ATOC5860 Presentations - Spring 2023 - FIRST COME FIRST SERVED

3/23/2023 - Homework #2, #3 (Regression and EOFs)

1	Genevieve Clow
2	Ethan Murray
3	
4	Clairy Reiher
5	
6	
7	
8	

4/26/2022 - Homework #4, #5, #6 (Timeseries analysis - power spectra, filtering, machine learning)

1	Katey
2	
3	Holly
4	
5	Yu-Wen
6	
7	
8	

Have you signed up for your date to present homework?? https://bit.ly/3FcVFkI

Today's plan…. Barnes 1.6-1.8

1) Compositing

2) Wilks Field Test – Motivating the importance of independent N

3) Other Common Distributions

- Chi-square distribution (test of variance – population vs. sample)
- F-statistic (test of variance – sample vs. sample)
- Binomial
- Normal approximation to Binomial

4) Non-parametric tests

- Kolmogorov-Smirnov Test (test if sample distribution matches population distribution)

Note: All code in lecture3 folder. Also - Check out Statistical_Tables.pdf – You might find the printed values useful for checking python functions. I sure do!! ☺

Compositing – What is it? Advantages/Disadvantages

Compositing, also sometimes called *superposed epoch analysis* when applied to time series, is one of the simplest analysis techniques imaginable. It is very powerful, but can also be misused.

Compositing: sorting data into categories and comparing the statistics for different categories.

The idea is by averaging the data in a smart way, you can isolate the signal and remove the background "noise" (unwanted/not interesting to you - signals)

Advantages over regression (fitting a line): can isolate nonlinear relationships, don't need to make assumptions about the underlying distributions

Disadvantages: does not use all of the data (more susceptible to sampling errors), tends to focus on "extremes" of each category

Compositing – Steps

1. Determine categories

   – for diurnal cycle, use time of day

   – impacts of ENSO, warm/cold SSTs

   – impacts of sea ice loss, high and low September sea ice years

   You should have an a priori hypothesis as to why the variable being composited should depend on the category

2. compute the statistics for each category

3. display results

4. validate results

   – calculate relevant statistics (most often z-test using a comparison of means), if N is big enough (i.e. enough data in each category to use the Central Limit Theorem)

   – OR, perform Monte Carlo or sub-sampling techniques

   – subdivide the data and show relationship exists in sub-samples of the data

Compositing Example: ATOC5860_applicationlab1_bootstrapping.ipynb

1. **Determine categories:** El Nino year SWE vs. All years SWE
2. **Compute Statistics:** for each category

```
### Calculate the average snowfall on April 1 at Berthoud Pass, Colorado
SWE_avg=data['BerthoudPass_April1SWE_inches'].mean()
SWE_std=data['BerthoudPass_April1SWE_inches'].std()
N_SWE=len(data.BerthoudPass_April1SWE_inches)
print(f'Average SWE (inches): {np.round(SWE_avg,2)}')
print(f'Standard Deviation SWE (inches): {np.round(SWE_std,2)}')
print(f'N: {np.round(N_SWE,2)}')
```

```
Average SWE (inches): 16.24
Standard Deviation SWE (inches): 3.43
N: 87
```

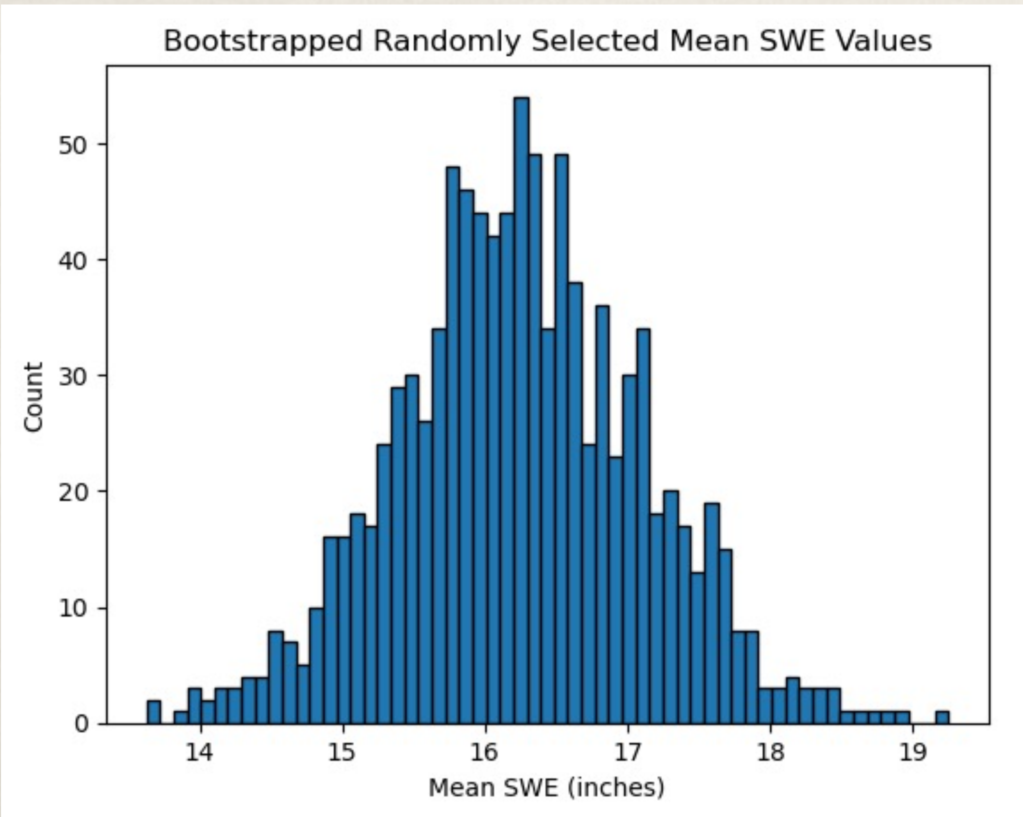
```
### Calculate the average SWE when it was an el nino year
SWE_avg_nino=data[data.Nino34_anomaly_prevDec>1.0]['BerthoudPass_April1SWE_inches'].mean()
SWE_std_nino=data[data.Nino34_anomaly_prevDec>1.0]['BerthoudPass_April1SWE_inches'].std()
N_SWE_nino=len(data[data.Nino34_anomaly_prevDec>1.0].BerthoudPass_April1SWE_inches)
print(f'Average SWE El Nino (inches): {np.round(SWE_avg_nino,2)}')
print(f'Standard Deviation SWE El Nino (inches): {np.round(SWE_std_nino,2)}')
print(f'N El Nino: {np.round(N_SWE_nino,2)}')
```

```
Average SWE El Nino (inches): 16.44
Standard Deviation SWE El Nino (inches): 3.29
N El Nino: 16
```

Compositing Example: ATOC5860_applicationlab1_bootstrapping.ipynb

3. Assess significance of differences using hypothesis testing

1) Confidence Level: 95%
2) H0 (null hypothesis): El Nino years mean SWE (16.44 inches) =All years mean SWE (16.24 inches)
3) Statistics to be used: bootstrap to obtain many randomly selected SWE means with N=16, Assess probability of obtaining 16.44 inches by chance. Assumptions: no distribution assumption required to bootstrap. Note: central limit theorem applies. I expect the distribution of means will be normally distributed.



Bootstrapped Distribution Statistics:

N=1000
Mean=16.24
Std=0.87
Min=13.63
Max=19.26

Compositing Example: ATOC5860_applicationlab1_bootstrapping.ipynb

3. Assess significance of differences using hypothesis testing (continued)

4) Critical region – probability that this happened by chance has to be less than 5%.
5) Calculate probability that snowfall during El Nino of 15.2 was less than non-ENSO occurred by chance?

```
## What is the probability that the snowfall El Nino mean differs from the mean by chance?
## Using Barnes equation (83) on page 15 to calculate probability using z-statistic
sample_mean=SWE_avg_nino
sample_N=1
population_mean=np.mean(P_Bootstrap_mean)
population_std=np.std(P_Bootstrap_mean)
xstd=population_std/np.sqrt(sample_N)
z_nino=(sample_mean-population_mean)/xstd

print(f'sample_mean - El Nino: {np.round(sample_mean,2)}')
print(f'population_mean: {np.round(population_mean,2)}')
print(f'population_std: {np.round(population_std,2)}')
print(f'Z-statistic (# standard errors that the sample mean deviates from the population mean: {np.round(z_nino,2)}')

prob=(1-stats.norm.cdf(np.abs(z_nino)))*2*100 ##this is a two-sided test
print(f'Probability happened by chance, two-tailed test (percent): {np.round(prob,0)}%')
```

```
sample_mean - El Nino: 16.44
population_mean: 16.24
population_std: 0.87
Z-statistic (# standard errors that the sample mean deviates from the population mean: 0.23
Probability happened by chance, two-tailed test (percent): 82.0%
```

Result: Cannot Reject H0 (null hypothesis)
H0: El Nino years mean SWE = All years mean SWE

Compositing Example:
ATOC5860_applicationlab1_bootstrapping.ipynb

We tested the Null Hypothesis:
El Nino years mean SWE = All years mean SWE

Another equally valid Null Hypothesis to test:
El Nino years mean SWE > All years mean SWE

One-tailed or two-tailed? If null hypothesis includes the sign – a one tailed should be applied.

If my null hypothesis doesn't include the sign – a two tailed test should be applied.

Compositing – Cautions:

If your underlying distribution isn't normal and your N is small you may not have enough data to composite or bootstrap!

Be also careful of data that are spatially and temporally correlated, which reduces your effective N. ***(More on this in unit 2…)***

"THE STIPPLING SHOWS STATISTICALLY SIGNIFICANT GRID POINTS"

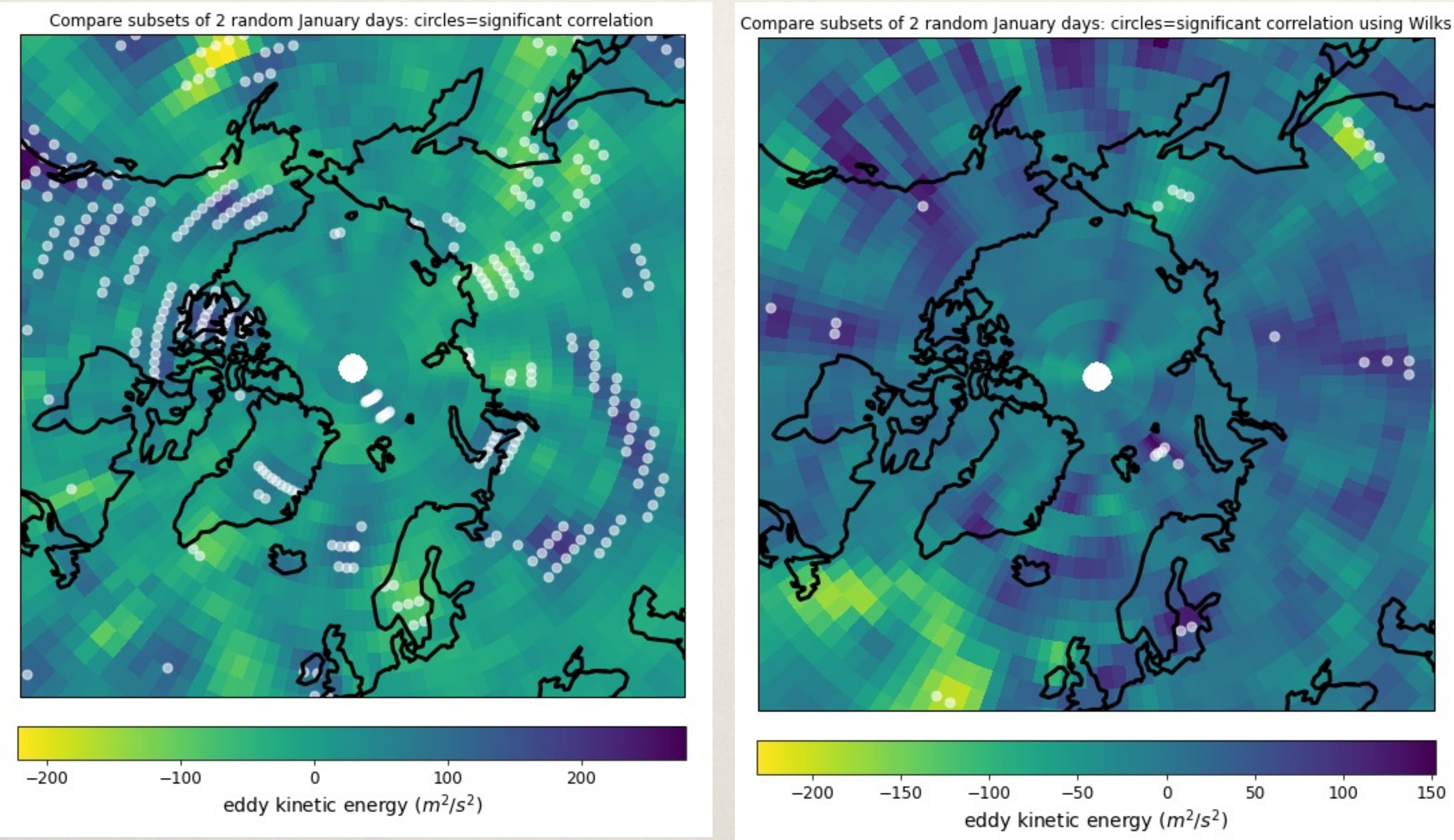
How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It

BY D. S. WILKS

Controlling the false discovery rate provides a computationally straightforward approach to interpretation of multiple hypothesis tests.

Wilks 2016 BAMS Available on Google Drive Under Lecture 3 (take a look!)

SHOW WILKS EXAMPLE…..



Try the python code – wilks_field_test_cartopy.ipynb

Other Common Distributions: Chi-square Distribution (chi2!)

Use chi2 to compare a sample variance to a population with known variance…

Before we get to the distribution, let's define a random variable χ^2:

$$\chi^2 = (N-1)\frac{s^2}{\sigma^2} \tag{118}$$

This quantity can be used to test if the sample variance s^2 is different from the population variance σ^2. Note we are using a ratio, rather than a difference, since a difference ends up being very complicated.

If the underlying distribution from which we draw N values to compute χ^2 is a normally distributed population with standard deviation σ, then the χ^2 values will be distributed as follows:

$$f(\chi^2) = f_0(\nu)(\chi^2)^{(\frac{1}{2}\nu-1)} \exp^{-\frac{1}{2}\chi^2}, \tag{119}$$

where f_0 is a normalization factor. This is the *Chi-square distribution*. The Chi-square distribution is a member of the Gamma family of continuous probability functions.

What chi2 can do for you – significance testing and confidence intervals for variance…

The Chi-square distribution can be used to estimate the significance of the ratio $\frac{s^2}{\sigma^2}$.

If you are trying to determine the "true" variance, you can move things around to get the confidence limits for the true variance given your sample variance. That is,:

$$\frac{s^2(N-1)}{\chi^2_{0.975}} \leqslant \sigma^2 \leqslant \frac{s^2(N-1)}{\chi^2_{0.025}} \tag{120}$$

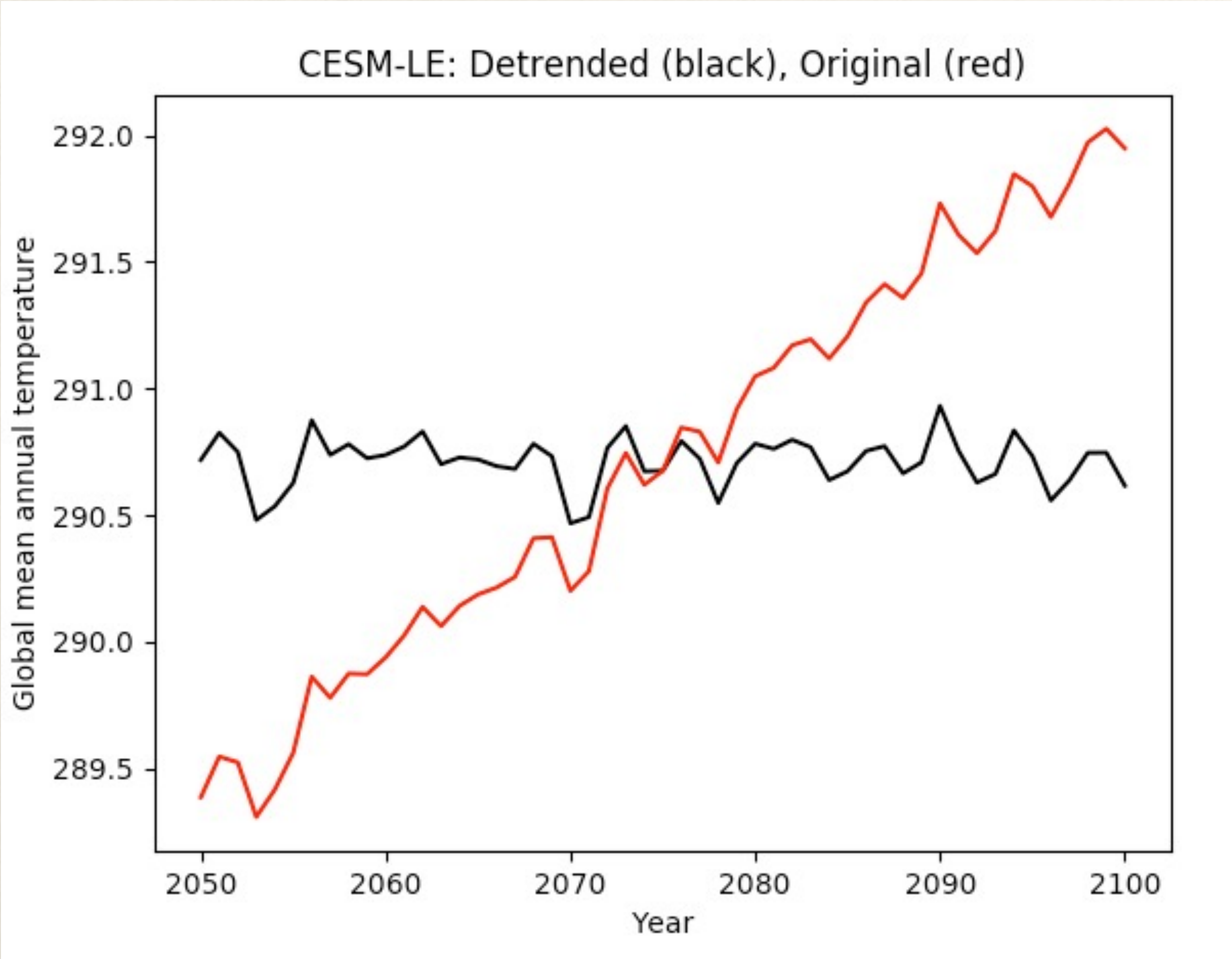
Notes on the chi2 distribution: tests of variance

- χ^2 is used to assess the confidence limits on σ
- like the t-distribution, the Chi-square distribution is a function of ν
- the Chi-square distribution is not symmetric about its mean: $\chi^2_{0.025} \neq \chi^2_{0.975}$
- for $\nu > 30$, the Chi-square distribution approaches the Normal distribution

Example to compare variances using chi-squared:

Does the global annual mean temperature variance in member 1 2050-2099 differ from the 1850 pre-industrial control?

Why should we detrend the member 1 CESM-LE data?



Try the python code – chisquared.ipynb

Other Common Distributions
F-statistic: tests of variance when comparing two sample variances that both have small N.

The F-statistic is used to assess the ratio between two sample standard deviations s_1 and s_2 - whereas the Chi-square is used when you have one sample and you are comparing it to a known population variance σ. In other words, the Chi-square only has a small sample N_1, while the F-statistic has a small sample N_1 and N_2.

If s_1 and s_2 are the variances of independent random samples of size N_1 and N_2, taken from two Normal populations having true variances σ_1^2 and σ_2^2, respectively, then we can define a random variable F as:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{121}$$

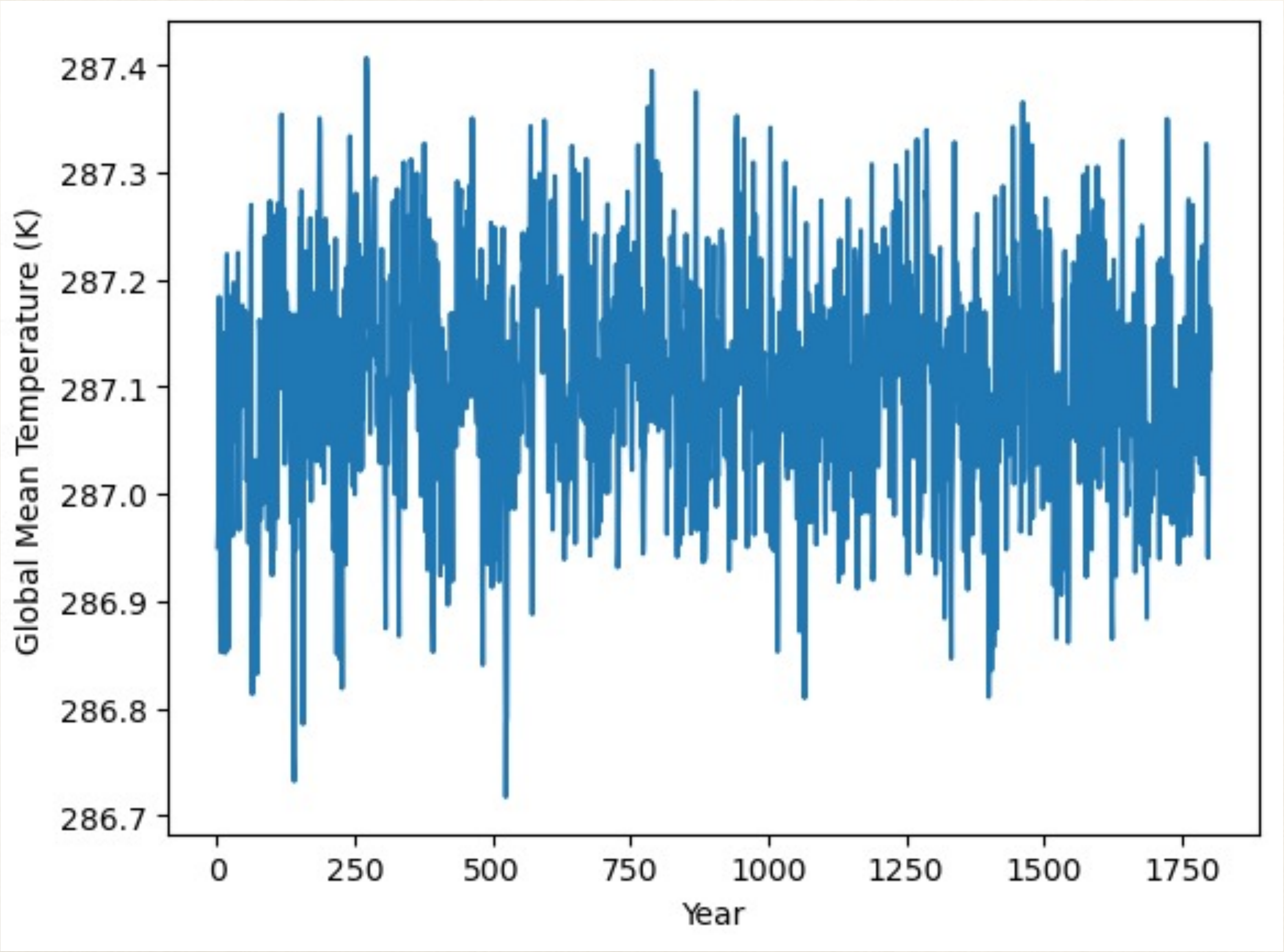
F is a random variable that follows an F-distribution with parameters $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$. Note that the F-statistic is the ratio of two chi-squared random variables (scaled appropriately due to the N-1).

Note that if both s_1 and s_2 come from Normal populations with the same variances, i.e. $\sigma_1 = \sigma_2$ then the F-statistics simplifies to:

$$F = \frac{s_1^2}{s_2^2} \tag{122}$$

Note: F-statistic will be used extensively when testing the significance of peaks in frequency spectra

Example using F statistic:
Does the variance in global mean temperature differ?



Try the python code – ftest.ipynb

WORK IN BREAKOUT ROOMS TO EXAMINE SAMPLE CODE TO TEST DIFFERENCES IN VARIANCE….

Try the python code:
1) ftest.ipynb
2) chisquared.ipynb

If there is time – Check out:
wilks_field_test_cartopy.ipynb

Other Common Distributions: Binomial Distribution

The binomial distribution is often one of the first taught in college statistics. Suppose you have a set of N trials where the outcome of each trial is either a "success" or "failure". These trials are called "Bernoulli trials".

The probability of a success in one trial is $p = \boldsymbol{Pr}(\mathsf{success})$. If X is the total number of successes in N trials, then

$$\boldsymbol{Pr}(\mathsf{X} = \mathsf{k}) = \binom{\mathsf{N}}{\mathsf{k}} \mathsf{p}^{\mathsf{k}} (1 - \mathsf{p})^{n-k} \tag{123}$$

Recall that

$$\binom{\mathsf{N}}{\mathsf{k}} = \frac{\mathsf{N}!}{\mathsf{k}!(\mathsf{N} - \mathsf{k})!} \tag{124}$$

Note that the right-hand-side looks complicated, but really is the probability of a certain number of successes times the probability of a certain number of failures, with a special factor in-front since the order of successes and failures does not matter.

Example From Barnes Notes – Binomial Distribution

1.7.4.2 Binomial Distribution II:

What is the probability of getting a score of 12 or more on a test of 30 true-false questions?

Let's define a success as getting the question right.

$$\begin{aligned} \boldsymbol{Pr}(\text{success}) &= 1/2 && (130) \\ \boldsymbol{Pr}(\text{failure}) &= 1 - \boldsymbol{Pr}(\text{success}) = 1/2 && (131) \end{aligned}$$

Now, the problem is a bit more tedious, because we have to add up the probabilities for $k = 12, 13, 14 \ldots 29, 30$. Instead, we can calculate the probability of getting $k = 0, 1, 2, 3 \ldots 11$ questions right, and then 1 minus this probability is the probability we want. Thus,

$$\begin{aligned} \boldsymbol{Pr}(X \geqslant 12) &= 1 - \boldsymbol{Pr}(X < 12) && (132) \\ &= 1 - \sum_{k=0}^{11} \binom{30}{k} (0.5)^k (0.5)^{30-k} && (133) \\ &= 1 - 0.1 = 90\% && (134) \end{aligned}$$

Other Common Distributions
Normal Approximation to Binomial Distribution

You can probably imagine the problems getting much more tedious as you have to sum more and more. However, there is a normal approximation to the Binomial distribution in the case of

1. large N
2. $Np \geqslant 10$ Number of trials (N) times the probability of success (rho) is greater than 10
3. $N(1-p) \geqslant 10$ Number of trials (N) times the probability of failure (1-rho) is greater than 10

In this case, the statistic

$$z = \frac{X - Np}{\sqrt{Np(1-p)}} \tag{125}$$

follows a Normal distribution with $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$.

An approximate two-tailed 95% confidence interval for the number of successes X is then,

$$Np - 1.96 \cdot \sqrt{Np(1-p)} \leqslant X \leqslant Np + 1.96 \cdot \sqrt{Np(1-p)} \tag{126}$$

Example – Normal Approximation to Binomial

48 CMIP5 models are discussed in the IPCC 5th Assessment Report. How many models must agree that global temperatures will increase by 2100 so that we can say with 95% certainty that the models do not agree purely by chance? What is the 95% confidence interval on the number of models with increasing temperatures under the null hypothesis?

Here, let a success be that the model says global temperatures will increase. Our null hypothesis is that the models randomly guess whether global temperatures will increase - thus, there is a 50% chance that any one model will predict a temperature increase ($p = 0.5$). We want to know k^* such that:

$$\boldsymbol{Pr}(X \geqslant k^* | H_0) \leqslant 0.05 \tag{135}$$

That is, k^* is the number of models that must show a temperature increase for us to believe it is more than chance (that the null hypothesis can be rejected).

$$\sum_{k=k^*}^{48} \binom{48}{k} (0.5)^k (1-0.5)^{48-k} \leqslant 0.05 \tag{136}$$

This would take a long time by hand, however, let's check if the Normal approximation to the Binomial applies.

1. N is large ($N = 48$) ✓
2. $Np = 24 \geqslant 10$ ✓
3. $N(1-p) = 24 \geqslant 10$ ✓

Note: Here N=48, p=0.5

Example – Normal Approximation to Binomial

48 CMIP5 models are discussed in the IPCC 5th Assessment Report. How many models must agree that global temperatures will increase by 2100 so that we can say with 95% certainty that the models do not agree purely by chance? What is the 95% confidence interval on the number of models with increasing temperatures under the null hypothesis?

So, we can use the Normal approximation for large N (here we choose a one-sided test):

$$\boldsymbol{Pr}\left(Z > \frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1 - 0.5)}}\right) = 0.05 \quad (137)$$

$$\frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1 - 0.5)}} = 1.645 \quad (138)$$

$$k^* \geqslant 30. \quad (139)$$

So, at least 30 models must show increasing temperatures to reject the null hypothesis that the warming is due to random chance. As expected, more than half of the models must show an increase.

The 95% confidence interval under the null hypothesis is:

$$Np \pm 1.96 \cdot \sqrt{Np(1-p)} \quad (140)$$

$$24 \pm 1.96 \cdot \sqrt{24(1-.5)} = 24 \pm 6.8 \quad (141)$$

$$17.2 \leqslant X \leqslant 30.8 \quad (142)$$

Note: N=48, p=0.5

Try the python code – binomial_example_cmip5.ipynb

Non-parametric Tests

Thus far, most of the statistics we have used assume an underlying distribution (typically normal). However, there may be instances where you do not think that the data is normally distributed - in which case, you might want to use a non-parametric statistical test.

Non-parametric Tests – Kolmogorov-Smirnov Test ("KS test") (we'll work through this one together)

Other Non-parametric Tests (read on your own if interested):

1) Signs Test (note: good example in Hartman Chapter 1)

2) Runs Test (see python code runs_test.ipynb for an example)

Non-parametric Tests – Kolmogorov-Smirnov Test ("KS test")

The Kolmogorov-Smirnov test (or "KS test") tests the equality of two continuous, one-dimensional probability distributions. The most standard version tests whether a particular sample distribution is the same as a specific reference distribution.

A few important notes about the ks-test:

- this is a non-parametric test, so you do not need to know what the true distribution of your data is or test against some null hypothesis of a particular distribution
- the test is sensitive to both *location* and *shape* and cannot tell you why the distributions are different (e.g. is the sample shifted compared to the reference? is the sample distribution wider than the reference?)

Non-parametric Tests – Kolmogorov-Smirnov Test ("KS test")

The test works by comparing the CDFs (cumulative density functions) of the sample of length N and reference distributions. Specifically, the difference between the two CDFs is computed, and the *maximum difference*, denoted as D, is used as the test statistic. The null hypothesis is rejected at the significance level α if

$$\sqrt{\mathsf{N}}\mathsf{D} > \mathsf{K}_\alpha \tag{160}$$

where K_α is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant \mathsf{K}_\alpha) = 1 - \alpha \tag{161}$$

and the probability density function of K is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant \mathsf{x}) = 1 - 2\sum_{\mathsf{j}=1}^{\infty}(-1)^{\mathsf{j}-1}e^{-2\mathsf{j}^2\mathsf{x}^2} \tag{162}$$

Let's do an example with python code – ks_test_onesample.ipynb

Example to test a sample distribution is normal using a KS test:

```
### Compare 1 sample with the TRUE normal
### Test the null hypothesis that the two distributions are identical

## Let's say you 1 sample with N observations having nu degrees of freedom
## (generate this sample below drawing randomly from a t-distribution)
nu = 5 ## nu=degrees of freedom
N = 1000  ## number of random variates
x2 = stats.t.rvs(nu,size=N) ## generate N random numbers with nu degrees of freedom
##print(np.shape(x2))
```

```
## Is your your sample normally distributed?  Use a KS test!
## Test the null hypothesis that your sample distribution and the normal distribution are identical.
D, p = stats.kstest(x2,'norm') ## https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html
print('KS test statistic:',np.round(D,3))
print('p-value (must be less than 0.05 to reject the null hypothesis at the 95% confidence level):',np.round(p,3))

x = np.arange(-10.,10.,.1)
z = stats.norm.pdf(x,0,1)
plt.plot(x,z,color = 'black', label = 'Standard Normal: theoretical', linewidth = 5)
hist, bin_edges = np.histogram(x2,x, density=True)
plt.plot(bin_edges[:-1],hist,color = 'red', label = 't-distribution, nu = ' + str(nu), linewidth = 1.75)
plt.legend(fontsize = 12)
if(p<0.05):
    plt.title('Reject the null. p = ' + str(round(p, 3)))
else:
    plt.title('Cannot reject the null. p = ' + str(round(p, 3)))
```

Try the python code – ks_test_onesample.ipynb

WORK IN GROUPS TO EXAMINE SAMPLE CODE….

Try the python code:
1) binomial_example_cmip5.ipynb
2) ks_test_onesample.ipynb

If there is time – Check out:
runs_test.ipynb

Next Tuesday:

1) Homework #1 due. Please write up your answers in a text document and include all relevant figures/data. Provide code as a separate attachment. More instructions are provided on slack.
Please name your files using the name convention at the top of the instructions:
"ATOC5860_HW1_LastName.pdf, .html, .ipynb".

2) We start Unit #2 – Regression.

Slack me if you have questions ☺