Chapter 2

Misuses of Statistical Analysis in Climate Research

by Hans von Storch

2.1 Prologue

The history of misuses of statistics is as long as the history of statistics itself. The following is a personal assessment about such misuses in our field, climate research. Some people might find my subjective essay of the matter unfair and not balanced. This might be so, but an effective drug sometimes tastes bitter.

The application of statistical analysis in climate research is methodologically more complicated than in many other sciences, among others because of the following reasons:

- In climate research only very rarely it is possible to perform real *independent experiments* (see Navarra's discussion in Chapter 1). There is more or less only one observational record which is analysed again and again so that the processes of building hypotheses and testing hypotheses are hardly separable. Only with dynamical models can independent

Acknowledgments: I thank Bob Livezey for his most helpful critical comments, and Ashwini Kulkarni for responding so positively to my requests to discuss the problem of correlation and trend-tests.

data be created - with the problem that these data are describing the real climate system only to some unknown extent.

- Almost all data in climate research are *interrelated* both in space and time - this spatial and temporal correlation is most useful since it allows the reconstruction of the space-time state of the atmosphere and the ocean from a limited number of observations. However, for statistical inference, i.e., the process of *inferring* from a limited sample robust statements about an hypothetical underlying "true" structure, this correlation causes difficulties since most standard statistical techniques use the basic premise that the data are derived in independent experiments.

Because of these two problems the fundamental question of how much *information* about the examined process is really available can often hardly be answered. Confusion about the amount of information is an excellent hotbed for methodological insufficiencies and even outright errors. Many such insufficiencies and errors arise from

- The *obsession with statistical recipes* in particular *hypothesis testing*. Some people, and sometimes even peer reviewers, react like Pawlow's dogs when they see a hypothesis derived from data and they demand a statistical test of the hypothesis. (See Section 2.2.)

- The use of statistical techniques as a *cook-book like recipe* without a real understanding about the concepts and the limitation arising from unavoidable basic assumptions. Often these basic assumptions are disregarded with the effect that the conclusion of the statistical analysis is void. A standard example is disregard of the serial correlation. (See Sections 2.3 and 9.4.)

- The *misunderstanding of given names*. Sometimes physically meaningful *names* are attributed to mathematically defined objects. These objects, for instance the *Decorrelation Time*, make perfect sense when used as prescribed. However, often the statistical definition is forgotten and the *physical meaning* of the *name* is taken as a definition of the object - which is then *interpreted* in a different and sometimes inadequate manner. (See Section 2.4.)

- The *use of sophisticated techniques*. It happens again and again that some people expect miracle-like results from advanced techniques. The results of such advanced, for a "layman" supposedly non-understandable, techniques are then believed without further doubts. (See Section 2.5.)

2.2 Mandatory Testing and the Mexican Hat

In the desert at the border of Utah and Arizona there is a famous combination of vertically aligned stones named the "Mexican Hat" which looks like a human with a Mexican hat. It is a random product of nature and not man-made ... really? *Can we test the null hypothesis "The Mexican Hat is of natural origin"?* To do so we need a *test statistic* for a pile of stones and a probability distribution for this test statistic under the null hypothesis. Let's take

$$t(p) = \begin{cases} 1 & \text{if } p \text{ forms a Mexican Hat} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

for any pile of stones p. How do we get a probability distribution of $t(p)$ for all piles of stones p not affected by man? - We walk through the desert, examine a large number, say $n = 10^6$, of piles of stones, and count the frequency of $t(p) = 0$ and of $t(p) = 1$. Now, the Mexican Hat is famous for good reasons - there is only one p with $t(p) = 1$, namely the Mexican Hat itself. The other $n - 1 = 10^6 - 1$ samples go with $t(p) = 0$. Therefore the probability distribution for p not affected by man is

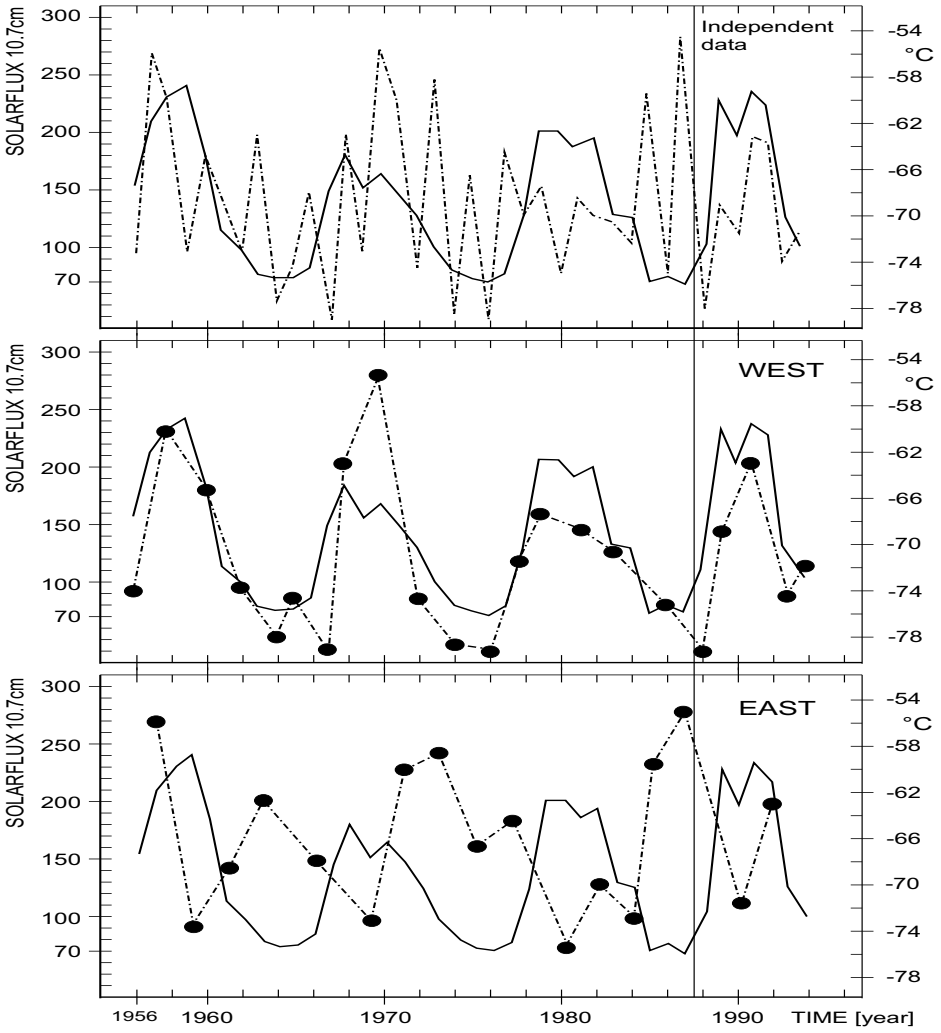
$$prob\,(t(p) = k) = \begin{cases} 10^{-6} & \text{for } k = 1 \\ 1 - 10^{-6} & \text{for } k = 0 \end{cases} \tag{2.2}$$

After these preparations everything is ready for the final test. We reject the null hypothesis with a risk of 10^{-6} if $t(\text{Mexican hat}) = 1$. This condition is fulfilled and we may conclude: *The Mexican Hat is not of natural origin but man-made.*

Obviously, this argument is pretty absurd - but where is the logical error? The fundamental error is that *the null hypothesis is not independent of the data which are used to conduct the test.* We know a-priori that the Mexican Hat is a rare event, therefore the impossibility of finding such a combination of stones cannot be used as evidence against its natural origin. The same trick can of course be used to "prove" that any rare event is "non-natural", be it a heat wave or a particularly violent storm - the probability of observing a rare event is small.

One might argue that no serious scientist would fall into this trap. However, they do. The hypothesis of a connection between solar activity and the statistics of climate on Earth is old and has been debated heatedly over many decades. The debate had faded away in the last few decades - and has been refueled by a remarkable observation by K. Labitzke. She studied the relationship between the solar activity and the stratospheric temperature at the North Pole. There was no obvious relationship - but she *saw* that during years in which the Quasibiennial Oscillation (QBO) was in its West Phase, there was an excellent *positive* correlation between solar activity and North Pole temperature whereas during years with the QBO in its East Phase

Figure 2.1: *Labitzke' and van Loon's relationship between solar flux and the temperature at 30 hPa at the North Pole for all winters during which the QBO is in its West Phase and in its East Phase. The correlations are 0.1, 0.8 and -0.5. (From Labitzke and van Loon, 1988).*



there was a good *negative* correlation (Labitzke, 1987; Labitzke and van Loon, 1988).

Labitzke's finding was and is spectacular - and obviously right for the data from the time interval at her disposal (see Figure 2.1). Of course it could be that the result was a coincidence as unlikely as the formation of a Mexican Hat. Or it could represent a real on-going signal. Unfortunately, the data which were used by Labitzke to formulate her hypothesis can no longer be used for the assessment of whether we deal with a signal or a coincidence. Therefore an answer to this question requires information unrelated to the data as for instance dynamical arguments or GCM experiments. However, physical hypotheses on the nature of the solar-weather link were not available and are possibly developing right now - so that nothing was left but to wait for more data and better understanding. (The data which have become available since Labitzke's discovery in 1987 support the hypothesis.)

In spite of this fundamental problem an intense debate about the "statistical significance" broke out. The reviewers of the first comprehensive paper on that matter by Labitzke and van Loon (1988) demanded a test. Reluctantly the authors did what they were asked for and found of course an extremely little risk for the rejection of the null hypothesis "The solar-weather link is zero". After the publication various other papers were published dealing with technical aspects of the test - while the basic problem *that the data to conduct the test had been used to formulate the null hypothesis* remained.

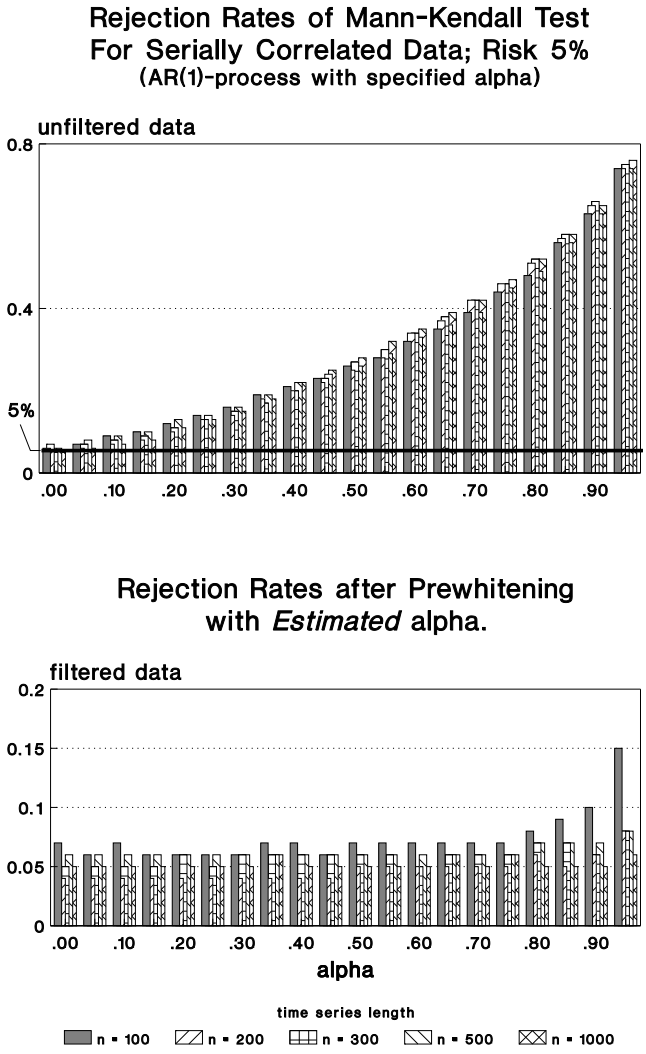
When hypotheses are to be derived from limited data, I suggest two alternative routes to go. If the time scale of the considered process is short compared to the available data, then split the full data set into two parts. Derive the hypothesis (for instance a statistical model) from the first half of the data and examine the hypothesis with the remaining part of the data.[1] *If the time scale of the considered process is long compared to the time series such that a split into two parts is impossible, then I recommend using all data to build a model optimally fitting the data. Check the fitted model whether it is consistent with all known physical features and state explicitly that it is impossible to make statements about the reliability of the model because of limited evidence.*

2.3 Neglecting Serial Correlation

Most standard statistical techniques are derived with explicit need for statistically independent data. However, almost all climatic data are somehow correlated in time. The resulting problems for testing null hypotheses is discussed in some detail in Section 9.4. In case of the t-test the problem is nowadays often acknowledged - and as a cure people try to determine the "equivalent sample size" (see Section 2.4). When done properly, the t-test

[1] An example of this approach is offered by Wallace and Gutzler (1981).

Figure 2.2: *Rejection rates of the Mann-Kendall test of the null hypothesis "no trend" when applied to 1000 time series of length n generated by an AR(1)-process (2.3) with prescribed α. The adopted nominal risk of the test is 5%.*
Top: results for unprocessed serially correlated data.
Bottom: results after pre-whitening the data with (2.4). (From Kulkarni and von Storch, 1995)



becomes conservative - and when the "equivalent sample size" is "optimized" the test becomes liberal[2]. We discuss this case in detail in Section 2.4.

There are, however, again and again cases in which people simply ignore this condition, in particular when dealing with more exotic tests such as the Mann-Kendall test, which is used to reject the null hypothesis of "no trends". To demonstrate that the result of such a test really depends *strongly* on the autocorrelation, Kulkarni and von Storch (1995) made a series of Monte Carlo experiments with AR(1)-processes with different values of the parameter α.

$$\mathbf{X}_t = \alpha \mathbf{X}_{t-1} + \mathbf{N}_t \tag{2.3}$$

with Gaussian "white noise" $\mathbf{N}_t$, which is neither auto-correlated nor correlated with $\mathbf{X}_{t-k}$ for $k \geq 1$. α is the lag-1 autocorrelation of $\mathbf{X}_t$. 1000 iid[3] time series of different lengths, varying form $n = 100$ to $n = 1000$ were generated and a Mann-Kendall test was performed. Since the time series have no trends, we expect a (false) rejection rate of 5% if we adopt a risk of 5%, i.e., 50 out of the 1000 tests should return the result "reject null hypothesis". The actual rejection rate is much higher (see Figure 2.2). For autocorrelations $\alpha \leq 0.10$ the actual rejection rate is about the nominal rate of 5%, but for $\alpha = 0.3$ the rate is already 0.15, and for $\alpha = 0.6$ the rate > 0.30. If we test a data field with a lag-1 autocorrelation of 0.3, we must expect that on average at 15% of all points a "statistically significant trend" is found even though there is no trend but only "red noise". This finding is mostly independent of the time series length.

When we have physical reasons to assume that the considered time series is a sum of a trend and stochastic fluctuations generated by an AR(1) process, and this assumption is sometimes reasonable, then there is a simple cure, the success of which is demonstrated in the lower panel of Figure 2.2. Before conducting the Mann-Kenndall test, the time series is "pre-whitened" by first estimating the lag-autocorrelation $\hat{\alpha}$ at lag-1, and by replacing the original time series $\mathbf{X}_t$ by the series

$$\mathbf{Y}_t = \mathbf{X}_t - \hat{\alpha} \mathbf{X}_{t-1} \tag{2.4}$$

The "pre-whitened" time series is considerably less plagued by serial correlation, and the same Monte Carlo test as above returns actual rejections rates close to the nominal one, at least for moderate autocorrelations and not too short time series. The filter operation (2.4) affects also any trend; however, other Monte Carlo experiments have revealed that the power of the test is reduced only weakly as long as α is not too large.

A word of caution is, however, required: If the process is *not* AR(1) but of higher order or of a different model type, then the pre-whitening (2.4)

[2] A test is named "liberal" if it rejects the null hypothesis more often than specified by the significance level. A "conservative" rejects less often than specified by the significance level.

[3] "iid" stands for "independent identically distributed".

is insufficient and the Mann-Kenndall test rejects still more null hypotheses than specified by the significance level.

Another possible cure is to "prune" the data, i.e., to form a subset of observations which are temporally well separated so that any two consecutive samples in the reduced data set are no longer autocorrelated (see Section 9.4.3).

When you use a technique which assumes independent data and you believe that serial correlation might be prevalent in your data, I suggest the following "Monte Carlo" diagnostic: Generate synthetic time series with a prescribed serial correlation, for instance by means of an AR(1)-process (2.3). Create time series without correlation ($\alpha = 0$) and with correlation ($0 < \alpha < 1$) and try out if the analysis, which is made with the real data, returns different results for the cases with and without serial correlation. In the case that they are different, you cannot use the chosen technique.

2.4 Misleading Names: The Case of the Decorrelation Time

The concept of "the" *Decorrelation Time* is based on the following reasoning:[4] The variance of the mean $\bar{\mathbf{X}}^n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k$ of n identically distributed and independent random variables $\mathbf{X}_k = \mathbf{X}$ is

$$\textsc{Var}\big(\bar{\mathbf{X}}^n\big) = \frac{1}{n}\textsc{Var}(\mathbf{X}) \tag{2.5}$$

If the $\mathbf{X}_k$ are autocorrelated then (2.5) is no longer valid but we may *define* a number, named the *equivalent sample size* n' such that

$$\textsc{Var}\big(\bar{\mathbf{X}}^n\big) = \frac{1}{n'}\textsc{Var}(\mathbf{X}) \tag{2.6}$$

The *decorrelation time* is then *defined* as

$$\tau_D = \lim_{n\to\infty} \frac{n}{n'} \cdot \Delta t = \left[1 + 2\sum_{\Delta=1}^{\infty} \rho(\Delta)\right] \Delta t \tag{2.7}$$

with the autocorrelation function ρ of $\mathbf{X}_t$.

The decorrelation times for an AR(1) process (2.3) is

$$\tau_D = \frac{1+\alpha}{1-\alpha}\Delta t \tag{2.8}$$

There are several conceptual problems with "the" Decorrelation Time:

[4] This section is entirely based on the paper by Zwiers and von Storch (1995). See also Section 9.4.3.

- The definition (2.7) of a decorrelation time makes sense *when dealing with the problem of the mean of n-consecutive serially correlated observations.* However, its arbitrariness in defining a *characteristic time scale* becomes obvious when we reformulate our problem by replacing the *mean* in (2.6) by, for instance, the *variance.* Then, the characteristic time scale is (Trenberth, 1984):

  $$\tau = \left[1 + 2\sum_{k=1}^{\infty} \rho^2(k)\right] \Delta t$$

  Thus characteristic time scales τ depends markedly on the *statistical problem* under consideration. *These numbers are, in general, not physically defined numbers.*

- For an AR(1)-process we have to distinguish between the physically meaningful processes with positive memory ($\alpha > 0$) and the physically meaningless processes with negative memory ($\alpha < 0$). If $\alpha > 0$ then formula (2.8) gives a time $\tau_D > \Delta t$ representative of the decay of the auto-correlation function. Thus, in this case, τ_D may be seen as a physically useful time scale, namely a "persistence time scale" (but see the dependency on the time step discussed below). If $\alpha < 0$ then (2.8) returns times $\tau_D < \Delta t$, even though probability statements for any two states with an even time lag are identical to probabilities of an AR(p) process with an AR-coefficient $|\alpha|$.

  Thus the number τ_D makes sense as a characteristic time scale when dealing with red noise processes. But for many higher order AR(p)-processes the number τ_D does not reflect a useful physical information.

- The Decorrelation Time depends on the time increment Δt: To demonstrate this dependency we consider again the AR(1)-process (2.3) with a time increment of $\Delta t = 1$ and $\alpha \geq 0$. Then we may construct other AR(1) processes with time increments k by noting that

  $$\mathbf{X}_t = \alpha^k \mathbf{X}_{t-k} + \mathbf{N}'_t \tag{2.9}$$

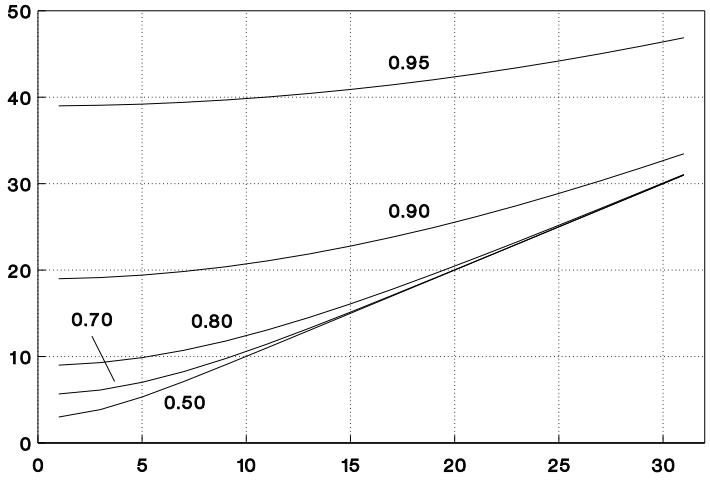
  with some noise term $\mathbf{N}'_t$ which is a function of $\mathbf{N}_t \ldots \mathbf{N}_{t-k+1}$. The decorrelation times τ_D of the two processes (2.3,2.9) are because of $\alpha < 1$:

  $$\tau_{D,1} = \frac{1+\alpha}{1-\alpha} \cdot 1 \geq 1 \quad \text{and} \quad \tau_{D,k} = \frac{1+\alpha^k}{1-\alpha^k} \cdot k \geq k \tag{2.10}$$

  so that

  $$\lim_{k\to\infty} \frac{\tau_{D,k}}{k} = 1 \tag{2.11}$$

Figure 2.3: *The dependency of the decorrelation time $\tau_{D,k}$ (2.10) on the time increment k (horizontal axis) and on the coefficient α (0.95, 0.90, 0.80, 0.70 and 0.50; see labels). (From von Storch and Zwiers, 1999).*



That means that the decorrelation time is as least as long as the time increment; in case of "white noise", with $\alpha = 0$, the decorrelation time is always equal to the time increment. In Figure 2.3 the dimensional decorrelation times are plotted for different α-values and different time increments k. The longer the time increment, the larger the decorrelation time. For sufficiently large time increments we have $\tau_{D,k} = k$. For small α-values, such as $\alpha = 0.5$, we have virtually $\tau_{D,k} = k$ already after $k = 5$. If $\alpha = 0.8$ then $\tau_{D,1} = 9$, $\tau_{D,11} = 13.1$ and $\tau_{D,21} = 21.4$. If the time increment is 1 day, then the decorrelation time of an $\alpha = 0.8$-process is 9 days or 21 days - if we sample the process once a day or once every 21 days.

We conclude that the absolute value of the decorrelation time is of questionable informational value. However, the relative values obtained from several time series sampled with the same time increment are useful to infer whether the system has in some components a longer memory than in others. If the decorrelation time is well above the time increment, as in case of the $\alpha = 0.95$-curve in Figure 2.3, then the number has some informational value whereas decorrelation times close to the time increment, as in case of the $\alpha = 0.5$-curve, are mostly useless.

We have seen that the name "Decorrelation Time" is not based on physical reasoning but on strictly mathematical grounds. Nevertheless the number is often incorrectly *interpreted* as the minimum time so that two consecutive observations $\mathbf{X}_t$ and $\mathbf{X}_{t+\tau_D}$ are independent. If used as a vague estimate with the reservations mentioned above, such a use is in order. However, the number is often introduced as crucial parameter in test routines. Probably the most frequent victim of this misuse is the conventional t-test.

We illustrate this case by a simple example from Zwiers and von Storch (1995): We want to answer the question whether the long-term mean winter temperatures in Hamburg and Victoria are equal. To answer this question, we have at our disposal daily observations for one winter from both locations. We treat the winter temperatures at both locations as random variables, say $\mathbf{T}_H$ and $\mathbf{T}_V$. The "long term mean" winter temperatures at the two locations, denoted as μ_H and μ_V respectively, are parameters of the probability distributions of these random variables. In the statistical nomenclature the question we pose is: do the samples of temperature observations contain sufficient evidence to reject the *null hypothesis* $H_0 : \mu_H - \mu_V = 0$.

The standard approach to this problem is to use to the *Student's t-test*. The test is conducted by imposing a statistical model upon the processes which resulted in the temperature samples and then, within the confines of this model, measuring the degree to which the data agree with H_0. An essential part of the model which is implicit in the t-test is the assumption that the data which enter the test represent a set of statistically independent observations. In our case, and in many other applications in climate research, this assumption is not satisfied. The Student's t-test usually becomes "liberal" in these circumstances. That is, it tends to reject that null hypothesis on weaker evidence than is implied by the significance level[5] which is specified for the test. One manifestation of this problem is that the Student's t-test will reject the null hypothesis more frequently than expected when the null hypothesis is true.

A relatively clean and simple solution to this problem is to form subsamples of approximately independent observations from the observations. In the case of daily temperature data, one might use physical insight to argue that observations which are, say, 5 days apart, are effectively independent of each other. If the number of samples, the sample means and standard deviations from these reduced data sets are denoted by n^*, $\tilde{\mathbf{T}}^*_H$, $\tilde{\mathbf{T}}^*_V$, $\tilde{\sigma}^*_H$ and $\tilde{\sigma}^*_V$ respectively, then the test statistic

$$t = \frac{\tilde{\mathbf{T}}^*_H - \tilde{\mathbf{T}}^*_V}{\sqrt{(\tilde{\sigma}^{*\,2}_H + \tilde{\sigma}^{*\,2}_V)/n^*}} \tag{2.12}$$

has a Student's t-distribution with n^* degrees of freedom provided that the null hypothesis is true[6] and a test can be conducted at the chosen signif-

[5]The significance level indicates the probability with which the null hypothesis will be rejected when it is true.

[6]Strictly speaking, this is true only if the standard deviations of $\mathbf{T}_H$ and $\mathbf{T}_V$ are equal.

icance level by comparing the value of (2.12) with the percentiles of the $t(n^*)$-distribution.

The advantage of (2.12) is that this test operates as specified by the user provided that the interval between successive observations is long enough. The disadvantage is that a reduced amount of data is utilized in the analysis. Therefore, the following concept was developed in the 1970s to overcome this disadvantage: The numerator in (2.12) is a random variable because it differs from one pair of temperature samples to the next. When the observations which comprise the samples are serially uncorrelated the denominator in (2.12) is an estimate of the standard deviation of the numerator and the ratio can be thought of as an expression of the difference of means in units of estimated standard deviations. For serially correlated data, with sample means $\tilde{\mathbf{T}}$ and sample standard deviations $\tilde{\sigma}$ derived from all available observations, the standard deviation of $\tilde{\mathbf{T}}_H - \tilde{\mathbf{T}}_V$ is $\sqrt{(\tilde{\sigma}_H^2 + \tilde{\sigma}_V^2)/n'}$ with the *equivalent sample size* n' as defined in (2.6). For sufficiently large samples sizes the ratio

$$t = \frac{\tilde{\mathbf{T}}_H - \tilde{\mathbf{T}}_V}{\sqrt{(\tilde{\sigma}_H^2 + \tilde{\sigma}_V^2)/n'}} \tag{2.13}$$

has a standard Gaussian distribution with zero mean and standard deviation one. Thus one can conduct a test by comparing (2.13) to the percentiles of the standard Gaussian distribution.

So far everything is fine.

Since $t(n')$ is approximately equal to the Gaussian distribution for $n' \geq 30$, one may compare the test statistic (2.13) also with the percentiles of the $t(n')$-distribution. The incorrect step is the heuristic assumption that this prescription - "compare with the percentiles of the $t(n')$, or $t(n'-1)$ distribution" - would be right for small ($n' < 30$) equivalent samples sizes. The rationale of doing so is the tacitly assumed fact that the statistic (2.13) would be $t(n')$ or $t(n'-1)$-distributed under the null hypothesis. However, *this assumption is simply wrong.* The distribution (2.13) is not $t(k)$-distributed for any k, be it the equivalent sample size n' or any other number. This result has been published by several authors (Katz (1982), Thiébaux and Zwiers (1984) and Zwiers and von Storch (1995)) but has stubbornly been ignored by most of the atmospheric sciences community.

A justification for the small sample size test would be that its behaviour under the null hypothesis is well approximated by the t-test with the equivalent sample size representing the degrees of freedom. But this is not so, as is demonstrated by the following example with an AR(1)-process (2.3) with $\alpha = .60$. The exact equivalent sample size $n' = \frac{1}{4}n$ is known for the process since its parameters are completely known. One hundred independent samples of variable length n were randomly generated. Each sample was used to test the null hypothesis $H_o : \mathrm{E}(\mathbf{X}_t) = 0$ with the t-statistic (2.13) at the 5% significance level. If the test operates correctly the null hypothesis should be (incorrectly) rejected 5% of the time. The actual rejection rate (Figure 2.4)

Sample Size n

is notably *smaller* than the expected rate of 5% for $4n' = n \leq 30$. Thus, the t-test operating with the true equivalent sample size is *conservative* and thus *wrong*.

More problems show up when the equivalent sample is unknown. In this case it may be possible to specify n' on the basis of physical reasoning. Assuming that conservative practices are used, this should result in underestimated values of n' and consequently even more conservative tests. In most applications, however, an attempt is made to estimate n' from the same data that are used to compute the sample mean and variance. Monte Carlo experiments show that the actual rejection rate of the t-test tends to be greater than the nominal rate when n' is estimated. Also this case has been simulated in a series of Monte Carlo experiments with the same AR(1)-process. The resulting rate of erroneous rejections is shown in Figure 2.4 - for small ratio sample sizes the actual significance level can be several times *greater* than the nominal significance level. Thus, the t-test operating with the estimated equivalent sample size is *liberal* and thus *wrong*.

Zwiers and von Storch (1995) offer a "table look-up" test as a useful alternative to the inadequate "t-test with equivalent sample size" for situations with serial correlations similar to red noise processes.

2.5 Use of Advanced Techniques

The following case is an educational example which demonstrates how easily an otherwise careful analysis can be damaged by an inconsistency hidden in a seemingly unimportant detail of an advanced technique. When people have experience with the advanced technique for a while then such errors are often found mainly by instinct ("This result cannot be true - I must have made an error.") - but when it is new then the researcher is somewhat defenseless against such errors.

The background of the present case was the search for evidence of bifurcations and other fingerprints of truly nonlinear behaviour of the dynamical system "atmosphere". Even though the nonlinearity of the dynamics of the planetary-scale atmospheric circulation was accepted as an obvious fact by the meteorological community, atmospheric scientists only began to discuss the possibility of two or more stable states in the late 1970's. If such multiple stable states exist, it should be possible to find bi- or multi-modal distributions in the observed data (if these states are well separated).

Hansen and Sutera (1986) identified a bi-modal distribution in a variable characterizing the energy of the planetary-scale waves in the Northern Hemisphere winter. Daily amplitudes for the zonal wavenumbers $k = 2$ to 4 for 500 hPa height were averaged for midlatitudes. A "wave-amplitude indicator" $\mathbf{Z}$ was finally obtained by subtracting the annual cycle and by filtering out all variability on time scales shorter than 5 days. The probability density function f_Z was estimated by applying a technique called the *maximum*

penalty technique to 16 winters of daily data. The resulting f_Z had two maxima separated by a minor minimum. This bimodality was taken as proof of the existence of two stable states of the atmospheric general circulation: A "zonal regime", with $\mathbf{Z} < 0$, exhibiting small amplitudes of the planetary waves and a "wavy regime", with $\mathbf{Z} > 0$, with amplified planetary-scale zonal disturbances.

Hansen and Sutera performed a "Monte Carlo" experiment to evaluate the likelihood of fitting a bimodal distribution to the data with the maximum penalty technique even if the generating distribution is unimodal. The authors concluded that this likelihood is small. On the basis of this statistical check, the found bimodality was taken for granted by many scientists for almost a decade.

When I read the paper, I had never heard about the "maximum penalty method" but had no doubts that everything would have been done properly in the analysis. The importance of the question prompted other scientists to perform the same analysis to further refine and verify the results. Nitsche et al. (1994) reanalysed step-by-step the same data set which had been used in the original analysis and came to the conclusion that the purportedly small probability for a misfit was *large*. The error in the original analysis was not at all obvious. Only by carefully scrutinizing the pitfalls of the maximum penalty technique did Nitsche and coworkers find the inconsistency between the Monte Carlo experiments and the analysis of the observational data.

Nitsche et al. reproduced the original estimation, but showed that something like 150 years of daily data would be required to exclude with sufficient certainty the possibility that the underlying distribution would be unimodal. What this boils down to is, that the null hypothesis according to which the distribution would be unimodal, is *not rejected by the available data* - and the published test was *wrong* . However, since the failure to reject the null hypothesis does not imply the acceptance of the null hypothesis (but merely the lack of *enough* evidence to reject it), the present situation is that the (alternative) hypothesis "The sample distribution does *not* originate from a unimodal distribution" is not falsified but still open for discussion.

I have learned the following rule to be useful when dealing with advanced methods: Such methods are often needed to find a signal in a vast noisy phase space, i.e., the needle in the haystack - but after having the needle in our hand, we should be able to identify the needle as a needle by simply looking at it.[7] *Whenever you are unable to do so there is a good chance that something is rotten in the analysis.*

[7] See again Wallace's and Gutzler's study who identified their teleconnection patterns first by examining correlation maps - and then by simple weighted means of few grid point values - see Section 12.1.

2.6 Epilogue

I have chosen the examples of this Chapter to advise users of statistical concepts to be aware of the sometimes hidden assumptions of these concepts. Statistical Analysis is not a *Wunderwaffe*[8] to extract a wealth of information from a limited sample of observations. More results require more assumptions, i.e., information given by theories and other insights *unrelated to the data under consideration.*

But, even if it is not a *Wunderwaffe* Statistical Analysis is an indispensable tool in the evaluation of limited empirical evidence. The results of Statistical Analysis are not miracle-like enlightenment but sound and understandable assessments of the consistency of concepts and data.

[8] Magic bullet.