8 Back to Bayes: knowledge as a probability distribution

In this chapter the reader is requested to sit back and think. Think about what you are doing and why, and what your conclusions really mean. You have a theory, containing a number of unknown – or insufficiently known – parameters, and you have a set of experimental data. You wish to use the data to validate your theory and to determine or refine the parameters in your theory. Your data contain inaccuracies and whatever you infer from your data contains inaccuracies as well. While the probability distribution of the data, given the theory, is often known or derivable from counting events, the *inverse*, i.e., the inferred probability distribution of the estimated parameters given the experimental outcome, is of a different, more subjective kind. Scientists who reject any subjective measures must restrict themselves to hypothesis testing. If you want more, turn to Bayes.

8.1 Direct and inverse probabilities

Consider the reading of a sensitive digital voltmeter sensing a constant small voltage – say in the microvolt range – during a given time, say 1 millisecond. Repeat the experiment many times. Since the voltmeter itself adds a random noise due to the thermal fluctuations in its input circuit, your observations y_i will be samples from a probability distribution $f(y_i - \theta)$, where θ is the real voltage of the source. You can determine f by collecting many samples. In some cases, when you know the physical process that adds the noise, you may even be able to predict the distribution function. For example, if you observe the number of light pulses in a given time interval Δt, knowing that they occur randomly at a given average rate θ, then the number k observed in a given time interval will obey the Poisson probability distribution $f(k, \theta\Delta t)$. Such (conditional) probabilities $f(y|\theta)$ are called *direct* probabilities. They result from direct counting of events or from considering symmetries in the random process. In this chapter the notation f is used for such direct probabilities which are also called *physical* probabilities.

Now consider the value of a physical constant, e.g. Avogadro's number N_{A}. It is the number of atoms in 1 gram of pure ^{12}C. According to CODATA its value is $(6.022\,141\,79 \pm 0.000\,000\,30) \times 10^{23}$. That is, the number given is not exact. At best a probability distribution $p(N_A)$ can be given for N_A, e.g. a normal distribution with mean $6.022\,141\,79 \times 10^{23}$ and standard deviation 3.0×10^{16}. But what does a probability of this kind mean? It is *not* a frequency distribution that you can find by counting the outcome of a large number of similar experiments, because – if there had been a large number of independent evaluations – the CODATA committee would have averaged those and proposed another mean and s.d. Similarly, the metereologist's prediction that "there is a 30 percent probability that it will rain tomorrow" or the surgeon's prediction that "the patient has a 95 percent chance of surviving the operation" says something about a unique event that cannot be repeated; such probabilities are more expressions of a belief based on earlier experience than the outcome of counting numbers in repeated experiments. Philosophers call such probabilities *epistemic*.[1] Other names are *subjective* and *inverse* probabilities.

The distinction between direct and inverse probabilities has been clear to Laplace and his followers since the late eighteenth century.[2] But the subjective nature of inverse probability has caused many scientists to shy away from using such concepts. The exponent of the critical school is the eminent statistician R. A. Fisher who developed a range of statistical tools in the first half of the twentieth century, all based on the frequency definition of direct probabilities. He circumvented the use of inverse probabilities by introducing the *likelihood* as a substitute.

There is a good reason to be critical to concepts in physics that are not entirely objective: subjective bias, arbitrariness and prejudice may easily creep into the interpretation of results. So, *if* you use inverse probabilities as an expression of your knowledge, it is essential that such probabilities are unbiased and do not include "information" that does not rest on verifiable knowledge. But with this restriction the use of inverse probabilities is very rich and very powerful to infer model parameters from experimental data.

Since the middle of the twentieth century the construction of inverse probabilities has gained ground over the critical school and is recently enjoying a real revival. It is called *the Bayesian approach*.

8.2 Enter Bayes

In two papers of Thomas Bayes (1763, 1764), published posthumously by R. Price, the principle of what is now called "Bayes' method" of constructing inverse probabilities was laid down within a context of combinatorial

[1] The word epistemic, from the Greek *epistèmè*: knowledge, was first used in the context of probabilities by Skyrms (1966).

[2] See Hald (2007) for a historical review of statistical inference.

problems. Ten years later the concept was worked out by Laplace. It is really very simple, once you agree to work with inverse probabilities.[3]

Consider two events, T and E (T stands for "theory": a parameter or set of parameters in a theory, and E for "experiment": an observed quantity or quantities). The probability of the joint event $p(\mathrm{T},\mathrm{E})$ can be expressed as the marginal probability of one event times the conditional probability of the other:

$$p(\mathrm{T},\mathrm{E}) = p(\mathrm{T})\,p(\mathrm{E}|\mathrm{T}) = p(\mathrm{E})p(\mathrm{T}|\mathrm{E}). \tag{8.1}$$

This implies that the *posterior* probability of T, $p(\mathrm{T}|\mathrm{E})$, i.e., the probability after the experiment is known, is proportional to the product of the *prior* probability of T, $p(\mathrm{T})$, i.e., the probability before the experiment is known, and the probability of the experimental outcome, given the theory:

$$p(\mathrm{T}|\mathrm{E}) \propto p(\mathrm{T})p(\mathrm{E}|\mathrm{T}). \tag{8.2}$$

The proportionality constant is really a normalization factor; it is simply the inverse of the sum or integral of the right-hand side over all possibilities of T.

In more specific terminology (and now the notation f is used for direct probabilities and p for inverse probabilities): you have a theory with a set of parameters $\boldsymbol{\theta}$ and you have a set of data $\boldsymbol{y}$. Now the posterior probability of your parameters is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = c f(\boldsymbol{y}|\boldsymbol{\theta})p_0(\boldsymbol{\theta}), \tag{8.3}$$

where $p_0(\boldsymbol{\theta})$ is the *prior* probability density function of your parameters. The latter expresses the knowledge you have about $\boldsymbol{\theta}$ before you know the experimental results. The constant c is given by

$$c^{-1} = \int f(\boldsymbol{y}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \tag{8.4}$$

with integration carried out over the full domain of possible values for $\boldsymbol{\theta}$. Here it is assumed that the parameters can take on a continuous range of values (with p being probability densities), but they can just as well be discrete, in which case the integration becomes a summation and the p's are probability mass functions. Likewise the direct probabilities $f(\boldsymbol{y}|\boldsymbol{\theta})$ can be either continuous or discrete.

[3] See, among many others, Box and Tiao (1973) and Lee (1989) for Bayesian treatments of statistical problems. Cox (2006) compares the frequentist and Bayesian approaches to statistical inference.

8.3 Choosing the prior

The prior distribution p_0 *must* be unbiased. It can only depend on previous experiments and derived by equations like (8.3). If no such experimental information is known, the prior must be as *uninformative* as possible: any information you put into the prior that does not rely on verifiable data introduces a form of prejudice.

The most uninformative prior is a constant: all values are equally possible. It seems a bit strange to propose a constant for a probability density function (pdf): a respectable pdf should be normalized, i.e., the integral over its domain should be unity. Probability densities that cannot be normalized are called *improper*. But you can make the constant pdf respectable if you cut its value to zero at the far ends beyond the range of possible values. Since the direct probability $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a peaked function with finite integral, the integral in (8.4) exists even for $p_0 \equiv 1$. So it is OK to allow improper priors.

There is one objective requirement for an acceptable prior: it should scale properly with transformations of the parameters. Consider a *location parameter* μ, occurring as an additive factor in the range $(-\infty, \infty)$. It could just as well be replaced by a linear transformation $\mu' = a\mu + b$; a uniform distribution of μ should also be uniform if expressed in μ'. That is indeed the case: since $p(\mu)\, d\mu = p'(\mu')\, d\mu'$ and $d\mu' = a d\mu, p' = p/a$, which is uniform if p is a constant. Now consider a *scale parameter* σ, occurring as a multiplicative factor in the range $(0, \infty)$. It could just as well be replaced by $c\sigma$ or by σ^2 or by σ^{-1}, or by the transformation $\sigma' = b\sigma^a$. It is clear that the variable $\log\sigma$ transforms linearly: $\log\sigma' = a\log\sigma + b$; therefore the distribution should be uniform in $\log\sigma$. This implies that the uncommitted (or *ignorant*) prior should be proportional to $1/\sigma$ since $d\log\sigma = d\sigma/\sigma$. Summarizing (this rule is due to Jeffreys, 1939):

The most uninformative (ignorant, uncommitted, unbiased) improper prior $p_0(\theta)$ *equals 1 if* θ *is a location parameter or* $1/\theta$ *if* θ *is a scale parameter.*

8.4 Three examples of Bayesian inference

Updating knowledge: Avogadro's number

CODATA suggests that we may believe the inverse probability density function of Avogadro's number to be

$$p_0(N_{\mathrm{A}}) \propto \exp\left[-\frac{(N_A - \mu_0)^2}{2\sigma_0^2}\right], \tag{8.5}$$

with $\mu_0 = 6.022\,141\,79 \times 10^{23}$ and $\sigma_0 = 3.0 \times 10^{16}$.

A scientist comes along with a reliable new measurement of N_A. She measured the value $y = 6.022\,141\,48 \times 10^{23}$ and asserts that her analysis of experimental errors indicates that her result y is a sample from a normal distribution $N(y - N_\mathrm{A}, \sigma_1)$, where $\sigma_1 = 7.5 \times 10^{16}$.

Inserting these data into (8.3) we find that

$$p(N_\mathrm{A}|y) \propto \exp\left[-\frac{(y-N_\mathrm{A})^2}{2\sigma_1^2}\right] \exp\left[-\frac{(N_\mathrm{A}-\mu_0)^2}{2\sigma_0^2}\right]. \tag{8.6}$$

Working out the exponent $\big($omitting a factor $-\frac{1}{2}$ for the time being$\big)$:

$$\frac{(y-N_\mathrm{A})^2}{\sigma_1^2} + \frac{(N_\mathrm{A}-\mu_0)^2}{\sigma_0^2} \tag{8.7}$$

$$= \left(\sigma_0^{-2} + \sigma_1^{-2}\right)\left[N_\mathrm{A}^2 - 2N_\mathrm{A}\frac{\mu_0\sigma_0^{-2} + y\sigma_1^{-2}}{\sigma_0^{-2} + \sigma_1^{-2}} + \cdots\right] \tag{8.8}$$

$$= \frac{(N_\mathrm{A}-\mu)^2}{\sigma^2} + \cdots, \tag{8.9}$$

where

$$\mu = \frac{\mu_0\sigma_0^{-2} + y\sigma_1^{-2}}{\sigma_0^{-2} + \sigma_1^{-2}}, \tag{8.10}$$

$$\sigma^{-2} = \sigma_0^{-2} + \sigma_1^{-2}. \tag{8.11}$$

Thus the posteriori inverse probability density of N_A is a normal distribution with weighted averages for the mean and variance (see also Exercise 5.7 on page 70):

$$p(N_\mathrm{A}|y) \propto \exp\left[-\frac{(N_\mathrm{A}-\mu)^2}{2\sigma^2}\right]. \tag{8.12}$$

The result is that the parameters μ_0 and σ_0 in the prior pdf (8.5) have been updated to μ and σ in the posterior pdf (8.12).

Inference from a series of normally distributed samples

Suppose your experimental data are n independent samples from a normal distribution with unknown μ and unknown σ. You have no prior knowledge of the mean and s.d., so you take the uninformative prior

$$p_0(\mu, \sigma) = 1/\sigma, \tag{8.13}$$

because μ is a location parameter and σ is a scale parameter. The probability of observing n values $y_i, i = 1, \ldots, n$ is the product of the probabilities of all measurements, because the data are independent:

$$f(\mathbf{y}|\mu,\sigma) = \Pi_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \tag{8.14}$$

$$\propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \tag{8.15}$$

This can be rewritten as

$$\sigma^{-n} \exp\left[-\frac{(\langle y\rangle-\mu)^2 + \langle(\Delta y)^2\rangle}{2\sigma^2/n}\right], \tag{8.16}$$

where

$$\langle y\rangle = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{8.17}$$

and

$$\langle(\Delta y)^2\rangle = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^2. \tag{8.18}$$

For the posterior probability density we can now write:

$$p(\mu,\sigma|\mathbf{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{(\mu-\langle y\rangle)^2 + \langle(\Delta y)^2\rangle}{2\sigma^2/n}\right]. \tag{8.19}$$

The proportionality constant can be obtained by integrating the right-hand side over both μ and σ. In this case there is an analytical expression for the integral, but it is often easier to determine the constant by numerical integration.

It is interesting to see that the probability (8.19) of the parameters is given by only two properties of the data set: the average and the mean-squared deviation from the average. Apparently these two properties are sufficient to know everything about the statistics of the data set (*sufficient statistics*). But this is only true if we already know that the samples come from a normal distribution!

The pdf of (8.19) is *bivariate*. It is plotted as a number of contours at various fractional heights in Fig. 8.1 for the example of 10 samples with $\langle y\rangle = 0$ and $\langle(\Delta y)^2\rangle = 1$. The values within a given contour represent a defined integrated probability.

In practice you will more often find use for one-dimensional distribution functions. First consider the pdf for μ (Fig. 8.2).

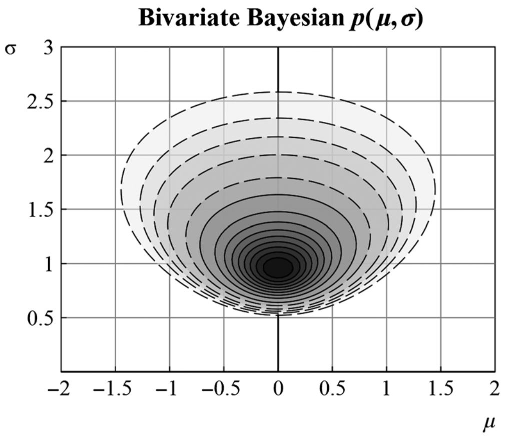

Figure 8.1 Contour plot of the Bayesian inverse bivariate pdf of the mean μ and s.d. σ given the value of 10 independent normally distributed experimental samples. The average equals zero and the rmsd equals 1. Contours – from inside out – are full-drawn at fractional heights 0.9, 0.8, . . . , 0.1; broken at 0.05, 0.02, 0.01, 0.005, 0.002.

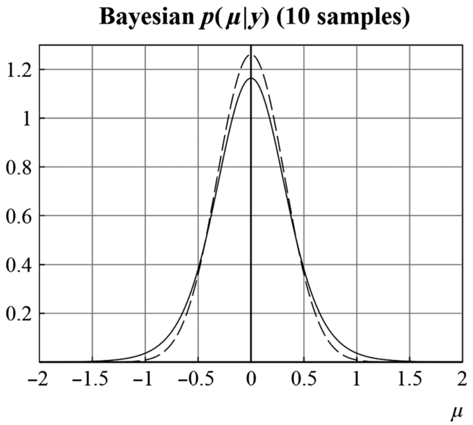

Figure 8.2 The Bayesian posterior pdf for the parameter μ, given the value of 10 independent normally distributed experimental samples with zero average and rmsd = 1. The drawn line is the marginal $p(\mu|\mathbf{y})$ for unknown σ; the broken line is $p(\mu|\mathbf{y}, \sigma)$ for known $\sigma = 1$.

For *known* σ you see from (8.19) that the posterior pdf of μ is a normal distribution around $\langle y \rangle$ with variance σ^2/n:

$$p(\mu|\mathbf{y},\sigma) \propto \exp\left[-\frac{(\mu - \langle y \rangle)^2}{2\sigma^2/n}\right]. \tag{8.20}$$

For *unknown* σ the probability must be integrated over all possible values of σ in order to obtain a *marginal* distribution of μ.

$$p(\mu|\mathbf{y}) = \int_0^\infty p(\mu, \sigma|\mathbf{y})\, d\sigma. \tag{8.21}$$

This integral can be written as proportional to

$$\int_0^\infty \sigma^{-(n+1)} \exp\left(-\frac{q}{\sigma^2}\right) d\sigma, \tag{8.22}$$

where

$$q = \frac{1}{2}n[(\mu - \langle y \rangle)^2 + \langle (\Delta y)^2 \rangle]. \tag{8.23}$$

By substituting q/σ^2 for a new variable, a Gamma-function is obtained. The integral appears to be proportional to

$$p(\mu|\mathbf{y}) \propto \left(1 + \frac{(\mu - \langle y \rangle)^2}{\langle (\Delta y)^2 \rangle}\right)^{-n/2}. \tag{8.24}$$

This is exactly Student's t-distribution density function $f(t|\nu)$ for $\nu = n - 1$ degrees of freedom, as a function of the variable t:

$$f(t|\nu) \propto \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \tag{8.25}$$

$$t = \sqrt{\frac{(n-1)(\mu - \langle y \rangle)^2}{\langle (\Delta y)^2 \rangle}} = \frac{\mu - \langle y \rangle}{\hat{\sigma}/\sqrt{n}}, \tag{8.26}$$

where $\hat{\sigma}^2 = [n/(n-1)]\, \langle (\Delta y)^2 \rangle$. See data sheet STUDENT'S T-DISTRIBUTION on page 213 for further details on the t-distribution.

Next consider the pdf for σ. If μ is known, the pdf is given by (8.19). Figure 8.3 shows $p(\sigma|\mathbf{y}, \mu = 0)$ for the example used above. It is more common that you do *not* know μ in advance; then your Bayesian posterior probability is the marginal probability:

$$p(\sigma|\mathbf{y}) = \int_{-\infty}^\infty p(\mu, \sigma|\mathbf{y})\, d\mu \tag{8.27}$$

$$\propto \sigma^{-n} \exp\left[-\frac{\langle (\Delta x)^2 \rangle}{2\sigma^2/n}\right]. \tag{8.28}$$

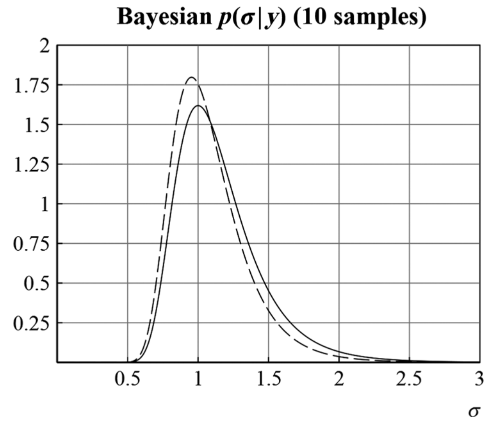

Figure 8.3 The Bayesian posterior pdf for the parameter σ, given the value of 10 independent normally distributed experimental samples with zero average and rmsd = 1. The drawn line is the marginal $p(\sigma|\boldsymbol{y})$ for unknown μ; the broken line is $p(\sigma|\boldsymbol{y}, \mu)$ for known $\mu = 0$.

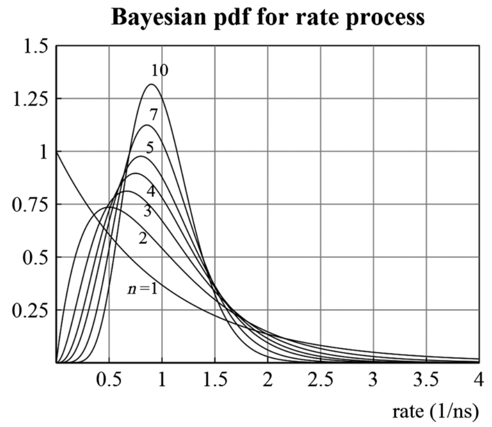

Figure 8.4 The Bayesian posterior pdf for the rate parameter k, given the value of n independent time intervals between events. For this example the average observed time equals 1 ns. The pdf's are drawn for $n = 1, 2, 3, 4, 5, 7, 10$.

As you see from Fig. 8.3, the value predicted for σ is slightly larger and slightly less accurate when μ is not known *a priori*.

Infer a rate constant from a few events

Consider the observation of single events that sample a rate process. This could be a pulse emitted from a source that is excited at $t = 0$; it could be the observation of a conformational change in the simulation of a protein that is made unstable by changing its environment at $t = 0$; it could be the time between two sightings of a meteor, or any other seldom event that you can observe only a few times. Your theory says that the event results from a simple rate process with constant probability $k\Delta t$ that the event occurs in any small time interval Δt. You observe n independent events at times or intervals $t_i, i = 1, \dots, n$. What can you say about the rate constant k?

In a Bayesian approach you wish to determine after the first event the inverse posterior probability

$$p_1(k|t_1) \propto f(t_1|k) p_0(k), \tag{8.29}$$

where $f(t|k)$ is the direct probability that an event occurs after a time t, given the rate constant k. This is easy to derive. Divide time in small intervals Δt; $t/\Delta t = m$. The probability that the pulse occurs at the m-th interval, and not before, is $(1 - k\Delta t)^{m-1} k\Delta t$. Taking the limit $\Delta t \to 0; m \to \infty$, you find

$$f(t|k) = ke^{-kt}. \tag{8.30}$$

The prior inverse probability $p_0(k)$ must be taken as $1/k$, since k is a scale parameter. So

$$p_1(k|t_1) \propto e^{-kt_1}. \tag{8.31}$$

After observing a second event at time t_2, you can update this probability:

$$p_2(k|t_1, t_2) \propto ke^{-kt_2} e^{-kt_1} \tag{8.32}$$

and after n events you obtain

$$p_n(k|t_1, \dots, t_n) \propto k^{(n-1)} \exp[-k(t_1 + \cdots, t_n)]. \tag{8.33}$$

In general, if the average of the observed time intervals is $\langle t \rangle$, and the proportionality constant is included by integrating this function, it is found that

$$p_n(k|t_1, \dots, t_n) = \frac{(n\langle t \rangle)^n}{(n-1)!} k^{n-1} \exp(-kn\langle t \rangle). \tag{8.34}$$

So you see that the average of the observation times is sufficient statistics: it determines all you can know about k. It is easily seen that the expectations over this distribution of k and its variance are given by

$$\hat{k} = E[k] = \frac{1}{\langle t \rangle}, \tag{8.35}$$

$$\hat{\sigma}^2 = E[(k - \langle k \rangle)^2] = \frac{1}{n \langle t \rangle^2}. \tag{8.36}$$

The latter equation implies that $\hat{\sigma} = \hat{k}/\sqrt{n}$. As always, the relative standard inaccuracy decreases with the square root of the number of observations.

The case $n = 7$ has been used before in this book: Fig. 2.5 on page 12. For that case three different *point estimates* are given: the mean (1.00), the median (0.95) and the mode (0.86). It is pointless to haggle about what is best, as all values are well within a s.d. (0.38) from the mean.

8.5 Conclusion

The examples above express your knowledge in terms of probability density functions. These have one disadvantage: they look much more exact than they are. Be aware that such probability distributions only express your degree of ignorance about the parameters derived from theory and experiment. Your best value is not necessarily the exact mean or the exact mode; it can be anywhere within the width of the distribution. Be careful to report the right number of digits!

How to proceed if you absolutely refuse to use inverse probabilities? First, you can fool yourself by defining the *likelihood* of a parameter as equal to the direct probability of the measured value:

$$l(\theta|y) = f(y|\theta). \tag{8.37}$$

This is of course equivalent to the posterior Bayesian probability when you assume a uniform prior. Inconsistencies are expected for scale parameters. Renaming a quantity does not solve your problem but hides it like an ostrich does.

Second, you can limit yourself to testing hypotheses rather than predicting values. It is useful if you wish to assess the effect of some agent that does or does not influence a sampled result. The *null hypothesis* usually assumes that the agent has no effect and you try to prove that the result you obtain is unlikely under the null hypothesis. If so, you accept the truth of the alternative hypothesis ("the agent does have an effect"). This procedure avoids any inverse probabilities, but it makes life quite poor: you also want to know *what* the effect of the agent is. Many questions you wish to be answered by your experiments remain out of bounds.

Summary *This chapter has taken a Bayesian point of view on statistics, accepting the notion of "inverse probability". Simple rules then allow you to express all the knowledge you have, including the outcome of your recent experiments, in probability functions of the parameters in your theory. In three examples it is shown that you can either update existing knowledge with new experimental data or – without prior knowledge – express the knowledge gained from limited experimental data in probability distributions. The introduction to this chapter invited you to sit back and think. Now, sit back and draw your conclusions.*