13

The law of large numbers

For many experiments and observations concerning natural phenomena—such as measuring the speed of light—one finds that performing the procedure twice under (what seem) identical conditions results in two different outcomes. Uncontrollable factors cause "random" variation. In practice one tries to overcome this as follows: the experiment is repeated a number of times and the results are averaged in some way. In this chapter we will see why this works so well, using a model for repeated measurements. We view them as a sequence of independent random variables, each with the same unknown distribution. It is a probabilistic fact that from such a sequence—in principle—any feature of the distribution can be recovered. This is a consequence of the law of large numbers.

13.1 Averages vary less

Scientists and engineers involved in experimental work have known for centuries that more accurate answers are obtained when measurements or experiments are repeated a number of times and one averages the individual outcomes.[1] For example, if you read a description of A.A. Michelson's work done in 1879 to determine the speed of light, you would find that for each value he collected, repeated measurements at several levels were performed. In an article in *Statistical Science* describing his work ([18]), R.J. MacKay and R.W. Oldford state: "It is clear that Michelson appreciated the power of averaging to reduce variability in measurement." We shall see that we can understand this reduction using only what we have learned so far about probability in combination with a simple inequality called Chebyshev's inequality.

Throughout this chapter we consider a sequence of random variables X_1, X_2, X_3, ... You should think of X_i as the result of the ith repetition of a particular measurement or experiment. We confine ourselves to the situation where

[1] We leave the problem of systematic errors aside but will return to it in Chapter 19.

experimental conditions of subsequent experiments are identical, and the outcome of any one experiment does not influence the outcomes of others. Under those circumstances, the random variables of the sequence are independent, and all have the same distribution, and we therefore call $X_1, X_2, X_3, \ldots$ an *independent and identically distributed sequence*. We shall denote the distribution function of each random variable X_i by F, its expectation by μ, and the standard deviation by σ.

The average of the first n random variables in the sequence is

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

and using linearity of expectations we find:

$$\mathrm{E}\left[\bar{X}_n\right] = \frac{1}{n}\mathrm{E}[X_1 + X_2 + \cdots + X_n] = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu.$$

By the variance-of-the-sum rule, using the independence of $X_1, \ldots, X_n$,

$$\mathrm{Var}\left(\bar{X}_n\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

This establishes the following rule.

EXPECTATION AND VARIANCE OF AN AVERAGE. If $\bar{X}_n$ is the average of n independent random variables with the same expectation μ and variance σ^2, then

$$\mathrm{E}\left[\bar{X}_n\right] = \mu \quad \text{and} \quad \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}.$$

The expectation of $\bar{X}_n$ is again μ, and its standard deviation is less than that of a single X_i by a factor $\sqrt{n}$; the "typical distance" from μ is $\sqrt{n}$ smaller. The latter property is what Michelson used to gain accuracy. To illustrate this, we analyze an example.

Suppose the random variables $X_1, X_2, \ldots$ are continuous with a $Gam(2, 1)$ distribution, so with probability density:

$$f(x) = x\mathrm{e}^{-x} \qquad \text{for } x \geq 0.$$

Recall from Section 11.2 that this means that each X_i is distributed as the sum of two independent $Exp(1)$ random variables. Hence, $S_n = X_1 + \cdots + X_n$ is distributed as the sum of $2n$ independent $Exp(1)$ random variables, which has a $Gam(2n, 1)$ distribution, with probability density

$$f_{S_n}(x) = \frac{x^{2n-1}\mathrm{e}^{-x}}{(2n-1)!} \qquad \text{for } x \geq 0.$$

Because $\bar{X}_n = S_n/n$, we find by applying the change-of-units rule (page 106):

$$f_{\bar{X}_n}(x) = n f_{S_n}(nx) = \frac{n\,(nx)^{2n-1}\,\mathrm{e}^{-nx}}{(2n-1)!} \qquad \text{for } x \geq 0.$$

This is the probability density of the $Gam(2n, n)$ distribution.

So we have determined the distribution of $\bar{X}_n$ explicitly and we can investigate what happens as n increases, for example, by plotting probability densities. In the left-hand column of Figure 13.1 you see plots of $f_{\bar{X}_n}$ for $n = 1$, 2, 4, 9, 16, and 400 (note that for $n = 1$ this is just f itself). For comparison, we take as a second example a so-called *bimodal* density function: a density with two bumps, formally called *modes*. For the same values of n we determined the probability density function of $\bar{X}_n$ (unlike the previous example, we are not concerned with the computations, just with the results). The graphs of these densities are given side by side with the gamma densities in Figure 13.1.

The graphs clearly show that, as n increases, there is "contraction" of the probability mass near the expected value μ (for the gamma densities this is 2, for the bimodal densities 2.625).

Quick exercise 13.1 Compare the probabilities that $\bar{X}_n$ is within 0.5 of its expected value for $n = 1$, 4, 16, and 400. Do this for the gamma case only by estimating the probabilities from the graphs in the left-hand column of Figure 13.1.

13.2 Chebyshev's inequality

The contraction of probability mass near the expectation is a consequence of the fact that, for any probability distribution, most probability mass is within a few standard deviations from the expectation. To show this we will employ the following tool, which provides a bound for the probability that the random variable Y is outside the interval $(\mathrm{E}[Y] - a, \mathrm{E}[Y] + a)$.

Chebyshev's inequality. For an arbitrary random variable Y and any $a > 0$:

$$\mathrm{P}(|Y - \mathrm{E}[Y]\,| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y)\,.$$

We shall derive this inequality for continuous Y (the discrete case is similar). Let f_Y be the probability density function of Y. Let μ denote $\mathrm{E}[Y]$. Then:

$$\begin{aligned}\mathrm{Var}(Y) &= \int_{-\infty}^{\infty} (y-\mu)^2 f_Y(y)\,\mathrm{d}y \geq \int_{|y-\mu|\geq a} (y-\mu)^2 f_Y(y)\,\mathrm{d}y \\ &\geq \int_{|y-\mu|\geq a} a^2 f_Y(y)\,\mathrm{d}y = a^2 \mathrm{P}(|Y-\mu| \geq a)\,.\end{aligned}$$

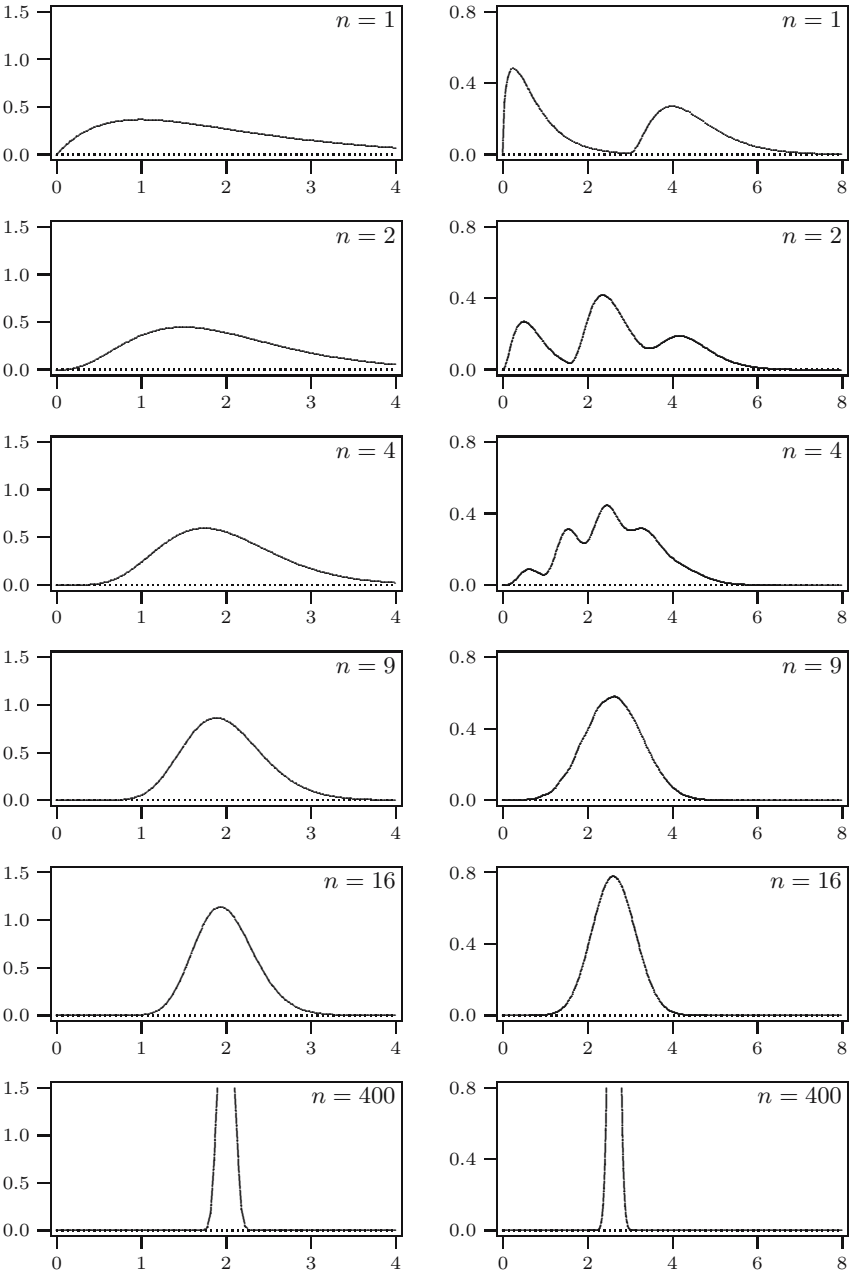

Fig. 13.1. Densities of averages. Left column: from a gamma density; right column: from a bimodal density.

Dividing both sides of the resulting inequality by a^2, we obtain Chebyshev's inequality.

Denote $\mathrm{Var}(Y)$ by σ^2 and consider the probability that Y is within a few standard deviations from its expectation μ:

$$\mathrm{P}(|Y-\mu| < k\sigma) = 1 - \mathrm{P}(|Y-\mu| \geq k\sigma)\,,$$

where k is a small integer. Setting $a = k\sigma$ in Chebyshev's inequality, we find

$$\mathrm{P}(|Y-\mu| < k\sigma) \geq 1 - \frac{\mathrm{Var}(Y)}{k^2\sigma^2} = 1 - \frac{1}{k^2}. \tag{13.1}$$

For $k = 2, 3, 4$ the right-hand side is $3/4$, $8/9$, and $15/16$, respectively. This suggests that with Chebyshev's inequality we can make very strong statements. For most distributions, however, the actual value of $\mathrm{P}(|Y-\mu| < k\sigma)$ is even *higher* than the lower bound (13.1). We summarize this as a somewhat loose rule.

THE "$\mu \pm$ A FEW σ" RULE. Most of the probability mass of a random variable is within a few standard deviations from its expectation.

QUICK EXERCISE 13.2 Calculate $\mathrm{P}(|Y-\mu| < k\sigma)$ exactly for $k = 1, 2, 3, 4$ when Y has an $Exp(1)$ distribution and compare this with the bounds from Chebyshev's inequality.

13.3 The law of large numbers

We return to the independent and identically distributed sequence of random variables $X_1, X_2, \ldots$ with expectation μ and variance σ^2. We apply Chebyshev's inequality to the average $\bar{X}_n$, where we use $\mathrm{E}\big[\bar{X}_n\big] = \mu$ and $\mathrm{Var}\big(\bar{X}_n\big) = \sigma^2/n$, and where $\varepsilon > 0$:

$$\mathrm{P}\big(\big|\bar{X}_n - \mu\big| > \varepsilon\big) = \mathrm{P}\big(\big|\bar{X}_n - \mathrm{E}\big[\bar{X}_n\big]\big| > \varepsilon\big) \leq \frac{1}{\varepsilon^2}\mathrm{Var}\big(\bar{X}_n\big) = \frac{\sigma^2}{n\varepsilon^2}.$$

The right-hand side vanishes as n goes to infinity, no matter how small ε is. This proves the following law.

THE LAW OF LARGE NUMBERS. If $\bar{X}_n$ is the average of n independent random variables with expectation μ and variance σ^2, then for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathrm{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0.$$

A connection with experimental work

Let us try to interpret the law of large numbers from an experimenter's perspective. Imagine you conduct a series of experiments. The experimental setup is complicated and your measurements vary quite a bit around the "true" value you are after. Suppose (unknown to you) your measurements have a gamma distribution, and its expectation is what you want to determine. You decide to do a certain number of measurements, say n, and to use their average as your estimate of the expectation.

We can simulate all this, and Figure 13.2 shows the results of a simulation, where we chose the same $Gam(2, 1)$ distribution, i.e., with expectation $\mu = 2$. We anticipated that you might want to do as many as 500 measurements, so we generated realizations for $X_1, X_2, \ldots, X_{500}$. For each n we computed the average of the first n values and plotted these averages against n in Figure 13.2.

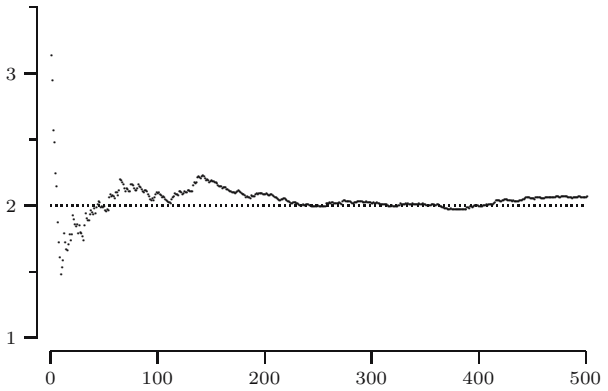

Fig. 13.2. Averages of realizations of a sequence of gamma distributed random variables.

If your decision is to do 200 repetitions, you would find (in this simulation) a value of about 2.09 (slightly too high, but you wouldn't know!), whereas with $n = 400$ you would be almost exactly correct with 1.99, and with $n = 500$ again a little farther away with 2.06. For another sequence of realizations, the details in the pattern that you see in Figure 13.2 would be different, but the general dampening of the oscillations would still be present. This follows from what we saw earlier, that as n is larger, the probability for the average to be within a certain distance of the expectation increases, in the limit even to 1. In practice it *may* happen that with a large number of repetitions your average is farther from the "true" value than with a smaller number of repetitions—if it is, then you had bad luck, because the odds are in your favor.

The averages may fail to converge

The law of large numbers is valid if the expectation of the distribution F is finite. This is not always the case. For example, the Cauchy and some Pareto distributions have heavy tails: their probability densities do go to 0 as x becomes large, but (too) slowly.[2] On the left in Figure 13.3 you see the result of a simulation with $Cau(2,1)$ random variables. As in the gamma case, the averages tend to go toward 2 (which is the point of symmetry of the $Cau(2,1)$ density), but once in a while a very large (positive or negative) realization of an X_i throws off the average.

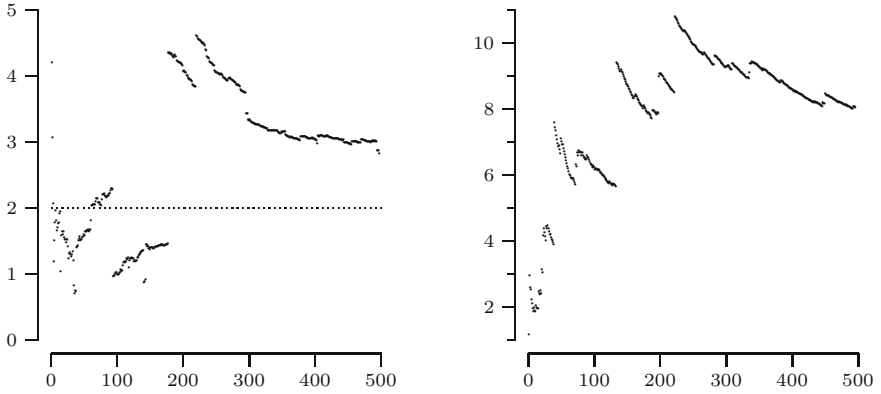

Fig. 13.3. Averages of realizations of a sequence of Cauchy (at left) and Pareto (at right) distributed random variables.

On the right in Figure 13.3 the result of a simulation with a $Par(0.99)$ distribution is shown. Its expectation is infinite. In the plot we see segments where the average "drifts downward," separated by upward jumps, which correspond to X_i with extremely large values. The effect of the jumps dominates: it can be shown that $\bar{X}_n$ grows beyond any level.

You might think that these patterns are phenomena that occur because of the short length of the simulation and that in longer simulations they would disappear after some value of n. However, the patterns as described will continue to occur and the results of a longer simulation, say to $n = 5000$, would not look any "better."

> **Remark 13.1 (There is a stronger law of large numbers).** Even though it is a strong statement, the law of large numbers in this paragraph is more accurately known as the *weak* law of large numbers. A stronger result holds, the *strong* law of large numbers, which says that:

[2] They represent two separate cases: the Cauchy expectation does not exist (see Remark 7.1) and the $Par(\alpha)$'s expectation is $+\infty$ if $\alpha \leq 1$ (see Section 7.2).

$$\mathrm{P}\Big(\lim_{n\to\infty} \bar{X}_n = \mu\Big) = 1.$$

This is also expressed as "as n goes to infinity, $\bar{X}_n$ converges to μ with probability 1." It is not easy to see, but it is true that the strong law is actually stronger. The conditions for the law of large numbers, as stated in this section, could be relaxed. They suffice for both versions of the law. The conditions can be weakened to a point where the weak law still follows from them, but the strong law does not anymore; the strong law requires the stronger conditions.

13.4 Consequences of the law of large numbers

We continue with the sequence $X_1, X_2, \ldots$ of independent random variables with distribution function F. In the previous section we saw how we could recover the (unknown) expectation μ from a realization of the sequence. We shall see that in fact we can recover any feature of the probability distribution. In order to avoid unnecessary indices, as in $\mathrm{E}[X_1]$ and $\mathrm{P}(X_1 \in C)$, we introduce an additional random variable X that also has F as its distribution function.

Recovering the probability of an event

Suppose that, rather than being interested in $\mu = \mathrm{E}[X]$, we want to know the probability of an event, for example,

$$p = \mathrm{P}(X \in C)\,, \quad \text{where } C = (a, b] \text{ for some } a < b.$$

If you do not know this probability p, you would probably estimate it from how often the event $\{X_i \in C\}$ occurs in the sequence. You would use the relative frequency of $X_i \in C$ among $X_1, \ldots, X_n$: the number of times the set C was hit divided by n. Define for each i:

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases}$$

The random variable Y_i indicates whether the corresponding X_i hits the set C; it is called an *indicator random variable*. In general, an indicator random variable for an event A is a random variable that is 1 when A occurs and 0 when A^c occurs. Using this terminology, Y_i is the indicator random variable of the event $X_i \in C$. Its expectation is given by

$$\mathrm{E}[Y_i] = 1 \cdot \mathrm{P}(X_i \in C) + 0 \cdot \mathrm{P}(X_i \notin C) = \mathrm{P}(X_i \in C) = \mathrm{P}(X \in C) = p.$$

Using the Y_i, the relative frequency is expressed as $(Y_1 + Y_2 + \cdots + Y_n)/n = \bar{Y}_n$. Note that the random variables $Y_1, Y_2, \ldots$ are independent; the X_i form an independent sequence, and Y_i is determined from X_i only (this is an application of the rule about propagation of independence; see page 126).

The law of large numbers, with p in the role of μ, can now be applied to $\bar{Y}_n$; it is the average of n independent random variables with expectation p and variance $p(1-p)$, so

$$\lim_{n\to\infty} \mathrm{P}\big(|\bar{Y}_n - p| > \varepsilon\big) = 0 \tag{13.2}$$

for any $\varepsilon > 0$. By reasoning along the same lines as in the previous section, we see that from a long sequence of realizations we can get an accurate estimate of the probability p.

Recovering the probability density function

Consider the continuous case, where f is the probability density function corresponding with F, and now choose $C = (a-h, a+h]$, for some (small) positive h. By equation (13.2), for large n:

$$\bar{Y}_n \approx p = \mathrm{P}(X \in C) = \int_{a-h}^{a+h} f(x)\,\mathrm{d}x \approx 2hf(a). \tag{13.3}$$

This relationship suggests to estimate the probability density in a as follows:

$$f(a) \approx \frac{\bar{Y}_n}{2h} = \frac{\text{the number of times } X_i \in C \text{ for } i \le n}{n \cdot \text{the length of } C}.$$

In Figure 13.4 we have done so for $h = 0.25$ and two values of a: 2 and 4. Rather than plotting the estimate in just one point, we use the same value for the whole interval $(a-h, a+h]$. This results in a vertical bar, whose area corresponds to $\bar{Y}_n$:

$$\text{height} \cdot \text{width} = \frac{\bar{Y}_n}{2h} \cdot 2h = \bar{Y}_n.$$

These estimates are based on the realizations of 500 independent $Gam(2,1)$ distributed random variables. In order to be able to see how well things came

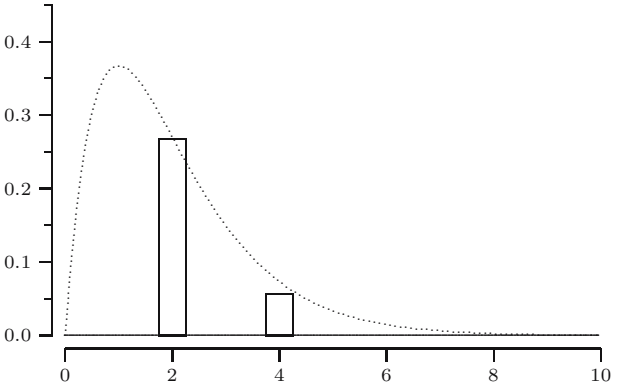

Fig. 13.4. Estimating the density at two points.

out, the $Gam(2,1)$ density function is shown as well; near $a = 2$ the estimate is very accurate, but around $a = 4$ it is a little too low.

There really is no reason to derive estimated values around just a few points, as is done in Figure 13.4. We might as well cover the whole x-axis with a grid (with grid size $2h$) and do the computation for each point in the grid, thus covering the axis with a series of bars. The resulting bar graph is called a *histogram*. Figure 13.5 shows the result for two sets of realizations.

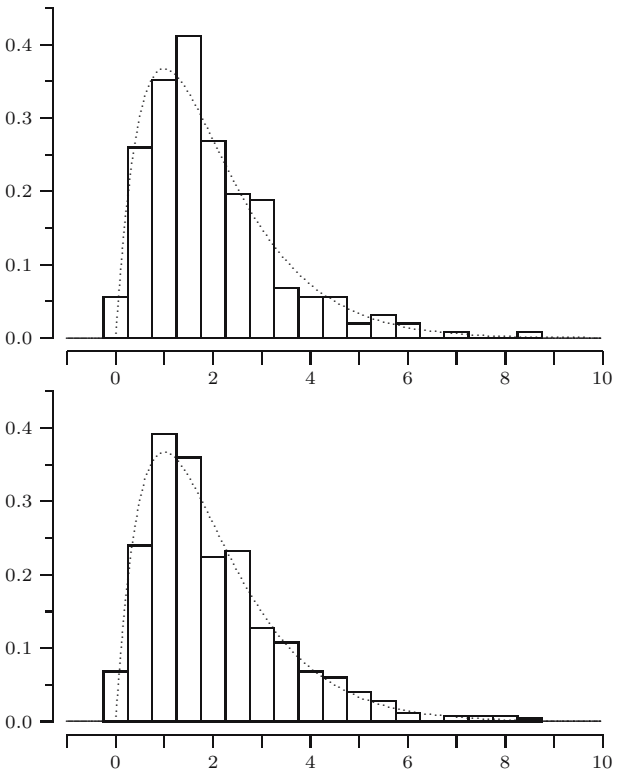

Fig. 13.5. Recovering the density function by way of histograms.

The top graph is constructed from the same realizations as Figure 13.4 and the bottom graph is constructed from a new set of realizations. Both graphs match the general shape of the density, with some bumps and valleys that are particular for the corresponding set of realizations. In Chapters 15 and 17 we shall return to histograms and treat them more elaborately.

QUICK EXERCISE 13.3 The height of the bar at $x = 2$ in the first histogram is 0.26. How many of the 500 realizations were between 1.75 and 2.25?

13.5 Solutions to the quick exercises

13.1 The answers you have found should be in the neighborhood of the following exact values:

n	1	4	16	400
$\mathrm{P}\big(\lvert\bar{X}_n - \mu\rvert < 0.5\big)$	0.27	0.52	0.85	1.00

13.2 Because Y has an $Exp(1)$ distribution $\mu = 1$ and $\mathrm{Var}(Y) = \sigma^2 = 1$; we find for $k \geq 1$:

$$\begin{aligned}\mathrm{P}(|Y - \mu| < k\sigma) &= \mathrm{P}(|Y - 1| < k) \\ &= \mathrm{P}(1 - k < Y < k + 1) = \mathrm{P}(Y < k + 1) = 1 - \mathrm{e}^{-k-1}.\end{aligned}$$

Using this formula and (13.1) we obtain the following numbers:

k	1	2	3	4
Lower bound from Chebyshev	0	0.750	0.889	0.938
$\mathrm{P}(\lvert Y - 1\rvert < k)$	0.865	0.950	0.982	0.993

13.3 The value of $\bar{Y}_n$ for this bar equals its area $0.26 \cdot 0.5 = 0.13$. The bar represents 13% of the values, or $0.13 \cdot 500 = 65$ realizations.

13.6 Exercises

13.1 Verify the "$\mu \pm$a few σ" rule as you did in Quick exercise 13.2 for the following distributions: $U(-1, 1)$, $U(-a, a)$, $N(0, 1)$, $N(\mu, \sigma^2)$, $Par(3)$, $Geo(1/2)$. Construct a table as in the answer to the quick exercise and enter a line for each distribution.

13.2 ⊞ An accountant wants to simplify his bookkeeping by rounding amounts to the nearest integer, for example, rounding €99.53 and €100.46 both to €100. What is the cumulative effect of this if there are, say, 100 amounts? To study this we model the rounding errors by 100 independent $U(-0.5, 0.5)$ random variables $X_1, X_2, \ldots, X_{100}$.

a. Compute the expectation and the variance of the X_i.

b. Use Chebyshev's inequality to compute an upper bound for the probability $\mathrm{P}(|X_1 + X_2 + \cdots + X_{100}| > 10)$ that the cumulative rounding error $X_1 + X_2 + \cdots + X_{100}$ exceeds €10.

13.3 Consider the situation of the previous exercise. A manager wants to know what happens to the mean absolute error $\frac{1}{n}\sum_{i=1}^{n}|X_i|$ as n becomes large. What can you say about this, applying the law of large numbers?

13.4 ⊞ Of the voters in Florida, a proportion p will vote for candidate G, and a proportion $1-p$ will vote for candidate B. In an election poll a number of voters are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for G." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables X_1, X_2,... are independent and have a $Ber(p)$ distribution.

a. Suppose we use $\bar{X}_n$ to predict p. According to Chebyshev's inequality, how large should n be (how many people should be interviewed) such that the probability that $\bar{X}_n$ is within 0.2 of the "true" p is at least 0.9?
Hint: solve this first for $p = 1/2$, and use that $p(1-p) \leq 1/4$ for all $0 \leq p \leq 1$.

b. Answer the same question, but now $\bar{X}_n$ should be within 0.1 of p.

c. Answer the question from part **a**, but now the probability should be at least 0.95.

d. If $p > 1/2$ candidate G wins; if $\bar{X}_n > 1/2$ you predict that G will win. Find an n (as small as you can) such that the probability that you predict correctly is at least 0.9, if in fact $p = 0.6$.

13.5 You are trying to determine the melting point of a new material, of which you have a large number of samples. For each sample that you measure you find a value close to the actual melting point c but corrupted with a measurement error. We model this with random variables:

$$M_i = c + U_i$$

where M_i is the measured value in degree Kelvin, and U_i is the occurring random error. It is known that $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = 3$, for each i, and that we may consider the random variables M_1, M_2, ... independent. According to Chebyshev's inequality, how many samples do you need to measure to be 90% sure that the average of the measurements is within half a degree of c?

13.6 ⊡ The casino La bella Fortuna is for sale and you think you might want to buy it, but you want to know how much money you are going to make. All the present owner can tell you is that the roulette game Red or Black is played about 1000 times a night, 365 days a year. Each time it is played you have probability 19/37 of winning the player's bet of €1 and probability 18/37 of having to pay the player €1.

Explain in detail why the law of large numbers can be used to determine the income of the casino, and determine how much it is.

13.7 Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with distributions function F. Define F_n as follows: for any a

$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}.$$

Consider a fixed and introduce the appropriate indicator random variables (as in Section 13.4). Compute their expectation and variance and show that the law of large numbers tells us that

$$\lim_{n\to\infty} \mathrm{P}(|F_n(a) - F(a)| > \varepsilon) = 0.$$

13.8 ⊡ In Section 13.4 we described how the probability density function could be recovered from a sequence X_1, X_2, X_3, ... We consider the $Gam(2,1)$ probability density discussed in the main text and a histogram bar around the point $a = 2$. Then $f(a) = f(2) = 2\mathrm{e}^{-2} = 0.27$ and the estimate for $f(2)$ is $\bar{Y}_n/2h$, where $\bar{Y}_n$ as in (13.3).

a. Express the standard deviation of $\bar{Y}_n/2h$ in terms of n and h.

b. Choose $h = 0.25$. How large should n be (according to Chebyshev's inequality) so that the estimate is within 20% of the "true value", with probability 80%?

13.9 ⊞ Let X_1, X_2, ... be an independent sequence of $U(-1,1)$ random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. It is claimed that for some a and any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|T_n - a| > \varepsilon) = 0.$$

a. Explain how this could be true.

b. Determine a.

13.10 ⊡ Let M_n be the maximum of n independent $U(0,1)$ random variables.

a. Derive the exact expression for $\mathrm{P}(|M_n - 1| > \varepsilon)$.
Hint: see Section 8.4.

b. Show that $\lim_{n\to\infty} \mathrm{P}(|M_n - 1| > \varepsilon) = 0$. Can this be derived from Chebyshev's inequality or the law of large numbers?

13.11 For some $t > 1$, let X be a random variable taking the values 0 and t, with probabilities

$$\mathrm{P}(X = 0) = 1 - \frac{1}{t} \quad \text{and} \quad \mathrm{P}(X = t) = \frac{1}{t}.$$

Then $\mathrm{E}[X] = 1$ and $\mathrm{Var}(X) = t - 1$. Consider the probability $\mathrm{P}(|X - 1| > a)$.

a. Verify the following: if $t = 10$ and $a = 8$ then $\mathrm{P}(|X - 1| > a) = 1/10$ and Chebyshev's inequality gives an upper bound for this probability of $9/64$. The difference is $9/64 - 1/10 \approx 0.04$. We will say that for $t = 10$ the Chebyshev gap for X at $a = 8$ is 0.04.

b. Compute the Chebyshev gap for $t = 10$ at $a = 5$ and at $a = 10$.

c. Can you find a gap smaller than 0.01, smaller than 0.001, smaller than 0.0001?

d. Do you think one could improve Chebyshev's inequality, i.e., find an upper bound closer to the true probabilities?

13.12 (A more general law of large numbers). Let $X_1, X_2, \ldots$ be a sequence of independent random variables, with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$, for $i = 1, 2, \ldots$. Suppose that $0 < \sigma_i^2 \leq M$, for all i. Let a be an arbitrary positive number.

a. Apply Chebyshev's inequality to show that

$$\mathrm{P}\left(\left|\bar{X}_n - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > a\right) \leq \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2 a^2}.$$

b. Conclude from **a** that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\bar{X}_n - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > a\right) = 0.$$

Check that the law of large numbers is a special case of this result.