Applied Statistics (ECS764P) - Lab 1: Probability theory

Fredrik Dahlqvist

11 October 2023

1 Theory

The following questions are meant to test your understanding of lectures 1 and 2. Answers to these questions will not be marked, but if you can solve these questions, you will be fine at the exam...

1. The *triangular distribution* is the distribution you get by summing two uniform distributions on $[0, 2]$. Its pdf is given by:

$$f(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1 \\ 2 - x & \text{if } 1 \leq x \leq 2 \\ 0 & \text{else} \end{cases} .$$

Plot this distribution (in Python or on a piece of paper). Compute its CDF. Use your plot to check that your answer makes sense.

2. Compute the following:
   (a) Consider the slightly modified Bernoulli distribution which is supported by $\{1, 2\}$ (instead of $\{0, 1\}$) and where the probability mass of $\{1\}$ is $(1 - p)$ and the probability mass of $\{2\}$ is p. Compute the variance of this distribution.
   (b) The mean of the triangular distribution defined above.
   (c) The standard deviation of the uniform distribution on an interval $[a, b]$,

3. Compute the pushforward of the uniform distribution on $[0, 1]$ through the map $f : [0, 1] \to [0, 1], x \mapsto x^2$.
Hint: compute the CDF from the definition of the pushforward, then compute the PDF by differentiating. (Think of the distributions support. What are the possible values?)

4. Recall that measures (and therefore probability measures) are σ-additive. This means that if μ is a measure, X is a set, and $(A_i)_{i\in\mathbb{N}}$ is a collection of disjoint subsets which partition X – that is to say

$$X = \bigcup_{i=0}^{\infty} A_i,$$

then it must be the case that

$$\mu(X) = \sum_{i=0}^{\infty} \mu(A_i). \tag{1}$$

In other words, the masses of the set A_i add up to the mass of X.
Consider the uniform distribution on $(0, 1]$. What is its pdf? Consider the collection of sets defined by

$$A_i = \left(\frac{1}{2^{i+1}}, \frac{1}{2^i} \right], \quad 0 \leq i$$

Show that it forms a partition of $(0, 1]$ (i.e. the A_is are pairwise disjoint and their union is the whole of $(0, 1]$. Show that the σ-additivity equation (1) holds for this partition. (*Hint: You might want to check out this page:* `https://en.wikipedia.org/wiki/Geometric_series`.)

2 Practice

General instructions Complete the following tasks in a Jupyter Notebook. This Jupyter Notebook will need to be submitted on QMPlus (follow Labs and Coursework→ Coursework 1 - submission) by 18 October 2023 at 18:00. This coursework will count for 10% of your final mark for the module.

The marks awarded for each sub-question are detailed below. However, note that your code must run without any bugs to get full marks. The person marking your worksheet will start by *running all cells*. If any error is thrown, your final grade will be halved (i.e. the maximum possible grade for a buggy notebook will be 5/10). There is not 'a correct way' to answer these questions!

1. **(2 mark)** Implement the counting measure in Python. Test that it satisfies additivity on the disjoint sets `{"a", "b", "c"}`, `{"d", "e", "f"}`.

   Hint: If you have never written a Python function, read `https://www.w3schools.com/python/python_functions.asp`, if you have never used Python sets, read `https://www.w3schools.com/python/python_sets.asp`.

   Bonus mark if your implementation of the counting measure checks that the input type is correct and raises an error otherwise.

2. **(2 marks)** Create a Python class which implements intervals. Use this new data type to write a function which implements the length measure on intervals. Test it on the interval $[1, 3.5]$.

   Hint: If you have never written a Python class, read `https://www.w3schools.com/python/python_classes.asp`.

   Bonus mark if your implementation of the length measure checks that the input type is correct and raises an error otherwise.

3. (**3 marks**) Import `scipy.stats` in order to access the `scipy.stats.expon` distribution. This implements the exponential distribution $\mathrm{Exp}(\lambda)$. Make sure you read the documentation `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.expon.html` to understand how it works and how the parameter λ is encoded. Using the cdf method of `scipy.stats.expon` define a function called `expon_measure` which will take as input an interval (defined in the previous question) and will return its probability mass under the probability measure $\mathrm{Exp}(2)$ (i.e. $\lambda = 2$). Test your function by computing the probability measure of the following intervals:

   (a) $[0, 1]$
   (b) $[1, 1]$
   (c) $[1, 10]$
   (d) $[0, \infty)$

   Plot the pdf of $\mathrm{Exp}(2)$ on comment on whether your answers seem to make sense visually.

4. (**3 marks**) Using the pdf method of `scipy.stats.expon`, define a function called `expon_pdf` which will take one argument `x` and return the pdf of the probability measure $\mathrm{Exp}(2)$ evaluated at `x`. Import the integration routine `quad` from `scipy.integrate`, and read the documentation `https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.quad.html` to see how it works. Use `quad` to compute and print the following integrals

   (a) $\int_0^1 \texttt{expon\_pdf}(x)\ dx$
   (b) $\int_1^1 \texttt{expon\_pdf}(x)\ dx$
   (c) $\int_1^{10} \texttt{expon\_pdf}(x)\ dx$
   (d) $\int_0^{\infty} \texttt{expon\_pdf}(x)\ dx$

   Compare your answers with those of the previous question. What do you see? Why is this the case?