Chapter 19
Data Analysis

19.1 Introduction

It is the aim of this chapter to present some of the most important techniques of statistical data analysis which is of interest for experimental as well as theoretical sciences. In particular, the superstition that numerically generated data sets do not need to be analyzed with statistical methods is certainly not justified if the data was generated by Monte Carlo methods. Some simple methods of statistical analysis have already been discussed in previous chapters. For instance, in Chap. 12 we discussed simple quality tests for random number generators, in Chap. 15 we calculated the errors associated with the observables of the ISING model. Here, these simple methods will be summarized and some more advanced techniques will be introduced on a basic level. For a more advanced discussion of this topic we refer the interested reader to Refs. [1–5].

19.2 Calculation of Errors

We repeat briefly the basics of simple estimators which we made use of previously. We approximate the expectation value $\langle x \rangle$ of some variable x

$$\langle x \rangle = \int \mathrm{d}x \, x p(x) \, , \tag{19.1}$$

where $p(x)$ is a pdf, by its arithmetic mean

$$\langle x \rangle \approx \overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \, , \tag{19.2}$$

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8_19

where the numbers x_i follow the distribution $p(x)$. It is of conceptual importance to distinguish between the expectation value $\langle x \rangle$ which is a c-number, while the estimator $\overline{x}$ is a random number fluctuating around $\langle x \rangle$. The error of approximating $\langle x \rangle$ by $\overline{x}$ can be estimated by calculating the variance

$$\text{var}\,(\overline{x}) = \frac{\text{var}\,(x)}{N} = \frac{\left\langle x^2 \right\rangle - \langle x \rangle^2}{N} , \tag{19.3}$$

if the random numbers x_i are uncorrelated (see Appendix E). In case of correlated data the treatment becomes more involved and this will be discussed in Sect. 19.3. The expectation values $\left\langle x^2 \right\rangle$ and $\langle x \rangle$ in Eq. (19.3) may again be replaced by the corresponding estimators $\overline{x^2}$ and $\overline{x}$ in order to obtain a reasonable estimate of the variance var $(\overline{x})$. In particular, we approximate

$$\left\langle x^2 \right\rangle \approx \overline{x^2} = \frac{1}{N} \sum_{i=1}^{N} x_i^2 . \tag{19.4}$$

This approximation has already been applied in our investigation of the Ising model, Chap. 15. When dealing with Markov-chain Monte Carlo simulations, the result (19.3) can be interpreted in a rather trivial way: Repeating the simulation under identical conditions results in roughly 68 % of all simulations to yield a mean value $\overline{x} \in [\overline{x} - \sigma_{\overline{x}}, \overline{x} + \sigma_{\overline{x}}]$, where $\sigma_{\overline{x}} = \sqrt{\text{var}\,(\overline{x})}$ is the standard error.

We consider now the, in the meanwhile, quite familiar situation in which the underlying pdf $p(x)$ of a sequence of random numbers $\{x_i\}$ is unknown. In such a case one cannot simply use a particular estimator without some knowledge of the particular form of $p(x)$. A common way to proceed is the *poor person's assumption*: The underlying distribution is symmetric. This assumption has its origin in the central limit theorem (see Appendix, Sect. E.8). However, some intuitive checks may be required if fatal misconceptions are to be avoided. Is the data set reasonably large one can retrieve essential information from collecting the data points in form of a histogram or, if the index i refers to time instances, by plotting a time sequence.

We can deduce a first idea about the form of the underlying pdf from a histogram. For instance, if the data set displays only one peak, as in Fig. 19.1, quantities like the mean or the variance could be useful. But if there are two (or more) separate peaks, as in Fig. 19.2, it does not necessarily make sense to calculate the mean or variance by summing over all the data points. Such a situation can, for instance, occur in statistical spin models, with two phases, as we observed it in the q-state Potts model, Fig. 18.7a, b.

Time series, in which the data points x_i are plotted as a function of discrete time instances t_i, can also reveal important information about the properties of the data set. For instance, systematic trends, outliers, or hints for correlations may be observed.

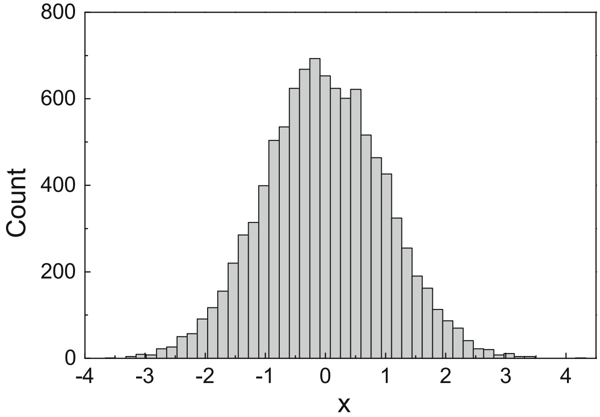

Fig. 19.1 Histogram generated by random sampling of a Gaussian of mean zero and variance one

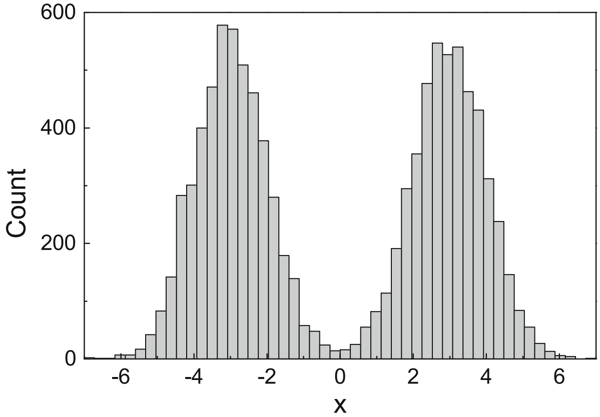

Fig. 19.2 Histogram generated by random sampling of two Gaussians of mean zero and variance one, displaced by $+3$ and -3, respectively

Let us turn our attention to some more advanced estimator techniques. So far we discussed the sample mean and sample variance as candidates for unbiased estimators.[1] In a more general context the calculation of observables from data sets might be more complex. In the following we assume a data set of N data points $(x_1, x_2, \ldots, x_N)$. Basically, we would like to estimate a quantity of the form $f(\langle x \rangle)$

[1] Since mean and variance are calculated from the same data points, they are usually not unbiased. Therefore a common choice is the so called *bias corrected variance* $\text{var}\,(\overline{x})_B$ which is given by $\text{var}\,(\overline{x})_B = \frac{N}{N-1}\text{var}\,(\overline{x})$ where N is the number of data points. A more detailed discussion can be found in any textbook on statistics [6–9].

where f is some particular function (for instance $\langle x\rangle^2$). A bad (*biased*) estimate would be to calculate

$$\bar{f} = \frac{1}{N}\sum f(x_i) \,, \tag{19.5}$$

which is definitely not the quantity we are interested in because for $N \to \infty$ we have $\bar{f} \to \langle f\rangle$ and not $f(\langle x\rangle)$. A better estimate would be to calculate

$$f(\bar{x}) = f\left(\frac{1}{N}\sum x_i\right) \,, \tag{19.6}$$

which converges to $f(\langle x\rangle)$ for $N \to \infty$. We discuss here two different methods to calculate the error attached to $f(\bar{x})$, namely the *Jackknife* method and the *statistical bootstrap* method.

We define Jackknife averages

$$x_i^J = \frac{1}{N-1}\sum_{j\neq i} x_j \,, \tag{19.7}$$

and x_i^J is the average of all values $x_j \neq x_i$. Moreover, we define

$$f_i^J \equiv f(x_i^J) \,, \tag{19.8}$$

and this opens the possibility to estimate $f(\langle x\rangle)$ following

$$f(\langle x\rangle) \approx \bar{f}^J = \frac{1}{N}\sum_i f_i^J \,, \tag{19.9}$$

with the statistical error

$$\sigma_{\bar{f}^J}^2 = (N-1)\left[\overline{(f^J)^2} - (\bar{f}^J)^2\right] \,, \tag{19.10}$$

which can be written as

$$\sigma_{\bar{f}^J}^2 = \frac{N-1}{N}\sum_i (f_i^J - \bar{f}^J)^2 \,, \tag{19.11}$$

for uncorrelated f_i^J (see Appendix E).

In the case of the statistical bootstrap we consider again a set of N data-points $\{x_i\}$. We randomly choose N elements from this data set *without removal* which constitutes the set $\{x_j^{(i)}\}$ and calculate for these N points the observable $f_i = f(1/N\sum_j x_j^{(i)})$. This procedure is repeated M-times and we get

$$f(\langle x\rangle) \approx \bar{f}_{BS} = \frac{1}{M}\sum_i f_i \,, \tag{19.12}$$

and

$$\sigma^2_{\bar{f}_{BS}} = \frac{1}{M}\sum_i \left(f_i - \bar{f}_{BS}\right)^2 \ . \tag{19.13}$$

This method was applied in Chap. 15 to determine estimates for the error-bars of the various observables as a function of temperature in Fig. 15.6. The methods discussed here can, of course, also be employed to derive estimates for the errors attached to the various observables studied in the POTTS model, Chap. 18.

Let us close this section with a short comment on systematic errors. As already highlighted within Chap. 1 one also has to be aware of possible systematic errors. Like in experimental data, these errors are more easily overlooked in numerical data since they are rather hard to identify. In general, there is no method available to investigate systematic errors. For instance, in the simulation of the ISING model, the main source of errors was that the MARKOV-chain was not allowed to completely equilibrate which would have been equivalent to running the simulation forever. The introduction of the concept of an auto-correlation time will, at least, allow for a systematic investigation of this fundamental problem.

19.3 Auto-Correlations

The situation becomes more involved whenever the random numbers of the sequence $\{x_i\}$ are correlated, i.e. $\text{cov}\left(x_i, x_j\right) \neq 0$ for $i \neq j$ [see Appendix, Eq. (E.16)], where the elements of the series $\{x_i\}$ are successive members of a time series. Hence, existing covariances between elements x_i and x_j account for auto-correlations of a certain observable between different time steps. We rewrite Eq. (19.3):

$$\begin{aligned}\text{var}\left(\bar{x}\right) &= \left\langle \overline{x^2}\right\rangle - \left\langle \bar{x}\right\rangle^2 \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\left\langle x_i x_j\right\rangle - \frac{1}{N^2}\sum_{i,j=1}^{N}\left\langle x_i\right\rangle\left\langle x_j\right\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\left\langle x_i^2\right\rangle - \left\langle x_i\right\rangle^2\right) \\ &\quad + \frac{1}{N^2}\sum_{i\neq j}\left(\left\langle x_i x_j\right\rangle - \left\langle x_i\right\rangle\left\langle x_j\right\rangle\right) \ . \end{aligned} \tag{19.14}$$

The first term on the right-hand side of Eq. (19.14) is identified as $\mathrm{var}\,(x_i)\,/N$ which is assumed to be identical for all i, i.e. $\mathrm{var}\,(x_i) \equiv \mathrm{var}\,(x)$. Furthermore, we rewrite the sum

$$\sum_{i \neq j} \cdot = 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \cdot \,,$$

and obtain

$$\mathrm{var}\,(\overline{x}) = \frac{1}{N}\left[\mathrm{var}\,(x) + \frac{2}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathrm{cov}\left(x_i, x_j\right)\right] . \qquad (19.15)$$

Let us assume time translational invariance:

$$\mathrm{cov}\left(x_i, x_j\right) \equiv C(j-i)\ , \quad \text{for}\ j > i. \qquad (19.16)$$

We apply this relation to Eq. (19.15) and obtain

$$\begin{aligned}
\mathrm{var}\,(\overline{x}) &= \frac{1}{N}\left[\mathrm{var}\,(x) + \frac{2}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} C(j-i)\right] \\
&= \frac{1}{N}\left[\mathrm{var}\,(x) + \frac{2}{N}\sum_{k=1}^{N} C(k)\,(N-k)\right] \\
&= \frac{1}{N}\left[\mathrm{var}\,(x) + 2\sum_{k=1}^{N} C(k)\left(1-\frac{k}{N}\right)\right] , \qquad (19.17)
\end{aligned}$$

which can be reformulated as:

$$\mathrm{var}\,(\overline{x}) = \frac{2\mathrm{var}\,(x)\,\hat{\tau}_x^i}{N} . \qquad (19.18)$$

We introduced here the (proper) *integrated auto-correlation time* $\hat{\tau}_x^i$

$$\hat{\tau}_x^i = \frac{1}{2} + \sum_{k=1}^{N} A(k)\left(1-\frac{k}{N}\right) , \qquad (19.19)$$

and the normalized auto-correlation function

$$A(k) = \frac{C(k)}{C(0)} = \frac{\mathrm{cov}\,(x_i, x_{i+k})}{\mathrm{var}\,(x_i)} . \qquad (19.20)$$

In most cases we are interested in the limit $N \to \infty$ of Eq. (19.19):

$$\tau_x^i = \lim_{N\to\infty} \hat{\tau}_x^i = \frac{1}{2} + \sum_{k=1}^{\infty} A(k) \ . \tag{19.21}$$

The form of the auto-correlation function $A(x)$ can be approximated using the results of Sect. 16.4. There, we observed that the stationary distribution π was the left-eigenvector of the transition matrix P with eigenvalue 1, Eq. (16.73). Let $\{\varphi_\ell\}$ denote the set of all left-eigenvectors of the matrix P with eigenvalues λ_ℓ, i.e. $\varphi_\ell P = \lambda_\ell \varphi_\ell$.[2] Then some arbitrary state $q(0)$ can be expressed in this basis as:

$$q(0) = \sum_i \alpha_i \varphi_i \ . \tag{19.22}$$

After n consecutive time-steps we arrive at state $q(n)$

$$q(n) = q(0)P^n = \sum_i \alpha_i \varphi_i P^n = \sum_i \alpha_i \lambda_i^n \varphi_i \ , \tag{19.23}$$

which follows from Eq. (16.62). We denote the observable we want to calculate by $O(n)$ and expand it according to Ref. [10]

$$O(n) = \sum_i [q(n)]_i o_i = \sum_i \alpha_i \lambda_i^n o_i \ , \tag{19.24}$$

where o_i stands for the expectation value of O in the i-th eigenstate φ_i. For large n the value of $O(n)$ will be dominated by the largest eigenvalue of P, say λ_0, and we denote this value by $O(\infty) = \alpha_0 o_0$. This allows us to rewrite Eq. (19.24) as

$$O(n) = O(\infty) + \sum_{i\neq 0} \alpha_i o_i \lambda_i^n \ . \tag{19.25}$$

Let $\lambda_1 \in \mathbb{R}$ be the second largest eigenvalue and let us define the *exponential auto-correlation time* τ_x^e via

$$\tau_x^e = -\frac{1}{\log(\lambda_1)} \ , \tag{19.26}$$

[2] Note that since P is a stochastic matrix, it follows that $|\lambda_\ell| \leq 1$ for all ℓ. Furthermore, it can be shown that the largest eigenvalue of a stochastic matrix is equal to 1.

and the value of $O(n)$ can, for large values of n, be approximated by

$$O(n) \approx O(\infty) + \beta \exp\left(-\frac{n}{\tau_x^e}\right) , \tag{19.27}$$

where β is some constant. Hence, the auto-correlation obeys

$$C(n) \propto [O(0) - O(\infty)] [O(n) - O(\infty)] \propto \beta \exp\left(-\frac{n}{\tau_x^e}\right) , \tag{19.28}$$

and we can simply set for the auto-correlation function $A(k)$

$$A(k) = \gamma \exp\left(-\frac{k}{\tau_x^e}\right) , \tag{19.29}$$

where γ is some constant. We use this result in the expression for the integrated auto-correlation time (19.21) and arrive at:

$$\begin{aligned} \tau_x^i &= \frac{1}{2} + \gamma \sum_{k=1}^{\infty} \left[\exp\left(-\frac{1}{\tau_x^e}\right)\right]^k \\ &= \frac{1}{2} + \gamma \frac{\exp\left(-\frac{1}{\tau_x^e}\right)}{1 - \exp\left(-\frac{1}{\tau_x^e}\right)} . \end{aligned} \tag{19.30}$$

For $\tau_x^e \gg 1$ the exponential function can be expanded into a Taylor series. Keeping terms up to first order results in:

$$\tau_x^i = \frac{1}{2} + \gamma \frac{1 - \frac{1}{\tau_x^e}}{\frac{1}{\tau_x^e}} = \frac{1}{2} + \gamma \left(\tau_x^e - 1\right) \propto \gamma \tau_x^e . \tag{19.31}$$

However, we note that in general relation (19.31) is only a poor approximation because usually the exponential auto-correlation time is very different from the integrated auto-correlation time.

Let us briefly discuss our results. A comparison between Eqs. (19.3) and (19.18) reveals that due to correlations in the time series, the number of effective (or useful) data points N_{eff} can be determined from

$$N_{\text{eff}} = \frac{N}{2\tau_x^i} . \tag{19.32}$$

In the limit $\tau_x^e \to 0$ we obtain $\tau_x^i = 1/2$ and, thus, recover Eq. (19.3). The effective number of measurements is the relevant quantity whenever the error of a Monte Carlo integration is calculated.

In another approach, one can determine the exponential auto-correlation time τ_x^e and use it to estimate the number of steps that should be neglected between two successive measurements. This can be achieved by fitting the auto-correlation $A(k)$ with an exponential function. (A brief introduction to least squares fits can be found in Appendix H.) We note that in one and the same system the auto-correlation times may be very different for different observables.

19.4 The Histogram Technique

The histogram technique is a method which allows to approximate the expectation value of some observable for temperatures near a given temperature T_0 without performing further MARKOV-chain Monte Carlo simulations. The basic idea is easily sketched. Suppose the observable O is solely a function of energy E. We perform a MARKOV-chain Monte Carlo simulation for a given temperature T_0 and measure the energy E several times. The resulting measurements are sorted in a histogram with bin width ΔE as was demonstrated in Sect. 18.3. If $n(E)$ denotes the number of configurations measured within the interval $(E, E + \Delta E)$, then the probability that some energy is measured to lay within the interval $(E, E + \Delta E)$ is given by

$$P_H(E, T_0) = \frac{n(E)}{M} \,, \tag{19.33}$$

where the index H refers to histogram and $M = \sum_E n(E)$ is the number of measurements. However, we note that this probability can also be expressed by the BOLTZMANN distribution

$$P(E, T) = \frac{N(E) \exp\left(-\frac{E}{k_\mathrm{B}T}\right)}{\sum_E N(E) \exp\left(-\frac{E}{k_\mathrm{B}T}\right)} \,, \tag{19.34}$$

where $N(E)$ denotes the number of micro-states within the interval $(E, E + \Delta E)$. $N(E)$ is independent of the temperature T and relation (19.34) is valid for all temperatures T. In particular, for $T = T_0$

$$P_H(E, T_0) = P(E, T_0) \,, \tag{19.35}$$

which immediately yields

$$N(E) = \alpha n(E) \exp\left(\frac{E}{k_\mathrm{B}T_0}\right) \,, \tag{19.36}$$

where α is some constant and we emphasize that $n(E)$ was measured at T_0. Inserting Eq. (19.36) into (19.34) yields

$$P(E,T) = \frac{n(E)\exp\left[-\left(\frac{1}{k_\mathrm{B}T} - \frac{1}{k_\mathrm{B}T_0}\right)E\right]}{\sum_E n(E)\exp\left[-\left(\frac{1}{k_\mathrm{B}T} - \frac{1}{k_\mathrm{B}T_0}\right)E\right]} , \tag{19.37}$$

for arbitrary T. The expectation value $\langle O\rangle_T$ of the observable O at some temperature T can now be determined from

$$\begin{aligned}\langle O\rangle_T &= \sum_E O(E)P(E,T)\\ &= \frac{\sum_E O(E)n(E)\exp\left[-\left(\frac{1}{k_\mathrm{B}T} - \frac{1}{k_\mathrm{B}T_0}\right)E\right]}{\sum_E n(E)\exp\left[-\left(\frac{1}{k_\mathrm{B}T} - \frac{1}{k_\mathrm{B}T_0}\right)E\right]} .\end{aligned} \tag{19.38}$$

This result implies, that it is not necessary to run an additional MARKOV-chain Monte Carlo simulation in an attempt to compute the expectation value $\langle O\rangle_T$ for temperature T if T is in the vicinity of T_0. However, if T deviates strongly from T_0, the above procedure (19.38) does not provide a good approximation because the relevant configurations at T may have been very improbable at T_0 and may, therefore, not have been reproduced sufficiently often in the original MARKOV-chain Monte Carlo simulation.

Summary

Data analysis is an important but often neglected part of natural sciences and in particular of numerical simulations. It consists mainly of consistency checks and error analysis. This chapter concentrated in a first step on error analysis. It discussed the most common methods to arrive at an estimate of the error involved whenever expectation values of some property are analyzed. These went beyond all those methods which have already been discussed in some detail throughout this book. In a second step auto-correlations have been discussed. They should be part of consistency checks and give valuable information about possible systematic errors. The auto-correlation analysis was of particular importance whenever the quality of the sequence of random numbers was crucial to a particular simulation. (Experiments in which the events are expected to be random, like radioactive decay, fall also into this category.) Nevertheless, this method proved to be very useful in MARKOV-chain Monte Carlo simulations as it allowed to define and determine an auto-correlation time which could serve as a measure of the number of sweeps which have to be neglected between two consecutive measurements. Finally, the histogram technique was introduced as a method of data interpolation. It allowed, in addition to applications which have already been presented within this book, to derive the expectation value of some property at some 'temperature' T from the already known expectation value of this same property at some other temperature T_0 if $T \sim T_0$ and if the equilibrium distribution was known.

Problems

1. Calculate the auto-correlation function for random numbers generated by the two linear congruential generators discussed in Sect. 12.2. Check also the random number generator provided by your system. Discuss the results.
2. POTTS model: Calculate the error attached to the specific heat c_h and the susceptibility χ using the Jackknife method for all values of $q = 1, \ldots, 8$. Plot the corresponding diagrams and discuss the results. Determine the exponential and integrated correlation time.

References

1. Gaul, W., Opitz, O., Schader, M. (eds.): Data Analysis. Springer, Berlin/Heidelberg (2000)
2. Sivia, D., Skilling, J.: Data Analysis, 2nd edn. Oxford University Press, Oxford (2006)
3. Adèr, H.J., Mellenbergh, G.J., Hand, D.J. (eds.): Advising on Research Methods: A Consultant's Companion, chap. 14, 15. Johannes van Kessel, Huizen (2008)
4. Brandt, S.: Data Analysis. Springer, Berlin/Heidelberg (2014)
5. von der Linden, W., Dose, V., von Toussaint, U.: Bayesian Probability Theory. Cambridge University Press, Cambridge (2014)
6. Iversen, G.P., Gergen, I.: Statistics. Springer Undergraduate Textbooks in Statistics. Springer, Berlin/Heidelberg (1997)
7. Wilcox, R.R.: Basic Statistics. Oxford University Press, New York (2009)
8. Monahan, J.F.: Numerical Methods of Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2011)
9. Wood, S.: Core Statistics. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge (2015)
10. Sokal, A.D.: Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. Department of Physics. New York University, New York (1996). www.stat.unc.edu/faculty/cji/Sokal.pdf