**AS5001 (SUPAAAA) : Advanced (Astronomical) Data Analysis**
**Problem Set 1 —— Due Mon 3 Oct 2022.**

**Problem 1. Histogram and Cumulative Distributions of Random Variables [25]**

(a) Be encouraged to build up a toolkit of subroutines, each doing a specific job with well defined inputs and outputs, that you can then use again and again in different contexts. Write a function subroutine that generates a random number distributed uniformly between specified limits $A$ and $B$. Document in the code the inputs and outputs of each subroutine. For example, in FORTRAN your function subroutine RANU might start with comments as follows:

```
      FUNCTION RANU( A, B, ISEED )
! Return on each call a new random number RANU sampling
! a uniform distribution between limits A and B, and
! a new value of the seed integer ISEED to be used on the next call.
! In:  A      R4 lower limit
!      B      R4 upper limit
!      ISEED  I4 seed integer for input to the random number generator
! Out: RANU   R4 uniform random number
!      ISEED  I4 seed integer to be used on the next call
```

Note: This should be easy, for example, using FORTRAN's intrinsic function RAN(ISEED), or equivalent, to obtain a random number distributed uniformly between 0 and 1. Design RANU so that ISEED (or equivalent) can be used to repeat the same sequence of random numbers, if required.

Use RANU to generate 10 random numbers uniformly distributed between -10 and +10. Print those 10 values. Now use RANU a second time to re-generate the same sequence of random numbers and print them to show that they are the same as the first sequence.

(b) Write a program that plots a binned histogram and the cumulative (staircase) distribution for $N$ samples drawn from your uniform random number generator RANU. The staircase should step up at each of the $N$ values, not just at the edges of the histrogram bins, so that the finite bin size of your histogram does not degrade the resolution of your cumulative distribution. Make plots for $N = 10, 100, 1000, 10^4$.

(c) Write similar function subroutines RANG(AVG,SIG,ISEED) for Gaussian and RANE(TAU,ISEED) for Exponential random numbers. These should transform values from RANU into samples of the Gaussian and Exponential distributions. Make plots as in (b) for $N = 10^4$.

**Problem 2. Mean and Variance, Median and MAD [25]**

(a) Write a subroutine that computes the the sample mean and sample variance of an input array of data values. Take care to ensure that your sample mean and variance are unbiased estimators. As always, describe clearly the input and output arguments of the subroutine in comments at the top, e.g.

```
      SUBROUTINE MEANVAR( N, X, AVG, VAR )
! compute the sample mean and sample variance of data X(N)
! In:  N     I4 number of data
!      X(N)  R4 array of data values
! Out: AVG   R4 sample mean of data
!      VAR   R4 sample variance of the data
```

(b) Write a subroutine that computes the sample median, and the mean-absolute-deviation (MAD) for an array of data. Note: To compute the median, sort the data values, find the value half-way through the sorted list, interpolating if needed.

```
      SUBROUTINE MEDMAD( N, X, XMED, XMAD )
! compute median and mean-absolute-deviation of data X(N)
! In:  N      I4 number of data
!      X(N)   R4 array of data values
! Out: XMED   R4 median of the data values
!      XMAD   R4 mean absolute deviation (relative to median)
```

Sanity-check MEANVAR and MEDMAD using e.g. (1,2,3) and (1,2,3,4) as input.

(c) I'll supply a dataset from part of a zero-exposure CCD frame. Use your routines MEANVAR and MEDMAD to compute statistics that estimate the CCD bias level (mean and median) and readout noise (standard deviation and mean-absolute-deviation). Summarise your results in a table, giving also a 1-$\sigma$ error bar for your sample mean, for the standard deviation, and (if possible) for the median.

**Problem 3. Central Moments of a Distribution** **[25]**

(a) Write a subroutine that computes statistics that estimate the mean, variance, and higher-order central moments of a dataset.

```
      SUBROUTINE MOMENTS( N, X, M, XMOM )
! compute first M central moments of the data points X(N)
! Inp: N      I4 number of data points
!      X(N)   R4 array of data points
!      M      I4 number of moments
! Out: XMOM(1) R4 sample mean
!      XMOM(2) R4 sample variance (unbiased estimate for sigma^2}
!      XMOM(M) R4 higher central moments (normalised by sigma^M)
```

(b) Use your subroutines RANU($a$,$b$), RANG($\mu$,$\sigma$) and RANE($\tau$) (Problem 2) to generate $N \sim 10^6$ random samples of Uniform, Gaussian, and Exponential distributions. Use MOMENTS to compute the first 4 central moments in each case. To check your results, derive analytic formulae for the first 4 central moments in terms of the parameters of each distribution. (The required integrals are not difficult. If you get stuck, for partial credit, use http://integrals.wolfram.com.) Prepare a table comparing your numerical and analytic results (test at least 2 sets of input parameters) to demonstrate the accuracy of your subroutines, and comment on reasons for any discrepancies.

**Problem 4. Monte-Carlo Simulation to compare Sample Means and Medians** **[25]**

(a) The median $X_{med}$ is a "robust" statistic, less sensitive to large outliers than the sample mean $\overline{X}$. For this reason, a "running median filter" is often used to "smooth" or remove "spikes" from a light curve or spectrum. Medians are also often used to combine a series of CCD frames so that the result is not badly affected by cosmic ray hits. However, $X_{med}$ is a noisier statistic than $\overline{X}$, as you will discover in the following Monte-Carlo calculation.

Use your subroutine RANG($\mu$,$\sigma$) to draw $N$ samples from a Gaussian distribution with $(\mu, \sigma) = (0, 1)$. Calculate the sample mean $\overline{X}$ and median $X_{med}$ for the set of $N$ samples. Repeat the above calculation $M$ times, and characterise the distributions of $X_{med}$ and $\overline{X}$ by calculating their mean values and standard deviations over the $M$ samples. Choose $M$ large enough to keep uncertainties below 1% of the standard deviations. Make a log-log plot showing from your Monte-Carlo simulation how the standard deviations of $X_{med}$ and $\overline{X}$ vary with $N$ for $N = 1, 2, 3, ..., 100$. (If this calculation takes a very long time, try to improve the efficiency of your code.) Which statistic has the larger variance, $X_{med}$ or $\overline{X}$? Based on your results, decide whether it is better to take medians with an odd or even number of data points. Explain your reasoning.

(b) For a sample of $N$ Gaussian random variables (each with mean 0 and standard deviation $\sigma$), derive an analytic formula in terms of $N$ and $\sigma$ for $\sigma^2(\overline{X})$, the variance of the sample mean. Derive a similar formula for $\sigma^2(X_{med})$, in the limit of large $N$. Which is larger, $\sigma^2(\overline{X})$ or $\sigma^2(X_{med})$? Compare your formulae with the large-$N$ results from your Monte-Carlo simulation.