Michelle Maclennan

31 August 2020

**ATOC7500 – Application Lab #1**

**Significance Testing Using Bootstrapping and Z/T-tests**

**in class Monday August 31 and Wednesday September 2, 2020**

**Notebook #1 – Statistical significance using Bootstrapping**

**ATOC7500\_applicationlab1\_bootstrapping.ipynb**

**LEARNING GOALS:**

1) Use an ipython notebook to read in csv file, print variables, calculate basic statistics, do a bootstrap, make histogram plot

2) Hypothesis testing and statistical significance testing using bootstrapping

**DATA and UNDERLYING SCIENCE:**

In this notebook, you will analyze the relationship between Tropical Pacific Sea Surface Temperature (SST) anomalies and Colorado snowpack. Specifically, you will test the hypothesis that December Pacific SST anomalies driven by the El Nino Southern Oscillation affect the total wintertime snow accumulation at Loveland Pass, Colorado. When SSTs in the central Pacific are anomalously warm/cold, the position of the mid-latitude jet shifts and precipitation in the United States shifts. This notebook will guide you through an analysis to investigate the connections between December Nino3.4 SST anomalies (in units of °C) and the following April 1 Loveland Pass, Colorado Snow Water Equivalence (in units of inches). Note that SWE is a measure of the amount of water contained in the snowpack. To convert to snow depth, you multiply by ~5 (the exact value depends on the snow density).

The Loveland Pass SWE data are from:

<https://www.wcc.nrcs.usda.gov/snow/>

The Nino3.4 data are from:

<https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Nino34/>

**Questions to guide your analysis of Notebook #1:**

For full credit: write answers to the questions and then upload this document to your github along with notebook #1 (including any edits that you make).

1) Composite Loveland Pass, Colorado snowpack. Fill out the following table showing the April 1 SWE in all years, in El Nino years (conditioned on Nino3.4 being 1 degree C warmer than average), and in La Nina years (condition on Nino3.4 being 1 degree C cooler than average).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean SWE** | **Std. Dev. SWE** | **N (# years)** |
| **All years** | 16.33 in | 4.22 in | 81 |
| **El Nino Years** | 15.29 in | 4.0 in | 16 |
| **La Nina Years** | 17.78 in | 4.11 in | 15 |

2) Use hypothesis testing to assess if the differences in snowpack are statistically significant. Write the 5 steps. Test your hypothesis using bootstrapping.

5 steps of hypothesis testing:

1. State the significance level (α)
2. State the null hypothesis (H0) and alternative (H1)
3. State the statistic to be used, and the assumptions required to use it
4. State the critical region
5. Evaluate the statistic and state the conclusion

Instructions for bootstrap: Say there are N years with El Nino conditions. Instead of averaging the Loveland SWE in those N years, randomly grab N Loveland SWE values and take their average. Then do this again, and again, and again 1000 times. In the end you will end up with a distribution of SWE averages in the case of random sampling, i.e., the distribution you would expect if there was no physical relationship between Nino3.4 SST anomalies and Loveland Pass SWE.

1. Plot a histogram of this distribution and provide basic statistics describing this distribution( (mean, standard deviation, minimum, and maximum).

A screenshot of a cell phone

Description automatically generated

Mean: 16.30 in

Standard deviation: 1.04 in

Minimum: 13.43 in

Maximum: 20.49 in

1. Quantify the likelihood of getting your value of mean El Nino and La Nina SWE by chance alone using percentiles of this bootstrapped distribution. What is the probability that differences between the El Nino composite and all years occurred by chance? What is the probability that differences between the La Nina composite and all years occurred by chance?

The null hypothesis we are testing is that the difference between the El Nino/La Nina composites and the mean SWE for all years occurred by chance. Here, because we are not testing whether SWE was higher or lower, we need to use a two-tailed test. The probability that the differences between the El Nino composite and all years occurred by chance is 34.16%. The probability that the differences between the La Nina composite and all years occurred by chance is 15.48%.

3) Test the sensitivity of the results obtained in 2) by changing the number of bootstraps, the statistical significance level, or the definition of El Nino/La Nina (e.g., change the temperature threshold so that El Nino is defined using a 0.5 degree C temperature anomaly or a 3 degree C temperature anomaly). In other words – play and learn something about the robustness of your conclusions.

If we decrease the number of bootstraps from 1000 to 100, the distribution of mean SWE no longer looks normally distributed due to the smaller number of sample means:

A screenshot of a cell phone

Description automatically generated

Here, the probability that the difference between mean and El Nino SWE occurred by chance is 35.4% (a very slight increase from the previous test of 1000 bootstraps), and for La Nina it is 26.15% (a large increase in probability). However, with such few sample means from which to draw our probabilities, this seems like a less-robust way to test. I would expect that the distribution becomes more robust around 1000 bootstraps, and that it’s probably unnecessary to increase the number beyond 3000-4000, as it likely converges on the same probabilities as the sample means approach a standard normal distribution. For example, if we perform 3500 bootstraps then we get the following distribution:

A screenshot of a cell phone

Description automatically generated

Corresponding to this distribution, the probability that the difference in mean and El Nino SWE occurred by chance is 30.74%, and for La Nina it is 17.2% - quite similar to the original probabilities for 1000 bootstraps.

4) Maybe you want to see if you get the same answer when you use a t-test… Maybe you want to set up the bootstrap in another way?? Another bootstrapping approach is provided by Vineel Yettella (ATOC Ph.D. 2018). Check these out and see what you find!!

The t-test for an ENSO anomaly threshold of 1 deg C yields the result that we cannot reject the null hypothesis that ENSO years have the same SWE mean as the full record. Even if we increase the temperature anomaly threshold used to define ENSO events to 3.0 deg C, or decrease it to 0.5 deg C, it is always the case that the null hypothesis cannot be rejected.

For Vineel’s method, the confidence interval for the difference in means with a significance level of alpha = 0.05 is -1.03 to 3.15 deg C. Because it contains zero, we cannot reject the null hypothesis. If we increase alpha to 0.35, the confidence interval becomes 0.03-2.03 deg C, and because 0 is no longer included in this interval we can reject the null hypothesis. That said, alpha = 0.35 is not a strong significance level so just because we can reject the null hypothesis doesn’t mean we should trust that result.

**Notebook #2 – Statistical significance using z/t-tests**

**ATOC7500\_applicationlab1\_ztest\_ttest.ipynb**

**LEARNING GOALS:**

1) Use an ipython notebook to read in a netcdf file, make line plots and histograms, and calculate statistics

2) Calculate statistical significance of the changes in a normalized mean using a z-statistic and a t-statistic

3) Calculate confidence intervals for model-projected global warming using z-statistic and t-statistic.

**DATA and UNDERLYING SCIENCE:**

You will be plotting *munged* 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 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/>

**Questions to guide your analysis of Notebook #2:**

For full credit: write answers to the questions and then upload this document to your github along with notebook #2 (including any edits that you make).

1) Use the 2600-year long 1850 control run to calculate population statistics with constant forcing (in the absence of climate change). Find the population mean and population standard deviation for CESM1 global annual mean surface temperature. Normalize the data and again find the population mean and population standard deviation. Plot a histogram of the normalized data. Is the distribution Gaussian?

The population mean for CESM1 global annual mean surface temperature is 287.11 K. The population standard deviation is 0.1 K. When the data are normalized, the mean becomes 0 and the standard deviation is 1.0, and the distribution looks Gaussian.

A close up of a logo

Description automatically generated

2) Calculate global warming in the first ensemble member over a given time period defined by the startyear and endyear variables. Compare the warming in this first ensemble member with the 1850 control run statistics and assess if the warming is statistically significant. Use hypothesis testing and state the 5 steps. What is your null hypothesis? Try using a z-statistic (appropriate for N>30) and a t-statistic (appropriate for N<30). What is the probability that the warming in the first ensemble member occurred by chance? Change the startyear and endyear variables – When does global warming become statistically significant in the first ensemble member?

1. We’re going to use a significance level of alpha = 0.05, meaning we have 95% confidence intervals or there is a 5% chance of a Type I error.
2. The null hypothesis H0 is that there is no significant difference between temperatures in the first ensemble and the 1850 control run.
3. The appropriate statistic to use is the t-statistic, since our number of years N=10 < 30. Is this case, we must assume that we know the underlying distribution (we assume it to be Gaussian). We also need the standard deviation of the sample.
4. The critical region is t > tc = 1.833, which corresponds to the 95% confidence interval for N-1 = 9 degrees of freedom in the one-tailed t-test. We can use a one-tailed t-test because we have a priori information, in that we expect global warming to have caused global mean temperatures to increase.
5. The probability that the warming in the first ensemble member occurred by chance is 0% when comparing the decade 2020-2030 to the 1850 control run. When we change the star year to 1930 and the end year to 1940, the probability that the warming occurred by chance increases to 53.87%. Global warming becomes statistically significant in the first ensemble member when the t-statistic is greater than 1.833, which first occurs in the decade 1980-1990.

Note that if you erroneously use the z-statistic for this hypothesis test, with a significance level of alpha = 0.05 and the null hypothesis that there is no significant difference between the temperatures, the critical value for the one-tailed z-test is 1.96, which is reached in the decade 1980-1990 as well with a probability of 0%.

3) Many climate modeling centers run only a handful of ensemble members for climate change projections. Given that the CESM Large Ensemble has lots of members, you can calculate the warming over the 21st century and place confidence intervals in that warming by assessing the spread across ensemble members. Calculate confidence intervals using both a z-statistic and a t-statistic. How different are they? Plot a histogram of global warming in the ensemble members – Is a normal distribution a good approximation? Re-do your confidence interval analysis by assuming that you only had 6 ensemble members or 3 ensemble members. How many members do you need? Look at the difference between a 95% confidence interval and a 99% confidence interval.

The confidence intervals for the t-statistic are very similar to the confidence intervals for the z-statistic. The t-statistic yields 3.61-3.66 deg C warming in the 95% confidence interval, and 3.60-3.70 deg C warming in the 99% confidence interval. The z-statistic yields 3.61-3.66 deg C warming in the 95% confidence interval, and 3.60-3.66 deg C warming in the 99% confidence interval.

A picture containing drawing

Description automatically generated

The histogram of global warming in the ensemble members has a Gaussian looking distribution in appearance, although it is slightly skewed to the left.

Assuming you only has six ensemble members, the confidence intervals in the t-statistic increase slightly. Now the 95% confidence interval has a range of 3.60-3.68 deg C, and the 99% confidence interval has a range of 3.51-3.77 deg C. This pattern continues if you decrease the number of ensemble members to three: the 95% confidence interval expands to 3.59-3.74 deg C, and the 99% confidence interval expands to 3.58-3.74 deg C.