Homework #2: Data Analysis and Geostatistics, Fall 2020

September 8, 2020

Overview and dataset

Download the high resolution DEM from Colorado from the course Google Drive, under the HW2 folder. The dataset is in an ascii file called *elevations.txt*.

Assume this dataset is the true elevation for this area, with no uncertainty. For modeling purposes, you need to know the mean, minimum, and maximum elevation of this 1km x 1km study site, and you are tasked with collecting point elevation measurements to get this information needed. Measurements are time consuming and you can only make a finite number of them. In this homework, you will estimate the mean, standard deviation, minimum and maximum elevation in this area using different sample sizes and sampling strategies. These approaches are not unique to elevation measurements; the spatial observation could be almost anything.

Using Monte-Carlo Simulation (*random sampling*) to test sampling strategies

1. Assuming no uncertainty and large enough sample size, we will use the complete dataset to define the true statistics of the underlying distribution of elevation. Calculate the "true" minimum q_0, maximum q_{100}, standard deviation σ and mean μ, using the entire dataset.

2. Plot a relative density histogram of the entire elevation dataset

3. Randomly sample 10 measurements from the *elevations.txt* dataset. Calculate the minimum, maximum, and mean elevation.

4. Repeat 1000 times, storing the mean, standard deviation, minimum, and maximum elevation each time.

5. Plot a relative density histogram of each statistic.

6. The Gaussian (normal) distribution has the form

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu).^2}{2\sigma^2}} \tag{1}$$

   for mean or expectation value μ and standard deviation σ. Fit the Gaussian distribution to the entire elevation data set, using the true mean μ and true standard deviation σ. Plot this Gaussian curve on the same plot as your relative density histogram.

7. Fit the Gaussian distribution to each sample statistic, using your new dataset from the 1000 random subsample datasets of elevation - you have a dataset of 1000 values of $\hat{\mu}, \hat{\sigma}, \hat{q}_0, \hat{q}_{100}$. Each of these sample statistics has a sample mean and sample standard deviation value that can be used as the parameters in the Gaussian distribution formula. Plot the Gaussian curves on your relative density histograms from above.

8. What is the probability of measuring a value less than the true mean?

9. What is the probability of measuring a minimum and maximum value within 1% of the true value?

10. What is the range of elevations that contains 68% of your measured mean values?

11. Vary your sample size and plot the sample statistics along with their uncertainties, as a function of sample size.

12. Plot the relative density histogram and normal pdf for uniform sampling with a spacing of 200 meters. Assume the dataset spans 1000m x 1000m, and sample uniformly in both directions.

13. Repeat for uniform sampling with a spacing of 30 meters.

14. Plot the sample statistics and their uncertainties for uniform sampling, as a function of sample size.