Let R_t denote the event "the particle has a residence time longer than t seconds." In Section 5.3 we will see how continuous stirring determines the probabilities; here we just use that in a particular continuously stirred tank, R_t has probability e^{-t}. So:

$$\mathrm{P}(R_3) = \mathrm{e}^{-3} = 0.04978\ldots$$
$$\mathrm{P}(R_4) = \mathrm{e}^{-4} = 0.01831\ldots.$$

We can use the definition of conditional probability to find the probability that a particle that has stayed more than 3 seconds will stay more than 4:

$$\mathrm{P}(R_4 \mid R_3) = \frac{\mathrm{P}(R_4 \cap R_3)}{\mathrm{P}(R_3)} = \frac{\mathrm{P}(R_4)}{\mathrm{P}(R_3)} = \frac{\mathrm{e}^{-4}}{\mathrm{e}^{-3}} = \mathrm{e}^{-1} = 0.36787\ldots.$$

Quick exercise 3.4 Calculate $\mathrm{P}(R_3 \mid R_4^c)$.

For more details on the subject of residence time distributions see, for example, the book on reaction engineering by Fogler ([11]).

3.2 The multiplication rule

From the definition of conditional probability we derive a useful rule by multiplying left and right by $\mathrm{P}(C)$.

The multiplication rule. For any events A and C:

$$\mathrm{P}(A \cap C) = \mathrm{P}(A \mid C) \cdot \mathrm{P}(C)\,.$$

Computing the probability of $A \cap C$ can hence be decomposed into two parts, computing $\mathrm{P}(C)$ and $\mathrm{P}(A \mid C)$ separately, which is often easier than computing $\mathrm{P}(A \cap C)$ directly.

The probability of no coincident birthdays

Suppose you meet two arbitrarily chosen people. What is the probability their birthdays are different? Let B_2 denote the event that this happens. Whatever the birthday of the first person is, there is only one day the second person cannot "pick" as birthday, so:

$$\mathrm{P}(B_2) = 1 - \frac{1}{365}.$$

When the same question is asked with *three* people, conditional probabilities become helpful. The event B_3 can be seen as the intersection of the event B_2,

This point of view also leads the way to how one should define the expected value of a continuous random variable. Let, for example, X be a continuous random variable whose probability density function f is zero outside the interval $[0, 1]$. It seems reasonable to approximate X by the *discrete* random variable Y, taking the values

$$\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$$

with as probabilities the masses that X assigns to the intervals $[\frac{k-1}{n}, \frac{k}{n}]$:

$$\mathrm{P}\left(Y = \frac{k}{n}\right) = \mathrm{P}\left(\frac{k-1}{n} \leq X \leq \frac{k}{n}\right) = \int_{(k-1)/n}^{k/n} f(x)\,\mathrm{d}x.$$

We have a good idea of the size of this probability. For large n, it can be approximated well in terms of f:

$$\mathrm{P}\left(Y = \frac{k}{n}\right) = \int_{k/n-1/n}^{k/n} f(x)\,\mathrm{d}x \approx \frac{1}{n} f\left(\frac{k}{n}\right).$$

The "center-of-gravity" interpretation suggests that the expectation $\mathrm{E}[Y]$ of Y should approximate the expectation $\mathrm{E}[X]$ of X. We have

$$\mathrm{E}[Y] = \sum_{k=1}^{n} \frac{k}{n} \mathrm{P}\left(Y = \frac{k}{n}\right) \approx \sum_{k=1}^{n} \frac{k}{n} f\left(\frac{k}{n}\right) \frac{1}{n}.$$

By the definition of a definite integral, for large n the right-hand side is close to

$$\int_0^1 x f(x)\,\mathrm{d}x.$$

This motivates the following definition.

DEFINITION. The *expectation* of a continuous random variable X with probability density function f is the number

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

We also call $\mathrm{E}[X]$ the *expected value* or *mean* of X. Note that $\mathrm{E}[X]$ is indeed the center of gravity of the mass distribution described by the function f:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \frac{\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x}{\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x}.$$

This is illustrated in Figure 7.2.

7.5 Solutions to the quick exercises

7.1 We have

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 8 \cdot \frac{1}{5} + 16 \cdot \frac{1}{5} = \frac{31}{5} = 6.2.$$

7.2 The probability density function f of U is given by $f(x) = 0$ outside $[2, 5]$ and $f(x) = 1/3$ for $2 \leq x \leq 5$; hence

$$\mathrm{E}[U] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_2^5 \frac{1}{3} x\,\mathrm{d}x = \left[\frac{1}{6}x^2\right]_2^5 = 3\tfrac{1}{2}.$$

7.3 Using the change-of-variable formula we obtain

$$\begin{aligned}\mathrm{E}\left[2^X\right] &= \sum_i 2^{a_i} \mathrm{P}(X = a_i) \\ &= 2^0 \cdot \mathrm{P}(X = 0) + 2^1 \cdot \mathrm{P}(X = 1) \\ &= 1 \cdot (1 - p) + 2 \cdot p = 1 - p + 2p = 1 + p.\end{aligned}$$

You could also have noted that $Y = 2^X$ has a distribution given by $\mathrm{P}(Y = 1) = 1 - p, \mathrm{P}(Y = 2) = p$; hence

$$\mathrm{E}\left[2^X\right] = \mathrm{E}[Y] = 1 \cdot \mathrm{P}(Y = 1) + 2 \cdot \mathrm{P}(Y = 2) = 1 \cdot (1 - p) + 2 \cdot p = 1 + p.$$

7.4 We have

$$\mathrm{Var}(Y_1) = \tfrac{1}{2}(450 - 500)^2 + \tfrac{1}{2}(550 - 500)^2 = 50^2 = 2500,$$

so Y_1 has standard deviation €50 and

$$\mathrm{Var}(Y_2) = \tfrac{1}{2}(0 - 500)^2 + \tfrac{1}{2}(1000 - 500)^2 = 500^2 = 250\,000,$$

so Y_2 has standard deviation €500.

7.6 Exercises

7.1 ⊡ Let T be the outcome of a roll with a fair die.

a. Describe the probability distribution of T, that is, list the outcomes and the corresponding probabilities.

b. Determine $\mathrm{E}[T]$ and $\mathrm{Var}(T)$.

7.2 ⊡ The probability distribution of a discrete random variable X is given by

$$\mathrm{P}(X = -1) = \tfrac{1}{5}, \quad \mathrm{P}(X = 0) = \tfrac{2}{5}, \quad \mathrm{P}(X = 1) = \tfrac{2}{5}.$$

Table 1.1 Joint probability distribution of (H, W)

	$w=1$	$w=2$	$w=3$	$w=4$
$h=0$	0.000	0.000	0.000	0.125
$h=1$	0.125	0.125	0.125	0.000
$h=2$	0.250	0.125	0.000	0.000
$h=3$	0.125	0.000	0.000	0.000

$$\mathbb{P}(H=0, W=n+1) = \mathbb{P}(H=0|W=n+1)\mathbb{P}(W=n+1) = 1\times(1-p)^n,$$

the probability to see n tails. Further, $\mathbb{P}(H=0, W=w)=0$ for any $w \leq n$, as we cannot have seen the first head somewhere in the sequence if we saw none at all.

What about the non-boundary cases? For $(H=h, W=w)$, we can use the following argument: to see h heads in total, given the first one in the wth flip, we know that the first $w-1$ flips are all tails and we need to place $h-1$ heads in the remaining $n-w$ positions (the first is already placed in position w):

$$\mathbb{P}(H=h|W=w) = \binom{n-w}{h-1}(1-p)^{(n-w)-(h-1)}p^{h-1},$$

the binomial probability of having $h-1$ heads in $n-w$ trials. The probability of first head at $w \leq n$ is the geometric distribution $\mathbb{P}(W=w) = (1-p)^{w-1}p$. Combining:

$$\mathbb{P}(H=h, W=w) = \binom{n-w}{h-1}(1-p)^{n-h}p^h.$$

The conditional distribution of waiting w flips, given we have h heads in total, is easily calculated as

$$\mathbb{P}(W=w|H=h) = \frac{\binom{n-w}{h-1}}{\binom{n}{h}},$$

the number of ways to place $h-1$ heads in $n-w$ positions over the number of ways to place h heads in n positions. Interestingly, this probability is independent of the probability p to see head. For $n=3$ and $p=0.5$, the full joint probability $\mathbb{P}(H=h, W=w)$ is given in Table 1.1.

We might also be interested in computing the waiting time distribution without referring to the number of heads. This *marginal distribution* can be derived by applying the law of total probability. For example,

$$\mathbb{P}(W=2) = \sum_{h=0}^{3}\mathbb{P}(H=h, W=2) = (1-p)p = \frac{1}{4}$$

is the marginal probability that we see the first head in the second flip.

Example **4** To contribute another example, let us consider the following problem, encountered in molecular biology: DNA molecules carrying the inheritance information of an organism can be modeled as a sequence of nucleotides. There are four different such nucleotides: arginine (abbreviated `A`), cytosine (`C`), guanine (`G`), and tyrosine (`T`). A common problem is to determine how closely related two DNA sequences are. To make things easier, let us assume both sequences have the same length n, that the nucleotides in any two positions in the sequence are independent, and that each nucleotide has a probability of 1/4 to occur in any position. Similarity of the sequences can then be established by counting in how many positions the two sequences have the same nucleotide. Each such case is called a *match*, the converse a *mismatch*, so the following two sequences have seven matches and three mismatches (underlined):

$$\begin{array}{cccccccccc} A & C & \underline{C} & G & T & T & \underline{G} & G & \underline{T} & A \\ A & C & \underline{G} & G & T & T & \underline{C} & G & \underline{A} & A \end{array}$$

If the two sequences have nothing in common, we would expect to see a match in about 1/4 of the cases, and the number of matches would follow a $\mathrm{Binom}(n,\, p = 1/4)$ distribution. Conversely, evolutionarily related DNA sequences would show a much higher proportion of matches.

In subsequent chapters, we will *estimate* the nucleotide frequencies p from data and *test the hypothesis* that sequences are related by comparing the observed and expected number of matches

Continuous Random Variables. We also need random variables that take values in a continuous set to describe, e.g., measured lengths or optical densities. Similar to events, we cannot cover these in all mathematical rigor. A nontrivial mathematical argument shows that for such a continuous random variables X, a probability mass function cannot be defined properly, because $\mathbb{P}(X = x) = 0$ for all x. Instead, most of these variables have a *probability density function* (*pdf*) $f_X(x)$ with the properties

$$\begin{aligned} f_X(x) &\geq 0\,, \\ \int_{-\infty}^{\infty} f_X(y)\mathrm{d}y &= 1. \end{aligned}$$

The density is a function such that the probability of X to take a value in any interval $[x_l, x_u]$ is given by the area under the density function on this interval, that is, $\mathbb{P}(x_l \leq X \leq x_u) = \int_{x_l}^{x_u} f_X(y)\mathrm{d}y$. This probability can also be written in terms of the cumulative distribution function

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y$$

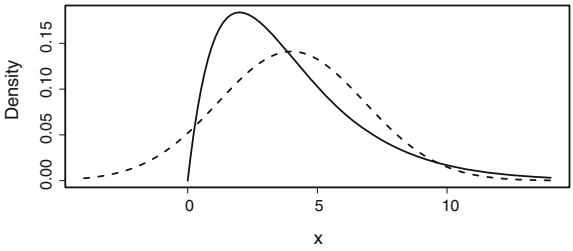

Fig. 1.8 Gamma (*solid*) and normal (*dashed*) densities both with mean $\mu = 4$ and variance $\sigma^2 = 8$, corresponding to Gamma parameters $k = 2$ and $\theta = 2$

Example **7** For introducing yet another continuous distribution on the go, let us consider the *Gamma distribution* with shape parameter k and scale parameter θ. It has density function $f(x; k, \theta) = x^{k-1} \exp(-x/\theta)/\theta^k \Gamma(k)$, is only defined for $x > 0$, and describes the distribution of the sum of k exponentially distributed waiting times, each with rate parameter θ (thus the time to wait for the kth event). This density function is usually not symmetric. For $k = 2$ and $\theta = 2$, the distribution has expectation $\mu = k\theta = 4$ and variance $\sigma^2 = k\theta^2 = 8$; its density is shown in Fig. 1.8 (solid line). For comparison, a normal distribution with the same expectation and variance is plotted by a dashed line. As we can see, the density functions look very different, although both have the same mean and variance. For additionally capturing their different shapes, higher moments are needed (see Sect. 1.6.5).

Example **8** Let us consider the following model of a random DNA sequence as introduced earlier: we assume independence among the nucleotides and in each position, the probabilities of having a particular nucleotide are p_A, p_C, p_G, p_T, respectively. We investigate two sequences of length n by comparing the nucleotides in the same position. Assume that the sequences are completely random and unrelated. At any position, the probability of a match is then $p := p_A^2 + p_C^2 + p_G^2 + p_T^2$, as both nucleotides have to be the same. Let us set $M_i = 1$ if the sequences match in position i and $M_i = 0$ else.

To decide whether two given sequences are related, we compute the number of matching nucleotides and compare it to the number of matches we expect just by chance. If the observed number is much higher than the expected number, we claim that the sequences are in fact related.[1]

The total number of matches in two random sequences of length n is given by $M := M_1 + \cdots + M_n$ and follows a binomial distribution: $M \sim \text{Binom}(n, p)$. Applying the linearity of the expectation and some algebra, we compute the expected number of matches:

[1] As a word of caution for the biological audience: this argument does not hold for *aligned* sequences, as the alignment maximizes the number of matches, and this maximum has a different distribution.

$$\begin{aligned}
\mathbb{E}(M) &= \sum_{k=-\infty}^{\infty} k\mathbb{P}(M=k) \\
&= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} np \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1}(1-p)^{(n-1)-(k-1)} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1}(1-p)^{(n-1)-(k-1)} \\
&= np,
\end{aligned}$$

where the last equality holds because we have the pmf of a Binom $(n-1, p)$ variable in the sum, which sums to one. The result also makes intuitive sense: the expected number of matches is the proportion p of matches times the number of nucleotides n. Consequently, for sequences of length $n = 100$, with nucleotide probabilities all equal to 0.25, the probability of a match is $p = 0.25$ and we expect to see 25 matches just by chance if the sequences are unrelated.

How surprised are we if we observe 29 matches? Would this give us reason to conclude that the sequences might in fact be related? To answer these questions, we would need to know how likely it is to see a deviation of 4 from the expected value. This information is captured by the variance, which we can calculate as

$$\mathrm{Var}(M) = \mathrm{Var}(M_1) + \cdots + \mathrm{Var}(M_n),$$

because we assumed that the nucleotides are independent among positions. Using the definition of the variance,

$$\mathrm{Var}(M_1) = \mathbb{E}((M_1)^2) - (\mathbb{E}(M_1))^2 = (0^2 \times (1-p) + 1^2 \times p) - p^2 = p(1-p),$$

and we immediately get

$$\mathrm{Var}(M) = n\mathrm{Var}(M_1) = np(1-p) = 18.75$$

and a standard deviation of 4.33 These values indicate that the deviation of the observed number of matches (=29) from the expected number of matches (=25) is within the range that we would expect to see in unrelated sequences, giving no evidence of the sequences being related. We will see in Chap. 3 how these arguments can be used for a more rigorous analysis.

1.6.3 Z-Scores

Using the expectation and variance of any random variable X, we can also compute a normalized version Z with expectation zero and variance one by

$$\mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \in \left[\frac{-l}{\mathrm{se}(\hat{\theta}_n)}, \frac{u}{\mathrm{se}(\hat{\theta}_n)}\right]\right) = 1 - \alpha. \tag{2.2}$$

Solving (2.2) requires that we find the two quantiles $q_{\alpha/2}$, $q_{1-\alpha/2}$ of the distribution of the normalized estimator, with $1-\alpha/2-\alpha/2 = 1-\alpha$. From these quantiles, we work out the upper value $u = q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)$ and thus the interval bound $\hat{\theta}_n + q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)$, and similar for the lower value l. For an unbiased estimator, the $(1-\alpha)$-confidence interval therefore takes the general form

$$C = \left[\hat{\theta}_n + q_{\alpha/2}\mathrm{se}(\hat{\theta}_n), \hat{\theta}_n + q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)\right],$$

which simplifies by $q_{\alpha/2} = -q_{1-\alpha/2}$ if the estimator additionally has a symmetric distribution around its mean. The two main remaining problems are then to establish the distribution of $\hat{\theta}_n$ to calculate the quantiles and to estimate its variance.

If $\hat{\theta}_n$ is an unbiased maximum-likelihood estimator, we already know that the estimator has a normal distribution and the correct quantiles are $z_{\alpha/2}$ and $z_{1-\alpha/2}$. The shifted and scaled interval is then symmetric around zero and the confidence interval is immediately given by

$$C = \left[\hat{\theta}_n - z_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n),\ \hat{\theta}_n + z_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)\right].$$

Before we deal with the more general case of estimators that are not ML, let us first look into two concrete examples and work out confidence intervals for the sequence matching problem and the estimates for the normal parameters.

Example 14 We would like to compute an interval $[\hat{p}_n - l, \hat{p}_n + u]$ around the estimate $\hat{p}_n$ of the matching probability, such that the interval contains the true value p with given probability $1 - \alpha$:

$$\mathbb{P}(p \in [\hat{p}_n - l, \hat{p}_n + u]) = \mathbb{P}(\hat{p}_n - l \leq p \leq \hat{p}_n + u) = 1 - \alpha.$$

Because $\hat{p}_n$ is the maximum-likelihood estimator of p, its distribution approaches a normal distribution for large n. Its normalized form has a standard normal distribution:

$$\frac{p - \hat{p}_n}{\mathrm{se}(\hat{p}_n)} \sim \mathrm{Norm}\ (0, 1).$$

We can therefore immediately solve the following equation by using the corresponding quantiles z_α for u and l

$$\mathbb{P}\left(\frac{-l}{\mathrm{se}(\hat{p}_n)} \leq \frac{p - \hat{p}_n}{\mathrm{se}(\hat{p}_n)} \leq \frac{u}{\mathrm{se}(\hat{p}_n)}\right) = 1 - \alpha.$$

By exploiting the symmetry of the normal distribution, we derive

$$\frac{u}{\mathrm{se}(\hat{p}_n)} = z_{1-\alpha/2} \iff u = z_{1-\alpha/2}\mathrm{se}(\hat{p}_n) \text{ and } l = z_{\alpha/2}\mathrm{se}(\hat{p}_n),$$

The standard error of $\hat{p}_n$ is $\mathrm{se}(\hat{p}_n) = \sqrt{p(1-p)/n}$, leading to the requested confidence interval

$$C = \left[\hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right],$$

where we replaced the unknown true parameter value p by its estimate $\hat{p}_n$.

An immediate caveat of the approximation of the true distribution of the estimator $\hat{p}_n$ by its asymptotic normal distribution is that this confidence interval is only valid for large sample sizes n and parameter values not too close to zero or one. For small p, for example, the confidence interval would also consider the case that $\hat{p}_n$ takes on a negative value, which is not possible. Hence, the approximations for this confidence interval are not always valid and more sophisticated intervals exist.

Example 15 Let us consider the estimator for the expectation of normally distributed data, i.e., $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ with $X_i \sim \text{ Norm } (\mu, \sigma^2)$. Being the ML-estimator, this random variable has a normal distribution. We already checked that it is unbiased, and we easily compute its variance $\mathrm{Var}(\bar{X})$ as

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$$

where we could take the sum outside the variance because we assumed the X_i to be independent. Thus, the normalized distribution of the difference in true and estimated mean is

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \text{ Norm } (0, 1).$$

Again, we do not know the true variance and need to estimate is using the unbiased estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, which leads to the normalized random variable

$$\frac{\bar{X}-\mu}{S/\sqrt{n}},$$

which does *not* have a standard normal distribution. We can derive its correct distribution by looking at the estimated variance in more detail. In particular, let us consider the quotient of the true and estimated variance:

$$(n-1)\frac{S^2}{\sigma^2} = (n-1)\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2.$$

Each summand is the square of a standard normal variable and there are $(n-1)$ independent such variables. Thus, from Sect. 1.4, we know that the sum has a χ^2-distribution with $(n-1)$ degrees of freedom. Replacing the true variance by its estimate, we derive the distribution

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma/\sqrt{n}}{S/\sqrt{n}} \sim \frac{\text{Norm}(0,1)}{\sqrt{\frac{1}{n-1}\chi^2(n-1)}},$$

which from Sect. 1.4 we recognize as a t-distribution with $(n-1)$ degrees of freedom. We therefore derive the correct $(1-\alpha)$-confidence interval

$$C = \left[\bar{X} - t_{1-\alpha/2}(n-1\frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

for the estimator $\bar{X}$ of the expected value. Again, this interval gets narrower if we increase the sample size n or decrease the variance σ^2 of the data.

As an example, let us repeatedly take 10 samples from a Norm(5,16) distribution and compute the corresponding 0.9-confidence interval for the estimated mean $\bar{X}$. For each such computation, we derive a slightly different interval, both in terms of the center of the interval (due to the estimated mean $\bar{X}$) and the length of the interval (due to the estimated variance of $\bar{X}$). For 25 repetitions, the confidence intervals are plotted next to each other in Fig. 2.2. Some intervals, such as the 5th and the 24th, do not cover the true value. To demonstrate the effect of estimating the variance, we compute the correct t-based and the incorrect normal confidence intervals, both using the estimated variance, for the 5th sample (which is too far away from the true mean) as

$$C^t = [1.396, 4.717] \quad \text{and} \quad C^{\text{norm}} = [0.631, 5.482].$$

The normal quantiles overestimate the width of the interval, such that the normal interval contains the true value, while the t-based does not.

2.3.1 The Bootstrap

For computing the confidence interval for a given estimate, we frequently encounter two problems: finding the variance of an estimator, and working out the distribution of an estimator that is not an MLE. In addition, the theory leading to normal (or t-based) confidence intervals is based on the asymptotic distribution of the estimator, which might be quite different than the distribution for small sample sizes. A very popular way for solving these problems is by using the *bootstrap* method, which aims at estimating all necessary quantities directly from the data themselves. While mainly used for computing the estimator's variance, the bootstrap method

experiment, and therefore see $x_i + a$ observations in category i after the experiment. Then, we can use the estimate

$$\tilde{p}_{n,i} = \frac{x_i + a}{sa + n}$$

for the categories' probabilities, which is simply the MLE for the modified data. With $a > 0$, each estimate is strictly larger than (but potentially very close to) zero.

Let us assume we observed $(x_A, x_C, x_G, x_T) = (13, 6, 0, 1)$. How large should we choose a? If we choose it too large, it would spoil the whole estimation and assign almost identical probabilities everywhere, independent of the data. For example, with $a = 1000$ the estimates are

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.252, 0.25, 0.249, 0.249).$$

If we choose a too small, it might not have an effect and we end up with non-zero, but extremely low probabilities. Indeed, for $a = 0.1$,

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.642, 0.299, 0.005, 0.054).$$

We can calculate a reasonable compromise by selecting a such that we minimize the maximal risk of the corresponding estimator. For parameters of the multinomial distribution, this minimax estimator is achieved by choosing

$$a = \frac{\sqrt{n}}{s}.$$

For the example, $a = 1.118$ and we estimate

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.577, 0.291, 0.046, 0.087),$$

which is fairly close to the correct values, taking into account that we do not have many data available.

The seemingly ad-hoc estimator in Sect. 2.5.2 for the binomial case was derived in this way.

2.6 Fisher-Information and Cramér-Rao Bound

We conclude the chapter by a brief discussion of the idea of Fisher-information, from which we can derive a theoretical lower bound for the variance of an estimator. This bound tells us how precise we can actually estimate a given parameter with a fixed number of samples.

Recall the definition $\ell_n(\theta) = \sum_i \log(f(x_i; \theta))$ of the log-likelihood function. The *Fisher-score* is simply the derivative of this function with respect to the parameter(s),

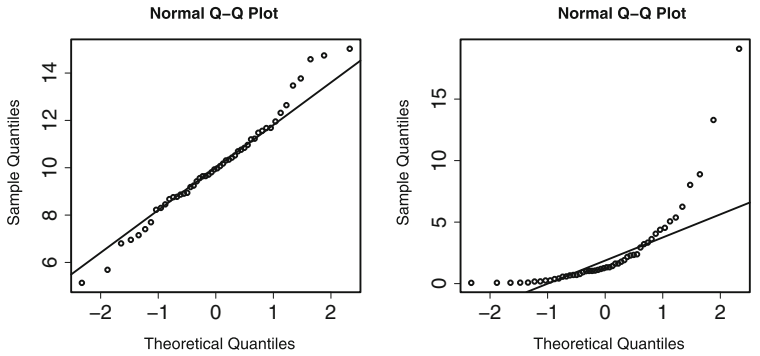

Fig. 1.10 Normal quantile–quantile plot of 50 Norm(10, 6) sample points (*left*) and 50 Exp(0.4) points (*right*). The solid line denotes the theoretical quantile of a normal distribution. While the normal data fit the line nicely, the exponential data deviate strongly from the expected quantiles of a normal distribution

A quantile-quantile plot for a normal sample is given in Fig. 1.10 (left) together with the theoretical quantiles (solid line). For comparison, sample points from an exponential distribution are plotted together with the normal distribution quantiles in Fig. 1.10 (right). As we can see, the agreement of the theoretical and the empirical quantiles is quite good for the normal sample, but it is poor for the exponential sample, especially in the tails.

Quantile plots can be generated in `R` using the function `qqplot()`. The functions `qqnorm()` and `qqline()` allow comparison to the normal distribution.

1.8.4 Barplots and Boxplots

It is still very popular to give data in terms of a bar with the height corresponding to the expectation and an additional error bar on top to indicate the standard deviation. However, this only shows two key numbers of the whole data (expectation and standard deviation), and does not allow to see how the data actually distribute. A much more informative alternative to plot data is to use the *boxplot*. It shows several parameters simultaneously: a rectangle denotes the positions of the 0.25- and 0.75-quantiles, with a horizontal line in the box showing the median (0.5-quantile). Thus, the middle 50% of the data are contained in that rectangle. On the top and bottom of the rectangle, the "whiskers" show the range of 1.5 times the distance between the 0.25- and 0.75-quantiles. Sample points outside this range are plotted individually. The previous normal and exponential data are both normalized to mean 2 and standard deviation 1 and the resulting data are shown as a barplot (left) and boxplot (right) in Fig. 1.11. In the barplot, no difference between the two samples can be noticed, while the different distributions of the data, the skewness of the exponential, and the symmetry of the normal are immediately recognized in the boxplot. Barplots with

Chapter 3 is devoted to hypothesis testing, with a main focus on the fundamental ideas and the interpretation of results. This chapter also contains sections on robust methods and correction for multiple testing, which become more and more important, especially in biology. Finally, Chap. 4 presents linear regression with one and several covariates and one-way analysis-of-variance. This chapter uses R more intensively to avoid tedious manual calculations, which the reader hopefully appreciates.

There surely is no shortage in statistics books. For further reading, I suggest to have a look at the two books by Wasserman: *All of Statistics* [1] and *All of Nonparametric Statistics* [2], which contain a much broader range of topics. The two books by Lehmann, *Theory of Point Estimation* [3] and *Testing Statistical Hypotheses* [4] contain almost everything one ever wanted to know about the material in Chaps. 2 and 3. For statistics using R, *Statistics—An Introduction using R* [5] by Crawley and *Introductory Statistics with R* [6] by Dalgaard are good choices, and *The R Book* [7] by Crawley offers a monumental reference. The *Tiny R Handbook* [8], published in the same series by Springer, might be a good companion to this book. For statistics related to bioinformatics, *Statistical Methods in Bioinformatics* [9] by Ewens and Grant provides lots of relevant information; the DNA sequence example is partly adapted from that book. Finally, for the german-speaking audience, I would recommend the two books by Pruscha *Statistisches Methodenbuch* [10], focusing on practical methods, and *Vorlesungen über mathematische Statistik* [11], its theory counterpart.

This script was typeset in LATEX, with all except the first two figures and all numerical data directly generated in R and included using Sweave [12].

I am indebted to many people that allowed this book to enter existence: I thank Jörg Stelling for his constant encouragement and support and for enabling me to work on this book. Elmar Hulliger, Ellis Whitehead, Markus Beat Dürr, Fabian Rudolf, and Robert Gnügge helped correcting various errors and provided many helpful suggestions. I thank my fiancée Elke Schlechter for her love and support. Financial support by the EU FP7 project UNICELLSYS is gratefully acknowledged. For all errors and flaws still lurking in the text, the figures, and the examples, I will nevertheless need to take full responsibility.

Basel, July 2011 Hans-Michael Kaltenbach

References

1. Wasserman, L.: All of Statistics. Springer, Heidelberg (2004)
2. Wasserman, L.: All of Nonparametric Statistics. Springer, Heidelberg (2006)
3. Lehmann, E.L., Casella, G.: Theory of Point Estimation, 2nd edn. Springer, Heidelberg (1998)
4. Lehmann, E.L., Romana, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, Heidelberg (2005)
5. Crawley, M.J.: Statistics—An Introduction using R. Wiley, New York (2005)

uncommonly large deviations of the observed test statistic t from the expected value under H_0. Using this level, we then compute the rejection region $\mathscr{R}_\alpha$ of T such that we reject the null hypothesis if the observed value t is inside the *rejection region* $\mathscr{R}_\alpha$, and do not reject if it is outside. This rejection region is computed from the distribution of T under H_0 and the two hypotheses. If the null hypothesis is rejected at a level α, we say that the test is *significant at level* α.

The p-value corresponds to the smallest level α such that the test would not yet reject. For example, if we observe a p-value of 0.034, the test would reject at the $\alpha = 0.05$ level, but not at the $\alpha = 0.01$ level. Indeed, no R-implementation of a statistical test requires a test level; they all report the p-value, so we can decide what to do.

Absence of evidence is not evidence of absence. Rejecting the null hypothesis hinges on the distribution of the test statistic *assuming* H_0 *is true*. The p-value gives the probability that we see a value at least as extreme as the observed one under this distribution. It is *not* the probability that H_0 is true. A very low p-value indicates that the test statistic is unlikely to take the observed value under H_0 and therefore provides good reason to reject it. On the contrary, a high p-value does not give proof that H_0 is actually true. In fact, it could be that the alternative just provides an even worse explanation, or the test has very low power to distinguish the two alternatives.

As a consequence, a statistical test should always be stated such that the null hypothesis defines the "status quo" and gets rejected if the desired result shows in the data. Thus, testing for a difference between a default and non-default assumption, we would set H_0 to state that there is no difference and reject this hypothesis in favor of the alternative that there is a difference. This way, we can quantify our level of confidence in the rejection by a low p-value. Would we set the no difference scenario as alternative, we would aim at "proving" H_0 with the data, which we can't.

Very informally, let T be our test statistic with value t for some specific data. Then, we compute $\mathbb{P}(T \geq t | H_0)$, the probability to see this data if H_0 is true, which is not $\mathbb{P}(H_0 | T \geq t)$, the probability of H_0 being true, given the data.

Stating the hypotheses. Stating a null hypothesis H_0 to reflect the default assumptions can become quite intricate for more involved problems. In particular, it might not be straightforward to formulate the hypothesis in terms of parameter regions or even to find the formal statistical hypothesis that correctly reflects our verbal hypothesis on the data. This lead some researchers to introduce a *type-III error* (see Sect. 3.5 for type-I/II errors), which is often stated as "asking the wrong question and using the wrong H_0", or "correctly rejecting H_0, but for the wrong reasons".

Additionally, we have also to take some care in stating a useful alternative hypothesis. As we already saw, the choice of the alternative partly determines the rejection region. For the case of one mean μ, the decision for a two- or a one-sided alternative can usually be decided from the problem itself. However, imagine we were to test two means at the same time with null hypothesis $H_0 : \mu_A > 0, \mu_B > 0$. A reasonable alternative is $H_1 : \mu_A < 0, \mu_B < 0$, but it is very strict and we may want to relax it to $H_1 : (\mu_A < 0 \text{ or } \mu_B < 0)$. Depending on which alternative we select, this leads to different rejection regions.

and the number of rejected hypotheses R is a random variable whose realization r can be observed. On the other hand, the number of false positives V is also a random variable, but its realization can not be observed. Intuitively, if we would pick one of the rejected hypotheses at random, the FDR can be interpreted as the expected chance that this hypothesis was falsely rejected. For 100 rejected hypotheses, an FDR of 0.05 would also mean that we expect that $100 \times 0.05 = 5$ of these are false positives. To achieve a certain FDR in a concrete situation, we need to choose the number of rejections such that the prescribed FDR is reached; we can do this by appropriately adjusting the p-value for which to reject.

Benjamini-Hochberg procedure. We would like to have strategy that guarantees a proportion of false positives of at most q^* among all rejected hypotheses. For this, we calculate at which p-value to reject a hypothesis such that the FDR stays below this desired threshold q^* as follows: we expect $\mathbb{E}(V) = \alpha k$ false positives in k tests for a given test level α. For a given set of data, the number of null hypotheses rejected at this level is $r(\alpha)$, a realization of R. The *Benjamini-Hochberg procedure* computes this number $r(\alpha)$ such that the maximal number of hypotheses are rejected while still keeping the expected proportion of false positives below the given threshold q^*, thus

$$\mathbb{E}\left(\frac{V}{R}\right) = \frac{\alpha k}{r(\alpha)} \leq q^*.$$

Let again p_i be the p-value of the ith hypothesis test and consider the ordered p-values $p_{(1)} \leq \cdots \leq p_{(k)}$. We then compute the largest index l such that

$$p_{(i)} \leq \frac{i}{k} q^*$$

for all $(i) < l$. The values $q_i = \frac{i}{k} q^*$ are sometimes called the q-values. One can show that if we reject those null hypotheses for which $p_i \leq q_l = \frac{l}{k} q^*$,

$$\text{FDR} \leq \frac{k_0}{k} q^* \leq q^*,$$

which guarantees the desired proportion of false positives.

Comparing p-values and q-values. Despite some similarities, p-values and q-values have some fundamental differences. The p-value gives the smallest test level at which not to reject and thus needs to be correct and exact to assess the data. On the other hand, q^* is a threshold for an expected ratio, so the actual ratio might be higher. It serves more as a "calling" tool that filters out uninteresting test results from a large number of performed tests. A proper analysis would then further investigate the remaining candidates, so it is usually not problematic if the desired q^* is not exactly achieved in the actual study.

Example 30 Let us consider the scenario that $k = 10$ individual tests were performed on data and that their ordered p-values $p_{(i)}$ are