EPS 659a — Time Series in Geoscience

Problem Set Five

due Friday, October 8, 2021

Problem 1: Statistical tests of data sets: t-tests for global warming

On the Canvas site you will find subdirectory `Probset5` with datasets that contain time series of monthly temperature anomalies from 1850 into 2021. From the UK Hadley Center temperature dataset, circa July 2021. Consider the global-and hemisphere-average monthly temperature anomalies, that is, files `Hadcrut_GlobalAverage.csv, Hadcrut_NHAverage.csv, Hadcrut_SHAverage.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. These are explained in the file `Notes_AvgData.txt`. You will be using the R command for the t-test for rejecting the hypothesis that the means of two datasets are equal. The R command `t.test` is described at

`https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test`.

You will be comparing data sets composed of each decade of the Hadley Center times series between 1850 and 2020. In the problems that follow, use R commands like `mean(SHtemps_1850s)` to compute the means and generate a time axis with

```
decades <- c(1855:2015:10)
```

(a) For the first and last 20th-century decades from the global-average data (1900s and 1990s), compute and report the sample means and the sample standard deviation of the data about its sample mean. Use the one-sided t-test to estimate the probability that each decadal mean of the temperature anomalies is distinct from zero. (The one-way t-test applies if the ”zero” of the time series is taken as known, not estimated from the data.) The formula for the one-way t-test is a combination of the sample mean $\bar{X}_N$ and the sample standard deviation $\hat{s}_N$ of each decade, meant to mimic a zero-mean Gaussian distribution with unit variance:

$$t = \frac{\sqrt{(N-1)}(\bar{X}_N - \mu)}{\hat{s}_N}$$

where μ is the expected mean. What is the value of t for this problem? Note that this estimate assumes that the underlying statistics of the null hypothesis are Gaussian normal. In this problem we take the expected decadal mean to be 0, and the degrees of freedom in the t-test is $(N-1)$.

(b) For the first and last 20th-century decades from the global-average data (1900s and 1990s), use the two-sided t-test to estimate the probability that the decadal means are distinct. (The two-way t-test applies if there is no fixed reference value for the mean, only two means estimated from the data.) The formula for the two-way t-test is a combination of the sample means $\bar{X}_N$ and $\bar{Y}_M$ and the sample variances $\hat{s^2}_N^{(X)}, \hat{s^2}_M^{(Y)}$ for the two decades, meant to mimic a zero-mean Gaussian distribution with unit variance:

$$t = \frac{(\bar{X}_N - \bar{Y}_M)}{\sqrt{\left(\frac{N\hat{s^2}^{(X)}_N + M\hat{s^2}^{(Y)}_M}{N+M-2}\right)\left(\frac{1}{N}+\frac{1}{M}\right)}} = \frac{\sqrt{(N-1)}(\bar{X}_N - \bar{Y}_N)}{\sqrt{\hat{s^2}^{(X)}_N + \hat{s^2}^{(Y)}_N}}$$

for the case where $N = M$. What is the value of t for this problem? Note that this estimate assumes that the underlying statistics of the null hypothesis are Gaussian normal, and the degrees of freedom in the t-test is $2(N-1)$.

(c) For each pair of adjacent decades (1850s/1860s, 1860s/1870s, ...2000s/2010s) of the global-average data series, use the two-way t-test to estimate the probability of a Type 1 error when rejecting the hypothesis that the anomaly-means of adjacent decades are identical. Plot these decadal-comparison probabilities through time, using a log probability to avoid hugging the plot axis. Change the `"alternative"` parameter in the R-language function `t.test` from `"two-sided"` to `"greater"` and `"less`, successively. For how many decadal pairs do successive decades increase average temperature at 99% confidence? For how many decadal pairs do successive decades decrease average temperature at 99% confidence? Based on this analysis, when did Global Warming begin, in your opinion?

(d) For each pair of adjacent decades (1850s/1860s, 1860s/1870s, ...2000s/2010s) of the Northern-Hemisphere data series, use the two-way t-test to estimate the probability of a Type 1 error when rejecting the hypothesis that the anomaly-means of adjacent decades are identical. Plot these decadal-comparison probabilities through time, using a log probability to avoid hugging the plot axis. Change the `"alternative"` parameter in the R-language function `t.test` from `"two-sided"` to `"greater"` and `"less`, successively. For how many decadal pairs do successive decades increase average temperature at 99% confidence? For how many decadal pairs do successive decades decrease average temperature at 99% confidence? Based on this analysis, when did Global Warming begin in the Northern Hemisphere, in your opinion?

(e) For each pair of adjacent decades (1850s/1860s, 1860s/1870s, ...2000s/2010s) of the Southern-Hemisphere data series, use the two-way t-test to estimate the probability of a Type 1 error when rejecting the hypothesis that the anomaly-means of adjacent decades are identical. Plot these decadal-comparison probabilities through time, using a log probability to avoid hugging the plot axis. Change the `"alternative"` parameter in the R-language function `t.test` from `"two-sided"` to `"greater"` and `"less`, successively. For how many decadal pairs do successive decades increase average temperature at 99% confidence? For how many decadal pairs do successive decades decrease average temperature at 99% confidence? Based on this analysis, when did Global Warming begin in the Southern Hemisphere, in your opinion?

Problem 2: Statistical tests of data sets: ANOVA-tests for global warming

The ANOVA technique is another way to determine if particular subsets of a dataset are behaving differently relative to the variance of all the data taken at once. The R language has some tools for computing ANOVA, but the principles of the technique are best understood by working through an example step by step. In this example, we would like to know if the decadal sample means of the Hadley Center global-average temperature anomalies deviate nonrandomly from the grand mean of all the temperature anomaly data. One compares the data variance between subsets of the larger data set to the data variances within each data subset. The simplest ANOVA test takes the ratio of "between" and "within" variances, per degree of freedom, and compares this ratio to the CDF of the F distribution with appropriate degrees of freedom.

For the Hadley Center temperature-anomaly time series, we can examine whether the variance among the decadal averages is larger than the variance of the temperature anomalies $\{\tau_i\}$ within each decade. Clearly, if the data had properties that were invariant over time, the decadal sample means would still differ from each other and from the grand-mean of the dataset, due to random fluctuations in Earth's climate system. The ANOVA analysis determines whether the differences among the decadal means are too large to arise from the null hypothesis of invariant temperature statistics. If we partition the 1850-2019 time period into $N = 17$ decades, we can compute sample mean $\bar{\tau}_j$ for the jth decade, for $j = 1, \ldots N$ and also the "grand mean" $\bar{\tau}_0$ for the total data set. The variances of the sample means about the grand mean can be summed to form an average variance among the subsets

$$Y_b = \frac{1}{N-1}\sum_{j=1}^{N}(\bar{\tau}_j - \bar{\tau}_0)^2$$

where the degrees-of-freedom (dof) for the average is $N - 1$, owning to the dof allocated to the grand mean of all data. Y_b is the average "between" variance for $N - 1$ degrees of freedom.

Let $\tau_{i,j}$ be the ith temperature anomaly in the jth decade. The variance of temperature anomalies $\tau_{i,j}$ within the jth decade about its sample average $\bar{\tau}_j$ is associated with $M = 120$ independent data and therefore with $M - 1$ degrees of freedom

$$Y_w = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{M-1}\sum_{i=1}^{M}(\tau_{i,j} - \bar{\tau}_j)^2$$

We average this "within" variance over the N decades, so that Y_w is the average variance among $N(M - 1)$ dof. This means that we should evaluate the variance ratio

$$F = Y_b/Y_w$$

with percentiles of the F distribution with $N - 1$ and $N(M - 1)$ degrees of freedom. Reference the F distribution in R at

`https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Fdist.html`

(a) Plot the PDF of the F variance-ratio distribution for three cases using the R function `df()` with non-centrality parameter equal to zero. Plot the PDF for $F \in [0, 10]$ for $(df1, df2) = (2, 8)$ degrees of freedom, for $(df1, df2) = (10, 20)$ degrees of freedom, and $(df1, df2) = (17, 2000)$ degrees of freedom. Plot the F PDFs on a linear-linear plot, then on a linear-log plot – the R functions offer the option of returning logarithmic values of the PDF.

(b) Evaluate the ANOVA decadal means and variances for the 1850-2019 interval of the Hadley Center global-average temperature anomalies from Problems 1. Evaluate the F variance ratio $F = Y_b/Y_w$ for the "between" and "within" variances and use the R function for the F CDF (R-language function `pf()`) to estimate the statistical probability for nonrandomness. The total variance about the sample mean of a decade of temperature anomalies is a multiple of the R-language `var()` command:

```
nn <- length(data)

variance_total <- (nn-1)*var(data)
```

You will need to sum together these variances for the $M = 17$ decades of the global-average data. What is the probability that the decadal averages represent real signal, and not random fluctuations? What is the complementary probability of a Type 1 error that the temperature data are all drawn from a simple Gaussian PDF?

(c) Rather than sort the temperature anomalies into decades, sort the values of the Hadley Center global-average temperature anomalies into months of the year using R commands such as

```
djan <- seq(1,N,by=12)

January <- anomalies[djan]

dfeb <- seq(2,N,by=12)

February <- anomalies[dfeb]
```

. . . and so on. Limit the following analysis to the years 1850 through 1970, a total of 120 values for each month. Evaluate the ANOVA month-based means and variances for the 1850-1970 interval. There will be $N = 12$ months and $M = 120$ values in each month-based mean anomaly. Evaluate the F variance ratio $F = Y_b/Y_w$ for the "between" and "within" variances and use the R function for the F CDF (to estimate the statistical probability for nonrandomness. What is the probability that the month-based averages represent real signal, and not random fluctuations? (NOTE: we will revisit this statistical estimate in a later problem set after we have fit for the long-term trends in the temperature data.)

Problem 3: Statistical tests of data sets: Correlation of Hemispheric averages

Apply the formula for correlation of data series to compare the decadal temperatures of the Northern- and Southern-Hemisphere data series from the Hadley Center. The sample correlation coefficient $\hat{r}$ between two ordered data sets $\{X_n\}$ and $\{Y_n\}$ for $n = 1, 2, \ldots N$ is

$$\hat{r} = \frac{\sum_{n=1}^{N}(X_n - \bar{X}_N)(Y_n - \bar{Y}_N)}{\left(\sum_{n=1}^{N}(X_n - \bar{X}_N)^2 \sum_{m=1}^{N}(Y_m - \bar{Y}_N)^2\right)^{1/2}}$$

This formula returns $\hat{r}$ as a real number in the range $[-1, -1]$. In the problems that follow, use R commands like `mean(DecadalMeans)` to compute the means and generate a time axis with

```
decades <- c(1855:2015:10)
```

(a) Compute and plot the sample correlation $\hat{r}_w$ between monthly temperature anomalies in the two Hemispheres in decade-long intervals, that is, correlate the 1850s between the Hemispheres, then the 1860s, and so on until the 2010s. Plot the signed correlations $\hat{r}$ against the midpoints of their decades: 1855, 1865, . . . 2015. For this problem, the total number of months in each correlation is $N = 120$.

(b) One way to estimate the confidence for nonrandom correlation is to treat the pattern of amplitudes in $\{X_n\}$ as a single degree of freedom, that is, as a single dimension in an $(N-1)$-dimensional vector space. (Recall that we have used one dof already, because we have subtracted the mean before correlation.) The variance of the data series $\{Y_n\}$ that is explained by the amplitude pattern of $\{X_n\}$ is a random variable. For the null hypothesis that no correlation exists, the expectation value of $\hat{r}^2$ is a variance ratio with expectation value $1/(N-2)$, and the random variable $(N-2)\hat{r}^2$ should follow the F distribution with 1 and (N-2) degrees of freedom. Using the CDF function in R `pf()`, compute and plot the confidence for non-randomness for the correlations in part (a) of this problem – this quantity is the cumulative probability $P(F)$ integrated from zero to $F = (N-2)\hat{r}^2$.

(c) Compute and plot the complementary probabilities of a Type 1 error $(1 - P(F))$ for these $M = 17$ decadal correlations against years 1855, 1865, $\ldots$ 2015. This is the probability of a Type 1 error, and is the integral of the F distribution from $F = (N-2)\hat{r}^2$ to infinity. Use a logarithmic y-axis to capture the smaller values properly.

(d) compute the correlation $\hat{r}_b$ *between* the decadal means of the hemispheric temperature series over the 1850-2020 interval by adapting the correlation formula to correlate the $M = 17$ decadal means. For this case the F statistic is based on the quantity $F = (M-1)\hat{r}_b$. Compute the confidence for nonrandomness of correlation between the sample means (this is the CDF of the F distribution, of course!). Compute the complementary probability of a Type 1 error $(1 - P(F))$ for the assumption that the sequences of decadal means have correlation $\hat{r}_b$ between the hemispheres.