EPFL, Spring 2023
Guillaume Obozinski

# EE-209 Eléments de Statistiques pour les Data Sciences

## Feuille d'exercices 3

**Exercise 3.1** A beta random variable is a continuous random variable taking values in the interval $[0, 1]$. It will become very useful when we will do Bayesian statistics. A beta random variable with parameters $\alpha, \beta > 0$ has the following probability density function

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} 1_{\{0\le x\le 1\}},$$

with $\Gamma$ defined by $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\, dx$. The only property of $\Gamma$ that you will need to use in this exercise is that $\forall a > 0$, $\Gamma(a + 1) = a\,\Gamma(a)$, and that $\Gamma(1) = 1$.

(a) Prove that for any integer $n \ge 0$, $\Gamma(n + 1) = n!$

(b) Prove that the uniform distribution is a particular case of a Beta distribution.

(c) Explain why the fact that $p(x; \alpha, \beta)$ is a probability distribution implies that $\forall \alpha, \beta > 0$, we have

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

(d) Compute $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ as a function of $(\alpha, \beta)$.

(e) Compute $\mathrm{Var}(X)$ as a function of $(\alpha, \beta)$.

(f) Sometimes the Beta distribution is parameterized as a function of the expectation $\mu = \mathbb{E}[X]$ and of $\varphi = \alpha + \beta$. Express $\alpha, \beta$ and $\mathrm{Var}(X)$ as a function of $\mu$ and $\varphi$. How does $\varphi$ modify the variance?

**Exercise 3.2** In this problem, we consider a pair of random variables $(X, Y)$ taking values on $[0, 1]^2$, with joint probability distribution

$$p(x, y) = c - c(y - x)^2,$$

on $[0, 1]^2$ and where $c > 0$ is a constant to be determined.

(a) Determine the form of $p_X(x)$ the marginal pdf of $X$ and use results from Exercise 3.1 to determine the value of the constant $c$

(b) Using the form of beta pdfs provided at the beginning of Exercise 3.1, show that $p_X(x)$ is actually a mixture of beta distributions of the form

$$p(x) = \pi\, p(x; \alpha_1, \beta_1) + (1-\pi)\, p(x; \alpha_0, \beta_0).$$

In particular, specify $\pi, \alpha_1, \beta_1, \alpha_0$, and $\beta_0$.

(c) Compute $\mathbb{E}[X]$.

(d) Compute $\text{cov}(X, Y)$. Comment on its sign. Is it surprizing given the form of the joint pdf?

(e) Compute $\text{corr}(X, Y)$.

**Exercise 3.3 The disoriented explorer.**
A disoriented explorer reaches a clearing in the jungle. There are three paths leading from the clearing.

- Path "f" will takes him into a deep forest that forces him to return to the clearing after a time $X_\text{f}$ that is distributed as a normal random variable with mean 2 hours and variance of 0.1 hour.
- Path "c" leads to a canyon (and is thus also a dead-end) that forces him to return after a time $X_\text{c}$ that is distributed as a normal random variable with mean 3 hours and variance of 0.2 hour.
- Path "v" takes him to the village exactly after $X_\text{v}$ equal to 1 hour.

In this exercise, we aim at understanding some properties of the random variable $X$ which is defined as the total number of hours to reach the village. We work under the following two essential assumptions:

$\mathcal{H}_1$: Although the explorer chooses the path at random, with equal probabilities, he will not choose a path that he has already established is a dead-end.

$\mathcal{H}_2$: The times spent on each path by the explorer are assumed to be independent.

(a) **The modeling step**. We will first formalize the problem mathematically. Let's denote by $\mathcal{Y}$ the set of all possible sequence of paths taken by the explorer to reach the village. For example, the sequence "fv" means that the explorer took first the path "f" before taking the correct path "v" leading to the village. In our setting, given that the choices that the explorer makes are random, the sequence of paths he will take is a random variable that we will call $Y$ and that takes the values in $\mathcal{Y}$ with some probability.

  ($i$) Using assumption $\mathcal{H}_1$, list all the possible sequences of paths that the explorer can take. This will specify all the elements in the set $\mathcal{Y}$.

  ($ii$) Using assumption $\mathcal{H}_2$, determine the probability that the explorer takes each of these sequences of paths.

  ($iii$) Determine the distribution of $X$ given that $Y =$ "fv".
  Hint: As a first step, you can try to express the distribution of $X$ given that $Y =$ "fv" as a function of the distributions of the random variables $X_\text{f}$ and $X_\text{v}$.

  ($iv$) Similarly, determine the distribution of $X$ given that $Y = y$ for any $y \in \mathcal{Y}$.

(b) **Expected number of hours to reach the village.**

$(i)$ Compute the conditional expectation $\mathbb{E}[X \mid Y = \text{“cv”}]$ which is the expected number of hours before reaching the village given that the explorer took first the path “c” before taking the correct path “v”.

$(ii)$ Using an analogous approach, compute the conditional expectations $\mathbb{E}[X \mid Y = y]$ for any $y \in \mathcal{Y}$.

$(iii)$ Using the law of total expectation proved in the exercise sheet of week 2, compute the expectation $\mathbb{E}[X]$ from all the conditional expectations $\mathbb{E}[X \mid Y = y]$ with $y \in \mathcal{Y}$.

(c) **Variance of the number of hours to reach the village.**

$(i)$ Recall the expression of the variance of $X$ as a function of $\mathbb{E}[X^2]$ and $\mathbb{E}[X]$.

In order to compute $\mathrm{Var}(X)$, we are only missing $\mathbb{E}[X^2]$.

$(ii)$ Using an approach analogous to question $(b)$, compute the conditional expectations $\mathbb{E}[X^2 \mid Y = y]$ for any $y \in \mathcal{Y}$.

$(iii)$ Using the law of total expectation proved in the exercise sheet of week 2, compute the expectation $\mathbb{E}[X^2]$ from all the conditional expectations $\mathbb{E}[X^2 \mid Y = y]$ with $y \in \mathcal{Y}$.

$(iv)$ Deduce the variance of $X$.