PART IB Paper 7: Mathematics

PROBABILITY

Examples paper 6

Elementary exercises are marked: †, Tripos standard, but not necessarily Tripos length, are marked: ∗.

Characterising Distributions

1. † On page 11 of handout #3, a number of quantities are defined that statisticians like to use to characterise probability density functions: mean, variance, standard deviation, mode, median, quartiles, interquartile range, and skewness.

   The following probability density function is known as the "Rayleigh distribution" (named after Lord Rayleigh, a former chancellor of this university):

   $$f_{\mathrm{X}}(x) = \frac{x}{s^2} e^{-x^2/(2s^2)} \text{ for } x \geq 0.$$

   Find the cumulative density and all the characterising quantities mentioned above for the Rayleigh distribution for $s = 1$. You may find these through tedious integration, which is always a good exercise since half of your exam questions across all papers typically test your speed and accuracy at computing integrals. Or, if you feel that you've had enough training in integration, it is perfectly acceptable for you to search for "Rayleigh distribution" on the internet and copy these quantities out of a table.

Combining and Manipulating Distributions

2. Revisit the IA Paper 1 Thermofluid Mechanics Examples Paper 2 Question 10, assuming that the ping pong balls have the same constant velocity of 1 m/s but that the angle of the velocity is a uniformly distributed random variable over the interval $[-\pi/2, \pi/2]$ (we could assume it to be uniformly distributed over all directions but we can ignore ping pong balls flying away from the wall.)

   We reproduce the original question to save you from having to go fumble through your pile of old examples papers from last year:
   Ping pong balls of 10 g of mass hit a wall with a velocity of 1 m/s, and bounce back in the opposite direction with the same absolute velocity.
   (a) Determine the overall change in momentum for each ball.
   (b) If the average number of ping pong ball hits per unit of time is 10 per second, determine the momentum flux into and out of the wall, and the average force exerted by the balls on the wall.

3. ∗ Show that for a discrete random variable X with expected value $\mathbb{E}[\mathrm{X}]$, and such that $\mathbb{P}[\mathrm{X} < 0] = 0$, then

   (a) for each value of $t > 0$

   $$\mathbb{E}[\mathrm{X}] \;\geq\; t \sum_{x \geq t} P_{\mathrm{X}}(x), \qquad \text{i.e. that } \mathbb{P}[\mathrm{X} \geq t] \;\leq\; \frac{\mathbb{E}[\mathrm{X}]}{t}.$$

   Show further that if the standard deviation σ is known then

   $$\mathbb{P}[|\mathrm{X} - \mathbb{E}[\mathrm{X}]| \geq t] \;\leq\; \frac{\sigma^2}{t^2}, \qquad \text{and hence that} \qquad \mathbb{P}[|\mathrm{X} - \mathbb{E}[\mathrm{X}]| \geq k\sigma] \;\leq\; \frac{1}{k^2},$$

   where $k > 0$.

   (b) find a lower bound for the probability of X falling within 2 standard deviations of the mean

   (c) for a proposed interactive computer system it is estimated that the response time $\mathbb{E}[\mathrm{T}]$ is 0.5 seconds. Find an upper bound on the probability that the response time T will be 2 seconds or more

   (d) the standard deviation of the response time is 0.1 seconds. Place bounds on the probability that it will be between 0.25 and 0.75 seconds.

   (e) a data line transmits binary data independently with a probability $p = 0.2$ of error. Hence, the number of errors X in a data block of length n is a binomial $\mathrm{B}(n, p)$ random variable. For $n = 10$, calculate upper bounds on $\mathbb{P}[\mathrm{X} \geq 4]$ using the inequalities above, then calculate $\mathbb{P}[\mathrm{X} \geq 4]$ exactly. Note that $\mathbb{P}[|\mathrm{X} - \mathbb{E}[\mathrm{X}]| \geq 2] = P_{\mathrm{X}}(0) + \mathbb{P}[\mathrm{X} \geq 4]$. Now consider $n = 1000$ and compute upper bounds on $\mathbb{P}[\mathrm{X} \geq 400]$. Note that $\mathbb{P}[\mathrm{X} \geq 400]$ cannot be computed exactly because sums of binomials are hard to compute for any but small dimensions.

   (These are known as the Markov and Chebyshev inequalities respectively, and allow bounds to be placed on random variables when the corresponding probability density functions are unknown or hard to compute.)

4. A 105 mm wide container is designed to hold 10 components arranged side by side. The widths of the components are assumed to be independent and Gaussian distributed with mean 10 mm and standard deviation 1.0 mm. Calculate

   (a) the probability that 10 components will not fit into the container

   (b) the probability that 11 components will fit in the container.

Probability and Moment Generating functions

5. $\ast$ A random variable has an exponential probability density function given by:

$$f_{\mathrm{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } \lambda > 0.$$

   (a) Calculate the moment generating function and find a relationship between the mean and the variance.

   (b) What is the relationship between the mean and the variance for a random variable with Poisson distribution?

   (c) A small sample of measured times (in minutes) between arrival times of calls attempting to use a certain telephone exchange are given below:

   3.73 0.07 8.25 2.79 0.42 6.45 0.77 1.51 0.36 7.53 5.90 5.70 2.08 10.11 10.09 1.80 2.55 2.23 0.46 8.92
   4.40 5.00 13.24 4.84 3.31 11.54 7.42 9.39 3.75 1.39 13.89 31.38 17.48 11.91 2.26 4.29 0.46 3.27 3.92 2.20

   From this sample, does it appear that the traffic using the exchange is Poisson distributed?
   Note: the measurements above are available as a `csv` file on the course moodle site to save you having to type the numbers in.

6. If X_1 and X_2 are independent Poisson variables with parameters λ_1 and λ_2 respectively

   (a) show that $\mathrm{X}_1 + \mathrm{X}_2$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

   (b) Assume in the following that X_1 and X_2 represent the numbers of emissions per minute from two radioactive sources which have means 4 and 6 respectively. Find the probability that in any minute the total number of emissions from the two sources is equal to 2.

   (c) Find the probability that in any minute the value of X_1 is exactly twice the value of X_2.

   (d) Determine the mean and variance of $\mathrm{Z} = 3\mathrm{X}_1 - 2\mathrm{X}_2$.

7. $\ast$ If X and Y are independent random variables with Gaussian distributions

$$\mathrm{X} \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{and} \quad \mathrm{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

   (a) show that the random variable $\mathrm{Z} = \mathrm{X} - \mathrm{Y}$ has a probability density function which is also Gaussian with distribution $\mathrm{Z} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

   (b) A manufactured product is sold in cans. The cans have a weight which is Gaussian distributed with mean 200 g and standard deviation 9 g. The filling machine is set to give a total weight (can plus contents) with mean W and (known) standard deviation of 12 g. What should be the least value of W to ensure that less than 0.25% of the filled cans have contents weighing less than 1000 g?

8. $\ast$ Let X_1 and X_2 be independent Gaussian random variables as in the previous question, and let

$$\begin{bmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1/2 \end{bmatrix} \begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{bmatrix}$$

   (a) Characterise the marginal probability density functions of Y_1 and Y_2.

   (b) Express the probability $\mathbb{P}[\mathrm{Y}_2 \leq y_2 | \mathrm{Y}_1 = y_1]$ in function of the cumulative densities of X_1 and/or X_2 and hence give the conditional density $f_{\mathrm{Y}_2|\mathrm{Y}_1}(y_2|y_1)$.

   (c) Express the joint density of Y_1 and Y_2 and verify for the case $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$ that it matches the definition of a multivariate Gaussian given (with vectors and matrices) in page 24 of the handout #4.

   (d) In general, $\mathbf{X}$ is a vector of independent Gaussian random variables $\mathrm{X}_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$ and $\boldsymbol{A}$ is a given (non-random) $n \times n$ matrix, are the components of the vector $\mathbf{Y} = \boldsymbol{A}\mathbf{X}$ Gaussian? Are they independent?

Testing and Statistical Significance

9. † A web site receives traffic according to a Poisson distribution with intensity of 10 hits per day. On a randomly chosen day, the web site receives only 4 hits.

   (a) Does the traffic on this day constitute statistically significant evidence that something is unusual?

   (b) A null hypothesis $\mathcal{H}_0$ has been rejected at the $p = 5\%$ significance level. Which of the following statements are true: i) "the probability that the null hypothesis $\mathcal{H}_0$ is true is less than 5%", ii) "if the null hypothesis were true, the probability of the observed data or something more extreme is less than 5%".

Previous Tripos questions

	2019, 6.	2017, 6.	2014, 4.	2011, 4. b)	2008, 4 a) - e).
2022, 6. c)	2018, 6.	2015, 5.	2012, 4.	2010, 4. b) - e)	2007, 5 a) - d) and 6 a) - c)

Answers

1. $\mathbb{E}[\mathrm{X}] = \sqrt{\pi/2}$, $\mathrm{Var}[\mathrm{X}] = 2 - \pi/2$, $\sigma = \sqrt{2 - \pi/2}$, Mode: $x = 1$, Quartiles: $\sqrt{2\log 4/3}$, $\sqrt{2\log 2}$ and $\sqrt{2\log 4}$, interquartile range 0.91, skewness$\sqrt{\pi/2}(\pi - 3)/(2 - \pi/2)^{3/2} \approx 0.63$.

2. $\Delta m\mathbb{E}[\mathrm{V}_x] = -0.0127\,\mathrm{kg/s}$ and $\mathbb{E}[\mathrm{F}_x] = 0.127\,\mathrm{N}$.

3. b) $\mathbb{P}[|\mathrm{X} - \mathbb{E}[\mathrm{X}]| \leq 2\sigma] \geq \frac{3}{4}$, c) $\mathbb{P}[\mathrm{T} \geq 2] \leq \frac{1}{4}$, d) $\mathbb{P}[0.25 < \mathrm{T} < 0.75] \geq 0.84$,
   e) $\mathbb{P}[\mathrm{X} \geq 4] \leq 1/2$ using first inequality and $\mathbb{P}[\mathrm{X} \geq 4] \leq 0.29$ using second inequality, $\mathbb{P}[\mathrm{X} \geq 4] = 0.12$ by exact calculation. For $n = 1000$, first inequality $\mathbb{P}[\mathrm{X} \geq 400] \leq 1/2$ and second inequality $\mathbb{P}[\mathrm{X} \geq 400] \leq 0.004$.

4. a) 0.0571, b) 0.0655.

5. a) $g_{\mathrm{X}}(s) = \frac{\lambda}{\lambda - s}$ and $\mathrm{Var}[\mathrm{X}] = \mathbb{E}[\mathrm{X}]^2$, b) $\mathrm{Var}[\mathrm{X}] = \mathbb{E}[\mathrm{X}]$, c) yes.

6. b) 2.27×10^{-3}, c) $e^{-10}\sum_{k=0}^{\infty}\frac{96^k}{k!(2k)!}$, d) 0 and 60.

7. b) $W \geq 1242\,\mathrm{g}$.

8. d) yes, no

9. a) p-value 0.03, we can thus reject the null hypothesis, only statement ii) is true.

Thierry Savin
Lent 2024