Part 1: Warming Up

Probability Theory: Math 605, Fall 2024

Luc Rey-Bellet

University of Massachusetts Amherst

2024-09-27

1 Axioms of Probability

- The axioms of probability were formalized by Andrey Kolmogorov in 1933 but probability theory started much earlier.
- The Doctrines of Chances by Abraham de Moivre in 1718 is usually considered as the first probability textbook and de Moivre also first proved a version of the central limit theorem in 1733.
- The Ars conjectandi by Jacob Bernoulli in 1713 has the first proof of the Law of Large numbers and the subsequent work by Pierre Simon Laplace Théorie analytique des probabilités in 1819 present a formalization of probability theory.
- Laplace's book the concept of conditional probability ("the" key idea in probability) is also presented (first introduced by Thomas Bayes in 1763 in An_Essay_towards_solving_a_Problem_in_the_Doctrine_of_Chances by Thomas Bayes.
- The mathematical foundations, at the very bottom, of Probability relies on measure theory developed earlier by the likes of Emile Borel and Henri Lebesgue and many others.
- According to a remark attributed to Marc Kac: **probability theory is measure theory with a soul**
- Probability is the new kid on the mathematics block (after Geometry, Algebra and Analysis) but of course the **coolest kid**.
- Probability theory provides the mathematical foundations and the language for Statistics, Ergodic Theory, Statistical Mechanics, Information Theory, and (a big part of) Machine Learning.

1.1 σ-algebras

The state space or sample space Ω is some abstract space and we denote subsets of Ω by capital letters $A, B, \cdots$. We use the language of set theory

$$A^c = \Omega \setminus A, \quad A \cap B, \quad A \cup B, \quad A \setminus B, \quad A \supset B, \quad \emptyset,$$

Definition 1.1 (partition) A **partition of** Ω is a collection of sets, $(A_i)_{i=1}^{\infty}$, such that

- $A_i \cap A_j = \emptyset$ (pairwise disjoint) for all $i \neq j$.
- $\bigcup_{i=1}^{\infty} A_i = \Omega$

Intuition:

Ω = collection of all possible outcomes of an experiment

A = event = collection of all outcomes compatible with the event A

Write 2^{Ω} for the collection of all subsets of Ω and we denote by calligraphic letter $\mathcal{A}, \mathcal{E}, \mathcal{F}, \ldots$ collections of subsets of Ω (that is subsets of 2^{Ω}).

We introduce natural collection of subsets

Definition 1.2 (algebra) A collection of set $\mathcal{E}$ is an **algebra** if

- $\emptyset \in \mathcal{E}$ and $\Omega \in \mathcal{E}$.
- $\mathcal{E}$ is closed under complement: $A \in \mathcal{E} \implies A^c \in \mathcal{E}$.
- $\mathcal{E}$ is closed under finite union and intersection: $A_1, \cdots, A_n \in \mathcal{E} \implies \begin{array}{l} \bigcup_{i=1}^n A_i \in \mathcal{E} \\ \bigcap_{i=1}^n A_i \in \mathcal{E} \end{array}.$

Definition 1.3 (σ-algebra) A collection of set $\mathcal{A}$ is an σ**-algebra** if

- $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$.
- $\mathcal{A}$ is closed under complement: $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
- $\mathcal{A}$ is closed under *countable* union and intersection: $A_1, A_2, \cdots, \in \mathcal{A} \implies \begin{array}{l} \bigcup_{i=1}^\infty A_i \in \mathcal{A} \\ \bigcap_{i=1}^\infty A_i \in \mathcal{A} \end{array}.$

Examples: $\mathcal{A} = \{\emptyset, \Omega\}, \mathcal{A} = 2^\Omega, \mathcal{A} = \{\emptyset, A, A^c, \Omega\}$

Definition 1.4 (σ-algebra generated by $\mathcal{C}$) If $\mathcal{C}$ is a collection of subsets of Ω, the **σ-algebra generated by $\mathcal{C}$**, denoted by $\sigma(\mathcal{C})$, is the smallest σ-algebra containing $\mathcal{C}$.

Remark: $\sigma(\mathcal{C})$ always exists since, alternatively, you can think of it as the intersection of all σ-algebras which contains $\mathcal{C}$. Indeed we always have $\mathcal{C} \subset 2^{\Omega}$ and arbitrary intersections of σ-algebras are σ-algebras (see Homework for more details).

If the space Ω has a topology (e.g $\Omega = \mathbb{R}$) then it natural to pick a σ-algebra compatible with open set.

Definition 1.5 (The Borel σ-algebra) If $\Omega = \mathbb{R}$ the **Borel σ-algebras $\mathcal{B}$** is the σ-algebra generated by the collection of all open sets (or, equivalently, by the closed sets).
The sets in the Borel σ-algebras $\mathcal{B}$ are called **Borel sets**.

Remark: The Borel σ-algebra of $\mathbb{R}$ is strictly smaller than $2^{\mathbb{R}}$ (see Math 623 for a proof (existence of non-measurable sets)). The σ-algebra $2^{\mathbb{R}}$ is too big to useful!

An important characterization of the Borel σ-algebra of $\mathbb{R}$ is the following. It will play a big role to characterize random variables

Theorem 1.1 The Borel σ-algebra of $\mathbb{R}$ is generated by the countable collection of intervals of the form

$$(-\infty, a] \quad \text{where} \quad a \in \mathbb{Q}$$

Proof. Let $\mathcal{C}$ be the collection of all open set.

- If A is an open set, then A can be written as the union of disjoint intervals $A = \cup_i (a_i, b_i)$.
- An interval (a, b) can be written as $(a, b) = (-\infty, b) \setminus (-\infty, a]$.
- Taking a sequence $b_n \in \mathbb{Q}$ with $b_n \nearrow b$ we have $(-\infty, b) = \bigcup_n (-\infty, b_n]$.
- Taking a sequence $a_n \in \mathbb{Q}$ with $a_n \searrow a$ we have $(-\infty, a] = \bigcap_n (-\infty, a_n]$.

If $\mathcal{D}$ is the collection of interval $(-\infty, a]$, with $a \in \mathbb{Q}$ then we have shown that $\mathcal{C} \subset \sigma(\mathcal{D})$ and therefore $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$. But since $\mathcal{D}$ consists of closed sets (which are Borel sets) we have $\mathcal{D} \subset \sigma(\mathcal{C})$ and thus $\sigma(\mathcal{D}) \subset \sigma(\mathcal{C})$. □

1.2 Axioms of Probability

Two simple axioms underlie probability theory.

Definition 1.6 (Probability measure) A **probability measure defined a on a σ-algebra $\mathcal{A}$** is a function

$$P : \mathcal{A} \to [0, 1]$$

with the following properties

- $P(\emptyset) = 0$ and $P(\Omega) = 1$
- (**Countable additivity**) For any **countable** collection of pairwise disjoint sets A_i (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$) we have

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

A **Probability space** $(\Omega, \mathcal{A}, P)$ consists of the set Ω a σ-algebra $\mathcal{A}$ and a probability measure P.

Remark: As we shall see, requiring finite additivity $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ for finite n would not be sufficient.

1.3 Consequence of the axioms

Theorem 1.2 (Elementary properties or probability measures)

- Finite additivity: $A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$
- Monotonicity: $A \subset B \implies P(A) \leq P(B)$
- Complement: $P(A^c) = 1 - P(A)$
- Union rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Inclusion-exclusion:

$$P\left(\bigcup_{i=1}^{N} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \\ + \cdots + (-1)^{N+1} P(A_1 \cap \cdots A_N)$$

Proof. Homework

1.4 Countable additivity and limits

Monotone limits for sets

$$A_n \searrow A \quad \text{means} \quad A_1 \supset A_2 \supset A_3 \;\text{ and }\; A = \bigcap_{n=1}^{\infty} A_n$$

$$A_n \nearrow A \quad \text{means} \quad A_1 \subset A_2 \subset A_3 \;\text{ and }\; A = \bigcup_{n=1}^{\infty} A_n$$

Theorem 1.3 (Sequential continuity)

- Countable additivity implies sequential continuity that is, if $A_1, A_2, \cdots \in \mathcal{A}$

$$\begin{aligned} A_n \searrow A &\implies P(A_n) \searrow P(A) \\ A_n \nearrow A &\implies P(A_n) \nearrow P(A) \end{aligned}$$

- Finite additivity + sequential continuity implies countable additivity

Proof. Let us assume first countable additivity. If $A_1 \subset A_2 \subset A_3 \subset \cdots$ define the disjoints set $B_1 = A_1, B_2 = A_2 \setminus A_1, B_3 = A_3 \setminus A_2, \cdots$. Then for any N, $B_1 \cup \cdots \cup B_N = A_N$ and $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$. By countable additivity

$$P(A) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_i) = \lim_{N\to\infty} \sum_{n=1}^{N} P(B_i) = \lim_{N\to\infty} P(A_N).$$

If $A_1 \supset A_2 \supset A_3 \supset \cdots$ then taking complement we have $A_1^c \subset A_2^c \subset A_3^c \subset \cdots$ and using the above and de Morgan's law

$$P(A) = P\left(\bigcap_{n=1}^{\infty} A_n\right) = 1 - P\left(\bigcup_{n=1}^{\infty} A_n^c\right) = 1 - \lim_{N\to\infty} P\left(\bigcup_{n=1}^{N} A_n^c\right) = 1 - \lim_{N\to\infty} P\left(A_N^c\right) = \lim_{N\to\infty} P(A_N)$$

For the converse statement suppose A_i are pairwise disjoint and set $B_N = \bigcup_{n=1}^{N} A_n$. Then $B_N \nearrow B = \bigcup_{n=1}^{\infty} A_n$. Using finite additivity and continuity we find

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P(B) = \lim_{N\to\infty} P(B_N) = \lim_{N\to\infty} \sum_{n=1}^{N} P(A_n) = \sum_{n=1}^{\infty} P(A_n)$$

which prove countable additivity.

□

1.5 More on limits of sets

limsup and liminf of sets: For an arbitrary collection of sets A_n we define the limsup and liminf by

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\ge n} A_m = \{\omega \in \Omega\,;\omega \in A_n \text{for infinitely many } n\}$$

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m\ge n} A_m = \{\omega \in \Omega\,;\omega \in A_n \text{for all but finitely many } n\}$$

The fact that the two definitions coincide requires some thoughts (see homework for more details). For example if $\omega \in \bigcap_{n=1}^{\infty} \bigcup_{m\ge n} A_m$ then $\omega \in \bigcup_{m\ge n} A_m$ for every n. Taking $n=1$ we find that there exist some $k_1 \ge 1$ such that $\omega \in A_{k_1}$. Taking next $n = k_1+1$ we see that there exists $k_2 > k_1$ such that $\omega \in A_{k_2}$, and so on. Therefore ω belongs to infinitely many A_n.

Characteristic function of a set A: $1_A(\omega) = \left\{ \begin{array}{ll} 1 & \omega \in A \\ 0 & \omega \notin A \end{array} \right.$

Limits of sets: We say that A_n converge to A if $\lim_n 1_{A_n}(\omega) = 1_A(\omega)$ for all ω.

Theorem 1.4 If A_n converges to A then $\lim_{n\to\infty} P(A_n) = P(A)$.

Proof. If $\lim_n 1_{A_n}(\omega) = 1_A(\omega)$ for some ω then the sequence is either 1 (if $\omega \in A$) or 0 (if $\omega \notin A$) for all but finitely many n. Therefore if A_n converges to A this means

$$A = \limsup_n A_n = \liminf_n A_n.$$

Set $B_n = \bigcap_{m\ge n} A_m$ and $C_n = \bigcup_{m\ge n} A_m$. Then we have $B_n \subset A_n \subset C_n$ and thus, by monotonicity

$$P(B_n) \le P(A_n) \le P(C_n).$$

Since the sequence B_n is increasing to $\liminf_n A_n$ and and and the sequence C_n is decreasing to $\limsup_n A_n$ and $A = \limsup_n A_n = \liminf_n A_n$ we have $\lim_{n\to\infty} P(A_n) = P(A)$. $\quad\square$.

1.6 Conditional Probability and Independence

The intuition behing conditional probabilities is as follows. Suppose you observe that the vent B has occurred (e.g. “it rained yesterday”). How does that influence the probability of another event A (e.g. there was wind yesterday or it is raining today)?

For $\omega \in A$ to be a possible outcome we must have $\omega \in A \cap B$. So only events of the form $A \cap B$ are relevant. Normalizing the probabilities leads to the following

Definition 1.7 Given $B \in \mathcal{A}$ with $P(B) > 0$, **the conditional probability of A given B** is defined by

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

If the occurence of B does **not** influence the probability that A occurs, that is if $P(A|B) = P(A)$ then we will call theses events independent. This means $P(A|B) = P(A) \iff P(A \cap B) = P(A)P(B) \iff P(B|A) = P(A)$.

Definition 1.8 (Independent events)

- Two events A and B are independent if $P(A \cap B) = P(A)P(B)$
- A collection of events $(A_i)_{i \in I}$ (I possibly be infinite) if for any *finite* $J \subset I$, $P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$

Theorem 1.5 (Properties of conditional probability)

- **Independence**: If A and B are independent so are A and B^c, A^c and B, and A^c and B^c.
- **Product rule**: For any events $A_1, A_2, \cdots , A_n$

  $$P(A_1 \cap \cdots \cap A_n) = P(A_1)\, P(A_2|A_1)\, P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

- **Conditioning**: If $(E_n))n$ is a finite or countable partition of Ω then

  $$P(A) = \sum_n P(A|E_n)P(E_n).$$

- **Bayes rule**: If $(E_n)_n$ is a finite or countable partition of Ω then

  $$P(E_m|A) = \frac{P(A|E_m)P(E_m)}{\sum_n P(A|E_n)P(E_n)}.$$

Proof of Theorem 1.5.

- **Independence**: $P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = (1 - P(B))P(A) = P(A)P(B^c)$.
- **Product rule**: Use repeatedly $P(A \cap B) = P(B|A)P(A)$ which is just the definition. Then

$$P(A_1 \cap \cdots \cap A_n) = P(A_n | A_1 \cap A_{n-1})P(A_1 \cap \cdots \cap A_{n-1}) = \cdots$$

- **Conditioning**: Use countable additivity

$$P(A) = P(A \cap \cup_n E_n) = P(\cup_n (A \cap E_n)) = \sum_n P(A \cap E_n) = \sum_n P(A|E_n)P(E_n).$$

- **Bayes**: Use the defintion and conditioning:

$$P(E_m|A) = \frac{P(A \cap E_m)}{P(A)} = \frac{P(A|E_m)P(E_m)}{P(A)} = \frac{P(A|E_m)P(E_m)}{\sum_n P(A|E_n)P(E_n)}.$$

1.7 Conditional probability model

Given a probability P and any (fixed) event A we can build a new probability

Theorem 1.6 (Conditional probability model) Given a probability on P on a σ-algebra $\mathcal{A}$ and a (fixed) set $B \in \mathcal{A}$ the map

$$A \mapsto P(A|B)$$

defines a probability measure on $\mathcal{A}$, *the conditonal probabilty given the event* B

Proof. Easy to verify the axioms.

One can then extend the concepts of independence to conditional independence.

Definition 1.9 The events A_1 **and** A_2 **are independent condtionally on the event** B if

$$P(A_1 \cap A_1|B) = P(A_1|B)P(A_2|B)$$

This concept is important in Markov chain, Markov and Bayesian networks (graphical models) (see e.g. https://en.wikipedia.org/wiki/Graphical_model and https://en.wikipedia.org/wiki/Bayesian_network, and later examples in the class).

Example Watches from companies A are defective with probbailities $\frac{1}{100}$ and watches from companies B are defective with probailities $\frac{2}{100}$. Pick two random watches. If the first watch works, what is the probability that the second watch work?

For the factory of origin Y we have $P(Y=A)=P(Y=B)=\frac{1}{2}$.

If we describe the state of the watches by X_1 and X_2 with $X_i=1$ is the watches works and if all watches are independently defective we have *conditional independence*

$$P(X_1=1,X_2=1|Y=A)=P(X_1=1|Y=A)P(X_2=1|Y=A)$$

and similarly for $Y=B$.

We want to compute $P(X_2=1|X_1=1)=\frac{P(X_2=1,X_1=1)}{P(X_1=1)}$

1.8 Take-home messages

- σ-algebra would not be necessary if we work only with discrete sample space.

- But to describe sample space like $\mathbb{R}$ or a countable collection of discrete models (think coin flip) they are necessary (recall that $2^{\mathbb{N}}$ has the same cardinality as $\mathbb{R}$). In the dark corners of real numbers various monsters are lurking that need to be tamed (and then ignored).

- At a deeper and more interesting level σ-algebra will occur later in conditional expectation, martingales and stochastic processes, they will be use as **information-theoretic tool** and will describe sets of questions or inquiries you can perform on a model.

1.9 Homework problems

Exercise 1.1

1. Suppose $\{\mathcal{A}_j\}_{j\in J}$ is an arbitrary collection of σ-algebras (J does not need to be countable). Show that the intersection $\cap_{j\in J}\mathcal{A}_j$ is a σ-algebra.
2. Is the union of a σ-algebras a σ-algebra? Prove or disprove.

Exercise 1.2 Suppose $\mathcal{A}$ is a σ-algebra and let $B \in \mathcal{A}$ be some set. Define $\mathcal{B} = \{A \cap B\,;\, A \in \mathcal{A}\}$.

1. Show that $\mathcal{B}$ is a σ-algebra of subsets of B.
2. Interpret the conditional probability $P(A|B)$ as a probability on the σ-algebra $\mathcal{B}$.

Exercise 1.3 Suppose $f : E \to F$ is an arbitrary map. For $B \subset F$ let $f^{-1}(B) = \{x \in E\,;\, f(x) \in B\}$ the inverse image of the set B.

1. Prove that $1_B(f(x)) = 1_{f^{-1}(B)}(x)$
2. Prove that if $\mathcal{A}$ is a σ-algebra of subsets of F then $f^{-1}(\mathcal{A}) = \{f^{-1}(B)\,;\, B \in \mathcal{A}\}$ is a σ-algebra of subsets of E

Exercise 1.4

- Prove the properties of probability measures given in Theorem 1.2
- Prove also the so-called Bonferronni inequality

$$\sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) \leq P(\cup_{i=}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$

Exercise 1.5 (Markov property and conditional independence) Three events A_1, A_2, A_3 satisfy the **Markov property** if we have $P(A_3|A_1 \cap A_2) = P(A_3|A_2)$. Show that A_1, A_2, A_3 satisfy the Markov property if and only if A_3 and A_1 are independent conditionally on A_2.

Exercise 1.6 (limsup and liminf for sequences) This exercise serves as a reminder from your analysis class.

1. Suppose $\{x_n\}$ is a bounded sequence of real numbers then define the limsup by

$$\limsup_{n} x_n = \lim_{n\to\infty} \sup_{k\geq n} x_k$$

   Show that $\limsup_n x_n$ is an accumulation point of the sequence x_n (i.e. there exists a convergent subsequence of x_n which converges to $\limsup_n x_n$) and that this is actually the greatest such accumulation point.

2. What is the corresponding statement for $\liminf$ (just state it, no proof)?

Exercise 1.7 (limsup and liminf of sets)

- Prove that two definitions for limsup and liminf of sets given in Section 1.5 are equivalent.
- Show that $\limsup_n A_n^c = (\liminf_n A_n)^c$
- I found online this pretty instructive illustration of limsup and liminf (slightly edited).

  A tech company is closing down and firing the set Ω of all his employees who become beggars and have to live on the street (in this story people never die). A local church decides to start to give out free food to them every day. On the n^{th} day A_n is the subset of fired employees who show up at church to get fed. There are three categories of people:

  1. Some of the people eventually get a new job and never show up at the church again.
  2. Others are too proud and try be seen around all the time, but they need to eat so they always come back eventually.
  3. Lastly there are the people who after trying everything else, eventually give up start to get their food from the church each day.

  Express the categories of people in 1., 2., 3. in terms of limsup and liminf? What needs to happen for $\lim_n A_n$ to exist?

Exercise 1.8 (Random permutations) Consider the following lottery game: On N lottery tickets (visibly) numbered 1 to N and the numbers 1 to N are randomly assigned hidden under the scratch pad. (Each number comes up exactly once, in other words one is picking a random permutation of the N numbers). If the randomly assigned number match the visible numbers then you win. This lottery has the remarkable property that the probability that nobody wins is essentially independent of N.

Consider the events $A_i = \{\text{ticket } i \text{ is a winner}\}$.

- Compute $P(A_i)$ and $P(A_i \cap A_j)$ for $i \neq j$ using the product rule.
- Use the inclusion-exclusion formula to compute the probability that there is at least one winner.
- Show that if N is even moderately big (maybe $N \geq 5$ or 6 will do) this probability that nobody wins is for all practical purpose independent of N and nearly equal to $\frac{1}{e}$.
- Compute now the probability that there are exactly k winners in that games. *Hint*: There are $\binom{n}{k}$ ways to have k winners and to have exactly k winner, we must have k matches and no matches among the $n-k$ others.

Exercise 1.9 (A series of liars) Consider a sequence of n “liars” $L_1, \cdots, L_n$. The first liar L_1 receives information about the occurrence of some event in the form “yes or no”, and transmits it to L_2, who transmits it to L_3, etc. . . Each liar transmits what he hears with probability $0 < p < 1$ and the contrary with probability $q = 1-p$. The decision of lying or not is made independently by each liar. What is the probability x_n of obtaining the correct information from L_n? What is the limit of x_n as n increases to infinity? *Hint*: For example you may want to define a random variable $X_i = \pm 1$ to describe the state of the liar (1=truthful, -1=lying).

2 Discrete random variables

2.1 Probabilities on countable state spaces

- If $\Omega = (\omega_1, \omega_2, \dots)$ is countable sample space take as σ-algebra $\mathcal{A} = 2^\Omega$ to the collection of all susbets of Ω.
- Specifying a probability on $\mathcal{A}$ is equivalent to choosing a collection of numbers $p_n = P(\{\omega_n\})$ which the probability of the event $\{\omega_n\}$ with

$$p_n \ge 0 \quad \text{and} \quad \sum_i p_n = 1$$

and then for any $A \subset \Omega$ we have

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

- Countable additivity reduces to the fact that for absolutely convergent series we can freely interchange of summations

$$P(\cup_i A_i) = \sum_{\omega \in \cup_i A_i} p(\omega) = \sum_i \sum_{\omega \in A_i} p(\omega) = \sum_i P(A_i)$$

if the A_i are pairwise disjoint.

We recall some standard probability models

Definition 2.1 (Poisson) A **Poisson distribution** with parameter $\lambda > 0$ is a probability distribution on $\{0, 1, 2, 3, \cdots\}$ with

$$p_n = e^{-\lambda} \frac{\lambda^n}{n!}.$$

The Poisson distribution (see https://en.wikipedia.org/wiki/Poisson_distribution) is ubiquitous, in particular because of its relation with the binomial distribution (sometimes called *The Law of small numbers* see Theorem 2.1). Typical examples are the number of typos in a page, the number of earthquakes hitting a region, the number of radiactive atoms decay, etc...

Definition 2.2 (Pareto or Zeta distribution) A **Pareto distribution** with decay rate $\alpha > 0$ is a probability distribution on $\{1, 2, 3, \cdots\}$ with

$$p_n = \frac{1}{\zeta(\alpha+1)} \frac{1}{n^{\alpha+1}} \quad \text{where} \quad \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \text{ Riemann zeta function}$$

The Pareto distribution is a model with polynomial tails (see https://en.wikipedia.org/wiki/Zeta_distribution). It is widely used in economics where polynomial tails often occurs. Classical example is the the wealth distribution in a population, or the size of cities, or the number of casualties in wars. See the Pareto principle ("20% of the population controls 80% of the wealth") and the examples in this paper

Definition 2.3 (Independent Bernoulli trials) The model is characterized by an integer n (= number of trials) and by a number $0 \leq p \leq 1$ (=the probability of success in each trials). We assume that the trials succeed or fails independently of each other. The state space is $\Omega = \{0,1\}^n$ and we write $\omega = (\omega_1, \cdots, \omega_n)$ with

$$\omega_i = \begin{cases} 0 & \text{if } i^{th} \text{ trial fails} \\ 1 & \text{if } i^{th} \text{ trial succeeds} \end{cases}$$

If we set $\|\omega\|_1 = \sum_{i=1}^n \omega_i$ which counts the number of 1's in ω then

$$P(\{\omega\}) = p^{\|\omega\|_1}(1-p)^{n-\|\omega\|_1}$$

Definition 2.4 (Binomial random variable) The **binomial distribution** describe the the number of success in n Bernoulli trials. It is a probability distribution on $\{0, 1, \cdots, n\}$ with

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways k successes can occur among n trials.

Definition 2.5 (Geometric distribution) The **geometric distribution** describes when the first succesful trial occurs in a series of Bernoulli trials. It is a probability distribution on $\{1, 2, 3 \cdots, \}$ with

$$p_n = (1-p)^{n-1} p$$

since we have unsuccesful trial before the first succesful ones.

Definition 2.6 (Negative binomial or Pascal's distribution) The **negative binomial distribution** describes when the k^{th} successful trial occurs in a series of Bernoulli trials. It is a probability distribution on $\{k, k+1, k+2 \cdots\}$ with

$$p_n = \binom{n-1}{k-1} (1-p)^{n-k} p^k$$

Sometimes the negative binomial is defined slightly differently and counts the number of failure until the k^{th} success occurs.

2.2 Poisson approximation

If the number of trial $n \gg 1$ is large and the probability of success if small $p \ll 1$, then the binomial distribution can be approximated by a Poisson distribution with parameter $\lambda = np$. A formal mathematical result is the following (proof in the homework)

Theorem 2.1 (The Law of small numbers) Suppose the probability of success, p_n, varies with n in such a way that $\lim_{n\to\infty} np_n = \lambda$ (take $p_n = \frac{\lambda}{n}$) then we have

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

Example: Birthday problem

- There are N students in class. Event $A =$“at least two students share the same birthday”.
- $p =$ probability that **one pair** of students have the same birthday? It is $p = \frac{1}{365}$.
- Number of trials here is the number of pair of students $n = \binom{N}{2} = \frac{N(N-1)}{2}$. Note:T rials are weakly dependent here.
- Poisson approximation $\lambda = np = \binom{N}{2}\frac{1}{365}$ and so $P(A^c) \approx e^{-\binom{N}{2}\frac{1}{365}}$.
- Exact value $P(\text{no pair share the same birthday}) = \frac{365}{365}\frac{364}{365}\cdots\frac{365-N+1}{365}$

See more in the nice book Understanding Probability by Henk Tijms

▼ Code

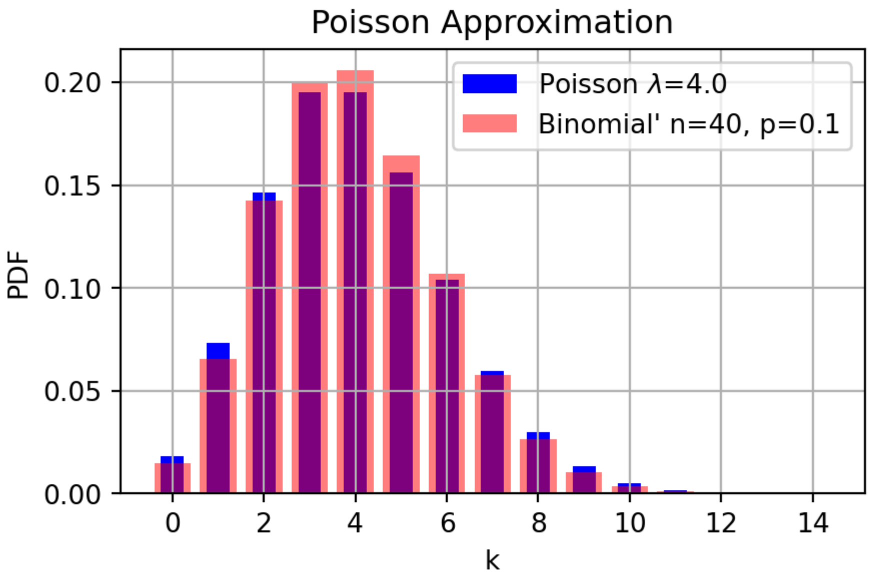
```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import poisson
from scipy.stats import binom

n = 40    # Number of trials
p = 0.1   # Probability of success
lam = n*p # paramter of the Poisson distri

# Values to evaluate the PDF at
x = np.arange(0, 15)

# Calculate the PDF of the Poisson distrib
pdfP = poisson.pmf(x, mu=lam)
pdfB = binom.pmf(x, n, p)

# Plot the PDF
plt.figure(figsize=(5,3))
plt.bar(x, pdfP, color='blue', width=0.5,
plt.bar(x, pdfB, color='red', alpha=0.5, l
plt.title('Poisson Approximation')
plt.xlabel('k')
plt.ylabel('PDF')
plt.legend()
plt.grid(True)
plt.show()
```



2.3 Random variable on countable state spaces

You should think as a random variable as associating a quantity (a number or maybe a vector) (hence the name “variable”) to an each outcomes $\omega \in \Omega$. The probability defined on Ω implies that this variable is random, hence the name “random variable”.

Definition 2.7 (Random Variables on discrete state space)

Suppose E is countable set with $E \subset \mathbb{R}$ (e.g $E = \mathbb{N}$ or $\mathbb{Z}$). A map

$$X : \Omega \to E$$

is called a **discrete random variable (RV)**. For example if Ω is itself countable then the image of Ω by X will always be discrete.

The distribution of X, also called **the law of X** is a probability measure on E called P^X, induced from the probability P on Ω and defined as follows: if $A \subset E$ then

$$P^X(A) = P(\{\omega\,;\, X(\omega) \in A\}) = P(X^{-1}(A)) = P(X \in A).$$

It particular for $j \in E$ we have $p_j^X = P(X = j)$.

2.4 Sampling with or without replacement

- A urn with N balls, b blue balls and $r = N - b$ red balls. We select n balls out of the N balls either with or without replacing balls after choosing them.
- Sample space $\Omega = \{0,1\}^n$ with $\omega = (\omega_1, \cdots, \omega_n)$ where $\omega_i = 0$ means getting a red ball and $\omega_i = 1$ means getting a blue ball on the i^{th} selection.
- Random variable $X(\omega) = \|\omega\|_1 = \sum_i \omega_i$ which is the number of blue balls and take values in $\{0, 1, \cdots, n\}$

Definition 2.8 (Sampling distributions: binomial and hypergeometric)

- If we **sample with replacement**, the probability to get a blue ball is always $p = \frac{b}{b+r}$ and we get a **binomial distribution**

$$p_k^X = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- If we **sample without replacement**, then we get a **hypergeometric distribution**

$$p_k^X = P(X = k) = \frac{\binom{b}{k}\binom{N-b}{n-k}}{\binom{N}{n}}$$

2.5 Expectation

Definition 2.9 (functions of random variables) If X is a discrete random variable taking value in E and $g : E \to F$ is a map then $Y = g(X)$ is a discrete random variables taking values in F. The distribution of Y, P^Y is such that

$$P_j^Y = P(Y = j) = P(\omega : Y(\omega) = j) = P(\omega : g(X(\omega)) = j) = P^X(i; g(i) = j)$$

Definition 2.10 If $X : \Omega \to E$ is a discrete random variable taking value in a countable set E and $g : E \to \mathbb{R}$ a function we define the expectation of **expectation of a random variable** $g(X)$

$$E[g(X)] = \sum_{i \in E} g(i) P(X = i)$$

The expectation is well defined if $g \geq 0$ (in which case the sum could possible infinite) or if the series is absolutely convergent $\sum_i |g(i)| P(X = i) < \infty$. In the latter case we say that $g(X)$ is **integrable**.

The following properties are easy to check using the corresponding properties of sum

Theorem 2.2 (Elementary properties of expectation)

1. **Linearity of expectation**: If $\lambda_1, \lambda_2 \in \mathbb{R}$ then $E[\lambda_1 g_1(X) + \lambda_2 g_2(X)] = \lambda_1 E[g_1(X)] + \lambda_2 E[g_2(X)]$
2. **Monotonicity**: If $g_1(x) \leq g_2(x)$ for all $x \in E$ then $E[g_1(X)] \leq E[g_2(X)]$
3. **Triangular inequality**: $|E[g(X)]| \leq E[|g(X)|]$.

- We can also write the expectation using the distribution of $Y = g(X)$ that is $E[g(X)] = E[Y] = \sum_{j \in F} jP(Y = j)$ (check this).
- We denote by $L^1 = \{X : \Omega \to \mathbb{R}, \sum_i |X(\omega_i)|P(X = i) \leq \infty\}$ the set of random variable with a finite expectation $E[X]$.
- L^1 is a vector space (see more on this later).

The following formula is useful (see later for a general version using Fubini-Tomelli)

Theorem 2.3 If X takes values in $\mathbb{N}$ then

$$E[X] = \sum_{n=1} P(X \geq n) = \sum_{n>0} P(X > n)$$

Proof.

$$\begin{aligned}
E[X] &= P(X=1) &&+2P(X=2) + 3P(X=3) &&+\cdots \\
&= P(X=1) &&+P(X=2) + P(X=3) &&+\cdots \\
& &&+P(X=2) + P(X=3) &&+\cdots \\
& && \qquad\qquad\quad\; + P(X=3) &&+\cdots
\end{aligned}$$

□

2.6 Examples

1. **Poisson RV:** $X : \Omega \to \mathbb{N}$ and $P(X = j) = e^{-\lambda}\frac{\lambda^j}{j!}$. Then

$$E[X] = \sum_{j=0}^{\infty} je^{-\lambda}\frac{\lambda^j}{j!}\sum_{j=1}^{\infty} e^{-\lambda}\frac{\lambda^j}{(j-1)!} = \lambda\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda$$

2. **Bernoulli (Indicator) Random variable:** Given an event A define $X_A : \omega \to \{0, 1\}$ by

$$X_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

Then

$$E[X_A] = 1P(A) + 0(1 - P(A)) = P(A)$$

3. **Binomial RV:** The Binomial RV Z with parameters (n, p) can be written as sum of n Bernoulli RV $Z = X_1 + \cdots + X_n$ where $X_i = 1$ if a success occurs. From the linarity of exectation and the previous example we see immediately that $E[X] = np$.

3. **Hypergeometric RV:** If we sample n balls out b blue balls and r red balls. The we can write the number of blue balls as $Z = X_1 + \cdots + X_n$ as for the binomial but this time the random variables are not independent. Clearly X_1 is Bernoulli with paramter $p = \frac{b}{b+r}$. What about X_2? We can argue by conditioning that

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 0)P(X_1 = 0) \\ &= \frac{b-1}{b+r-1}\frac{b}{b+r} + \frac{b}{b+r-1}\frac{r}{b+r} = \frac{b}{b+r} \end{aligned}$$

and so X_2 has the same distribution as X_1 (they are not independent). Rather than computing on it is easier to regroup and argue that immediately that, by symmetry, all X_i must have the same distribution since it does not matter in which order the balls are drawn. So $E[Z] = n\frac{b}{b+r}$.

4. **Zeta RV:** If X is has zeta distribution with parameter $\alpha > 0$ then

$$E[X] = \sum_{n=1}^{\infty} n \frac{1}{\zeta(\alpha+1)} \frac{1}{n^{\alpha+1}} = \frac{1}{\zeta(\alpha+1)} \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} = \begin{cases} \frac{\zeta(\alpha)}{\zeta(\alpha+1)} & \alpha > 1 \\ +\infty & 0 < \alpha \leq 1 \end{cases}$$

5. **Geometric RV**: We could use Theorem 2.3 and note that

$$P(X \geq n) = q^{n-1}p + q^n p + \cdots = q^{n-1}p(1+q+q^2+\cdots) = q^{n-1}p\frac{1}{1-q} = q^{n-1}$$

   and therefore $E[X] = \sum_{n=1}^{\infty} P(X \geq n) = \sum_{n=1}^{\infty} q^{n-1} = \frac{1}{1-q} = \frac{1}{p}$.

6. **Negative binomial RV**: Since a negative binomial with paramter k is the sum of k geometric RV we have $E[X] = \frac{k}{p}$.

7. **Uniform distribution** on $\Omega = \{1, 2, \cdots, N\}$: we have $p_j = \frac{1}{N}$ for every j and thus

$$E[X] = \frac{1}{N}(1+2+\cdots+N) = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$$

8. **A RV without expectation**: Suppose $p_n = \frac{3}{\pi^2}\frac{1}{n^2}$ for $n \in \mathbb{Z} \setminus \{0\}$ and $p_0 = 0$ and for a suitable normalization c_n. Then the series for the expecation

$$\sum_{n=-\infty}^{\infty} np_n$$

   is undefined because $\sum_{n>0} np_n = \infty$ and $\sum_{n<0} np_n = -\infty$.

2.7 Independence of Random Variables

We can easily lift the notion of independence from events ($P(AB) = P(A) = P(B)$) to random variables.

Definition 2.11 (Independence of random variables)

1. Two discrete random variables X and Y taking values in E and F are independent if $P(X = i, Y = j) = P(X = i)P(Y = j)$ for all $i \in E, j \in F$.
2. The random variables $X_1, \cdots, X_n$ taking values in $E_1, \cdots, E_n$ are independent if $P(X_1 = i_1, \cdots, X_n = i_n) = P(X_1 = i_1) \cdots P(X_n = i_n)$ for all $i_1 \in E_1, \cdots i_n \in E_n \in F$.
3. The collection of random variables $(X_n)_{n \geq 1}$ is independent if the RVs $X_{i_1}, \cdots, X_{i_k}$ are independent for any k and $I_1, \cdots, i_k$.

It is not hard to check that

Theorem 2.4 If X and Y are independent discrete RVs and $h : E \to \mathbb{R}$ and $g : F \to \mathbb{R}$ are such that $g(X)$ and $h(Y)$ are either non-negative or integrable then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]. \tag{2.1}$$

Conversely if Equation 2.1 holds for all h, g (non-negative or integrable) then X and Y are independent.

2.8 Variance of a RV

Definition 2.12 (Variance of a random variables) If X is a discrete random variable with finite mean $\mu = E[X]$ then the variance of X, denoted by $V[X]$ is

$$V[X] = E[(X-\mu)^2] = \sum_j (j-\mu)^2 P(X=j)$$

An alternative formula is

$$E[X^2] - E[X]^2$$

- If $E[X^2]$ is finite then X has finite mean and thus finite variance because of the elementary inequality $|x| \leq 1 + x^2$.
- If X and Y are independent random variables then

$$V[X+Y] = V[X] + V[Y]$$

- Suppose $X_1, \cdots, X_n$ are independent random variables with common mean $E[X_i] = \mu$ and common variance $V[X_i] = \sigma^2$. Let $S_n = X_1 + \cdots X_n$ be their sum and $\frac{S_n}{n}$ their average. Then

$$E\left[\frac{S_n}{n}\right] = \mu \quad \text{and} \quad V\left[\frac{S_n}{n}\right] = \frac{\sigma^2}{n}$$

2.9 Conditional expectation

Conditioning on event, random variables, and eventually σ-algebras is at the center of probability theory. We start with afew examples for discrete random variables and we will extend to the general case later on.

Definition 2.13 (conditional expectation with respect to an event) If B is a (fixed) event with $P(B) > 0$ and $g(X)$ is integrable (or non-negative) then the **conditional expectation of $g(X)$ conditioned on the event B** is given by

$$E[g(X)|B] = \sum_i g(i)P(X = i|B) \tag{2.2}$$

Note that this is simply the expectation with respect to the probability $P(\cdot|B)$ which we have seen in Theorem 1.6 is a probability in its own right.

Moreover since $P(X = i|B) \le \frac{P(X=i)}{P(B)}$ the sum in Equation 2.2 well defined if $g(X)$ is integrable wioth respect to P.

We can push this a little more by considering two discrete RVs X and Y and condition on the event $Y = j$.

Definition 2.14 (conditional expectation with respect to a random variable) If $g(X,Y)$ is integrable or non-negative then **conditional expectation of** $g(X,Y)$ **conditioned on the event** $Y=j$ is given by

$$E[g(X,Y)|Y=j] = \sum_i g(i,j)P(X=i|Y=j) \tag{2.3}$$

Note that the right hand side of Equation 2.3 is a function of j, say $h(j)$ and thus we can define a random variable $h(Y)$ which is called the **conditional expectation of** $g(X,Y)$ **conditioned on** Y and is denoted by

$$E[g(X,Y)|Y].$$

Example Suppose X_1 and X_2 are two independent binomial RV each with paramters (N,p). We show that $E[X_1|X_1+X_2] = \frac{X_1+X_2}{2}$. We have

$$\begin{aligned} P(X_1=k|X_1+X_2=n) &= \frac{P(X_1=k)P(X_2=n-k)}{P(X_1+X_2=n)} = \frac{\binom{N}{k}p^k(1-p)^{N-k}\binom{N}{n-k}p^{n-k}(1-p)^{N-(n-k)}}{\binom{2N}{n}p^n(1-p)^{2N-n}} \\ &= \frac{\binom{N}{k}\binom{N}{n-k}}{\binom{2N}{n}} \end{aligned}$$

and thus X_1 conditioned on $X_1+X_2=n$ has an hypergeometric distribution when sampling n balls out of the urns with N red balls and N black balls. As we have seen before (or can seen by symmetry) the mean of such distribution is $\frac{n}{2}$. This implies that $E[X_1|X_1+X_2] = \frac{X_1+X_2}{2}$.

Note that we have

$$\begin{aligned} E[h(Y)] &= \sum_j h(j)P(Y=j) = \sum_j \sum_i g(i,j)P(X=i|Y=j)P(y=j) \\ &= \sum_j \sum_i g(i,j)P(X=i,Y=j) = E[g(X,Y)] \end{aligned}$$

To show thus we use the fact (from Math 523) that if the sequences a_{ij} is integrable (that is $\sum_{ij} |a_{ij}| < \infty$) then $\sum_i \sum_j a_{ij} = \sum_j \sum_i a_{ij}$. This is a special case of Fubini Theorem which we will prove later on in full generality.

From this we have obtained the formula for conditioning

Theorem 2.5 If X and Y are discrete random variables and g is integrable then we have

$$E[g(X,Y)] = E[E[g(X,Y)|Y]]$$

Note that is a generalization of the conditioning formula. If we take $Y = 1_B$ and $X = 1_A$ then have for example $P(1_A = 1|1_B = 1) = P(A|B)$. Thus

$$E[1_A|1_B = 1] = P(A|B), \quad [1_A|1_B = 0] = P(A|B^c)$$

2.10 Homework problems

Exercise 2.1 (Poisson approximation) Prove Theorem 2.1

Exercise 2.2 (Applications of the Poisson approximation)

- In the german lottery 6 balls are drawn at random out of 49 balls. As it happened within a year or so the exact same 6 number were drawn twice which seems to be an amazing coincidence since the probability to to draw a sequence of number of is $1/\binom{49}{6} = 1/13,983,816$. Or is it?

  When this event happened suppose the german lottery had been played twice a week for 28 years. Compute the probability that the same numbers appear twice during those 28 years? (If you want you may use a Poisson approximation.)
- Use a Poisson approximation to estimate the probability that at least 3 people in a room of N share the same birthday.
- Revisit the lottery problem of Exercise 1.8. Use a Poisson approximation to estimate the probability that there are exactly k winners and compare with the exact result.

Exercise 2.3 (Hypergeometric and lottery)

- In the powerball at each drawing 5 balls are selected at random among 69 white balls numbered 1 to 69 and 1 ball is seleccted among 26 red balls numbered 1 to 26. A powerball ticket costs \$2 and consists of selected 5+1 numbers. You get a prize if the balls selected math the winning balls see the prizes here.
  - Express the probability of each prize using the hypergeometric distribution.
  - All the prizes are fixed except the jackpot obtained for 5 correct white ball plus the red powerball. You will observe that most people will play only when the jackpot is big enough. As a (rational) mathematician determine the minimal jackpot for which it makes sense to buy a powerball ticket?

Exercise 2.4 (Variance of the hypergeometric random variable) Show that for an hypergeometric random variable Z when you sample n balls out of an urn containing b blue balls\$ and $r = N - b$ balls and Z is the number of blue balls, the variance of Z is given by

$$V[Z] = n\frac{b}{N}\frac{N-b}{N}\frac{N-n}{N-1}$$

Note that if we were sampling the with replacement the term in red would not be present.
Hint: Write $Z = X_1 + \cdots + X_n$ and compute $E[Z^2]$ by expanding the square and using the interchangability of the random variables X_i.

Exercise 2.5 Suppose X is a geometric random variable with success parameter p.

- Show that $P(X > m + n|X > n) = P(X > m)$? What does that mean?
- Show that $E\left[\frac{1}{X}\right] = \log\left(p^{\frac{p}{p-1}}\right)$
- Show that $E\left[X(X-1)(X-2)\cdots(X-r+1)\right] = \frac{r!(1-p)^{r-1}}{p^r}$. Use this to compute $E[X^2]$ and $E[X^3]$.

Exercise 2.6 (More on conditioning) Suppose X and Y are two discrete random variables. Show that

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X]$$

Exercise 2.7 (Conditional Variance) A natural way to define conditional variance is to start with the definition

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

which is simply the variance with respect to the conditional probability $P(\cdot|Y=j)$.
This allows to define $V[X|Y]$ as the function of the random variable Y whose value is equal to $V[X|Y=j]$ when $Y=j$. Prove the **conditional variance formula** sometimes also called the **Law of total variance**

$$V[X] = E[V[X|Y]] + V[E[X|Y]]$$

Exercise 2.8 (Random sum of random variables) Suppose N is a random variables taking value in $\mathbb{N}$. If Y_j are IID (discrete) random variables

$$S_N = \sum_{n=1}^{N} Y_j \quad \text{with the convention } S_0 = 0$$

Use the conditioning formula for mean and variance to compute

$$E[S_N] \quad \text{and } V[S_N]$$

in terms of the mean and variance of Y and N.

Exercise 2.9 (Eggs) The chickens in my backyard lays a random number T of eggs, each egg is green with probability p and brown with probability $1-p$ indepedently of the other eggs and of T.

- Show that if T has a Poisson distribution then the number of green eggs layed has a Poisson distribution with parameter θp. *Hint*: Write the number of green eggs can be written as the random sum $X_1 + \cdots + X_T$ of Bernoulli random variables. Use conditioning
- Show that the the number of green eggs and brown eggs layed are independent random variables.

3 Borel Stong Law of Large numbers

3.1 Frequentist vs Bayesian

There are two broad interpretation of probability frequentist versus Bayesian (see e.g https://en.wikipedia.org/wiki/Probability_interpretations)

- In the Bayesian approach (subjective probability) the probability of an event $P(A)$ measure the degree of belief assigned by the probabilist to the ocurrence of the event. A good example of that approach in sports betting: How do you compute the probability that the Kansas City Chiefs will win the superbowl? It really measure the degree of belief assigned by a bettor (or by a bookmaker) to that event.
  This approach works well with Bayes formula

  $$\underbrace{P(A|B)}_{\text{Posterior}} = \underbrace{\frac{P(B|A)}{P(B)}}\underbrace{P(A)}_{\text{prior}}$$

  which describes how information (here the occurence of B) modify your belief about the occurence of A.

- In the frequentist approach the probability of an event is assigned by observing its frequency over time when repeating the same experiment multipl times. This works well if you want to compute the probability that the roulette wheel lands on the number 23. But this work not so well to compute the probbaility that an earthquake of magnituyde more than 6 will hit San Francisco area.
  Underlying the frequentist approach is the strong law of large number which we prove in this section: if we repeat the same experiment N times in the same conditions, and independetly of each other then

  $$\frac{\#\text{ of occurrences of the event } A}{N} \longrightarrow P(A) \quad \text{as } N \to \infty$$

3.2 Borel-Cantelli Lemma

Recall that $\limsup_n A_n = \cap_{n\geq 1} \cup_{k\geq n} A_k = \{\omega \in \Omega\,;\ \omega \text{ belongs to infinitely many} A_n\}$

Theorem 3.1 (The first Borel Cantelli Lemma) Let A_n be a sequence of events. Then

$$\sum_n P(A_n) < \infty \implies P\left(\limsup_n A_n\right) = 0$$

Proof. The sets $B_n = \cup_{k\geq n} A_k$ decrease to $\limsup_n A_n$ and so by sequential continuity

$$P(\limsup_n A_n) = \lim_{n\to\infty} P(B_n)\,.$$

By countable additivity we have

$$P(B_n) = P(\cup_{k\geq n} A_k) \leq \sum_{k\geq n} P(A_k)\,.$$

Since, by assumption, $\sum_n P(A_n) < \infty$ the tail sum $\sum_{k\geq n} P(A_k)$ must go to 0 as $n \to \infty$ and this concludes the proof. □

For the converse statement we need to add an assumption of independence.

Theorem 3.2 (The second Borel Cantelli Lemma) Let A_n be a sequence of **independent**. Then

$$\sum_n P(A_n) = \infty \implies P\left(\limsup_n A_n\right) = 1$$

Proof. Before we do the proof we need to remind the reader about some elementary fact about series. For two sequences of number (a_n) and (b_n) we sat that $a_n \sim b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. A result from analysis states that

$$\text{If } a_n \sim b_n \text{ then } \sum_n a_n \text{ converges } \iff \sum_n b_n \text{ converges}$$

By L'Hospital rule $\lim_{x\to 0} \frac{-\ln(1-x)}{x} = 1$ and thus $\sum_n a_n$ converges if and only if $-\sum_n \ln(1-a_n)$ converges.

Since $\sum_{k\ge n} P(A_k) = \infty$ for every n we must have then

$$\infty = -\sum_{k\ge n} \ln(1-P(A_k)) = -\ln \prod_{k\ge n} (1-P(A_k))$$

Thus, using the independence assumption $0 = \prod_{k\ge n}(1-P(A_k)) = \prod_{k\ge n} P(A_k^c) = P\left(\bigcap_{k\ge n} A_k^c\right)$.
This implies that $P(\bigcup_{k\ge n} A_k) = 1$ and thus $P(\limsup_n A_n) = 1$. □

3.3 Markov inequality and the weak law of large numbers

Let X_i be a sequence of IID discrete random variables with expected value p ((e.g take independent copies of the indicator function $X = 1_A$ for some set). Then we set

$$S_n = X_1 + \cdots + X_n$$

Theorem 3.3 (Markov inequality) If $Z \ge 0$ is a non-negative random variable then for any $a > 0$

$$P(Z \ge a) \le \frac{E[Z]}{a} .$$

Proof. We have the equality

$$Z \ge a1_{Z \ge a},$$

and taking expectation gives

$$E[Z] \ge aP(Z \ge a).$$

□

Using the Markov inequality we can then prove Chebyshev

Theorem 3.4 (Chebyshev inequality) Suppose X is a random variable with mean μ and variance σ^2, then

$$P\left(|X-\mu| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof. This simply Markov inequality applied to the random variable $(X-\mu)^2$ and with $a=\epsilon^2$. □

Theorem 3.5 (Weak Law of Large Numbers) Suppose X_i are IID random variable with common mean μ and variance σ^2, then for any $\epsilon \geq 0$

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

Proof. The random variable $\frac{S_n}{n}$ has mean μ and variance $\frac{\sigma^2}{n}$ and applying Chebyshev gives

$$P\left(\left|\frac{S_n}{n}-\mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

and this proves the statement. □

3.4 Strong Law of Large numbers

We prove now the strong Law of Large numbers for sum of independent discrete random variables. Once we have develop the theory of integration a bit more we will see that the same proof applies to general random variables (with finite mean and variance).

Theorem 3.6 (Strong Law of Large numbers) Let X_i be a sequence of IID (indendent and identically distributed) discrete random variables with common means $E[X_i] = \mu$ and variance $V[X_i] = \sigma^2$. Then

$$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu \right) = 1$$

The proof is based on the following Lemma which is a consequence of Borel Cantelli Lemma

Lemma 3.1 Suppose that for any $\epsilon > 0$ we have

$$\sum_n P(|Z_n - \mu| \geq \epsilon) < \infty$$

then Z_n converges to μ almost surely, that is

$$P\left(\omega\,;\, \lim_n Z_n(\omega) = \mu\right) = 1$$

Proof. If $Z_n(\omega)$ converges to μ then for any $\epsilon > 0$ there exists N (which may depends on ω) such that $|Z_n(\omega) - \mu| < \epsilon$ for all $n \ge N$. Therefore

$$\begin{aligned}\left\{\omega\,;\, \lim_n Z_n(\omega) = \mu\right\} &= \{\omega\,;\, |Z_n(\omega) - \mu| \ge \epsilon \text{ for finitely many } n\} \\ &= \{\omega\,;\, |Z_n(\omega) - \mu| \ge \epsilon \text{ for infinitely many } n\}^c\end{aligned}$$

By our assumption and the first Borel-Cantelli Lemma Theorem 3.1 we have that

$$P\left(\omega\,;\, |Z_n(\omega) - \mu| \ge \epsilon \text{ for infinitely many } n\right) = 0$$

and thus

$$P\left(\omega\,;\, \lim_n Z_n(\omega) = \mu\right) = 1$$

□

Proof of Theorem 3.6. Chebyshev inequality tells us that $P\left(\left|\frac{S_n}{n} - \mu\right| \ge \epsilon\right) \le \frac{\sigma^2}{n\epsilon^2}$ which unfortunately is not summbale in n so we cannot use the Lemma directly. There are various way around this and here is an elegant one: consider the random variable

$$Z_k = \frac{S_{k^2}}{k^2}$$

that is, we consider a subsequence of $\frac{S_n}{n}$.

Then

$$P\left(|Z_k - \mu| \geq \epsilon\right) = P\left(\left|\frac{S_{k^2}}{k^2} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{k^2\epsilon^2}$$

which is summable in k and thus by Lemma Lemma 3.1 we have that $\frac{S_{k2}}{k^2}$ converges to μ with probability 1. We now use a squeezing argument.

Assume next that the random variable X_i are non-negative then for any n let $k = k(n)$ such that $k^2 \leq n \leq (k+1)^2$. Since all the terms are positive we have

$$\frac{X_1 + \cdots + X_{k^2}}{k^2}\frac{k^2}{n} \leq \frac{X_1 + \cdots + X_n}{n} \leq \frac{(k+1)^2}{n}\frac{X_1 + \cdots + X_{(k+1)^2}}{(k+1)^2}$$

Now as $n \to \infty, k = k(n) \to \infty$ and since $\lim_{k\to\infty}\frac{(k+1)^2}{k^2} = 1$

$$1 \leq \frac{n}{k^2} \leq \frac{(k+1)^2}{k^2} \implies \lim_n \frac{n}{k^2} = 1\,.$$

Simlarly $\lim_{n\to\infty}\frac{(k+1)^2}{n} = 1$ and this shows that $\frac{S_n}{n}$ converges almost surely.

Finally for the general case note that if a sequence $a_n \to a$ then $|a_n| \to |a|$ and so if we write $a^+ = \max\{a, 0\}$ and $a^- = -min\{a, 0\}$ we have $a = a^+ - a^-$ and $|a| = a^+a^-$. Therefore a_n converges if and only if a_n^+ and a_n^- converges and we can now apply the previous argument. $\square$.

3.5 Homework problems

Exercise 3.1 Consider a hypergeometric random variable $Z = Z_{n,b,N}$ with parameters N b, $r = N - b$ and sampling size n. Consider the following limits

$$n, N \to \infty, \quad n \le N, \quad p = \frac{b}{N}$$

Argue that under these condition as $n, N \to \infty$, $\frac{1}{n} Z_{n,b,N}$ converges almost surely to p.
Hint: Look at the formula for the variance in Exercise 2.4 and study the proof of the strong law of large numbers. Your argument should not be long!

Exercise 3.2 (The infinite Monkey theorem) The short story *The Library of Babel* by Jorge Luis Borges https://en.wikipedia.org/wiki/The_Library_of_Babel is about a vast library which contains all possible 410-page books of a certain format and character sets. Give now a Monkey a typewriter and show, under suitable assumption on the typing abilities of the monkey, that the monkey will eventually type every book in the Libray of Babel infinitely many times.

Exercise 3.3 (Bob and Alice do online dating) On a certain day, Alice decides to look for a potential life partner on an online dating portal. Being a very thorough and confident person, she decides that everyday she will pick a guy uniformly at random from among the male members of the dating portal, and go out on a date with him. Unbeknownst to Alice her neighbor Bob, an exceedingly shy mathematician, is interested in dating her. Bob decides that he will go out on a date with Alice only on the days that Alice happens to pick him from the dating portal, of which he is already a member.

For the first two parts, assume that 50 new male members and 40 new female members join the dating portal everyday.

1. What is the probability that Alice and Bob would have a date on the nth day? Do you think Bob and Alice would eventually stop meeting? Justify your answer, clearly stating any additional assumptions.
2. Now suppose that Bob also picks a girl uniformly at random everyday, from among the female members of the portal, and that Alice behaves exactly as before. Assume also that Bob and Alice will meet on a given day if and only if they both happen to pick each other. In this case, do you think Bob and Alice would eventually stop meeting?
3. For this part, suppose that Alice and Bob behave as in part 1., i.e.,Alice picks a guy uniformly at random, but Bob is only interested in dating Alice. However, the number of male members in the portal increases by 1 percent everyday. Do you think Bob and Alice would eventually stop meeting?