EAS372 Assignment 3 Due: 19 March, 2013

Format: Please submit a tidy, organized report *in hard copy*, covering the exercise below. Report should be single-sided, double spaced with font size 12 pt. The page limit is **five**, not counting figures and tables.

Preamble: A working knowledge of basic statistics is extremely useful to meteorologists (one might almost say, imperative), and this assignment is intended to reinforce and to some extent develop your background in this vital subject. If the properties asked for are unfamiliar, please do your own reading (there are many textbooks on statistics in Cameron Library, and I have given a few Wikipedia links). For the calculations, I have no preferences as to the tool you adopt: you may write your own programs, or use Excel, or R or MATLAB or any other package of your choice.

Acknowledgement: Many thanks to Andrew Giles (Environment Canada, Kelowna) for having provided the data enabling this assignment .

Data: A three column text file gives the entire record of February minimum and maximum surface temperatures observed at your assigned weather station, with the entries in sequential order. Symbolically, one might refer to the data as $[i,\, T_i^{\vee},\, T_i^{\wedge},\ (i = 1, 2..., N)]$ where i is the row label (ranging up to N, the total number of days of record), and $T_i^{\vee}$, $T_i^{\wedge}$ are respectively the minimum and the maximum temperatures on that particular day of that particular February. The temperature data are given as integers, to be interpreted as $10\times(T^{\circ}\mathrm{C})$, i.e. $117 \longmapsto +11.7^{\circ}\mathrm{C}$ and $-11 \longmapsto -1.1^{\circ}\mathrm{C}$. Thus your file looks something like

$$\begin{array}{ccc} 1 & -117 & -65 \\ 2 & -143 & -94 \\ 3 & \mathrm{M} & -100 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1680 & 76 & 10 \end{array} \qquad (1)$$

("M" means "missing data" – you may delete any such line, or find some other way to accommodate missing data that does not upset your computing method).

Task: For each series, please

1. for each series (column), determine the extreme minimum, extreme maximum, mean and standard deviation. As to symbols, in the case of the daily minimums, use the terminology $T^{\vee,min}$ (the extreme lowest daily minimum), $T^{\vee,max}$ (the extreme highest daily minimum), $\overline{T^{\vee}}$ (mean of the daily minimums) and $\sigma_{T^{\vee}}$ (standard deviation of the daily minimums). Adopt corresponding symbols for attributes of the series of daily maxima

2. sort your two temperature series from smallest to largest member and, for each, determine the 33rd percentile, the median and the 66th percentile (the 1/3 and 2/3 percentiles should equate to the values you are using for assignment 2 to define your three equi-probable classes)

3. give the histogram (`http://en.wikipedia.org/wiki/Histogram`) of each variable, using temperature bins of a sufficient width ($\Delta T^{\vee}$) and number that your histogram covers the interval from the extreme lowest to the extreme highest value, and is reasonably smooth. Each bin is centred on a temperature $T_j^{\vee}$ and spans $T_j^{\vee} \pm \Delta T^{\vee}/2$. (Here j labels your bins, and has nothing to do with the row label i. As a starting point you might take 30 bins with $\Delta T^{\vee}$ specified as $(T^{\vee,max} - T^{\vee,min})/30$, i.e. one 30th of the difference between the extreme max and the extreme min.) Your histogram is the set of relative frequencies $F_j = n_j/N$ associated with your bins, n_j being the sample count in bin j and N being your total number of samples N (where $N \approx 60 \times 28$, if you had a station with a 60 year record)

4. convert your histogram to an empirical probability density function[1] (or "PDF") $f_j = F_j/\Delta T$ (the unit, obviously, is K^{-1}). Tabulate and plot these PDFs, i.e. (taking the case of the series of minimums) tabulate and plot the values of $f_j = F_j/\Delta T$ versus the bin centre-point value $T_j^{\vee}$. On the same axes, plot *as a continuous line* the Gaussian (i.e.

[1] In their textbook titled "Intro Stats," which you may have used for STAT 141 and/or STAT 151, De Veau, Velleman & Bock use the term "probability model" for the PDF (Intro Stats, 3rd edition, p445). A useful "calculus-literate" overview is given by Wikipedia, Section 1 at `http://en.wikipedia.org/wiki/Probability_density_function`, and the relationship between the PDF and the moments is covered at `http://en.wikipedia.org/wiki/Moment_mathematics`.

Normal) PDF, e.g. for the minimums

$$f_{\vee}(T^{\vee}) = \frac{1}{\sqrt{2\pi}\ \sigma_{T\vee}}\ \exp\left(-\frac{(T_{\vee} - \overline{T^{\vee}})^2}{2\sigma_{T\vee}^2}\right)\ . \tag{2}$$

Compute the values you would get for the 33rd and 66th percentiles if the distribution actually *were* Normal (adopting the true observed mean and standard deviation), and compare with the true percentiles

5. the moments (about the mean) of a distribution whose PDF is $f(x)$ may be computed as integrals

$$\overline{(x-\overline{x})^n} = \int_{x=-\infty}^{\infty} (x-\overline{x})^n\ f(x)\ dx\ , \tag{3}$$

so that for instance (taking $n = 2$) the variance is

$$\sigma_x^2 = \int_{x=-\infty}^{\infty} (x-\overline{x})^2\ f(x)\ dx\ . \tag{4}$$

From your empirical PDFs, deduce the implied variances for $T^{\vee}, T^{\wedge}$ by summation[2], e.g. for the minima

$$\sigma_{T\vee}^2 \approx \sum_j (T_j^{\vee} - \overline{T^{\vee}})^2\ f(T_j^{\vee})\ \Delta T^{\vee} \tag{5}$$

where the $T_j^{\vee}$ are your bin centre-point values, $f(T_j^{\vee})$ is the PDF you determined above (effectively, if you like, a table of numbers), and $\Delta T^{\vee}$ is your bin width. Compare your answers for $\sigma_{T\vee}$, $\sigma_{T^{\wedge}}$ with your directly computed values.

[2]This summation is a *numerical* integration, i.e. a discrete (thus, inexact) approximation to the integral.