CHAPTER 2

EXERCISES FOR STATISTICAL INFERENCE

Exercise 2.1. A widely used measure of the state of the Pacific Ocean is the Pacific Decadal Oscillation (PDO) index. This index is a certain linear combination of sea surface temperatures in the Pacific Ocean poleward of 20N. More details and background information can be found at http://jisao.washington.edu/pdo/, but this information is not necessary in order to complete this homework assignment. The January-March mean value of this index is plotted in fig. 2.1. A glance at the figure reveals that the PDO index was predominantly negative during the period 1950-1976 and predominantly positive 1977-2017. In this homework set, you will address the following questions about the PDO:

1. Are the samples independent?
2. Has the mean PDO changed in recent decades?

To address the above questions, you need to download the data file `PDO.latest.txt` *from the class website*. The R tutorial discussed how to download this data set, but the data file from the PDO webpage changed slightly, so I have updated the data set and included R code for reading it called `Chapter02.exercise.pdo.R`. Both files can be downloaded from the class website[1] (click “data” and ”repository of R programs”). This code should run without generating any errors, so please contact the instructor if you have trouble running this code.

[1] http://cola.gmu.edu/delsole/clim762/webpage/clim762_frontpage.html

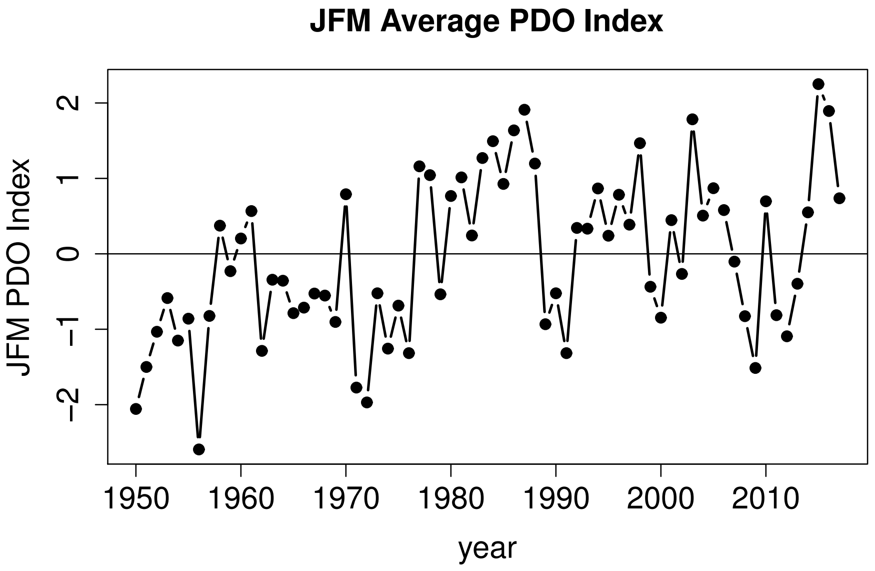

Figure 2.1 The January-March mean Pacific Decadal Oscillation (PDO) index. Positive values are indicated by thick bars. Statistics of the January-March mean PDO Index are given in the table.

For this homework set, you may use the built-in functions `mean` and `var`, but do not use the built-in functions `cov` and `cor`; rather, you should compute these quantities explicitly using the `sum` command (as shown in the tutorial). The purpose of requiring you to calculate these directly is to ensure you understand how to do these calculations yourself. You can, however, use these functions to check your results. You may also use these built-in functions after this homework set.

You will write a *function* that performs a hypothesis test. You should read the appropriate sections of the tutorial to become familiar with functions. In particular, the tutorial illustrates a complete example for one particular hypothesis test (i.e., equality-of-variance test), which in turn can serve as a template for the other hypothesis tests in this homework.

In each problem, you should explicitly state your null hypothesis, test statistic, the observed value of the statistic, rejection region, and final decision about the hypothesis. You should also state all the assumptions that were made in performing the hypothesis test.

(a) **Are the Samples Independent?** Recall that one assumption in both the F-test and t-test is that the samples are independent. Climate time series have the property that the degree of dependence decays with time separation. For instance, 1-day forecasts are more accurate than 5-day forecasts. One approach to checking the independence assumption is by testing if the correlation *between two consecutive time steps* vanishes. If no correlation can be detected between consecutive time steps, then it is unlikely it can be detected for larger time separations. Write an all-purpose function called `cor.equal.test` that tests the hypothesis that the correlation between two given data sets vanishes. Turn in a copy of your code. The call to this function should start as follows:

```
cor.equal.test = function(data1,data2,alpha=0.05) {
## THIS FUNCTION TESTS VANISHING CORRELATION BETWEEN
## TWO BI-VARIATE, NORMALLY DISTRIBUTED RANDOM VARIABLES
##
# INPUT:
#   DATA1: [N]-DIMENSIONAL VECTOR OF DATA
#   DATA2: [N]-DIMENSIONAL VECTOR OF DATA
#   ALPHA: SIGNIFICANCE LEVEL OF THE TEST (DEFAULT = 5%)
# OUTPUT LIST:
#   RHO: SAMPLE CORRELATION BETWEEN DATA1 AND DATA2
#   RHO.CRIT: 100*ALPHA% CRITICAL VALUE FOR THE CORRELATION
#   PVAL: P-VALUE OF THE STATISTIC RHO
```

(b) Using this function, test the hypothesis that during 1950-1977 the correlation between consecutive years is zero. This means that whatever you use for the time series in data1, data2 is the *same time series shifted by one year*. This also means you cannot input the entire time series, but instead the length of the time series is short by one year. Do the same test for 1978-2017. What is your answer to these questions? Turn in a copy of the output of your function, which should give numerical values for all quantities like $\hat{\rho}$, ρ_{crit}, p-value.

Incidentally, an R function called cor.test also performs this test. Check that your function agrees with cor.test.

(c) Repeat the above test, but this time for the whole period 1950-2017. You should find that $\rho = 0$ is rejected more strongly than for the two separate periods individually. Explain why. Hint: draw a scatter diagram for the three separate cases.

(d) Answer the other questions posed at the beginning (i.e., define your test statistic, state the value of the statistic, specify the rejection region, etc.).

(e) **Has the mean PDO changed in recent decades?** To address this question, we test the hypothesis of no difference in population mean between the periods 1950-1977 and 1978-2017? Write a function called mean.equal.test that performs the hypothesis test. Run your function on the PDO data. Turn in a copy of your code and its output. The function should begin as follows:

```
mean.equal.test = function(data1,data2,alpha=0.05) {
## THIS FUNCTION TESTS EQUALITY OF MEANS OF TWO IID
## NORMALLY DISTRIBUTED RANDOM VARIABLES
##
# INPUT:
#   DATA1: [N1]-DIMENSIONAL VECTOR OF DATA
#   DATA2: [N2]-DIMENSIONAL VECTOR OF DATA
#   ALPHA: SIGNIFICANCE LEVEL OF THE TEST (DEFAULT = 5%)
# OUTPUT LIST:
#   DIFF.MEAN: DIFFERENCE IN MEANS (MEAN1 - MEAN2)
#   DIFF.MEAN.CRIT: 100*ALPHA% LEVEL CRITICAL VALUE OF THE DIFFERENCE IN MEANS
#   PVAL: P-VALUE OF THE T-STATISTIC
#   T: T-STATISTIC FOR DIFFERENCE IN MEANS
#   T.CRIT: 100*ALPHA% LEVEL CRITICAL VALUE OF THE T STATISTIC
#   MEAN1: ESTIMATE OF THE MEAN OF DATA1
#   MEAN2: ESTIMATE OF THE MEAN OF DATA2
#   SPOOL: POOLED ESTIMATE OF THE STANDARD DEVIATION
```

(f) An output quantity not discussed in class (or the notes) is `diff.mean.crit`: this is the threshold value of the absolute difference in means for deciding whether to reject the null hypothesis. Write an equation that gives this threshold value. State the numerical value for this problem.

(g) State your test statistic, the value of the statistic, and specify the rejection region. What is your decision? Be sure to explicitly state all assumptions made in your hypothesis test.

(h) Explain whether the $\rho = 0$ test and equality-of-variance tests, which examined above and in the R tutorial, are relevant to testing equality of means.

□