Chapter 3
Basic Statistical Concepts

3.1 Introduction

A key scientific challenge is to better understand the functioning of the environment. Informed analysis of observations can make a strong contribution to this goal. The most insightful analysis requires knowledge of the relevant environmental processes and of statistical methodologies, that can lead the analyst towards a true understanding.

Compared to other aspects of the environment, climatology has a rich archive of direct observations. This has created an opportunity for the application of a wide range of statistical methods. This chapter reviews some of the basic statistical concepts that have been applied to better understand climate processes and to represent physically based predictability in the climate system. Like many environmental datasets, climate observations are sampling processes that evolve in space and time; the analysis of spatial patterns in time series of fields is the core of this book. Most reference is made to the application of special statistical techniques to study the fluctuation of climate from year to year. An additional special challenge is given by the size of the historical record. Typically, an analyst is faced with about 30–40 years of reliable data, which is sufficiently long to tease out some clues about the functioning of the climate system, but sufficiently short to lend itself to considerable imaginative interpretation. Thus, it becomes important to have a good appreciation for the effective sample size, so as to apportion the appropriate weight to the result in the overall investigation. When estimated properly, statistical significance allows us to have the correct degree of surprise at the statistical outcome, and therefore allows us to give the correct weight to this clue in our attempt to understand the big picture.

3.2 Climate Datasets

Climate observations were traditionally made at a known location. On land, this would be a climate station; over the ocean, this would normally be a ship, such that the exact location of the observations needed to be reported in addition to the climate state. The raw climate datasets from satellites can take a different form,

A. Navarra and V. Simoncini, *A Guide to Empirical Orthogonal Functions for Climate Data Analysis*, DOI 10.1007/978-90-481-3702-2_3,

being samples of space–time averages across the domain covered by the satellite. The blending of traditional and satellite datasets is therefore a critical step. In the context of climate analysis, data usually represent distinct observations at different times, possibly but not necessarily obtained at constant time intervals. In this case the term time series is usually employed for the data.

Since climate evolves in a continuous field, climate observations are often interpolated to a regular grid before analysis. These interpolation schemes may take in mind the processes and scales in the physical environment, and the ability of the data to resolve those scales. A good example are datasets of SST (monthly mean Sea Surface Temperatures). More recently, physically based interpolation schemes have been used to generate complete fields that are dynamically and physically consistent. These datasets have become known as the *reanalysis datasets*. They represent an ambitious advance in the creation of environmental datasets. In many ways, the user of such datasets needs more than ever to be aware of the types of data that were used in the study. Yet with careful analysis, they can provide an extremely powerful tool to deepen the understanding of the climate system.

The family of methods that are described in the following chapters are often applied to gridded datasets, such that the vectors derived from the analysis can be plotted as spatial patterns. However, there is no need to restrict the analysis to the gridded datasets. Analysis can equally be made of individual station time series. If the network of stations is sufficiently dense, contours of the weights can again be constructed to better communicate the meaning of the derived pattern. Alternatively, regional indices of climate, or regional indices of other environmental indicators can be used.

3.3 The Sample and the Population

An important concept in statistics is the relation between sample and population. Applying this concept to the analysis of short environmental series is not straightforward. It is assumed that the sample is taken from an infinite size population. The challenge is to infer characteristics of the population from the sample. The problem for climate science is that most properties of the system are not stationary. The problems of decadal climate variability have been mentioned above. In addition, the relationship between two variables need not be stationary. It can depend on the background climate state that prevailed over the analysis period. In fact, the degree of association between two variables may actually have varied during the 30 year period itself – though the sample size will likely be too small to deduce with any certainty that a real change took place. Let us pause to ask what we would mean by “a real change”. Assume that we find a run of 10 years when the correlation is lower than during the whole historical record. What we want to know is the following: in case the interannual variability were repeatedly run with the prevailing background climate state of those 10 years, would that low correlation be maintained? or, would the 10 years of low correlation be merely due to the inevitable sampling fluctuations that occur even when the correlation between two variables is statistically stationary?

In addition, how does the situation change when we take a sample of a correlation coefficient over 30 years? The question we are trying to answer by taking that sample is: if the same background climate state were to continually operate and generate an infinite number of years of interannual variability, what would the correlation between the two variables be? In other words, the population is an imaginary infinite set of realizations generated from a given background climate state. For the purposes of making inferences (see statistical significance section below), we must assume that the correlation coefficient was stationary over the 30 year period itself.

3.4 Estimating the Mean State and Variance

A critical step in climate analysis is nearly always the estimation of the background mean state. Given the data $x_1, \ldots, x_n$, the mean, or average, is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

The computation of the mean is crucial to allow estimation of climate anomalies, given by the deviation from the mean, that is,

$$x_i' = x_i - \bar{x}.$$

The climate anomaly represents the departure from the assumed population mean at a given time for a given time series. If there is a systematic bias in the estimation of the mean from one location to another, this can introduce bias in the covariance of anomalies between the two series; see later chapters for a more detailed discussion. Most widely available datasets have given careful consideration to the estimation of the mean from which anomalies are calculated.

For the background mean state for a dataset, the dataset creator will have considered such features as the period with best data coverage. If one is working with the subsequent anomaly dataset, one still has to make a choice over which years to run your analysis. This requires careful consideration and some experimentation, because of multi-year (decadal and beyond) variability in the climate system. Choice of period can greatly impact the amount of variance represented by a decadal mode of variation. For example, an analysis over West Africa for 1971–2000 contains little decadal variability, whereas 1950–1980 is dominated by a decadal fluctuation.

The sample variance of the observed data is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

In particular, we note the multiplicative factor $\frac{1}{n-1}$, as opposed to the more intuitive factor $\frac{1}{n}$. The new factor allows the variance defined above to represent an unbiased estimator of the population variance; see Clarke and Cooke (1998). To intuitively explain this fact, we note that in s^2 there are $n-1$ degrees of freedom, rather than n, because $x_1, \ldots, x_n$ are related via the mean $\bar{x}$. Therefore, roughly speaking, the division by $n-1$ takes into account the actual number of degrees of freedom in the data.

The variance provides a measure of dispersion of data around the mean. The larger the variance, the more spread the data. It is important to remark that the variance is expressed in the square of the data measure unit. For this reason, its square root s, called the standard deviation, is also referenced. Both statistics introduced above are dimensional quantities. To be able to perform a meaningful comparison among data expressed in different measure units, observational data are usually standardized to adimensional numbers. This is achieved by using the following transformation:

$$z_i = \frac{x_i - \bar{x}}{s}. \tag{3.1}$$

The standardized variable has mean zero and standard deviation equal to one.

Exercises and Problems

1. Given the data $\{1.2, -1.0, 1.1, 0.8, -0.4, 0.95, -0.2\}$, determine their mean, variance and standard deviation. Then, standardize the variables by means of (3.1).
   We have $n = 7$. *Simple computation gives* $\bar{x} = 0.35$, $s^2 = 0.75583$ *and* $s = 0.86939$. *Standardization using (3.1) provides the following new data (final results rounded to the first five decimal digits),*

   $$\{0.97770, -1.5528, 0.86268, 0.51761, -0.86268, 0.69014, -0.63263\},$$

   for which we obtain $\bar{z} = 0$ *and* $s(z) = 1$.
2. Given the data $\{1.2, -19, 2.68, 0.8 - 3.0, 20.0, -0.2\}$, compute mean, variance, standard deviation. Compare the results with those of the previous exercise.
   We have $n = 7$. *Simple computation gives* $\bar{x} = 0.3542$, $s^2 = 129.7$ *and* $s = 11.39$. *Although the mean is basically the same as for the previous data, the variance and the standard deviation are much larger in this case. This shows that these data are more spread around the mean, as it can be clearly noticed by directly inspecting the data.*

3.5 Associations Between Time Series

The basis for applying EOF methods derives from the realization that the evolution of climate processes in time leads to relationships between time series of different atmospheric variables at nearby and remote locations. There are many ways to measure the nature and extent of a relationship between two time series. One of the most common is the Pearson correlation coefficient. This is closely related to the concept of least squares linear regression. To illustrate this concept, we first do the simplest thing possible to explore the relationship between two time series – we make a scatter plot of the observation pairs (x_i, y_i) (see the symbols "$*$" in Fig. 3.1).

Making an assumption of a linear relationship, we try to draw a straight line through the data points. We can fit the line to minimize the sum of squared errors in the Y variable. This line captures some of the variance in the independent series. In mathematical terms, this line yields the "best approximation" straight line, in the least squares sense, and it is given by the equation $y = b_1 x + b_0$, with

$$b_1 := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad b_0 = \bar{y} - b_1 \bar{x}.$$

The fraction of variance represented, corresponds to the degree of association (cf. Fig. 3.1). Analogously, the line $x = c_1 y + c_0$ can be drawn to minimize the sum of squared errors in the X variable. The fraction of variance explained is the same as for the Y variable. The combination of the two coefficients b_1 and c_1 yields the

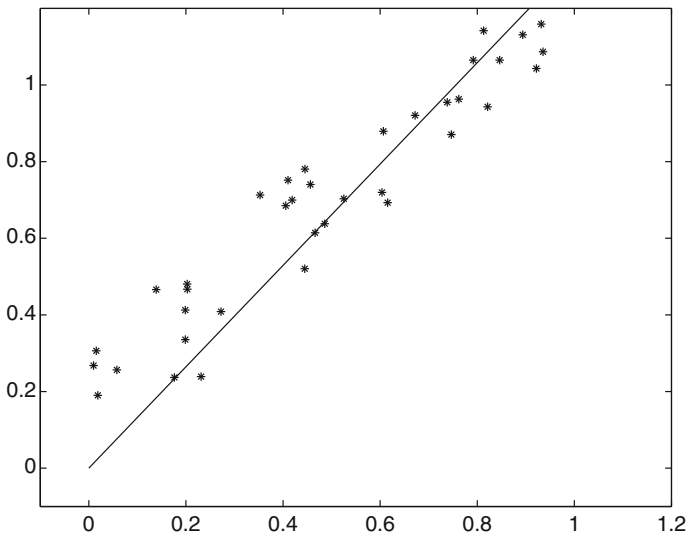

Fig. 3.1 Scatter plot of observations and fitting line

correlation coefficient r, which provides a measure of association among the two variables, and it is defined as

$$r^2 = b_1 \cdot c_1 = \frac{\left(\sum_i (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_i (y_i - \bar{y})^2 \sum_i (x_i - \bar{x})^2}.$$

To generalize this concept to multidimensional data, assume now that a set of $m \times n$ data $x_{1,1}, x_{1,2}, \ldots, x_{1,n}, \ldots, x_{m,n}$ is given. Here we are considering m variables and n observations (time series of length n for each of the m variables). Let $\bar{x}_j, \bar{x}_k$ be the means associated with the time series j and k. Analogously, we define the standard deviations s_j, s_k. For each pair of variables, the associated correlation coefficient is given by

$$r_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_{j,i} - \bar{x}_j)(x_{k,i} - \bar{x}_k)}{s_j s_k}.$$

For the ith observation, $i = 1, \ldots, n$, the sum above multiplies the standardized jth and kth variables. The coefficient associated with these two variables is small (large) in absolute value, if both standardized variables are small (large), in all n observations. The normalization operates such that the correlation takes values between -1 (all points would lie on a backward sloping line) and 1 (all points would fall on a forward sloping line, cf. Fig. 3.1). Note that $r_{j,j} = 1$ for all j. In case standardization is not used, a related measure of association between deviations is the covariance coefficient, which can be viewed as a non-normalized correlation:

$$s_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{j,i} - \bar{x}_j)(x_{k,i} - \bar{x}_k). \tag{3.2}$$

Here $s_{j,j} = s_j^2$ is the variance of the jth variable. The matrix $\mathbf{S} = (s_{j,k})$ of all coefficients above is called the (cross-)covariance matrix and is symmetric, that is the covariance between the jth and kth variables is the same as the covariance between the kth and jth variables. The total variance of the field is then given by

$$T = \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{i,j} - \bar{x}_i)^2 = \sum_{i=1}^{m} s_{ii} = \text{trace}(\mathbf{S}), \tag{3.3}$$

showing that the total field variance is just the trace of the covariance matrix.

Both the above are related to the squared error departures from a linear relation. There are other ways to measure association. An example is the rank order Spearman correlation coefficient; see, e.g., Clarke and Cooke (1998).

Other measures could be determined based on absolute error. For instance, the Linear Error in Probability Space (LEPS) works on the mean absolute difference in the ranking, or cumulative probability. Associations could be measured in terms of the extent to which variance is explained by some specified non-linear relationship, such as quadratic or log linear. The correlation coefficients introduced above can be collected in one matrix, that more clearly visualizes the association of each time series with all others.

For instance, the correlation matrix is given by

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \vdots & \ddots & \ddots & \vdots \\ r_{m,1} & r_{m,2} & \cdots & r_{m,m} \end{pmatrix}.$$

As will be shown, there are special properties of correlation and covariance matrices that can be uncovered by a Principal Component Analysis. Matrix properties of other measures of association have not been as much investigated; they will be alluded to in the following chapters.

Exercises and Problems

1. Given the data $x = \{-1.1, 0.8, 1.2\}$ and $y = \{0.6, 0.9, 2.4\}$, determine the covariance and correlation matrices.
   We have $m = 2$ variables, and $n = 3$ observations. Therefore, $\bar{x} = 0.3$, $\bar{y} = 1.3$, $s(x) = 1.2288$ and $s(y) = 0.9643$, so that the standardized variables are $z(x) = \{-1.1393, 0.40689, 0.73241\}$ and $z(y) = \{-0.72587, -0.41478, 1.1406\}$. The correlation coefficient is given by $r_{1,2} = 1/2(z(x)_1 z(y)_1 + z(x)_2 z(y)_2 + z(x)_3 z(y)_3) = 0.74$ (note that the computation of $r_{1,2}$ is actually done with full accuracy and only the first 2 decimals are reported). Hence, the corresponding matrix is

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0.74 \\ 0.74 & 1 \end{pmatrix}.$$

   The value of $r_{1,2}$ shows a significant positive correlation between the two variables. Analogously, the covariance is given by $s_{1,2} = 0.885$.
2. Given the data $x = \{-1.1, 0.8, 1.2\}$, $y = \{0.6, 0.9, 2.4\}$ and $z = \{4.2, -1.1, 6.8\}$, determine the covariance and correlation matrices.
   We have $m = 3$ variables, and $n = 3$ observations. The first two sets are as in the previous example. We have, $z = 3.3$, $s(z) = 4.0262$, so that the new standardized variable is $z(x) = \{0.22354, -1.0929, 0.86931\}$. We obtain $r_{1,3} = -0.00313$ and $r_{2,3} = -0.64$. The correlation between the y and z variables is significant, whereas that between x and z is negligible. Analogously, we obtain $s_{3,3} = 16.21$, $s_{1,3} = -0.155$ and $s_{2,3} = 2.49$.

3. Given the data $x = \{-1.1, 0.8, 1.2\}$, $y = \{0.6, 0.9, 2.4\}$ and $z = \{104.2, -100.1, 126.8\}$, determine the covariance coefficients. Comment on the role of dimensionality.
   Only the third variable has changed. We have $s_{3,3} = 1562.2$, $s_{1,3} = -40.90$ *and* $s_{2,3} = 53.29$. *Note that the larger variability is due to the significantly different unit of z, which is also reflected in the covariance coefficients.*

3.6 Hypothesis Testing

In an attempt to infer conclusions on an unobservable population, we can set about estimating the extent to which our available sample can tell us something about that population. Let us consider the simple example of testing whether the population mean is zero. Statistical significance is estimated by formally expressing two possibilities that we need to choose between. Here, the first one is that the population mean is zero. As an alternative possibility, we can say that the population mean is not zero (other options may be more significant in some cases, such as "mean greater than zero"). Formally, the original hypothesis is termed the null hypothesis (H_0), whereas the second one is called the alternative hypothesis (H_1), and these are written as

$$H_0 : \quad \mu = 0, \qquad H_1 : \quad \mu \neq 0.$$

We want to distinguish between these two possibilities in a way that allows us to know the likelihood that our choice is in fact wrong (i.e. how surprised we should be if our decision turns out to be the wrong one). We start out by assuming that H_0 is true. If H_0 is true, then the sample should obey certain statistical properties. If the sample does not reflect these properties, then we start to doubt H_0. For example, we can define a test statistic whose distribution we know under the assumption that H_0 is true and we explore to what extent our sample obeys this distribution.

A particularly popular distribution is the *normal* distribution, as it represents an effective model for data stemming from a variety of applications. Data following a normal distribution distribute around their mean with a probability that decreases significantly as data move away from the mean. The set of normally distributed variables with mean μ and variance σ^2 is usually denoted by $N(\mu, \sigma^2)$. The probability of normal data distributes along a bell-shaped curve, as described in the plots of Fig. 3.2 for various values of μ and σ. In other words, the probability that a sample taken from an $N(\mu, \sigma^2)$ normal population has mean in the interval $[\mu - d, \mu + d]$ equals the area of the region below the bell-shaped curve, with extremes on the ascissa at $\mu - d$ and $\mu + d$. A normally distributed variable x with mean μ and variance σ^2 can be transformed into a standardized normally distributed variable in $N(0, 1)$ by means of the change of variable $z = (x - \mu)/\sigma$. Reference values for a variable z in $N(0, 1)$ are tabulated and can be used for hypothesis tests. Most

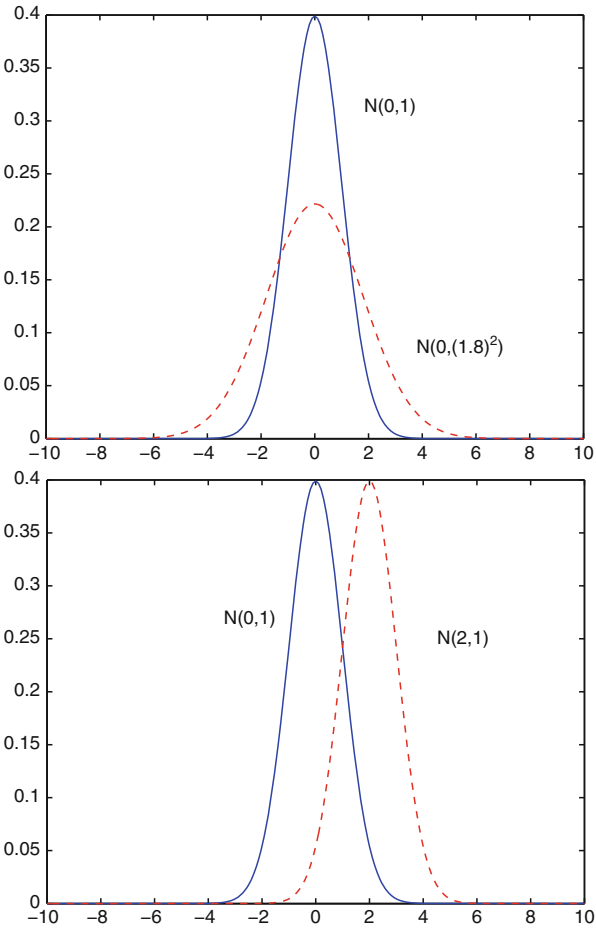

Fig. 3.2 Normal distributions for various values of μ and σ

statistical computer software provides a pretty accurate evaluation of the probability and other quantities associated with the normal distribution.

The trick in hypothesis testing is to define powerful test statistics, such as the standardized statistic

$$z = \frac{x - \mu}{se},$$

where se is the standard error of x, given by

$$se = \frac{\sigma}{\sqrt{n}},$$

σ is the population standard deviation and n is the sample size. The standard error represents the standard deviation of the sample mean distribution. In other words,

imagine estimating the mean of the population ten times, each time taking a sample of n individuals from the population. The ten resulting sample means will all be slightly different. The expected standard deviation of the sampled means is what we refer to as the standard error of the estimated mean.

Assume that H_0 is true. If the sample is sufficiently large, namely if n is sufficiently large, then z approximately behaves as if it were normally distributed with mean 0 and standard deviation 1.

Now, if the distribution of z departs substantially from 0, then we may start to doubt H_0. If the variable z were exactly normally distributed with zero mean and unit variance, then only on 2.5% of occasions z would take a value of $+1.96$ or higher. Likewise, on a further 2.5% of occasions, z would take a value not greater than -1.96 (cf. Fig. 3.3). That is, there is a 5% chance of the absolute value of z being greater than 1.96. So, if z takes an absolute value greater than 1.96, such a result is certainly quite surprising if in fact, the true mean is zero (surprising because we only expect it to happen on 5% of occasions when we sample a population mean with mean 0). Note that we can never be certain that H_0 is wrong. For statistical significance, we may decide that something that would only happen by chance on 5% of occasions is just too surprising, and that the wisest choice to make in this situation is to conclude that the available evidence does not support H_0. That is, at the 5% level of significance, we reject H_0 and accept the alternative hypothesis H_1, that the mean is not equal to zero. However, in terms of acquiring clues about the overall functioning of the environment, we may prefer not to work in the discrete terms of rejection or acceptance of H_0. Rather, acknowledging that using statistics alone, we can never distinguish between the two hypotheses with certainty, we may prefer to note the likelihood that H_0 can be rejected based on statistics alone, and absorb this information into broader evidence based on physical theories and physically based models.

The above approach estimates the probability of rejecting H_0 by starting with the assumption that H_0 is actually true. This is the usual way to frame a statistical

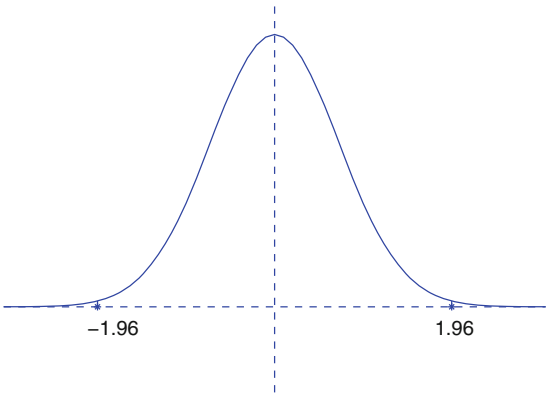

Fig. 3.3 Normal distribution. The area of the region below the curve and absissas in $[-1.96, 1.96]$ is equal to 0.95

significance test, on the premise that the analyst is usually interested in whether H_0 can be rejected, such as with whether a correlation coefficient is non-zero. Here, if we choose to reject the idea that the true correlation is zero, what is the probability that we are wrong (and in fact, there is a linear association between the two variables). This probability of wrongly rejecting H_0 is often termed the probability of making a Type I error, and is the statistical significance probability, alpha. However, there is another error type that can be made, usually referred to as Type II error, that of accepting H_0 when in fact H_0 should be rejected. This probability can also be estimated assuming the distribution of test statistics. However, it is generally not considered as useful as the Type I error probability, that focuses on whether we can reject H_0.

The distribution of the reference statistic z is easy to derive and work with. In many instances the test statistic is more complex. A typical complication appears when the standard deviation of the population is not known (of course, this is usually the situation we find ourselves in). In this situation, we can use the Student statistics, or t-statistic, in which the population standard deviation is replaced by the sample standard deviation, that is

$$t_0 = \frac{x - \mu}{\bar{s}} \sqrt{n} \tag{3.4}$$

The new variable t_0 depends on n, more precisely on $n - 1$, and for each value of n, t_0 follows a specific distribution. Is it important to stress that to be able to employ the Student distribution as test statistic, we need to assume that the given sample comes from a normal distribution.

As n grows, the Student distribution increasingly resembles the normal distribution. The likelihood of t_0 exceeding a reference value is tabulated, for different values of $n - 1$, called the degrees of freedom, Df, which is related to the size of the available sample. The degrees of freedom is a complicated issue for many climate analyses. The above holds if each term in the sample is independent. However, in many climate time series, adjacent observations are correlated in time, and this reduces the effective degrees of freedom (and can complicate the distribution of the test statistic). This is particularly a challenge for estimating the significance of the relationship between two variables. The correlation coefficient significance is very difficult to estimate because of this effect; see von Storch and Zwiers (1999). This problem transfers into the estimation of significance for EOFs, since they themselves are summaries of the cross-correlations/covariances in datasets.

Exercises and Problems

1. Assume that a sample of 100 units is taken from a population which was in the past known to have mean $\mu = 12.3$ and standard deviation $\sigma = 15$. The computed sample mean is $x = 14.2$. Carry out a hypothesis test with 5% level of significance, to analyze whether the population mean has changed.

We set H_0*:* $\mu = 12.3$ *and* $H_1 : \mu \neq 12.3$*. We have* $z = (x - \mu)/\sigma = 0.12$*. The critical region for 5% level of significance would be* $|z| > 1.96$*, therefore the new variable z is well away from the critical region. We do not reject the null hypothesis.*

2. What would happen in the example above if the standard deviation were $\sigma = 0.9$? What if the significance level were 1%?
   With the same framework as before, we have $z = (x - \mu)/\sigma = 2.11$*, hence this variable falls within the critical region* $|z| > 1.96$*. We have to reject the null hypothesis in favor of the alternative hypothesis* H_1 *for a 5% level of significance. For a significance level equal to* 1%*, the corresponding critical region is* $|z| > 2.57$*, so that the null hypothesis would not be rejected.*

The inherent difficulty associated with the effective number of degrees of freedom in the Student statistics is one of the reasons why alternatives such as Monte Carlo estimates of significance are attractive. To illustrate the concept, consider that we have two time series of length 30 years. Each time series has serial correlation and can be represented by an autoregressive process:

$$x_t = a x_{t-1} + z_t. \tag{3.5}$$

We can use random number generators in combination with the above model to simulate 500 pairs of time series with the same serial correlation properties as the original two series. The distribution of the 500 correlations between each randomly generated pair of series is now constructed empirically. We expect the mean of the correlations calculated to be zero, but the spread will depend on the degree of autocorrelation in the two series. If the pair of series are highly auto-correlated, the location of the correlation magnitude that occurs on 5% of occasions will be much higher than if the pair of series were uncorrelated. Now we are using the correlation itself as the test statistic, knowing the distribution of the sample correlations under the assumption that the true population correlation is 0. The correlation magnitude corresponding to the 5% significance level can be found by identifying the threshold above which were found only 5% of the sample correlations. The temporal d.f. problem is also present for methods devised to estimate the statistical significance of EOFs. Higher percentage of variance explained are expected by chance, when time series used in the EOF analysis contain serial correlation. Thus caution is needed not to place excessive weight on significance estimates of EOFs when series have serial correlation.

3.7 Missing Data

Dealing with missing data is an important aspect for application of EOF methods. In some datasets, the fields will have been made complete for the analyst, in which case the analyst should investigate carefully the way the data were interpolated and

possible consequences for EOF analysis as discussed below. The EOF methods are applied to correlation and covariance matrices. It is tempting to calculate correlations using the available data for each pair of series, assuming that this gives the best estimate of the correlation between each series, even if some correlations are based on a smaller sample than others. However, this approach can lead to problems with the inversion of the correlation/covariance matrices to derive the EOF solutions. It is usually best to make all series complete in some way over the analysis period.

Usually, the analyst decides on a fixed analysis period (say, 1961–1990) and decides on the maximum number of missing values that is acceptable for a series to be included (say, at least 25 out of 30 values must have data). A simple and quite robust solution to missing data is to set all missing values in a series equal to the mean of the available data for that series. This will ensure the missing values are all zero anomalies when the correlation/covariance matrices are calculated. Zero anomalies have least impact on the correlation/covariances. While it can reduce some genuine cross correlations between time series and this can distort the EOF solutions, it is nonetheless a cautious conservative approach and as such, is an attractive solution. Application of more sophisticated interpolation methods requires care for any increase in correlations/covariances that it may introduce into the datasets.