Part 3: Measures on Product Spaces and Conditional Expectation

Probability Theory: Math 605, Fall 2024

Luc Rey-Bellet
University of Massachusetts Amherst

2024-11-22

1 Independence and product measures

In this section we study the concept of independence in a general form: independence of random variables and independence of σ-algebra. This leads to the concept of product measures and the classic Fubini Theorem. We illustrate these ideas with vector valued random variables and some simulation algorithm. Dependent random variables will be considered in next section.

1.1 Independence

- Two events $A, B \in \mathcal{A}$ are independent if $P(A|B) = P(A)$ or equivalently $P(A \cap B) = P(A)P(B)$ and this generalizes to arbitrary collection of events.
- For two random variable X and Y to be independent it should means that *any* information or knowledege derived from the RV Y should not influence the RV X. All the information encoded in the RV X taking values in $(E, \mathcal{E})$ is the σ-algebra generated by X that is $\sigma(X) = X^{-1}(\mathcal{E}) = \{X^{-1}(B), B \in \mathcal{E}\}$. This motivates the following.

Definition 1.1 (Independence) Let $(\Omega, \mathcal{A}, P)$ be a probability space.

1. *Independence of σ-algebras*: Two sub-σ-algebra $\mathcal{A}_1 \subset \mathcal{A}$ and $\mathcal{A}_2 \subset \mathcal{A}$ are **independent** if

$$P(A_1 \cap A_2) = P(A_1)P(A_2) \quad \text{for all } A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2 \,.$$

   A collection (not necessarily countable) of σ-algebras $\{\mathcal{A}_j\}_{j \in J}$ is independent if, for any finite subset $I \subset J$,

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i) \quad \text{for all } A_i \in \mathcal{A}_i \,.$$

2. *Independence of random variables*: The collection of random variables $X_j : (\Omega, \mathcal{A}, P) \to (E_j, \mathcal{E}_j)$ for $j \in J$ are indedpendent if the collection of σ-algebras $\sigma(X_j)$ are independent.

We consider from now on only two random variables X and Y but all of this generalizes easily to arbitrary finite collections. Our next theorem makes the definition of indepdence a bit more easy to check.

Theorem 1.1 (Characterization of independence) Two random variables X (taking values in $(E, \mathcal{E})$) and Y taking values in $(F, \mathcal{F})$) to be independent if and only if any of the following equivalent conditions holds.

1. $P(X \in A, Y \in B) = P(X \in A)P(X \in B)$ for all $A \in \mathcal{E}$ and for all $B \in \mathcal{F}$.
2. $P(X \in A, Y \in B) = P(X \in A)P(X \in B)$ for all $A \in \mathcal{C}$ and for all $B \in \mathcal{D}$ where $\mathcal{C}$ and $\mathcal{D}$ are p-systems generating $\mathcal{E}$ and $\mathcal{F}$.
3. $f(X)$ and $g(Y)$ are independent for any measurable f and g.
4. $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all bounded and measurable (or all non-negative) f, g.
5. If $E = F = \mathbb{R}$ (or $\mathbb{R}^d$), $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all bounded and continuous functions f, g.

Proof.

- Item 1. is merely a restatement of the definition and clearly item 1. $\implies$ item 2.

- To see that Item 2. $\implies$ item 1. we use the monotone class theorem. Fix $B \in \mathcal{D}$ then the collection

$$\{A \in \mathcal{E} \,:\, P(X \in A, Y \in B) = P(X \in A)P(Y \in B)\}$$

contains the p-system $\mathcal{C}$ and is a d-system (contains Ω, closed under complement, closed under increasing limits, check this yourself please). Therefore by the monotone class theorem it contains $\mathcal{E}$. Analgously fix now $A \in \mathcal{E}$, then the set
$\{B \in \mathcal{F} \,:\, P(X \in A, Y \in B) = P(X \in A)P(Y \in B)\}$ contains $\mathcal{F}$.

- To see that item 3. $\implies$ item 1. take $f(X) = 1_A(X)$ and $g(Y) = 1_B(Y)$. If these two random variable are independent, this simply means that the event $X \in A$ and $Y \in B$ are independent. Conversely we note that $V = f(X)$ is measurable with respect to $\sigma(X)$: since $V^{-1}(B) = X^{-1}(f^{-1}(B)) \in \sigma(X)$ this shows that $\sigma(f(X)) \subset \sigma(X)$. Likewise $\sigma(g(Y)) \subset \sigma(Y)$.
Since $\sigma(X)$ and $\sigma(Y)$ are independent so are $\sigma(f(X))$ and $\sigma(g(Y))$.

- To see that item 4. $\implies$ item 1. take $f = 1_A$ and $g = 1_B$. To show that item 1. $\implies$ item 4. note that item 1. can be rewritten that $E[1_A(X)1_B(Y)] = E[1_A(X)]E[1_B(Y)]$. By linearity of the expectation then item 4. holds for all simple functions g and g. If f and g are non negative then we choose sequences of simple functions such that $f_n \nearrow f$ anf $g_n \nearrow g$. We have then $f_n g_n \nearrow fg$ and using the monotone convergence theorem twice we have

$$\begin{aligned} E[f(X)g(Y)] &= E[\lim_n f_n(X)g_n(Y)] = \lim_n E[f_n(X)g_n(Y)] = \lim_n E[f_n(X)]E[g_n(Y)] \\ &= \lim_n E[f_n(X)] \lim_n E[g_n(Y)] = E[f(X)]E[g(Y)] \end{aligned}$$

If f and g are bounded and measurable then we write $f = f_+ - f_-$ and $g = g_= - g_-$. Then $f_\pm$ and $g_\pm$ are bounded and measurable and thus the product of $f_\pm(X)$ and $g_\pm(Y)$ are also integrable. We have

$$\begin{aligned} E[f(X)g(Y)] &= E[(f_+(X) - f_-(X))(g_+(Y) - g_-(Y))] \\ &= E[f_+(X)g_+(Y)] + E[f_-(X)g_-(Y)] - E[f_+(X)g_-(Y)] - E[f_-(X)(g_+(Y)] \\ &= E[f_+(X)]E[g_+(Y)] + E[f_-(X)]E[g_-(Y)] - E[f_+(X)]E[g_-(Y)] - E[f_-(X])E[g_+(Y)] \\ &= E[f_+(X) - f_-(X)]E[g_+(Y) - g_-(Y)] \end{aligned}$$

• Clearly item 1 $\implies$ item 4 $\implies$ item 5 For the converse we show that item 5. $\implies$ item 2. Given an interval (a, b) consider an increasing sequence of piecewise linear function continuous function $f_n \nearrow 1_{(a,b)}$.

$$f_n(t) = \begin{cases} 0 & t \leq a + \frac{1}{n} \text{ or } t \geq b - \frac{1}{n} \\ 1 & a + \frac{1}{n} \leq t \leq b - \frac{1}{n} \\ \text{linear} & \text{otherwise} \end{cases}$$

Let us consider the p-system which contains all intervals of the form (a, b) and which generate the Borel σ-algebra $\mathcal{B}$. By using a monotone convergence argument like for item 1. $\implies$ item 4. we see that for $f_n \nearrow 1_{(a,b)}$ and $g_n \nearrow 1_{(c,d)}$

$$E[f_n(X)g_n(Y)] = E[f_n(X)]E[g_n(Y)] \implies P(X \in (a, b))P(Y \in (c, d))$$

and so item 2. holds.
For separable metric space, the so-called Urysohn lemma can be used to prove the same results. □

1.2 Independence and product measures

- While we have expressed so far independence of X and Y as a property on the probability spacew $(\Omega, \mathcal{A}, P)$ we can also view it as a property of the distributions P^X and P^Y.

- Example: Suppose X and Y are discrete random variables then if they are independent we have

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

  and thus $P^{(X,Y)}(i, j) = P^X(i)P^Y(j)$. That is the distribution of the random variable $Z = (X, Y)$ factorizes into the product of P^X and P^Y.

- Product spaces. In order to build-up more examples we need the so-called Fubini theorem. Given two measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ we consider the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ where $\mathcal{E} \otimes \mathcal{F}$ is the sigma-algebra generated by the rectangles $A \times B$ (see **?@exr-45**)

- Measurable functions on product spaces. As we have seen in **?@exr-45** for any measurable function $f : E \times F \to \mathbb{R}$ the sections $g(y) = f(x, y)$ (for any x) and $h(x) = f(x, y)$ (for any y) are measurable.

Theorem 1.2 (Tonelli-Fubini Theorem) Suppose P is a probability on $(E, \mathcal{E})$ and Q is a probability on $(F, \mathcal{F})$.

1. The function $R : \mathcal{E} \times \mathcal{F}$ defined by

$$R(A \times B) = P(A)Q(B)$$

extends to a unique probability measure on $\mathcal{E} \otimes \mathcal{F}$. This measure is denoted by $P \otimes Q$ and is called the product measure of P and Q.

2. Suppose f is measurable with respect to $\mathcal{E} \otimes \mathcal{F}$, either non-negative, or integrable with respect to $P \otimes Q$. Then the functions

$$x \mapsto \int f(x, y) dQ(y) \quad \text{and} \quad y \mapsto \int f(x, y) dP(x)$$

are integrable with respect to P and Q respectively and we have

$$\int_{E \times F} f(x, y) d(P \otimes Q)(x, y) = \int_E \left(\int_F f(x, y) dQ(y) \right) dP(x) = \int_F \left(\int_E f(x, y) dP(x) \right) dQ(y)$$

Proof. For item 1. we need to extend the R to arbitray element in $\mathcal{E} \otimes \mathcal{F}$. For $C \in \mathcal{E} \otimes \mathcal{F}$ consider the slice of C along x given by

$$C(x) = \{y \in F \, : (x, y) \in C\} \, .$$

If $C = A \times B$, then $C(x) = B$ for all $x \in A$ and $C(x) = \emptyset$ for $x \notin A$ and we have then

$$R(C) = P(A)Q(B) = \int_E Q(C(x))dP(x)$$

Now we define

$$\mathcal{H} = \{C \in \mathcal{E} \otimes \mathcal{F} \, : \, x \to Q(C(x)) \text{ is measurable}\}$$

it is not difficult to check that $\mathcal{H}$ is a σ-algebra and $\mathcal{H} \supset E \times F$ and therefore $\mathcal{H} = \mathcal{E} \otimes \mathcal{F}$.
We now *define*, for any $C \in \mathcal{H} = \mathcal{E} \otimes \mathcal{F}$,

$$R(C) = \int_E Q(C(x))dP(x)$$

and we check this is a probability measure. Clearly $R(E \times F) = P(E)Q(F) = 1$. Let $C_n \in \mathcal{E} \otimes \mathcal{F}$ be pairwise disjoint and $C = \cup_{n=1}^{\infty} C_n$. Then the slices $C_n(x)$ are pairwise disjoint and by the monotone convergence theorem $Q(C(x)) = \sum_{n=1}^{\infty} Q(C_n(x))$.

Applying MCT again to the function $g_n(x) = \sum_{k=1}^{n} Q(C_n(x))$ we find that

$$\sum_{n=1}^{\infty} R(C_n) = \sum_{n=1}^{\infty} \int_E Q(C_n(x))dP(x) = \int_E \sum_{n=1}^{\infty} Q(C_n(x))dP(x) = \int_E Q(C(x))dP(x) = R(C).$$

and this shows that R is a probability measure. Uniqueness of R follows from the monotone class theorem.

For item 2. note that in item 1, we have proved it in the case where $f(x,y) = 1_C(x,y)$. By linearity the result then holds for simple functions. If f is nonnegative and measurable then pick an increasing sequence such that $f_n \nearrow f$. Then by MCT

$$\int f(x,y)dP \otimes Q(x,y) = \lim_n \int f_n(x,y)dP \otimes Q(x,y) = \lim_n \int_E (\int_F f_n(x,y)dQ(y))dP(x)$$

But the function $x \to \int_F f_n(x,y)dQ(y)$ is increasing in n and by MCT again, and again.

$$\begin{aligned}\int f(x,y)dP \otimes Q(x,y) &= \int_E \lim_n (\int_F f_n(x,y)dQ(y))dP(x) = \int_E (\int_F \lim_n f_n(x,y)dQ(y))dP(x) \\ &= \int_E (\int_F f(x,y)dQ(y))dP(x)\end{aligned}$$

Simlarly one shows that $\int f(x,y)dP \otimes Q(x,y) = \int_F (\int_E f(x,y)dP(x))dQ(y)$. The result for integrable f follows by decomposing into postive and negative part. $\square$

Applying this to random variables we get

Corollary 1.1 Suppose $Z = (X, Y) : (\Omega, \mathcal{A}, P) \to (E \times F, \mathcal{E} \otimes \mathcal{F})$. Then the random variables X and Y are independent if and only if the distribution $P^{(X,Y)}$ of the pair (X, Y) is equal to $P^X \otimes P^Y$.

Proof. The random variables X and Y are independent if and only if

$$P((X, Y) \in A \times B) = P(X \in A)P(X \in B)\,.$$

This is equivalent to saying that

$$P^{(X,Y)}(A \times B) = P^X(A) \times P^Y(B)$$

By the uniqueness in Fubini Theorem we have $P^{(X,Y)} = P^X \otimes P^Y$. $\square$

1.3 Constructing a probability space for independent random variables

We can construct a probability model for $X_1, \cdots, X_n$ which are independent (real-valued) random variables with given distribution $P^{X_1}, \cdots, P^{X_n}$.

We know how to construct the probability space for each random variable separately, for example,

$$X_i : (\Omega_i, \mathcal{B}_i, P_i) \to (\mathbb{R}, \mathcal{B})$$

where $\Omega_i = [0,1]$, P_i is Lebesgue measure on $[0,1]$ and $X_i = Q_i$ where Q_i is a quantile function for X_i.

We know take

$$\Omega = \prod_i \Omega_i = [0,1]^n, \quad \mathcal{B} = \otimes_{i=1}^n \mathcal{B}_i, \quad P = \otimes_{i=1}^n P_i$$

and define the map $X : \Omega \to \mathbb{R}^n$ by $X = (X_1, \cdots X_n)$.

Fubini-Tonelli Theorems shows that $P^X = P \circ X^{-1}$ is the distribution of $X = (X_1, \cdots X_n)$ on $\mathbb{R}^n$ with the product σ-algebra and that the random variables are independent

Fubini-Tonelli for countably many RV:

We extend this result to countable many independent random variables: this is important in practice where we need such models: for example flipping a coin as many times as needed!

This can be seen as an extension of Fubini theorem and is also a specical case of the so-called Kolmogorov extension theorem which us used to construct general probability measures on infinite product space. No proof is given here.

Infinite product σ-algebras

Given σ-algebras $\mathcal{A}_j$ on Ω_j we set $\Omega = \prod_{j=1}^{\infty} \Omega_j$ and define rectangles for $n_1 < n_2 < \cdots < n_k$ and k finite but arbitrary

$$A_{n_1} \times A_{n_2} \times \cdots \times A_{n_k} = \{\omega = (\omega_1, \omega_2, \omega_3, \cdots) \in \Omega \ : w_{n_j} \in A_{n_j}\}$$

where $A_{n_j} \in \mathcal{A}_{n_j}$. The product σ-algebra $\mathcal{A} = \bigotimes_{j=1}^{\infty} \mathcal{A}_j$ the σ-algebras generated by all the rectangles.

Theorem 1.3 Given probaility spaces $(\Omega_i, \mathcal{A}_i, P_i)$ and with $\Omega = \prod_{j=1}^{\infty} \Omega_i$ and $\mathcal{A} = \bigotimes_{j=1}^{\infty} \mathcal{A}_j$ there exists a unique probability P on $\Omega(A)$ such that

$$P(A_{n_1} \times \cdots \times A_{n_k}) = \prod_{j=1}^{k} P_{n_j}(A_{n_j})$$

for all $A_{n_j} \in \mathcal{A}_{n_j}$, all $n_1 < \cdots < n_k$ and all k.

If we have a RV $X_n : \Omega_n \to \mathbb{R}$ then we define its extension to Ω by

$$\tilde{X}_n(\omega) = X_n(\omega_n)$$

and the distribution of $\tilde{X}_n$ is the same as the distribution of X_n because

$$\tilde{X_n}^{-1}(B_n) = \Omega_1 \times \cdots \times \Omega_{n-1} \times X_n^{-1}(B_n) \times \Omega_{n+1} \times \cdots$$

and thus

$$P(\tilde{X}_n \in B_n) = P_n(X_n \in B_n)\,.$$

A similar computation shows that $\tilde{X}_n$ and $\tilde{X}_m$ are independent for $n \neq m$ or more generally any finite collection of $X_j's$ are independent.

1.4 Kolmogorov zero-one law

We consider $(X_n)_{n=1}^{\infty}$ to be RVs defined on some probability space Ω. We may think of n as "time" and we consider the following σ-algebras

$$\begin{aligned} \mathcal{B}_n &= \sigma(X_n) && \text{the } \sigma\text{-algebra generated by } X_n \\ \mathcal{C}_n &= \sigma\left(\cup_{p\geq n}\mathcal{B}_n\right) && \text{the } \sigma\text{-algebra describing the "future" after time } n \\ \mathcal{C}_\infty &= \cap_{n=1}^{\infty}\mathcal{C}_n && \text{the "tail" } \sigma\text{-algebra or } \sigma\text{-algebra "at infinity"} \end{aligned}$$

Theorem 1.4 (Zero-one law) Suppose X_n is a sequence of *independent* random variables and let $\mathcal{C}_\infty$ be the corresponding tail σ-algebra. Then we have

$$C \in \mathcal{C}_\infty \implies P(C) = 0 \text{ or } P(C) = 1$$

Proof. The σ algebras $\{\mathcal{B}_1, \cdots, \mathcal{B}_n, \mathcal{C}_n\}$ are independent and therefore $\{\mathcal{B}_1, \cdots, \mathcal{B}_n, \mathcal{C}_\infty\}$ are independent for every n since $\mathcal{C}_\infty \subset \mathcal{C}_n$. Therefore $\mathcal{C}_0 = \sigma\left(\cup_{n\geq 0}\mathcal{B}_n\right)$ and $\mathcal{C}_\infty$ are independent. So for $A \in \mathcal{C}_0$ and $B \in \mathcal{C}_\infty$ we have $P(A \cap B) = P(A)P(B)$. This holds also for $A = B$ since $\mathcal{C}_\infty \subset \mathcal{C}_0$. Therefore we have $P(A) = P(A)^2$ which is possible only if $P(A) = 0$ or 1 $\square$.

Examples Given independent random variable $X_1, X_2, \cdots$ we define $S_n = X_1 + X_2 + \cdots + X_n$.

- The event $\{\omega \ : \ \lim_n X_n(\omega) \text{ exists } \}$ belongs to every $\mathcal{C}_n$ and thus belong to $\mathcal{C}_\infty$. Therefore X_n either converges a.s or diverges a.s.
- A random variable which is measurable with respect to $\mathcal{C}_\infty$ must be constant almost surely. Therefore

$$\limsup_n X_n \, , \quad \liminf_n X_n \, , \quad \limsup_n \frac{1}{n} S_n \, , \quad \liminf_n \frac{1}{n} S_n$$

  are all constant almost surely.
- The event $\limsup_n \{X_n \in B\}$ (also called $\{X_n \in B \text{ infinitely often } \}$) is in $\mathcal{C}_\infty$.
- The event $\limsup_n \{S_n \in B\}$ is not in $\mathcal{C}_\infty$.

1.5 Exercises

Exercise 1.1

- Suppose $X \geq 0$ is a non-negative random variable and $p > 0$. Show that

$$E[X^p] = \int_0^\infty pt^{p-1}(1 - F(t))dt = \int_0^\infty pt^{p-1}P(X > t)dt$$

  In particular $E[X] = \int P(X > t)dt$.
  Hint: Use Fubini on the product measure P times Lebesgue measure on $[0, \infty)$.

- Deduce from this that if X takes non-negative integer values we have

$$E[X] = \sum_{n>0} P(X > n)\,, \quad E[X^2] = 2\sum_{n>0} nP(X > n) + E[X]\,.$$

Exercise 1.2 Find three random variable X, Y, and Z taking values in $\{-1, 1\}$ which are pairwise independent but are not independent.

Exercise 1.3

- A random variable is a Bernoulli RV if X takes only values 0 or 1 and then $E[X] = P(X = 1)$ (alternatively you can think of $X(\omega) = 1_A(\omega)$ for some measurable set A.) Show that $Y = 1 - X$ is also Bernoulli and that the product of two Bernoulli random variables is again a Bernoulli RV (no independence required).
- Suppose $X_1, X_2, \cdots, X_n$ are Bernoulli random variables on some probability space $(\Omega, \mathcal{A}, P)$ (they are *not* assumed to be independent) and let $Y_k = 1 - X_k$. Show that

$$P(X_1 = 0, X_2 = 0, \cdots, X_n = 0) = E\left[\prod_{i=1}^{n} Y_i\right]$$

  and

$$P(X_1 = 0, \cdots, X_k = 0, X_{k+1} = 1, \cdots, X_n = 1) = E\left[\prod_{i=1}^{k} Y_i \prod_{j=k+1}^{n} X_j\right]$$

- Show that the Bernoulli random variables $X_1, \cdots, X_n$ are independent if and only $E[\prod_{j\in J} X_j] = \prod_{j\in J} E[X_j]$ for all subset J of $\{1, \cdots, n\}$.

Exercise 1.4 Consider the probability space $([0,1), \mathcal{B}, P)$ where $\mathcal{B}$ is the Borel σ-algebra and P is the uniform distribution on $[0,1)$. Expand each number ω in $[0,1)$ in dyadic expansion (or bits)

$$\omega = \sum_{n=1}^{\infty} \frac{d_n(\omega)}{2^n} = 0.d_1(\omega)d_2(\omega)d_3(\omega)\cdots \qquad \text{with} \quad d_n(\omega) \in \{0,1\}$$

- For certain numbers the dyadic expansion is not unique. For example we have

  $$\frac{1}{2} = 0.100000 = \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = 0.01111111111\cdots$$

  Show that P almost surely a number ω has a unique dyadic expansion.

- Prove that each $d_n(\omega)$ is a random variable and that they are identically distributed (i.e. each d_n has the same distribution).

- Prove that the $d_n, n = 1, 2, 3, \cdots$ are a collection of independent and identically distributed random variables.

Remark: this problem shows that you can think of the Lebesgue measure on $[0,1)$ as the infinite product measure of independent Bernoulli trials.

Exercise 1.5 Suppose X and Y are RV with finite variance (or equivalently $X, Y \in L^2$. The *covariance of X and Y* is defined by

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

1. Show that $\mathrm{Cov}(X,Y)$ is well defined and bounded by $\sqrt{\mathrm{VarX}}\sqrt{\mathrm{VarY}}$.
2. The correlation coefficient $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{VarX}}\sqrt{\mathrm{VarY}}}$ measure the correlation between X and Y. Given a number $\alpha \in [-1,1]$ find two random variables X and Y such that $\rho(X,Y) = \alpha$.
3. Show that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(X+Y) + 2\mathrm{Cov}(X,Y)$.
4. Show that if X and Y are independent then $\mathrm{Cov}(X,Y) = 0$ and so $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
5. The converse statement of 3. is in general not true, that is $\mathrm{Cov}(X,Y) = 0$ does not imply that X and Y are independent. ` *Hint* For example take X to be standard normal Z discrete with $P(Z = \pm 1) = \frac{1}{2}$ (Z is sometimes called a Rademacher RV) and $Y = ZX$.

2 Probability kernels and measures on product spaces

In this section we introduce the concept of probability kernels to construct probability measures on product spaces and we give several examples and applications of distributions on product spaces.

2.1 Probability kernels

How do we build "general" measures on some product space $E \times F$ (e.g. $\mathbb{R}^2$ or $\mathbb{R}^n$)?

Definition 2.1 Given 2 measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ a **probability kernel** $K(x, B)$ **from** $(E, \mathcal{E})$ **into** $(F, \mathcal{F})$ (also often called a **Markov kernel**) is a map

$$K : E \times \mathcal{F} \to \mathbb{R}$$

such that

1. For any $B \in \mathcal{F}$ the map $x \to K(x, B)$ is a measurable map.
2. For any $x \in E$ the map $B \to K(x, B)$ is a probability measure on $(F, \mathcal{F})$.

Examples

The following theorem is a generalization of Fubini-Tonelli

Theorem 2.1 Let P be a probability measure on $(E, \mathcal{E})$ and $K(x, B)$ a probability kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then a probability measure R on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is defined by

$$\int f(x,y) dR(x,y) = \int_X \left(\int_F f(x,y) K(x, dy) \right) dP(x)$$

for any measurable non-negative f. In particular we have for any $A \in \mathcal{E}$ and $B \in \mathcal{F}$

$$R(A \times B) = \int 1_A(x) 1_B(y) dR(x,y) = \int 1_A(x) K(x,B) dP(x) = \int_A K(x,B) dP(x)\,.$$

We write

$$R(dx, dy) = P(dx) K(x, dy)$$

Proof. The proof is very similar to Fubini-Tonelli theorem and is thus omitted.

- This theorem is intimately related to the concepts of conditional probability and conditional expectations which we will study later.
- Roughly speaking, on nice probability space (e.g. separable metric spaces), *every* probaility measure on the product space $E \times F$ can be constructed in the way described in Theorem 2.1 (this is a deep result).

Definition 2.2 (marginals of a probability measure) Given a probability measure R on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ the marginals of R are on E and F are defined defined to be the measures given by

$$P(A) = R(A \times F) \quad A \in \mathcal{E}$$
$$Q(B) = R(E \times B) \quad B \in \mathcal{F}$$

Alternatively we can think of the marginals as the image measure

$$P = R \circ \pi_E^{-1}, \qquad P = R \circ \pi_F^{-1}$$

where π_E and π_F are the projection maps $\pi_E(x, y) = x$ and $\pi_F(x, y) = y$

- If $R = P \otimes Q$ is a product measure then the marginal of R are P and Q. This is for the kernel $K(x, A) = Q(A)$.
- If $R(dx, dy) = P(dx)K(x, dy)$ then we have

$$R(A \times F) = \int_A K(x, F) dP(x) = P(A) \qquad R(E \times B) = \int_E K(x, B) dP(X)$$

so its marginals are P and and Q given by $Q(B) = \int_E K(x, B) dP(x)$.

2.2 Conditional distributions

On nice probability space (e.g when E and F are separable metic spaces like $\mathbb{R}^n$) one can show that every probability measure R can be written using a kernel. There exists probability spaces where the conclusions of the following theorem do not hold but in practice most examples are covered.

Theorem 2.2 (Existence of conditional distributions) Suppose E and F are separable metric spaces equipped with their respective Borel σ-algebra $\mathcal{E}$ and $\mathcal{F}$. Suppose R is probabilty measure on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F}$ with marginals P and Q. Then there exists a kernel $k(x, B)$ such that

$$R(A \times B) = \int K(x, B) dP(x)$$

The kernel $K(x, B)$ is unique P a.s.

Proof. The proof has technical difficulties and we only content ourselves with some comments on the proof. Consider the marginal measure on E

$$P(A) = R(A \times F)$$

and for fixed given $B \in \mathcal{F}$ with $Q(B) > 0$ consider the probability measure on E given by

$$P_B(A) = \frac{R(A \times B)}{R(E \times B)} = \frac{R(A \times B)}{Q(B)}$$

It is easy to verify that if $P(A) = 0$ then $P_B(A) = 0$ and so $P_B \ll P$ and by Radon-Nikodym theorem there exists a random variable Y_B such that

$$P_B(A) = \int 1_A(x) Y_B(x) dP(x) \quad \Longrightarrow \quad R(A \times B) = \int 1_A(x) \underbrace{Y_B(x) Q(B)}_{\equiv K(x,B)} dP(x)$$

It is important to note that for any choice of B, $K(x, B)$ is defined P almost surely. The technical issue is that for a kernel we want $K(x, B)$ to be defined for all B, P-almost surely! If we ignore this point we note that we have

$$R(A \times F) = P(A) = \int 1_A(x) K(x, F) dP(x)$$

and thus $K(x, F) = 1$, for P almost all x. We also have for pairwise disjoint B_i, using the monotone convergence theorem

$$\int 1_A(x) K(x, \cup_{n=1}^{\infty} B_n) dP(x) = R(A \times \cup_{n=1}^{\infty} B_n)$$
$$= \sum_{n=1}^{\infty} R(A \times B_n) = \sum_{n=1}^{\infty} \int 1_A(x) K(x, B_n) dP(x) = \int 1_A(x) \sum_{n=1}^{\infty} K(x, B_n) dP(x)$$

and since A is abribtrary this shows that $K(x, \cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} K(x, B_n)$ for P almost all x. The rest of the proof consists in showing first for $F = \mathbb{R}$ that we can define the kernel P-almost surely. This uses the fact the the Borel σ-algebra is countably generated, the monotone class theorem, and that the CDF is right-continuous. The general case is then proved by using the fact that complete separable metric spaces are isomorphic (in the sense of measure theory) to subsets of $\mathbb{R}$.

2.3 Lebesgue measure on $\mathbb{R}^n$ and densities

In previous chapter we have constructed the Lebesgue probability measure P_0 on $[0,1]$ as the unique measure such that $P_0[(a,b]] = (b-a)$. Using this and other uniformly distributed random variables we define the Lebesgue measure on $\mathbb{R}$ and on $\mathbb{R}^n$

Definition 2.3 The Lebesgue measure on $\mathbb{R}$ is a set function m such that

1. $m(\emptyset) = 0$.
2. $m\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} m(A_i)$
3. $m((a,b]) = b - a$ for any $a < b$.

The Lebsgue measure on $\mathbb{R}$ is *not* a probability measure since $m(\mathbb{R}) = +\infty$, it is an example of an infinite measure. We can easily construct it using uniform random variables on $(n, n+1]$ with distribution $P_{n,n+1}$ namely we set

$$m(A) = \sum_{n=-\infty}^{\infty} P_{n,n+1}(A)$$

The uniqueness of the measure $P_{n,n+1}$ implies that the measure m is unique as well.

By doing a Fubini-Tonelli theorem type argument one can construct the Lebesgue measure on $\mathbb{R}^n$.

Definition 2.4 If we equip $\mathbb{R}^n$ with the product σ-algebra $\mathcal{B} \otimes \cdots \otimes \mathcal{B}$ the Lebesgue measure m_n on $\mathbb{R}^n$ is the product of n Lebesgue measure on $\mathbb{R}$. We have

$$m_n\left(\prod_{i=1}^{n}[a_i, b_i]\right) = \prod_{i=1}^{n}(b_i - a_i)$$

Notations we often use the notation dx or $dx_1 \cdots dx_n$ for integration with respect to m_n.

Definition 2.5 A probability measure on $(\mathbb{R}^n, \mathcal{B}_n)$ (where $\mathcal{B}_n = \mathcal{B} \otimes \cdots \otimes \mathcal{B}$) has a **density** f if f is a nonnegative Borel measurable function and

$$P(A) = \int_A f(x)dx = \int 1_A(x)f(x)dx = \int f(x_1, \cdots, x_n)1_A(x_1, \cdots, x_n)dx_1 \cdots dx_n$$

Theorem 2.3 A non-negative Borel measurable function $f(x)$ is the density of a Borel probability measure if and only if $\int f(x)dx = 1$ and it determines the probability P. Conversely the probability measure determines its density (if it exists!) up to a set of Lebesgue measure 0.

Proof. Given $f \geq 0$ it is easy to check that

$$P(A) = \int 1_A f(x)dx$$

defines a probbaility measure (same proof as in **?@exr-63**).
Conversely assume that f and f' are two densities for the measure P, then for any measurable set A we have $P(A) = \int_A f(x)dx = \int_A f'(x)dx$. Consider now the set

$$A_n = \left\{ x : f'(x) \geq f(x) + \frac{1}{n} \right\}$$

Then we have

$$P(A_n) = \int_{A_n} f'(x)dx \geq \int_{A_n} \left(f(x) + \frac{1}{n} \right) dx = P(A_n) + \frac{1}{n} m(A_n)$$

and therefore $m(A_n) = 0$ and since A_n increases to $A = \{f < f'\}$ we have shown that $m(\{f < f'\}) = 0$. By symmetry we have $f = f'$ a.s. $\square$.

Theorem 2.4 Suppose the random variable (X, Y) has a probability distribution $R(dx, dy)$ with density $f(x, y)$. Then

1. Both X and Y have densities given respectively by

$$f_X(x) = \int f(x, y)dy\,, \qquad f_Y(y) = \int f(x, y)dx\,.$$

2. X and Y are independent if and only if

$$f(x, y) = f_X(x)f_Y(y)\,.$$

3. If we set

$$k(x, y) = \frac{f(x, y)}{f_X(x)} \text{ if } f_X(x) \neq 0$$

and this defines a kernel $K(x, B) = \int_B k(x, y)dy$ and the probability distribution(X, Y) is given by $R(dx, dy) = f_X(x)k(x, y)dxdy$.

Remark It does not matter how $K(x, B)$ is defined for such x where $f_X(x) = 0$ and so we have left it undefined. There are in general many kernels densities which will give the same probability aloowing for changes on sets of zero probbaility.

Proof.

1. For $A \in \mathcal{B}$ we have

$$P(X \in A) = P(X \in A, Y \in \mathbb{R}) = \int_A \left(\int_R f(x,y) dy \right) dx = \int_A f_X(x) dx$$

and since this holds for all A, $f_X(x)$ is a density for the distribution of X.

2. If $f(x,y) = f_X(x) f_Y(y)$ then

$$\begin{aligned} P(X \in A, Y \in B) &= \int 1_{A \times B}(x,y) f_X(x) f_Y(y) dx dy = \int 1_A(x) f_X(x) dx \int 1_B(x) f_Y(y) dy \\ &= P(X \in A) P(Y \in B) \end{aligned}$$

Conversely assume that X and Y are independent. Consider the collection of sets

$$\mathcal{H} = \left\{ C \in \mathcal{B} \otimes \mathcal{B} \, : \, \int_C f(x,y) dx dy = \int_C f_X(x) f_Y(y) dx dy \right\}$$

The independence and Fubini Theorem implies that any set $C = A \times B$ belongs to $\mathcal{H}$. Since this is a p-system generating the σ-algebra the monotone class theorem shows that $\mathcal{H} = \mathcal{B} \otimes \mathcal{B}$.

To prove item 3, we have

$$\int k(x,y)dy = \int \frac{f(x,y)}{f_X(x)}dy = \frac{f_X(x)}{f_X(x)} = 1$$

and thus $k(x,y)$ is a density. Furthermore we have

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A K(x,B)dP(x) = \int_A f_X(x)\left(\int_B k(x,y)dy\right)dx = \int_A f_X(x)\left(\int_B \frac{f(x,y)}{f_X(x)}dy\right)dx \\ &= \int_A \int_B f(x,y)dxdy \end{aligned}$$

and this concludes the proof since the measure is uniquely determined by its value on rectangles (by a monotone class theorem argument).

□

2.4 Example: Box Muller algorithms

We derive here a method to generate two independent normal random variables using two independent random number. This is a different algorithm that the one using the quantile for the normal RV which is known only numerically.

Theorem 2.5 Suppose that U_1 and U_2 are two independent random numbers then

$$X_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \qquad X_2 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$$

are two independent normal random variables with mean 0 and variance 1.

Proof. We use the expectation rule together with polar coordinates. For any nonnegative function $h(x_1, x_2)$ we have, using polar coordinate $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$ and then the change of variable $s = r^2/2$

$$\begin{aligned} E[h(X_1, X_2)] &= \int h(x_1, x_2) f(x_1, x_2) dx_1 dx_2 = \int_{\mathbb{R}^2} h(x_1, x_2) \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 \\ &= \int_{(0,\infty]\times[0,2\pi]} h(r\cos\theta, r\sin\theta) \frac{1}{2\pi} d\theta r e^{-\frac{r^2}{2}} dr \\ &= \int_{(0,\infty]\times[0,2\pi]} h(\sqrt{2s}\cos\theta, \sqrt{2s}\sin\theta) \frac{1}{2\pi} d\theta e^{-s} ds \end{aligned}$$

This computation shows that if S is exponential with parameter 1 and Θ is uniform on $[0, 2\pi]$ then $\sqrt{2S}\cos(\Theta)$ and $\sqrt{2S}\sin(\Theta)$ are independent standard normal. But we can write also $S = -\ln(U_1)$ and $\Theta = 2\pi U_2$. $\square$.

2.5 Exponential mixture of exponential is polynomial

Let us consider an exponential random variable Y whose parameter is itself an exponential random variable X with parameter $\lambda > 0$. That is X has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases}$$

and consider the kernel $K(x, dy) = k(x, y)dy$ with

$$k(x, y) = \begin{cases} xe^{-xy} & x > 0, y > 0 \\ 0 & \text{else} \end{cases}$$

Thenthe random variables (X, Y) have the joint density

$$f(x, y) = \begin{cases} \lambda e^{-\lambda x} x e^{-xy} = \lambda x e^{-(\lambda+y)x} & x > 0, y > 0 \\ 0 & \text{else} \end{cases}$$

Then, using that the mean of an exponential RV is the reciprocal of the paramter, the density of Y is

$$f(y) = \int_0^\infty f(x, y)dx = \frac{\lambda}{\lambda + y} \int_0^\infty x(\lambda + y)e^{-(\lambda+y)x} = \frac{\lambda}{(\lambda + y)^2}$$

which decays polynomially! In particular $E[Y] = \infty$.

2.6 gamma and beta random variables

Recall that a gamma RV has density $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x \geq 0$.

Consider now two independent gamma random RV X_1 and X_2 with paramters (α_1, β) and (α_2, β). We prove the following facts

1. $Z = X_1 + X_2$ is a gamma RV with parameters $(\alpha_1 + \alpha_2, \beta)$
2. $U = \frac{X_1}{X_1+X_2}$ is a beta distribution with paramters α_1 and α_2 which has the density

$$f(u) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1 - 1}(1-u)^{\alpha_2} \quad 0 \leq u \leq 1$$

3. $X_1 + X_2$ and $\frac{X_1}{X_1+X_2}$ are independent RV

We use the expectation rule **?@thm-expectationrule** and the change of variable $z = x_1 + x_2$ and $u = \frac{x_1}{x_1+x_2}$ or $x_1 = uz$ and $x_2 = (1-u)z$. This maps $[0, \infty) \times [0, \infty)$ to $[0, 1] \times [0, \infty)$ and the Jacobian of this transformation is equal z.

We have then for any nonnegative h

$$\begin{aligned}
E[h(Z,U)] &= E\left[h\left((X_1+X_2, \frac{X_1}{X_1+X_2}\right)\right] \\
&= \int_0^\infty \int_0^\infty h\left(x_1+x_2, \frac{x_1}{x_1+x_2}\right) \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} e^{-\beta(x_1+x_2)} dx_1 dx_2 \\
&= \int_0^1 \int_0^\infty h(z,u) \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (uz)^{\alpha_1-1} ((1-u)z)^{\alpha_2-1} e^{-\beta z} z du dz \\
&= \int_0^1 h(z,u) \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1-1}(1-u)^{\alpha_2-1} du \int_0^\infty \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} z^{\alpha_1+\alpha_2-1} e^{-\beta z} dz
\end{aligned}$$

and this proves all three statements at once.

Remark This is a nice, indirect way, to compute the normalization for the density of the β distribution which is proportional to $u^{\alpha_1-1}(1-u)^{\alpha_2-1}$.

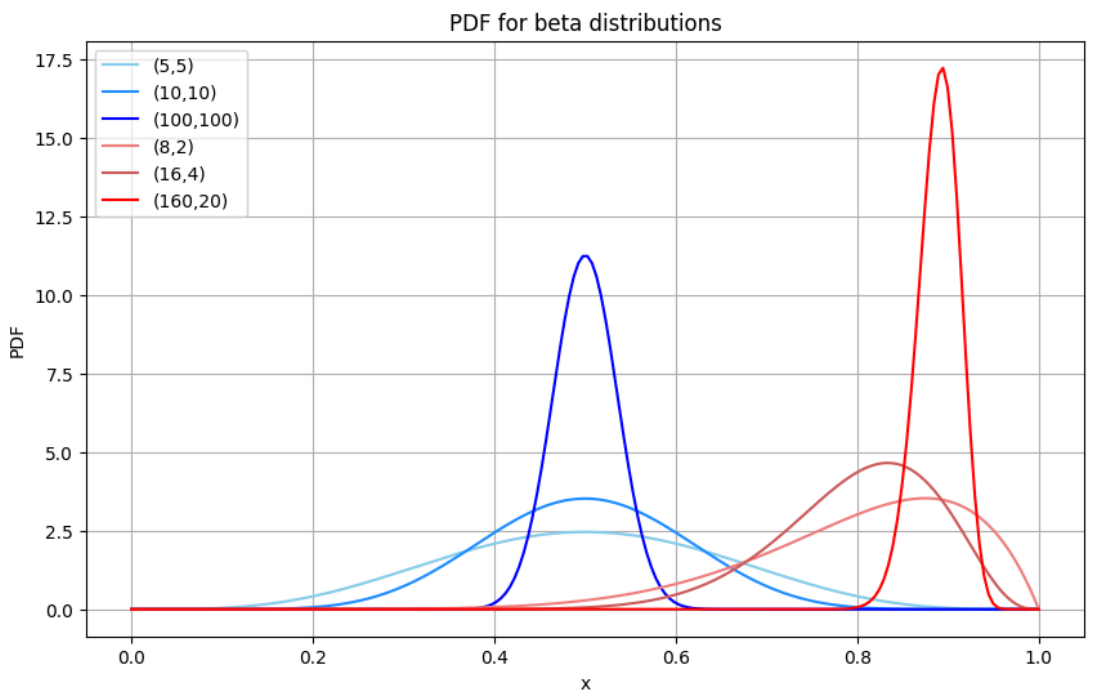
2.7 The beta binomial model and a whiff of Bayesian statistics

A beta random variable U has mean

$$E[U] = \int_0^1 u \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1-1}(1-u)^{\alpha_2-1} du = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1+1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2+1)} = \frac{\alpha_1}{\alpha_1+\alpha_2}$$

and proceeding similarly one finds that

$$\mathrm{Var}(U) = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)} = \frac{\alpha_1}{\alpha_1+\alpha_2}\frac{\alpha_2}{\alpha_1+\alpha_2}\frac{1}{\alpha_1+\alpha_2+1}$$



There is a natural connection between a binomial random variable (discrete) and the beta random variable (continuous) (let us call it P for a change). The pdf looks strangely similar

$$\binom{n}{j} p^j (1-p)^{n-j} \qquad \text{versus } \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1-1}(1-p)^{\alpha_2-1}$$

But p is a parameter for the binomial while it is a variable for the second!

Model:

- Make n independent trials, each with a (random) probability P.
- Take P to have a beta distribution with suitably chosen parameter α_1, α_2.
- The mean $\frac{\alpha_1}{\alpha_1+\alpha_2}$ is your average guess for the "true" probability p and by adjusting the scale you can adjust the variance (uncertainty) associated with your guess.

This leads to considering a random variable (X, P) taking value in $\{0, 1, \cdots, n\} \times [0, 1]$ with a "density"

$$f(j,p) = \underbrace{\binom{n}{j} p^j (1-p)^{n-j}}_{=k(p,j) \text{ kernel}} \underbrace{\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1-1}(1-p)^{\alpha_2-1}}_{=f(p)}$$

which is normalized $\sum_{j=0}^{n} \int_0^1 f(k,p)dp = 1$

The beta-binomial distribution with parameters (n, α_1, α_2) is the marginal distribution of X on $\{0, 1, \cdots, n\}$ given by

$$\begin{aligned} f(j) = \int_0^1 f(j,p)dp &= \binom{n}{j} \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 p^{\alpha_1+n-1}(1-p)^{\alpha_2+n-k-1} \\ &= \binom{n}{j} \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1+j)\Gamma(\alpha_2+n-j)}{\Gamma(\alpha_1+\alpha_2+n)} \end{aligned}$$

Bayesian statistics framework

- We interpret the distribution $f(p)$ with parameter α_1 and α_2 as the prior distribution which describe our beliefs *before* we do any experiment. For example $\alpha_1 = \alpha_2 = 1$ correspond to a uniform distribution on p (that is we are totally agnostic, a fine choice if you know nothing).
- The beta-binomial which is the marginal $f(j) = \int_0^1 f(j,p)dp$ describe the distribution of the independent trials under this model. It is called the evidence.
- The kernel $k(p, j)$ is called the likelihood which describes the number of success, j, given a certain probability of success, p. It is called the likelihood when we view it as a function of p and think of j as a parameter.
- Now we can write the distribution using kernels in two ways:

$$f(j,p) = k(p,j)f(p) = k(j,p)f(j)$$

and the kernel $k(j, p)$ is called the posterior distribution. It is interpreted as the distribution of p given that j trials have occured.

- We can rewrite this (this is just a version of *Bayes theorem*) as

$$\text{posterior} = k(j,p) = \frac{k(p,j)f(p)}{f(j)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

- For the particular model at hand

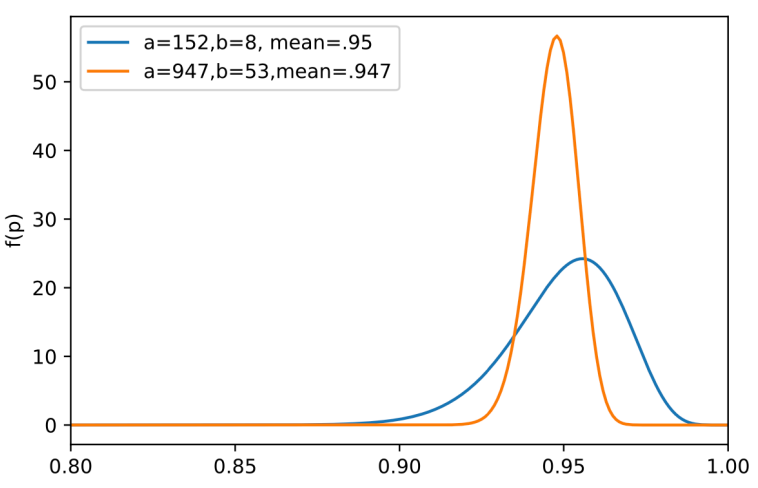
$$k(j,p) \propto p^j(1-p)^{n-j}p^{\alpha_1-1}(1-p)^{\alpha_2-1} \propto p^{\alpha_1+j-1}(1-p)^{\alpha_2+n-j-1}$$

  and therefore $k(j,p)$ has a binomial distribution with parameter $\alpha_1 + j$ and $\alpha_2 + n - j$.

- **This is special: the prior and posterior distribution belong to the same family**. We say that we have conjugate priors and this is a simple model to play. with. In general we need Monte-Carlo Markov chains to do the job.

- Example: Amazon seller ratings You want to buy a book online

Vendor 1 has 151 positive rating and 7 negative rating (95.5%).
Vendor 2 has 946 positive ratings and 52 negative ratings (94.7%).
Uniform prior with $\alpha_1 = \alpha_2 = 1$ gives two beta posterior with $\alpha_1 = 152$ and $\alpha_2 = 8$ and $\alpha_1 = 947$ and $\alpha_2 = 53$.



2.8 Exercises

Exercise 2.1

1. Suppose that X is a gamma RV with parameter α and β and Z is an exponential RV with parameter 1 and suppose further that Z and X are independent. Find the CDF (and then the PDF) of $Y = \frac{Z}{X}$.
2. Proceeding as in Section 2.5 consider an exponential RV Y whose parameter X has a gamma distribution with parameters α and β. Find the marginal distribution of Y.
3. Compare 1. and 2. and explain why they give the same result.

Exercise 2.2 (Poisson-Gamma model) Consider a Poisson RV X with a random parameter Λ which itself has a gamma distribution (for some parameters (α, β).)

1. What are the joint density and the kernel, $f(j, \lambda) = k(\lambda, j) f(\lambda)$, for the pair (X, Λ).
2. What is the density $f(j)$ of X? (the "evidence")?
3. What is the "posterior distribution" (that is what is the kernel $k(j, \lambda)$ if we write $f(j, \lambda) = k(j, \lambda) f(j)$)?

Exercise 2.3 (Ratio distribution) Suppose (X, Y) have a joint pdf $f(x, y)$. We want to compute the distribution of the ratio $Z = \frac{X}{Y}$. Show that Z has the pdf given by

$$f(z) = \int_{-\infty}^{\infty} |v| f(zv, v) dv$$

Hint; Consider the transformation $Z = \frac{X}{Y}$ and $V = Y$ and then compute either $E[h(Z)]$ (the expectation rule) or $P(Z \leq t)$ (the CDF) using this change of variables.

Use this to show that the ratio of standard normal random variable has a Cauchy distribution.

Exercise 2.4 ($Z = XY$)

1. Suppose X is a real-valued random variable with pdf $f(x)$. Let Y be a random variable taking value in $\{1, 2, 3, \cdots\}$ with $P(Y = k) = p_k$. Show that the random variable Z has a density and compute it.
2. Same setting as part 1. except that $P(Y = 0) = p_0 > 0$. What is the CDF of Z?

Exercise 2.5 (Order statistics) Suppose $(X_1, X_2, \cdots, X_n)$ are independent random variables, each with common density $f(x)$. The order statistics of $(X_1, \cdots, X_n)$ is given by

$$X_{(1)} = \min_k X_k, \quad X_{(2)} = \text{ second smallest of } X_1, \cdots, X_n, \quad \cdots\cdots X_{(n)} = \max_k X_k$$

that the $X_{(k)}$ are the same as X_k but arranged in increasing order.

- Show that the joint density of $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ is given by

$$g(x_1, \cdots, x_n) = \begin{cases} n! f(x_1) \cdots f(x_n) & \text{if } x_1 \leq x_2 \leq \cdots \leq x_n \\ 0 & \text{else} \end{cases}$$

- Show that the density of $X_{(k)}$ is given by

$$g_{(k)}(x) = k \binom{n}{k} f(x)(1 - F(x))^{n-k} F(x)^{k-1}$$

  where F denotes the CDF of X_i.

- What is the distribution of $X_{(k)}$ if the X_i are uniform on $[0, 1]$?

3 Conditional expectation

3.1 Conditioning on a random variable

Recall the definition of conditional expectation for discrete random variables X, Y. We define for integrable function $g(X, Y)$

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

and define $E[g(X,Y)|Y]$ has the random variable of the form $\psi(Y)$ such that $\psi(j) = E[g(X,Y)|Y=j]$.

It is not difficult to check that for any h which is bounded and non-negative we have

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

This property of the conditional expectation $E[E[g(X,Y)|Y]h(Y)] = E[g(X,Y)h(Y)]$ will be our starting point to define the conditional expectation.

First we prove a simple but important result

Theorem 3.1 Suppose Y is random variable taking value in $(E, \mathcal{E})$. A random variable Z taking value in $(\mathbb{R}, \mathcal{B})$ is measurable with respect to $\sigma(Y)$ if and only if $Z = h(Y)$ for some measurable function $h : E \to \mathbb{R}$.

Proof. If $Z = h(Y)$ then since $\sigma(Z) = \sigma(h(Y)) \subset \sigma(Y)$, Z is measurable with respect to $\sigma(Y)$.

Let us now consider a set $A \in \sigma(Y)$ then $A = Y^{-1}(B)$ for some $B \in \mathcal{E}$. Then $1_A = 1_{Y^{-1}(B)} = 1_B(Y)$. Therefore any simple function, measurable with respect to $\sigma(Y)$ has the form $\sum_{k=1}^M a_k 1_{A_k} = \sum_{k=1}^M a_k 1_{B_k}(Y) = h(Y)$ with $h = \sum_{k=1}^M a_k 1_{B_k}$.

Suppose Z is measurable with respect to $\sigma(Y)$. Then we can write $Z = Z_+ - Z_-$ where $Z_\pm$ are non-negative and measurable with respect to $\sigma(Y)$ so without loss of generality we can assume Z is non-negative. We can write then $Z = \sup_n Z_n$ where Z_n is an increasing sequence of simple function which are measurable with respect to $\sigma(Y)$.

To conclude we have $Z_n(\omega) = h_n(Y(\omega)) \leq Z_{n+1} = h_{n+1}(\omega))$ and so with $h = \sup_n h_n$ we can write $Z = h(Y)$. □

Definition 3.1 (Conditional expectation) Suppose that either $Z \geq 0$ or $Z \in L^1(P)$ and let be Y some random variable. The **conditional expectation** expectation $E[Z|Y]$ is a random variable which is measurable with respect to $\sigma(Y)$ (i.e. it can be written as $\psi(Y)$ for some measurable ψ) such that

$$E[E[Z|Y]U] = E[ZU] \tag{3.1}$$

for all bounded and $\sigma(Y)$-measurable function U.
The random variable $E[Z|Y]$ is essentially unique in the sense that if both $\psi_1(Y)$ and $\psi_2(Y)$ satisfy the requirements then $\psi_1(Y) = \psi_2(Y)$ almost surely.

We have the following theorem which shows that the previous definition makes sense.

Theorem 3.2 For $Z \geq 0$ or $Z \in L^1(P)$ the conditional expectation $E[Z|Y]$ exists and is essentiall unique.

Proof. The existence will be established later in a slighlty more general framework. As for uniqeness let us assume that $\psi_1(Y)$ and $\psi_2(Y)$ do satisfy the requirement. Then any bounded function of Y, $U = h(Y)$ we have we have

$$E[\psi_1(Y)U] = E[\psi_2(Y)U]$$

We now pick $U = 1_{\{\psi_1(Y) > \psi_2(Y)|\}}$ and thus we obtain $E[(\psi_1(Y) - \psi_2(Y))1_{\psi_1(Y)\psi_2(Y)>0}] = 0$.
Therefore the random variable $(\psi_1(Y) - \psi_2(Y))1_{\psi_1(Y)\psi_2(Y)>0}$, which is non-negative, must be equal to 0, almost surely and thus $\psi_1(Y) \leq \psi_2(Y)$, almost surely. By symmetry $\psi_1(Y) = \psi_2(Y)$, almost surely. $\square$.

3.2 Examples of conditional expectations: conditional distributions

Theorem 3.3 Suppose (X, Y) are random variables talking values in $E \times F$ with probability distribution $P^{(X,Y)}$ on $(E \times F, \mathcal{E} \times \mathcal{F})$. Write $P^{(X,Y)}$ using a probability $P(A)$ on E and a kernel $K(x, B)$. Then, if $g(Y)$ is integrable, we have

$$E[g(Y)|X] = \int_F g(y) dK(X, y)$$

For example if $P^{(X,Y)}$ has a joint density $f(x, y)$ then $E[g(Y)|X] = \int_E g(y) \frac{f(X,y)}{f(X)} dy$.

Proof. Let $\psi(x) = \int_F g(y) dK(x, y)$. Since $g \in L^1$ is integrable we have

$$E[|\psi(X)|] = \int_E |\int_F g(y) dK(x, y)| dP(x) \leq \int_E \int_F |g(y)| dK(x, y) dP(x) = E[|g(Y)|] < \infty$$

and thus the RV $\psi(X)$ is integrable.

If $U = h(X)$ to be bounded, then $\psi(X)h(X)$ is integrable. We have

$$E[\psi(X)h(X)] = \int_E \int_F g(y) dK(x, y) h(x) dP(x) = \int g(y) h(x) dP^{(X,Y)}(x, y)$$

and this proves that $\psi(X) = E[g(Y)|X]$. $\square$

3.3 More examples

Example Suppose Y is a discrete random variable taking value in $\mathbb{N}$ with pdf $p(n)$ and Z is a real-valued continous random variable with density $f(z)$ and Z is independent of Y. Let

$$X = Y + Z$$

You can think of X has the discrete random variable Y corrupted by the "noise" Z. We want to compute $E[h(X)|Y]$.

We can apply Theorem 3.3. We compute the kernel

$$K(n, A) = P(X \in A|Y = n) = P(Y + Z \in A|Y = n) = P(Z + n \in A) = P(Z \in A - n) = \int_A f(z + n$$

so that the kernel is described by the density $k(n, x) = f(n + x)$. Therefore

$$E[h(X)|Y = n] = \int h(x)f(n + x)dx = \int h(x - n)f(x)dx = E[h(Z - n)]$$

and thus $E[h(X)|Y] = E[h(Z - Y)]$ where the expectation on the r.h.s is with respect to Z.

3.4 Properties of conditional expectations

Here are the basic rules to manipulate conditional expectations

Theorem 3.4 (rules of conditional expectation) Suppose Y, Y_1, Y_2 are random variables and Z, Z_1, Z_2 are non-negative or integrable random variables. Then we have

1. **Linearity:** $E[\alpha_1 Z_1 + \alpha_2 Z_2|Y] = \alpha_1 E[Z_1|Y] + \alpha_2 E[Z_2|Y]$ almost surely.
2. **Independence:** If Z is independent of Y then $E[Z|Y] = E[Z]$ almost surely.
3. **Y-measurability**: If Z is measurable with respect to $\sigma(Y)$ then $E[Z|Y] = Z$ almost surely.
4. **Monotonicity:** If $Z_1 \leq Z_2$ then $E[Z_1|Y] \leq E[Z_2|Y]$ almost surely. In particular if $Z \geq 0$ then $E[Z|Y] \geq 0$ almost surely.
5. **Tower property:** We have $E[E[Z|(Y_1, Y_2)]|Y_2] = E[Z|Y_2]$ almost surely.
6. **Conditional Jensen**: If ϕ is a convex function then $E[\phi(Z)|Y] \geq \phi(E[(Z)|Y])$ almost surely.
7. **Product property** : If $Z = \tilde{Z}h(Y)$ where $\tilde{Z}$ is integrable (resp. non-negative) and $h(Y)$ is bounded (resp. non-negative) then $E[\tilde{Z}h(Y)|Y] = h(Y)E[\tilde{Z}|Y]$ almost surely.

Proof. All these proofs are similar and follow from the fact the conditional expectation is the (essentially) unique random variable satisfying Equation 3.1

Item 1. (Linearity): By definition $E[\alpha_1 Z_1 + \alpha_2 Z_2 | Y]$ is the unique, $\sigma(Y)$ measurable RV such that for $U = h(Y)$

$$E\left[E[\alpha_1 Z_1 + \alpha_2 Z_2 | Y]U\right] = E[(\alpha_1 Z_1 + \alpha_2 Z_2)U]$$

But by linearity of expectation and using the definition of conditional expectation $E[Z_i|U]$ we have

$$\begin{aligned} E[(\alpha_1 Z_1 + \alpha_2 Z_2)U] = \alpha_1 E[Z_1 U] + \alpha_2 E[Z_2 U] &= \alpha_1 E[E[Z_1|Y]U] + \alpha_2 E[E[Z_2|Y]U] \\ &= E[(\alpha_1 E[Z_1|Y]U + \alpha_2 E[Z_2|Y])U] \end{aligned}$$

From the uniqueness of condtional expectation we have $E[\alpha_1 Z_1 + \alpha_2 Z_2|Y] = \alpha_1 E[Z_1|Y]U + \alpha_2 E[Z_2|Y]$ almost surely.

Item 2. (Independence): If Z and Y are independent then Z and $U = h(Y)$ are also independent and so

$$E[E[Z|Y]U] = E[ZU] = E[Z]E[U] = E[E[Z]U]$$

From the uniqueness of condtional expectation we obtain that $E[Z|Y] = E[Z]$.

Item 3. ($\sigma(Y)$-measurability): If Z is measurable with respect to $\sigma(Y)$ then $E[Z|Y] = Z$ by definition.

Item 4. (Montonicity): If $Z_1 \leq Z_2$ take $U = h(Y) \geq 0$. We have then

$$E[E[Z_1|Y]U] = E[Z_1U] \leq E[Z_2U] = E[E[Z_2|Y]U]$$

Take now $U = 1_{\{E[Z_2|Y]<E[Z_1|Y]\}}$ and so we have

$$E[(E[Z_2|Y] - E[Z_1|Y])1_{\{E[Z_2|Y]<E[Z_1|Y]\}}] \geq 0$$

which implies that $(E[Z_2|Y] - E[Z_1|Y])1_{\{E[Z_2|Y]<E[Z_1|Y]\}} = 0$ almost surely and thus $E[Z_2|Y] > E[Z_1|Y]$ almost surely.

Item 5. (Tower property): If U is measurable with respect to $\sigma(Y_2)$ then $U = h(Y_2)$ and thus U is also measurable with respect to (Y_1, Y_2). Then we have, using the definition of conditional expectation three times,

$$E[E[E[Z|(Y_1, Y_2)]|Y_2]h(Y_2)] = E[E[Z|(Y_1, Y_2)]h(Y_2)] = E[Zh(Y_2)] = E[E[Z|Y_2]h(Y_2)]$$

and thus $E[E[Z|(Y_1, Y_2)]|Y_2] = E[Z|Y_2]$ almost surely.

Item 6. (Conditional Jensen): Left as a homework. Revisit the roof of Jensen inequality.

Item 7. (Product) If $\tilde{Z}$ is integrable and $g(Y)$ is bounded then $\tilde{Z}g(Y)$ is integrable and using the definition with $U = g(Y)h(Y)$

$$E[E[\tilde{Z}g(Y)|Y]h(Y)] = E[\tilde{Z}g(Y)h(Y)] = E[E[\tilde{Z}|Y]g(y)h(Y)] = E[g(y)E[\tilde{Z}|Y]h(Y)]$$

and thus $E[\tilde{Z}g(Y)|Y] = g(Y)E[g(Y)|Y]$ almost surely.

3.5 Conditioning on a σ-algebra

The general theory deals with conditional expectation with respect to σ-algebra.

Definition 3.2 Let Z be an integrable (resp. finite non-negative) random variable, and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. A version of the conditional expectation of Z given $\mathcal{G}$ is any integrable (resp. finite non-negative) $\mathcal{G}$-measurable random variable, denoted by $E[Z|\mathcal{G}]$ such that

$$E[ZU] = E[E[Z|\mathcal{G}]U]$$

for all bounded (resp. bounded non-negative) $\mathcal{G}$-measurable random variables U.

Theorem 3.5 Let Z be an integrable (resp. finite non-negative) random variable, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{A}$. The the conditional expectation $E[Z|\mathcal{G}]$ exists and is essentially unique, that is, two versions of the conditional expectation of Z given $\mathcal{G}$ are equal, almost surely.

Proof. The uniqueness part of the statement is proved as in Theorem 3.2 and will not be repeated. The existence part, which we had not been proved in Theorem 3.2 is now established as a consequence of Radon-Nikoym theorem.

To do this we consider firt the case where Z is non-negative and integrable. Define now the measure

$$Q(A) = \int_A Z dP \quad \text{for } A \in \mathcal{G}$$

on the measure space $(\Omega, \mathcal{G})$. Clearly Q is absolutely continuous with respect to P (more properly the restriction of P on the sub-σ-algebra $\mathcal{G}$)) and therefore there exists a random variable on $(\Omega, \mathcal{G})$, we we which denote by $E[Z|\mathcal{G}]$ such that

$$Q(A) = \int_A E[Z|\mathcal{G}] dP$$

Using the two representations of $Q(A)$ we see that for any non-negative U measurable with respect to $\mathcal{G}$ we have

$$\int U dQ = E[UZ] = E[E[Z|\mathcal{G}]U]$$

This is extended to integrable random variables by decomposing Z into positive and negative part and for general non-negative random variable by a monotone convergence argument.

Theorem 3.6 (rules of conditional expectation) Let Z, Z_1, Z_2 are non-negative or integrable random variables and $\mathcal{G}$ and $\mathcal{H}$ some sub-σ-algebras.
Then we have

1. **Linearity:** $E[\alpha_1 Z_1 + \alpha_2 Z_2|\mathcal{G}] = \alpha_1 E[Z_1|\mathcal{G}] + \alpha_2 E[Z_2|\mathcal{G}]$ almost surely.
2. **Independence:** If Z is independent of $\mathcal{G}$ then $E[Z|\mathcal{G}] = E[Z]$ almost surely.
3. **Y-measurability**: If Z is measurable with respect to $\mathcal{G}$ then $E[Z|\mathcal{G}] = Z$ almost surely.
4. **Monotonicity:** If $Z_1 \leq Z_2$ then $E[Z_1|\mathcal{G}] \leq E[Z_{\mathcal{G}}|Y]$ almost surely. In particular if $Z \geq 0$ then $E[Z|Y] \geq 0$ almost surely.
5. **Tower property:** If $\mathcal{H} \subset \mathcal{G}$ we have $E[E[Z|\mathcal{G}]|\mathcal{H}] = E[Z|\mathcal{H}]$ almost surely.
6. **Conditional Jensen**: If ϕ is a convex function then $E[\phi(Z)|\mathcal{G}] \geq \phi(E[(Z)|\mathcal{G}])$ almost surely.
7. **Product property** : If $Z = \tilde{Z}V$ where $\tilde{Z}$ is integrable (resp. non-negative) and V is bounded (resp. non-negative)and measurable with respect to $\mathcal{G}$ then $E[\tilde{Z}V|\mathcal{G}] = VE[\tilde{Z}|\mathcal{G}]$ almost surely.

Proof. Same as before.

3.6 Example

Suppose X_i is a sequence of IID random variables where X_i are integrable with $E[X_i]\bar{\mu}$ and let $S_n = X_1 + \cdots + X_n$.

We now consider the σ-algebras

$$\mathcal{F}_n = \sigma(X_1, \cdots, X_n)$$

and the σ-algebras $\mathcal{F}_n$ are increasing $\mathcal{F}_n \subset \mathcal{F}_{n+m}$.

Let us compute $E[S_{n+m}|\mathcal{F}_n]$.

We have

$$\begin{aligned} E[S_{n+m}|\mathcal{F}_n] = E[S_n + X_{n+1} + \cdots + X_{n+m}|\mathcal{F}_n] &= E[S_n|\mathcal{F}_n] + E[X_{n+1} + \cdots + X_{n+m}|\mathcal{F}_n] \\ &= S_n + E[X_{n+1} + \cdots + X_{n+m}] = S_n + m\mu \end{aligned}$$

since S_n is measurable with respect to $\mathcal{F}_n$ and $X_{n+1}, \cdots, X_{n+m}$ are independent of $\mathcal{F}_n$.

This shows that

$$E[S_{n+m} - (n+m)\mu|\mathcal{F}_n] = S_n - n\mu.$$

This is the martingale property which we will explore later on. If set $M_n = S_n - n\mu$ then we have

$$E[M_{n+m}|\mathcal{F}_n] = M_n\,.$$

3.7 Conditional expectation as a projection

Another way to think of the conditional expectation is to see it as acting on vector space of random variables (defined almost surely). For example let us consider the vector spaces

$$
\begin{aligned}
L^+(\Omega, \mathcal{A}, P) &= \{X : \Omega \to [0, \infty) \text{ measurable}\} \\
L^1(\Omega, \mathcal{A}, P) &= \{X : \Omega \to \mathbb{R} \text{ measurable } : \ E[X] < \infty)\} \\
L^2(\Omega, \mathcal{A}, P) &= \{X : \Omega \to \mathbb{R} \text{ measurable } : \ E[X^2] < \infty)\}
\end{aligned}
$$

We can think as the conditional expectation as a map. For example

$$
\begin{aligned}
\Pi_{\mathcal{G}} : & L^1(\Omega, \mathcal{A}, P) \to L^1(\Omega, \mathcal{G}, P) \\
& Z \mapsto E[Z|\mathcal{G}]
\end{aligned}
$$

which projects $L^1(\Omega, \mathcal{A}, P)$ to its linear subspace $L^1(\Omega, \mathcal{A}, P)$.

Here the meaning of projection means that $\Pi_{\mathcal{G}}$ is a linear map and satisfies $\Pi_{\mathcal{G}}^2 = \Pi_{\mathcal{G}}$. The range of $\Pi_{\mathcal{G}}$ is the space onto which it projects, in this case $L^1(\Omega, \mathcal{G}, P)$.

Let us consider the (smaller) space $L^2(\Omega, \mathcal{A}, P)$ which is equiped with the inner product $\langle X\ ,\ Y\rangle = E[XY]$. We can give a proof of the existence of conditional expectation (for $Z \in L^2$) using Riesz-Fisher Theorem (the same theorem we used to prove Radon-Nokodym theorem). The agument goes as follows: fix $Z \in L^2(\Omega, \mathcal{A}, P)$ and consider the map ϕ_Z acting on $L^2(\Omega, \mathcal{G}, P)$.

$$\begin{aligned} \phi_Z\ &: L^2(\Omega, \mathcal{G}, P) \to \mathbb{R} \\ &U \mapsto \phi_Z(U) = E[ZU] \end{aligned}$$

We have, by Cauchy-Schwartz, in the space $L^2(\Omega, \mathcal{A}, P)$,

$$|\phi_Z(U)| = |E[ZU]| = \langle Z\ ,\ U\rangle_{L^2(\Omega,\mathcal{A},P)} \leq \|Z\|_{L^2(\Omega,\mathcal{A},P)}\|U\|_{L^2(\Omega,\mathcal{A},P)} = \|Z\|_{L^2(\Omega,\mathcal{A},P)}\|U\|_{L^2(\Omega,\mathcal{G},P)}$$

where we have used that $L^2(\Omega, \mathcal{G}, P)$ is a linear subspace of $L^2(\Omega, \mathcal{A}, P)$.

We can now apply Riesz-Fisher theorem, in $L^2(\Omega, \mathcal{A}, P)$ to conclude that there exists an element $L^2(\Omega, \mathcal{G}, P)$ such that

$$E[ZU] = E[E[Z|\mathcal{G}]U] \quad \text{for all } U \in L^2(\Omega, \mathcal{G}, P)$$

In that case the map $Z \mapsto \Pi_{\mathcal{G}}(Z) = E[Z|U]$ is an orthogonal projection in $L^2(\Omega, \mathcal{G}, P)$. Indeed using the property of conditional expectation we find that

$$\langle \Pi_{\mathcal{G}}(Z)\ ,\ Z'\rangle_{L^2(\Omega,\mathcal{G},P)} = E[Z'E[Z|\mathcal{G}]] = E[E[Z'|\mathcal{G}]E[Z|\mathcal{G}]] = E[ZE[Z'|\mathcal{G}]] = \langle Z\ ,\ \Pi_{\mathcal{G}}(Z')\rangle_{L^2(\Omega,\mathcal{G},P)}.$$

3.8 Rejection sampling

The rejection sampling is a basic simulation method to generate random variables taking values in $\mathbb{R}^n$ (discrete or continuous). We will prove the continuous case but the discrete case is similar and is left to the reader. The basic insight is that you assume that there exists a random variable X with density $f(x)$ that you already know how to simulate. You use this simulate a random variable with density $g(y)$ by suitable reweighting.

Theorem 3.7 (Rejection Method) Suppose that the random variable X has pdf $f(x)$ and Y has pdf $g(x)$, both taking values in $\mathbb{R}^n$, and that there exists a constant C such that

$$\frac{f(y)}{g(y)} \leq C \qquad \text{for all } y$$

To generate a sample from X do

- Step 1: Generate the random variable Y.
- Step 2: Generate a random number U.
- Step 3: If $U \leq \frac{f(Y)}{g(Y)C}$ set $X = Y$ otherwise reject and go back to Step 1.

The number of times the algorithm runs until a value for X is accepted is geometric with parameter $\frac{1}{C}$.

Proof. To obtain one value of X we need iterate the algorithm a random number of times, let us call it N, until the value is accepted. That is we generate independent random variables $Y_1, \cdots, Y_N$ until Y_N is accepted and then set $X = Y_N$.

Let us compute the CDF of Y_N. We have, by conditioning on Y, for any measurable A

$$P\{Y_N \in A\} = P\left\{Y \in A \,|\, U \leq \frac{f(Y)}{Cg(Y)}\right\} = \frac{P\left\{Y \in A, U \leq \frac{f(Y)}{Cg(Y)}\right\}}{P\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} \tag{1}$$

$$= \frac{E[E[1_{\{Y\in A\}} 1_{\{U \leq \frac{f(Y)}{Cg(Y)}\}} | Y]]}{P\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} = \frac{E[1_{\{Y\in A\}} E[1_{\{U \leq \frac{f(Y)}{Cg(Y)}\}} | Y]]}{P\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} = \frac{E[1_{\{Y\in A\}} \frac{f(Y)}{Cg(Y)}]}{P\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} \tag{2}$$

$$= \frac{\int_A \frac{f(y)}{Cg(y)} g(y) dy}{CP\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} = \frac{\int_A f(y) dy}{CP\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} = \frac{P(X \in A)}{CP\left\{U \leq \frac{f(Y)}{Cg(Y)}\right\}} \tag{3}$$

If we take $A = \mathbb{R}^n$ we find that the denominator is $CP\left(U \leq \frac{f(Y)}{Cg(Y)}\right) = 1$ and thus, as desired, that Y_N has the same distribution as X.

The above argument also shows that at each iteration, a value for X is accepted with probability $P\left(U \leq \frac{f(Y)}{Cg(Y)}\right) = \frac{1}{C}$ independently of the other iterations. Therefore the number of iterations needed is a geometric random with mean C.

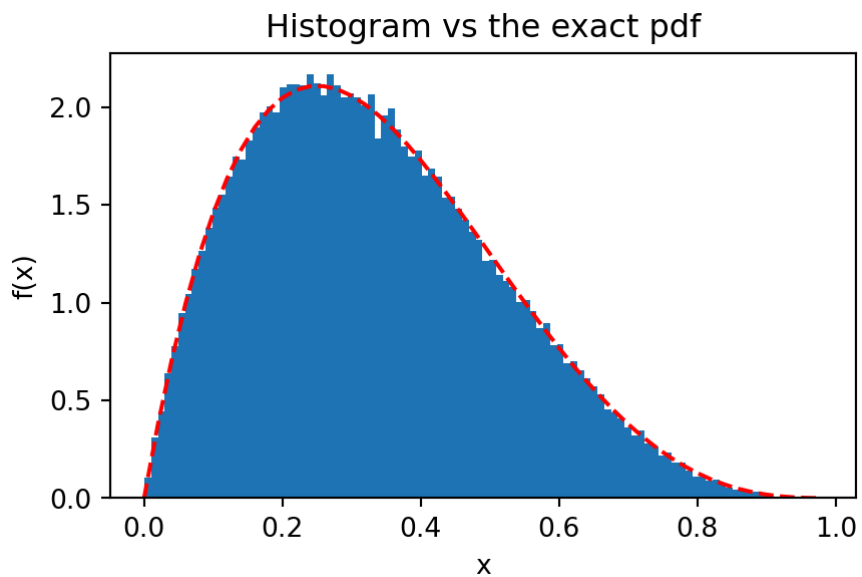
□.

3.9 Example

Suppose X has pdf $f(x) = 20x(1-x)^3$ for $0 \leq x \leq 1$. Since X is supported on $[0,1]$ we pick Y uniform on $[0,1]$ with $g(y) = 1$. Then $C = \max \frac{f(x)}{g(x)} = \max_{x \in [0,1]} 20x(1-x)^3 = \frac{135}{64}$. So generate two random numbers U_1 and U_2 and if $U_1 \leq \frac{256}{27} U_2 (1-U_2)^3$ set $X = U_2$. The average proportion of accepted values is $\frac{64}{135} = .4747..$

▼ Code

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3  def accept_reject(N):   # Generate N sampl
4      n_accept=0
5      x_list = []
6      while n_accept < N:
7          a=np.random.rand(2)
8          if a[0] < (256/27)*a[1]*(1-a[1])**
9              n_accept += 1
10             x_list.append(a[1])
11     return x_list
12 plt.figure(figsize=(5,3))
13 plt.hist(accept_reject(100000), bins=100,
14 t = np.arange(0., 1., 0.02)
15 plt.plot(t, 20*t*(1-t)**3, 'r--' )
16 plt.xlabel('x')
17 plt.ylabel('f(x)')
18 plt.title('Histogram vs the exact pdf')
19 plt.show()
```



3.10 Exercises

Exercise 3.1 (Conditional Jensen inequality)

- Show that if ϕ is a convex function and Z is integrable then

$$E[\phi(Z)|Y] \geq \phi(E[(Z)|Y])$$

  almost surely.

- Show that $\Pi_{\sigma(Y)} : Z \mapsto E[Z|Y]$ defines is a projection on $L^p(\Omega, \mathcal{A}, P)$ for any $1 \leq p \leq \infty$, i.e. $\Pi_{\sigma(Y)}$ maps L^p into L^p with $\|E[Z|Y]\|_p \leq \|Z\|_p$ and $\Pi^2_{\sigma(Y)} = \Pi_{\sigma(Y)}$.

Exercise 3.2 (Monotone convergence)

- Show that if the random variables Z_n are non-negative and Z_n converges to Z almost surely then $\lim_n E[Z_n|Y] = E[Z|Y]$ almost surely.

- Formulate a conditional version for Fatou's Lemma and the Dominated convergence theorem. You do not need to prove them, as they are derived from the monotoneconvergence theorem exactly as their non-conditioned counterpart.

Exercise 3.3

- Suppose the random variable X has a density $f(X)$. What is the density of X^2?
- Show that $E[X|X^2] = |X|\frac{f(-|X|)-f(|X|)}{f(-|X|)+f(|X|)}$

Exercise 3.4 (More on L^2-projections) All random variables are in L^2 so all the expectations make sense.

Exercise 3.5 (Generating uniform RV on the unit ball and the unit sphere) We denote by Y a RV uniformly distributed on the $(n-1)$ sphere $S_{n-1} = \{x_1^2 + \cdots + x_n^2 = 1\}$. We denote by X a RV uniformly distributed on the n unit ball $B_n = \{x_1^2 + \cdots + x_n^2 \leq 1\}$. We want to write algorithms to generate X and Y, in an efficient manner.

1. Show that to generate Y it is enough to generate n IID standard normal $Z = (Z_1, \cdots, Z_n)$ (e.g using the Box-Muller algorithm) and set $Y = \frac{Z}{\|Z\|}$.

2. Build a rejection algorithm for X by using the random variable $V = (V_1, \cdots, V_n)$ where V_i are IID uniform on $[-1, 1]$.
   Compute and analyze the acceptance rate. Show that this algorithm is totally useless in practice if n is not very very small. *Hint*: The volume of the n-unit ball is $\frac{\pi^{n/2}}{n\Gamma(n/2)}$.

3. Show X can be generated in a rejection-free in the following manner. Generate Y as in 1., pick a random number U and set $X = U^{1/n}Y$. *Hint:* What is the distribution of $\|X\|$?

Exercise 3.6 The beta random variable with parameter (n, m) has the pdf

$$f(x) = \frac{(n+m-1)!}{(n-1)!(m-1)!} x^{n-1}(1-x)^{m-1} \qquad \text{for } 0 \leq x \leq 1$$

It is a good model of a unimodal distribution supported on $[0, 1]$.

1. Write down a rejection algorithm (with a uniform Y) to simulate a beta random variable. Compute the acceptance probability.
2. Consider the results in Section 2.6. Use this to write down an algorithm to generate beta random variables.
3. Which one of the two algorithms is more efficient? To do this compare the number of random numbers needed to generate one sample and use Stirling formula to obtaimn a large n asymptotics

4 The characteristic and moment generating function for a random variable

One of the oldest trick in the mathematician toolbox is to obtain properties of a mathematical object by perfomring a transformation on that object to map it into another space. In analysis (say for ODE's and PDE's) the Fourier transform and the Laplace transform play a very important role. Both play an equally important role in probability theory!

- Fourier transform of a probbaility measure leads to a proof of the *central limit theorem*!
- Laplace transform (via Chernov bounds) leads to *concentration inequalities* and performance guarantees for Monte-Carlo methods and statistical learning.

4.1 Fourier transform and characteristic function

Notation For vectors $x, y \in \mathbb{R}^n$ we use the notation

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

for the scalar product in $\mathbb{R}^n$

Definition 4.1 (Fourier transform and characteristic function)

- For a probability measure P on $\mathbb{R}^n$ the Fourier transform of P is a function $\widehat{P}(t) : \mathbb{R}^n \to \mathbb{C}$ given by

$$\widehat{P}(t) = \int_{\mathbb{R}^n} e^{i\langle t,x \rangle} dP(x)$$

- For a random variable X taking value in $\mathbb{R}^n$ the characterictic function of X is the Fourier transorm of P^X (the distribution of X): we have

$$\phi_X(t) = E\left[e^{i\langle t,x \rangle}\right] = \int_{\mathbb{R}^n} e^{i\langle t,x \rangle} dP^X(x)$$

Remarks

- We have not talked explicitly about integration of complex valued function $h = f + ig$ where f and g are the real and imaginary part. It is simply defined as

$$\int (f + ig) dP = \int f dP + i \int g dP$$

  provided f and g are integrable. A complex function h is integrable iff and only if $|h|$ is intergable if and only if f and g are integrable. (The only thing a bit hard to prove is the triangle inequality $|\int h dP| \leq \int |h| dP$.)

- The function

$$e^{i\langle t,x\rangle} = \cos(\langle t, x\rangle) + i \sin(\langle t, x\rangle)$$

  is integrable since $\sin(\langle t, x\rangle)$ and $\cos(\langle t, x\rangle)$ are bounded function (thus in $L^\infty \subset L^1$) or by noting that $|e^{i\langle t,x\rangle}| = 1$.

- Suppose the measure P has a density $f(x)$ then we have

$$\widehat{P}(t) = \int e^{i\langle t,x\rangle} f(x) dx$$

  which is simply the Fourier transform (usually denoted by $\hat{f}(t)$) of the function $f(x)$. (Diverse conventions are used for the Fourier, e.g. using $e^{-i2\pi\langle k,x\rangle}$ instead $e^{i\langle t,x\rangle}$ but these differ only by trivial rescaling).

4.2 Analytic properties of the Fourier transform

We turn next to the properties of the Fourier transform. A *very useful* thing to remember of the Fourier transform

$$\text{the smoother the Fourier transform is the faster the function (or the measure) decay and vice versa}$$

The next two theorem makes this explicit in the context of measures. The first one is very general and simply use that we dealing with probaility measure

Theorem 4.1 The Fourier transform $\widehat{P}(t)$ of a probability measure P is uniformly continuous on $\mathbb{R}$, and satisfies $\widehat{P}(0) = 1$ and $|\widehat{P}(t)| \leq 1$.

Proof. Clearly $\widehat{P}(0) = \int_{\mathbb{R}^n} dP(x) = 1$ since P is a probability measure and since $|e^{i\langle t,x\rangle}| = 1$, by the triangle inequality $|\widehat{P}(t)| \leq 1$.

For the uniform contintuity we have

$$|\widehat{P}(t+h) - \widehat{P}(t)| \leq \int |e^{i\langle (t+h),x\rangle} - e^{i\langle t,x\rangle}| dP(x) = \int |e^{i\langle t,x\rangle}||e^{i\langle h,x\rangle} - 1| dP(x) = \int |e^{i\langle h,x\rangle} - 1| dP(x)$$

The right hand side is independent of t which is going to show uniformity. To conclude we need to show that the right hand-side goes to 0 as $h \to 0$. We can use dominated convergence since

$$\lim_{h\to 0} e^{i\langle h,x\rangle} = 1 \text{ for all } x \qquad \text{and} \qquad |e^{i\langle h,x\rangle} - 1| \leq 2 \quad \square$$

As we have seen the L^p spaces are form a dcereasing sequence $L^1 \supset L^2 \supset \cdots L^\infty$ and the next theorem show that if some random variable belongs to L^m (for some integer n) then its chararcteristic function will be m-times continuously differentiable.

Theorem 4.2 Suppose X is RV taking value in $\mathbb{R}^n$ and is such that $E\left[|X|^m\right] < \infty$. Then the characteristic function $\phi_X(t) = E[e^{i\langle t, X\rangle}]$ has continuous partial derivative up to oder m, and for any $k \leq m$,

$$\frac{\partial^k \phi_X}{\partial x_{i_1} \cdots \partial x_{i_k}}(t) = i^k E\left[X_{i_1} \cdots X_{i_k} e^{i\langle t, X\rangle}\right]$$

Proof. We will prove only the $m = 1$ case, the rest is proved by a tedious induction argument. Denoting by e_i the basis element

$$\frac{\partial \phi_X}{\partial x_i}(t) = \lim_{h\to 0} \frac{\phi_X(t + he_i) - \phi_X(t)}{h} = \lim_{h\to 0} E\left[\frac{1}{h}\left(e^{i\langle t+he_i, X\rangle} - e^{i\langle t, X\rangle}\right)\right] = \lim_{h\to 0} E\left[e^{i\langle t, X\rangle}\frac{e^{ihX_i} - 1}{h}\right]$$

To exchange the limit and expectation we use a DCT argument and the bound

$$|e^{i\alpha} - 1| = \left|\int_0^\alpha \frac{d}{ds} e^{is}\right| \leq \int_0^\alpha |ie^{is}| \leq |\alpha|\,.$$

From this we see that $\left|e^{i\langle t, X\rangle}\frac{e^{ihX_i}-1}{h}\right| \leq |X_i|$ which is integrable and independent of h. The DCT concludes the proof. □

4.3 More properties

Two simple but extremly useful properties:

Theorem 4.3 If X takes value in $\mathbb{R}^n$, $b \in \mathbb{R}^m$ and A is a $m \times n$ matrix then

$$\phi_{AX+b}(t) = e^{i\langle t,b\rangle}\phi_X(A^T t)$$

Proof. This simply follows from the equality

$$e^{i\langle t,AX+b\rangle} = e^{i\langle t,b\rangle}e^{i\langle A^T t,X\rangle}. \quad \square$$

Theorem 4.4 Suppose X and Y are independent RV taking values in $\mathbb{R}^n$ then

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$$

Proof. By independence

$$E\left[e^{i\langle t,X+Y\rangle}\right] = E\left[e^{i\langle t,X\rangle}e^{i\langle t,Y\rangle}\right] = E\left[e^{i\langle t,X\rangle}\right]E\left[e^{i\langle t,Y\rangle}\right] \quad \square$$

4.4 Examples

- Bernoulli with parameter p:

$$\phi_X(t) = E[e^{itX}] = e^{it}p + e^{i0}(1-p) = (1-p) + e^{it}p\,.$$

- Binomial with paramters (n, p): using the binomial theorem

$$\phi_X(t) = E[e^{itX}] = \sum_{k=0}^{n} \binom{n}{k} e^{itk} p^k (1-p)^{n-k} = (e^{it}p + (1-p))^n$$

- Poisson with paramters λ:

$$\phi_X(t) = E[e^{itX}] = e^{-\lambda} \sum_{k=0}^{\infty} e^{itk} \frac{\lambda^k}{k!} = e^{\lambda(e^{it}-1)}$$

- Normal with paramters μ, σ^2: We start with the special case $\mu = 0$ and $\sigma^2 = 1$ and we need to compute the complex integral

$$\phi_X(t) = E[e^{itX}] = \frac{1}{\sqrt{2\pi}} \int e^{itx} e^{-x^2} 2dx$$

You can do it via contour integral and residue theorem (complete the square!). Instead we use an ODE's argument.

First we note that by symmetry

$$\phi_X(t) = \frac{1}{\sqrt{2\pi}} \int \cos(tx) e^{-x^2/2} dx$$

By Theorem 4.2 we can differentiate under the integral and find after integrating by part

$$\phi'_X(t) = \frac{1}{\sqrt{2\pi}} \int -x \sin(tx) e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int -t \cos(tx) e^{-x^2/2} dx = -t\phi_X(t)$$

a separable ODE with initial condition $\phi_X(0) = 1$. The solution is easily found to be $e^{-t^2/2}$ and thus we have

$$\phi_X(t) = E\left[e^{itX}\right] = e^{-t^2/2}.$$

Finally noting that $Y = \sigma X + \mu$ is a normal is with mean μ and σ we find by Theorem 4.3 that

$$\phi_Y(t) = e^{i\mu t} E\left[e^{i\sigma t X}\right] = e^{i\mu t - \sigma^2 t^2/2}.$$

- Exponential with parameter β: For any (complex) z we have $\int_a^b e^{zx} dx = \frac{e^{zb} - e^{za}}{z}$. From this we deduce that

$$\phi_X(t) = \beta \int_0^\infty e^{(it-\beta)x} dx = \frac{\beta}{\beta - it}$$

4.5 Uniqueness Theorem

We show now that the Fourier transform determines the probability measures uniquely, that is if two probability measures P and Q have the same Fourier transforms $\widehat{P} = \widehat{Q}$ then they must coincide $P = Q$. For simplicity we only consider the 1-d case but the proof extends without problem.

There exists several version of this proof (see your textbook for one such proof). We give here a direct proof which also gives an explcit formula on how to reconstruct the measure from the its Fourier transform.

Our proof relies on the following computation of the so-called Dirichlet integral

Lemma 4.1 (Dirichlet integral) For $T > 0$ let $S(T) = \int_0^T \frac{\sin t}{t}\, dt$. We have then

$$\lim_{T \to \infty} S(T) = \frac{\pi}{2}$$

This is a fun integral to do and can be done using a contour integral in the complex plane, or by a Laplace transform trick, or by the so-called Feynmann trick (add a parameter and differentiate). See for example the Wikipedia page.

We will take this result for granted and note that we have

$$\int_0^T \frac{\sin(\theta t)}{t} dt = \mathrm{sgn}(\theta) \mathrm{S}(|\theta| \mathrm{T})$$

where $\mathrm{sgn}(\theta)$ is $+1, 0$ or -1 is θ is positive, 0, or negative.

Theorem 4.5 (Fourier inversion formula) If a and b are not atoms for P we have

$$P((a,b]) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \widehat{P}(t)\, dt \tag{4.1}$$

In particular distinct probability measures cannot have the same characteristic function.

Proof. The inversion formula implies uniqueness. The collections of $(a,b]$ such that a and b are not atoms is a p-system which generates $\mathcal{B}$ so the monotone class theorem implies the result, see **?@thm-uniquenesspm**. (See exercise for more on atoms).

Let I_T denote the integral in Equation 4.1. Using the bound $|e^{iz} - e^{iz'}| \leq |z - z'|$s we see that that the integrand is bounded and thus by Fubini's theorem we have

$$I_T = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt \right) dP(x)$$

Using Euler formula and the fact that $\cos$ is even $\sin$ is odd we find

$$\begin{aligned} I_T &= \int_{-\infty}^{\infty} \left(\int_{0}^{T} \frac{\sin(t(x-a)) - \sin(t(x-b))}{\pi t}\, dt \right) dP(x) \\ &= \int_{-\infty}^{\infty} \frac{\operatorname{sgn}(x-a)}{\pi} S(T|x-a|) - \frac{\operatorname{sgn}(x-b)}{\pi} S(T|x-b|) dP(x) \end{aligned} \tag{4.2}$$

The integrand in Equation 4.2 is bounded and converges as $T \to \infty$ to the function

$$\psi_{a,b}(x) = \begin{cases} 0 & x < a \\ \frac{1}{2} & x = a \\ 1 & a < x < b \\ \frac{1}{2} & x = b \\ 0 & x > b \end{cases}$$

By DCT we have that $I_T \to \int \psi_{a,b} dP = P((a,b])$ if a and b are not atoms. $\square$

You can use the Fourier inversion formula to extract more information.

Theorem 4.6 Suppose the Fourier transform $\widehat{P}(t)$ is integrable, $\int |\widehat{P}(t)| dt < \infty$ then P has a density $f(x) = \int e^{-itx} \widehat{P}(t) dt$.

Proof. Using that $\left|\frac{e^{-ita} - e^{-itb}}{it}\right| \leq |b-a|$ the fact that $|\widehat{P}(t)|$ is integrable means we can extend the integral in Equation 4.1 to an integral from $-\infty$ to ∞. As a consequence we get $P((a,b)) \leq |b-a| \int_{-\infty}^{\infty} |\widehat{P}(t)| \, dt$ and thus P has no atoms. Furthermore for the CDF $F(t)$ of P we have for h negative or positive

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-it(x+h)}}{ith} \widehat{P}(t) dt$$

The integrand is dominated by $|\widehat{P}(t)|$ and by DCT F is differentiable and P has density $F'(x) = f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \widehat{P}(t) dt$. $\square$

4.6 Gaussian vectors

First we recall some basics about random vectors

Definition 4.2 (mean and covariance of a random vector) For random vector $X = (X_1, X_2, \cdots, X_n)$ where $X_i \in L^2$ we define

- The **mean** $\mu = \mu_X$ **of** X is the vector

$$\mu = E[X] \equiv (E[X_1], \cdots, E[X_n])\,.$$

- The **covariance matrix** $\Sigma = \Sigma_X$ **of** X is the $n \times n$ matrix

$$\Sigma_{ij} = \mathrm{Cov}[\mathrm{X_i X_j}] = E[(X_i - E[X_i])(X_j - E[X_j])]$$

Theorem 4.7 The covariance matrix Σ is symmetric (i.e. $\Sigma^T = \Sigma$) and positive definite, i.e., for any $\alpha \in \mathbb{R}^n$

$$\langle \alpha, \Sigma\alpha \rangle \geq 0$$

Proof.

$$\begin{aligned}\langle \alpha\,,\,\Sigma\alpha\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbb{E}[(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])]\\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])\right]\\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\alpha_i(X_i-\mathbb{E}[X_i])\right)\left(\sum_{j=1}^{n}\alpha_j(X_j-\mathbb{E}[X_j])\right)\right] = \mathbb{E}\left[|\langle\alpha,(X-\mathbb{E}[X]\rangle|^2\right]\geq 0.\end{aligned}$$

□

A positive define symmetric matrix has non-negative eigenvalues and can be diagonalized by an orthonormal matrix. If the matrix Σ is degenerate (i.e. some eigenvalues are 0) or equivalently $\langle\alpha,\Sigma\alpha\rangle = 0$ for some $\alpha\in\mathbb{R}^n$. Then X almost surely lies in some hyperplane of $\mathbb{R}^n$ of dimension strctly less than n. Indeed by the previous calculation $\langle\alpha,\Sigma\alpha\rangle = 0$ implies that

$$\langle\alpha,(X-\mathbb{E}[X]\rangle = 0 \text{ almost surely}$$

In that case X certainly cannot have a density.

Definition 4.3 (Gaussian vectors)

- A (extended) Gaussian random variable X is a real random variable with the charatceristic function, if for some $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$

$$\phi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$$

  Note $\sigma = 0$ is allowed which correspond to $X = \mu$ a.s.

- A n-dimensional random vector X is called a Gaussian random vector if for any $\alpha \in \mathbb{R}^n$ the random variable $\langle \alpha, X \rangle$ is a (extended) Gaussian random variable

Another, common way to define Gaussian vectors in terms of their characteristic functions.

Theorem 4.8 (Characterization of Gaussian vectors) A random vector is Gaussian if and only its characteric function is of the form

$$\phi_X(t) = e^{i\langle \mu_X, t\rangle - \frac{1}{2}\langle t, \Sigma_X t\rangle} \tag{4.3}$$

and $\mu_X = E[X]$ is the mean vector for X and Σ_X is the covariance matrix of X.

Proof. If X is gaussian vector then $Z = \langle t, X \rangle$ is Gaussian random variable and then we have

$$\phi_Z(1) = E[e^{iZ}] = E[e^{i\langle t, X\rangle}] = e^{i\mu_Z - \frac{1}{2}\sigma_Z^2}$$

where

$$\mu_Z = E[\langle t, X\rangle] = \langle t, E[X]\rangle = \langle t, \mu_X\rangle$$

and

$$\sigma_Z^2 = E[(\langle t, X\rangle - E[\langle t, X\rangle])^2] = E[\langle t, X - E[X]\rangle^2] =< t, \Sigma_X t >$$

where, in the last equality, we have repeated to computation done in the proof of Theorem 4.7. This proved the desired formula for the characteristic function of X.

Conversely suppose X is a random vector with characteristic function given in Equation 4.3 then for $Z = \langle \alpha, X\rangle$ we have

$$\phi_Z(u) = E[e^{iuZ}] = E[e^{i\langle u\alpha, X\rangle}] = e^{i\langle u\alpha, \mu_X\rangle - \frac{1}{2}\langle u\alpha, \Sigma_X u\alpha\rangle} = e^{iu\langle \alpha, \mu_X\rangle - \frac{1}{2}u^2\langle \alpha, \Sigma_X \alpha\rangle}$$

which is the characteristic function of Gaussian random variable. □

Constructing and simulating a Gaussian random vector: We can construct a Gaussian random vector with prescribed mean μ_X and covaraince Σ_X as followed. Since Σ is symmetric and non-negative definite we can write it as

$$\Sigma = Q \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} Q^T \equiv AA^T \quad \text{with} \quad A = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} Q$$

If $Z = (Z_1, \cdots, Z_n)$ is a vector of n independent standard normal then we set $X = \mu_X + AZ$ and since $\phi_Z(t) = e^{-\frac{1}{2}\langle t,t\rangle}$ by Theorem 4.3 we have $\phi_X(t) = e^{i\langle t,\mu_X\rangle}\phi_Z(A^T t) = e^{i\langle t,\mu_X\rangle - \frac{1}{2}\langle t,\Sigma_X t\rangle}$ as desired.

Density of a non-generate gaussian random vector: Using the same notation as above, if $\Sigma_X = AA^T$ is invertible then the matrix A is also invertible and consider the vector $Z = A^{-1}(X - \mu_X)$. It is also a Gaussian vector with $\mu_Z = 0$ and $\Sigma_Z = E[ZZ^T] = E[A^{-1}(X-\mu_x)(A^{-1}(X-\mu_x))^T] = A^{-1}\Sigma_X(A^T)(-1) = I$. Therefore the characteristic function of Z is $\phi_Z(t) = e^{-\frac{1}{2}\sum_{i=1}^n t_i^2}$ which is the characteristic function for a vector or n independent standard normal whose density is

$$f_Z(z) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\langle z,z\rangle}$$

Using the expectation rule and the change of variable $x = Az + \mu_X$ we find

$$f_X(x) = \frac{1}{|\det(\mathrm{A})|} f_Z(A^{-1}(x-\mu_X)) = \frac{1}{\sqrt{\det(\Sigma_X)}} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\langle x,\Sigma_X^{-1}x\rangle}$$

2d gaussian vector For $n = 2$ we can write the positive definite covariance matrix as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

for some $0 \leq \rho \leq 1$. The case $\rho = 0$ corresponds to two independent and the case $\rho = 1$ corresponds to the degenerate case $\det(\Sigma_X) = 0$ (e.g. take $X_2 = \frac{\sigma_2}{\sigma_1} X_1$. For $\rho < 1$, the density is then given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2} \right) \right\}.$$

The marginal X_1 and X_2 are normal with mean 0 and variance σ_1 and σ_2 and the density of X_1 conditioned on $X_2 = x_2$ (i.e. the kernel $k(x_2, x_1)$) is given by

$$k(x_2, x_1) = f_{X_1|X_2=x_2}(x_1) = \frac{1}{\sqrt{2\pi\sigma_1\sqrt{1-\rho^2}}} \exp\left\{ -\frac{1}{2\sigma_1^2(1-\rho^2)} \left(x_1 - \rho\frac{\sigma_1}{\sigma_2} x_2 \right)^2 \right\}$$

which is a Gaussian vector with mean $\rho\frac{\sigma_1}{\sigma_2} x_2$ and variance $\sigma_1^2(1-\rho^2)$.

In the degenerate case, using the notation δ_x for the probability measure concentrated at x we have the kernel $K(x_2, \cdot) = \delta_{\frac{\sigma_1}{\sigma_2} x_2}(\cdot)$.

4.7 Examples

We can use the inversion theorem in creative ways.

Examples:

1. The Laplace RV Y is a two-sided version of the exponential RV. Its density is $f(x) = \frac{\beta}{2}e^{-\beta|x|}$. You can think of the Laplace distribution as the mixture (with mixture parameters $\frac{1}{2}, \frac{1}{2}$) of an exponential RV X and $-X$ where X is exponential. Its characteristic function is then

$$\phi_Y(t) = E[e^{itY}] = \frac{1}{2}E[e^{itX}] + \frac{1}{2}E[e^{-itX}] = \frac{1}{2}\frac{\beta}{\beta - it} + \frac{\beta}{\beta + it} = \frac{\beta^2}{\beta^2 + t^2}$$

2. The Cauchy RV Z had density $f(x) = \frac{\beta}{\pi(x^2+\beta^2)}$ and its characteristic function is given by

$$\phi_Y(t) = E[e^{itY}] = \int_\infty^\infty e^{itx}\frac{\beta}{\pi(x^2 + \beta^2)}$$

a priori not an easy integral. However notice that the Fourier transform of the Laplace looks (up to constants) exactly like the density of a Cauchy! So we using Theorem 4.6 for the Laplace distribution shows that

$$\frac{\beta}{2}e^{-\beta|x|} = \frac{1}{2\pi}\int e^{-itx}\phi_Y(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\beta^2}{\beta^2 + t^2}dt = \frac{\beta}{2}\int e^{itx}\frac{\beta}{\pi(\beta^2 + t^2)}dt = \frac{\beta}{2}\phi_Z(x)$$

from which conclude that $\phi_Z(t) = e^{-\beta|t|}$.

4.8 Sum of independent random variables

Suppose X and Y are independent random variables. We wish to understand what is the distribution of $X+Y$. The first tool is to use the characteristic function and the fact that if X and Y are independent

$$E[e^{it(X+Y)}] = E[e^{itX}]E[e^{itY}]$$

together with the uniquness theorem.

Examples

1. Suppose X is normal with paramter μ and σ^2 and Y normal with paramter ν and η^2. Then if X and Y are independent then $X+Y$ is normal with paramter $\mu+\nu$ and $\sigma^2+\eta^2$. This follows form the uniqueness theorem and

$$E[e^{it(X+Y)}] = E[e^{itX}]E[e^{itY}] = e^{i\mu t-\sigma^2t^2/2}e^{i\nu t-\eta^2t^2/2} = e^{i(\mu+\nu)t-(\sigma^2+\eta^2)t^2/2}$$

2. Suppose $X_1,\cdots,X_n$ are independent Bernoulli RV with paramters p then $X_1+\cdots+X_n$ is a binomial RV. Indeed we have

$$E[e^{it(X_1+\cdots+X_n)}] = E[e^{itX_1}]\cdots E[e^{itX_n}] = (e^{it}p+(1-p))^n$$

Another tool is the following convolution theorem

Theorem 4.9 (Convolution of probability measures) Assume X and Y are independent random variables.

- If X and Y have distribution P^X and P^Y then $X+Y$ has the distribution

$$P^X \star P^Y(A) = \iint 1_A(x+y) dP^X(x) dP^Y(y) \quad \text{convolution product}$$

- If X and Y have densities $f_X(x)$ and $f_Y(y)$ then $X+Y$ has the density

$$f_{X+Y}(z) = \int f_X(z-y) f_Y(y) dy = \int f_X(x) f_Y(z-x) dx$$

Proof. For the first part let us take a non-negtive function h and set $Z = X+Y$. We have then

$$E[h(Z)] = E[h(X+Y)] = \int h(x+y) P^X(dx) P^Y(dy)$$

Taking $h = 1_A$ give the result.

For the second part if X and Y have a density we have

$$\begin{aligned} P(Z \in A) = E[1_A(Z)] &= \iint 1_A(x+y) f_X(x) f_Y(y) dx dy \\ &= \iint 1_A(z) f_X(z-y) f_Y(y) dz dy \quad \text{change of variables } z = x+y, dz = dx \\ &= \int \left(\int f_X(z-y) f_Y(y) dy \right) 1_A(z) dz \quad \text{Fubini} \end{aligned}$$

and since this holds for all A, Z has the claimed density. The second formula is proved in the same way. $\square$.

Example: triangular distribution

Suppose X and Y are independent and uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]$. Then $Z = X + Y$ is between -1 and $+1$ for $z \in [-1, 1]$ we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \begin{cases} \int_{-\frac{1}{2}}^{z+\frac{1}{2}} dy & -1 \le z \le 0 \\ \int_{z-\frac{1}{2}}^{\frac{1}{2}} dy & 0 \le z \le 1 \end{cases} = 1 - |z|$$

4.9 Moment and moment generating functions

The moment problem is the question whether a probability distribution P is uniquely determined by all its moments $E[X^n]$. It is in general not true as the following examples shows.

- Recall the log-normal distribution is the distribution of e^X is X is a normal distribution. Its pdf is given, for $\mu = 0$ and $\sigma^2 = 1$

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

  and all moments exists $E[X^r] = \int_0^\infty x^k f(x) dx = e^{r^2/2}$.

- Now consider

$$g(x) = f(x)(1 + \sin(2\pi \ln(x)))$$

  Then for $k = 0, 1, 2, \cdots$ we have with the change of variable $\ln(x) = s + k$

$$\int_0^\infty x^k f(x) \sin(2\pi \ln(x)) dx = \frac{1}{\sqrt{2\pi}} e^{k^2/2} \int_{-\infty}^\infty e^{-s^2/2} \sin(2\pi s) ds = 0\,.$$

  This shows that g is the density of a RV Y and that all moments of Y coincide with the moments of the log-normal!

A stronger condition on the moments do imply uniqueness: if all moments exists and $E[X^n]$ do not grow to fast with n then the moments do determine the distribution. This use a analytic continuation argument and relies on the uniqueness theorem for the Fourier transform.

Theorem 4.10 Suppose X and Y are RV such that the moment generating functions $M_X(t) = M_Y(t)$ in $[-t_0, t_0]$ and are finite in that interval. Then X and Y have the same distribution.

Proof.

- Since $e^{t|x|} \leq e^{tx} + e^{-tx}$ and the right hand-side is integrable, the function $e^{t|X|} = \sum_{k=0}^{\infty} \frac{s|X|^k}{k!}$ is integrable. By the DCT (for sums of RVs, in the form of **?@exr-62**) we have $E[e^{tX}] = \sum_{k=0}^{\infty} \frac{t^k E[X^k]}{k!}$.
- This implies that $\frac{t^k E[X^k]^k}{k!} \to 0$ as $k \to \infty$ (for $|t| \leq t_0$). We claim that this implies that $\frac{s^k E[|X|^k]^k}{k!} \to 0$ as $k \to \infty$ as long as $s < t_0$. If k is even $E[|X|k] = E[X^k]$ and there is nothing to do. For k odd we use on one hand that $|X|^{2k-1} \leq 1 + |X|^2 k$ as well that $s < t$ we have (for k sufficiently large) $2ks^{2k-1} < t^2 k$. Together this shows

$$\frac{s^{2k-1} E[|X|^k]^k}{k!} \leq \frac{t^{2k} E[X^k]^k}{k!} \qquad \text{for } k \text{ large enough.}$$

- The next piece is the Taylor expansion theorem with reminder for function which are n-times continuously differentiable

$$f(x) = \sum_{k=0}^{n} f^{(k)}(x_0)\frac{(x-x_0)^k}{k!} + \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-s)^n dt$$

  from which we obtain

$$\left| e^{itx}\left(e^{ihx} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right) \right| \leq \frac{|hx|^{n+1}}{(n+1)!}$$

- Integrating with respect to P and taking $n \to \infty$ together with Theorem 4.2 gives

$$\phi_X(t+h) = \sum_{k=0}^{\infty} \phi_X^{(k)}(t) \tag{4.4}$$

  for $|h| < t_0$.

- Now suppose X and Y have the same moment generating function in a neighborhood of 0. Then all their moments coincide and thus by Equation 4.4 (with $t = 0$), $\phi_X(t) = \phi_Y(t)$ on the interval $(-r_0, r_0)$. By Theorem 4.2 their derivatives must also be equal on $(-t_0, t_0)$. Using now Equation 4.4 (with $t = -t_0 + \epsilon$ and $t = t_0 - \epsilon$ shows that $\phi_X(t) = \phi_Y(t)$ on the interval $(-2t_0, 2t_0)$. Repeating this argument $\phi_X(t) = \phi_Y(t)$ for all t and thus by Theorem 4.5 X and Y must have the same distribution. $\square$

4.10 Exercises

Exercise 4.1

- Show that a characteristic function $\phi_X(t)$ satifies $\overline{\phi_X(t)} = \phi_X(-t)$ (complex conjugate).
- Show that a characteristic function $\phi_X(t)$ is real if and only if the random variable X is symmetric (i.e X and $-X$ have the same distribution)
- Show that if ϕ_X is the characteristic function for some RV X then $\phi_X^2(t)$ and $|\phi_X(t)|^2$ are characteristic function as well. What are the corresponding RVs?

Exercise 4.2 (Independence and correlation for Gaussian random vectors)

- Consider two Gaussian random vectors X and Y (of dimension n and m) which are jointly Gaussian, that is $Z = (X, Y)$ is also a Gaussian random vector. Show that X and Y are independent if and only they are uncorelated that is

$$E[(X - \mu_X)(Y - \mu_Y)^T] = 0$$

  Hint: Use the characteristic function.

- Suppose $X_1, \cdots, X_n$ are IID normal RV with mean μ and variance σ^2 and let $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$ be the emprical mean. Show that

$$\overline{X}_n \quad \text{and } (X_1 - \overline{X}_n, X_2 - \overline{X}_n, \cdots, X_n - \overline{X}_n)$$

  are independent. *Hint*: Use part 1.

- The empirical variance of IID RV $X_1, \cdots, X_n$ is given by $V_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$. Show that if the X_i;s are IID normal RV with mean μ and variance σ^2 then $\overline{X_n}$ and V_n are independent random variable

Exercise 4.3 In this problem we study the characteristic function for a Gamma random variable with parameter α and β and density $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. In particular you will prove that $\phi_X(t) = \left(\frac{\beta}{\beta - it}\right)^\alpha$.

- First show that it is enough to consider the case $\beta = 1$ (change scale.)
- Use the moments $E[X^n]$ to show that $\phi_X(t) = \sum_{n=0}^{\infty} (it)^n \frac{\Gamma(\alpha+n)}{\Gamma(\alpha) n!}$ and use then the binomial series for $(1 - it)^{-\alpha}$.
- Use your result to show that
  - The sum of two independent gamma random variables with parameters (α_1, β) and (α_2, β) is a gamma random variable.
  - If $X_1, X_2, \cdots, X_n$ are independent normal random variable with mean 0 and variance σ^2. show that $X_1^2 + \cdots + + X_n^2$ is a gamma random variable and find the parameters.

Exercise 4.4 Show that if X and Y are RV values taking in the positive integers with distributions $P^X(n)$ and $P^Y(n)$ and are independent then $X+Y$ has distribution $P^{X+Y}(n) = \sum_{k=0}^{k} P^X(k)P^Y(n-k)$ (this is called the convolution product of the two sequences P^X and P^Y).

Exercise 4.5

- In Theorem 4.5, modify the statement of the theorem if a or b are atoms.
- Show that

$$P(\{a\}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita} \widehat{P}(t) dt$$

  Hint: Imitate the proof of the inversion formula in Theorem 4.5

- Suppose a RV X takes integer values in $\mathbb{Z}$, show that

$$P(X=n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itn} \phi_X(t) dt$$

  Hint: Show that $\phi_X(t)$ is periodic