# EPS 659a — Time Series in Geoscience

## Problem Set Three

*due Friday, Sept 24, 2021*

*Problem 1: Means and Medians of Datasets*

Consider the following functions of the sequence $\{X_n\}_{n=1}^N$, where $X_n = n^2$ and $N = 9$. Use R to compute these quantities:

a) the sample mean $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_n$

b) The sample median of the $\{X_n\}_{n=1}^N$

c) The sample harmonic mean $\mu_H$, where $(\mu_H)^{-1} = \frac{1}{N}\sum_{i=1}^N (X_n)^{-1}$

d) The sample geometric mean $\mu_G$, where $\ln \mu_G = \frac{1}{N}\sum_{i=1}^N \ln X_n$

*Problem 2: Means and Medians of continuous PDFs*

Consider the $\chi^2$-distribution with 4 degrees of freedom (unit variance):

$$p(X) = \frac{1}{4} X \exp(-X/2)$$

defined for $X \in [0, \infty)$. Compute the following functions (a) thru d) via Calculus, part e) via a numerical integration in R.

a) the cumulative density function (CDF) $P(X) = \int_0^X p(y)dy$ for the probability that the random variable $< X$.

b) the mean $\mu = \int_0^\infty X p(X) dX$

c) The median of the PDF, the value $\mu_M$, for which $\frac{1}{2} = \int_0^{\mu_M} p(X)dX$

d) The harmonic mean $\mu_H$, where $(\mu_H)^{-1} = \int_0^\infty \frac{p(X)}{X} dX$

e) The geometric mean $\mu_G$, where $\ln \mu_G = \int_0^\infty (\ln X) p(X) dX$. This integral lacks a closed-form solution, so evaluate it numerically. You cannot integrate to infinity in a computer, of course, so integrate from 0 to 50, and from 0 to 100 and report if the integral appears to be converging. Because $X \ln X$ is formally undefined at $X = 0$, you cannot evaluate it there. However, convergence of the numerical integral is bolstered by $\lim_{X\to 0}(X \ln X) = 0$, suggesting that the integrand remains bounded.[1] You can loop the sum by evaluating the integrand for

[1] You can verify this limit by evaluating a sequence of terms $\{X_k \ln X_k\}_{k=1}^\infty$ where $X_k = 2^{-k}$. A little algebra will show that the ratio of the $(k+1)$th term to the $k$th term approaches 1/2 as $k \to \infty$. This implies that the sequence $\{X_k \ln X_k\}_{k=1}^\infty$ shrinks geometrically as $k$ increases and $X_k \to 0$. You can run the R loop command "`for (k in 1:20){print(2^{-k}*log(2^{-k}))}`" to verify this.

$\Delta X = 0.01$ and $X_k = (k - 1/2)\Delta X$, for $k = 1, \ldots M$ where $M = 5000$ or $10000$:

$$\ln \mu_G \approx \sum_{k=1}^{M} \ln X_k p(X_k) \Delta X$$

Plot the integrand to clarify your understanding of the function that you are integrating here.

*Problem 3: Statistical tests of data sets: Sunspots*

On the Canvas site you will find the file `sunspots_annual_1700_2008.csv`. These are annual-average sunspots from 1700AD to 2008AD, distributed as part of the Anaconda software distribution. (Anaconda doesn't give an academic citation for the time series.). Treating the observed sunspot number $S$ as a random variable, let us test its statistics.

a) Find the maximum and minimum values for $\{S_k\}_{k=1700}^{2008}$ and use R to plot a histogram of sunspot frequency using 20 equally-spaced bins.

b) compute and report the sample mean and standard deviation (about the mean – use the sample variance) for the sunspot data.

c) plot a Gaussian PDF with mean and variance from part b). Scaling the PDF by a factor to compare with the histogram from part a) involves some normalization factors, so first plot the PDF and Histograms in different subplots, using the same X-axis between the minimum and maximum sunspot numbers. Then plot them on the same graph once you figure out the normalization. (The notebook template `Pset2_template.Rmd` has an example of normalizing this type of comparison plot.) Does the PDF look OK?

d) Find the maximum and minimum values for $\{\ln(10 + S_k)\}_{k=1700}^{2008}$ and use R to plot a histogram of adjusted-log(sunspot) frequency using 50 equally-spaced bins. Compute and report the sample mean and standard deviation (about the mean – use the sample variance) for the natural logarithm of the sunspot data ($\ln(10 + S_k)$).

e) plot a Gaussian PDF with mean and variance from part d). Plot the hypothetical PDF in a different subplot and compare with the histogram from part d). Does the PDF resemble the histogram better than in Part (c)?

f) Sort the values $S_k$ from smallest to largest. Compute and plot the empirical CDF.

g) Compute and plot the CDF numerically for $S > 0$ for a modified log-normal PDF

$$p(S) = \frac{1}{\sqrt{2\pi}\sigma(10 + S)} \exp\left(-(\ln(10 + S) - \mu)^2/2\sigma^2\right)$$

by guessing three reasonable values for the mean $\mu$ and standard deviation $\sigma$ of $\ln S$, and plotting these against the CDF for the sunspot data $\{S_k\}_{k=1700}^{2008}$. Because $\ln S$ has a singularity at $S = 0$, the damping constant 10 makes the integral easier to integrate numerically over the range of sunspot observations $S$. Because we avoid that part of the random variable that lies close to the singularity, however, the CDF never reaches unity as $S \to \infty$. You can renormalize the PDF empirically by multiplying it by $1/\max(cdf)$ where $\max(cdf)$ is the limiting value, if you like.

*Problem 4: Negative Binomial Distribution*

What is the probability that an event with probability $s$ occurs $k$ times in $N \geq k$ trials? This probability is given by the $k$th term in the binomial distribution

$$binom(k,N) = \binom{N}{k} s^k (1-s)^{N-k}$$

The binomial coefficient $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ can be computed in R with the command `choose(N,k)`. You can verify this function with $\binom{4}{2} = 6$. The negative binomial distribution $nbin(k,N)$ addresses a related question. What is the probability that an event with probability $s$ occurs for exactly the $k$th time in the $N$th trial, where $N \geq k$? This probability is equivalent to there being $k-1$ events in the first $N-1$ trials, and then the $N$th trial being the desired event. The first probability is the binomial term from $N-1$ trials, and the second probability is $s$. Therefore the "negative" binomial probability of reaching $k$ events on the $N$th trial is

$$\begin{aligned} nbin(k,N) &= binom(k-1,N-1) \times s = \tbinom{N-1}{k-1} s^{k-1}(1-s)^{N-k} \times s = \tbinom{N-1}{k-1} s^k (1-s)^{N-k} \\ &= \frac{(N-1)!}{(k-1)!(N-k)!} s^k (1-s)^{N-k} = \tfrac{k}{N} \times binom(k,N) \end{aligned}$$

a) Use the above formulas to compute the probabilities $nbin(k,N)$ explicitly in R for $k=4$, $s=0.5$ and $N \in [4,20]$, and plot the results.

b) Use the R function for the negative binomial distribution (`dnbinom`) to compute the same probabilities for the same $k, s$ and range of $N$. Compute and plot the cumulative probabilities (`pnbinom`) and confirm that the CDF approaches unity as $N \to \infty$.

c) Use the R function for the negative binomial distribution (`dnbinom`) to compute the probabilities for $k=4$, $s=0.1$ and a range of $N \in [k,100]$. Compute and plot the cumulative probabilities (`pnbinom`) and confirm that the CDF approaches unity as $N \to \infty$.