Chapter 3
Bootstrap Confidence Intervals

Abstract In statistical analysis of climate time series, our aim (Chap. 1) is to estimate parameters of $X_{\text{trend}}(T), X_{\text{out}}(T), S(T)$ and $X_{\text{noise}}(T)$. Denote in general such a parameter as θ. An estimator, $\hat{\theta}$, is a recipe how to calculate θ from a set of data. The data, discretely sampled time series $\{t(i), x(i)\}_{i=1}^{n}$, are influenced by measurement and proxy errors of $x(i)$, outliers, dating errors of $t(i)$ and climatic noise. Therefore, $\hat{\theta}$ cannot be expected to equal θ. The accuracy of $\hat{\theta}$, how close it comes to θ, is described by statistical terms such as standard error, bias, mean squared error and confidence interval (CI). These are introduced in Sect. 3.1.

With the exploration of new archives or innovations in proxy, measurement and dating techniques, new $\hat{\theta}$ values, denoted as estimates, become available and eventually join or replace previous estimates. A telling example from geochronology is where θ is the time before present when the Earth's magnetic field changed from reversed polarity during the Matuyama epoch to normal polarity during the Brunhes epoch, at the beginning of the late Pleistocene. Estimates published over the past decades include 690 ka (Cox, Science 163(3864):237–245, 1969) and 730 ka (Mankinen and Dalrymple, J Geophys Res 84(B2):615–626, 1979), both based on K/Ar dating, and 790 ka (Johnson, Quat Res 17(2):135–147, 1982) and 780 ka (Shackleton et al., Trans R Soc Edinb Earth Sci 81(4):251–261, 1990), both based on astronomical tuning. The currently accepted value is 779 ka with a standard error of 2 ka (Singer and Pringle, Earth Planet Sci Lett 139(1–2):47–61, 1996), written as 779 ± 2 ka, based on $^{40}\text{Ar}/^{39}\text{Ar}$ dating (a high-precision variant of K/Ar dating). An example with a much greater uncertainty regards the case where θ is the radiative forcing (change in net vertical irradiance at the tropopause) of changes in atmospheric concentrations of mineral dust, where even the sign of θ is uncertain (Penner et al., Aerosols, their direct and indirect effects. In: Houghton et al. (eds) Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 289–348, 2001; Forster et al., Changes

M. Mudelsee, *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, Atmospheric and Oceanographic Sciences Library 51,
DOI 10.1007/978-3-319-04450-7_3,

in atmospheric constituents and in radiative forcing. In: Solomon et al. (eds) Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 129–234, 2007). It is evident that the growth of climatological knowledge depends critically on estimates of θ that are accompanied by error bars or other measures of their accuracy.

Bootstrap resampling (Sects. 3.2 and 3.3) is an approach to construct error bars and CIs. The idea is to draw random resamples from the data and calculate error bars and CIs from repeated estimations on the resamples. For climate time series, the bootstrap is potentially superior to the classical approach, which relies partly on unrealistic assumptions regarding distributional shape, persistence and spacing (Chap. 1). However, the bootstrap, developed originally for data without serial dependence, has to be adapted before applying it to time series. Two classes of adaptions exist for taking persistence into account. First, nonparametric bootstrap methods resample sequences, or blocks, of the data. They preserve the dependence structure over the length of a block. Second, the parametric bootstrap adopts a dependence model. As such, the AR(1) model (Chap. 2) is our favourite.

It turns out that both bootstrap resampling types have the potential to yield acceptably accurate CIs for estimated climate parameters. A problem for the block bootstrap arises from uneven time spacing. Another difficult point is to find optimal block lengths. This could make the parametric bootstrap superior within the context of this book, especially for small data sizes (less than, say, 50). The block bootstrap, however, is important when the deviations from AR(1) persistence seem to be strong. Various CI types are investigated. We prefer a version (so-called BCa interval) that automatically corrects for estimation bias and scale effects. Computing-intensive calibration techniques can further increase the accuracy.

Keywords Error bar • Confidence interval • Standard error • Standard deviation • Expectation value • Root mean squared error • Coefficient of variation • Bias • Monte Carlo experiment

3.1 Error Bars and Confidence Intervals

Let θ be the parameter of interest of the climatic process $\{X(T)\}$ and $\hat{\theta}$ be the estimator. Extension to a set of parameters is straightforward. Any meaningful construction lets the estimator be a function of the process, $\hat{\theta} = g(\{X(T)\})$. That means $\hat{\theta}$ is a random variable with statistical properties. The standard deviation of $\hat{\theta}$, denoted as standard error, is

$$\mathrm{se}_{\hat{\theta}} = \left[VAR\left(\hat{\theta}\right)\right]^{1/2}. \tag{3.1}$$

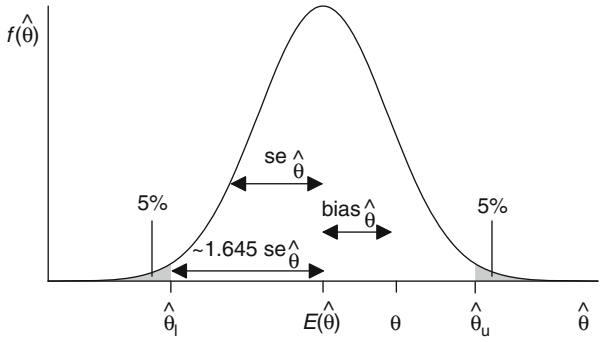


Fig. 3.1 Standard error ($\mathrm{se}_{\hat{\theta}}$), bias ($\mathrm{bias}_{\hat{\theta}}$) and equi-tailed confidence interval ($\mathrm{CI}_{\hat{\theta},1-2\alpha} = [\hat{\theta}_\mathrm{l}; \hat{\theta}_\mathrm{u}]$) for a Gaussian distributed estimator, $\hat{\theta}$. The true parameter value is θ; the confidence level is $1 - 2\alpha = 90\,\%$

The bias of $\hat{\theta}$ is

$$\mathrm{bias}_{\hat{\theta}} = E\left(\hat{\theta}\right) - \theta. \tag{3.2}$$

$\mathrm{bias}_{\hat{\theta}} > 0$ ($\mathrm{bias}_{\hat{\theta}} < 0$) means a systematic overestimation (underestimation). $\mathrm{se}_{\hat{\theta}}$ and $\mathrm{bias}_{\hat{\theta}}$ are illustrated in Fig. 3.1. Desirable estimators have small $\mathrm{se}_{\hat{\theta}}$ and small $\mathrm{bias}_{\hat{\theta}}$. In many estimations, a trade-off problem between $\mathrm{se}_{\hat{\theta}}$ and $\mathrm{bias}_{\hat{\theta}}$ occurs. A convenient measure is the root mean squared error:

$$\begin{aligned} \mathrm{RMSE}_{\hat{\theta}} &= \left\{ E\left[\left(\hat{\theta} - \theta \right)^2 \right] \right\}^{1/2} \\ &= \left(\mathrm{se}_{\hat{\theta}}{}^2 + \mathrm{bias}_{\hat{\theta}}{}^2 \right)^{1/2}. \end{aligned} \tag{3.3}$$

The coefficient of variation is

$$\mathrm{CV}_{\hat{\theta}} = \mathrm{se}_{\hat{\theta}} \Big/ \left| E\left(\hat{\theta}\right) \right|. \tag{3.4}$$

While $\hat{\theta}$ is a best guess of θ or a point estimate, a CI is an interval estimate that informs how good a guess is (Fig. 3.1). The CI for θ is

$$\mathrm{CI}_{\hat{\theta},1-2\alpha} = \left[\hat{\theta}_\mathrm{l}; \hat{\theta}_\mathrm{u} \right], \tag{3.5}$$

where $0 \leq 1 - 2\alpha \leq 1$ is a prescribed value, denoted as confidence level. The practical examples in this book consider 90 % ($\alpha = 0.05$) or 95 % ($\alpha = 0.025$) CIs, which are reasonable choices for climatological problems. $\hat{\theta}_\mathrm{l}$ is the lower and

$\hat{\theta}_\mathrm{u}$ the upper endpoint of the CI. $\hat{\theta}_\mathrm{l}$ and $\hat{\theta}_\mathrm{u}$ are random variables and have statistical properties such as standard error or bias. The properties of interest for CIs are the coverages

$$\gamma_\mathrm{l} = \mathrm{prob}\left(\theta \leq \hat{\theta}_\mathrm{l}\right), \tag{3.6}$$

$$\gamma_\mathrm{u} = \mathrm{prob}\left(\theta \geq \hat{\theta}_\mathrm{u}\right) \tag{3.7}$$

and

$$\gamma = \mathrm{prob}\left(\hat{\theta}_\mathrm{l} < \theta < \hat{\theta}_\mathrm{u}\right) = 1 - \gamma_\mathrm{l} - \gamma_\mathrm{u}. \tag{3.8}$$

Exact CIs have coverages, γ, equal to the nominal value $1 - 2\alpha$. Construction of exact CIs requires knowledge of the distribution of $\hat{\theta}$, which can be achieved only for simple problems. In more complex situations, only approximate CIs can be constructed (Sect. 3.1.3). As regards the division of the nominal coverage between the CI endpoints, this book adopts a practical approach and considers only equi-tailed CIs, where nominally $\gamma_\mathrm{l} = \gamma_\mathrm{u} = \alpha$. As a second CI property besides coverage, we consider interval length, $\hat{\theta}_\mathrm{u} - \hat{\theta}_\mathrm{l}$, which is ideally small.

Preceding paragraphs considered estimators on the process level. In practice, on the sample level, we plug in the data $\{t(i), x(i)\}_{i=1}^n$ for $\{T(i), X(i)\}_{i=1}^n$. Following the usual convention, we denote also the estimator on the sample level as $\hat{\theta}$. An example is the autocorrelation estimator (Eq. 2.4).

3.1.1 *Theoretical Example: Mean Estimation of Gaussian White Noise*

Let the process $\{X(i)\}_{i=1}^n$ be given by

$$X(i) = \mathcal{E}_{\mathrm{N}(\mu,\sigma^2)}(i), \qquad i = 1, \ldots, n, \tag{3.9}$$

which is called a Gaussian purely random process or Gaussian white noise. There is no serial dependence, and the times $T(i)$ are not of interest. Consider as estimator $\hat{\theta}$ of the mean, μ, the sample mean, written on process level as

$$\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X(i)/n. \tag{3.10}$$

Let also σ be unknown and estimated by the sample standard deviation, $\hat{\sigma} = S_{n-1}$, given in the next example (Eq. 3.19). The properties of $\bar{X}$ readily follow as

$$\mathrm{se}_{\bar{X}} = \sigma \cdot n^{-1/2}, \tag{3.11}$$

$$\mathrm{bias}_{\bar{X}} = 0, \tag{3.12}$$

$$\mathrm{RMSE}_{\bar{X}} = \mathrm{se}_{\bar{X}} \tag{3.13}$$

and

$$\mathrm{CV}_{\bar{X}} = \sigma \cdot n^{-1/2} \cdot \mu^{-1}. \tag{3.14}$$

An exact CI of level $1 - 2\alpha$ can be constructed by means of the Student's t distribution of $\bar{X}$ (von Storch and Zwiers 1999):

$$\mathrm{CI}_{\bar{X},1-2\alpha} = \left[\bar{X} + t_{n-1}(\alpha) \cdot S_{n-1} \cdot n^{-1/2}; \bar{X} + t_{n-1}(1-\alpha) \cdot S_{n-1} \cdot n^{-1/2}\right]. \tag{3.15}$$

$t_\nu(\beta)$ is the percentage point at β of the t distribution function with ν degrees of freedom (Sect. 3.9).

On the sample level, we write the estimated sample mean,

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x(i)/n, \tag{3.16}$$

the estimated standard error,

$$\widehat{\mathrm{se}}_{\bar{x}} = \left\{\sum_{i=1}^{n} \left[x(i) - \bar{x}\right]^2 / n^2\right\}^{1/2}, \tag{3.17}$$

and the confidence interval,

$$\mathrm{CI}_{\bar{x},1-2\alpha} = \left[\bar{x} + t_{n-1}(\alpha) \cdot s_{n-1} \cdot n^{-1/2}; \bar{x} + t_{n-1}(1-\alpha) \cdot s_{n-1} \cdot n^{-1/2}\right], \tag{3.18}$$

where s_{n-1} is given by Eq. (3.25).

The performance of the CI in Eq. (3.18) for Gaussian white noise is analysed by means of a Monte Carlo simulation experiment. The CI performs excellent in coverage (Table 3.1), as expected from its exactness. The second CI property, length, decreases with data size. It can be further compared with CI lengths for other location measures.

Table 3.1 Monte Carlo experiment, mean estimation of a Gaussian purely random process. $n_{\text{sim}} = 4{,}750{,}000$ random samples of $\{X(i)\}_{i=1}^{n}$ were generated after Eq. (3.9) with $\mu = 1.0, \sigma = 2.0$ and various n values. An exact confidence interval $\text{CI}_{\bar{x},1-2\alpha}$ was constructed for each simulation after Eq. (3.18) with $\alpha = 0.025$. Average CI length, empirical $\text{RMSE}_{\bar{X}}$ and empirical coverage were determined subsequently. The entries are rounded

n	$\text{RMSE}_{\bar{x}}^{\text{a}}$	Nominal[b]	$\langle\text{CI length}\rangle^{\text{c}}$	Nominal[d]	$\gamma_{\bar{x}}^{\text{e}}$	Nominal
10	0.6327	0.6325	2.7832	2.7832	0.9499	0.9500
20	0.4474	0.4472	1.8476	1.8476	0.9498	0.9500
50	0.2828	0.2828	1.1310	1.1310	0.9501	0.9500
100	0.2000	0.2000	0.7916	0.7917	0.9499	0.9500
200	0.1415	0.1414	0.5570	0.5571	0.9499	0.9500
500	0.0894	0.0894	0.3513	0.3513	0.9500	0.9500
1000	0.0633	0.0632	0.2482	0.2482	0.9499	0.9500

[a]Empirical $\text{RMSE}_{\bar{X}}$, given by $\left[\sum_{i=1}^{n_{\text{sim}}} (\bar{x} - \mu)^2 / n_{\text{sim}}\right]^{1/2}$

[b]$\sigma \cdot n^{-1/2}$

[c]Average value over n_{sim} simulations

[d] $2 \cdot t_{n-1}(1-\alpha) \cdot \sigma \cdot c \cdot n^{-1/2}$, where c is given by Eq. (3.24)

[e]Empirical coverage, given by the number of simulations where $\text{CI}_{\bar{x},1-2\alpha}$ contains μ, divided by n_{sim}. Standard error of $\gamma_{\bar{x}}$ is (Efron and Tibshirani 1993) nominally $[2\alpha(1-2\alpha)/n_{\text{sim}}]^{1/2} = 0.0001$

3.1.2 *Theoretical Example: Standard Deviation Estimation of Gaussian White Noise*

Consider the Gaussian white-noise process (Eq. 3.9) with unknown mean and as estimator of σ the sample standard deviation, written on process level as

$$\hat{\sigma} = S_{n-1} = \left\{ \sum_{i=1}^{n} \left[X(i) - \bar{X}\right]^2 / (n-1) \right\}^{1/2}. \tag{3.19}$$

The properties of S_{n-1} are as follows:

$$\text{se}_{S_{n-1}} = \sigma \cdot \left(1 - c^2\right)^{1/2}, \tag{3.20}$$

$$\text{bias}_{S_{n-1}} = \sigma \cdot (c - 1), \tag{3.21}$$

$$\text{RMSE}_{S_{n-1}} = \sigma \cdot [2(1-c)]^{1/2} \tag{3.22}$$

and

$$\text{CV}_{S_{n-1}} = \left(1/c^2 - 1\right)^{1/2}, \tag{3.23}$$

where

$$c = [2/(n-1)]^{1/2} \cdot \Gamma(n/2) \,/\, \Gamma((n-1)/2). \tag{3.24}$$

Table 3.2 Monte Carlo experiment, standard deviation estimation of a Gaussian purely random process. $n_{\mathrm{sim}} = 4{,}750{,}000$ random samples of $\{X(i)\}_{i=1}^{n}$ were generated after Eq. (3.9) with $\mu = 1.0, \sigma = 2.0$ and various n values. An exact confidence interval $\mathrm{CI}_{s_{n-1},1-2\alpha}$ was constructed for each simulation after Eq. (3.26) with $\alpha = 0.025$. Average CI length, empirical $\mathrm{RMSE}_{S_{n-1}}$ and empirical coverage were determined subsequently

n	$\mathrm{RMSE}^{\mathrm{a}}_{s_{n-1}}$	*Nominal*[b]	$\langle CI\ length \rangle^{\mathrm{c}}$	*Nominal*[d]	$\gamma^{\mathrm{e}}_{s_{n-1}}$	*Nominal*
10	0.4677	0.4677	2.2133	2.2133	0.9500	0.9500
20	0.3232	0.3233	1.3818	1.3819	0.9500	0.9500
50	0.2018	0.2018	0.8174	0.8174	0.9499	0.9500
100	0.1421	0.1420	0.5659	0.5659	0.9499	0.9500
200	0.1002	0.1002	0.3960	0.3960	0.9500	0.9500
500	0.0633	0.0633	0.2489	0.2489	0.9500	0.9500
1000	0.0447	0.0447	0.1757	0.1757	0.9501	0.9500

[a]Empirical $\mathrm{RMSE}_{S_{n-1}}$, given by $\left[\sum_{i=1}^{n_{\mathrm{sim}}} (s_{n-1} - \sigma)^2 / n_{\mathrm{sim}}\right]^{1/2}$

[b]$\sigma \cdot [2(1-c)]^{1/2}$

[c]Average value over n_{sim} simulations

[d]$\left[\left(\chi^2_{n-1}(1-\alpha)\right)^{-1/2} - \left(\chi^2_{n-1}(\alpha)\right)^{-1/2}\right] \cdot \sigma \cdot c \cdot (n-1)^{1/2}$

[e]Empirical coverage, given by the number of simulations where $\mathrm{CI}_{s_{n-1},1-2\alpha}$ contains σ, divided by n_{sim}. Standard error of $\gamma_{s_{n-1}}$ is nominally $[2\alpha(1-2\alpha)/n_{\mathrm{sim}}]^{1/2} = 0.0001$

On the sample level, we write

$$\hat{\sigma} = s_{n-1} = \left\{ \sum_{i=1}^{n} [x(i) - \bar{x}]^2 / (n-1) \right\}^{1/2} \tag{3.25}$$

and use the chi-squared distribution of S^2_{n-1} (von Storch and Zwiers 1999) to find

$$\mathrm{CI}_{s_{n-1},1-2\alpha} = \left[s_{n-1} \left[(n-1) \,/\, \chi^2_{n-1}(\alpha) \right]^{1/2} ; \right.$$
$$\left. s_{n-1} \left[(n-1) \,/\, \chi^2_{n-1}(1-\alpha) \right]^{1/2} \right], \tag{3.26}$$

where $\chi^2_\nu(\beta)$ is the percentage point at β of the chi-squared distribution function with ν degrees of freedom (Sect. 3.9).

The performance of the CI in Eq. (3.26) for Gaussian white noise is analysed by means of a Monte Carlo simulation experiment. The CI performs excellent in coverage (Table 3.2), as expected from its exactness. The CI property length can be compared with CI lengths for other measures of spread or variation.

3.1.3 *Real World*

The two theoretical examples (Sects. 3.1.1 and 3.1.2) presented convenient settings. $X(i)$ was normally distributed and persistence was absent; for the latter reason the

Table 3.3 Monte Carlo experiment, mean and median estimation of a lognormal purely random process. $n_{\text{sim}} = 4{,}750{,}000$ random samples of $\{X(i)\}_{i=1}^{n}$ were generated after $X(i) = \exp\left[\mathcal{E}_{\text{N}(\mu,\sigma^2)}(i)\right]$, $i = 1,\ldots,n$, with $\mu = 1.0, \sigma = 1.0$ and various n values. The density function is skewed (Fig. 3.2). Analysed as estimators of the centre of location of the distribution were the sample mean (Eq. 3.16) and the sample median, $\hat{m}$ (see background material, Sect. 3.8). $\text{CI}_{\bar{x},1-2\alpha}$ was constructed after Eq. (3.18) with $\alpha = 0.025$

n	$\text{RMSE}_{\hat{m}}$	$\text{RMSE}_{\bar{x}}$	$\gamma_{\bar{x}}^{\text{a}}$	*Nominal*	C^{b}
10	1.1647	1.8575	0.8392	0.9500	−0.1108
20	0.7893	1.3140	0.8670	0.9500	−0.0830
50	0.4884	0.8309	0.8991	0.9500	−0.0509
100	0.3430	0.5880	0.9170	0.9500	−0.0330
200	0.2418	0.4155	0.9296	0.9500	−0.0204
500	0.1526	0.2627	0.9399	0.9500	−0.0101
1000	0.1078	0.1858	0.9442	0.9500	−0.0058

[a]Standard error of $\gamma_{\bar{x}}$ is nominally 0.0001
[b]Empirical coverage error of $\text{CI}_{\bar{x},1-2\alpha}$, given by $\gamma_{\bar{x}}$ minus nominal value

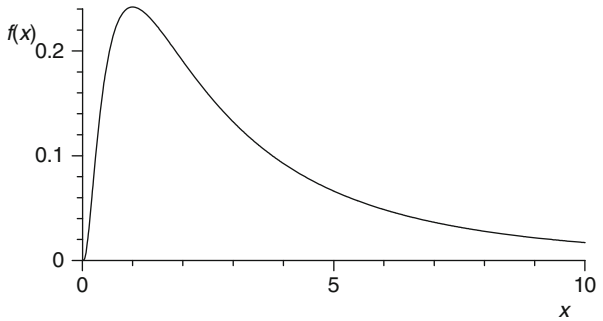


Fig. 3.2 Lognormal density function from Example 3 (Table 3.3), with $\mu = 1.0$ and $\sigma = 1.0$. The expression for $f(x)$ is given by Eq. (3.64)

spacing was not relevant. The simple estimators $\hat{\mu}$ and $\hat{\sigma}$ could then be applied for mean and standard deviation estimation, which allowed to deduce their distributions as Student's t and chi-squared, respectively. Finally, exact CIs were obtained using the percentage points of the distributions of the estimators.

In the real climatological world, however, such simple assumptions regarding distributional shape, persistence and spacing cannot be expected to be fulfilled (Chap. 1). In the practical setting, further questions than just after mean and standard deviation are asked, leading to more complex parameters, θ. The major part of the rest of this book is devoted to such problems. Also the estimators of those parameters have commonly more complex distributions, $f(\hat{\theta})$.

Example 3 (Table 3.3) goes a small step from the theoretical in the direction of the real world. This case illustrates the effects of violations of the distributional assumption. Example 3 assumes that $X(i)$ are Gaussian distributed, although the prescribed true distribution is lognormal. This leads to a Student's t CI with an

empirical coverage that deviates from the nominal value by several standard errors (Table 3.3). The difference is the coverage error (see next paragraph); its absolute value decreases with the data size. This CI is not exact but only approximate. Table 3.4 summarizes theoretical and practical settings.

Coverage error, C, is defined by means of a single-sided CI endpoint (Efron and Tibshirani 1993), for example,

$$C = \gamma_1 - \alpha. \tag{3.27}$$

If C decreases with sample size as $\mathcal{O}\left(n^{-1/2}\right)$, that is, if C is composed of terms of powers of $1/n$ that are greater than or equal to $1/2$, then the CI is called first-order accurate; if C is of $\mathcal{O}\left(n^{-1}\right)$, then the CI is called second-order accurate; and so forth. The same CI accuracy applies also to two-sided CIs. Desirable approximate CIs have a high-order accuracy. Coverage accuracy is the major criterion employed in this book for assessing the quality of a CI. As a second property, we consider interval length, $\hat{\theta}_\mathrm{u} - \hat{\theta}_\mathrm{l}$, which is ideally small. Related to CI accuracy is CI correctness (Efron and Tibshirani 1993: Sect. 22.2 therein), which refers to the difference between an exact CI endpoint (which has $C = 0$) and an approximate CI endpoint, expanded in terms of powers of n.

For practical situations it is conceivable that different estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the same parameter, θ, exist. Consider, for example, parameter estimation of the AR(p) model, for which Priestley (1981: Sect. 5.4.1 therein) gives four sets of estimators, namely, exact likelihood, least squares, approximate least squares and Yule–Walker. Each estimator has its own properties such as standard error, bias, RMSE, CI length or CI coverage accuracy.

An important attribute of an estimator is robustness, which means that the $\hat{\theta}$ properties depend only weakly on made assumptions (shape, persistence and spacing). Robust estimators perform better (e.g. have smaller RMSE or higher coverage accuracy) than non-robust in nonideal situations. Example 3 shows that the sample median as an estimator of the centre of location of a distribution is more robust (with regard to $\mathrm{RMSE}_{\hat{\theta}}$) than the mean. In essence, because of the complexity of the setting in the real world and the dependence on the situation and the aims of the analysis, there is no general rule how to construct best an estimator. It has something of an art, which is not meant negatively. In this light, the growth of climatological knowledge does not only depend on more and better data but also on improved methods to analyse them.

Table 3.4 shows also how real-world climatological estimation problems may be tackled. The classical approach comes from theory and aims to extend the applicability by introducing countermeasures. Regarding distributional shape, a measure may be to estimate the shape of the noise data (Sect. 1.6). Then one looks and applies the estimator for the parameter θ that performs for this particular shape best in terms of a user-specified property, say RMSE. The CI follows from the estimator's distribution. The problem is that only for simple shapes and parameters, knowledge is available that would allow this procedure. (In this regard,

Table 3.4 Estimation settings (theoretical and practical) and approaches (classical and bootstrap) to solve practical problems[a]

	Distributional shape	Persistence	Spacing	Estimator, $\hat{\theta}$	Distribution of $\hat{\theta}$, $f(\hat{\theta})$	Confidence interval, $\mathrm{CI}_{\hat{\theta},1-2\alpha}$
Setting						
Theoretical[b]	Known, normal	No (yes)	Not relevant (even)	Tractable	Deducible	Exact
Example 1	Normal	No	Not relevant	$\hat{\mu}$ (Eq. 3.10)	t distribution	Exact
Example 2	Normal	No	Not relevant	$\hat{\sigma}$ (Eq. 3.19)	χ^2 distribution	Exact
Practical	Nonnormal	Yes	Uneven	More complex than $\hat{\mu}$ or $\hat{\sigma}$	Often not deducible	Exact only if $f(\hat{\theta})$ deducible
Example 3	Lognormal	No	Not relevant	$\hat{\mu}$ (Eq. 3.19), $\hat{m}$	Ignored	Student's t approximation
Approach						
Classical	Find shape, apply suitable $\hat{\theta}$ or transform x	Effective data size	Ignore		Assume normality	Approximate, based on assumptions
Bootstrap	Not very relevant[c]	Block bootstrap or parametric	Not relevant[d]		Not very relevant[c]	Approximate, based on fewer assumptions

[a]Indicated are the main lines of settings and approaches. Exceptions exist; for example, theory deals also with uneven spacing (Parzen 1984)
[b]Theory must necessarily impose restrictions to shape, persistence and spacing to obtain tractable problems
[c]Distributional properties can influence bootstrap CI accuracy (see text), but this is a minor effect[a]
[d]Restriction: parametric persistence models more complex than AR(1) are not considered for uneven spacing because of the embedding problem (Sect. 2.3)

the lognormal without is clearly simpler than the lognormal with shift parameter (Sect. 3.8).) Transformations of the data, such that the noise part has a simple shape, can also be tried, but then the problem is that the systematic part of the model (Eq. 1.2) can take intractable forms; see Atkinson and Cox (1988) on this dilemma. (The double-logarithmic transformation described in Sect. 2.6 was in the converse direction. It produced a simpler systematic part and a more complex noise part.)

Regarding persistence, the effective data size, n', can be used instead of n for CI calculation. The problem here is that n' depends on the persistence model and on which estimator is used (Chap. 2). One may take n'_{μ} (Eq. 2.7), n'_{σ^2} (Eq. 2.36) or n'_{ρ} (Eq. 2.38) for the AR(1) process and hope that deviations to the problem at hand are small. Regarding spacing, it is fair to say that the classical approach mostly ignores unevenness because its influence on n' and the distribution of $\hat{\theta}$ can in the general case not be deduced. As a result, the classical approach often contents itself with approximate normality, that is, with $f(\hat{\theta})$ approaching normal shape as $n \to \infty$. For many theoretical estimations, approximate normality can be proven. However, the point is that in practice n is limited and it is mostly unknown how accurate the normal approximation of the CI is.

3.2 Bootstrap Principle

Table 3.4 lists also the bootstrap approach to solve practical estimation problems. These tasks include constructing CIs for estimators more complex than the mean, and this in the presence of nonnormal distributions, persistence and uneven spacing. The main idea behind the bootstrap is to use the data to mimic the unknown distribution function, which is now replaced by the empirical distribution function (Eq. 3.46). Mimicking the data generating process is achieved by drawing random samples from the data set. The simplest form is the ordinary bootstrap, that is, drawing one by one with replacement. Preserving the persistence properties of time series data requires adaptions of the ordinary bootstrap, which are explained in Sect. 3.3. Reapplying the estimation procedure to the new random samples, called resamples, yields new estimates, called replications. Section 3.4 explains CI construction using the replications. Figure 3.3 shows the bootstrap principle and the workflow. It gives also a simple bootstrap CI variant (bootstrap normal CI).

The bootstrap means that numerical simulation replaces theoretical derivation of the distribution of an estimator. This can be an improvement, especially if the complexity of the problem defies obtaining an exact theoretical result. However, also the bootstrap is not free of assumptions. The main requirement is that the properties distributional shape and persistence are preserved by the bootstrap resampling. There is also "simulation noise", but this can be made arbitrarily small by using a large number of resamples, B. Assumptions made at CI construction add to the fact that in complex situations, bootstrap CIs, like classical CIs, are not exact but approximate. In complex cases, for small sample size, non-smooth functionals such

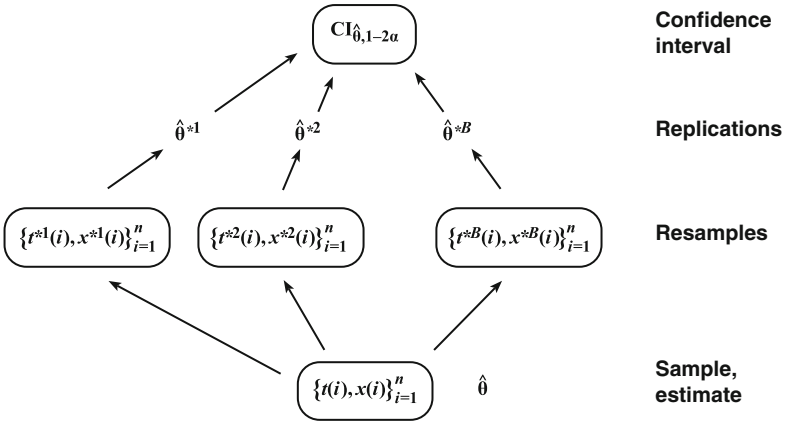


Fig. 3.3 Bootstrap principle for constructing confidence intervals. Given is a sample of data and an estimate of a parameter of interest. Using bootstrap resampling (Sect. 3.3), new data sets—resamples—are formed. The resamples ideally preserve fully the statistical properties of the process that generated the data. For convenience of presentation, we assume that this process (Eq. 1.2) consists only of the noise part; the following chapters analyse bootstrap resampling where the model has also a systematic part. In the simple case where $t(i)$ are perfectly known and also persistence is absent, $t^*(i) = t(i), i = 1, \ldots, n$, and $\{x^*(i)\}_{i=1}^{n}$ is obtained by drawing randomly, one by one and with replacement, n elements from the set of sample values, $\{x(i)\}_{i=1}^{n}$. The resamples are marked with an *asterisk* and numbered with an index, $b = 1, \ldots, B$. The number of resamples, B, is typically a few thousand. The estimator is applied to each of the resamples, yielding B new estimates—the replications. The set of replications $\{\hat{\theta}^{*b}\}_{b=1}^{B}$ is then used for CI construction. Several methods exist for that purpose (Sect. 3.4), which can, for example, correct for estimation bias. In the simple case of normal bootstrap confidence intervals, henceforth denoted briefly as normal CIs, $CI_{\hat{\theta},1-2\alpha} = [\hat{\theta} + z(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}^*}; \hat{\theta} - z(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}^*}]$, where $\widehat{\mathrm{se}}_{\hat{\theta}^*}$ is the sample standard error of the replications, denoted as estimated bootstrap standard error, and $z(\alpha)$ is the percentage point of the normal distribution (Sect. 3.9)

as the median and without underlying theory, even the bootstrap may fail to yield acceptable results (LePage and Billard 1992). However, bootstrap CIs seem to be more flexible and require less strict assumptions than classical CIs (Table 3.4). A word on usage of "simulation": henceforth we reserve this for Monte Carlo experiments, where statistical methods are tested by means of artificial data from models with predefined properties. The bootstrap procedure, on the other hand, is referred to as "resampling".

3.3 Bootstrap Resampling

The ordinary bootstrap, resampling one by one with replacement, is a nonparametric method because it can virtually be applied to data from any continuous PDF without involvement of distributional parameters. By resampling one by one, the

serial dependence in $\{X(i)\}_{i=1}^{n}$ is lost. For the analysis of time series, the ordinary bootstrap has therefore to be adapted to take serial dependence into account. This can be done nonparametrically, by resampling block by block of data. Alternatively, persistence can be modelled. The preferred model in the case of climate time series is the AR(1) process (Chap. 2).

For convenience of presentation, this chapter omits the effects of errors in the timescale, $t(i)$, that is, it sets $t^*(i) = t(i), i = 1, \ldots, n$, or briefly $\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$. Bootstrap adaptions for solving estimation problems associated with an uncertain timescale, which are relevant for climatology, seem not to have been developed yet in the statistical literature. The subsequent chapters present some possible bootstrap adaptions. These are steps into new territory.

3.3.1 Nonparametric: Moving Block Bootstrap

The moving block bootstrap algorithm, denoted as MBB, divides the time series values $\{x(i)\}_{i=1}^{n}$ into sequences or blocks of l consecutive points (Algorithm 3.1). The blocks may overlap and their number is $n - l + 1$. MBB draws randomly a block and inserts the contained values as the first l resample values, $\{x^*(i)\}_{i=1}^{l}$. The following randomly drawn block yields $\{x^*(i)\}_{i=l+1}^{2l}$ and so forth. When the last point, $x^*(n)$, has been inserted, the algorithm stops; remaining block values are discarded. The resampled times are unchanged (Algorithm 3.1). One indexes the first resample as $\left\{t^{*1}(i), x^{*1}(i)\right\}_{i=1}^{n}$ and repeats MBB until B resamples exist.

A possible adaption of the MBB to uneven spacing is introduced later in this section. Other nonparametric bootstrap algorithms are described briefly in the background material (Sect. 3.8).

Block Length Selection

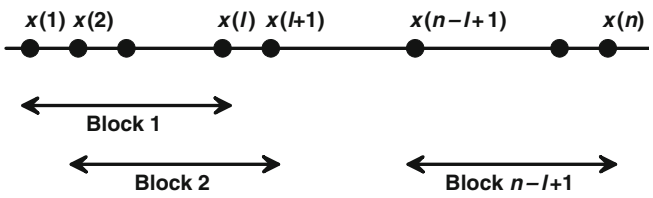
Selection of the block length, l, is a crucial step because it determines properties like bootstrap standard error or bootstrap CI coverage accuracy. Berkowitz and Kilian (2000: p. 20 therein) describe the trade-off problem involved as follows:

> As the block size becomes too small, the [MBB] destroys the time dependency of the data and its average accuracy will decline. As the block size becomes too large, there are few blocks and [resamples] will tend to look alike. As a result, the average accuracy of the [MBB] also will decline. This suggests that there exists an optimal block size [l_{opt}] which maximizes accuracy.

A simple block length selector can be derived from Sherman et al. (1998), who adapted a formula from Carlstein (1986), to the MBB:

$$l_{\text{opt}} = NINT\left\{\left[6^{1/2} \cdot \hat{\tilde{a}} \Big/ \left(1 - \hat{\tilde{a}}^2\right)\right]^{2/3} \cdot n^{1/3}\right\}, \tag{3.28}$$

Algorithm 3.1 Moving block bootstrap algorithm (MBB). Note: An equation such as $\{t^*(i)\}_{i=1}^n = \{t(i)\}_{i=1}^n$ is used to denote $t^*(i) = t(i), i = 1, \ldots, n$



Step 1	Data	$\{t(i), x(i)\}_{i=1}^n$
Step 2	Resampled times unchanged	$\{t^*(i)\}_{i=1}^n = \{t(i)\}_{i=1}^n$
Step 3	Blocks j (see above)	$\{x(i)\}_{i=j}^{j+l-1}, \quad j = 1, \ldots, n-l+1$
Step 4	Set counter	$c = 1$
Start resampling		
Step 5	Draw random block j^*	$j^* \in \{1, \ldots, n-l+1\}$
Step 6	Insert block data	$\{x^*(i)\}_{i=c}^{c+l-1} = \{x(i)\}_{i=j^*}^{j^*+l-1}$
	If $x^*(n)$ has been inserted	Stop inserting and exit
Step 7	Increase counter	$c \to c+l$
Step 8	Go to Step 5	
End resampling		

where $NINT(\cdot)$ is the nearest integer function and $\hat{\tilde{a}} = \exp(-\bar{d}/\hat{\tau})$ is the estimated "equivalent autocorrelation coefficient" (Fig. 2.3) of an AR(1) process fitted to the data with uneven spacing. (If $\hat{\tilde{a}} \to 0$ and $\hat{\tilde{a}} \to 1$, then take $l_{\text{opt}} = 1$ and $l_{\text{opt}} = n-1$, respectively.) In the case of even spacing, $\hat{\tilde{a}}$ can be taken from Eq. (2.4). Instead of $\hat{\tilde{a}}$, also a bias-corrected version, $\hat{\tilde{a}}'$, can be used; see Sect. 2.6. Employing this block length selector for real-world problems is evidently a simplification because it was developed for normal shape, AR(1) persistence, even spacing and bootstrap standard error estimation. Hall et al. (1995a) show that for bootstrap CI estimation, l_{opt} should increase at a slower rate with n. On the other hand, in practice some simplification is inevitable, and the formula may yield acceptable results. This can be assessed by means of Monte Carlo simulations of real-world conditions, as is done in subsequent parts of this book.

Bühlmann and Künsch (1999) presented a fully data-driven block length selector (Algorithm 3.2). They showed the equivalence of l_{opt} selection and smoothing in spectral estimation (Chap. 5).

Berkowitz and Kilian (2000) presented a brute-force block length selector:

1. Approximate the data generating process by a parametric model (e.g. ARMA).
2. Generate Monte Carlo samples from this fitted model.

Algorithm 3.2 Block length selector after Bühlmann and Künsch (1999). Notes: $\widehat{\mathrm{IF}}(X(i))$ is the estimated influence function (Efron and Tibshirani 1993: Sect. 21.3 therein). $\hat{\theta}_{(j)}$ is the delete-one, jackknife value of $\hat{\theta}$, that is, the $\hat{\theta}$ value calculated from the data with the jth point removed; see Sect. 3.4.4. w_{SC} is the split-cosine window; $w_{\mathrm{SC}}(z) = 1$ for $|z| \leq 0.8$, $w_{\mathrm{SC}}(z) = [1 + \cos(5(z - 0.8)\pi)]/2$ for $0.8 < |z| \leq 1$ and $w_{\mathrm{SC}}(z) = 0$ for $|z| > 1$. w_{TH} is the Tukey–Hanning window; $w_{\mathrm{TH}}(z) = [1 + \cos(\pi z)]/2$ for $|z| \leq 1$ and $w_{\mathrm{TH}}(z) = 0$ for $|z| > 1$

Step 1 Calculate $\{Y(i)\}_{i=1}^{n} = \left\{\widehat{\mathrm{IF}}(X(i))\right\}_{i=1}^{n}$, where

$$\widehat{\mathrm{IF}}(X(j)) = n \cdot \left(\hat{\theta} - \hat{\theta}_{(j)}\right)$$

Step 2 Calculate $\hat{R}(h) = n^{-1} \sum_{i=1}^{n-|h|} Y(i) \cdot Y(i + |h|), \quad h = -n + 1, \ldots, n - 1$

Step 3 Calculate iteratively:

$$b_0 = n^{-1},$$

$$b_k = n^{-1/3} \left[\left(\sum_{h=-n+1}^{n-1} \hat{R}(h)^2 \right) \times \left(6 \sum_{h=-n+1}^{n-1} w_{\mathrm{SC}}(h \cdot b_{k-1} \cdot n^{4/21})^2 \cdot h^2 \cdot \hat{R}(h)^2 \right)^{-1} \right]^{1/3},$$

$$k = 1, 2, 3, 4,$$

$$\hat{b} = n^{-1/3} \cdot (2/3)^{1/3} \left[\left(\sum_{h=-n+1}^{n-1} w_{\mathrm{TH}}(h \cdot b_4 \cdot n^{4/21}) \cdot \hat{R}(h) \right) \times \left(\sum_{h=-n+1}^{n-1} w_{\mathrm{SC}}(h \cdot b_4 \cdot n^{4/21}) \cdot |h| \cdot \hat{R}(h) \right)^{-1} \right]^{2/3}$$

Step 4 Set $l_{\mathrm{opt}} = NINT(\hat{b}^{-1})$

3. Select the parameter of interest, θ, and an estimation property of interest, say, bootstrap CI accuracy.
4. Prescribe a search grid. For example, l_{search} runs from a start to an end value with some spacing.
5. Calculate the empirical bootstrap CI coverage error (or another property) using the Monte Carlo samples and MBB with l_{search}.
6. Select l_{search} with best performance.

Other block length selectors are described briefly in the background material (Sect. 3.8).

Uneven Spacing

Applying the MBB to unevenly spaced time series increases the estimation uncertainty because the time spacing values within the inserted block, $\{d(i)\}_{i=j^*}^{j^*+l-2}$, need not equal the spacing values at the insertion place, $\{d^*(i)\}_{i=c}^{c+l-2}$. This may reduce the ability to preserve serial dependence.

An attempt to adapt MBB to this situation could be to resample only blocks with spacing similar to the spacing at the insertion place. For example, only the $\beta\%$

blocks with nearest spacing could be made drawable. The unevenness in a block could be quantified by the coefficient of variation of the spacing, CV_d, similarly as was done in Fig. 2.3. In the case of equidistance, one would have $\mathrm{CV}_d = 0$ and take $\beta = 100\,\%$, that is, one would use MBB. It is, however, unclear which β value to take for $\mathrm{CV}_d > 0$. A second measure could be to decrease l when reducing the number of drawable blocks.

A Monte Carlo experiment (Sect. 3.8) tested a rather simple MBB adaption: $\beta = 50\,\%$ for $\mathrm{CV}_d > 0$. This was applied to mean estimation of a Gaussian AR(1) process. It turned out, however, that the accuracy of the BCa CI was lower compared to usage of the ordinary MBB under the same block length selector (Eq. 3.28). More Monte Carlo studies of β choices in dependence on CV_d and other spacing properties have to be carried out to find more accurate MBB adaptions to uneven spacing.

The practical conclusion is that for small CV_d and large deviations from AR(1) persistence, one may use MBB. On the other hand, large CV_d and minor deviations from the AR(1) model indicate to employ the parametric autoregressive bootstrap (next section). This resampling method could have a higher relevance than MBB for practical applications because the AR(1) persistence model is generally a suitable first-order approximation for weather and climate time series (Chap. 2). Such a combined approach should yield acceptable results also for small data sizes. For that purpose, we tend to prefer the ARB over the MBB resampling type on the basis of the Monte Carlo experiments of mean estimation (Tables 3.5 and 3.7). If CV_d is large and also the deviations from AR(1) dependence are large, both the MBB and the parametric autoregressive bootstrap may be tried and results compared. This difference should indicate the size of the difference of the approximate bootstrap CIs to the exact CI.

Systematic Model Parts and Nonstationarity

For explaining the bootstrap principle (Fig. 3.3), we assumed for convenience of presentation $x(i) = x_{\mathrm{noise}}(i)$. Realistic climate processes contain more parts, such as trend, outliers and variability (Eq. 1.2). The MBB can be applied to such processes by resampling from the residuals. Plugging in the estimates into the climate equation (Eq. 1.2) yields

$$r(i) = [x(i) - \hat{x}_{\mathrm{trend}}(i) - \hat{x}_{\mathrm{out}}(i)] \Big/ \hat{S}(i), \qquad i = 1, \ldots, n, \tag{3.29}$$

where $\hat{x}_{\mathrm{trend}}(i)$, $\hat{x}_{\mathrm{out}}(i)$ and $\hat{S}(i)$ are estimated trend, outlier and variability components, respectively. The following chapters explain such estimations. The residuals, $r(i)$, are realizations of the noise process. (Analogously, the residuals, $\epsilon(i)$, in Chap. 2 are realizations of a white-noise process.) The MBB for realistic climate processes is listed as Algorithm 3.3.

Algorithm 3.3 MBB for realistic climate processes, which comprise trend, outlier and variability components

Step 1	Data	$\{t(i), x(i)\}_{i=1}^{n}$
Step 2	Resampled times	$\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$
	unchanged	
Step 3	Residuals (Eq. 3.29)	$r(i) = [x(i) - \hat{x}_{\text{trend}}(i) - \hat{x}_{\text{out}}(i)] \Big/ \hat{S}(i)$
Step 4	Apply MBB	
	(Algorithm 3.1)	
	to residuals	$\{r(i)\}_{i=1}^{n}$
Step 5	Resampled residuals	$\{r^*(i)\}_{i=1}^{n}$
Step 6	Use resampled residuals	
	to produce resamples	$x^*(i) = \hat{x}_{\text{trend}}(i) + \hat{x}_{\text{out}}(i) + \hat{S}(i) \cdot r^*(i)$

The trend, outlier and variability components allow to describe nonstationary climate processes. A further type of nonstationarity regards persistence. Consider as example ice-volume fluctuations over the past 4 Ma. In the early part (Pliocene), the persistence was weaker than in the late part (Pleistocene), when huge continental ice sheets had been built up (Mudelsee and Raymo 2005). Such nonstationarity can be accounted for by the local block bootstrap (Paparoditis and Politis 2002), where, in the example, Pliocene resamples, $x^*(i)$, are restricted to come from the Pliocene data, $x(i)$, analogously for Pleistocene resamples. The local block bootstrap could also be applied, as an alternative to using MBB and the residuals, to produce nonparametric trend and variability estimates with CIs (Bühlmann 1998). The cited paper applies smoothing to an ozone time series from Switzerland, 1932–1996. Evidently, the size of the locality region should be chosen taking prior knowledge about the data generating process into account.

3.3.2 Parametric: Autoregressive Bootstrap

The autoregressive bootstrap algorithm (ARB) is the ordinary bootstrap applied to the white-noise residuals, $\epsilon(i)$. We first take the residuals, $r(i)$, from the climate equation as in Eq. (3.29). Using the persistence model for $r(i)$, the residuals $\epsilon(i)$ are then formed. $\epsilon(i)$ are treated as realizations of a white-noise process; see Eq. (2.5). We employ the AR(1) persistence model as a suitable description for climate processes (Chap. 2). Advantageously, the distributional shape need not be Gaussian. Even and uneven spacings are treated separately.

Even Spacing

The ARB for even spacing is listed as Algorithm 3.4. Although the bias correction (Step 7) is only approximate (Sect. 2.6), this is considered an important step because ignoring bias can lead to a bad bootstrap performance (Stine 1987). Scaling, as done in Step 8 using a factor $\left[1-(\hat{a}')^2\right]^{-1/2}$, is non-standard. It has the computational advantage that no transient behaviour is required in Step 11. Centering (Step 9) achieves that the resample generating process has expectation zero, as the white-noise process is supposed to have. After Step 9, a further scaling with a factor $[(n-1)/(n-2)]^{1/2}$ (Stine 1987) is omitted. This factor is in the general case only approximate (Peters and Freedman 1984) and its effect is considered negligible compared with the other uncertainties. Lahiri (2003) explains the "traditional" method to generate a number of samples that is very much larger than n at Step 10 and use those at Step 11 for extracting $r^*(i)$ from the transient sequence. The advantage of the non-standard formulation (Step 8) corresponds to the advantage of strict stationarity of the non-standard formulation of the AR(1) model (Chap. 2).

Uneven Spacing

The ARB for uneven spacing is listed as Algorithm 3.5. It corresponds basically to the ARB for even spacing, where the persistence parameter, a, is replaced by $\exp\{-[t(i)-t(i-1)]/\tau\}$. Bias correction for $\hat{\tau}$ at Step 7 goes via $\hat{\bar{a}}' = \exp(-\bar{d}/\hat{\tau}')$.

3.3.3 Parametric: Surrogate Data

The surrogate data approach (Algorithm 3.6), related to ARB, is a simulation rather than a resampling method. No residuals are drawn as in the ARB. Instead, climate equation residuals $\{r^*(i)\}_{i=1}^n$ are obtained by numerical simulation (Step 8) from the persistence model with estimated (and bias-corrected) parameters. Because also the distributional shape is specified, the surrogate data approach is bounded stronger by parametric restrictions than the ARB. Therein lies its danger: it is more prone than the ARB to systematic errors from violated assumptions.

3.4 Bootstrap Confidence Intervals

Estimation of θ is repeated for the resamples, $\{t^{*b}(i), x^{*b}(i)\}_{i=1}^n, b = 1, \ldots, B$. This yields the bootstrap replications, $\{\hat{\theta}^{*b}\}_{b=1}^B$. The replications are used to construct equi-tailed $(1-2\alpha)$ confidence intervals, $\mathrm{CI}_{\hat{\theta},1-2\alpha}$; see Fig. 3.3.

Algorithm 3.4 Autoregressive bootstrap algorithm (ARB), even spacing

Step 1	Data	$\{t(i), x(i)\}_{i=1}^{n}$
Step 2	Resampled times	$\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$
	unchanged	
Step 3	Estimated trend,	$\{\hat{x}_{\text{trend}}(i)\}_{i=1}^{n}, \{\hat{x}_{\text{out}}(i)\}_{i=1}^{n}, \left\{\hat{S}(i)\right\}_{i=1}^{n}$
	outliers, variability	
Step 4	Climate equation	$\{r(i)\}_{i=1}^{n}$
	residuals (Eq. 3.29)	
Step 5	Assume $\{r(i)\}_{i=1}^{n}$ to	
	come from AR(1)	
	model for even	
	spacing (Eq. 2.1)	
Step 6	Estimate AR(1)	
	parameter (Eq. 2.4)	$\hat{a}$
Step 7	Bias correction	$\hat{a}'$
Step 8	White-noise residuals	$\epsilon(i) = [r(i) - \hat{a}' \cdot r(i-1)] \times \left[1 - (\hat{a}')^2\right]^{-1/2}, \quad i = 2, \ldots, n$
Step 9	Centering	$\tilde{\epsilon}(i) = \epsilon(i) - \sum_{i=2}^{n} \epsilon(i)/(n-1)$
Step 10	Draw $\tilde{\epsilon}^*(j)$,	
	$j = 2, \ldots, n$,	
	with replacement from	$\{\tilde{\epsilon}(i)\}_{i=2}^{n}$
Step 11	Resampled climate	$r^*(1)$ drawn from $\{r(i)\}_{i=1}^{n}$,
	residuals	$r^*(i) = \hat{a}' \cdot r^*(i-1) + \left[1 - (\hat{a}')^2\right]^{1/2} \cdot \tilde{\epsilon}^*(i), \quad i = 2, \ldots, n$
Step 12	Resampled data	$x^*(i) = \hat{x}_{\text{trend}}(i) + \hat{x}_{\text{out}}(i) + \hat{S}(i) \cdot r^*(i), \quad i = 1, \ldots, n$

Algorithm 3.5 Autoregressive bootstrap algorithm (ARB), uneven spacing

Step 1	Data	$\{t(i), x(i)\}_{i=1}^{n}$
Step 2	Resampled times	$\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$
	unchanged	
Step 3	Estimated trend,	$\{\hat{x}_{\text{trend}}(i)\}_{i=1}^{n}, \{\hat{x}_{\text{out}}(i)\}_{i=1}^{n}, \left\{\hat{S}(i)\right\}_{i=1}^{n}$
	outliers, variability	
Step 4	Climate equation	$\{r(i)\}_{i=1}^{n}$
	residuals (Eq. 3.29)	
Step 5	Assume $\{r(i)\}_{i=1}^{n}$ to	
	come from AR(1)	
	model for uneven	
	spacing (Eq. 2.9)	
Step 6	Estimate persistence	
	time (Eq. 2.11)	$\hat{\tau}$
Step 7	Bias correction	$\hat{\tau}'$
Step 8	Abbreviation	$\hat{a}'(i) = \exp\{-[t(i) - t(i-1)]/\hat{\tau}'\}, \quad i = 2, \ldots, n$
Step 9	White-noise residuals	$\epsilon(i) = [r(i) - \hat{a}'(i) \cdot r(i-1)] \times \left\{1 - [\hat{a}'(i)]^2\right\}^{-1/2}, \; i = 2, \ldots, n$
Step 10	Centering	$\tilde{\epsilon}(i) = \epsilon(i) - \sum_{i=2}^{n} \epsilon(i)/(n-1)$
Step 11	Draw $\tilde{\epsilon}^*(j)$,	
	$j = 2, \ldots, n$,	
	with replacement from	$\{\tilde{\epsilon}(i)\}_{i=2}^{n}$
Step 12	Resampled climate	$r^*(1)$ drawn from $\{r(i)\}_{i=1}^{n}$,
	residuals	$r^*(i) = \hat{a}'(i) \cdot r^*(i-1) + \left\{1 - [\hat{a}'(i)]^2\right\}^{1/2} \times \tilde{\epsilon}^*(i), \quad i = 2, \ldots, n$
Step 13	Resampled data	$x^*(i) = \hat{x}_{\text{trend}}(i) + \hat{x}_{\text{out}}(i) + \hat{S}(i) \cdot r^*(i), \quad i = 1, \ldots, n$

Algorithm 3.6 Surrogate data approach

Step 1	Data	$\{t(i), x(i)\}_{i=1}^{n}$
Step 2	Resampled times	$\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$
	unchanged	
Step 3	Estimated trend,	$\{\hat{x}_{\text{trend}}(i)\}_{i=1}^{n}$,
	outliers,	$\{\hat{x}_{\text{out}}(i)\}_{i=1}^{n}$,
	variability	$\left\{\hat{S}(i)\right\}_{i=1}^{n}$
Step 4	Climate equation	$\{r(i)\}_{i=1}^{n}$
	residuals (Eq. 3.29)	
Step 5	Assume $\{r(i)\}_{i=1}^{n}$ to	
	come from	
	specific model	
	(shape, persistence)	
Step 6	Estimate model	
	parameters	
Step 7	Bias correction	
Step 8	Simulate climate	
	equation residuals	$\{r^*(i)\}_{i=1}^{n}$
	from estimated model	
Step 9	Simulated data	$x^*(i) = \hat{x}_{\text{trend}}(i) + \hat{x}_{\text{out}}(i) + \hat{S}(i) \cdot r^*(i)$,
		$i = 1, \ldots, n$

Two approaches, standard error based and percentile based, dominate theory and practice of bootstrap CI construction. The estimated bootstrap standard error is the sample standard error of the replications,

$$\widehat{\text{se}}_{\hat{\theta}^*} = \left\{ \sum_{b=1}^{B} \left[\hat{\theta}^{*b} - \left\langle \hat{\theta}^{*b} \right\rangle \right]^2 \Big/ (B-1) \right\}^{1/2}, \tag{3.30}$$

where $\left\langle \hat{\theta}^{*b} \right\rangle = \sum_{b=1}^{B} \hat{\theta}^{*b}/B$. The percentiles result from the empirical distribution function (Eq. 3.46) of the replications. The accuracy of bootstrap CIs depends

critically on the similarity (in terms of standard errors or percentiles) of the distribution of the bootstrap replications and the true distribution, $f(\hat{\theta})$. Various concepts exist for accounting for the deviations between the two distributions.

Suppressing "simulation noise" requires more resamples for percentile estimation than for bootstrap standard error estimation. This book follows the recommendation of Efron and Tibshirani (1993) and sets throughout $B = 2000$ (or 1999 for percentile CIs). For a reasonable α value such as 0.025, this means that a number of 50 replications are outside the percentile bound. An own simulation study, analysing the coefficient of variation of a CI endpoint in dependence of B, confirmed that this choice is sufficient also in a bivariate setting (Mudelsee and Alkio 2007).

3.4.1 Normal Confidence Interval

The bootstrap normal confidence interval, already given in Fig. 3.3, is

$$\mathrm{CI}_{\hat{\theta},1-2\alpha} = \left[\hat{\theta} + z(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}*}; \hat{\theta} - z(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}*}\right], \tag{3.31}$$

where $z(\alpha)$ is the percentage point of the normal distribution (Sect. 3.9).

3.4.2 Student's t Confidence Interval

The bootstrap Student's t confidence interval is

$$\mathrm{CI}_{\hat{\theta},1-2\alpha} = \left[\hat{\theta} + t_\nu(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}*}; \hat{\theta} - t_\nu(\alpha) \cdot \widehat{\mathrm{se}}_{\hat{\theta}*}\right], \tag{3.32}$$

where $t_\nu(\alpha)$ is the percentage point of the t distribution function with ν degrees of freedom (Sect. 3.9). It is in practice presumably always more accurate to prefer, as this book does, Student's t CIs over normal CIs because they recognize the reduction of degrees of freedom. (For data sizes above, say, 30, the difference becomes negligible.)

3.4.3 Percentile Confidence Interval

The bootstrap percentile confidence interval is

$$\mathrm{CI}_{\hat{\theta},1-2\alpha} = \left[\hat{\theta}^*(\alpha); \hat{\theta}^*(1-\alpha)\right], \tag{3.33}$$

that is, it is the interval between the 100αth percentage point and the $100(1-\alpha)$th percentage point of the empirical distribution of $\left\{\hat{\theta}^{*b}\right\}_{b=1}^{B}$. Because of finite

B, "simulation noise" is introduced in estimating percentile-based CIs. $B = 1999$ sufficiently reduces this effect; see the introduction to this section. One takes this value instead of 2000 because then commonly used percentage points can be evaluated without interpolation (e.g. 95th percentage point $= 0.95 \cdot (1999 + 1)$th = 1900th largest replication value).

3.4.4 BCa Confidence Interval

The bootstrap bias-corrected and accelerated (BCa) confidence interval is

$$\mathrm{CI}_{\hat{\theta},1-2\alpha} = \left[\hat{\theta}^*(\alpha 1); \hat{\theta}^*(\alpha 2)\right], \tag{3.34}$$

where

$$\alpha 1 = F\left(\hat{z}_0 + \frac{\hat{z}_0 + z(\alpha)}{1 - \hat{a}\left[\hat{z}_0 + z(\alpha)\right]}\right) \tag{3.35}$$

and

$$\alpha 2 = F\left(\hat{z}_0 + \frac{\hat{z}_0 + z(1-\alpha)}{1 - \hat{a}\left[\hat{z}_0 + z(1-\alpha)\right]}\right). \tag{3.36}$$

$F(\cdot)$ is the standard normal distribution function (Eq. 3.52). $\hat{z}_0$, the bias correction, is computed as

$$\hat{z}_0 = F^{-1}\left(\frac{\#\left\{\hat{\theta}^{*b} < \hat{\theta}\right\}}{B}\right), \tag{3.37}$$

where $\#\{\hat{\theta}^{*b} < \hat{\theta}\}$ means the number of replications where $\hat{\theta}^{*b} < \hat{\theta}$ and $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. The acceleration, $\hat{a}$, is computed (Efron and Tibshirani 1993) as

$$\hat{a} = \frac{\sum_{j=1}^{n}\left[\left\langle\hat{\theta}_{(j)}\right\rangle - \hat{\theta}_{(j)}\right]^3}{6\left\{\sum_{j=1}^{n}\left[\left\langle\hat{\theta}_{(j)}\right\rangle - \hat{\theta}_{(j)}\right]^2\right\}^{3/2}}, \tag{3.38}$$

where $\hat{\theta}_{(j)}$ is the jackknife value of $\hat{\theta}$. Consider the original sample with the jth point removed, that is, $\{t(i), x(i)\}, i = 1, \ldots, n, i \neq j$. The jackknife value is then the value of $\hat{\theta}$ calculated using this sample of reduced size. The average, $\left\langle\hat{\theta}_{(j)}\right\rangle$, is given by $\left[\sum_{j=1}^{n}\hat{\theta}_{(j)}\right] \Big/ n$.

$\hat{z}_0$ corrects for the median estimation bias; for example, if just half of the replications have $\hat{\theta}^{*b} < \hat{\theta}$, then $\hat{z}_0 = 0$. The acceleration, $\hat{a}$, takes into account scale effects, which arise when the standard error of $\hat{\theta}$ itself depends on the true parameter value, θ.

3.5 Examples

In the first, theoretical example, we compare classical and bootstrap CIs in terms of coverage accuracy (Table 3.5). The mean of AR(1) processes with uneven spacing was estimated for two distributional shapes, normal and lognormal. The classical CI employed the effective data size for mean estimation; the bootstrap CI used the ARB algorithm and the BCa method.

The classical CI performed better for the normal than for the lognormal shape. This is because the normal assumption made at CI construction is violated in the case of the lognormal shape. With increasing data size, the lognormal approaches the normal distribution (Johnson et al. 1994: Chap. 14 therein) and the difference in performance decreases. However, this difference is still significant for $n = 1000$ in the example.

Also the bootstrap CI performed better for the normal than for the lognormal shape. This may be because persistence time estimation ($\hat{\tau}$) and persistence time bias correction ($\hat{\tau}'$) are less accurate for nonnormally distributed data.

Table 3.5 Monte Carlo experiment, mean estimation of AR(1) noise processes with uneven spacing, normal and lognormal shape. $n_{\text{sim}} = 47{,}500$ random samples were generated from the Gaussian AR(1) process, $\{X(i)\}_{i=1}^{n}$, after Eq. (2.9) with $\tau = 1$. The samples from the lognormal AR(1) process were generated by taking $\exp[X(i)]$. The start was set to $t(1) = 1$; the time spacing, $d(i)$, was drawn from a gamma distribution (Eq. 2.48) with order parameter 16, that is, a distribution with a coefficient of variation equal to $(16)^{-1/2} = 0.25$, and subsequently scaled to $\bar{d} = 1$. Two CI types for the estimated mean were constructed, classical and bootstrap. The classical CI employed n'_{μ} calculated from Eq. (2.7) with $\hat{\tilde{a}}' = \exp\left(-\bar{d}/\hat{\tau}'\right)$ plugged in for a and the t distribution (Eq. 3.18). The bootstrap CI used the ARB (Algorithm 3.5) and the BCa method (Sect. 3.4.4) with $B = 1999$ and $\alpha = 0.025$

	$\gamma_{\bar{x}}^{\text{a}}$ Distribution				
	Normal CI type		Lognormal CI type		
n	Classical	Bootstrap	Classical	Bootstrap	Nominal
10	0.918	0.863	0.835	0.789	0.950
20	0.929	0.903	0.845	0.845	0.950
50	0.938	0.929	0.876	0.888	0.950
100	0.943	0.941	0.897	0.909	0.950
200	0.942	0.943	0.914	0.922	0.950
500	0.947	0.948	0.926	0.930	0.950
1000	0.947	0.949	0.933	0.937	0.950

[a]Standard error of $\gamma_{\bar{x}}$ is nominally 0.001

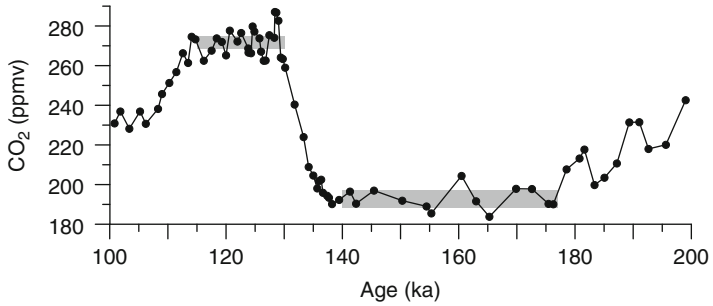


Fig. 3.4 Determination of mean CO_2 levels in the Vostok record (Fig. 1.4b) during a glacial and an interglacial. The interval from 140 to 177 ka represents the glacial (MIS 6), the interval from 115 to 130 ka the interglacial (marine isotope substage 5.5). The 95 % bootstrap CIs for the estimated means are shown as *shaded bars*

For small sample sizes ($n \lessapprox 50$ (normal distribution) or $n \lessapprox 20$ (lognormal distribution)), the classical CI performed better than the bootstrap CI. This advantage is likely in part owing to the fact that a formula for the effective data size for mean estimation is known; it may disappear for more complex estimators, where no formula for the effective data size exists. For larger sample sizes ($n \gtrapprox 100$ (normal distribution) or $n \gtrapprox 50$ (lognormal distribution)), the bootstrap CI is as good as the classical CI (normal shape) or better (lognormal shape).

In the second, practical example, Fig. 3.4 shows the transition from a glacial (MIS 6) to the last interglacial (MIS 5) in the Vostok CO_2 record. The mean CO_2 concentration was estimated for the time intervals from 140 to 177 ka (glacial) and from 115 to 130 ka (interglacial). Student's t CIs (Sect. 3.4.2) were constructed using nonparametric stationary bootstrap resampling, a variant of the MBB, where the block length is not constant (Sect. 3.8). The number of resamples was $B = 2000$. The average block length was set to $NINT\left(4 \cdot \tau/\bar{d}\right)$.

The mean glacial CO_2 level was determined as 192.8 ppmv with 95 % CI [188.3 ppmv; 197.3 ppmv]; the mean interglacial CO_2 level was 271.9 ppmv with 95 % CI [268.8 ppmv; 275.0 ppmv]. Because of the reduced data sizes in the intervals (glacial, $n = 13$; interglacial, $n = 24$), also the accuracies of the CIs may be reduced. The enormous glacial–interglacial amplitude in CO_2 documents the importance of this greenhouse gas for late Pleistocene climate changes, the ice age. The relation between CO_2 and temperature changes is analysed in Chaps. 7 and 8.

3.6 Bootstrap Hypothesis Tests

By the analysis of climate time series, $\{t(i), x(i)\}_{i=1}^{n}$, we make, generally speaking, a statistical inference of properties of the climate system. One type of inference is the estimation of a climate parameter, θ. In addition to a point estimate, $\hat{\theta}$, an

interval estimate, $\mathrm{CI}_{\hat{\theta},1-2\alpha}$, helps to assess how accurate $\hat{\theta}$ is. The bootstrap is used to construct CIs in complex situations regarding data properties shape, persistence and spacing. The second type of inference is testing a hypothesis, a statement about the climate system, using the data sample. Again, this can be a difficult task (shape, persistence and spacing), and again, the bootstrap can be a powerful tool in such a situation. Hypothesis tests are also called significance tests or statistical tests.

A hypothesis test involves the following procedure. A null hypothesis (or short: null), H_0, is formulated. H_0 is tested against an alternative hypothesis, H_1. The hypotheses H_0 and H_1 are mutually exclusive. H_0 is a simple null hypothesis if it completely specifies the data generating process. An example would be "$X(i)$ is a Gaussian white-noise process with zero mean and unit standard deviation." H_0 is a composite null hypothesis if some parameter of $X(i)$ is unspecified, for example, "Gaussian white-noise process with zero mean." Next, a test statistic, U, is calculated. Any meaningful construction lets U be a function of the data generating process, $U = g\left(\{T(i), X(i)\}_{i=1}^{n}\right)$. On the sample level, $u = g\left(\{t(i), x(i)\}_{i=1}^{n}\right)$. In the example H_0: "Gaussian white-noise process with $\mu = 0$", one could take $U = \bar{X} = \sum_{i=1}^{n} X(i)/n$, the sample mean. U is a random variable with a distribution function, $F_0(u)$, where the index "0" indicates that U is computed "under H_0", that is, as if H_0 were true. $F_0(u)$ is the null distribution. In the example, $F_0(u)$ would be Student's t distribution function (Sect. 3.9). If in the example the alternative were H_1: "$\mu > 0$", then a large, positive u value would speak against H_0 and for H_1. Using $F_0(u)$ and plugging in the data $\{t(i), x(i)\}_{i=1}^{n}$, the one-sided significance probability or one-sided P-value results as

$$\begin{aligned} P &= \mathrm{prob}\,(U \geq u \mid H_0) \\ &= 1 - F_0(u). \end{aligned} \tag{3.39}$$

The P-value is the probability that under H_0 a value of the test statistic greater than or equal to the observed value, u, is observed. If P is small, then H_0 is rejected and H_1 accepted; otherwise, H_0 cannot be rejected and H_1 cannot be accepted. The two-sided P-value is

$$P = \mathrm{prob}\,(|U| \geq |u| \mid H_0)\,. \tag{3.40}$$

In the example, a two-sided test would be indicated for H_1: "Gaussian white noise with $\mu \neq 0$". Besides the P-value, a second result of a statistical test is the power. In the one-sided test example:

$$\mathrm{power} = \mathrm{prob}\,(U \geq u \mid H_1)\,. \tag{3.41}$$

A type-2 error is accepting H_0, although it is a false statement and H_1 is true. The probability of a type-2 error is $\beta = 1 - \mathrm{power}$. A type-1 error is rejecting H_0 against H_1, although H_0 is true. P, the significance probability, is therefore denoted also as type-1-error probability or false-alarm probability; u is denoted also as false-alarm level.

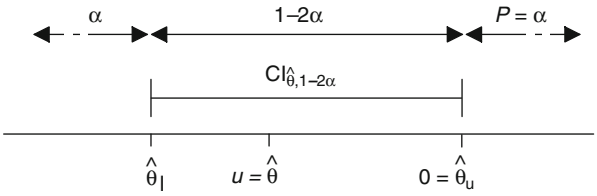


Fig. 3.5 Hypothesis test and confidence interval. The parametric null hypothesis H_0: "$\theta < 0$" cannot be rejected against H_1: "$\theta \geq 0$" with a P-value equal to α

Although H_0 can be a composite null, it is usually more explicit than H_1. In climatological practice, the selection of H_1 should be guided by prior climatological knowledge. H_1 determines also whether a test should be designed as one- or two-sided. For example, if H_0 were "no temperature change in a climate model experiment studying the effects of doubled CO_2 concentrations, $\Delta T = 0$", then a one-sided test against H_1: "$\Delta T > 0$" would be appropriate because physics would not let one expect a temperature decrease. Because H_1 is normally rather general, it is difficult to quantify the test power. Therefore, more emphasis is put on accurate P-values. Various test statistics, $U_1, U_2, \ldots$, may be appropriate for testing H_0 against H_1. The statistic of choice has for a given data set a small type-1-error probability (small P-value) as first quality criterion. The second quality criterion is a small type-2-error probability (large power), preferably calculated for some realistic, explicit alternative. We can say that a test does not intend to prove that a hypothesis is true but rather that it does try to reject a null hypothesis. A null hypothesis becomes more "reliable" after it has been tested successfully against various realistic alternatives using various data samples; see Popper (1935). It is important that H_0 and H_1 are established independently of the data to prevent circular reasoning; see von Storch and Zwiers (1999: Sect. 6.4 therein). As a final general remark, it is more informative to give P-values than to report merely whether they are below certain specified significance levels, say $P < 0.1, 0.05$ or 0.01.

When H_0 concerns a particular parameter value ($U = \theta$), a CI can be used to derive the P-value (Efron and Tibshirani 1993: Sect. 15.4 therein). Suppose that a test observes $u = \hat{\theta} < 0$. Then select α such that the upper CI bound equals zero. Nominally, prob$(\theta \geq 0) = \alpha$ (Fig. 3.5). This gives a P-value of α for the test of H_0: "$\theta < 0$" against H_1: "$\theta \geq 0$". An example from a bivariate setting with data $\{x(i), y(i)\}_{i=1}^{n}$ would be the comparison of means μ_X and μ_Y. If the CI at level $1 - 2\alpha$ for the absolute value of the difference of means, $|\mu_X - \mu_Y|$, does contain zero, then H_0: "$\mu_X = \mu_Y$" cannot be rejected against H_1: "$\mu_X \neq \mu_Y$" at the level $p = 1-2\alpha$ in this two-sided test. A criticism to this CI method of hypothesis testing would be that the CIs are not necessarily constructed as if H_0 were true. There might be scale changes and $F_0(u)$ depend on H_0. However, the BCa CI provides a correction to this effect (Efron and Tibshirani 1993: p. 216 therein). Another option would be to construct a test statistic, U, such that $F_0(u)$ is the same for all H_0. Such a statistic is called a pivot.

Davison and Hinkley (1997: Chap. 4 therein) explain the construction of hypothesis tests by approximating $F_0(u)$ with $\hat{F}_0(u)$ obtained from bootstrap resampling or the bootstrap surrogate data approach (Sect. 3.3.3). The permutation test, developed in the 1930s (Edgington 1986), is the bootstrap test with the difference that no replacement is done for drawing the random samples. This book here puts more emphasis on bootstrap CIs than on bootstrap hypothesis test because CIs contain more quantitative information. We subscribe to Efron and Tibshirani's (1993: p. 218 therein) view that "hypothesis tests tend to be overused and confidence intervals underused in statistical applications." We also agree with Yates (1951: p. 32 therein), who assessed the influence of Fisher's (1925) classic on the practice of "scientific research workers" as not entirely positive because it had caused them "to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating".

An illustrative example is the case where θ is the anthropogenic signal proportion in the increase of the global temperature over the past 150 years. Specifically, θ can be defined as $\Delta T_{\text{with}} - \Delta T_{\text{without}}$, where ΔT_{with} is the temperature change calculated using an AOGCM and taking human activities such as fossil fuel consumption into account and $\Delta T_{\text{without}}$ is the temperature change without the effects of human activities ("control run"). Hasselmann (1993) and Hegerl et al. (1996) developed the "fingerprint" approach to derive a powerful test statistic from the high-dimensional, gridded AOGCM output and showed that H_0: "$\theta = 0$" can be rejected against H_1: "$\theta > 0$". One task was to quantify the natural temperature variability in the temporal and spatial domains, in order to derive the null distribution. This is difficult because the observed variability contains both natural and anthropogenic portions. It was solved using AOGCM experiments without simulated anthropogenic forcings and a surrogate data approach (Sect. 3.3.3), that is, several control runs with perturbed initial conditions. It is evident that an estimate, $\hat{\theta}$, with confidence interval, $\text{CI}_{\hat{\theta},1-2\alpha}$, for the anthropogenic signal proportion would mean a step further towards quantification.

3.7 Notation

Table 3.6 summarizes the notation.

3.8 Background Material

We use **RMSE** instead of the mean squared error (given by $\text{RMSE}^2_{\hat{\theta}}$). RMSE, with the same units as the data, is a handy parameter.

We use a **coefficient of variation** operator (Eq. 3.4) with the absolute value of the mean to avoid negative values; the standard formulation does not do this.

Table 3.6 Notation

$X(T)$	Climate variable, continuous time, process level
$X_{\text{trend}}(T)$	Trend component, continuous time, process level
$X_{\text{out}}(T)$	Outlier component, continuous time, process level
$S(T)$	Variability, continuous time
$X_{\text{noise}}(T)$	Noise component, continuous time, process level
T	Continuous time
$X(i)$	Climate variable, discrete time, process level
$X_{\text{trend}}(i)$	Trend component, discrete time, process level
$X_{\text{out}}(i)$	Outlier component, discrete time, process level
$S(i)$	Variability, discrete time
$X_{\text{noise}}(i)$	Noise component, discrete time, process level
$T(i)$	Discrete time
i	Index
j	Index
$\mathcal{E}_{\text{N}(\mu,\,\sigma^2)}(i)$	Gaussian noise process with mean μ and standard deviation σ, discrete time
$x(i)$	Climate variable, discrete time, sample level
$t(i)$	Discrete time, sample level
$\{t(i), x(i)\}_{i=1}^{n}$	Data or sample, discrete time series
$d(i)$	Time spacing, sample level
$\bar{d}$	Average time spacing, sample level
n	Data size
θ	(Climate) parameter
$\hat{\theta}$	Estimator of (climate) parameter, process and sample levels, estimate
$\hat{\theta}_1, \hat{\theta}_2$	Other estimators
PDF	Probability density function
$f(\hat{\theta})$	PDF of $\hat{\theta}$
$F(\cdot)$	Probability distribution function
$F^{-1}(\cdot)$	Inverse probability distribution function
$F_{\text{emp}}(\cdot)$	Empirical distribution function
$E(\cdot)$	Expectation operator
$VAR(\cdot)$	Variance operator
$g(\cdot)$	Function
$\Gamma(\cdot)$	Gamma function
$NINT(\cdot)$	Nearest integer function
$\text{se}_{\hat{\theta}}$	Standard error of $\hat{\theta}$
$\text{bias}_{\hat{\theta}}$	Bias of $\hat{\theta}$
$\text{RMSE}_{\hat{\theta}}$	Root mean squared error of $\hat{\theta}$
$\text{CV}_{\hat{\theta}}$	Coefficient of variation of $\hat{\theta}$
CI	Confidence interval
$\text{CI}_{\hat{\theta},1-2\alpha}$	Confidence interval for $\hat{\theta}$ of level $1-2\alpha$
$\hat{\theta}_{\text{l}}$	Lower bound of CI for $\hat{\theta}$
$\hat{\theta}_{\text{u}}$	Upper bound of CI for $\hat{\theta}$
γ_{l}	Coverage, below lower CI bound

(continued)

Table 3.6 (continued)

γ_{u}	Coverage, above upper CI bound
γ	Coverage of CI
μ	Mean
$\hat{\mu}$	Mean estimator
$\bar{X}$	Sample mean, process level
$\bar{x}$	Sample mean, sample level
$\gamma_{\bar{x}}$	Coverage of $\mathrm{CI}_{\bar{x},1-2\alpha}$
σ	Standard deviation
$\hat{\sigma}$	Standard deviation estimator
S_{n-1}	Sample standard deviation, process level
s_{n-1}	Sample standard deviation, sample level
$\gamma_{s_{n-1}}$	Coverage of $\mathrm{CI}_{s_{n-1},1-2\alpha}$
$z(\beta) = z_\beta$	Percentage point at β of the standard normal distribution
$t_\nu(\beta)$	Percentage point at β of the t distribution function with ν degrees of freedom
$\chi^2_\nu(\beta)$	Percentage point at β of the chi-squared distribution function with ν degrees of freedom
β	Probability
n_{sim}	Number of (Monte Carlo) simulations
c	Constant
c	Counter
C	Coverage error
$\mathcal{O}(\cdot)$	Order of
$\langle\cdot\rangle$	Average
AR(1)	Autoregressive process of order 1
AR(p)	Autoregressive process of order p
MA(q)	Moving average process of order q
ARMA(p,q)	Mixed autoregressive moving average process
n'	Effective data size
n'_μ	Effective data size for mean estimation
n'_{σ^2}	Effective data size for variance estimation
n'_ρ	Effective data size for correlation estimation
a	AR(1) autocorrelation parameter (even spacing)
$\hat{a}$	AR(1) autocorrelation parameter (even spacing) estimator
$\hat{a}'$	AR(1) autocorrelation parameter (even spacing) estimator, bias-corrected
τ	AR(1) persistence time (uneven spacing)
$\hat{\tau}$	AR(1) persistence time (uneven spacing) estimator
$\hat{\tau}'$	AR(1) persistence time (uneven spacing) estimator, bias-corrected
$\bar{a}$	AR(1) equivalent autocorrelation parameter (uneven spacing)
$\hat{\bar{a}}$	AR(1) equivalent autocorrelation parameter (uneven spacing) estimator
$\hat{\bar{a}}'$	AR(1) equivalent autocorrelation parameter (uneven spacing) estimator, bias-corrected
$t^*, t^*(i)$	Bootstrap version of discrete time, sample level
$t^{*b}(i)$	Indexed bootstrap version of discrete time, sample level

(continued)

Table 3.6 (continued)

$b = 1, \ldots, B$	Index
B	Number of bootstrap resamples
$x^*, x^*(i)$	Bootstrap version of climate variable, discrete time, sample level
$x^{*b}(i)$	Indexed bootstrap version of climate variable, discrete time, sample level
$d^*(i)$	Bootstrap version of time spacing, sample level
$\{t^*(i), x^*(i)\}_{i=1}^n$	Bootstrap resample
$\hat{\theta}^*$	Bootstrap replication
$\hat{\theta}^{*b}$	Indexed bootstrap replication
MBB	Moving block bootstrap
ARB	Autoregressive bootstrap
NBB	Nonoverlapping block bootstrap
CBB	Circular block bootstrap
SB	Stationary bootstrap
MaBB	Matched-block bootstrap
TaBB	Tapered block bootstrap
l	Block length
l_{opt}	Optimal block length
l_{search}	Block length search value
$Y(i)$	Variable (l_{opt} selector after Bühlmann and Künsch (1999))
$\widehat{\text{IF}}(X(i))$	Estimated influence function
$\hat{R}(h)$	Function (l_{opt} selector after Bühlmann and Künsch (1999))
$\hat{R}(h)$	Autocovariance estimator (Chap. 2)
$\hat{\rho}(h)$	Autocorrelation estimator (Chap. 2)
h	Lag
$b_0, b_1, b_2, b_3, b_4, \hat{b}$	Parameters (l_{opt} selector after Bühlmann and Künsch (1999))
$w_{\text{SC}}(\cdot)$	Split-cosine window
$w_{\text{TH}}(\cdot)$	Tukey–Hanning window
z	Auxiliary variable
CV_d	Coefficient of variation of the spacing
β	Percentage of drawable blocks (adaption of MBB to uneven spacing)
$\hat{x}_{\text{trend}}(i)$	Estimated trend component, discrete time, sample level
$\hat{x}_{\text{out}}(i)$	Estimated outlier component, discrete time, sample level
$\hat{S}(i)$	Estimated variability, discrete time
$r(i)$	Residual of climate equation, discrete time (Eq. 1.2)
$r^*(i)$	Bootstrap version of residual of climate equation, discrete time (Eq. 1.2)
$\epsilon(i)$	White-noise residual, discrete time
$\widetilde{\epsilon}(i)$	Centred white-noise residual, discrete time
$\widetilde{\epsilon}^*(i)$	Bootstrap version of centred white-noise residual, discrete time
$\hat{a}'(i)$	Abbreviation (ARB algorithm)
BCa CI	Bias-corrected and accelerated CI
ABC CI	Approximate BCa CI
$\widehat{\text{se}}_{\hat{\theta}^*}$	Estimated bootstrap standard error

(continued)

Table 3.6 (continued)

$\hat{\theta}^*(\alpha)$	Percentage point at α of the empirical distribution of $\hat{\theta}^*$
$\alpha 1, \alpha 2$	Other α values
$\hat{z}_0$	Bias correction
$\hat{a}$	Acceleration
#{}	Number of cases
$\hat{\theta}_{(j)}$	Jackknife value of $\hat{\theta}$
H_0	Null hypothesis
H_1	Alternative hypothesis
U	Test statistic, process level
u	Test statistic, sample level (u is also denoted as false-alarm level)
U_1, U_2	Other test statistics, process level
$F_0(u)$	Null distribution
$\hat{F}_0(u)$	Estimated null distribution
P	P-value, probability of a type-1 error or false-alarm probability
β	Probability of a type-2 error
$\bar{x}_\mathrm{w}$	Weighted mean
$\mathrm{se}_{\bar{x}_\mathrm{w},\mathrm{ext}}$	External error of the weighted mean
$\mathrm{se}_{\bar{x}_\mathrm{w},\mathrm{int}}$	Internal error of the weighted mean
M	Median
$\hat{M}$	Sample median, process level
$\hat{m}$	Sample median, sample level
$X'(i)$	Size-sorted $X(i)$
ϵ	Small value
$\hat{\theta}_1^{*b}(\lambda)$	Indexed lower bootstrap CI bound over a grid of confidence levels
λ	Variable, determines confidence level
$\hat{p}(\lambda)$	Empirical probability (bootstrap calibration)
$y, p_0, p_1, p_2, p_3, p_4, q_0, q_1, q_2, q_3, q_4$	Parameters ($z(\beta)$ approximation)
u, v, w	Parameters (error function approximation)
b, δ	Parameters (lognormal distribution)
p, q	Parameters (geometric distribution)
$\mathbf{Z}$	Set of whole numbers
$\mathbf{S}$	Set of numbers
$\mathbf{S}^*$	Set of permuted elements of $\mathbf{S}$
AOGCM	Atmosphere–Ocean General Circulation Model
MIS	Marine isotope stage (sometimes also loosely used for marine isotope substage)
ΔT	Modelled temperature change
ΔT_with	Modelled temperature change, with fossil fuel consumption
$\Delta T_\mathrm{without}$	Modelled temperature change, without fossil fuel consumption

The **weighted mean** of a sample of data points (e.g. measurements) with known individual standard deviations (e.g. measurement errors), $\{x(i), S(i)\}_{i=1}^{n}$, is a combined summary estimate (Birge 1932; Bevington and Robinson 1992):

$$\bar{x}_{\mathrm{w}} = \left[\sum_{i=1}^{n} x(i) \,/ S(i)^2\right] \Bigg/ \left[\sum_{i=1}^{n} 1 \,/ S(i)^2\right]. \tag{3.42}$$

The internal error of the weighted mean is given by

$$\mathrm{se}_{\bar{x}_{\mathrm{w}},\mathrm{int}} = \left[\sum_{i=1}^{n} 1 \,/ S(i)^2\right]^{-1/2}. \tag{3.43}$$

The external error of the weighted mean is given by

$$\mathrm{se}_{\bar{x}_{\mathrm{w}},\mathrm{ext}} = \left\{\sum_{i=1}^{n} [\,(x(i) - \bar{x}_{\mathrm{w}})/\, S(i)]^2\right\}^{1/2} \times \left\{(n-1)\left[\sum_{i=1}^{n} 1 \,/ S(i)^2\right]\right\}^{-1/2}. \tag{3.44}$$

The internal error measures the variation via the average statistical error from the individual (measurement) errors. The external error measures via the spread of the individual data values. A deviation between internal and external errors indicates violated assumptions; a smaller external error may point to overestimated individual standard deviations, and a larger external error may point to hidden systematic influences that are not included in the individual standard deviations. Researchers should report both internal and external errors and, adopting a conservative approach (Birge 1932), should consider the maximum of both for the interpretation of results. The weighted mean is a special case of weighted linear least-squares regression (Sect. 4.1.1), where the slope is prescribed as zero.

Standard deviation estimation for Gaussian white noise seems to have raised more interest in previous decades than today, as the discussion from 1968 in the journal The American Statistician illustrates (Cureton 1968a,b; Bolch 1968; Markowitz 1968a,b; Jarrett 1968). For example, the choice $\hat{\sigma} = c \cdot S_{n-1}$, with c given by Eq. (3.24), yields minimal $\mathrm{RMSE}_{\hat{\sigma}}$ among all σ estimators for Gaussian white noise (Goodman 1953). Or, $\hat{\sigma} = c^{-1} \cdot S_{n-1}$ yields $\mathrm{bias}_{\hat{\sigma}} = 0$ for Gaussian white noise; see, for example, Holtzman (1950). Today, it appears for practical purposes rather arbitrary whether or not to scale S_{n-1}, or whether to use $n-1$ or n. The resulting differences are likely much smaller than the effects of violations of the Gaussian assumption.

The **median** of a distribution is defined via $F(M) = 0.5$. ($F(\cdot)$ is the distribution function; see Eq. (3.52).) The sample median as estimator of M is on the process level

$$\hat{M} = \begin{cases} X'((n+1)/2) & \text{for uneven } n, \\ 0.5 \cdot [X'(n/2) + X'(n/2+1)] & \text{for even } n, \end{cases} \tag{3.45}$$

where $X'(i)$ are the size-sorted $X(i)$. On the sample level, $\hat{m}$ results from using $x(i)$.

A **robust estimation** procedure "performs well not only under ideal conditions, the model assumptions that have been postulated, but also under departures from the ideal" (Bickel 1988). In the context of this book, the assumptions regard distributional shape, persistence and spacing; the performance regards an estimator and its properties such as RMSE or CI coverage accuracy. Under ideal conditions, robust estimation procedures can be less efficient (have higher $\mathrm{se}_{\hat{\theta}}$) than non-robust procedures. For example, for Gaussianity and $n \to \infty$, $\mathrm{se}_{\hat{m}} \to (\pi/2)^{1/2} \cdot \mathrm{se}_{\hat{\mu}}$ (Chu 1955). Robust estimators can require sorting operations, which makes it often difficult to deduce their distribution. The term "robust" was coined by Box (1953) and Box and Andersen (1955); relevant papers on robust location estimation include Huber (1964) and Hampel (1985); for more details, see Tukey (1977) or Huber (1981). Unfortunately, today's usage of "robust" in the climate research literature is rather arbitrary.

The **empirical distribution function** of a sample $\{x(i)\}_{i=1}^{n}$ is given by

$$F_{\mathrm{emp}}(x) = \frac{\text{number of values } \leq x}{n}. \tag{3.46}$$

$F_{\mathrm{emp}}(x)$ is the sample analogue of the theoretical distribution function, for example, Eq. (3.52).

Bootstrap resampling was formally introduced by Efron (1979); this article summarizes also earlier work. Singh (1981) soon recognized that the ordinary bootstrap yields inconsistent results in a setting with serial dependence. A consistent estimator, $\hat{\theta}$, converges in probability to θ as n increases. Convergence in probability means

$$\lim_{n \to \infty} \mathrm{prob}\left(|\hat{\theta} - \theta| > \epsilon\right) = 0 \; \forall \; \epsilon > 0. \tag{3.47}$$

Textbooks on bootstrap resampling include those written by Efron and Tibshirani (1993), Davison and Hinkley (1997), and Good (2005). Statistical point estimation is covered by Lehmann and Casella (1998).

The **moving block bootstrap** or MBB was introduced by Künsch (1989) and Liu and Singh (1992). The MBB resamples overlapping blocks. Carlstein (1986) had earlier suggested a method (denoted as NBB) that resamples nonoverlapping

blocks and does not truncate the final block. This may lead to resamples with data size less than n, that is, subsampling (see below). Hall (1985) had already considered overlapping and nonoverlapping block methods in the context of spatial data. Bühlmann (1994) showed that if

1. $X(i)$ is a stationary Gaussian process with short-range dependence
2. $\hat{\theta}$ is a smooth function $g(\{x(i)\})$ of the data (e.g. the mean is a smooth function, but the median not) and
3. The block length, l, increases with the data size, n, within bounds, $l = \mathcal{O}\left(n^{1/2-\epsilon}\right), 0 < \epsilon < 1/2$

then the MBB produces resamples from a process that converges to the data generating process. The MBB is then called asymptotically valid. The questions after the validity and other properties of the MBB and other bootstrap methods under relaxed assumptions (non-Gaussian processes, long-range dependence, etc.) are currently extensively studied in statistical science. For long-range dependence and the sample mean as estimator with an asymptotically Gaussian distribution, MBB can be modified to provide a valid approximation (Lahiri 1993). For long-range dependence and non-Gaussian limit distributions, MBB has to be changed to subsampling one single block (Hall et al. 1998). Block length selection is less explored for long-range dependence; intuitively, a larger length should be used than for short-range dependence. See Berkowitz and Kilian (2000), Bühlmann (2002), Politis (2003), Lahiri (2003) and references cited in these overviews.

Other **block length selectors** for the MBB and also for other nonparametric bootstrap methods have been proposed. Hall et al. (1995a) gave an iterative method based on subsamples and cross-validation. As regards the subsample size, consult Carlstein et al. (1998: p. 309 therein). Although the convergence properties in the general case are unknown, the method performed well in the Monte Carlo simulations shown. Politis and White (2004) developed a rule that selects block length as two times the smallest integer lag, after which the autocovariance function (Eq. 2.18) "appears negligible". A related rule, based on the persistence time, τ, of the AR(1) process for uneven spacing (Sect. 2.1.2), would set $l = NINT\left(4 \cdot \tau/\bar{d}\right)$; Mudelsee (2003) suggested this rule for correlation estimation of bivariate, unevenly spaced time series (Chap. 7).

An **MBB adaption to uneven spacing** was analysed using a Monte Carlo experiment. The following simple rule was employed. Instead of allowing all $n - l + 1$ blocks to be drawn for insertion, only the 50 % blocks closest (plus ties) in the coefficient of variation of the spacing, CV_d, were made drawable. This was applied to mean estimation of a Gaussian AR(1) process. The comparison between this MBB adaption and the ordinary MBB was made in terms of coverage accuracy and average CI length (Table 3.7). The experiment used the BCa CI and employed the block length selector after Eq. (3.28) for the MBB and its adaption. The result (Table 3.7) exhibits a reduced coverage accuracy of the MBB adaption. The following deficit outweighed the advantage of the adaption (increased similarity of CV_d between sample and resample). Reducing the drawable blocks to 50 %

Table 3.7 Monte Carlo experiment, moving block bootstrap adaption to uneven spacing. $n_{\text{sim}} = 47{,}500$ random samples were generated from the Gaussian AR(1) process, $\{X(i)\}_{i=1}^{n}$, after Eq. (2.9) with $\tau = 1$. The start was set to $t(1) = 1$; the time spacing, $d(i)$, was drawn from a gamma distribution (Eq. 2.48) with order parameter 16, that is, a distribution with a coefficient of variation equal to $(16)^{-1/2} = 0.25$, and subsequently scaled to $\bar{d} = 1$. Bootstrap BCa CIs for the estimated mean were constructed with $B = 1999$ and $\alpha = 0.025$. The ordinary MBB resampling algorithm was compared with an MBB adaption to uneven spacing. The adaption made drawable only the 50 % blocks closest (plus ties) in the coefficient of variation of the spacing. Both the MBB and its adaption to uneven spacing yield clearly larger coverage errors than the ARB because in that Monte Carlo experiment (Table 3.5), the prescribed AR(1) dependence matches the assumption made by the ARB (Sect. 3.3.2)

	$\gamma_{\bar{x}}$[a] Resampling method			⟨CI length⟩[b] Resampling method	
n	MBB	Adapted MBB	Nominal	MBB	Adapted MBB
10	0.591	0.623	0.950	0.836	0.864
20	0.799	0.788	0.950	0.915	0.890
50	0.874	0.861	0.950	0.685	0.672
100	0.901	0.888	0.950	0.510	0.505
200	0.913	0.903	0.950	0.374	0.372
500	0.929	0.920	0.950	0.244	0.244
1000	0.935	0.923	0.950	0.176	0.175

[a]Standard error of $\gamma_{\bar{x}}$ is nominally 0.001
[b]Average value over n_{sim} simulations

reduced, in comparison with the ordinary MBB, the variation between resamples. This in turn reduced the variation between the replications (sample means of resamples). This led to narrower CIs from the adapted MBB algorithm (last two columns in Table 3.7). The CIs from the adapted MBB, finally, contained the true μ value less often than the CIs from the ordinary MBB. This means a reduced accuracy because the empirical coverages were in this case of mean estimation always less than the nominal value.

Other **nonparametric bootstrap resampling methods** than the MBB have been proposed. The circular block bootstrap (CBB) (Politis and Romano 1992a) "wraps" the data $\{x(i)\}_{i=1}^{n}$ around a circle such that $x(n)$ (Algorithm 3.1) has a successor, $x(1)$. The CBB then resamples overlapping blocks of length l from this periodic structure. That overcomes the deficit of the MBB that data near the edges, $x(1)$ or $x(n)$, have a lower probability to be resampled than data in the centre. Also the stationary bootstrap (SB) (Politis and Romano 1994) uses the periodic structure to ensure stationarity of the resampling process. Also the SB uses overlapping blocks—however, the block length is not constant but geometrically distributed. Similar selectors as for the MBB (Sect. 3.3.1) can be used for adjusting the average block length. As regards the choice among MBB, NBB, CBB and SB, Lahiri (1999) showed that (1) overlapping blocks (MBB, CBB, SB) are better than nonoverlapping blocks (NBB) in terms of RMSE of estimation of variance and related quantities like bootstrap standard error and (2) nonrandom block lengths

(MBB, CBB) are, under the same criterion, at least as good as random block lengths (SB). For estimation of the distribution function and related quantities like CI points, less is known, but there are indications that also here MBB and CBB perform better (Lahiri 2003: Chap. 5 therein). Some recent developments are the following. The matched-block bootstrap (MaBB) (Carlstein et al. 1998) introduces dependence between blocks to reduce bias in the bootstrap variance by imposing probability rules. One rule prefers resampling blocks such that block values at the endpoints, where the blocks are concatenated, show a higher agreement than under the MBB. The tapered block bootstrap (TaBB) (Paparoditis and Politis 2001) tapers (weights) data by means of a function before concatenating blocks. The idea is to give reduced weight to data near the block endpoints. This could make the TaBB have lower estimation bias than MBB or CBB (Paparoditis and Politis 2001). Advanced block bootstrap methods could be better than MBB for analysing equidistant climate time series, especially in the case of the MaBB, which shows good theoretical and simulation results when $X(i)$ is an AR(p) process (Carlstein et al. 1998). For uneven spacing, it could be more important to enhance MBB by matching blocks in terms of their spacing structure. This point deserves further study by means of Monte Carlo experiments. Subsampling refers to a procedure where the bootstrap resample size is less than the data size. NBB can lead to subsampling. Also the jackknife (Efron 1979), where $l = n - 1$ and one block only is resampled, is a subsampling variant. A detailed account is given by Politis et al. (1999). We finally mention the wild bootstrap, which attempts to reconstruct the distribution of a residual $r(i)$ (Eq. 3.29) by means of a two-point distribution (Wu 1986; Härdle and Marron 1991). The adaption of the wild bootstrap to nonparametric autoregression by Neumann and Kreiss (1998) has not yet been extended to uneven spacing, however.

The **autoregressive bootstrap** or ARB has been developed in the early 1980s; relevant early papers include Freedman and Peters (1984), Peters and Freedman (1984), Efron and Tibshirani (1986), and Findley (1986). Then Bose (1988) showed second-order correctness of the ARB method for estimating stationary AR(p) models—not necessarily Gaussian—and even spacing. Validity of the ARB for nonstationary AR(p) models (e.g. random walk or unit-root processes) requires subsampling, that is, drawing less than n resamples at Step 12 of the ARB (Algorithm 3.4); see Lahiri (2003: Chap. 8 therein). The ARB was extended to stationary ARMA(p, q) models with even spacing by Kreiss and Franke (1992). It seems difficult to generate theoretical knowledge about ARB performance for time series models with uneven spacing.

Other **parametric bootstrap resampling methods** than the ARB have been proposed. The sieve bootstrap (Kreiss 1992; Bühlmann 1997) assumes an AR(∞) process. Because of the high number of terms, this model is highly flexible and can approximate other persistence models than the AR(p) with $p < \infty$. Therefore, the sieve bootstrap could also be called a semi-parametric bootstrap method. The deficit of this method regarding application to climate time series is that it is restricted to even spacing. The parametric bootstrap for Gaussian ARFIMA processes was shown to yield similar asymptotic coverage errors of CIs for covariance estimation as in the case of independent processes (Andrews and Lieberman 2002).

The **frequency-domain bootstrap** is explained in Chap. 5.

The **surrogate data** approach comes from dynamical systems theory in physics (Theiler et al. 1992). Contrary to the assertion in the review on surrogate time series by Schreiber and Schmitz (2000: p. 352 therein), this approach is *not* the common choice in the bootstrap literature. The same as the surrogate data approach is the so-called Monte Carlo approach (Press et al. 1992: Sect. 15.6 therein).

Bootstrap CIs, their construction and statistical properties are reviewed in the above-mentioned textbooks and by DiCiccio and Efron (1996) and Carpenter and Bithell (2000). The challenging question "why not replace [CIs] with more informative tools?" has been raised by Hall and Martin (1996: p. 213 therein). This is based on their criticism that "the process of setting confidence intervals merely picks two points off a bootstrap histogram, ignoring much relevant information about shape and other important features." It has yet to be seen whether graphical tools such as those described by Hall and Martin (1996) will be accepted by the scientific communities. The percentile CI was proposed by Efron (1979), the BCa CI by Efron (1987). A numerical approximation to the BCa interval, called ABC interval, was introduced by DiCiccio and Efron (1992). See Sect. 3.9 on numerical issues concerning construction of BCa intervals. Götze and Künsch (1996) show the second-order correctness of BCa CIs for various estimators and the MBB for serially dependent processes. Hall (1988) determined theoretical coverage accuracies of various bootstrap CI types for estimators that are smooth functions of the data. Bootstrap-t CIs are formed using the standard error, $\mathrm{se}^*_{\hat{\theta}}$, of a single bootstrap replication (Efron and Tibshirani 1993). For simple estimators like $\hat{\mu} = \bar{X}$, plug-in estimates can be used instead of $\mathrm{se}^*_{\hat{\theta}}$. However, for more complex estimators, no plug-in estimates are at hand. A second bootstrap loop (bootstrapping from bootstrap samples) had to be invoked, which would increase computing costs.

Bootstrap calibration can strongly increase CI coverage accuracy. Consider that a single CI point is sought, say, the lower bound, $\hat{\theta}_\mathrm{l}$, for an estimate, $\hat{\theta}$. Let the bound be calculated for each bootstrap sample, $b = 1, \ldots, B$, and over a grid of confidence levels, for example,

$$\hat{\theta}_\mathrm{l}^{*b}(\lambda), \qquad \lambda = 0.01, \ldots, 0.99. \tag{3.48}$$

For each λ, compute

$$\hat{p}(\lambda) = \frac{\#\left\{\hat{\theta} \le \hat{\theta}_\mathrm{l}^{*b}(\lambda)\right\}}{B}. \tag{3.49}$$

Finally, solve $\hat{p}(\lambda) = \alpha$ for λ. In case a two-sided, equi-tailed CI is sought, the calibration curve $\hat{p}(\lambda) = 1 - 2\alpha$, where

$$\hat{p}(\lambda) = \frac{\#\left\{\hat{\theta}_\mathrm{l}^{*b}(\lambda) < \hat{\theta} < \hat{\theta}_\mathrm{u}^{*b}(\lambda)\right\}}{B}, \tag{3.50}$$

is solved for λ. To calculate the CI points for a bootstrap sample requires to perform a second bootstrap–estimation loop. Analysing second-loop bootstrap methods like calibration or bootstrap-t interval construction may require enormous computing costs. Relevant papers on calibrated bootstrap CIs include Hall (1986), Loh (1987, 1991), Hall and Martin (1988), Martin (1990), and Booth and Hall (1994). Regarding the context of resampling data from serially dependent processes, Choi and Hall (2000) report that the sieve or AR(∞) bootstrap has a significantly better performance than blocking methods in CI calibration. However, the sieve bootstrap is not applicable to unevenly spaced time series. This book presents a Monte Carlo experiment on calibrated bootstrap CIs for correlation estimation (Chap. 7), with satisfying coverage performance despite the used MBB resampling.

Bootstrap hypothesis tests are detailed by Davison and Hinkley (1997: Chap. 4 therein); see also Efron and Tibshirani (1993: Chap. 15 therein) and Lehmann and Romano (2005: Chap. 15 therein). The relation between making a test statistic pivotal and bootstrap CI calibration is described by Beran (1987, 1988). Guidelines for bootstrap hypothesis testing are provided by Hall and Wilson (1991). An extension of MBB hypothesis testing of the mean from univariate to multivariate time series has been presented by Wilks (1997). The dimensionality may be rather high, and the method may therefore be applicable to time-dependent climate fields such as gridded temperature output from a mathematical climate model. Beersma and Buishand (1999) compare variances of bivariate time series using jackknife resampling. They find significantly higher variability of future northern European precipitation amounts in the computer simulation with elevated greenhouse gas concentrations than in the simulation without (control run). Huybers and Wunsch (2005) test the hypothesis that Earth's obliquity variations influence glacial terminations during the late Pleistocene using parametric resampling of the timescale (Sect. 4.1.7). Huybers (2011) extends this work to Earth's precession variations.

Multiple hypothesis tests may be performed when analysing a hypothesis that consists of several sub-hypotheses. This situation arises in spectrum estimation (Chap. 5), where a range of frequencies is examined. The traditional method is adjusting the P-values of the individual tests to yield the desired overall P-value. A recent paper (Storey 2007: p. 347 therein) states "that one can improve the overall performance of multiple significance tests by borrowing information across all the tests when assessing the relative significance of each one, rather than calculating P-values for each test individually".

The **anthropogenic warming signal** has stimulated much work applying various types of hypothesis tests using measured and AOGCM temperature data. More details on the fingerprint approach are contained in the following papers: Hasselmann (1997), Hegerl and North (1997) and Hegerl et al. (1997). Correlation approaches to detect the anthropogenic warming signal are described by Folland et al. (1998) and Wigley et al. (2000). A recent overview is given by Barnett et al. (2005).

3.9 Technical Issues

The **standard normal (Gaussian) distribution** has the following PDF:

$$f(x) = (2\pi)^{-1/2} \exp\left(-x^2/2\right). \tag{3.51}$$

Figure 3.1 shows the distributional shape. The distribution function,

$$F(x) = \int_{-\infty}^{x} f(x')dx', \tag{3.52}$$

cannot be expressed in closed analytical form. We use

$$F(x) = 1 - 0.5 \operatorname{erfcc}\left(x \Big/ \sqrt{2}\right), \tag{3.53}$$

where for $x \geq 0$ the complementary error function, erfcc, is approximated (Press et al. 1992: Sect. 6.2 therein) via

$$\begin{aligned} \operatorname{erfcc}(u) \approx v \exp(&-w^2 - 1.26551223 + v\,(1.00002368 + v\,(0.37409196 \\ &+ v\,(0.09678418 + v\,(-0.18628806 + v\,(0.27886807 \\ &+ v\,(-1.13520398 + v\,(1.48851587 \\ &+ v\,(-0.82215223 + v\,0.17087277))))))))), \end{aligned} \tag{3.54}$$

$$v = 1/(1 + w/2), \tag{3.55}$$

$$w = |u|. \tag{3.56}$$

For $x < 0$, use the symmetry, $F(-x) = 1 - F(x)$. For all x, this approximation has a relative error of less than $1.2 \cdot 10^{-7}$ (Press et al. 1992). The inverse function of $F(x)$ defines the percentage point on the x axis, $z(\beta)$, with $0 \leq \beta \leq 1$. Approximations are used for calculating $z(\beta)$; for the Monte Carlo simulation experiments in this book, the formula given by Odeh and Evans (1974) is employed:

$$z(\beta) \simeq -y - \frac{\{[(y \cdot p_4 + p_3) \cdot y + p_2] \cdot y + p_1\} \cdot y + p_0}{\{[(y \cdot q_4 + q_3) \cdot y + q_2] \cdot y + q_1\} \cdot y + q_0}, \qquad 0 < \beta < 0.5, \tag{3.57}$$

where

$$y = \left[\ln\left(\beta^{-2}\right)\right]^{1/2} \tag{3.58}$$

and

$$\begin{aligned} p_0 &= -0.322232431088, & p_1 &= -1.0, \\ p_2 &= -0.342242088547, & p_3 &= -0.0204231210245, \\ p_4 &= -0.453642210148 \cdot 10^{-4}, & q_0 &= 0.0993484626060, \\ q_1 &= 0.588581570495, & q_2 &= 0.531103462366, \\ q_3 &= 0.103537752850, & q_4 &= 0.38560700634 \cdot 10^{-2}. \end{aligned} \tag{3.59}$$

If $0.5 < \beta < 1$, then $z(\beta) = -z(1-\beta)$. This approximation produces, for example, the values $z(1-0.025) \approx 1.959964$ and $z(1-0.05) \approx 1.644854$. For $10^{-20} \leq \beta \leq 1 - 10^{-20}$, Eq. (3.57) yields an approximation that is accurate to seven decimal places (Odeh and Evans 1974). The percentage point of the standard normal distribution can be used to calculate approximate percentage points of other distributions such as Student's t and chi-squared (see the following paragraphs). See the following for more details on the Gaussian distribution: Johnson et al. (1994: Chap. 13 therein) and Patel and Read (1996).

Student's t distribution with ν degrees of freedom has the following PDF:

$$f(x) = \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\,\Gamma(\nu/2)} \left(1 + x^2/\nu\right)^{-(\nu+1)/2}, \qquad \nu = 1, 2, \ldots. \tag{3.60}$$

Approximations have to be used for calculating the percentage point, $t_\nu(\beta)$. For the Monte Carlo simulation experiments in this book, the following formula (Abramowitz and Stegun 1965: p. 949 therein) is employed:

$$\begin{aligned} t_\nu(\beta) \simeq z_\beta &+ \frac{z_\beta^3 + z_\beta}{4\nu} + \frac{5z_\beta^5 + 16z_\beta^3 + 3z_\beta}{96\nu^2} \\ &+ \frac{3z_\beta^7 + 19z_\beta^5 + 17z_\beta^3 - 15z_\beta}{384\nu^3} \\ &+ \frac{79z_\beta^9 + 776z_\beta^7 + 1482z_\beta^5 - 1920z_\beta^3 - 945z_\beta}{92{,}160\nu^4}, \end{aligned} \tag{3.61}$$

where $z_\beta = z(\beta)$ is the percentage point of the standard normal distribution. For $\nu \geq 10$ and $0.0025 \leq \beta \leq 0.9975$, this approximation has a relative accuracy of less than 0.015 % (own determination using Johnson et al. 1995: Table 28.7 therein). See Johnson et al. (1995: Chap. 28 therein) for more details on the t distribution.

The **chi-squared distribution** with ν degrees of freedom has the following PDF:

$$f(x) = \exp(-x/2)x^{\nu/2-1} \Big/ \left[2^{\nu/2} \cdot \Gamma(\nu/2)\right], \qquad x \geq 0,\ \nu > 0. \tag{3.62}$$

It has mean ν and variance 2ν. Approximations are used for calculating the percentage point, $\chi^2_\nu(\beta)$. For the Monte Carlo simulation experiments in this book, the following formula (Goldstein 1973) is employed:

$$\chi^2_\nu(\beta) \simeq \nu \left\{ 1 - \frac{2}{9\nu} + \frac{4z_\beta^4 + 16z_\beta^2 - 28}{1215\nu^2} + \frac{8z_\beta^6 + 720z_\beta^4 + 3216z_\beta^2 + 2904}{229{,}635\nu^3} + (2/\nu)^{1/2} \left[\frac{z_\beta}{3} - \frac{z_\beta^3 - 3z_\beta}{162\nu} - \frac{3z_\beta^5 + 40z_\beta^3 + 45z_\beta}{5832\nu^2} + \frac{301z_\beta^7 - 1519z_\beta^5 - 32{,}769z_\beta^3 - 79{,}349z_\beta}{7{,}873{,}200\nu^3} \right] \right\}^3, \tag{3.63}$$

where $z_\beta = z(\beta)$ is the percentage point of the standard normal distribution. For $\nu \geq 10$ and $0.001 \leq \beta \leq 0.999$, this approximation has a relative accuracy of less than 0.05 % (Zar 1978). See Johnson et al. (1994: Chap. 18 therein) for more details on the chi-squared distribution.

The **lognormal distribution** can be defined as follows. If $\ln[X(i)]$ is distributed as $\mathrm{N}(\mu, \sigma^2)$, then $X(i)$ has a lognormal distribution with parameters μ and σ (shape). It has the PDF

$$f(x) = (2\pi)^{-1/2} \cdot \sigma^{-1} \cdot x^{-1} \cdot \exp\left\{ -[\ln(x/b)]^2 \Big/ \left(2\sigma^2\right) \right\}, \qquad x > 0, \tag{3.64}$$

where $b = \exp(\mu)$. The lognormal has expectation $\exp(\mu + \sigma^2/2)$ and variance $\left\{\exp(2\mu) \cdot \exp(\sigma^2) \cdot [\exp(\sigma^2) - 1]\right\}$. Other definitions with an additional shift parameter ($(X(i) - \delta)$ instead of $X(i)$) exist. See Aitchison and Brown (1957), Antle (1985), Crow and Shimizu (1988) or Johnson et al. (1994: Chap. 14 therein) for more details on the lognormal distribution.

The **geometric distribution** is a discrete distribution with

$$\mathrm{prob}\,(X = x) = p \cdot q^x, \qquad x = 0, 1, 2, \ldots, \tag{3.65}$$

where $q = 1 - p$ and $0 < p < 1$. It has expectation q/p. See Johnson et al. (1993: Chap. 5 therein) for more details on the geometric distribution.

BCa CI construction has numerical pitfalls. Regarding the bias correction, $\hat{z}_0$, in the case of a discretely distributed, unsmooth estimator, $\hat{\theta}$, own experiments with median estimation and $x(i) \in \mathbf{Z}$ (whole numbers) have shown that a higher CI accuracy is achieved when using instead of Eq. (3.37) the following formula:

$$\hat{z}_0 = F^{-1}\left(\frac{\#\left\{\hat{\theta}^{*b} < \hat{\theta}\right\}}{B} + \frac{\#\left\{\hat{\theta}^{*b} = \hat{\theta}\right\}}{2B} \right). \tag{3.66}$$

Because only a finite number, B, of $\hat{\theta}^*$ values are computed, $\hat{\theta}^*(\alpha 1)$ and $\hat{\theta}^*(\alpha 2)$ are calculated by interpolation. If now B is too small, the acceleration, $\hat{a}$, too large and α too small, then $\alpha 1$ may become too small or $\alpha 2$ too large to carry out the interpolation. The choice of values for this book ($B = 2000, \alpha \geq 0.025$), however, prohibits this problem. See Efron and Tibshirani (1993: Sect. 14.7 therein) and Davison and Hinkley (1997: Sect. 5.3.2 therein) on the interpolation pitfall, and further Andrews and Buchinsky (2000, 2002) on the choice of B. Refer to Polansky (1999) on the finite sample bounds on coverage for percentile-based CIs. As regards estimation of the acceleration, possible alternatives to Eq. (3.38) are analysed by Frangos and Schucany (1990).

The **balanced bootstrap** (Davison et al. 1986) is a bootstrap variant where over all $n \cdot B$ resampling operations, each of the values $\{x(i)\}_{i=1}^n$ is prescribed to be drawn equally often (B times). This can increase the accuracy of bootstrap estimates or, instead, allow to reduce B with the same accuracy as when using the "unbalanced" bootstrap with a higher number of resamples. In the case of a process without serial dependence, a simple algorithm for a balanced version of the ordinary bootstrap is as follows (Davison and Hinkley 1997: Sect. 9.2.1 therein). Step 1. Concatenate B copies of $\{x(i)\}_{i=1}^n$ into a single set $\mathbf{S}$ of size $n \cdot B$. Step 2. Permute the elements of $\mathbf{S}$ at random and call this set $\mathbf{S}^*$. Step 3. For $b = 1, \ldots, B$, take successive sets of n elements of $\mathbf{S}^*$ as balanced resamples $\{x^{*b}(i)\}_{i=1}^n$. In the case of serial dependence, a balanced version of the MBB would permute blocks of elements of $\mathbf{S}$. A reduced number of resamples, B, means reduced computing costs for the balanced bootstrap. How large this gain is depends on the type of estimation. The gain may not be large for quantile estimation (Davison and Hinkley 1997), which is required in BCa CI construction (Sect. 3.4.4).

2SAMPLES (Mudelsee and Alkio 2007) is a Fortran 90 program for performing comparisons of location measures (mean and median) and variability measures (standard deviation and MAD) between two samples. The difference measures are estimated with BCa CI. It is freely available from the web site for this book (9 November 2013).

boot is an R package implementing the functions (and datasets) from the book by Davison and Hinkley (1997). It is available from the site http://cran.r-project.org/web/packages/boot/ (9 November 2013).

Resample is a Windows software that is freely available for download on http://woodm.myweb.port.ac.uk/nms/resample.htm (9 November 2013).

Good (2005) is a reference where routines for bootstrap resampling, BCa and bootstrap-t CI construction can be found. Also two- and multi-sample comparisons are included. The following languages/environments are supported: C++, EViews, Excel, GAUSS, Matlab, R, Resampling Stats, SAS, S-Plus and Stata.

A **Matlab/R** computer code for practical implementation of the block length selector of Politis and White (2004) can be downloaded from http://econ.duke.edu/~ap172/ (9 November 2013).

Resampling Stats is a resampling software purchasable as standalone, Excel and Matlab versions from http://www.resample.com (9 November 2013).

Shazam is a commercial econometrics software that includes bootstrap resampling (http://shazam.econ.ubc.ca, 9 November 2013).

SPSS is a tool that includes bootstrap resampling and CI construction (http://www-01.ibm.com/software/analytics/spss/products/statistics/, 9 November 2013).