Applied Statistics (ECS764P) - Lab 1: Probability theory

Fredrik Dahlqvist

11 October 2023

1 Theory

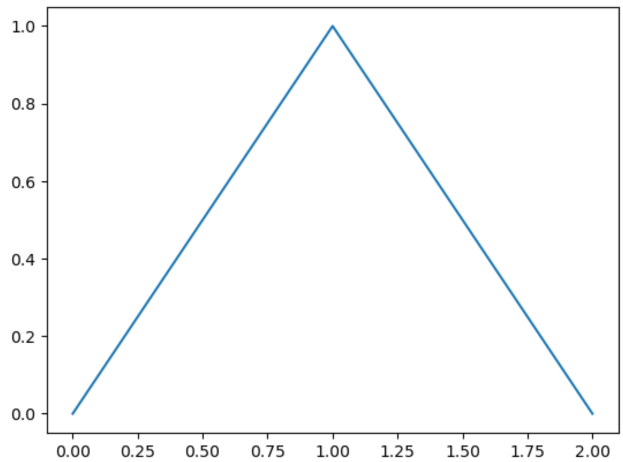
The following questions are meant to test your understanding of lectures 1 and 2. Answers to these questions will not be marked, but if you can solve these questions, you will be fine at the exam...

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.

Answer: The distribution looks like this (hence the name):



The CDF $F(t)$ is computed by integrating the pdf from $-\infty$ to t. If $0 \leq t \leq 1$ we get

$$\begin{aligned} F(t) &= \int_{-\infty}^{t} f(x)\ \mathrm{d}x && \text{Definition of the CDF} \\ &= \int_{0}^{t} x\ \mathrm{d}x && \text{definition of } f \text{ on } 0 \leq t \leq 1 \\ &= \frac{x^2}{2}\Big|_0^t && \text{Standard calculus} \\ &= \frac{t^2}{2} \end{aligned}$$

Looking at the graph above, this makes perfect sense: we're computing the area of a right-angle triangle which is precisely half of the area of a t-by-t square.

If $1 \le t \le 2$ we get:

$$\begin{aligned}
F(t) &= \int_{-\infty}^{t} f(x) \, \mathrm{d}x && \text{Definition of the CDF} \\
&= \int_0^1 x \, \mathrm{d}x + \int_1^t (2-x) \, \mathrm{d}x && \text{definition of } f \text{ on } 1 < t \le 2 \\
&= \frac{1}{2} + 2x - \frac{x^2}{2}\Big|_1^t && \text{Standard calculus} \\
&= \frac{1}{2} - \frac{t^2}{2} + 2t - 2 + \frac{1}{2} \\
&= -\frac{t^2}{2} + 2t - 1
\end{aligned}$$

Again, this makes sense when looking at the graph: we want the area of the triangle below the ascending part of the function ($\frac{1}{2}$) plus the difference between the triangle below the descending part of the function ($\frac{1}{2}$) and the triangle starting at $2 - t$. If you work this out you get the answer above.

Putting everything together we get:

$$F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \frac{t^2}{2} & \text{if } 0 < t \le 1 \\ -\frac{t^2}{2} + 2t - 1 & \text{if } 1 < t \le 2 \\ 1 & \text{else} \end{cases}.$$

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 $[a, b]$,

   Answer:

   (a) First we need to compute the mean of this distribution. It is given by

$$\mu\left(\text{Bern}(p)\right) = \sum_{x \in \{1,2\}} x \cdot \text{Bern}(p)\left(\{x\}\right) = 1.(1-p) + 2.p = 1 + p.$$

   Now we can compute the variance

$$\begin{aligned}
Var\left(\text{Bern}(p)\right) &= (1-p)(1 - \mu\left(\text{Bern}(p)\right))^2 + p(2 - \mu\left(\text{Bern}(p)\right))^2 \\
&= (1-p)p^2 + p(1-p)^2 \\
&= (1-p)(p^2 + p(1-p)) \\
&= (1-p)p
\end{aligned}$$

   (b) We simply compute the integral

$$\begin{aligned}
\int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x &= \int_0^1 x.x \, \mathrm{d}x + \int_1^2 (2-x)x \, \mathrm{d}x && \text{By definition of } f \\
&= \frac{x^3}{3}\Big|_0^1 + x^2 - \frac{x^3}{3}\Big|_1^2 && \text{Simple calculus} \\
&= \frac{1}{3} + \left(4 - \frac{8}{3} - \left(1 - \frac{1}{3}\right)\right) \\
&= 1
\end{aligned}$$

(c) Recall that the pdf is the constant function $\frac{1}{b-a}$. We start by computing the mean of the distribution:

$$\begin{aligned}\mu(\text{Uniform}(a,b)) &= \int_a^b \frac{x}{b-a}\ \mathrm{d}x \\ &= \left.\frac{x^2}{2(b-a)}\right|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} \\ &= \frac{b+a}{2}.\end{aligned}$$

To compute the definition we can simply expand $(x-\mu(\text{Uniform}(a,b)))^2$ and integrate the corresponding polynomial. Alternatively, we can integrate by substitution: we put $u = x - \mu(\text{Uniform}(a,b))$ and simply get $\mathrm{d}u = \mathrm{d}x$.

$$\begin{aligned}Var(\text{Uniform}(a,b)) &= \int_a^b \frac{(x-\mu(\text{Uniform}(a,b))])^2}{b-a}\ \mathrm{d}x && \text{Definition of the variance}\\ &= \frac{1}{b-a}\int_{a-\mu(\text{Uniform}(a,b))}^{b-\mu(\text{Uniform}(a,b))} u^2\ \mathrm{d}u && \text{Using the substitution above}\\ &= \frac{1}{b-a}\left.\frac{u^3}{3}\right|_{a-\mu(\text{Uniform}(a,b))}^{b-\mu(\text{Uniform}(a,b))} && \text{Simple calculus}\\ &= \frac{1}{b-a}\left.\frac{u^3}{3}\right|_{\frac{a-b}{2}}^{\frac{b-a}{2}} && \text{Plug in value of } \mu(\text{Uniform}(a,b))\\ &= \frac{1}{b-a}\left(\frac{(b-a)^3}{24} - \frac{(a-b)^3}{24}\right) \\ &= \frac{(b-a)^2}{12} && \text{Using } (a-b) = (-1)(b-a)\end{aligned}$$

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?)

Answer: The support is $[0,1]$ as well since the square of any number in $[0,1]$ is also in $[0,1]$. From the definitions, the CDF is given by

$$\begin{aligned}F(t) &= f_*\mathbb{P}((-\infty,t]) \\ &= \mathbb{P}(\{x\in[0,1] \mid 0\le f(x) < t\}) \\ &= \mathbb{P}\left(\{x\in[0,1] \mid 0\le x^2 < t\}\right) \\ &= \mathbb{P}\left(\{x\in[0,1] \mid 0\le x \le \sqrt{t}\}\right) \\ &= \sqrt{t}\end{aligned}$$

If we now differentiate we get

$$\begin{aligned}PDF(t) &= \frac{d}{dt}\sqrt{t} \\ &= \frac{1}{2\sqrt{t}}\end{aligned}$$

It is not hard to check that this is indeed a PDF since

$$\int_0^1 \frac{1}{2\sqrt{t}} = \sqrt{t}\Big|_0^1 = 1$$

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

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

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

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

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`.)

Answer: The pdf of the uniform distribution is the function

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1] \\ 0 & \text{else} \end{cases}.$$

This means that the probability mass of an interval $(a,b]$ (or $[a,b)$ or $[a,b]$ or (a,b), the end points make no difference) contained in $(0,1]$ is given by

$$\int_a^b 1 \,\mathrm{d}x = b - a.$$

It's simply the usual length.

Suppose $i \neq j$ then either $i < j$ or $j < i$. Suppose $i < j$ (the argument is exactly the same of $j < i$), if $x \in A_i$ then by definition

$$\frac{1}{2^{i+1}} < x < \frac{1}{2^i}$$

but then we cannot have $x \in A_i$ since

$$\frac{1}{2^{j+1}} < \frac{1}{2^j} \le \frac{1}{2^{i+1}} < x \le \frac{1}{2^i}.$$

So $A_i \cap A_j = \emptyset$. To see that $(0,1] = \bigcup_{i=0}^{\infty} A_i$, pick any $x \in (0,1]$, then we can always find an i such that

$$\frac{1}{2^{i+1}} < x \le \frac{1}{2^i}$$

and so $x \in A_i$ for some i.

Finally, let us check that (1) holds in this case. Using the fact that we're just measuring lengths, we get

$$\sum_{i=0}^{\infty} \mathbb{P}(A_i) = \sum_{i=0}^{\infty} \left(\frac{1}{2^i} - \frac{1}{2^{i+1}}\right) = \sum_{i=0}^{\infty} \frac{1}{2^{i+1}} = 1 = \mathbb{P}((0,1])$$

as desired.

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!

Marking Scheme: Cap maximum total at 10 marks!

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.

   Marking Scheme: 1 mark for implementing the function using `len()`. The function must have the correct type: i.e. take a set as an input and return a number as an output. 1 mark for testing (computing the measure of the union and the sum of the measures which should be equal).

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.

   Marking Scheme: 1 mark for writing a sensible class, 0.5 marks for writing a sensible length function, 0.5 mark for testing it and getting the correct answer (2.5).

3. (**3 marks**) Import `scipy.stats` in order to access the `scipy.stats.expon` distribution. This implements the exponential distribution $\text{Exp}(\lambda)$. 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 $\text{Exp}(2)$ (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 $\text{Exp}(2)$ on comment on whether your answers seem to make sense visually.

   Marking Scheme:

   - 1 mark for implementing the function `expon_measure` as the difference of the cdf evaluated at the upper bound of the interval and the cdf evaluated at the lower bound of the interval.
   - 1.5 mark for (a)-(d) being correct (subtract 0.5 mark per mistake until you reach 0).
   - 0.5 mark for plotting the pdf correctly and making a sensible comment.

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 $\text{Exp}(2)$ 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?

   Marking Scheme:

- 2 marks for computing the integrals (a)-(d) correctly (0.5 mark per answer).
- 0.5 mark for identifying that the answers are the same as in (1), 0.5 mark for giving a sensible explanation of why.