Applied Statistics (ECS764P) - Lab 2

Fredrik Dahlqvist

1 Nov 2023

1 Theory

1. Normal distributions have the following two properties:

   - the sum of two normals is normal: $\text{Normal}(\mu_1, \sigma_1) + \text{Normal}(\mu_2, \sigma_2) = \text{Normal}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$
   - re-scaling a normal gives a normal: for any $\alpha > 0$, $\alpha \cdot \text{Normal}(\mu, \sigma) = \text{Normal}(\alpha\mu, \alpha\sigma)$

   Use these two facts to compute the distribution of sample means for identically and normally distributed independent samples of length n. Specifically, compute the distribution of

   $$\frac{1}{n}\sum_{i=1}^{n} \text{Normal}(\mu, \sigma)$$

   Answer: By the first property we get that

   $$\begin{aligned}\sum_{i=1}^{n} \text{Normal}(\mu, \sigma) &= \text{Normal}\left(\sum_{i=1}^{n} \mu, \sqrt{\sum_{i=1}^{n} \sigma^2}\right) \\ &= \text{Normal}\left(n\mu, \sqrt{n}\sigma\right)\end{aligned}$$

   By the second property we now have

   $$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \text{Normal}(\mu, \sigma) &= \frac{1}{n}\text{Normal}\left(n\mu, \sqrt{n}\sigma\right) \\ &= \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

2. Consider the array [3,4,2,5]. Find the sample mean and the sample median. Suppose we add an additional observation $x \geq 5$ to this array. What is the smallest value of x for which the mean will be larger or equal to the median?

   Answer: The mean is given by
   $$\bar{X} = \frac{3+4+2+5}{4} = \frac{7}{2}.$$
   After ordering the array we get that the conventional value of the mean is given by
   $$m = \frac{3+4}{2} = \frac{7}{2}.$$
   Since we're adding a new observation $x \geq 5$, the median will necessarily be the third entry in the ordered list, i.e. 4, not matter what x is. We must therefore find the smallest x such that
   $$\frac{3+4+2+5+x}{5} \geq 4$$
   which means that we want $x \geq 6$, and the smallest such value is clearly $x = 6$.

3. Using the definition of the sum of two probability measures given during the lectures, show that the sum of two identical and independent Bernoulli distributions $\mathrm{Bern}(p)$ is given by a binomial distribution $\mathrm{Binom}(2, p)$. Formally show that

$$\mathrm{Bern}(p) + \mathrm{Bern}(p) = \mathrm{Binom}(2, p)$$

(Hint: What is the support of $\mathrm{Bern}(p) + \mathrm{Bern}(p)$*? What is the support of* $\mathrm{Binom}(2,p)$*? Do the two probability measures agree on every element of their support? If yes, then they are equal.)*

Answer: The support of $\mathrm{Binom}(2,p)$ is, by definition, the set $\{0, 1, 2\}$. To see that this is also the support of $\mathrm{Bern}(p) + \mathrm{Bern}(p)$, imagine taking a sample x_1 from the first copy of $\mathrm{Bern}(p)$ and a sample x_2 from the second copy of $\mathrm{Bern}(p)$. If you add them up, then there are three possible outcomes: either $x_1 = x_2 = 0$ in which case $x_1 + x_2 = 0$, or $x_1 = 0, x_2 = 1$ in which case $x_1 + x_2 = 1$, or $x_1 = 1, x_2 = 0$ in which case $x_1 + x_2 = 1$ also, and finally if $x_1 + x_2 = 1$ then $x_1 + x_2 = 2$. So the support of $\mathrm{Bern}(p) + \mathrm{Bern}(p)$ is also $\{0, 1, 2\}$. Now let's show that the two probability measures agree on every element of their support.

$$\begin{aligned}
& (\mathrm{Bern}(p) + \mathrm{Bern}(p))(0) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(x_1, x_2) \mid x_1 + x_2 = 0\} && \text{Definition of } \mathrm{Bern}(p) + \mathrm{Bern}(p) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(0, 0)\} && \text{Only combination summing to 0} \\
= \ & \mathrm{Bern}(p)(0) \mathrm{Bern}(p)(0) && \text{Definition of the product measure} \\
= \ & (1 - p)^2 && \text{Definition of } \mathrm{Bern}(p) \\
= \ & \mathrm{Binom}(2, p)(0) && \text{Definition of } \mathrm{Binom}(2, p)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& (\mathrm{Bern}(p) + \mathrm{Bern}(p))(1) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(x_1, x_2) \mid x_1 + x_2 = 1\} && \text{Definition of } \mathrm{Bern}(p) + \mathrm{Bern}(p) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(0, 1), (1, 0)\} && \text{Combinations summing to 1} \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(0, 1)\} + \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(1, 0)\} && \text{Additivity of measures} \\
= \ & \mathrm{Bern}(p)(0) \mathrm{Bern}(p)(1) + \mathrm{Bern}(p)(1) \mathrm{Bern}(p)(0) && \text{Definition of the product measure} \\
= \ & (1 - p)p + p(1 - p) && \text{Definition of } \mathrm{Bern}(p) \\
= \ & 2p(1 - p) \\
= \ & \mathrm{Binom}(2, p)(1) && \text{Definition of } \mathrm{Binom}(2, p)
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
& (\mathrm{Bern}(p) + \mathrm{Bern}(p))(2) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(x_1, x_2) \mid x_1 + x_2 = 1\} && \text{Definition of } \mathrm{Bern}(p) + \mathrm{Bern}(p) \\
= \ & \mathrm{Bern}(p) \otimes \mathrm{Bern}(p) \{(1, 1)\} && \text{Only combination summing to 2} \\
= \ & \mathrm{Bern}(p)(1) \mathrm{Bern}(p)(1) && \text{Definition of the product measure} \\
= \ & p^2 && \text{Definition of } \mathrm{Bern}(p) \\
= \ & \mathrm{Binom}(2, p)(2) && \text{Definition of } \mathrm{Binom}(2, p)
\end{aligned}$$

Which concludes the proof that $\mathrm{Bern}(p) + \mathrm{Bern}(p) = \mathrm{Binom}(2, p)$

4. Using the definition of the multiplication of a probability measure by a positive real number, compute the PMF of the probability measure $\frac{1}{2}\mathbb{P}^{dice}$, where $\mathbb{P}^{dice}$ is the uniform distribution on $\{1, 2, 3, 4, 5, 6\}$.

Answer: The support of $\frac{1}{2}\mathbb{P}^{dice}$ is $\{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3\}$. So for $i \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3\}$.

$$\begin{aligned}
\frac{1}{2}\mathbb{P}^{dice}(i) &= \left(m_{\frac{1}{2}}\right)_* \mathbb{P}^{dice}(i) && \text{Definition of multiplication by number} \\
&= \mathbb{P}^{dice}\left(\left\{x \mid m_{\frac{1}{2}}(x) = \frac{x}{2} = i\right\}\right) && \text{Definition of pushforward} \\
&= \mathbb{P}^{dice}(\{2i\}) && \text{Only possibility} \\
&= \frac{1}{6} && \text{Definition of } \mathbb{P}^{dice}
\end{aligned}$$

2 Practice

1. (**Visualisation, 1.5 mark**) Using `scipy.stats`'s `rvs` method, sample 30 tuples $(x_i^1, x_i^2, x_i^3, x_i^4)_{1 \le i \le 30}$ s.th.

$$
\begin{aligned}
x_i^1 &\sim \text{Normal}(0, 1) \\
x_i^2 &\sim \text{Normal}(2, 4) \\
x_i^3 &\sim \text{Uniform}(0, 1) \\
x_i^4 &= x_i^3 \cdot z \text{ where } z \sim \text{Uniform}(0, 1)
\end{aligned}
$$

   Using one of the visualisation techniques discussed in the lectures, plot this 4-D data. (*Hint: you may find that you need to adjust some parameter(s) for your plot to be legible; if so please do it.*). The four dimensions are not all independent of one another. How does this manifest itself on your plot?

2. (**Visualisation, 1.5 mark**) Display a QQ plot for the following probability measures: the standard normal Normal$(0, 1)$ on the x-axis and the standard Cauchy distribution Cauchy$(0, 1)$ on the y-axis. What does the QQ plot tell us about the tails of these distributions?

3. (**Independent sum of two probability measures, 3 marks**) Recall from the lectures that if we have two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ with respective densities f_1 and f_2, then the density of the sum[1] $\mathbb{P}_1 + \mathbb{P}_2$ is given by the convolution of the two densities, viz.

$$f_{1+2}(t) = \int_{-\infty}^{\infty} f_1(x) f_2(t - x)\ dx.$$

   In this question we consider the sum of Beta $(2, 8) +$ Beta $(8, 2)$. What is the support of Beta $(2, 8)$? What is the support of Beta $(8, 2)$? Therefore, what is the support of Beta $(2, 8) +$ Beta $(8, 2)$?

   Write a function which implements the integrand of the integral above, that is to say that implements $f_1(x)f_2(t-x)$, where f_1 is the density of Beta $(2, 8)$ and f_2 is the density of Beta $(8, 2)$. (*Hint: this function will need two arguments.*)

   Next, generate 100 points $(t_1, \ldots, t_{100})$ along the support of Beta $(2, 8) +$ Beta $(8, 2)$ (using `numpy`'s `linspace` function), and using a `for` loop, compute the pdf $f_{1+2}(t_i)$ at these 100 points using `quad`. (*Hint: the documentation of `quad` has an example showing how to integrate a function with two arguments along its first argument.*) Plot your result.

   Finally, generate 10000 samples from Beta $(2, 8)$, 10000 samples from Beta $(8, 2)$, add them, and plot the histogram of these sums along with the pdf computed in the previous step. What do you observe?

4. (**Sample mean process and sample mean distribution, 4 marks**)

   - Write a function called `sample_mean` taking as inputs two integers `m` and `n`. The function should return an array of length `n` containing samples each obtained by taking `m` samples from the standard normal distribution and computing their sample mean. Call `sample_mean(m=10, n=10000)`, `sample_mean(m=100, n=10000)` , and `sample_mean(m=1000, n=10000)` and plot a histogram for each of these outputs.
   - By solving the first question of the Theory part, write a class called `sample_mean_distribution` whose constructor takes an integer `m` as input and implements the probability measure

$$\overline{\text{Normal}(0, 1)}_m \triangleq \frac{1}{m} \sum_{i=1}^{m} \text{Normal}(0, 1)$$

     in other words, the distribution of the length-m estimator of the mean. Instantiate the objects `sample_mean_distribution(10)`, `sample_mean_distribution(100)`, `sample_mean_distribution(1000)` and plot their PDFs.

[1] Recall that the sum $\mathbb{P}_1 + \mathbb{P}_2$ is *defined* as the pushforward of the product measure $\mathbb{P}_1 \otimes \mathbb{P}_2$ under the operation $+ : \mathbb{R}^2 \to \mathbb{R}$. This distribution models the following random process: (a) sample from $\mathbb{P}_1$, (b) sample (independently) from $\mathbb{P}_2$, (c) add the two samples.

- Compare (a) the 3 histograms, (b) the 3 PDFs and (c) the histograms with the PDF. What conclusions do you draw?