Homework 1 [100 pts.]

Bayes Theorem, introduction to the central limit theorem, moments, introduction to subsampling and bootstrapping, LaTeX

This homework will have 10 points go toward good figures (i.e. captions, legends, clear, high-resolution, etc.)

Reminder of how to earn extra credit

You will receive 10 extra-credit homework points if you use LaTeX to typeset the solutions to at least one homework. A LaTeX template and a list of beginner's guides can be found on the course webpage. The LaTeX code used to create this homework is also on the website so you can see how it was made.

Problem 1: Distributions and probabilities [30 pts.]

On the website, you will find two files called *homework_1_data_Y1.csv* and *homework_1_data_Y2.csv*. Within these files are 2 time series (Y_1, Y_2 respectfully).

- Y_1 is of hourly air temperature at Christman Field (2013-2014) in units of °F,
- Y_2 is of hourly wind speeds at Christman Field (2013-2014) in units of mph,

1. Find your own data set with at least 1000 samples (talk to me or the TA if you don't have any data that fits this requirement). Let's call this data set Y_3.
2. Standardize the three time series (Y_1, Y_2, Y_3) and plot estimates of their probability density functions[1] on the same plot. Use a bin size of 0.2 standard deviations and give the y-axis units of density.
3. Create a 4th time series, called Z, that has a length of 100,000 and is created from 100,000 samples of a standard normal distribution. Plot the probability density function of Z on the same plot as that of Y_1, Y_2, Y_3 using the same bin increment. Which of the Y time series look like the normal distribution Z? Which do not?
4. For each of the four *standardized* time series, answer the following question: What is the probability that any one particular measurement is +2 standard deviations or greater? What physical value does a standardized value of +2 standard deviations correspond to for the Y_is?
5. Imagine you go out today and measure the wind speed (Y_2) and you obtain a standardized value of +3 standard deviations. How rare of an event is this (how often does a value of +3 or more occur)? How rare would you think this event was if you erroneously assumed that wind speed followed a normal distribution?
6. Provide a few sentences about your data (Y_3) and what you have learned about its distribution and what this may tell you about the processes at play.

[1] To be a true PDF, the integral under the curve must equal 1.0. So, although your data is not actually continuous, you still want to make sure that the sum of your frequencies over all possible x values, multiplied by the spacing between your x values, is equal to 1.0.

Problem 2: Central Limit Theorem and N [30 pts.]

Now, we will explore the Central Limit Theorem[2] in the context of the four time series Z, Y_1, Y_2, Y_3. The Central Limit Theorem states that for samples of large enough length (N), the distribution of *sample means* will follow a normal distribution and thus we can apply gaussian statistics. For this problem, please use standardized[3] versions of the four time series (Z, Y_1, Y_2, Y_3) from Problem 1.

Over the next few weeks, you go out and collect a sample of N measurements, one sample for each time series. You find that the mean of each sample is -0.35 standard deviations. You wonder: How rare is a value of -0.35? Your goal is to quantify the "rarity" of getting a sample mean as extreme as -0.35 for the four time series as a function of the sample length N (letting N vary between 5 and 60 in increments of 5).

1. Plot estimates of the probability density functions of the sample means for the four time series for $N = 20$.
2. Plot the frequency of the sample mean being $\leqslant -0.35$ as a function of N for the four time series.
3. Do the approximate probability density functions of the different time series look the same? Different? Why?
4. What has this exercise shown you about assuming gaussian statistics for sample means?

Hint

To explore this question, you will need to "resample" the time series over and over again in order to obtain the distribution of the sample means. To start, let $N = 20$ and write code that loops through 10,000 different "experiments". An experiment consists of the following: randomly select N (in this case, 20) values from each time series, calculate the sample means, and store the values in a vector. When the loop is complete, you will have four vectors of 10,000 sample means and you can use this data to calculate the probability of getting a sample mean as extreme as -0.35 when $N = 20$. Repeat the process for the other values of N.

Problem 3: Application to climate data [30 pts.]

You will be plotting climate model output from the Community Earth System Model (CESM) Large Ensemble Project. The Large Ensemble Project includes a 42-member ensemble of fully coupled climate model simulations for the period 1920-2100 (note: only the original 30 are provided here). Each individual ensemble member is subject to the same radiative forcing scenario (historical up to 2005 and high greenhouse gas emission scenario (RCP8.5) thereafter), but begins from a slightly different initial atmospheric state (created by randomly perturbing temperatures at the level of round-off error). In the notebook, you will compare the ensemble remembers with a 2600-year-long model simulation (you have 1800 years) having constant pre-industrial (1850) radiative forcing conditions (perpetual 1850). By comparing the ensemble members to each other and to the 1850 control, you can assess the climate change in the presence of internal climate variability. More information on the CESM Large Ensemble Project can be found at: `http://www.cesm.ucar.edu/projects/community-projects/LENS/`

On the website, you will find two additional files called

- *TS_timeseries_cesmle_1850.nc*
- *TS_timeseries_cesmle_1920_2100.nc*

[2] `http://en.wikipedia.org/wiki/Central_limit_theorem`

[3] i.e. subtract their means and divide by their standard deviations, making $\overline{x} = 0$ and $\sigma_x = 1$.

The first is an 1850 control simulation of the NCAR CESM1 climate model. The second is the CESM large ensemble run from 1920-2100 under RCP8.5 conditions. Please use the variable called "gts_annual".

1. Use the 1850 control run to calculate population statistics (e.g. mean and variance) in the absence of climate change. Plot a histogram of the data. Do you think that this distribution is Gaussian?

2. Plot the time series from the control data as well as a time series of the global mean temperature from the first ensemble member.

3. Estimate present-day global warming in the first ensemble member by calculating the global mean temperature over 1990-2019 (30 years).

4. Under the assumption that there is no global warming, that is, the 1850 control run and the climate change simulations are drawn from the same climate, how odd is a 30-year mean temperature as the one you calculated in the step above?

   To explore this question, you will want to "resample" the control simulation over and over again in order to obtain the distribution of the sample means of length 30 years under the assumption of no global warming. Write code that grabs 30 year consecutive chunks and calculates their means, and stores these values in a vector. When the loop is complete, you will have a vector of ~ 59 sample means from the control run, and you can compare the distribution under the 1850 control climate to what you actually calculated from the climate change simulation.

5. It is always possible to draw a sample with a statistic that is at the extreme end. That is, even if global warming is not real, it is still possible to draw a single sample of a 30-year period with a very large temperature (even if very unlikely). Your analysis above only involved one ensemble member, and perhaps was just an unlikely fluke. Perform the same calculation in part 3 but now for the other 29 ensemble members. How much more confident are you that your null hypothesis that global warming is not real (i.e. all simulations are drawn from a control, unforced climate) can be rejected? Why? Try and be as quantitative as possible.[4]

[4]There are some simple statistics you can calculate with this - come see us in our office hours to discuss!