

# S1: Principles of Data Science

## Problem Sheet 2

### MPhil in Data Intensive Science

Matt Kenzie  
mk652@cam.ac.uk

Michaelmas Term 2023

## Problem Sheet 2

*Lectures 7 – 12*

Topics covered: More distributions, generating from distributions, limit theorems, propagation of errors, estimates, minimum variance bound, the likelihood, maximum likelihood estimation, profile likelihood, extended likelihood, binned likelihood, least-squares, Wilks' theorem

1. For a two-dimensional normal distribution with parameters,  $\mu_1 = 1$ ,  $\mu_2 = 4$ ,  $\sigma_1 = 3$ ,  $\sigma_2 = 2$ ,  $\rho = 0.5$  make plots of the conditional and marginal probabilities. Extension (don't waste too much time on this), what if this is now a 3D Gaussian? Can you think of ways of presenting the equivalent information?
2. Show that the mean and variance of the exponential distribution with p.d.f.

$$p(X; \lambda) = \lambda e^{-\lambda X} \tag{1}$$

are given by

$$\text{Mean: } \mu = \frac{1}{\lambda} \tag{2}$$

$$\text{Variance: } V(X) = \frac{1}{\lambda^2} \tag{3}$$

3. Show that the mean and variance of the  $\chi^2$  distribution with p.d.f.

$$p(X; k) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} X^{k/2-1} e^{-X/2} \tag{4}$$

are given by

$$\text{Mean: } \mu = k \tag{5}$$

$$\text{Variance: } V(X) = 2k. \tag{6}$$

4. Write a simple accept-reject generator in python which can generate from an arbitrary function of a random variable,  $f(X)$ . Generate some samples for the following distributions:

(a)  $f(x) = \cos^2(x)$

(b)  $f(x) = \sin(x) + \cos(x) + 2$

(c)  $f(x) = \frac{\sin(x) + \cos(x)}{\sinh(x) + \cosh(x)} + 25$

Think about how you might adapt this to work for two-dimensions and then  $n$ -dimensions. Can you think about ways which would speed up the generation?

Can you now make a comparison of generation efficiencies with a 1D normal distributions. Let's take a standard normal distribution (*i.e.*  $\mu = 0$ ,  $\sigma = 1$ ). Make a plot of the acceptance efficiency as a function of the generation range in terms of standard deviations. What is the approximate accept efficiency if you generate all the way out to 8 standard deviations?

5. Using the Jacobian matrix derive the standard error propagation formula for  $\sigma_f$  given  $\sigma_x$  and  $\sigma_y$  for the following transformations

(a)  $f = x + y$

(b)  $f = xy$

(c)  $f = x/y$

(d)  $f = \sin(x)$

(e)  $f = \cos(x)$

assuming that  $x$  and  $y$  are independent.

6. Using the minimum variance bound show that the sample mean,  $\bar{X}$ , is an efficient estimate of the distribution mean,  $\mu$ .
7. Use the maximum likelihood method to mathematically show that an estimate for the lifetime,  $\tau$ , of a decay-time distribution,  $(1/\tau)e^{-t/\tau}$ , is given by the average of the observed decay times,  $t$ .
8. Show that when fitting a straight line to pairs of points  $(x_i, y_i)$  with the least squares method that estimates of the slope and intercept are given by  $\hat{c} = \bar{y} - \hat{m}\bar{x}$  and  $m = \text{cov}(x, y)/V(x)$ .