# S1: Principles of Data Science

# Problem Sheet 3

## MPhil in Data Intensive Science

Matt Kenzie
mk652@cam.ac.uk

Michealmas Term 2023

## Problem Sheet 3

*Lectures 13 – 18*
Topics covered: Method of moments, goodness of fit tests, confidence intervals, hypothesis testing, the Neymann-Pearson lemma, limit setting, resampling methods

1. Finish off the method of moments example we started in the lectures. Start from the probability distribution:

$$f(X;\alpha,\beta) = N(1+\alpha x+\beta x^2) \quad \text{for} \quad a \leq x \leq b \tag{1}$$

   (a) Show that the normalisation parameter $N$ can be written in the following form:

$$N = \frac{1}{d_1+\alpha d_2+\beta d_3} \quad \text{where} \quad d_k = \frac{1}{k}(b^k - a^k). \tag{2}$$

   (b) There are two unknowns so let's take the first two central moments. Compute the expectation values of the first and second central moments.

   (c) Solve these two equations to express $\alpha$ and $\beta$ in terms of the expectations of the first and second central moments. The method of moments estimates $\hat{\alpha}$ and $\hat{\beta}$ can then be computed by plugging in the sample estimates of the first and second central moments.

   (d) Generate a sample from the p.d.f. using accept-reject (in a previous problem sheet you should already have a function which can do this). Plot your sample along with the p.d.f. to check it looks ok.

   (e) Now compute your method of moments estimates for $\hat{\alpha}$ and $\hat{\beta}$ from the sample along with the covariance and compare it to the MLE method.

   (f) To make sure your method is working then you can use the example dataset provided in the `gitlab` repo under `datasets/mom_data.npy` (this is just an array

of 2000 points I have generated from the relevant function). For this dataset I found the following: Method of Moments Estimate:

$$\hat{\alpha} = 0.51 \pm 0.05 \tag{3}$$

$$\hat{\beta} = 0.50 \pm 0.11 \tag{4}$$

Maximum Likelihood Estimate:

$$\hat{\alpha} = 0.53 \pm 0.05 \tag{5}$$

$$\hat{\beta} = 0.54 \pm 0.11 \tag{6}$$

2. If 1000 measurements are grouped in 25 bins and fitted to a curve which is the sum of an arbitary Gaussian on an arbitary flat background, how many degrees of freedom are there?

3. Below are the number of neutrino events detected in 10 second intervals by the Irvine-Michigan-Brookhaven experiment on 23rd February 1987. For context the "background" should give a Poisson distribution of neutrino counts, but some kind of "signal" producing excess neutrinos (*e.g.* a supernova like S1987a which was discovered around this time) would provide an excess of neutrinos in a short interval.

| No. of events | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| No. of intervals | 1042 | 860 | 307 | 78 | 15 | 3 | 0 | 0 | 0 | 1 |

   (a) Assuming these are decribed by a Poisson distribution (with a floating mean) find an estimate of the value of the mean, and then compute the $\chi^2$ and $p$-value of this particular set of observations?

   (b) Now compute the same but excluding the data in the last bin (*i.e.* excluding the interval with 9 events)?

   (c) Use these results to justify whether this data describes a Poisson background only or not.

4. Explain with a short proof, why the Neyman-Pearson test is the statistically most powerful hypothesis test.

5. Imagine a measurement of four parameters yields the following result:

$$x^+ = (-9.3 \pm 8.2) \tag{7}$$

$$y^+ = (-1.3 \pm 8.4) \tag{8}$$

$$x^- = (\ \ 5.7 \pm 8.2) \tag{9}$$

$$y^- = (\ \ 6.5 \pm 8.3) \tag{10}$$

$$\tag{11}$$

with the following correlation matrix

| | $x^+$ | $y^+$ | $x^-$ | $y^-$ |
|---|---|---|---|---|
| $x^+$ | 1 | -0.1 | -0.05 | 0.1 |
| $y^+$ | -0.1 | 1 | 0.1 | -0.05 |
| $x^-$ | -0.05 | 0.1 | 1 | 0.1 |
| $y^-$ | 0.1 | -0.05 | 0.1 | 1 |

(a) Assuming the observations of $(x^{\pm}, y^{\pm})$ are multivariate normally distributed, use the ML method to estimate the values, uncertainties and correlations of the parameters $(\gamma, r, \delta)$ where

$$x^{\pm} = r\cos(\delta \pm \gamma) \tag{12}$$

$$y^{\pm} = r\sin(\delta \pm \gamma) \tag{13}$$

(b) Make a plot of the profile likelihood scan for the parameter $r$. Do you see any issues here? Can you use this is to make an estimate of the uncertainty on your estimate of $r$? What about if you want to quote the interval containing 95%?

(c) Can you compute (or at least explain how you would compute) the coverage of your interval on $\gamma$?

(d) Explain how you could set an upper limit on the parameter $r$.

6. Imagine a set of $N$ counting experiments which all determine an estimated number of background, $b_i$, an estimated number of signal, $s_i$, and observe $d_i$ candidates in data. Write down a mathematical expression for (the most) powerful test-statistic in this case.