Statistical Data Analysis
Problem sheet 4
Due Monday 9 November 2020

Please turn in a copy of the parts of the source code you wrote or modified and relevant output. Please do not turn in any large parts of source code that you use but did not modify.

Exercise 1(a) [5 marks] Suppose the independent random variables r_i are uniformly distributed between zero and one. Using the code (either C++ or Python) from

`http://www.pp.rhul.ac.uk/~cowan/stat/root/mc/`

as a starting point, write a computer program to make histograms of

(a) $x = r_1 + r_2 - 1$

(b) $x = r_1 + r_2 + r_3 + r_4 - 2$

(c) $x = \sum_{i=1}^{12} r_i \quad - 6$

1(b) [5 marks] Calculate exactly (i.e., in closed form using the exact values $E[r_i] = 1/2$, $V[r_i] = 1/12$) the means and variances of the variables defined in (a)–(c) and compare these to the values you obtain from the histograms of generated numbers. (If you are using python you can get the mean and standard deviation of a numpy array with `array.mean()` and `array.std()`. With C++ the information is displayed when you plot the histograms using ROOT; the value labeled "RMS" is the sample standard deviation.) Comment on the connection between your histograms and the central limit theorem. Remember to adjust the minimum and maximum x values of the histogram so that you cover all of the generated values.

Exercise 2 Consider a sawtooth p.d.f.,

$$f(x) = \begin{cases} \frac{2x}{x_{\max}^2} & 0 < x < x_{\max} \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{1}$$

2(a) [5 marks] Use the transformation method to find the function $x(r)$ to generate random numbers according $f(x)$. Implement the method in a short computer program and make a histogram of the results. (Use $x_{\max} = 1$.)

2(b) [5 marks] Write a program to generate random numbers according to the sawtooth p.d.f. using the acceptance-rejection technique. (Use $x_{\max} = 1$.) Plot a histogram of the results.

Exercise 3 Suppose $\vec{x} = (x_1, \ldots, x_n)$ follows an n-dimensional Gaussian distribution $f(\vec{x}; \vec{\mu}, V)$ with $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix $V_{ij} = \mathrm{cov}[x_i, x_j]$. (In the formulas below regard $\vec{x}$ and $\vec{\mu}$ to be column vectors.) Consider two hypotheses for the vector of means, $\vec{\mu}_0$ and $\vec{\mu}_1$, where for both one uses the same covariance matrix V.

3(a) [4 marks] Consider the test statistic

$$t(\vec{x}) = \ln \frac{f(\vec{x}|\vec{\mu}_1)}{f(\vec{x}|\vec{\mu}_0)} \,.$$

Show that this $t(\vec{x})$ can be written in the form

$$t(\vec{x}) = w_0 + \sum_{i=1}^{n} w_i x_i \ ,$$

or equivalently $t(\vec{x}) = w_0 + \vec{w}^T \vec{x}$, where $\vec{w}$ is a column vector of coefficients w_i, $i = 1, \ldots, n$.

3(b) [**4 marks**] Show that the coefficients in the vector $\vec{w}$ found in (a) satisfy the property of a Fisher discriminant, i.e.,

$$\vec{w} \propto W^{-1}(\vec{\mu}_1 - \vec{\mu}_0) \ ,$$

where W is the sum of the covariance matrices for the two hypotheses.

3(c) [**4 marks**] The prior probabilities for hypotheses of $\vec{\mu}_0$ and $\vec{\mu}_1$ are π_0 and π_1, respectively. Show that the posterior probability for $\vec{\mu}_0$ is

$$P(\vec{\mu}_0 | \vec{x}) = \frac{1}{1 + \frac{\pi_1}{\pi_0} e^t} \ ,$$

where $t = \ln[f(\vec{x}|\vec{\mu}_1)/f(\vec{x}|\vec{\mu}_0)]$ as before.

G. Cowan