Statistical Analysis of Data
Problem sheet #6

Please send by the announced due date to Glen Cowan, Physics Dept., Royal Holloway, University of London, Egham, Surrey, TW20 0EX, or e-mail to `g.cowan@rhbnc.ac.uk`.

Exercise 6.1: Galileo's studies of motion included experiments with a ball and an inclined ramp. The ball's trajectory is horizontal just before it falls over the edge, as shown in Fig. 1. The horizontal distance d from the edge to the point of impact is measured for different values of the initial height of the ball h. Five data points obtained by Galileo in 1608 are shown in Table 1.[1]

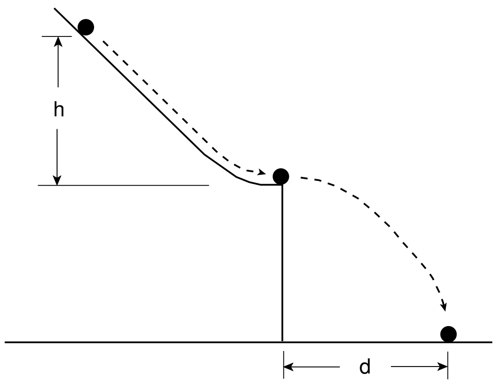

Figure 1: The configuration of the ball and ramp experiment performed by Galileo.

Table 1: Galileo's data on horizontal distance before impact d for five values of the starting height h. The units are punti (points); one punto is slightly less than 1 mm.

h	d
1000	1500
828	1340
800	1328
600	1172
300	800

Assume the heights h are known with negligible error, and that the horizontal distances d can be regarded as independent Gaussian random variables with standard deviations of $\sigma = 15$ punti. (It is not actually known how Galileo estimated the measurement uncertainties, but 1–2% is plausible.) In addition, we know that if $h = 0$, then the horizontal distance d will be zero, i.e. if the ball is started at the very edge of the ramp, it will fall straight down to the floor.

(a) Consider relations between h and d of the form

$$d = \alpha h \tag{1}$$

[1]See Stillman Drake and James Maclachlan, Galileo's discovery of the parabolic trajectory, *Scientific American* **232** (March 1975) 102; Stillman Drake, *Galileo at Work*, University of Chicago Press, Chicago (1978).

and

$$d = \alpha h + \beta h^2. \tag{2}$$

Find the least-squares estimators for the parameters α and β in closed form and evalute the estimates with the data from Table 1. (It will probably help to use the vector notation discussed in the lecture.) What are the values of the minimized χ^2 and the P-values for the two hypotheses?

(b) Assume a relation of the form

$$d = \alpha h^{\beta}\,. \tag{3}$$

Write a computer program to perform a least squares fit of α and β. Note that this is a nonlinear function of the parameters and must be solved numerically, e.g. with MINUIT.

(c) Galileo regarded the motion as the superposition of horizontal and vertical components, where the horizontal motion is of constant speed, and the vertical speed is zero at the lower edge of the ramp, but then increases in direct proportion to the time. This leads to a relation of the form

$$d = \alpha\sqrt{h}\,. \tag{4}$$

Find the least squares estimate for α and the value of the minimized χ^2. What is the P-value for this hypothesis?

Exercise 6.2: Consider an LS fit to a histogram with y_i entries in bins $i = 1, \dots, N$, with predicted values $\lambda_i(\boldsymbol{\theta})$. Suppose the total number of entries n is treated as a constant, so that the y_i are multinomially distributed.

(a) What is the covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_j]$? Why does the inverse of this matrix not exist?

(b) Consider the fit using only the first $N-1$ bins. Find the inverse covariance matrix, and show that this is equivalent to fitting to all N bins but without consideration of the correlations.

Exercise 6.3: Consider again Perrin's data on the number of mastic particles as a function of height (Exercise 6.5). Determine the LS estimates for Boltzmann's constant k (and equivalently Avogadro's number $N_{\mathrm{A}} = R/k$) and the coefficient ν_0 by minimizing

$$\chi^2(k, \nu_0) = \sum_{i=1}^{N} \frac{(n_i - \nu_i(k, \nu_0))^2}{\sigma_i^2}. \tag{5}$$

(a) Take the standard deviation σ_i of n_i to be $\sqrt{\nu_i}$ (the usual method of least squares).

(b) Take σ_i to be $\sqrt{n_i}$ (the modified method of least squares).

Compare the results from (a) and (b) to the estimates obtained by maximum likelihood in problem sheet #5.

Exercise 6.4: Suppose an estimator $\hat{\theta}$ is Gaussian distributed about the parameter's true value θ with a standard deviation $\sigma_{\hat{\theta}}$. Assume that $\sigma_{\hat{\theta}}$ is known.

(a) Sketch the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ defining the confidence belt (cf. SDA Section 9.2).

(b) Show that the central confidence interval for θ at a confidence level $1-\gamma$ is given by

$$[\hat{\theta} - \sigma_{\hat{\theta}} \Phi^{-1}(1-\gamma/2),\, \hat{\theta} + \sigma_{\hat{\theta}} \Phi^{-1}(1-\gamma/2)], \tag{6}$$

where Φ^{-1} is the quantile of the standard Gaussian (inverse of the cumulative distribution).

Exercise 6.5 (optional): Show that the upper and lower limits for the parameter p of a binomial distribution are

$$\begin{aligned} p_{\rm lo} &= \frac{n F_F^{-1}[\alpha; 2n, 2(N-n+1)]}{N-n+1 \;+\; n F_F^{-1}[\alpha; 2n, 2(N-n+1)]} \\ p_{\rm up} &= \frac{(n+1) F_F^{-1}[1-\beta; 2(n+1), 2(N-n)]}{(N-n) \;+\; (n+1) F_F^{-1}[1-\beta; 2(n+1), 2(N-n)]}\,. \end{aligned} \tag{7}$$

Here the confidence levels for the upper and lower limits are $1-\alpha$ and $1-\beta$, respectively, n is the number of successes observed in N trials, and F_F^{-1} is the quantile of the F distribution. This is defined by the p.d.f.

$$f(x; n_1, n_2) = \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{\Gamma(\frac{1}{2}(n_1+n_2))}{\Gamma(\frac{1}{2}n_1)\Gamma(\frac{1}{2}n_2)}\, x^{n_1/2-1} \left(1 + \frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2}, \tag{8}$$

where $x > 0$ and n_1 and n_2 are integer parameters (degrees of freedom). Use the fact that the cumulative binomial distribution is related to the cumulative distribution $F_F(x)$ for $n_1 = 2(n+1)$ and $n_2 = 2(N-n)$ degrees of freedom by [2]

$$\sum_{k=0}^{n} \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} \;=\; 1 \;-\; F_F\left[\frac{(N-n)p}{(n+1)(1-p)}; 2(n+1), 2(N-n)\right]. \tag{9}$$

Quantiles of the F distribution can be obtained from standard tables or computed with the routine `ffinv`. Equations (7) are implemented in the routines `binomlo`, `binomup` and `binomint`.

Exercise 6.6: At LEP II, the reaction $e^+e^- \to W^+W^-$ is used to study properties of the W boson. Suppose 10 W bosons are found (i.e. in 5 events), and out of the 10, 2 are found to decay to a muon and a neutrino.

(a) Find the 68.3% central confidence interval for the binomial parameter p for a W to decay to $\mu\nu$ (i.e. $p =$ branching ratio ($W \to \mu\nu$)). Express the answer as $p = \hat{p}^{+d}_{-c}$ where $\hat{p}$ is the ML estimate for p and $[\hat{p}-c, \hat{p}+d]$ is the confidence interval. (The routine `binomint` can be used.)

(b) Compare the interval from (a) to $\hat{p} \pm \hat{\sigma}_{\hat{p}}$, where $\hat{\sigma}_{\hat{p}}$ is the estimate of the standard deviation of $\hat{p}$.

(c) A common mistake is to regard the number of W bosons itself as a random variable and to include its variance in the error for $\hat{p}$ (e.g. using error propagation). Why is this not the correct approach for the error of a branching ratio?

[2] Use of the F distribution for evaluating binomial confidence intervals is due to A. Hald, *Statistical Theory with Engineering Applications*, John Wiley, New York, 1952.