov chains. They are widely diffused in statistics, mathematics, physics, and many other fields. Why are they so useful?
Let's say that you need to compute the expectation of some function g with respect to the distribution P admitting density function f with support $\mathcal{S}$, that is

$$E[g(X)] = \int_{\mathcal{S}} g(X)f(X)dx.$$

As we have seen, the integral might be too hard to compute or might simply not admit a closed-form solution. In this case, we say that it is *intractable*. Then one can rely either on numerical or on stochastic approximations. MCMC algorithms fall into the second class.

Say, for instance, that you would like to compute the area of a circle with radius 1. We know that the area of the circle is $\pi r^2 = \pi \cdot 1/4$. We also know that the are of the square in which the circle is inscribed is 4. How can we estimate the value of π? The general Monte Carlo recipe is the following: sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} f_{X_1,X_2}$ where

$$f_{X_1,X_2} = \mathbb{1}(-1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1).$$

that is, we sample points from the square. Then let $Y_i = \mathbb{1}(X_{i,1}^2 + X_{i,2}^2 \leq 1)$. The MC estimator for the value of π is given by

$$\hat{\pi} = \frac{4}{n} \sum_{i=1}^{n} Y_i.$$

What guarantees do we have for such an estimator? By the strong law of large numbers we know that, since $\hat{\pi}$ is unbiased, then $\hat{\pi} \xrightarrow{p} \pi$ as $n \to \infty$.

Lecture 20

An introduction to Markov Chain Monte Carlo[13]

Markov Chain Monte Carlo (MCMC) is a class of algorithms that provide approximate sampling techniques by means of Markov chains. They are widely diffused in statistics, mathematics, physics, and many other fields. Why are they so useful?
Let's say that you need to compute the expectation of some function g with respect to the distribution P admitting density function f with support $\mathcal{S}$, that is

$$E[g(X)] = \int_{\mathcal{S}} g(X)f(X)dx.$$

As we have seen, the integral might be too hard to compute or might simply not admit a closed-form solution. In this case, we say that it is *intractable*. Then one can rely either on numerical or on stochastic approximations. MCMC algorithms fall into the second class.

In more generality, MCMC's in machine learning are mainly used in (1) Bayesian inference for the computation of normalizing of constants in the posterior, marginalization, or expectation; (2) statistical mechanics; (3) optimization; (4) penalized likelihood model selection.

Motivating example

Say, for instance, that you would like to compute the area of a circle with radius 1. We know that such an area is $\pi r^2 = \pi \cdot 1$. We also know that the are of the square in which the circle is inscribed is 4. How can we estimate the value of π? The general Monte Carlo recipe is the following: sample $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{iid}{\sim} f_{X_1,X_2}$ where

$$f_{X_1,X_2} = \mathbb{1}(-1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1).$$

that is, we sample points from the square. Then let $Y_i = \mathbb{1}(X_{i,1}^2 + X_{i,2}^2 \leq 1)$. The MC estimator for the value of π is given by

$$\hat{\pi} = \frac{4}{n} \sum_{i=1}^{n} Y_i.$$

What guarantees do we have for such an estimator? By the weak law of large numbers we know that, since $\hat{\pi}$ is unbiased, then $\hat{\pi} \xrightarrow{p} \pi$ as $n \to \infty$.

[13] These lecture notes are roughly based on a class that I took at the University of Torino.

Moreover, if $\sigma^2 = V(Y)$, we also know that $\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by the central limit theorem as $n \to \infty$.

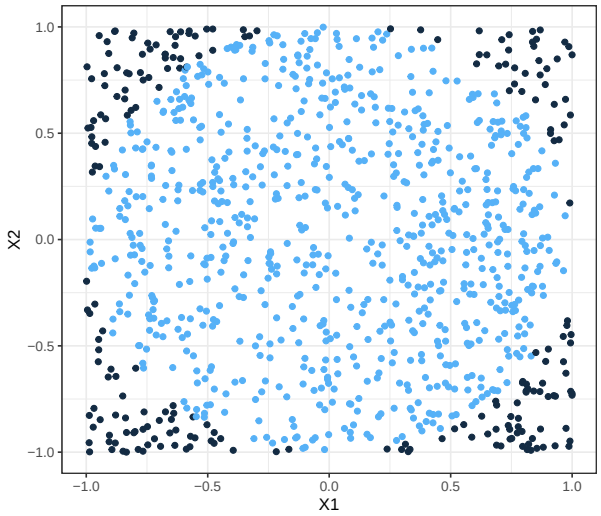


Figure 8: 1000 samples from a bivariate uniform distribution on $[-1, 1]^2$. The blue points fall inside the circle of radius 1, the black points fall outside. The estimate value of π from these samples is 3.1444.

```
n <- 1e3
x1 <- runif(n,-1,1)
x2 <- runif(n,-1,1)
lab <- if_else(x1^2+x2^2<=1,1,0)

df <- data.frame(X1 = x1, X2 = x2, label = lab)

ggplot(df, aes(X1, X2, col = label)) +
  geom_point() +
  xlim(-1,1) +
  ylim(-1,1) +
  theme_bw() +
  theme(legend.position="none")
```

General Monte Carlo recipe

Let's assume that we are interested in computing a numerical approximation to the following integral

$$I(g) = \int_{\mathcal{S}} g(x) f(x) dx$$

and that we are able to draw samples from the density function f. Then

- draw n samples $\{x_i\}_{i=1}^n$ from the density f with support $\mathcal{S}$;
- compute $I_n(g) = \frac{1}{n}\sum_{i=1}^n g(x_i)$ to approximate the integral.

What guarantees of convergence to $I(g)$ does $I_n(g)$ have? As above,

- by weak LLN, $I_n(g) \xrightarrow{p} \int_{\mathcal{S}} g(x)f(x)dx$ as $n \to \infty$;
- let $\sigma^2 = V(g(X))$. By CLT, $\sqrt{n}(I_n(f) - I(f)) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$.

Although the MC recipe might appear to be general enough to tackle any kind of problem, most of the times we do not know how to sample from density f. The techniques in the MCMC framework provide a way to do this, that is to sample .
While the MCMC class of algorithms is very large, we will only focus on one them for the sake of time. I chose Metropolis Hastings algorithm to motivate *why* we are so much interested in Markov chains. If you are more interested in the topic, you should first look into MC algorithms such as importance sampling or rejection sampling.

The basic idea of MCMC algorithms is to produce samples from a Markov chain having f as stationary distribution. This allows, at any step of the chain, to obtain samples that come from f. Of course, we require the Markov chain to be irreducible and ergodic: this allows us to recover in asymptotics the limiting probabilities, and therefore also the stationary distribution without computing the it directly.

The Metropolis-Hastings algorithm

The MH (Metropolis-Hastings) algorithm was first developed by Metropolis and then generalized by Hastings in 1970. Let us denote with f the target density we want to draw samples from. Moreover, let

- Q be the *proposal* distribution represented by an irreducible and aperiodic TPM Q;
- A be the matrix of *acceptance* probabilities.

The algorithm works in the following way:

1. given the current state of the chain at time $n = i$, $X_n = i$, a proposal X_{n+1}^* from the $i - th$ row of Q is made such that $P(X_{n+1} = j | X_n = i) = Q_{ij}$.

2. then the move of the chain occurs with probability

$$X_{n+1} = \begin{cases} X^*_{n+1} & \text{with probability } A_{ij} \\ X_n & \text{with probability } 1 - A_{ij} \end{cases}$$

Equivalently, the resulting TPM for this chain is

$$P_{ij} = \begin{cases} Q_{ij} A_{ij} & \text{for } i \neq j \\ 1 - \sum_{i \neq j} Q_{ij} A_{ij} & \text{for } i = j \end{cases}$$

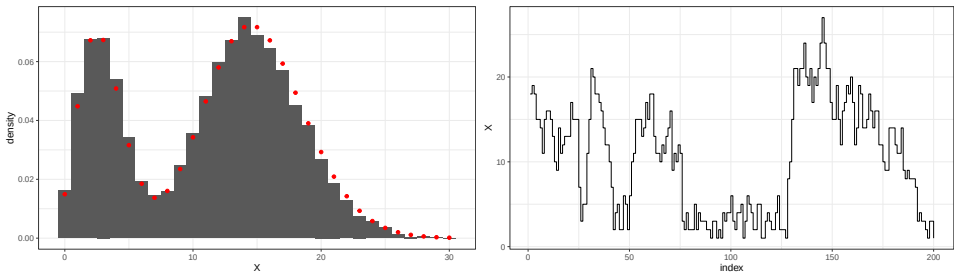
The MH algorithm with acceptance probabilities of the form

$$a_{ij} = \min \left\{ 1, \frac{f_j}{f_i} \frac{q_{ji}}{q_{ij}} \right\}$$

generate a *reversible* Markov chain, that is a chain that satisfied the *detailed balance* condition. Although we have not seen such a property, this ensures that f is a stationary distribution for this chain. Consequently, after a certain *burning time* of the chain, that you might imagine as a necessary number of steps for the chain to get into the asymptotic regime after forget the initial state, the chain will provide sample from the target density. The samples are not independent, but one cam both generate several separate Markov chains and consider samples only k steps apart, where k is determined looking at the autocorrelation function.The matrix Q typically needs careful design.

Two notable cases of the MH algorithms are

- the independent sampler: Q_{ji} does not depend on i, but only on j;
- the Metropolis algorithm: $Q_{ij} = Q_{ji}$.



(a) (b)

Figure 9: Metropolis-Hastings for sampling from the Poisson mixture P=0.3 ·Pois(3)+0.7·Pois(15). The length of the chain is $n = 10^5$ and burn-in time is 10^3 steps. The proposal is taken to be a symmetric random walk. Out of the $99 \cdot 10^3$ steps considered, only samples every $k = 10$ steps are considered. Figure (a): the black bars represent the normalized histogram of the samples, while the red dots denote the true density. Figure (b): 200 consecutive (every $k = 10$, as explained above) samples from the MC.

Some code to generate the example in figure 9:

```
n <- 1e5
i <- 1
path <- c(4)
while(i <= n){
  X_prop <- max(path[i]+sample(c(-1,1),1),0) # proposal
  if(runif(1,0,1)<=min(1, dens(X_prop)/dens(path[i]))){ # acceptance
    path <- c(path, X_prop)} else{
      path <- c(path, path[i])} #
  i <- i + 1
}

burn_in <- 1e3 # burn in time
path <- path[burn_in:length(path)] # discard burn in time
k_seq <- seq(1,length(path),by=10) # for uncorrelated samples
path <- path[k_seq] # discard correlated samples

df_dens <- data.frame(path = c(0:30), y = dens(c(0:30)))
df_path <- data.frame(X = path)

ggplot(df_path, aes(X)) +
  geom_histogram(binwidth=1, aes(y = ..density..)) +
  theme_bw() +
  geom_point(data=df_dens, aes(x = path, y = y ),
        col = 'red', inherit.aes = FALSE)
```

```
df_path <- data.frame(index = c(1:200), X = path[1:200])
ggplot(df_path, aes(index, X)) +
  geom_step() +
  theme_bw()
```

This is only one out of the 435 estimated bridges present in Venice, and it's not even Rialto's or Calatrava's. If you are more interested about these methods, check out the following introductory paper and many tutorials online.

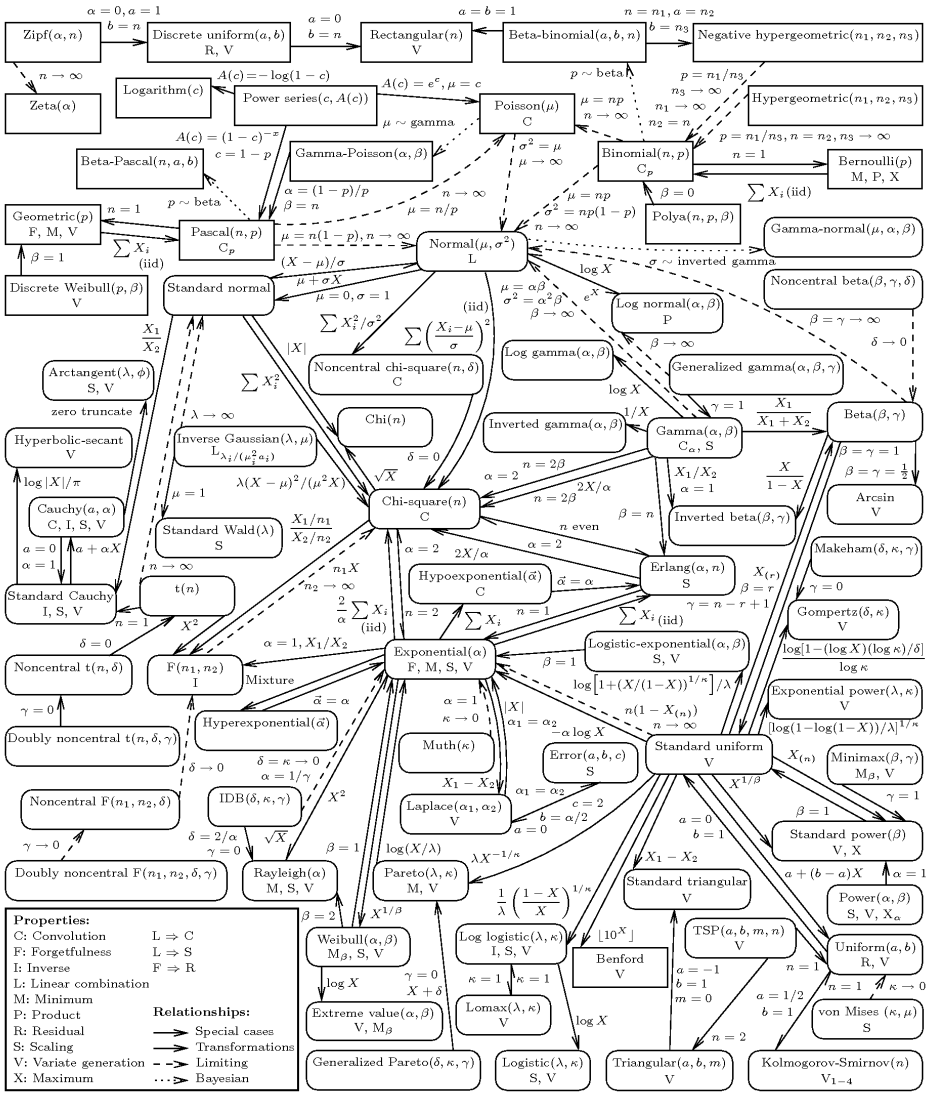


Figure 1: Relationships between probability distributions. Taken from http://www.math.wm.edu/~leemis/chart/UDR/UDR.html.