EPS 659a — Time Series in Geoscience

Problem Set Four

due Friday, October 1, 2021

Problem 1: Sample Means and Sample Variances, and Empirical CDFs of Global Temperature Anomalies

On the Canvas site you will find subdirectory `Probset4` with datasets that contain time series of monthly temperature anomalies from 1850 into 2021. The source of data is the UK Hadley Center temperature dataset, circa July 2021. Consider the global-average monthly temperature anomalies, that is, the file `Hadcrut_GlobalAverage.csv`. The Hadley Center datasets are described by

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *J. Geophys. Res., 117*, D08101, doi:10.1029/2011JD017187.

The dataset can be read into R as a dataframe, using the headers for the different columns of data. The columns are explained in the file `Notes_AvgData.txt`. Commands for reading the .csv file are in the template R markdown file `probset4_template.Rmd`. The time ordinate t can be constructed with $t = Y + (M - 0.5)/12.0$ where Y is the year and M is the month.

(a) Plot the entire data series against time from 1850AD to 2021AD in black, and atop the same graph plot the running average of the dataseries, summing over the previous five months and the following five months. If T_i are the temperature anomaly data, the running average will be

$$\tilde{T_k} = \frac{1}{11} \sum_{j=-5}^{5} T_{k+j}$$

this sum is ill-defined at the ends of the time series, so only compute and plot it (in light green atop the black curve of raw data) from June 1850 thru January 2021.

(b) For the full 1850-2021 data series, compute the sample mean $\bar{T}_N = \frac{1}{N} \sum_{n=1}^{N} T_n$ and sample variance $var(T - \bar{T}_N) = \frac{1}{N-1} \sum_{n=1}^{N} (T_n - \bar{T}_N)^2$. (Note that the R functions `mean()` and `var()` perform these calculations.) Plot a histogram of the data and superimpose a Gaussian PDF calculated for the sample mean and sample variance and scaled to match the histogram (e.g. by multiplying the PDF by N, the number of datapoints, divided by the binning interval).

(c) Draw a Q-Q plot that compares the empirical CDF $\hat{F}(T)$ of the Hadley-Center global average temperature anomalies, where T is the temperature anomaly, against the Gaussian CDF $F(T)$ parameterized by the sample mean and sample variance. Comment qualitatively on the deviations. Find the largest deviation $|\hat{F}(T_n) - F(T_n)|$ between the CDFs, among the data values T_n.

(d) Use the Beta distribution to estimate the relative probability of the median of the data set deviating from the theoretical prediction of the best-fitting Gaussian PDF. The Beta PDF for the median T_M of the Gaussian for N data involves the midpoint of the Gaussian CDF,

that is, at the value where $F(T_M) = 1 - F(T_M) = 0.5$.

$$beta(F(T_M)) = \frac{\Gamma(N+2)}{\Gamma(N/2+1)\Gamma(N/2+1)} F(T_M)^{N/2}(1-F(T_M))^{N/2} = \frac{\Gamma(N+2)}{\Gamma(N/2+1)\Gamma(N/2+1)}(\frac{1}{4})^{N/2}$$

If the data series has median $\hat{T}_M$, a temperature value for which the Gaussian CDF is $F(\hat{T}_M) = 0.5 + \rho$, then the Beta distribution PDF will be

$$beta(F(\hat{T}_M)) = \frac{\Gamma(N+2)}{\Gamma(N/2+1)\Gamma(N/2+1)} F(\hat{T}_M)^{N/2}(1-F(\hat{T}_M))^{N/2} = \frac{\Gamma(N+2)}{\Gamma(N/2+1)\Gamma(N/2+1)}(\frac{1}{4}-\rho^2)^{N/2}$$

so that the Gaussian-based PDF at the sample median will be smaller than at the theoretical median by a factor of

$$factor = (1 - 4\rho^2)^{N/2}$$

If $N = 10$ and $\rho = 0.05$, this factor is 0.951 for $N = 10$, which is fairly close to one. The factor is 0.605 for $N = 100$, which suggests that $\rho = 0.05$ is a significant deviation, but not extreme. The factor is 0.00657 for $N = 1000$, suggesting an extreme departure from expected behavior. What is the relative Beta-distribution probability factor for the sample median of the Hadley Center data series?

(e) Compute the sample mean and variance for temperature anomalies in successive decades of the time series, i.e., Jan-1850 to Dec-1859, Jan-1860 to Dec-1869, etc. Plot both of these quantities against the decadal midpoints (1855, 1865, etc). Convert the sample variances of the temperature anomalies into standard deviations of the sample-means and plot as error bars on the decadal-average sample means. (Google the `arrows` function in R to plot error bars.) Is global warming evident in this exercise?

(f) Consider the global-average temperature anomalies of the first and last decades of the 20th century, that is, 120 data points each. Use the sample means and variances of these anomalies to generate Gaussian PDFs for each decade, and plot the theoretical $F(X)$ and empirical $\hat{F}(X)$ CDFs for the two decades against each other. Comment on whether global warming seems to have occurred in the 20th century.

Problem 2: Do Large Earthquakes Follow Poisson Statistics?

Consider the Poisson distribution with event rate λ for an integer K number of events in a unit time interval

$$p(K) = \frac{\lambda^K}{K!} e^{-\lambda}$$

defined for $K \in [0, \infty)$. Note that unit normalization of this discrete PDF is assured by summing the infinite series

$$\sum_{K=0}^{\infty} \frac{\lambda^K}{K!} = e^{\lambda}$$

(a) Verify mathematically that the expectation value of events $\langle K \rangle = \lambda$.

(b) Verify that the expectation value $\langle K^2 \rangle = \lambda^2 + \lambda$.

(c) On the Canvas site you will find subdirectory `Probset4` that contains the dataset `GlobalEarthquakes.csv`, a list of all earthquakes with Richter magnitudes $M \geq 6.0$ that can

be obtained from IRIS for the 1980-2020 time interval via the ObsPy command `get_events`. Read this dataframe into R and plot the empirical CDF $F(M)$ of events against Richter magnitude M.

(d) Make a second plot of M against $(1 - F(M))$ on a logarithmic y-axis to demonstrate the power-law behavior of earthquake activity.

(e) Create a 41-point time series for the years 1980, 1981, ... 2020 with the number of earthquakes each year with $M \geq 8.0$. Compute the sample mean for this data set, set it equal to λ and compute the theoretical CDF of a Poission distribution with this value using R. Plot the theoretical PDF as a function of K against the empirical PDF based on your 41-point dataset. Qualitatively, would you be convinced that "great" earthquakes on Earth follow Poisson statistics?

(f) Create a 41-point time series for the years 1980, 1981, ... 2020 with the number of earthquakes each year with $M \geq 7.5$. Compute the sample mean for this data set, set it equal to λ and compute the theoretical CDF of a Poission distribution with this value using R. Plot the theoretical PDF as a function of K against the empirical PDF based on your 41-point dataset. Qualitatively, would you be convinced that "very large" earthquakes on Earth follow Poisson statistics?