1 Statistical Inference

This first chapter will briefly review basic statistical concepts, ways of thinking, and ideas that will reoccur throughout the book, as well as some general principles and mathematical techniques for handling these. In this sense it will lay out some of the ground on which statistical methods developed in later chapters rest. It is assumed that the reader is basically familiar with core concepts in probability theory and statistics, such as expectancy values, probability distributions like the binomial or Gaussian, Bayes' rule, or analysis of variance. The presentation given in this chapter is quite condensed and mainly serves to summarize and organize key facts and concepts required later, as well as to put special emphasis on some topics. Although this chapter is self-contained, if the reader did not pass through an introductory statistics course so far, it may be advisable to consult introductory chapters in a basic statistics textbook first (very readable introductions are provided, for instance, by Hays, 1994, or Wackerly et al., 2008; Kass et al., 2014, in particular, give a highly recommended introduction specifically targeted to a neuroscience readership). More generally, it is remarked here that the intention of the first six chapters was more to extract and summarize essential points and concepts from the literature referred to.

Statistics and statistical inference, in its essence, deals with the general issue of inferring in a defined sense the most likely state of affairs in an underlying population from a usually much smaller sample. That is, we would like to draw valid conclusions about a much larger unobserved population from the observation of just a tiny fraction of its members, where the 'validity' of the conclusions is formally judged by certain statistical criteria to be introduced below. It is clear that this endeavor rests in large part on probability theory which forms the fundament of all statistics: Empirical observations are essentially a collection of random variables from which we compute certain functions (called *statistics*) like the mean or variance which should be 'maximally informative' (see sect. 1.2) about the underlying population. Probability theory is usually treated in any introductory textbook on statistics and will not be covered here (see Hays 1994, Wackerly et al. 2008, Kass et al. 2014).

There is a huge body of work in *theoretical* (also sometimes called mathematical) statistics which deals with properties of probability distributions such as the distribution of functions of random variables (like *statistics*) and methods of how these could be derived. There are also a number of important theorems and lemmata (like the Rao-Blackwell theorem or the Neyman-Pearson lemma) which establish which kind of statistics and hypothesis tests possess 'optimal' (see sects. 1.2, 1.5) properties with regards to inference about the population. A very readable and mathematically low-key introduction to this whole field is provided by Wackerly et al. (2008; a mathematically more sophisticated presentation is given in Keener 2010).

While most of this book is focused on applied statistics, this first chapter will review some important results, concepts, and definitions from theoretical statistics. We will start with a discussion of statistical models which are at the heart of many of the most commonly applied statistical procedures.

1.1 Statistical models

In statistics we often formulate (commonly simple) mathematical models of experimental situations to infer some general properties of the underlying population or system which generated the data at hand. Statistical inference denotes this process by which we infer from a sample $\mathbf{X}=\{\mathbf{x}_i\}$ of observations, (population) parameters of a (supposedly) underlying model or distribution (Fig. 1.1), or test hypotheses about model parameters or other statistics. In classical statistics, the basic currency in model fitting (estimation) and testing is most commonly *variance* (a consequence of the normal distribution assumption commonly employed): Statistical models consist of a *structural* (*systematic*) part that is supposed to explain observed variation in the data, and a *random* (*error*) part that captures all the influences the model cannot account for. In this first section, we will walk through a set of quite different examples, motivated by specific experimental questions, to illustrate the concept of a statistical model from various angles.

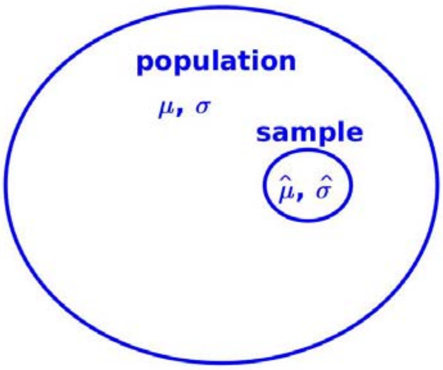

Fig. 1.1. In statistics we are usually dealing with small samples from (vastly) larger populations, and the task is to infer trustable properties (parameters) of the population from just the small sample at hand. **MATL1_1**.

Example 1.1. In a one-factor univariate analysis of variance (ANOVA) setting we observe data x_{ij} under different treatment conditions j from different subjects i, as illustrated in Table 1.1. To give a concrete example, assume we want to pursue the question of what role different synaptic receptor types (like NMDA, GABA$_A$, etc.) in hippocampus play in spatial learning. Learning performance could be measured, e.g., by the number of trials it takes an animal to reach a defined performance criterion (dependent variable), or by the time it takes the animal to find some hidden, to-be-memorized target, like the underwater platform in a Morris water maze (Morris 2008). Experimentally, one may manipulate synaptic receptors through genetic engineering (independent variable/factor), e.g. by knocking out or down genes coding for subcomponents of receptors of interest. We may now postulate that our sample observations $\{x_{ij}\}$, i.e. the memory scores as defined above for subjects i from genetic strain j, are composed as follows (Winer 1971):

$$(1.1)\quad \begin{array}{ll} (structural\ part) & x_{ij}=\mu+\tau_j+\varepsilon_{ij}\ ,\ for\ i=1..n, j=1..K, \\ (random\ part) & \varepsilon_{ij}\sim N(0,\sigma^2)\ ,\ E(\varepsilon_{ij}\varepsilon_{kl})=0\ for\ (i,j)\neq(k,l)\ , \end{array}$$

where the tilde '~' reads as 'distributed according to', and $N(\mu,\sigma^2)$ denotes the normal distribution with parameters μ (mean) and σ (standard deviation). That is, we assume that each observation x_{ij} is given by the sum of a grand (population) mean μ, a treatment effect τ_j specific for each of the K treatment conditions (but general across individuals within a treatment group), and an individual error term (random variable) ε_{ij} (with common variance σ^2 across individuals and conditions). The treatment effects τ_j account for the systematic (explainable) variation of the x_{ij}, i.e., in the example above, the systematic deviation from the grand mean caused by the manipulation of gene *j* (these terms, weighted by the relative number of observations in each treatment group, have thus to sum up to zero), while the ε_{ij} represent the unaccountable (noise) part.

	Treatment condition (e.g. pharmacological treatment)		
Subject/ observation	A	B	C
1	x_{11}	x_{12}	x_{13}
2	x_{21}	x_{22}	x_{23}
3	x_{31}	x_{32}	x_{33}
4	x_{41}	x_{42}	x_{43}
…	…	…	…
n	x_{n1}	x_{n2}	x_{n3}
	$E[x_{\cdot 1}] = \mu + \tau_1$	$E[x_{\cdot 2}] = \mu + \tau_2$	$E[x_{\cdot 3}] = \mu + \tau_3$

Table 1.1. One-factor ANOVA setting. Bottom row expresses the model assumptions.

A key to parametric statistical inference and hypothesis testing is to formulate specific distributional assumptions for the unknown error terms (or, more generally, the involved random variables). In ANOVA settings we usually assume, as in (1.1) above, that the error terms follow a normal distribution with mean 0 and standard deviation σ which needs to be estimated from the data. We furthermore assume that the individual error terms are mutually *uncorrelated*, i.e. E($\varepsilon_{ij}\varepsilon_{kl}$)=0 for $(i,j) \neq (k,l)$, which under normal distribution assumptions is equivalent to assuming independence (cf. sect. 6.6; this is because a normal distribution is completely specified by its first two statistical moments, the mean and the variance, or covariance matrix in the multivariate case). Random variables which fulfill these conditions, i.e. come all from the *same* distribution and are *independent*, are said to be '*identically and independently distributed*' (*i.i.d.*).

The justification for the frequent assumption of normally distributed errors comes from the *central limit theorem* (see sect. 1.5.2 below) which states that a sum of random variables converges to the normal distribution for large *n*, almost *regardless* of the form of the distribution of the random variables themselves. The error terms ε_{ij} may be thought of as representing the sum of many independent error sources which on average cancel out, thus $\varepsilon_{ij} \sim N(0,\sigma^2)$ (Winer 1971). However, to draw conclusions from a sample, it is crucially important to be aware of the fact that the inferences we make are usually based on a specific model with specific assumptions that could well be violated. In the ANOVA case, for instance, these include the linearity of model (1.1) and the assumption of independently and normally distributed errors: Errors may, for instance, be multiplicative or some more complex function of treatment condition, and either way they may not be normal or i.i.d..

<u>*Example 1.2.*</u> As another example, for a two-factor ANOVA design with observed sample $\{x_{ijk}\}$, i=1…n, j=1…J, k=1…K, we may formulate the model (Winer 1971, Hays 1994)

$$(1.2)\quad x_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}\ , \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

with $\sum_j \alpha_j = \sum_k \beta_k = \sum_j \alpha\beta_{jk} = \sum_k \alpha\beta_{jk} = 0$, **I** an identity matrix of size $n \times J \times K$ (number of subjects/group times number of factor level combinations), $\boldsymbol{\varepsilon} = (\varepsilon_{111},\ldots,\varepsilon_{ijk},\ldots,\varepsilon_{nJK})^T$ the vector of subject-specific error terms, and **0** a vector of zeros of same size as $\boldsymbol{\varepsilon}$. Thus in this case we assume that the deviations from the grand mean μ are caused by the sum of two different treatment conditions α_j and β_k, plus a term $\alpha\beta_{jk}$ that represent the *interaction* between these two specific treatments, and of course the error terms again [the distributional assumption for $\boldsymbol{\varepsilon}$ in (1.2) summarizes both the Gaussian as well as the independence assumption]. For instance, in the empirical situation from Example 1.1, we may further divide our group of animals by gender (factor β), enabling us to look for gender-specific effects of the genetic manipulations (factor α), where the gender-specificity (the differential impact of the genetic change on gender) would be expressed through the interaction term $\alpha\beta$.

Example 1.3. Instead of the categorical (factorial) models above, we may perhaps have observed pairs $\{(x_i, y_i)\}$ where both x_i and y_i are continuous variables. For instance, y_i may be the firing rate of a neuron which we would like to relate to the spatial position x_i of a rat on a linear track, the running speed in a treadmill, or the luminance of a visual object. (For now we will largely leave aside issues about the scale and appropriate distributional assumptions for the random variables in the empirical examples. For instance, firing rates empirically are often given as positive count [integer-scale, histogram] variables, although one may also define them as interval-scale variables based on averages across trials, or based on the inverse of inter-spike-intervals.) More specifically, we may want to postulate that the x_i and y_i are *linearly* related via

(1.3) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ ,} \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$,

where β_0, β_1 are model parameters. This brings us into the domain of linear regression. In neuroscience, the question of how spike rate y relates to environmental or mental variables x is also called an 'encoding' problem, while the reverse case, when sensory, motor, or mental attributes are to be inferred (predicted) from recorded neural variables, is commonly called a 'decoding' problem.

Example 1.4. Or perhaps, based on inspection of the data, it seems more reasonable to express the y_i in terms of powers of the x_i, for instance:

(1.4) $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \text{ ,} \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

Taking the linear track example from Example 1.3, the firing rate y of the neuron may not monotonically increase with position x on the track, but may exhibit a bell-shaped dependence as in hippocampal place fields (Fig. 1.2; O'Keefe 1976; Buzsaki & Draguhn 2004). As shown later, without introducing much additional computational burden, we could in fact express the y_i in terms of arbitrary functions of the x_i, called a *basis expansion* (sect. 2.6), as long as the right hand side stays *linear in the parameters* β_i.

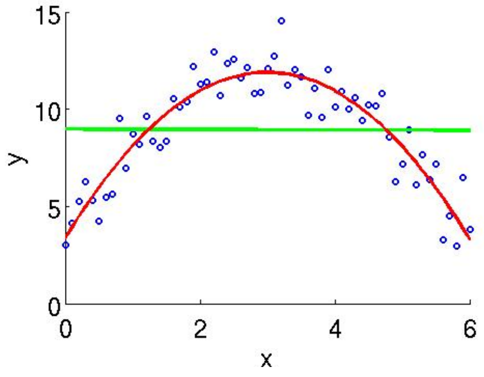

Fig. 1.2. Linear (green) vs. quadratic (red) model fit to data (blue circles) exhibiting a bell-shaped dependence. **MATL1_2**.

Models of the form (1.1), (1.2) and (1.3), (1.4), are combined and unified within the framework of *general linear models* (GLM). In a general-linear-model setting, categorical variables would be incorporated into the regression model by dummy-coding them through binary vectors (e.g. with a '1'-entry indicating that a particular experimental condition was present, while '0' indicating its absence; see sect. 2.1).

Example 1.5. Instead of taking nonlinear functions of just regressors x_i on the right hand side, let us assume that we have a function *f nonlinear* but invertible *in parameters* β_i themselves. Suppose our observations y_i are furthermore behavioral error *counts*, e.g. from a recognition memory task, thus not well captured by a Gaussian distribution. The regressors x_i may, for instance, represent the concentration of an administered drug hypothesized to affect memory performance. If the error probability *p* is generally small, the $y_i \in \mathrm{N}_0$ could be approximated by a Poisson distribution with mean μ_i depending on drug concentration x_i in a nonlinear way:

$$(1.5)\quad \begin{aligned} &f(\mu_i) = \beta_0 + \beta_1 x_i \\ &pr(Y = y_i \mid x_i) = \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}. \end{aligned}$$

For the latter expression we will adopt the notation $y_i \mid x_i \sim Poisson(\mu_i)$ in this book. If we assume the regressors x_i to be fixed (constant), a common choice in regression models, strictly, the terms $p(y_i \mid x_i)$ would not have the interpretation of a conditional probability. Function *f* in the first expression (the structural part) is also called a *link* function in statistics and extends the GLM class into the framework of general*ized* linear models (in addition to more flexible assumptions on the kind of distribution as in the example above; McCullagh & Nelder 1989; Fahrmeir & Tutz 2010). (Confusingly, the abbreviation 'GLM' is often used for both, general and generalized linear models. Here we will use it only for the *general* LM.) In this case, a particular *function of the response variable* or its conditional expectancy value $\mu_i := E[y_i \mid x_i]$ is still linearly related to the predictors, although overall the regression in parameters β_i becomes itself a nonlinear problem through f^{-1}, and explicit (analytical) solutions to (1.5) may no longer be available.

Example 1.6. There are also many common situations where the data $\{x_i\}$ are generally not i.i.d., time series for example (i.e. consecutive observations of some variable in time), like measurements of the membrane potential of a cell or of the local field potential (LFP) where consecutive values are strongly correlated (other example: spatial correlations in fMRI signals). A model for such observations could take the form

$$(1.6)\quad x_t = \alpha x_{t-1} + \varepsilon_t \text{ },\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

where *t* indexes time and α is a parameter. Time series models of this type which connect consecutive measurements in time by a linear function fall into a class called autoregressive (AR) models in the statistics literature (or, in this case, a linear map in the terminology of dynamical systems; see sects. 7.2, 9.1).

Example 1.7. Finally, as a simple multivariate example, we may collect N (= number of observations) calcium imaging frames from p (= number of variables) regions of interest (ROI) within an (N x p) data matrix **X**, and may propose that rows of **X** follow a multivariate normal distribution with mean (row) vector **μ** and covariance matrix **Σ**, written as

$$(1.7)\quad \mathbf{x}_{i\cdot} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_{i\cdot}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{i\cdot}-\boldsymbol{\mu})^T},$$

where |·| indicates the determinant of the matrix. Henceforth we will use the notation ‘ $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ ‘ not only to indicate the distribution object, but – as above – to refer to the density function itself.

1.2 Goals of model-based analysis and basic definitions

Having defined a model, one may have several goals in mind:

First, one may take the model as a compact description of the ‘state of affairs’ or ‘empirical laws’ in the population, and obtain *point estimates* like $\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\hat{\boldsymbol{\tau}},\hat{\boldsymbol{\beta}}$, of the unknown population or model parameters such as **μ**, **Σ**, **τ**, **β** in the examples above. Or one may want to establish *interval estimates* of an unknown parameter θ such that

$$(1.8)\ \theta \in [\hat{\theta}-c_L,\hat{\theta}+c_H]_{1-\alpha} := \{\theta \mid pr(\hat{\theta}-c_L \le \theta \le \hat{\theta}+c_H) \ge 1-\alpha\}$$

where $[\cdot]_{1-\alpha}$ is called the 1-α *confidence interval* with lower and upper bounds $\hat{\theta}-c_L$ and $\hat{\theta}+c_H$, respectively (cf. Winer 1971; Hays 1994; Wackerly et al. 2008).

Second, one may use such a model then for *prediction*, i.e. to predict properties of a previously unobserved, novel individual, for instance its response to a particular drug treatment, or, e.g., in *Example* 1.6 of an auto-regressive model, to perform forecasting in time. In that case, for instance, with x_t observed, $E[x_{t+1} \mid x_t] = E[\alpha x_t] + E[\varepsilon_t] = \alpha x_t$.

Third, one may want to test a hypothesis about model parameters τ_1, τ_2 (or just any statistic obtained from the data) like

(1.9) H_1: $\tau_1 \neq \tau_2$ (*alternative hypothesis*) vs. H_0: $\tau_1 = \tau_2$ (*null hypothesis*).

For instance, in the regression model (1.3) one may want to assess H_0: β_1=0, i.e. firing rate is not (linearly) related to spatial position or running speed in the particular example given, and contrast it with H_1: β_1>0, i.e. firing rate and spatial position are positively related. Say our empirical estimate for β_1 is $\hat{\beta}_1^{obs}$, then, in such a test scenario, we may define the one-tailed decision rule ‘accept H_1 if $p(\hat{\beta}_1 \ge \hat{\beta}_1^{obs}$ |‘H_0 true’) ≤ α’, where the α- (or type-I) error (*significance level*) is the probability of wrongly accepting the H_1 (although the H_0 is true; other acceptance or rejection regions, respectively, may be specified of course, depending on the precise form of our H_0). Conversely, the probability β := p(‘accept H_0’ | ‘H_1 true’) associated with our decision rule is called the β- (or type-II) error. The quantity 1-β is called the *power* (or sensitivity) of a test, and obviously it should be large. Fixing the α-level and desired power 1-β, for some hypothesis tests (e.g. those based on normal or t-distributions) one can, under certain conditions, derive the sample size required to perform the test with the requested power (see Winer 1971; Hays 1994, Wackerly et al. 2008).

Generally, throughout this book, we will use – as common practice in statistics – roman letters like t to denote a statistic obtained from a sample, Greek letters like θ to indicate the corresponding population parameter, and Greek letters with a 'hat' like $\hat{\theta}$ to denote empirical *estimates* of the true population parameter. The following definitions capture some basic properties of such parameter estimates, i.e. give criteria of what constitutes a 'good' estimate of a statistic (Fisher 1922; Winer 1971, Wackerly et al. 2008):

Def. 1.1, bias. Suppose we have $E(\hat{\theta}) = \theta + c$, where θ is the true population parameter and $\hat{\theta}$ its estimate, then c is called the *bias* of estimator $\hat{\theta}$. If c=0, then $\hat{\theta}$ is called *unbiased*. Thus, the bias reflects the systematic deviation of our average estimator from the true population parameter.

Def. 1.2, consistency. An estimator $\hat{\theta}$ is called *consistent* if it 'converges in probability' to the true population parameter θ (Wackerly et al. 2008): $\lim_{N\to\infty} \Pr(|\theta - \hat{\theta}_N| \le \varepsilon) = 1$ for any $\varepsilon > 0$, where we indicate the sample size dependence of the estimator by subscript N. Thus, for a consistent estimator, any bias should go away eventually as the sample size is increased (it should be 'asymptotically unbiased'), but at the same time the variation around the true population parameter should shrink to zero.

Def. 1.3, sampling distribution. The distribution $F_N(\hat{\theta})$ of parameter estimate $\hat{\theta}$ when drawing repeatedly samples of size N from the underlying population is called its *sampling distribution*.

Def. 1.4, standard error. The standard error of an estimator $\hat{\theta}$ is the standard deviation of its sampling distribution, defined as a function of sample size N: $SE_{\hat{\theta}}(N) := E[(\hat{\theta}_N - E(\hat{\theta}_N))^2]^{1/2}$.

For the standard error of the mean (*SEM*) we have the analytical expression $SE_{\hat{\mu}}(N) = \frac{\sigma}{\sqrt{N}}$. For the mean, since the sample mean $\bar{x}$ is an unbiased estimate of the population mean μ, one has $E[\hat{\mu}_N] = E[\bar{x}_N] = E[\sum x_i / N] = \mu$. Using this and the i.i.d. assumption in the expression above, one sees where the factor $1/\sqrt{N}$ in the SEM comes from: $\text{Var}[\bar{x}_N] = \text{Var}[\sum x_i / N] = 1/N^2 \sum \text{Var}[x_i] = N\sigma^2 / N^2$. The unbiased estimate for the variance from a sample with unknown population mean is $\hat{\sigma}^2 = N/(N-1)s^2$, with s being the sample standard deviation. Loosely, this is because the sample mean occurring in the expression for the sample variance is a random variable itself, and hence its own variance σ^2 / N contributes variation not accounted for in the sample estimate, so that the sample variance represents an underestimate (see Wackerly et al., 2008, for a derivation).

An overall measure of the accuracy of an estimate which accounts for *both* its (squared) *bias* and *variance* would be $E[(\hat{\theta}_N - \theta)^2]$, i.e. the total variation around the true population parameter (also called the 'mean squared error', MSE).

Def. 1.5, sufficiency. Loosely, a statistic (or set of statistics) is called *sufficient* if it contains all the information there is about a population in the sample, i.e. if we cannot learn anything else about the population distribution by calculating yet other sample statistics. More formally, a (set of) statistic(s) $t(\mathbf{X})$ is sufficient for θ if $p(\mathbf{X}|t,\theta) = p(\mathbf{X}|t)$, i.e. if the conditional probability of the data given t does not depend on parameters θ specifying the population distribution (Duda & Hart 1973; Berger 1985; Wackerly et al. 2008). There are usually different sets of statistics which may accomplish this, and the set which achieves this in the shortest way possible (minimum number of estimators) is called *minimally sufficient*. For instance, for a normally distributed population the sample mean and variance together are minimally sufficient, as the normal distribution is completely specified by these two parameters.

Def. 1.6, efficiency. The efficiency of some estimator $\hat{\theta}_k$ is defined with respect to the optimal estimator $\hat{\theta}_{opt}$ for which one achieves the lowest variance theoretically possible (Winer 1971):

$$Eff_{\hat{\theta}_k} = \frac{SE^2_{\hat{\theta}_{opt}}}{SE^2_{\hat{\theta}_k}} \in [0,1] .$$

The Rao-Blackwell theorem (Wackerly et al. 2008) establishes one important result about such estimators, namely that efficient estimators can be represented as expectancy values of unbiased (cf. Def. 1.1) estimators $\hat{\theta}$ given (conditional on) a sufficient (cf. Def. 1.5) statistic *t* for the parameter θ, i.e. $E[\hat{\theta} \mid t]$. The reciprocal $1/SE^2_{\hat{\theta}}$ defines the *precision* of an estimator, and for unbiased estimators is bounded from above by the so-called *Fisher information*. As shown by Fisher (1922), the method of maximum likelihood (see sect. 1.3.2 below) will return such efficient estimators (which in this sense contain the most information about the respective population parameter).

Obviously, a 'good' estimator should be unbiased (at least asymptotically so for sufficiently large *N*), should be consistent, should have low standard error, i.e. should be efficient, and should be (minimally) sufficient. We will return to these issues in sect. 2.4 & Ch. 4.

1.3 Principles of statistical parameter estimation

Having defined a statistical model as in the examples of sect. 1.1, how do we determine its parameters? There are three basic principles which have been introduced to estimate parameters of models or distributions from a sample: least-squared error (LSE), maximum likelihood (ML), and Bayesian inference (BI). Each of them will be discussed in turn.

1.3.1 Least-squares error (LSE) estimation

The principle of LSE estimation requires no distributional assumptions about the data and in this sense is the most general and easiest to apply (Winer 1971). However, it may not always give the best answers in terms of the definitions in sect. 1.2 above, in particular if the error terms are non-additive and/ or non-Gaussian. As the name implies, the LSE estimate is defined as the set of parameters that yields the smallest squared model errors, which in case of a linear model with additive error terms are equal to the squared deviations (residuals) of the predicted or estimated values from the observed data (Berger, 1985, discusses error or loss functions from a more general, decision-theoretical perspective). (Note that if the model errors are not additive, but for instance multiplicative, minimizing the squared deviations between predicted and observed values may not be the same as minimizing the error variation.) Say, for instance, our data set consists of univariate pairs $\{x_i, y_i\}$, as in Example 1.3, and we propose the model

$$(1.10)\quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) ,$$

with parameters β_0, β_1. Then the LSE estimates of β_0, β_1 are defined by

$$(1.11)\ \hat{\beta}_0, \hat{\beta}_1 := \underset{\beta_0,\beta_1}{\arg\min}\, Err(\boldsymbol{\beta}) = \underset{\beta_0,\beta_1}{\arg\min} \sum_i \hat{\varepsilon}_i^2 = \underset{\beta_0,\beta_1}{\arg\min} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 ,$$

that is, the estimates that minimize the squared residuals, equal to the squared estimated error terms $\hat{\varepsilon}_i^2$ under model eq. 1.10 (or, equivalently, which maximize the amount of variance in y_i *explained* by the deterministic part $\beta_0 + \beta_1 x_i$). Note that a solution Err($\boldsymbol{\beta}$)=0 typically does not exist as we usually have much more observations than free parameters!

We obtain these estimates by setting

$$(1.12)\ \frac{\partial Err(\boldsymbol{\beta})}{\partial \hat{\beta}_0} = \sum_i -2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \quad and \quad \frac{\partial Err(\boldsymbol{\beta})}{\partial \hat{\beta}_1} = \sum_i -2x_i[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 ,$$

which yields

$$(1.13)\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} ,$$

where $\bar{x}$ and $\bar{y}$ denote the respective sample means. (More generally, if the loss function were not quadratic in the parameters, we would have to check the second derivatives as well.)

1.3.2 Maximum likelihood (ML) estimation

The likelihood function $L_{\mathbf{X}}(\boldsymbol{\theta})$ is defined as the probability (or density) p of a data set **X**={**x**$_i$} given parameters $\boldsymbol{\theta}$, i.e. it tells us how likely it was to obtain the actually observed data set **X** as a function of model parameters $\boldsymbol{\theta}$. Unlike LSE estimation therefore, distributional assumptions with regards to the data are needed. On the positive side, ML estimates have theoretical properties which LSE estimates may lack, e.g. they provide *consistent* (Def. 1.2) and *efficient* (Def. 1.6) estimates (e.g. Myung 2003).

The likelihood factorizes into the product of the likelihoods of the individual observations if these are *independently and identically distributed* (*i.i.d.*):

$$(1.14)\ L_{\mathbf{X}}(\boldsymbol{\theta}) := p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_i p(\mathbf{x}_i \mid \boldsymbol{\theta}) .$$

Thus, the idea of ML inference (largely put forward by Ronald Fisher, 1922, 1934) is to choose parameters such that the likelihood of obtaining the observed data is maximized. In the classical, 'frequentist' view, these parameters are assumed to be (unknown) constants, hence $p(\mathbf{X} \mid \boldsymbol{\theta})$ is, strictly speaking, not a conditional probability (density). This is different from the Bayesian view (sect. 1.3.3) where the parameters are treated as random variables themselves (e.g. Duda & Hart 1973). For mathematical convenience, usually a maximum of the *log*-likelihood $l_{\mathbf{X}}(\boldsymbol{\theta}) := \log L_{\mathbf{X}}(\boldsymbol{\theta})$ is sought (as this converts products as in eq. 1.14 into sums, and, furthermore, may help with exponential distributions as illustrated below).

Example 1: ML estimation of the population mean μ under the univariate normal model

$$(1.15)\ x_i \sim N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} .$$

In this case, the log-likelihood function is given by

$$(1.16)\ \begin{aligned} l_{\mathbf{X}}(\mu) &= \log\left[\prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}\right] \\ &= \sum_i \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}\right) = \sum_i\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - (x_i-\mu)^2/2\sigma^2\right] \end{aligned}$$

Differentiating with respect to μ and setting to 0 gives

$$(1.17)\ \sum_i 2(x_i - \hat{\mu})/2\sigma^2 = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N}\sum_i x_i = \bar{x}.$$

Thus, the ML estimator of μ is the sample mean, and this estimate is unbiased, in contrast to the ML estimator of the standard deviation which underestimates σ^2 by a factor $(N$-$1)/N$ (although with $N\to\infty$, this bias vanishes, and the ML estimator is still consistent!).

ML estimators agree with LSE estimators if the data are independently normally distributed with equal (common) variance, for instance with regards to μ in this case, but this is not true more generally.

Example 2: ML estimation (MLE) of the parameters of the linear regression model (1.10). In this model, usually one assumes predictor variables x_i to be fixed (constant), and hence (assuming i.i.d. data) seeks a maximum of the log-likelihood

$$(1.18)\ l_{\{\mathbf{y}|\mathbf{x}\}}(\boldsymbol{\beta}) = \log\prod_i p(y_i\,|\,x_i,\boldsymbol{\beta}) = \sum_i \log p(y_i\,|\,x_i,\boldsymbol{\beta}) .$$

Since the errors ε were assumed to be Gaussian distributed with mean zero and variance σ^2, according to model (1.10) observations y should themselves follow a Gaussian distribution with mean $\beta_0+\beta_1 x$ (the constant part) and variance σ^2. Thus, the log-likelihood for this model becomes

$$(1.19)\ l_{\{\mathbf{y}|\mathbf{x}\}}(\boldsymbol{\beta}) = \sum_i \log\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\beta_0-\beta_1 x_i)^2/2\sigma^2}\right] = \sum_i\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right].$$

For simplicity we will focus on β_0 here and assume $\sigma^2>0$ known. Differentiating with respect to β_0 and setting to 0 gives

$$(1.20)\ \sum_i \frac{2(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)}{2\sigma^2} = 0 \quad\Rightarrow\quad N\hat{\beta}_0 = \sum_i y_i - \hat{\beta}_1\sum_i x_i \quad\Rightarrow\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} .$$

Hence we see that once again, under the present assumptions, the ML estimate β_0 agrees with the LSE estimate derived in 1.3.1. (Note that more generally one may have to assure that one is dealing with a *maximum* of the log-likelihood function, not a minimum or saddle, which requires the second derivatives to be less than 0.) A very readable introduction into ML estimation with examples from psychological models and the binomial distribution, including Matlab code, is provided in Myung (2003).

1.3.3 Bayesian inference

In MLE, we seek the parameter set $\boldsymbol{\theta}$ which most likely produced the data $\mathbf{X}$ at hand, maximizing $p(\mathbf{X}\,|\,\boldsymbol{\theta})$. Ideally, however, we might want to establish a (posterior) probability distribution directly about the unknown parameters $\boldsymbol{\theta}$, i.e. we would prefer to know $p(\boldsymbol{\theta}\,|\,\mathbf{X})$, rather than – the other way round – $p(\mathbf{X}\,|\,\boldsymbol{\theta})$ as in MLE (Duda & Hart 1973; Berger 1985). The term Bayesian inference comes from the fact that Bayes' rule is used to compute this posterior distribution

$$(1.21)\quad p(\boldsymbol{\theta}\,|\,\mathbf{X})=\frac{p(\mathbf{X}\,|\,\boldsymbol{\theta})p_{\alpha}(\boldsymbol{\theta})}{p(\mathbf{X})}=\frac{p(\mathbf{X}\,|\,\boldsymbol{\theta})p_{\boldsymbol{\alpha}}(\boldsymbol{\theta})}{\sum_{\theta}p(\mathbf{X}\,|\,\boldsymbol{\theta})p_{\boldsymbol{\alpha}}(\boldsymbol{\theta})}$$

of the model parameters $\boldsymbol{\theta}$ given the data, where we have written $p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}):=p(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})$ for short. In case of a density the sums in the denominator have to be replaced by integrals. The *prior distribution* $p_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$, governed by a set of *hyper-parameters* $\boldsymbol{\alpha}$, is the crucial additional ingredient in Bayesian inference, as it enables to incorporate prior knowledge about parameters $\boldsymbol{\theta}$ into our statistical inference (Duda & Hart 1973). Thus, in addition to distributional assumptions about the data (as was the case for ML estimation) *we also have to specify a prior distribution with hyper-parameters* $\boldsymbol{\alpha}$ which may summarize all the information about the parameters we may have in advance. This also allows for iterative updating, since once established knowledge (in form of a probability distribution above $\boldsymbol{\theta}$ with parameters $\boldsymbol{\alpha}$) can serve as a new prior on subsequent runs when new data become available.

For analytical tractability (or simply because it is a natural choice), the prior distribution is often taken to be a so-called *conjugate prior*, which is one which returns the same distributional form for the posterior as was assumed for the prior (e.g., the beta density is a conjugate prior for a Bernoulli process). As a point estimator for $\boldsymbol{\theta}$ one may simply take the largest mode (global maximum) of the posterior distribution, $\hat{\boldsymbol{\theta}}:=\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}\,|\,\mathbf{X})$ (called the maximum-a-posteriori, MAP, estimator, usually easiest to compute), the median, the expectancy value $E(\boldsymbol{\theta}\,|\,\mathbf{X})$, or one may work with the whole posterior distribution. Since we do have the full posterior in this case, we are in the strong position to compute probabilities for parameters $\boldsymbol{\theta}$ to assume values within any possible range of interest (the so-called *credible intervals*, sort of the Bayesian equivalent to the classical statistical concept of a confidence interval; Berger 1985), or to directly compute the probability of the H_0 or H_1 being true given the observed data (which is quite different from just computing the likelihood for a statistic to assume values within a certain range or set under the H_0, as in a typical α-level test). In fact, statistical tests in the Bayesian framework are often performed by just computing the posteriors for the various hypotheses of interest, and accepting the one with the highest posterior probability (Berger 1985; see also Wackerly et al. 2008). One advantage one may see in this is that one gets away from always taking the "devil's advocate" H_0 point-of-view which one tries to refute, and which has led to quite some publication bias. Rather, by directly pitching different hypotheses against each other through their posteriors the H_0 is, so to say, put on 'equal grounds' with all other hypotheses.

If reasonable prior information is available, Bayesian inference may yield much more precise estimates than MLE, since effectively the variation can be considerably reduced by constraining the range of plausible parameter values a priori (Duda & Hart 1973; Berger

1985). The possibility to integrate prior information with observed data in the Bayesian framework may also be of advantage in low-sample-size situations as the lack of data may be partially offset by what is known in advance (e.g., from other studies). However, obviously this can also be dangerous if the prior information is not reliable or incorrectly specified. Moreover, Bayesian estimates are biased in the classical statistical definition (Def. 1.1) toward the information provided by the prior (Wackerly et al. 2008), although this bias will usually vanish as the sample size increases and thus will dominate the prior more and more (i.e., Bayesian estimates may nevertheless be consistent from the "frequentist's" point of view).

On the down side, Bayesian inference is the method mathematically and computationally most involved. First, as noted above, to establish an analytical expression for the posterior distribution, the prior should match up with the likelihood function in a convenient way, e.g. the conjugate prior which leads to the same functional form for the posterior. If it does not, the (nested) integrals in the denominator may become a major obstacle to a full analytical derivation, even if an explicit expression for the likelihood and prior is available, and numerical schemes like Markov Chain Monte Carlo (MCMC) samplers may have to be called upon. In these samplers, at each step a new candidate estimate $\boldsymbol{\theta}^*$ is proposed from a 'proposal distribution' given the previous sample, and accepted or rejected according to how much more (or less) likely it was to obtain this new estimate compared to the previous one given the product of likelihood and prior. This way a chain of estimates $\{\boldsymbol{\theta}^*\}$ is generated which ultimately converges to the true posterior (see Bishop 2006, for more details). For many interesting cases we may not even be able to come up with a closed-form expression for the numerator or the likelihood function. For these cases, numerical sampling schemes like 'particle filters' have been suggested which work with a whole population of samples $\boldsymbol{\theta}^*$ ('particles') simultaneously, which are then moved around in parameter space to approximate the posterior (see sect. 9.3). Each of these samples $\boldsymbol{\theta}^*$ has to overcome the hurdle that it can indeed generate the data at hand with some nonzero probability (all candidate estimates $\boldsymbol{\theta}^*$ have to be consistent with the observed data). See Turner & Van Zandt (2012) for a very readable introduction into this field. In general, there is some debate as to whether the additional computational burden involved in Bayesian inference really pays off in the end (see, e.g., Hastie et al. 2009), at least from a more applied point of view.

1.4 Solving for parameters in analytically intractable situations

In the previous section we have discussed examples for which estimates could be obtained analytically, by explicit algebraic manipulations. However, what do we do in scenarios where (unlike the examples in 1.3.1 and 1.3.2) an analytical solution for our estimation problem is very difficult or impossible to obtain? We will have to resort to numerical techniques for solving the minimization, maximization, or integration problems we encounter in the respective LSE, likelihood, or Bayesian functions, or for at least obtaining a decent approximation. Function optimization is in itself a huge topic (cf. Bishop 2006; Press et al. 2007), and the three next paragraphs are merely to give an idea of some of the most commonly employed approaches.

1.4.1 Gradient descent and Newton-Raphson

One important class of techniques for this situation is called *gradient descent* (or *ascent*, if the goal is maximization). In this case we approximate a solution numerically by moving our estimate $\hat{\boldsymbol{\theta}}$ into a direction opposite to the *gradient* of our criterion (or cost) function, e.g. the negative log-likelihood $-l_{\mathbf{X}}(\boldsymbol{\theta})$, thus attempting to minimize it iteratively (Fig. 1.3). For instance, with *n* denoting the iteration step, the simple forward-Euler scheme reads

(1.22) $\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \gamma \frac{\partial l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}_n}$,

with learning rate $\gamma > 0$. Starting from some initial guess $\hat{\boldsymbol{\theta}}_0$, (1.22) is iterated until the solution converges up to some precision (error tolerance). Note that if γ is too small, it may take very long for the estimate $\hat{\boldsymbol{\theta}}$ to converge, while if it is too large, the process may overshoot and/ or oscillate and miss the solution. This is a problem especially for cost functions with strong local variations in slope, such as in the example in Fig. 1.3. In any case the process will converge only to the nearest *local optimum* which may be significantly worse than the global optimum, and this can be a serious problem if the criterion function is very rough with widely varying slopes and very many optima (Fig. 1.4). A partial remedy can be to start the process from many different initial estimates $\{\hat{\boldsymbol{\theta}}_0\}$, and then select the optimum among the final estimates $\{\hat{\boldsymbol{\theta}}_n\}$.

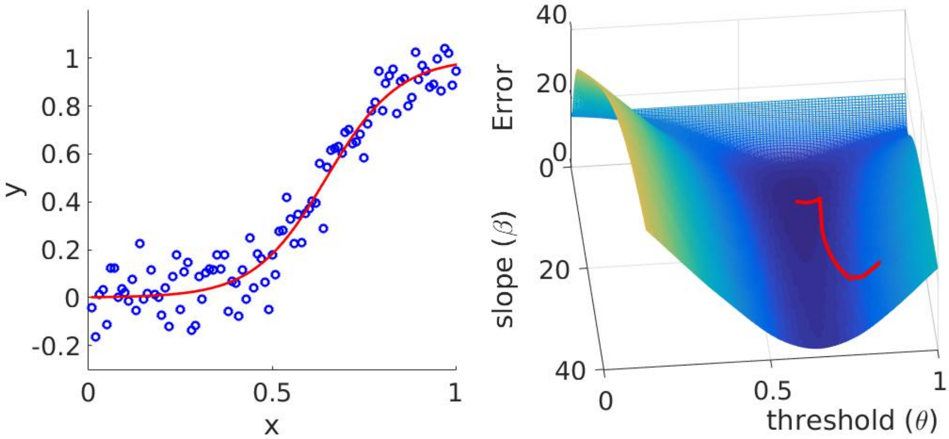

Fig. 1.3. Data (left; blue circles) were drawn from $y = [1 + \exp(\beta(\theta - x))]^{-1} + 0.1\varepsilon$, $\varepsilon \sim N(0,1)$, and parameters $\hat{\beta}$ and $\hat{\theta}$ of the sigmoid (red curve fit) were recovered by gradient descent on the LSE surface illustrated on the right (shown in red is the trajectory of the iterative gradient descent algorithm). In this example, the gradient was weighted with the inverse *absolute* Hessian matrix of second derivatives, similar as in the Newton-Raphson procedure, only that abs(**H**) was taken in (1.23). This was to account for the strong differences in gradient along the β- and θ-directions (note the elongated almost flat valley), while still ensuring that the procedure is strictly *descending* on the error surface. The reader is encouraged to compare this to how the 'standard' gradient descent algorithm (1.22) would perform on this problem for different settings of γ. **MATL1_3**.

A related numerical technique is the Newton-Raphson procedure which is aimed at finding the roots $f(\boldsymbol{\theta}) = 0$ of a function (Press et al. 2007). Since in LSE or ML problems we are interested in minima or maxima, respectively, we would go for the roots of the first derivative $f'(\boldsymbol{\theta}) = 0$. Taking the log-likelihood function $l_{\mathbf{X}}(\boldsymbol{\theta})$ as an example, a Newton-Raphson step in the multivariate case would be defined by

(1.23) $\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n - \mathbf{H}^{-1} \nabla l_{\mathbf{X}}(\boldsymbol{\theta})$

with the vector of partial derivatives $\nabla l_{\mathbf{X}}(\boldsymbol{\theta}) = \left(\frac{\partial l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,1}} \cdots \frac{\partial l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,k}} \right)^T$

and the *Hessian matrix* of second derivatives $\mathbf{H} = \begin{pmatrix} \frac{\partial^2 l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,1}^2} & \cdots & \frac{\partial^2 l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,1} \partial \theta_{n,k}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,k} \partial \theta_{n,1}} & \cdots & \frac{\partial^2 l_{\mathbf{X}}(\boldsymbol{\theta})}{\partial \theta_{n,k}^2} \end{pmatrix}$.

One may think of this scheme as a form of gradient ascent or descent on the *original* function $l_{\mathbf{X}}(\boldsymbol{\theta})$ with the simple learning rate γ replaced by an 'adaptive rate' which a) automatically adjusts the gradient with respect to the size of the local change in slope (the second derivatives), b) adjusts in sign depending on whether a minimum or a maximum is approached. Note that Newton-Raphson works well only if we are dealing with a single maximum or minimum (in fact, only if the function is convex or concave over the interval of interest) – otherwise we may, for instance, end up in a minimum while we were really looking for a maximum! Different such numerical schemes can be derived from Taylor series expansions of $f(\boldsymbol{\theta})$.

1.4.2 Expectation-Maximization (EM) algorithm

The idea of EM (popularized by Dempster et al. 1977) is to solve hard ML problems iteratively by determining the log-likelihood, averaged across a set of auxiliary or unobserved (latent) variables given a current estimate of the parameters $\boldsymbol{\theta}$, in a first step (E-step), and then in the second step (M-step) obtain a new estimate of unknown parameters $\boldsymbol{\theta}$ by maximizing the expected log-likelihood from the preceding E-step (McLachlan & Krishnan 1997). Thus the optimization problem is split into two parts, each of them easier to solve on its own, and either introducing auxiliary (latent) variables which, if they were known, would strongly simplify the problem, or for dealing with models which naturally include unobserved variables to begin with. More formally, the EM algorithm in general form is defined by the following steps, given data **X**, unobserved variables **Z**, and to be estimated parameters $\boldsymbol{\theta}$ (McLachlan & Krishnan 1997; Bishop 2006):

1) Come up with an initial estimate $\hat{\boldsymbol{\theta}}_0$.
2) *Expectation-step*: Compute expectation of joint (or 'complete' in the sense of being completed with the unobserved data) log-likelihood $\log L_{\mathbf{X},\mathbf{Z}}(\hat{\boldsymbol{\theta}})$ across latent or auxiliary variables **Z** given current estimate $\hat{\boldsymbol{\theta}}_k$: $\mathrm{Q}(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}_k) := E_{\mathbf{Z}|\mathbf{X},\hat{\boldsymbol{\theta}}_k}[\log L_{\mathbf{X},\mathbf{Z}}(\hat{\boldsymbol{\theta}})]$.
3) *Maximization-step*: Maximize $\mathrm{Q}(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}_k)$ with respect to $\hat{\boldsymbol{\theta}}$, yielding new estimate $\hat{\boldsymbol{\theta}}_{k+1}$.
4) Check for convergence of $\hat{\boldsymbol{\theta}}_k$ or the log-likelihood. If not converged yet, return to step 2.

In general, if the E- and M-steps are exact (and some other conditions hold, e.g. Wu 1983), the EM algorithm is known to converge (with each EM-cycle increasing the log-likelihood; McLachlan & Krishnan 1997; Bishop 2006), but – like the gradient-based techniques discussed in sect. 1.4.1 – it may find only local maxima (or potentially saddle points; Wu 1983).

We will postpone a specific example and Matlab-implementation to sect. 5.1.1 where parameter estimation in Gaussian Mixture Models, which relies on EM, is discussed. Many further examples of EM estimation will be provided in chapters 7-9.

1.4.3 Optimization in rough territory

Numerical methods like gradient descent work fine if the optimization function is rather smooth and has only one global (*convex* problems) or a few local minima. However, in the most nasty optimization situations, analytical solutions will not be available and the optimization surface may be so rough, fractal, and riddled with numerous local minima that numerical methods like gradient descent will hopelessly break down as well (Fig. 1.4; Wood 2010). In such scenarios, often optimization methods are utilized which contain a strong probabilistic component and may find the global optimum as $t \rightarrow \infty$. Examples are genetic algorithms (Mitchell 1996) which iterate loops of random parameter set variation and deterministic selection until some convergence is reached, simulated annealing (Aarts & Korst 1988) which gradually moves from completely probabilistic to completely deterministic search according to a specified scheme, or numerical samplers (Monte Carlo methods; e.g. Bishop 2006). For instance, Markov Chain Monte Carlo (MCMC) samplers perform a kind of "random walk" through parameter space, accepting or rejecting steps according to their relative likelihood or some other criterion. Other ways to deal with such situations will be provided in sect. 9.3.3.

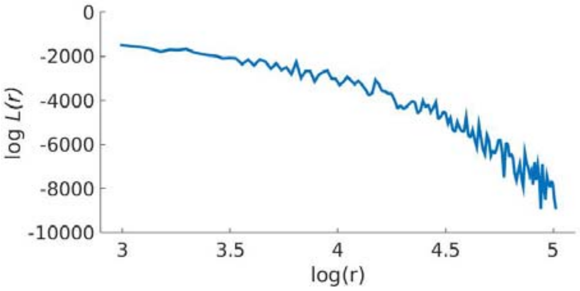

Fig. 1.4. Highly rugged log-likelihood function over variables (y,ε) of the nonlinear time series model $y_t \sim Poisson(0.1x_t)$, $x_t = rx_{t-1}\exp(x_{t-1}+\varepsilon_t)$, $\varepsilon_t \sim N(0,0.3^2)$, where only y_t but not x_t is directly observed (see Wood 2010 for details). MATL1_4.

1.5 Statistical hypothesis testing

Statisticians have produced a wealth of different hypothesis tests for many different types of questions and situations over the past century. In the frequentist framework, the idea of these tests often is to calculate the probability for obtaining the observed or some more extreme value for a specific statistic, or more generally to find the statistic within a specified range or set, given that the null hypothesis is true. The goal here is not to provide a comprehensive introduction into this area for which there are many good textbooks available (e.g. Hays 1994; Freedman et al. 2007; or Kass et al., 2014, for a neuroscience-oriented introduction), but rather to outline the *basic principles and logic of statistical test construction*. This hopefully will provide a) a generally better understanding of the limitations of existing tests, their underlying assumptions, and possible consequences of their violation, and b) some general ideas about how to construct tests in situations for which there are no out-of-the-box methods. There are three basic types of statistical test procedures, that is, ways of deriving the probability of events given a hypothesis: exact tests, asymptotic tests, and bootstrap methods.

1.5.1 Exact tests

An exact test is one for which the underlying probability distribution is exactly known, and the likelihood of an event can therefore, at least in theory, be precisely computed (i.e., does not rely on some approximation or assumptions). This is why these tests are also called *non-parametric* (no parameters need to be estimated) or sometimes 'distribution-free', a terminology I personally find a bit confusing as these tests still entail probability distributions specified by a set of parameters.

Example: Sign-test. Perhaps the simplest and oldest exact statistical test is the so-called sign test which is for *paired* observations $\{x_i \in X, y_i \in Y\}$, for example animals tested before (x_i) and after (y_i) an experimental treatment like a drug application, or investigating gender differences in preferences for food items among couples. More generally, assume we have such paired observations and would like to test the hypothesis that observations from X are larger than those from Y (or vice versa). Let us ignore or simply discard ties for now (cases for which $x_i = y_i$). For each pair we define $T_i = sgn(x_i - y_i)$, and test against the null hypothesis H_0: $E(T) = 0$ or, equivalently, H_0: $p(T=+1) = 0.5$. If we define $k_0 := \frac{1}{2}\sum_i (T_i + 1)$ (which simply counts the number of positive signs), then $k_0 \sim B(k_0, N, 0.5)$, the binomial distribution with $p = 0.5$ under the H_0. Hence we obtain the *exact* probability of observing k_0 or an even more extreme event as

$$(1.24)\quad p(k \geq k_0) = \sum_{i=k_0}^{N} \binom{N}{i} \left(\frac{1}{2}\right)^N .$$

Note that the sign test needs (and uses) only binary information from each pair, i.e. whether one is larger than the other for interval- or ordinal-scaled measurements, so the precise differences do not matter (and hence more detailed assumptions about their distribution are not necessary).

On the side, we further note that the binomial distribution can of course be employed more generally whenever hypotheses about binary categorical data are to be tested. For instance, we may want to know whether there are significantly more vegetarians among townspeople than among people living in the countryside. For this we may fix the probability for the binomial distribution at $pr_c = k_c/N_c$, the proportion of vegetarians among N_c interviewed country people, and ask whether $p(k \geq k_t) \leq \alpha$ according to the binomial with parameters pr_c and N_t, where k_t is the number of vegetarians among the studied sample of N_t townspeople.

Example: Mann-Whitney U-test (Wilcoxon ranksum test). Assume we have *unpaired* (independent) observations $X = \{x_1 ... x_{N_X}\}$ and $Y = \{y_1 ... y_{N_Y}\}$, with $N = N_X + N_Y$ the total number of observations in the two sets, e.g. rats from two different strains tested on some memory score (for instance, number of correctly recalled maze arms). Suppose only the rank order information among all the x_i and y_j is reliable or available, and we hence rank-transform all data $R(z)$: $z \rightarrow \{1..N\}$, that is combine all x_i and y_j into a sorted list to which we assign ranks in ascending order. The null hypothesis assumes that the two sets X and Y were drawn from the same distribution and hence, *on average*, for each $x_i \in X$ there are about as many values in Y preceding as there are succeeding it (Hays 1994), or – more formally – $H_0 : pr(x \geq y) = pr(y \geq x) = 1/2$ for randomly drawn x and y. So for all i we simply count the number of cases for which a rank in group Y exceeds the rank for $x_i \in X$, i.e. $U = \sum_i \#\{y_j \in Y \mid R(y_j) > R(x_i)\}$, and define that as our test statistic (ignoring ties here; but see Hays, 1994). This can be re-expressed in terms of the rank-sums $\overline{R}_X = \sum_i R(x_i)$ and $\overline{R}_Y = \sum_j R(y_j)$, yielding (Hays 1994; Wackerly et al. 2008)

$$(1.25)\quad U = N_X N_Y - \left(\overline{R}_X - \frac{N_X(N_X+1)}{2}\right) = N_X N_Y - U',$$

where the smaller one of U and U' is used; let's call this $U_{obs} = \min(U,U')$. Now, there are a total of $\binom{N}{N_X} = \binom{N}{N_Y}$ possible assignments of ranks to the observations from samples X and Y, and hence from purely combinatorial considerations, for small sample sizes N we may simply count the number of assignments that give a value $\min(U,U')$ (or, equivalently, $\overline{R}_X$ or $\overline{R}_Y$) as observed or a more extreme result, i.e. we compute the exact probability

$$(1.26)\quad p = \#\{\min(U,U') \le U_{obs}\} / \binom{N}{N_X}.$$

Since exact tests return an exact probability, no assumptions or approximations involved, it may seem like they are always the ideal thing to do. Practically speaking, however, they usually have less statistical power than parametric tests since they throw away information if the data are not inherently of rank, count, or categorical type. Moreover, it should be noted that for large N, the distribution of many test statistics from exact tests converge to known parametric distributions like χ^2 or Gaussian. For instance, for $N \to \infty$ the binomial distribution of counts converges to the Gaussian, and sums of squares of properly standardized counts (as used in frequency-table based tests) will converge to the χ^2-distribution as defined in (1.28) below (see Wackerly et al. 2008). These parametric approximations are in fact commonly used in statistical packages for testing, at least when N becomes larger (thus, strictly speaking, moving from exact/ non-parametric to asymptotic/ parametric tests, as they will be discussed next).

1.5.2 Asymptotic tests

Asymptotic tests are usually based on the *central limit theorem* (rooted in work by Laplace, Lyapunov, Lindeberg, and Lévy, among others; Fisz 1970) which states that a sum of random variables converges to the normal distribution for large N, almost (with few exceptions) *regardless* of the form of the distribution of the random variables themselves (Fig. 1.5):

Central limit theorem (CLT). Let X_i, i=1..N, be independent random variables with variance $\sigma^2 > 0$ and finite expectation $\mu := E(X_i) < \infty$. Then

$$(1.27)\quad \lim_{N\to\infty} \frac{\frac{1}{N}\sum_i X_i - \mu}{\sigma / \sqrt{N}} \sim N(0,1).$$

Hence, the *central limit theorem* can be applied when we test hypotheses about means, or for instance when we assume our observations to follow a model like eq. 1.1 or eq. 1.2, where we assume the error terms ε_i to represent sums of many independent error sources which on average cancel out, thus $\varepsilon_i \sim N(0,\sigma^2)$.

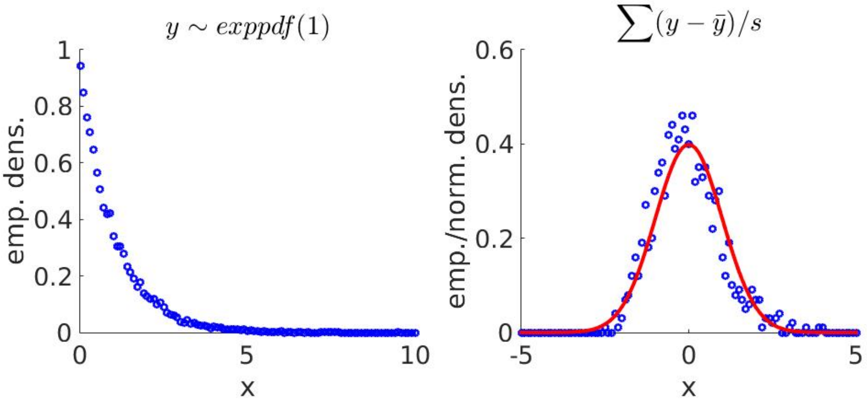

Fig. 1.5. Sums of random variables *y* drawn from a highly-non-Gaussian, exponential distribution (left) will converge to the Gaussian in the limit (right; blue: empirical density of sums of standardized exponential variables; red: Gaussian probability density). MATL1_5.

Sums of squares of normally distributed random variables are also frequently encountered in statistical test procedures based on variances. A further important distribution therefore is defined as (Fisz 1970; Winer 1971):

(1.28) Let $Z_i \sim N(0,1)$, then $\sum_{i=1}^{N} Z_i^2 \sim \chi_N^2$,

the χ^2 distribution with *N* degrees of freedom [*df*, the number of independent random variables (observations) not constrained by the to be estimated parameters; Fig. 1.6, left].

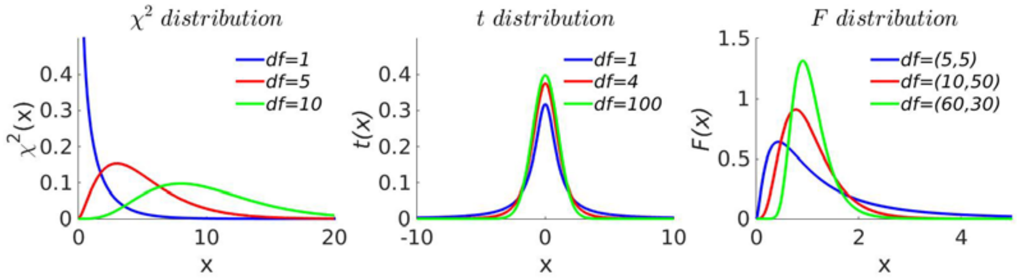

Fig. 1.6. Different parametric distributions. MATL1_6.

Further, the ratio between an independent standard normal and χ^2 distributed random variable with *N* degrees of freedom (divided by *N* and taken to power ½) defines the t-distribution with *N* degrees of freedom (Fig. 1.6, center; Fisz 1970; Winer 1971):

(1.29) $$\frac{z}{\sqrt{\chi_N^2 / N}} \sim t_N\ , \quad z \sim N(0,1)\,.$$

Example: t-test. Student's t-test (due to W.S. Gosset, see Hays 1994) is for the situation where we want to check whether a sample comes from a population with known mean parameter μ, or where we want to test whether two different samples come from populations with equal means (Winer 1971). Returning to Example 1.1, for instance, we may want to test whether the genetic knockout of a certain receptor subunit causes memory deficits compared to the wild-type condition, measured e.g. by the time it takes the animal to find a target or reach criterion. For this *independent* two-sample case we can test the H_0: $\mu_1 = \mu_2$ by the following asymptotically *t*-distributed statistic (e.g. Winer 1971; Hays 1994):

$$(1.30)\ t=\frac{(\bar{x}_1-\mu_1)-(\bar{x}_2-\mu_2)}{\hat{\sigma}_{\bar{x}_1-\bar{x}_2}}=\frac{\bar{x}_1-\bar{x}_2}{\hat{\sigma}_{pool}\sqrt{1/N_1+1/N_2}},\quad \hat{\sigma}_{pool}=\sqrt{\frac{(N_1-1)\hat{\sigma}_1^2+(N_2-1)\hat{\sigma}_2^2}{N_1+N_2-2}},$$

where $\hat{\sigma}_{pool}$ is the *pooled* standard deviation assuming equal variances for the two populations, and the population parameters μ_1, μ_2 in the numerator cancel according to the H_0.

Since the *t*-value (1.30) compares two sample averages $\bar{x}_1$ and $\bar{x}_2$ (which will be normally distributed for large *N* according to the CLT), and a sum of independent Gaussian variables is itself a Gaussian variable, one may assume that one could directly consult the normal distribution for significance levels. This is indeed true for sufficiently large *N*, but for smaller *N* we have to take into account the fact that $\hat{\sigma}_{pool}$ in the denominator is itself a random variable estimated from the sample, and hence the joint distribution of the sample means and $\hat{\sigma}_{pool}$ has to be considered. For i.i.d. Gaussian random variables these two distributions are independent, and the whole expression under these assumptions will in fact come down to a standard normal variable divided by the square root of a χ^2 distributed variable (divided by $\sqrt{df}$), as defined in eq. 1.29. To see this, suppose we had just $\bar{x}-\mu$ in the numerator (as in a one-sample *t*-test), and note that a sample variance is a sum of squares of centered random variables (divided by *N*). Then standardizing the numerator by multiplication with $\sqrt{N}/\sigma$ and doing the same for the denominator will get you this result. Hence *t* is distributed according to the *t*-distribution introduced above with $df = N_1 + N_2$ -2 degrees of freedom in the 2-independent-sample case.

One more note here is that, although *asymptotically* convergence is guaranteed (Hays 1994), at least for smaller samples it may help matters if – in the case of *non*-normally distributed data – we first transform these to bring them in closer relation to the Gaussian. For instance, reaction times or interspike-intervals, although measured with interval-scale precision, are typically *not* normally distributed (apart from the fact that they sharply cut off at zero or some finite positive value, lacking the normal tail). The Box-Cox (1964) class of transformations defined by

$$(1.31)\ \tilde{x}=\begin{cases}(x^q-1)/q & for\ q\neq 0\\ \log(x) & for\ q=0\end{cases}$$

could ease the situation in such cases (Fig. 1.7). Parameter *q* in this transform is usually determined through ML to bring the observed data into closest agreement with the normal assumption.

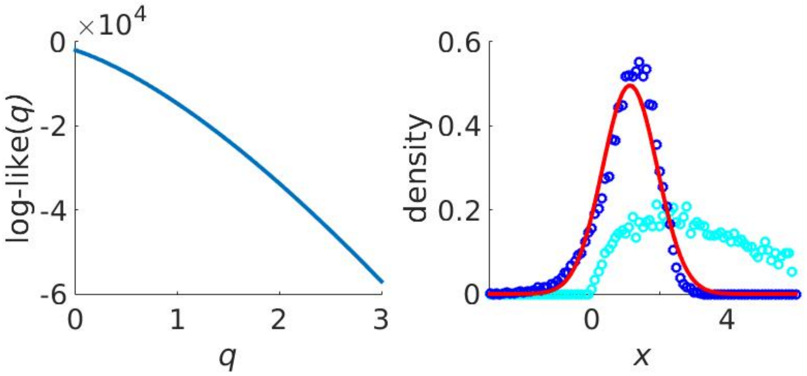

Fig. 1.7. Box-Cox transform for gamma-distributed random variables. Left: Log-likelihood for transform (1.31) as function of exponent q. Right: Original distribution (cyan circles), transformed distribution for the ML estimate q=0 (blue circles), and normal pdf (red curve) for comparison. MATL1_7.

Finally, the ratio between two independently χ^2-distributed quantities divided by their degrees of freedom yields an F-distribution as described in the following example (Fig. 1.6, right).

Example: F-test. The F-test (named so for Ronald Fisher; cf. Hays 1994) compares two sources of variance. Taking one-factor analysis of variance (ANOVA) as an example, we split up the total variation (sum of squares) in a set of grouped samples into a between-group component which captures effects produced by some treatment applied to the groups as in Table 1.1, and a within-group (error) component which represents unexplained error sources (Winer 1971; Hays 1994). For instance, coming back to the experimental setup above, we may have examined not just two but several groups of animals, defined by different genetic modifications which we would like to compare for their impact on memory (cf. example 1.1). The so-called treatment variance, or mean treatment sum of squares (MS_{treat}), captures the differences among the group means (variation of the group means around the grand average $\bar{x}$), while the mean error sum of squares (MS_{err}) adds up the variations of the individual observations within each group from their respective group averages $\bar{x}_k$ as (e.g. Winer 1971, Hays 1994)

$$(1.32)\quad MS_{err}=\frac{\sum_k^P\sum_i^{n_k}(x_{ik}-\bar{x}_k)^2}{n-P},\quad MS_{treat}=\frac{\sum_k^P n_k(\bar{x}_k-\bar{x})^2}{P-1},$$

with n_k the number of observations in the k^{th} group, and $n=\sum_{k=1}^P n_k$ the total number of observations. The F-ratio in this case is the ratio between these two sources of variance. Under the H_0 (no differences among group means), and assuming normally distributed error terms, it follows a ratio of two χ^2 distributions with n-P and P-1 degrees of freedom, respectively (Winer 1971):

$$(1.33)\quad \frac{MS_{treat}}{MS_{err}}\sim\frac{\chi^2_{P-1}/(P-1)}{\chi^2_{n-P}/(n-P)}=:F_{P-1,n-P}\ .$$

(Assuming a common error variance σ_ε^2, the standardization needed to make the terms in the sums standard normal cancels out in the numerator and denominator.) In analyses of variance the decomposition of the total sum of squared deviations from the grand mean is always such that the χ^2-terms in (1.33) are independent by construction.

Likelihood ratio tests. It turns out that many standard statistical tests, like those based on the general linear model (GLM; M/ANOVA), can be obtained from a common principle, the likelihood-ratio test principle (see Wackerly et al. 2008). The likelihood function as defined in sect. 1.3.2 gives the probability or density for obtaining a certain set of observations under a specific model. Given two hypotheses according to which the observed data could have been generated, it thus seems reasonable to favor the hypothesis for which we have a higher likelihood. Let Ω be the set (space) of all possible parameter values which could have generated the data at hand, i.e. $\boldsymbol{\theta} \in \Omega$, then the null hypothesis places certain constraints on this parameter set (e.g. by demanding μ=0), and we may denote this reduced set by $\Omega_0 \subset \Omega$. If the observed data X={x_i} was truly generated by the H_0 parameters, then the maximum likelihood $L_X(\hat{\boldsymbol{\theta}}_{\max} \in \Omega_0)$ should be about as large as $L_X(\hat{\boldsymbol{\theta}}_{\max} \in \Omega)$, since $\boldsymbol{\theta}$ truly comes from Ω_0. Thus, the likelihood ratio test statistic is defined as (see Wilks, 1938, and references to Neyman's and Pearson's earlier work therein)

$$(1.34) \quad \lambda = \frac{L_X(\hat{\boldsymbol{\theta}}_{\max} \in \Omega_0)}{L_X(\hat{\boldsymbol{\theta}}_{\max} \in \Omega)} \in [0,1] \text{ since } \Omega_0 \subset \Omega.$$

The maximum likelihood for the constrained set can never be larger than that for the full set, so $\lambda \to 1$ speaks for the H_0, while $\lambda \to 0$ dismisses it. Conveniently, as shown by Wilks (1938), in the large sample limit

$$(1.35) \quad D := -2\ln(\lambda) \sim \chi^2_{k-k_0},$$

where df=k-k_0 is the number of parameters fixed by the H_0, i.e. the difference in the number of free parameters between the full (k) and the constrained (k_0) model.

This gives us a general principle by which we can construct parametric tests, as long as the parameter sets and thus models specified by the H_0 are *nested* within (i.e., are true subsets of) the parameter set capturing the full space of possibilities. As a specific example, in the linear model eq. (1.4) the null hypothesis may claim that the higher order terms x_i^2 and x_i^3 do not contribute to explaining y, i.e. reduces the space of possible values for $\{\beta_2, \beta_3\}$ to $H_0 : \beta_2 = \beta_3 = 0$. One would proceed about this by obtaining the maximum likelihood (c.f. sect. 1.3.2) under the full model eq. (1.4), and the one for the reduced model in which $\beta_2 = \beta_3 = 0$ has been enforced (i.e., $y_i = \beta_0 + \beta_1 x_i$), and from these compute D as given in (1.35). Since k-k_0=2 in this case, one could check the evidence for the H_0 by looking up D in the χ^2-table for df=2.

1.5.3 Bootstrap (BS) methods

BS (or resampling) methods are a powerful alternative to exact and asymptotic tests if the population distribution of a (perhaps complicated) statistic is unknown, or common assumptions of asymptotic tests are already known to be strongly violated (Efron 1983, 1987; Efron & Tibshirani 1993). While in an exact test the distribution function $\mathrm{F}(\boldsymbol{\theta})$ is known, and in an asymptotic test F is assumed to be of a particular form with parameters estimated from the data, $\mathrm{F}(\hat{\boldsymbol{\theta}})$, in non-parametric BS tests the distributional form itself is commonly estimated, $\hat{\mathrm{F}}(\boldsymbol{\theta})$:

Def. 1.6: Empirical distribution function (EDF). Assume we have observed data $\{\mathbf{x}_1 .. \mathbf{x}_N\}$ from some underlying population distribution *F*, then the EDF is simply defined as the distribution function which puts equal weight $p(\mathbf{X}=\mathbf{x}_i)=1/N$ at each observation, i.e. (Davison & Hinkley 1997)

$$(1.36)\quad \hat{\mathrm{F}}(\mathbf{x}) = \sum_{x_i \le x} p(\mathbf{X}=\mathbf{x}_i) = \frac{\#\{x_i \le x\}}{N}.$$

However, the basic bootstrap exists in both *parametric* and *non-parametric* forms: In the *parametric* case, we indeed assume that the true population distribution comes from a family of functions F(**x**) parameterized by **θ**, where we employ the EDF just for estimating the parameters **θ**. The difference to the fully parametric case lies with the much higher flexibility in choosing the functional form of F (for which powerful basis function series may be employed, cf. sect 5.1.2). We then draw *with replacement* samples of size *N* from $\mathrm{F}_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$.

In the *non-parametric* case, we draw samples of size *N with replacement* directly from $\hat{\mathrm{F}}(\mathbf{x})$ (or some transformation of it supposed to reflect the H_0; Efron & Tibshirani 1993). Thus, having observed a sample $\{\mathbf{x}_1 .. \mathbf{x}_N\}$, we obtain a set of *B* BS samples $\{\mathbf{x}_1^* .. \mathbf{x}_N^*\}$. For instance, for a sample $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ we may have BS replications like $\{\mathbf{x}_1^*=\mathbf{x}_3, \mathbf{x}_2^*=\mathbf{x}_4, \mathbf{x}_3^*=\mathbf{x}_4, \mathbf{x}_4^*=\mathbf{x}_1, \mathbf{x}_5^*=\mathbf{x}_6, \mathbf{x}_6^*=\mathbf{x}_3\}$ or $\{\mathbf{x}_1^*=\mathbf{x}_2, \mathbf{x}_2^*=\mathbf{x}_6, \mathbf{x}_3^*=\mathbf{x}_3, \mathbf{x}_4^*=\mathbf{x}_2, \mathbf{x}_5^*=\mathbf{x}_2, \mathbf{x}_6^*=\mathbf{x}_4\}$. Putting this into concrete numbers, we may have observed the sample X={4,1,6,6,5,5,5,3,4,6} of dice throws. Drawing from this list 10 numbers at random with replacement, we may obtain bootstrap replicates like $X_1^*=\{3,4,6,4,4,4,5,6,6,4\}$ or $X_2^*=\{5,6,3,3,1,5,5,5,3,3\}$. Note that a '2' cannot occur in the BS samples since it was not present in the original sample either. Fig. 1.8 illustrates the convergence of the EDF from normally distributed random numbers to the normal cumulative density. The graph already highlights one potential problem with bootstrapping: If the *N* is too low, the EDF from a sample may severely misrepresent the true distribution, an issue that can only be circumvented with exact tests. In this case we may still be better off with distributional assumptions even if slightly violated.

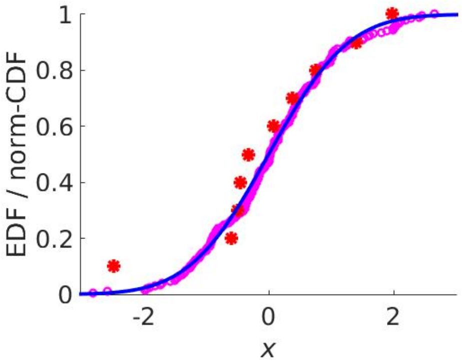

Fig. 1.8. Convergence of empirical distribution function for *n*=10 (red stars) and *n*=200 (magenta circles) to normal CDF (blue). MATL1_8.

Example: Correlation coefficient. Assume we have observed pairs $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ from which we compute the correlation coefficient $r = \mathrm{corr}(x, y)$. We would like to test whether *r* is significantly larger than 0 but perhaps cannot rely on normal distribution assumptions. Strictly, correlation coefficients are not distributed normally anyway, at least for smaller samples, let alone because they are confined in [-1,+1], but Fisher's simple transformation $\tilde{r} = \log(1 + r/1 - r)/2$ would make them approximately normal, *provided* the underlying (x,y)-data are jointly normal (e.g. Hays 1994; recall that in the case of normally distributed errors we could also test for a linear relation between x and y through the regression coefficients in a model like (1.10)). However, perhaps our observations come from some multimodal distribution (or the observations may not be independent from each other, as in time series, a situation which may require more specific bootstrapping methods as discussed later in sect. 7.7).

Let us first use the BS to construct confidence intervals for the respective population parameter ρ, which we estimate by $\hat{\rho} = r$. Assume, for instance, our data are the firing rates of two neurons. These may follow some bimodal, exponential, or gamma distribution, but for the sake of simplicity of this exposition, for the *parametric* scenario let's just assume the data come from a bivariate Gaussian. We may then draw B (e.g. 1000) samples from $N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ estimated from the original sample, and compute BS replications r_b^* of the correlation coefficient for each of the BS data sets. From these we may construct the BS 90-percentile interval which cuts off 5% of the BS values at each tail:

$$(1.37) \quad \rho \in [\hat{\rho} - c_L, \hat{\rho} + c_H]_{0.9}^{*} := \{\rho \mid \hat{F}_*^{-1}(0.05) \leq \rho \leq \hat{F}_*^{-1}(0.95)\},$$

where $\hat{F}_*^{-1}$ denotes the inverse of the BS cumulative distribution function, i.e. $\hat{F}_*^{-1}(0.05)$ is the value r_α^* such that $r_\alpha^* \geq r_b^*$ for 5% of the B BS values (or, in other words, $\hat{F}_*^{-1}(0.05)$ is the (0.05 × B)-th largest value from the B BS samples). (Note that strictly speaking the inverse of $\hat{F}_*$ may not exist, but we could just define $\hat{F}_*^{-1}(p) := \min\{x \mid \hat{F}_*(x) = p\}$.)

Alternatively, in the *non-parametric* case, we draw with replacement B BS replications $\{(x_1^*, y_1^*), \ldots, (x_N^*, y_N^*)\}$ (note that we draw *pairs*, not each of the x_i^*, y_i^*, independently). From these we compute BS replications r_b^*, sort them in ascending order as before, and determine the BS 90-percentile interval. We can also obtain BS estimates of the standard error and bias of the population correlation estimator $\hat{\rho}$ as $\widehat{SE}_{\hat{\rho}} = \text{avg}[(r^* - \bar{r}^*)^2]^{1/2}$ and $\widehat{bias}_{\hat{\rho}} = \text{avg}(r^*) - r$, respectively.

We point out that the confidence limits from BS percentiles as obtained above may be biased and usually underestimate the tails of the true distribution (simply because values from the tails are less likely to occur in an empirical sample). These shortcomings are at least partly alleviated by the BC_a (*bias-corrected* and *accelerated*) procedure (Efron 1987). With this procedure, in determining the BS confidence limits, the α-levels in $\hat{F}_*^{-1}(\alpha)$ are obtained from a standard normal approximation with variables z corrected by a bias term (which could be derived from the deviation of the BS median from the sample estimate) and an acceleration term which corrects the variance and for skewness in using the normal approximation (see Efron 1987, or Efron & Tibshirani, 1993, for details).

If the so-defined confidence limits include the value $r^* = 0$, we may perhaps infer that our empirical estimate $\hat{\rho}$ is not significantly different from 0. However, in doing so we would have made the implicit assumption that the H_0 distribution is just a translated (shifted) version of the H_1 distribution (on which the BS samples were based), centered around $\rho = 0$, and that it is either symmetrical or a mirror-image of the H_1 distribution. If it is not, then this inference may not be valid as illustrated in Fig. 1.9, because the confidence intervals were estimated based on the H_1, not the H_0 distribution! Rather, to directly obtain a bootstrapped

H_0 distribution, instead of drawing BS *pairs* (x_i^*, y_i^*), we may – under the H_0 – in fact compile new pairs (x_i^*, y_j^*) with x_i^* and y_j^* drawn *independently* from each other such that we can have $i \neq j$.

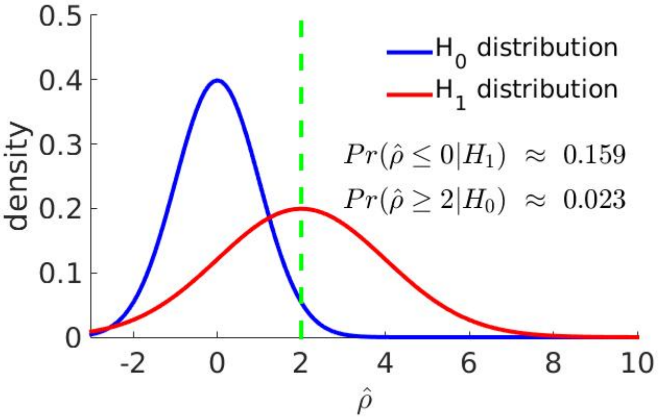

Fig. 1.9. The value $\hat{\rho}=0$ may be well contained within the 90% confidence limits around the mean $\rho=2$ of the H_1 distribution, yet values $\hat{\rho} \geq 2$ would still be highly unlikely to occur under the H_0. MATL1_9.

Alternatively, we may construct a BS H_0 distribution by *rotating* the original (x, y)-coordinate system so as to *de-correlate* the two variables (using, e.g., *principal component analysis*, see sect. 6.1 for details), and then draw randomly pairs with replacement from these transformed coordinates (i.e., from the projections of points (x, y) onto the rotated axes; Fig. 1.10), for which we then compute BS replications r_b^*. Or, if we opt for the parametric BS setting, one may simply set the off-diagonal elements $\sigma_{ij}=0$ for $i \neq j$ in $\hat{\mathbf{\Sigma}}$ for testing the H_0. If $r_{obs} \geq \hat{F}_*^{-1}(0.95)$, i.e. if the empirically observed correlation is larger than 95% of the BS values, we would conclude that the empirically observed r_{obs} significantly deviates from our expectations under the H_0.

There are important differences between the three approaches just outlined (independent drawing, de-correlation, parametric): For independent drawings *any* potential relation between the *x* and *y* would be destroyed – the H_0 assumes that the *x* and *y* are completely independent. In the other two approaches (decorrelation, bivariate Gaussian), in contrast, only the linear relations as captured by the standard Pearson correlation should be destroyed, while higher-order relationships (as those in Fig. 1.2) may still be intact. Hence these two procedures (de-correlating the axes or setting $\sigma_{ij}=0$ for $i \neq j$) test against a more specific H_0. *This example shows that we have to think carefully about which aspects of the data are left intact and which are destroyed in constructing an* H_0 *BS distribution, i.e. what the precise* H_0 *is that we would like to test*!

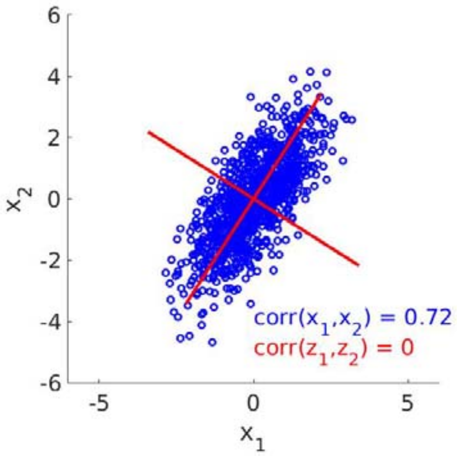

Fig. 1.10. Producing H_0 bootstraps for correlation coefficients by projecting the original data (blue circles) onto a de-correlated axes system (red). **MATL1_10**.

Two-sample bootstraps and permutation tests. Assume we have two samples X={x_1 .. x_{N1}} and Y={y_1 .. y_{N2}} drawn from underlying distributions F and G, respectively, and we would like to test whether these two population distributions differ in one or more aspects. We could think of this again in terms of the wild-type/knockout-comparison used as an example above in sect. 1.5.2 in connection with the *t*-test. In fact, as a test statistic we could still use Student's two-sample *t* as defined in that section, only that this time we will use bootstraps to check for significant differences (here the term 'bootstrap' will be used in a bit of a wider sense than sometimes in the literature, for any situation where we use the original observations to construct an EDF, rather than relying on exact or asymptotic distributions). One way to approach this question from the BS perspective is to combine all observations from both samples into a common set {x_1, …, x_{N1}, y_1, …, y_{N2}} from which we randomly draw N_1 and N_2 values to form new sets X* and Y*, respectively (Efron & Tibshirani 1993). We do this *B* times, and for each two BS sets X$_b^*$ and Y$_b^*$ we compute $t_b^* = |\bar{x}_b^* - \bar{y}_b^*| / (\hat{\sigma}_{pool}^* \sqrt{1/N_1 + 1/N_2})$. With smaller samples one may actually try out all possible assignments [*permutations*] of the N_1+N_2 observations or class labels, an idea going back to Ronald Fisher and Edwin J.G. Pitman (see Ernst 2004), also a kind of exact test. Finally, we check whether the value t_{obs} obtained from the original sample ranks in the top 5-percentile of the BS distribution. Note that this procedure tests the strong H_0: F = G (Efron & Tibshirani 1993), since in constructing the BS data *we ignore the original assignments to distributions F and G completely*.

Alternatively, for just testing H_0: $\mu_1 = \mu_2$ (equality of the population means), we could first subtract off the sample means $\bar{x}$ and $\bar{y}$, respectively, from the two original samples, add on the common mean $(N_1\bar{x} + N_2\bar{y})/(N_1 + N_2)$, and draw BS replications {X*, Y*} with each $\mathrm{X}^* \subseteq \mathrm{X}$ and $\mathrm{Y}^* \subseteq \mathrm{Y}$, i.e. draw separately with replacement from $\hat{F}$ and $\hat{G}$, not from $\mathrm{X} \cup \mathrm{Y}$ as above (Efron & Tibshirani 1993). Again we compute t_b^* for each BS replication, and take our significance level to be $\#\{t_b^* \geq t_{obs}\}/B$. Obviously these BS strategies could easily be extended to more than two samples.

Both these BS methods, along with an example of an exact test (Wilcoxon rank-sum) and the (asymptotic) *t*-test, are illustrated and compared in **MATL1_11** using a hypothetical data set. Two final remarks on bootstrapping: First, as already pointed out above (cf. Fig. 1.8), a *larger* sample size is often required for nonparametric bootstrapping than for parametric tests, since in the latter case the distributional form itself is assumed to be already known (and thus has not to be derived from the data). Second, a high-quality random number generator is needed for implementing bootstrapping methods to avoid biased results.

1.5.4 Multiple testing problem

In high-dimensional settings we may find ourselves pretty quickly in situations where we would like to test multiple hypotheses simultaneously, as typical for instance in gene-wide association studies where some 1000 gene variants are to be linked to different phenotypes. When testing 100 (independent) null hypotheses at α=0.05, then just by chance on average 5 of them will get a star (significant) although in reality the H_0 is true. The family-wise error rate (FWER) is defined as the probability of obtaining at least one significant result just by chance (Hastie et al. 2009), and for fixed α and K independent hypothesis tests is given by

$$(1.38)\quad \Pr(\#\{'accept\ H_1'|'H_0\ true'\} \geq 1) = 1-(1-\alpha)^K .$$

In general, if we had obtained k significances out of K tests at level α, we may take the cumulative binomial distribution $\sum_{r=k}^{K} B(r,K,\alpha)$ to check whether we could have achieved this or an even more extreme result just by chance.

We could also attempt to explicitly control the FWER, for which the Bonferroni correction $\alpha^*=\alpha/K$ is probably the most famous remedy. A less conservative choice is the Holm-Bonferroni procedure (Holm 1979) which arranges all probability outcomes in increasing order, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$, and rejects all $H_0^{(r)}$ for which

$$(1.39)\quad r < k^* ,\ k^* := \min\left\{ k \middle| p_{(k)} > \frac{\alpha}{K-(k-1)} \right\} .$$

Instead of controlling the FWER, one may want to specify the false discovery rate (FDR), which is the *expected relative number* of H_0 among the set of all H_0 *rejected* that were *falsely* called significant (Hastie et al. 2009). The FDR could be set by the Benjamini & Hochberg (1995) procedure, which similar to the Holm-Bonferroni method first arranges all $p_{(k)}$ in increasing order, and then rejects all $H_0^{(r)}$ for which

$$(1.40)\quad r < k^* ,\ k^* := \min\left\{ k \middle| p_{(k)} > \frac{k\alpha}{K} \right\} .$$

Note that both (1.39) and (1.40) yield the Bonferroni-corrected α level for k=1, and the nominal α level for k=K, but in between (1.39) (rising hyperbolically with k) is the more conservative choice for any given nominal significance level α.