9

Estimation Of Population Parameters

Overview

Estimation of parameters describing large populations is a fundamental challenge in many scientific disciplines. Section 9.1 describes random sampling, a simple yet powerful approach that enables us to obtain very accurate estimates from limited data. Sections 9.2 and 9.3 introduce the bias and standard error, which measure the average error and the standard deviation of an estimator. In Section 9.4 we describe deviation bounds, which characterize the probabilistic behavior of a random variable just based on its mean and variance. In Section 9.5, we prove the celebrated law of large numbers, which states that averages of independent samples converge as the number of samples tends to infinity, and use it to derive theoretical guarantees for estimators based on averaging. Section 9.6 provides a word of caution, describing several situations in which the law of large numbers does not hold. Section 9.7 discusses another celebrated result: the central limit theorem, according to which averages of independent quantities tend to have Gaussian distributions. In Section 9.8, we define confidence intervals, which quantify our uncertainty about parameter estimates, and show how to construct them using the central limit theorem. Finally, in Section 9.9 we introduce the bootstrap, a computational technique to estimate standard errors and build confidence intervals.

9.1 Random Sampling

Imagine that you want to estimate the average weight of the pigeons in New York. In theory, you could compute the average exactly by catching and weighing every pigeon. However, there are more than one million pigeons in the city,* so this would be impossible in practice. Statistics that describe entire populations, such as the mean weight of New York pigeons, are called *population parameters*. Estimating population parameters is a fundamental problem in many applications, such as economics, healthcare, sociology, and political science.

As in our pigeon example, measuring a whole population is often intractable or too costly. Consequently, population parameters are usually estimated from a subset or *sample* of the population: we catch a few pigeons and average their weights to estimate the mean. A crucial question is how to choose what individ-

*According to a quick Internet search. There seems to be considerable debate about the exact number.

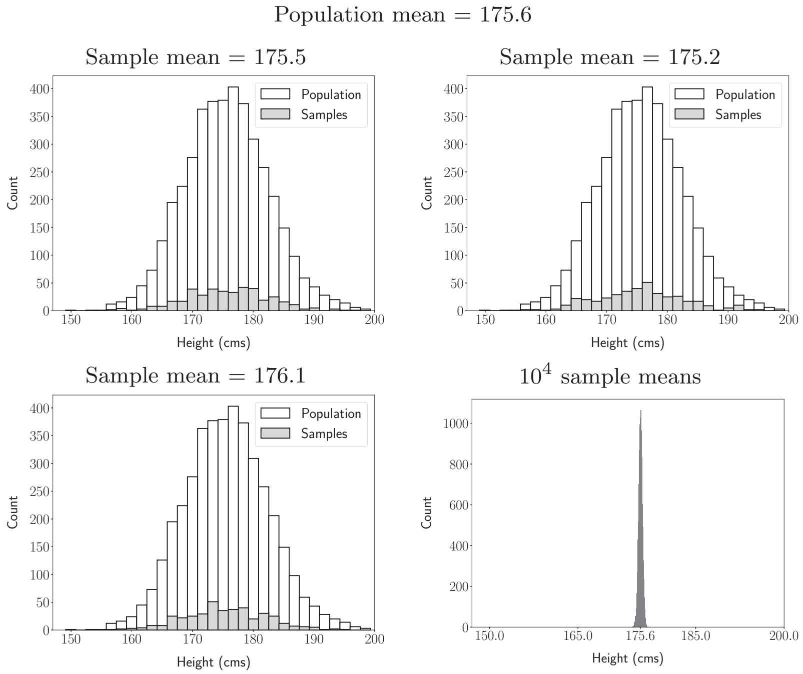


Figure 9.1 Estimating the mean height in a population via random sampling. In the top row and the bottom left, the white histogram represents the heights of a population of 4,082 individuals, extracted from Dataset 5 (see Figures 3.6 and 3.13). The gray histograms represent the values of 400 random samples measured uniformly at random with replacement from the population. The sample mean of the random samples is indicated above each graph. The plot on the bottom right shows a histogram of the sample means of 10,000 random subsets of size 400 obtained in the same way. The histogram concentrates tightly around the population mean, which equals 175.6 cm.

uals to measure. A simple strategy is to just pick random individuals from the population. As illustrated by the following two examples, this strategy, known as *random sampling*, often provides very accurate estimates from surprisingly small amounts of data.

Example 9.1 (Estimating the mean height)**.** We consider the problem of estimating the mean height of a population in a controlled scenario, where we know the true population mean. We use the data in Figures 3.6 and 3.13, extracted from Dataset 5, as the complete ground-truth population. The population mean equals

$$\begin{aligned}\mu_{\text{pop}} &:= \frac{1}{N}\sum_{k=1}^{N} h_i && (9.1)\\ &= 175.6, && (9.2)\end{aligned}$$

where h_1, h_2, …, h_N denote the heights of the $N := 4{,}082$ individuals in the population. To perform random sampling, we select $n := 400$ data points x_1, …, x_n independently and uniformly at random with replacement. In more detail, x_j for $1 \leq j \leq n$ is obtained by picking a random index k from $\{1, 2, \ldots, N\}$ and setting $x_j := h_k$. The population mean is then estimated by computing the sample mean of the samples x_1, …, x_n,

$$m := \frac{1}{n} \sum_{j=1}^{n} x_j. \tag{9.3}$$

The value of the sample mean changes depending on the selected samples. In the three examples shown in Figure 9.1, the sample means are very close to the population mean. The figure also includes a histogram of sample means computed by repeating the sampling process 10,000 times. The histogram is tightly concentrated around the population mean, indicating that the estimate is very likely to be accurate.

. . . .

Example 9.2 (Prevalence of COVID-19)**.** During the COVID-19 pandemic, determining the prevalence of the disease was a key challenge. In this example, we consider the problem of estimating the prevalence of COVID-19 in New York in a hypothetical scenario where 5% of New Yorkers have the virus. Our strategy is based on random sampling: we test 1,000 individuals chosen independently and uniformly at random with replacement from the entire population of 8.8 million people. Our estimate of the prevalence is simply the proportion of positive tests. For the sake of simplicity, we assume that the tests are perfect; if a person has the disease, they test positive. Figure 9.2 shows the result of simulating this procedure multiple times. The prevalence estimates are extremely accurate.

. . . .

In the remainder of this chapter, we study the estimates of population parameters obtained via random sampling. Our analysis has to be probabilistic because the measurements are random: repeating them would yield different data, and therefore a different estimate of the population parameter. To account for this, we model the measurements and the corresponding estimates as random variables. This enables us to encode our assumptions about the sampling process more formally.

Definition 9.3 (Probabilistic model of random samples)**.** *Let a_1, a_2, …, a_N denote the values of a quantity of interest in a population of size N. The indices $\tilde{k}_1$, $\tilde{k}_2$, …, $\tilde{k}_n$ are said to be drawn independently and uniformly at random with replacement, if they are independent random variables such that*

$$\mathrm{P}\left(\tilde{k}_j = i\right) = \frac{1}{N}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n. \tag{9.4}$$

The corresponding dataset of n random samples $\tilde{x}_1$, $\tilde{x}_2$, …, $\tilde{x}_n$, where

$$\tilde{x}_j = a_{\tilde{k}_j}, \qquad 1 \leq j \leq n, \tag{9.5}$$

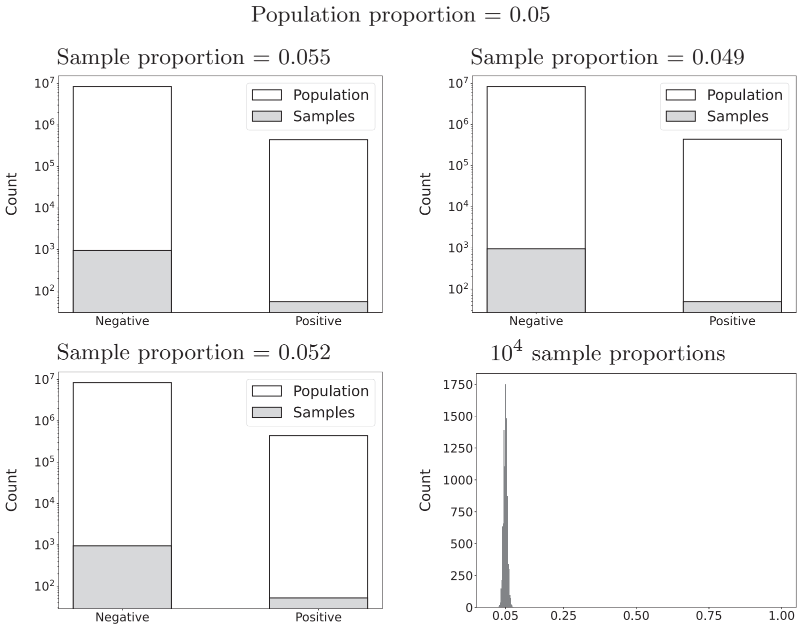


Figure 9.2 Estimating the prevalence of COVID-19 via random sampling. In the top row and the bottom left, the white bar plot indicates the number of people in New York that have COVID-19 (positive) and those that do not (negative) for a simulated scenario where the COVID-19 prevalence is 5%. The gray bar plots represent the number of positive and negative tests obtained from 1,000 random samples measured uniformly at random with replacement out of the total population of 8.8 million people. The sample proportion of positive tests in the random samples is indicated above each graph. The plot on the bottom right shows a histogram of the sample proportions of 10,000 random subsets of size 1,000 obtained in the same way. The histogram concentrates tightly around the true prevalence.

is then said to be obtained independently and uniformly at random with replacement.

In previous chapters, we define estimators of different quantities of interest, such as the sample mean (Definition 7.14) or the sample variance (Definition 7.36), as deterministic functions of the available data. Here we interpret the data as random variables, following Definition 9.3, so the corresponding estimators are also random variables. For instance, if we apply Definition 9.3 to Example 9.1 by setting $a_i := h_i$, $1 \leq i \leq N$, then the sample mean is the random variable

$$\widetilde{m} := \frac{1}{n} \sum_{j=1}^{n} \tilde{x}_j, \tag{9.6}$$

where $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ represent the n height measurements.

Example 9.4 (Prevalence of COVID-19: Probabilistic analysis)**.** In Example 9.2, we can encode the status of the ith individual, $1 \leq i \leq N$, in the notation of Definition 9.3 by setting $a_i := 1$ if they have COVID-19, and $a_i := 0$ if they don't. The data obtained via independent, uniform random sampling with replacement is represented by n random variables $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$, which are equal to one if the corresponding individual has COVID-19 and zero otherwise. The fraction of positive tests among the n selected individuals is equal to the sample mean of these Bernoulli random variables,

$$\widetilde{m} := \frac{1}{n} \sum_{j=1}^{n} \tilde{x}_j. \tag{9.7}$$

. . . .

In the following sections, we study the probabilistic behavior of estimators of population parameters, with a particular focus on the sample mean. In our analysis, we interpret the parameters of interest such as the population mean μ_{pop} in Example 9.1 as fixed deterministic quantities. In statistics, this is known as a *frequentist* viewpoint, in contrast to Bayesian approaches, in which the parameters are also modeled as random variables, as described in Section 6.7. We compare the two perspectives in Example 9.47.

9.2 The Bias

The bias of an estimator is its mean error. If it is zero, then the estimator is said to be *unbiased*. The distribution of an unbiased estimator is centered at the population parameter.

Definition 9.5 (Bias)**.** *Let $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ denote random variables representing measurements associated to a population parameter $\gamma \in \mathbb{R}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ denote an estimator designed to approximate γ from these measurements. The bias of the estimator is the mean of the difference between the random variable $h(\tilde{x}_1, \ldots, \tilde{x}_n)$ and γ. If the mean is equal to γ,*

$$\mathrm{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] = \gamma, \tag{9.8}$$

so that the bias is zero, then the estimator is unbiased.

The following theorem establishes that the sample mean is an unbiased estimator of the population mean, when it is computed using random samples. This explains why the histogram of sample means is centered exactly at the population mean in the bottom right graph of Figure 9.1.

Theorem 9.6 (The sample mean is unbiased)**.** *Let a_1, a_2, ..., a_N denote a dataset of size N with population mean*

$$\mu_{\text{pop}} := \frac{1}{N} \sum_{i=1}^{N} a_i \tag{9.9}$$

and let $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ be independent, uniform random samples following Definition 9.3. The sample mean

$$\widetilde{m} := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j \tag{9.10}$$

is an unbiased estimator of μ_{pop},

$$\operatorname{E}[\widetilde{m}] = \mu_{\text{pop}}. \tag{9.11}$$

Proof By Definition 9.3, the mean of each individual sample is equal to the population mean

$$\operatorname{E}[\tilde{x}_j] = \sum_{k=1}^{N} a_k p_{\tilde{k}_j}(k) \tag{9.12}$$

$$= \frac{1}{N}\sum_{k=1}^{N} a_k \tag{9.13}$$

$$= \mu_{\text{pop}}. \tag{9.14}$$

By linearity of expectation, this implies

$$\operatorname{E}[\widetilde{m}] = \operatorname{E}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j\right] \tag{9.15}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \operatorname{E}[\tilde{x}_j] \tag{9.16}$$

$$= \mu_{\text{pop}}. \tag{9.17}$$

■

Example 9.7 (Prevalence of COVID-19: Bias)**.** As explained in Example 9.4, we can model the data in Example 9.2 as n Bernoulli random variables $\tilde{x}_1$, ..., $\tilde{x}_n$. Our prevalence estimator is the sample proportion of positive tests, which equals the sample mean of these random variables. In this case, the population mean is equal to the ground-truth population prevalence θ_{pop},

$$\frac{1}{N}\sum_{i=1}^{N} a_i = \frac{\text{Number of COVID-19 cases}}{N} := \theta_{\text{pop}}. \tag{9.18}$$

By Theorem 9.6 the mean of the sample proportion is therefore equal to θ_{pop}. This is consistent with the bottom right graph of Figure 9.2, where the histogram of sample proportions is centered at the ground-truth prevalence.

....

In Definition 7.36 we define the sample variance as an average of n squared deviations from the sample mean, where we divide by $n-1$ instead of n. This is to ensure that the estimator is unbiased, under the assumption of independent, uniform random sampling with replacement.

Theorem 9.8 (The sample variance is unbiased)**.** *Let $a_1, a_2, \ldots, a_N$ denote a dataset of size N with population mean μ_{pop}, defined as in (9.9), and population variance*

$$\sigma_{\text{pop}}^2 := \frac{1}{N} \sum_{i=1}^{N} (a_i - \mu_{\text{pop}})^2, \tag{9.19}$$

and let $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ be independent, uniform random samples following Definition 9.3. The sample variance

$$\tilde{v} := \frac{1}{n-1} \sum_{j=1}^{n} \left(\tilde{x}_j - \widetilde{m}\right)^2, \tag{9.20}$$

where the sample mean $\widetilde{m}$ is defined as in (9.10), is an unbiased estimator of σ_{pop}^2,

$$\mathrm{E}\left[\tilde{v}\right] = \sigma_{\text{pop}}^2. \tag{9.21}$$

Proof We decompose the mean of the sample variance as follows

$$\begin{aligned} \mathrm{E}\left[\tilde{v}\right] &= \mathrm{E}\left[\frac{1}{n-1} \sum_{j=1}^{n} \left(\tilde{x}_j - \widetilde{m}\right)^2\right] && (9.22) \\ &= \frac{1}{n-1} \left(\sum_{j=1}^{n} \mathrm{E}\left[\tilde{x}_j^2\right] - 2 \sum_{j=1}^{n} \mathrm{E}\left[\widetilde{m}\tilde{x}_j\right] + \sum_{j=1}^{n} \mathrm{E}\left[\widetilde{m}^2\right]\right). && (9.23) \end{aligned}$$

By linearity of expectation,

$$\sum_{j=1}^{n} \mathrm{E}\left[\widetilde{m}^2\right] = n\mathrm{E}\left[\widetilde{m} \frac{1}{n} \sum_{j=1}^{n} \tilde{x}_j\right] = \sum_{j=1}^{n} \mathrm{E}\left[\widetilde{m}\tilde{x}_j\right], \tag{9.24}$$

which in turn equals

$$\begin{aligned} \sum_{j=1}^{n} \mathrm{E}\left[\widetilde{m}\tilde{x}_j\right] = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathrm{E}\left[\tilde{x}_j\tilde{x}_k\right] &= \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}\left[\tilde{x}_j^2\right] + \frac{1}{n} \sum_{j=1}^{n} \sum_{k \neq j} \mathrm{E}\left[\tilde{x}_j\tilde{x}_k\right] && (9.25) \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\mathrm{E}\left[\tilde{x}_j^2\right] + (n-1)\,\mu_{\text{pop}}^2\right), && (9.26) \end{aligned}$$

because $\mathrm{E}\left[\tilde{x}_j\tilde{x}_k\right] = \mathrm{E}\left[\tilde{x}_j\right]\mathrm{E}\left[\tilde{x}_k\right] = \mu_{\text{pop}}^2$ by the independence assumption. Plugging (9.26) and (9.24) into (9.23) yields

$$\begin{aligned}
\mathrm{E}\left[\tilde{v}\right] &= \frac{1}{n-1}\left(\sum_{j=1}^{n}\mathrm{E}\left[\tilde{x}_j^2\right] - \sum_{j=1}^{n}\mathrm{E}\left[\tilde{m}\tilde{x}_j\right]\right) && (9.27)\\
&= \frac{1}{n-1}\left(\sum_{j=1}^{n}\mathrm{E}\left[\tilde{x}_j^2\right] - \frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{E}\left[\tilde{x}_j^2\right] + (n-1)\,\mu_{\text{pop}}^2\right)\right) && (9.28)\\
&= \frac{1}{n-1}\frac{n-1}{n}\sum_{j=1}^{n}\left(\mathrm{E}\left[\tilde{x}_j^2\right] - \mu_{\text{pop}}^2\right) && (9.29)\\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{E}\left[\tilde{x}_j^2\right] - \mu_{\text{pop}}^2\right) && (9.30)\\
&= \frac{1}{n}\sum_{j=1}^{n}\mathrm{Var}\left[\tilde{x}_j\right], && (9.31)
\end{aligned}$$

since $\mathrm{E}\left[\tilde{x}_j\right] = \mu_{\text{pop}}$ by Theorem 9.6. This implies that $\mathrm{E}\left[\tilde{v}\right] = \sigma_{\text{pop}}^2$, because it follows from Definition 9.3 that the variance of each individual sample is equal to the population variance:

$$\begin{aligned}
\mathrm{Var}\left[\tilde{x}_j\right] &:= \mathrm{E}\left[\left(\tilde{x}_j - \mathrm{E}\left[\tilde{x}_j\right]\right)^2\right] && (9.32)\\
&= \mathrm{E}\left[\left(\tilde{x}_j - \mu_{\text{pop}}\right)^2\right] && (9.33)\\
&= \sum_{k=1}^{N}(a_k - \mu_{\text{pop}})^2 p_{\tilde{k}_j}(k) && (9.34)\\
&= \frac{1}{N}\sum_{k=1}^{N}(a_k - \mu_{\text{pop}})^2 && (9.35)\\
&= \sigma_{\text{pop}}^2. && (9.36)
\end{aligned}$$

■

9.3 The Standard Error

As explained in the previous section, the distribution of an unbiased estimator is centered at the population parameter. However, this does not necessarily imply a good approximation: if the distribution is very spread out, the value of the estimator can be far from the parameter. The average variation of the estimator is captured by its standard deviation, which is commonly known as its standard error.

Definition 9.9 (Standard error)**.** *Let $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ denote random variables representing measurements associated to a population parameter $\gamma \in \mathbb{R}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ denote an unbiased estimator of γ. The standard error of h is the* *standard deviation of the random variable* $h(\tilde{x}_1, \ldots, \tilde{x}_n)$*,*

$$\text{se}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] := \sqrt{\text{Var}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]}. \tag{9.37}$$

The following simple lemma justifies calling the standard error an error: it is equal to the root mean square error between the estimator and the parameter of interest, as long as the estimator is unbiased.

Lemma 9.10. *Let* $\tilde{x}_1$*,* $\tilde{x}_2$*, ...,* $\tilde{x}_n$ *denote random variables representing measurements associated to a population parameter* $\gamma \in \mathbb{R}$*, and let* $h : \mathbb{R}^n \to \mathbb{R}$ *denote an unbiased estimator of* γ*. The standard error of the estimator is equal to the root mean square difference between the estimator and* γ*,*

$$\text{se}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] = \sqrt{\text{E}\left[\left(h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma\right)^2\right]}. \tag{9.38}$$

Proof Since the estimator is unbiased $\text{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] = \gamma$,

$$\text{se}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] := \sqrt{\text{Var}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]} \tag{9.39}$$

$$= \sqrt{\text{E}\left[\left(h(\tilde{x}_1, \ldots, \tilde{x}_n) - \text{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]\right)^2\right]} \tag{9.40}$$

$$= \sqrt{\text{E}\left[\left(h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma\right)^2\right]}. \tag{9.41}$$

■

In order to determine the standard error of estimators that involve averaging, we establish a key result: the variance of a sum of independent random variables is equal to the sum of their variance.

Theorem 9.11 (Variance of the sum of independent random variables)**.** *Let* $\tilde{a}_1$*,* $\tilde{a}_2$*, ...,* $\tilde{a}_n$ *be independent random variables with finite variance belonging to the same probability space. The variance of their sum is equal to the sum of their variances,*

$$\text{Var}\left[\sum_{k=1}^{n} \tilde{a}_k\right] = \sum_{k=1}^{n} \text{Var}\left[\tilde{a}_k\right]. \tag{9.42}$$

Proof By the independence assumption, for any $1 \leq t \leq n$, the random variables $\tilde{a}_t$ and $\sum_{k=t+1}^{n} \tilde{a}_k$ are independent, and hence uncorrelated by Lemma 8.28. We can therefore apply Corollary 8.23 $n-1$ times to prove the result:

$$\text{Var}\left[\sum_{k=1}^{n} \tilde{a}_k\right] = \text{Var}\left[\tilde{a}_1\right] + \text{Var}\left[\sum_{k=2}^{n} \tilde{a}_k\right] \tag{9.43}$$

$$= \text{Var}\left[\tilde{a}_1\right] + \text{Var}\left[\tilde{a}_2\right] + \text{Var}\left[\sum_{k=3}^{n} \tilde{a}_k\right] \tag{9.44}$$

$$= \sum_{k=1}^{n} \text{Var}\left[\tilde{a}_k\right]. \tag{9.45}$$

■

A byproduct of this result is an expression for the variance of a binomial random variable.

Lemma 9.12 (Variance of a binomial random variable)**.** *The variance of a binomial random variable $\tilde{a}$ with parameters n and θ equals* $\operatorname{Var}[\tilde{a}] = n\theta(1-\theta)$.

Proof Recall that a binomial random variable can be represented as the sum of n independent Bernoulli random variables $\tilde{b}_1, \ldots, \tilde{b}_n$ (see Example 2.16). By Lemma 7.38 and Theorem 9.11,

$$\begin{aligned}\operatorname{Var}[\tilde{a}] &= \operatorname{Var}\left[\sum_{k=1}^{n} \tilde{b}_k\right] && (9.46)\\ &= \sum_{k=1}^{n} \operatorname{Var}[\tilde{b}_k] && (9.47)\\ &= n\theta(1-\theta). && (9.48)\end{aligned}$$

■

The number of data in a sample is often known as the *sample size*. The following theorem establishes that the standard error of the sample mean decreases proportionally to the square root of the sample size, assuming independent and uniform random sampling with replacement. Interestingly, the standard error does *not* depend at all on the size of the population N; it just depends on the sample size.

Theorem 9.13 (Standard error of the sample mean)**.** *Let $a_1, a_2, \ldots, a_N$ denote a dataset of size N with population mean μ_{pop}, defined as in* (9.9)*, and population variance σ^2_{pop}, defined as in* (9.19)*. Let $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ be independent, uniform random samples following Definition 9.3. The standard error of the sample mean $\widetilde{m} := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j$ equals*

$$\operatorname{se}[\widetilde{m}] = \frac{\sigma_{\text{pop}}}{\sqrt{n}}. \qquad (9.49)$$

Proof By (9.36) the variance of each sample, $\operatorname{Var}[\tilde{x}_j]$ for $1 \leq j \leq n$, equals σ^2_{pop}. The result then follows from Lemma 7.35 and Theorem 9.11,

$$\begin{aligned}\operatorname{se}[\widetilde{m}]^2 = \operatorname{Var}[\widetilde{m}] &= \operatorname{Var}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j\right] && (9.50)\\ &= \frac{1}{n^2}\operatorname{Var}\left[\sum_{j=1}^{n} \tilde{x}_j\right] && (9.51)\\ &= \frac{1}{n^2}\sum_{j=1}^{n} \operatorname{Var}[\tilde{x}_j] && (9.52)\\ &= \frac{\sigma^2_{\text{pop}}}{n}. && (9.53)\end{aligned}$$

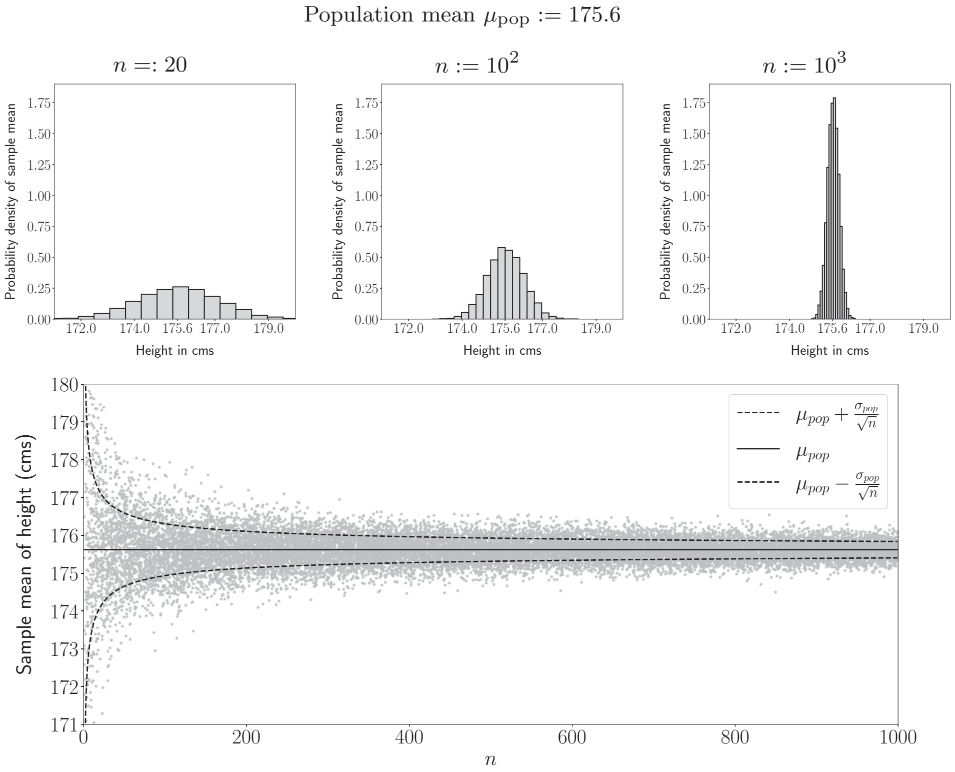


Figure 9.3 Bias and standard error of the sample mean. Analysis of the sample mean estimator applied to independent, uniform samples from the dataset in Example 9.1. The top row shows normalized histograms of the sample mean of n height measurements for different values of n. Each histogram is computed using 10^4 independent instances of the sample mean. The histograms are centered at the true population mean $\mu_{\text{pop}} = 175.6$ cm, indicating that the estimates are unbiased, in accordance with Theorem 9.6. The scatterplot below shows that as n increases, the standard error decreases proportionally to $\sigma_{\text{pop}}/\sqrt{n}$, where $\sigma_{\text{pop}} = 6.85$ cm is the population standard deviation, as established in Theorem 9.13.

■

Figure 9.3 shows the behavior of the sample mean estimator in Example 9.1 as a function of the sample size n. As predicted by Theorem 9.13, the sample means are increasingly concentrated around the population mean, and their average spread decreases proportionally to $\sqrt{n}$.

Example 9.14 (Prevalence of COVID-19: Standard error)**.** In this example we analyze the standard error of the prevalence estimator in Example 9.2, based on the data model described in Example 9.4. In Example 9.7 we show that the pop-

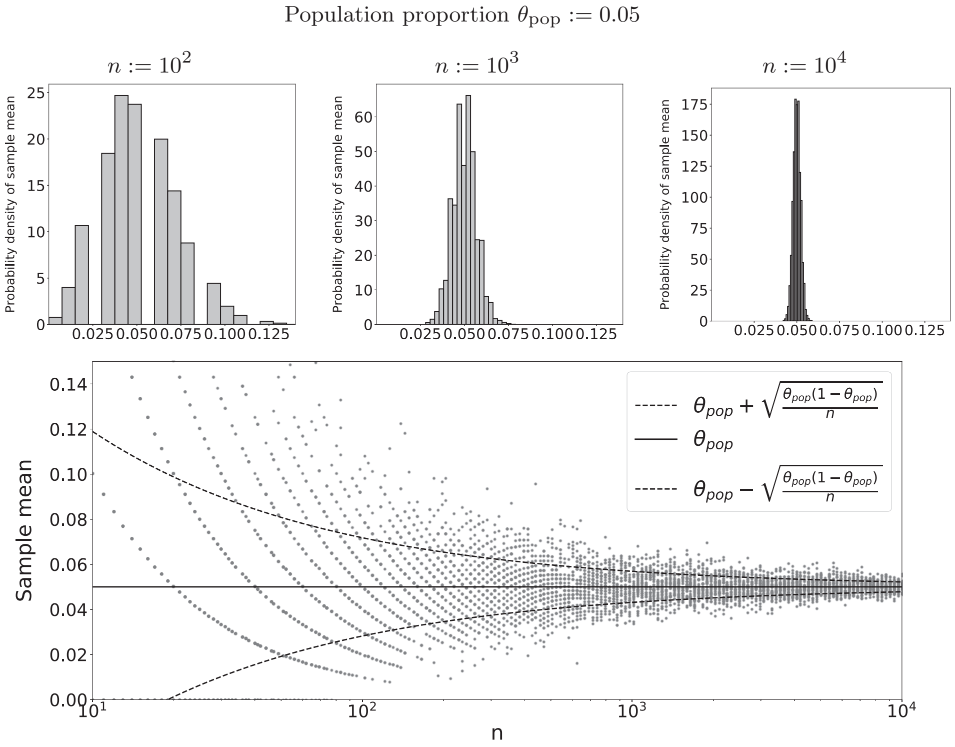


Figure 9.4 Bias and standard error of the sample proportion. Analysis of proportion estimates obtained from independent, uniform samples in the scenario described in Example 9.2. The top row shows normalized histograms of the sample proportion of positives out of n COVID-19 tests, for different values of n. Each histogram is computed using 10^4 independent instances of the sample mean. The histograms are centered at the true proportion of COVID-19 cases $\theta_{\text{pop}} = 0.05$ indicating that the estimates are unbiased, in accordance with Theorem 9.6. The scatterplot below shows that as n increases, the standard error decreases proportionally to $\sqrt{\theta_{\text{pop}}(1-\theta_{\text{pop}})}/\sqrt{n}$, as derived in Example 9.14.

ulation mean is equal to the population proportion θ_{pop}. The population variance therefore equals

$$\sigma_{\text{pop}}^2 := \frac{1}{N}\sum_{k=1}^{N}(a_k - \theta_{\text{pop}})^2 \tag{9.54}$$

$$= \frac{1}{N}\sum_{k=1}^{N} a_k^2 - \frac{2\theta_{\text{pop}}}{N}\sum_{k=1}^{N} a_k + \frac{1}{N}\sum_{k=1}^{N}\theta_{\text{pop}}^2 \tag{9.55}$$

$$= \theta_{\text{pop}} - 2\theta_{\text{pop}}^2 + \theta_{\text{pop}}^2 \tag{9.56}$$

$$= \theta_{\text{pop}}(1 - \theta_{\text{pop}}). \tag{9.57}$$

Since the sample proportion is equal to the sample mean, by Theorem 9.13 its standard error equals

$$\frac{\sigma_{\text{pop}}}{\sqrt{n}} = \sqrt{\frac{\theta_{\text{pop}}(1 - \theta_{\text{pop}})}{n}}. \tag{9.58}$$

Figure 9.4 shows that this formula indeed captures the right scaling of the estimation error for different values of n.

....

An important consequence of Theorem 9.13 is that the mean squared error of the sample mean (with respect to the population mean) tends to zero asymptotically as the sample size tends to infinity. In probability theory, this is known as convergence in mean square.

Corollary 9.15 (Convergence of the sample mean in mean square)**.** *Let a_1, a_2, ..., a_N denote a dataset of size N with population mean μ_{pop}, defined as in (9.9). Let $\tilde{x}_1$, $\tilde{x}_2$, ..., be a sequence of independent, uniform random samples following Definition 9.3. The mean squared error between μ_{pop} and the sample mean $\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n}\tilde{x}_j$ converges to zero as the number of measurements tends to infinity,*

$$\lim_{n\to\infty} \mathrm{E}\left[(\widetilde{m}_n - \mu_{\text{pop}})^2\right] = 0. \tag{9.59}$$

Proof By Theorem 9.6 the mean of $\widetilde{m}_n$ is μ_{pop}, so Theorem 9.13 implies

$$\mathrm{E}\left[(\widetilde{m}_n - \mu_{\text{pop}})^2\right] = \mathrm{Var}\left[\widetilde{m}_n\right] \tag{9.60}$$

$$= \frac{\sigma_{\text{pop}}^2}{n}, \tag{9.61}$$

which tends to zero as $n \to \infty$. ■

9.4 Deviation Bounds: Markov's And Chebyshev's Inequalities

In this section we take a small detour to derive deviation bounds, which allow us to make statements about the probabilistic behavior of a random variable just based on its mean and variance. These bounds are a fundamental tool in the theoretical analysis of statistical estimators. In Section 9.5 we use them to characterize the asymptotic behavior of the sample mean and prove the law of large numbers.

Markov's inequality states that if a random variable is nonnegative and small, then it cannot take large values with high probability.

Theorem 9.16 (Markov's inequality)**.** *Let $\tilde{a}$ be a nonnegative random variable with a pmf or pdf that is zero for negative values. For any positive constant $c > 0$,*

$$\mathrm{P}\left(\tilde{a} \geq c\right) \leq \frac{\mathrm{E}\left[\tilde{a}\right]}{c}. \tag{9.62}$$

Proof We prove the result assuming $\tilde{a}$ is continuous and has a pdf. The proof for discrete variables is the same replacing integrals by sums and the pdf by the pmf. Since the pdf is zero for negative values of a,

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \int_{a \in \mathbb{R}} a f_{\tilde{a}}(a) \, \mathrm{d}a & (9.63)\\
&= \int_{a=0}^{c} a f_{\tilde{a}}(a) \, \mathrm{d}a + \int_{a=c}^{\infty} a f_{\tilde{a}}(a) \, \mathrm{d}a & (9.64)\\
&\geq \int_{a=0}^{c} a f_{\tilde{a}}(a) \, \mathrm{d}a + c \int_{a=c}^{\infty} f_{\tilde{a}}(a) \, \mathrm{d}a & (9.65)\\
&\geq c \, \mathrm{P}\left(\tilde{a} \geq c\right). & (9.66)
\end{aligned}$$

■

Example 9.17 (Age of students)**.** The mean age of students at a university is 20 years. We decide to bound the fraction of students above 30 based on this information. Modeling age as a nonnegative random variable $\tilde{a}$, by Markov's inequality,

$$\mathrm{P}(\tilde{a} \geq 30) \leq \frac{\mathrm{E}\left[\tilde{a}\right]}{30} = \frac{2}{3}. \tag{9.67}$$

At most two thirds of the students are over 30.

....

The variance of a random variable is defined in Section 7.7 as the mean squared deviation from the mean. Consequently, when the variance is small, it seems plausible that the random variable cannot be far from its mean with high probability. This is indeed the case. The corresponding deviation bound is known as Chebyshev's inequality.

Theorem 9.18 (Chebyshev's inequality)**.** *For any positive constant $c > 0$ and any random variable $\tilde{a}$ with mean μ and bounded variance,*

$$\mathrm{P}\left(|\tilde{a} - \mu| \geq c\right) \leq \frac{\mathrm{Var}\left[\tilde{a}\right]}{c^2}. \tag{9.68}$$

Proof We apply Markov's inequality to the random variable $\tilde{b} := (\tilde{a} - \mu)^2$, which yields

$$\mathrm{P}\left(|\tilde{a}-\mu| \geq c\right) = \mathrm{P}\left((\tilde{a}-\mu)^2 \geq c^2\right) \tag{9.69}$$

$$\leq \frac{\mathrm{E}\left[(\tilde{a}-\mu)^2\right]}{c^2} \tag{9.70}$$

$$= \frac{\mathrm{Var}\left[\tilde{a}\right]}{c^2}. \tag{9.71}$$

■

A corollary to Chebyshev's inequality is that if the variance of a random variable is zero, then the random variable is constant (the probability that it deviates from its mean is zero). This result is key in our geometric interpretation of the covariance as an inner product in Section 8.7 (see Lemma 8.34). To prove it, we first need to introduce the union bound, also known as Boole's inequality, which states that the probability of a union of events is always smaller than the sum of the individual probabilities.

Theorem 9.19 (Union bound)**.** *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space, and let A_1, A_2, $\ldots A_k$ be k events in $\mathcal{C}$. Then,*

$$\mathrm{P}\left(\cup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathrm{P}\left(A_i\right). \tag{9.72}$$

The result also holds for a countably infinite union of events, when $k \to \infty$.

Proof We define the events B_1, B_2, $\ldots B_k$, setting $B_1 := A_1$ and

$$B_i := A_i \cap \left(\cup_{j=1}^{i-1} A_j^c\right) \tag{9.73}$$

for $2 \leq i \leq k$. In words, B_i is the part of A_i that does not already belong to any A_j for $j < i$. Notice that these events are all disjoint, B_i is a subset of A_i and $\cup_{i=1}^{k} B_i = \cup_{i=1}^{k} A_i$. Consequently, by Axiom 3 in Definition 1.9 and Lemma 1.12,

$$\mathrm{P}\left(\cup_{i=1}^{k} A_i\right) = \mathrm{P}\left(\cup_{i=1}^{k} B_i\right) \tag{9.74}$$

$$= \sum_{i=1}^{k} \mathrm{P}\left(B_i\right) \tag{9.75}$$

$$\leq \sum_{i=1}^{k} \mathrm{P}\left(A_i\right). \tag{9.76}$$

The result still holds when $k \to \infty$ because Axiom 3 in Definition 1.9 holds for countably infinite sequences of disjoint events. ■

Corollary 9.20. *If the variance of a random variable $\tilde{a}$ is zero,* $\mathrm{Var}\left[\tilde{a}\right] = 0$*, then $\tilde{a}$ is equal to its mean μ with probability one,* $\mathrm{P}\left(\tilde{a} = \mu\right) = 1$*. If the mean square is zero, then $\tilde{a}$ equals zero with probability one.*

Proof We express the event $\tilde{a} \neq \mu$ as the union of the events $|\tilde{a} - \mu| > 1/i$ for $i = 1, 2, \ldots$ If the variance is zero, by Chebyshev's inequality each of these events has probability zero:

$$\mathrm{P}\left(|\tilde{a} - \mu| \geq \frac{1}{i}\right) \leq i^2 \mathrm{Var}\left[\tilde{a}\right] = 0. \tag{9.77}$$

By the union bound in Theorem 9.19, this implies that the union also has probability zero,

$$\mathrm{P}\left(|\tilde{a} - \mu| \neq 0\right) = \mathrm{P}\left(\cup_{i=1}^{\infty}\left\{|\tilde{a} - \mu| > \frac{1}{i}\right\}\right) \tag{9.78}$$

$$\leq \sum_{i=1}^{\infty} \mathrm{P}\left(|\tilde{a} - \mu| \geq \frac{1}{i}\right) \tag{9.79}$$

$$= 0, \tag{9.80}$$

so the probability that $\tilde{a}$ equals μ is one.

By Lemma 7.33 the mean square is equal to the sum of the variance and the squared mean. Since these are nonnegative quantities, if their sum is zero, each of them must also equal zero. Consequently, if a random variable has zero mean square, its variance is zero. It is therefore equal to its mean, which is zero, with probability one. ■

Example 9.21 (Age of students: Improved bound)**.** We are not very satisfied with our bound on the number of students older than 30 years in Example 9.17. Further research reveals that the standard deviation of student age is 3 years. This allows us to obtain a tighter bound using Chebyshev's inequality:

$$\mathrm{P}(\tilde{a} \geq 30) \leq \mathrm{P}\left(|\tilde{a} - \mathrm{E}\left[\tilde{a}\right]| \geq 10\right) \tag{9.81}$$

$$\leq \frac{\mathrm{Var}\left[\tilde{a}\right]}{100} \tag{9.82}$$

$$= \frac{9}{100}. \tag{9.83}$$

At least 91% of the students are under 30 years (and above 10).

9.5 The Law Of Large Numbers

In Section 7.1 we define the mean of a random variable as an averaging operation that can be applied to random variables. This motivates our definition of the sample mean in Section 7.1.4, which estimates the mean of a random variable by averaging its samples. In this section, we show that this estimator approximates the mean of a random variable with arbitrary accuracy as long as we use enough samples, and these samples are independent. To make this statement precise, we introduce the concept of convergence in probability.

Let us consider a sequence of i.i.d. random variables $\tilde{x}_1$, $\tilde{x}_2$, $\ldots$ with mean μ

belonging to the same probability space. We denote the running average of the first n variables in the sequence by

$$\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j. \tag{9.84}$$

Consider the probability that $\widetilde{m}_n$ is at a distance of more than ϵ from μ, where ϵ is a small positive constant:

$$p_n := \mathrm{P}\left(|\widetilde{m}_n - \mu| > \epsilon\right). \tag{9.85}$$

The sequence p_1, p_2, p_3, ... is a deterministic sequence of real numbers. If this sequence converges to zero for any ϵ, we say that $\widetilde{m}_n$ converges to μ *in probability.* This means that no matter how small ϵ is, we can increase n so that $\widetilde{m}_n$ is ϵ-close to μ with arbitrarily high probability. The celebrated law of large numbers states that this is guaranteed to occur, as long as the i.i.d. random variables have finite variance.

Theorem 9.22 (The law of large numbers)**.** *Let $\tilde{x}_1$, $\tilde{x}_2$, ... be a countably infinite sequence of i.i.d. random variables with mean μ and finite variance belonging to the same probability space. The running average or sample mean $\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j$ converges in probability to μ as $n \to \infty$, in the sense that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\left(|\widetilde{m}_n - \mu| > \epsilon\right) = 0. \tag{9.86}$$

Proof By linearity of expectation, the mean of the sample mean is μ,

$$\begin{aligned}
\mathrm{E}\left[\widetilde{m}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j\right] && (9.87)\\
&= \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left[\tilde{x}_j\right] && (9.88)\\
&= \mu. && (9.89)
\end{aligned}$$

By Lemmas 7.35 and Theorem 9.11, the variance equals

$$\begin{aligned}
\mathrm{Var}\left[\widetilde{m}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j\right] && (9.90)\\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{j=1}^{n} \tilde{x}_j\right] && (9.91)\\
&= \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{Var}\left[\tilde{x}_j\right] && (9.92)\\
&= \frac{\sigma^2}{n}. && (9.93)
\end{aligned}$$

Consequently, for any $\epsilon > 0$, Chebyshev's inequality implies

$$\mathrm{P}\left(|\widetilde{m}_n - \mu| > \epsilon\right) \leq \frac{\mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{x}_j\right]}{\epsilon^2} \tag{9.94}$$

$$= \frac{\sigma^2}{n\epsilon^2}. \tag{9.95}$$

The limit when $n \to \infty$ is zero because ϵ is fixed (even if it is very small). ■

An immediate consequence of the law of large numbers is that the sample mean converges in probability to the population mean as the sample size increases, if the samples are measured independently and uniformly at random with replacement. Since sample proportions can be interpreted as sample means (see Example 9.4), this implies that they converge to the true population proportion under the same assumptions. In statistics, estimators that converge in probability to the parameter of interest are said to be *consistent*. Figure 9.5 illustrates the consistency of the sample mean. For any ϵ (in the figure, $\epsilon := 0.25$ cms), the probability that the sample mean deviates from the population by more than ϵ eventually converges to zero as the sample size grows.

Theorem 9.23 (Consistency of the sample mean)**.** *Let a_1, a_2, ..., a_N denote a dataset of size N with population mean μ_{pop}, defined as in* (9.9). *Let $\tilde{x}_1$, $\tilde{x}_2$, ..., be a sequence of independent, uniform random samples following Definition 9.3. The sample mean $\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n}\tilde{x}_j$ converges to μ_{pop} in probability as the number of measurements tends to infinity. For any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\left(|\widetilde{m}_n - \mu_{\text{pop}}| > \epsilon\right) = 0. \tag{9.96}$$

Proof The result follows directly from the law of large numbers (Theorem 9.22), because by (9.36) the variance of each random sample equals the population variance, which is finite since N is finite. ■

The law of large numbers enables us to establish the consistency of the empirical-probability estimator in Definition 1.22.

Theorem 9.24 (Consistency of the empirical-probability estimator)**.** *Let $\mathrm{P}(A)$ be the probability of an event A belonging to the collection of events of a probability space with sample space Ω. Consider a dataset $X := \{x_1, x_2, \ldots, x_n\}$ of outcomes in Ω. If the data are generated according to the probability measure P of the probability space, then the probability that each data point is in A equals $\mathrm{P}(A)$. To capture this, we define the Bernoulli random variables*

$$\tilde{b}_i = \begin{cases} 1, & \textit{if the ith data point is in } A, \\ 0, & \textit{otherwise}, \end{cases} \tag{9.97}$$

for $1 \leq i \leq n$, such that $\mathrm{P}(\tilde{b}_i = 1) = \mathrm{P}(A)$ and $\mathrm{P}(\tilde{b}_i = 0) = 1 - \mathrm{P}(A)$. If these random variables are independent, then the empirical-probability estimator

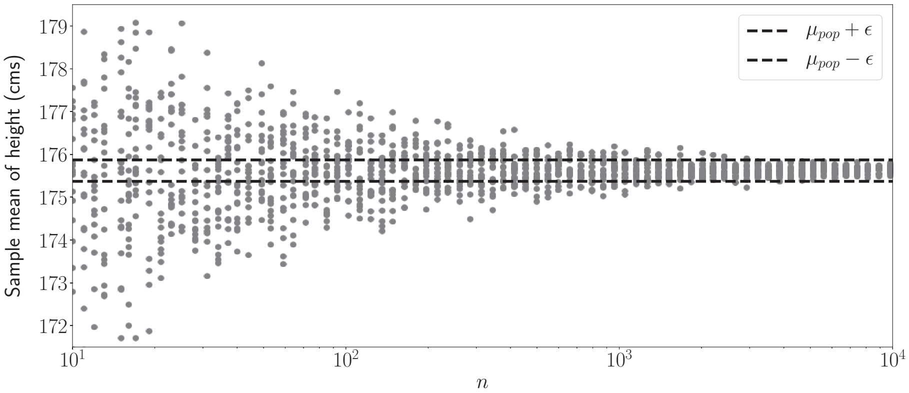


Figure 9.5 Consistency of the sample mean. Each point in the scatterplot corresponds to a sample mean computed using n independent, uniform samples from the dataset in Example 9.1. If the sample mean is in the region between the dashed lines, the difference with the population mean $\mu_{\text{pop}} = 175.6$ cm is less than $\epsilon := 0.25$ cm. As predicted by the law of large numbers, the sample converges to the population mean in probability, so as n grows, the fraction of sample means outside of the region eventually vanishes.

in Definition 1.22

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{b}_i \tag{9.98}$$

converges to $\mathrm{P}(A)$ *in probability as* $n \to \infty$.

Proof The result follows immediately from the law of large numbers, because the Bernoulli random variables have finite variance, and their mean is $\mathrm{P}(A)$. ■

Our proof of the law of large numbers provides a general strategy to establish that an estimator converges to its mean. We just need to prove that it is unbiased and that its variance vanishes as the sample size tend to infinity. Convergence in probability then follows from Chebyshev's bound. As an illustration, we apply this strategy to show that the sample variance is a consistent estimator of the variance.

Theorem 9.25 (The sample variance converges to the variance)**.** *Let* $\tilde{x}_1, \tilde{x}_2, \ldots$ *be a countably infinite sequence of i.i.d. random variables with finite variance* σ^2 *and finite fourth central moment, belonging to the same probability space. The sample variance*

$$\tilde{v}_n := \frac{1}{n-1}\sum_{j=1}^{n}\left(\tilde{x}_j - \widetilde{m}_n\right)^2, \qquad \widetilde{m}_n := \frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k, \tag{9.99}$$

converges in probability to σ^2 *as* $n \to \infty$.

In particular, if $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ are independent, uniform random samples from a dataset $a_1, a_2, \ldots, a_N$ of size N with population variance σ^2_{pop} (defined as in (9.19)), then the sample variance is a consistent estimator of σ^2_{pop}.

Proof We use the following expression for the variance of the sample variance, which is straightforward (but painful) to derive by repeatedly applying linearity of expectation:

$$\operatorname{Var}[\tilde{v}_n] = \frac{1}{n}\left(\kappa - \frac{(n-3)\sigma^4}{n-1}\right), \qquad \kappa := \operatorname{E}\left[(\tilde{x}_i - \mu)^4\right], \tag{9.100}$$

where μ and κ are the mean and fourth central moment of the i.i.d. variables.* The sample variance is an unbiased estimator of the variance, $\operatorname{E}[\tilde{v}] = \sigma^2$ (we omit the derivations which are identical to those in the proof of Theorem 9.8), so by Chebyshev's inequality, for any $\epsilon > 0$

$$\operatorname{P}\left(\left|\tilde{v}_n - \sigma^2\right| > \epsilon\right) \leq \frac{\operatorname{E}\left[(\tilde{v}_n - \sigma^2)^2\right]}{\epsilon^2} \tag{9.101}$$

$$= \frac{\operatorname{Var}[\tilde{v}_n]}{\epsilon^2} \tag{9.102}$$

$$= \frac{1}{n\epsilon^2}\left(\kappa - \frac{(n-3)\sigma^4}{n-1}\right), \tag{9.103}$$

which converges to zero as $n \to \infty$.

If $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ are independent, uniform random samples from a dataset $a_1, a_2, \ldots, a_N$, then they are i.i.d. by Definition 9.3 with mean equal to the population mean μ_{pop} (see (9.14)) and variance equal to the population variance σ^2_{pop} (see (9.36)). The fourth central moment equals $\frac{1}{N}\sum_{k=1}^{N}(a_k - \mu_{\text{pop}})^4$, which is finite, so the result applies.

■

To end this section, we show that the Chebyshev bound used in our proof of the law of the large numbers is usually very loose and consequently does not provide an accurate quantitative description of the behavior of the sample mean for fixed n.

Example 9.26 (Estimation of the mean height: Chebyshev bound)**.** The proof of the law of large numbers (Theorem 9.22) is based on the Chebyshev bound

$$\operatorname{P}\left(|\widetilde{m}_n - \mu| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{9.104}$$

Here we evaluate this bound for the height data in Example 9.1. We fix $\epsilon := 1$ cm and set μ and σ^2 equal to the population mean and population variance, respectively (recall that this is a controlled example, where we know the population parameters). For different values of n, the probability is estimated via the

*In the literature, κ is sometimes also used to denote the kurtosis, which is the fourth central moment divided by the square of the variance.

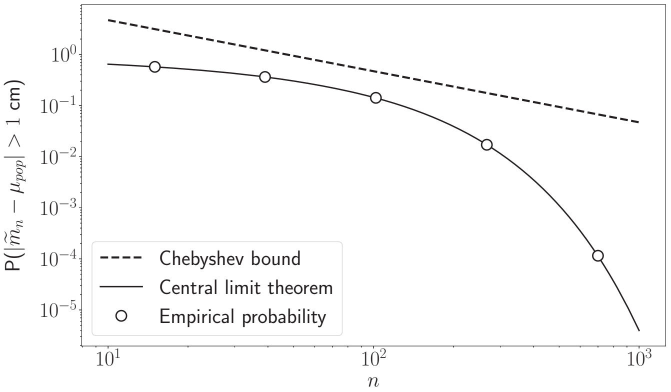


Figure 9.6 Chebyshev bound evaluated on real data. The dashed line depicts the Chebyshev bound $\sigma^2_{\text{pop}}/n\epsilon^2$ in the proof of Theorem 9.22 for $\epsilon := 1$ cm and different values of n, using the height data in Example 9.1. The white markers indicate the empirical probability (computed via 10^6 Monte Carlo simulations) that the sample mean is at a distance of more than 1 cm from the population mean. Both the bound and the empirical probability converge to zero as n increases, but the bound is very loose. The black curve depicts a much more accurate approximation of the probability based on the central limit theorem (see Section 9.7).

Monte Carlo method (see Section 1.7). We repeatedly select n samples independently and uniformly at random with replacement and compute the fraction of times that the difference between the sample mean and the population mean is greater than 1 cm. This yields an empirical probability that we can compare to the bound.

Figure 9.6 shows the results of our analysis. Reassuringly, the Chebyshev bound is indeed greater than the empirical probability, and both converge to zero as n increases. However, the bound is much larger than the empirical probability. The central limit theorem, presented in Section 9.7, provides a much more accurate approximation.

9.6 Some Averages Are Not To Be Trusted

The law of large numbers, described in Section 9.5, states that the average of a sequence of i.i.d. random variables converges to the mean. However, there are situations where the law of large numbers does not hold, as illustrated by the following two examples.

Example 9.27 (St Petersburg paradox)**.** A friend offers to play a game with you. They will flip a fair coin until it lands on tails and then pay you 2^k dollars, where k is the number of coin flips. She asks you how much money you are willing to pay

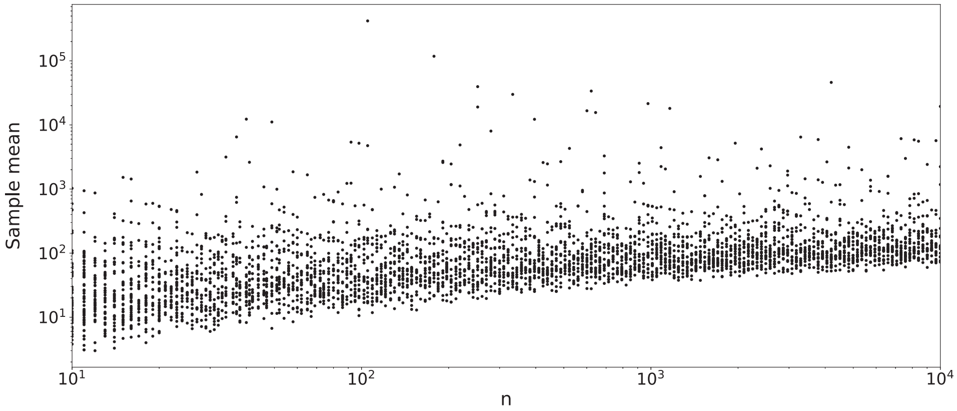


Figure 9.7 Average winnings in the St Petersburg paradox. Scatterplot of the sample mean of the winnings computed from n independent simulations of the game described in Example 9.27. Instead of converging to a finite value, as in Figures 9.3 and 9.4, the sample mean increases indefinitely with n. The reason is that the mean of the underlying distribution is infinite, so the law of large numbers does not hold.

to participate in the game. In order to decide, you simulate the game n times and compute the sample mean of the winnings. To make sure that you have a good estimate, you gradually increase n, expecting the sample mean to converge to a fixed value, once you perform a large enough number of simulations. Surprisingly, this never happens! As shown in Figure 9.7, the sample mean just keeps growing.

You are quite puzzled. Shouldn't the sample mean converge to the mean of the winnings by the law of large numbers? Let us model the winnings as a random variable $\tilde{w} := 2^{\tilde{k}}$, where $\tilde{k}$ represents the number of coin flips. The simulations can be modeled as n i.i.d. random variables with this distribution. If the mean and variance of the distribution are finite, then the sample mean should indeed converge to the mean by Theorem 9.22. As explained in Section 2.3.3, assuming the coin flips are independent, $\tilde{k}$ is a geometric random variable with parameter $1/2$, so its pmf is $p_{\tilde{k}}(k) = 1/2^k$. The mean of $\tilde{w}$ equals

$$\begin{aligned} \mathrm{E}[\tilde{w}] &= \mathrm{E}\left[2^{\tilde{k}}\right] && (9.105)\\ &= \sum_{k=1}^{\infty} 2^k p_{\tilde{k}}(k) && (9.106)\\ &= \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} && (9.107)\\ &= \infty. && (9.108) \end{aligned}$$

The catch is that the mean of the winnings is infinite! Consequently, the law of large numbers is not applicable.

This example is known as the St Petersburg paradox. If our criterion is to maximize the mean of the winnings, we should pay as much money as we can afford to play the game! Of course that makes no sense: the mean is driven to infinity by a vanishingly small fraction $(1/2^k)$ of extremely long sequences, which yield astronomical winnings (2^k). This explains the behavior of the average in your simulations. As the number of simulations grows, there is an increasing chance of observing a very long sequence that single-handedly blows up the average. Apart from these rare events, the winnings are rather modest.

As suggested in Section 7.5, we can use the median instead of the mean to describe the typical winnings. Half of the time, the first coin flip lands on tails, in which case you just earn two dollars. You should probably not invest your life savings in the game.

. . . .

Example 9.28 (Locating a radioactive source)**.** A physicist is carrying out an experiment where a rod of material containing a single radioactive source is placed in front of a line of radiation detectors. The goal is to determine the location of the source on the rod. The source emits particles from time to time. When the particle hits a detector, she records the location at which it hits. She reasons that since the radiation is isotropic– the particle is equally likely to be emitted in any direction– the average of the locations should provide a reasonable estimate for the position of the source on the rod. However, when she computes the sample mean of her data, she is surprised to find that it does not converge to a fixed value, as shown in Figure 9.8. It just fluctuates wildly!

To understand what is going on, we model the angle of the particle trajectory as a random variable $\tilde{u}$ that is uniformly distributed between $-\pi/2$ and $\pi/2$. The sensors are one meter away from the rod, so the location $\tilde{\ell}$ at which the particle is detected is equal to the tangent of $\tilde{u}$. The diagram at the top of Figure 9.8 depicts the probabilistic model of the experiment. The cdf of $\tilde{\ell}$ is equal to

$$\begin{aligned}
F_{\tilde{\ell}}(\ell) &= \mathrm{P}(\tilde{\ell} \leq \ell) && (9.109)\\
&= \mathrm{P}(\tan \tilde{u} \leq \ell) && (9.110)\\
&= \mathrm{P}(\tilde{u} \leq \arctan \ell) && (9.111)\\
&= \frac{1}{\pi} \int_{-\pi/2}^{\arctan \ell} \mathrm{d}u && (9.112)\\
&= \frac{1}{2} + \frac{\arctan \ell}{\pi}, && (9.113)
\end{aligned}$$

where (9.111) holds because the tangent is a monotonic function between $-\pi/2$ and $\pi/2$. Differentiating the cdf yields the pdf of the location,

$$f_{\tilde{\ell}}(\ell) = \frac{1}{\pi(1+\ell^2)}. \qquad (9.114)$$

The experimental data correspond to i.i.d. samples from this distribution. If the mean and variance are finite, then by the law of large numbers, the sample mean of the data should converge to the mean.

The mean of $\tilde{\ell}$ is equal to

$$\mathrm{E}[\tilde{\ell}] = \int_{-\infty}^{\infty} \frac{\ell}{\pi(1+\ell^2)} \, \mathrm{d}\ell \tag{9.115}$$

$$= \int_{0}^{\infty} \frac{\ell}{\pi(1+\ell^2)} \, \mathrm{d}\ell + \int_{-\infty}^{0} \frac{\ell}{\pi(1+\ell^2)} \, \mathrm{d}\ell \tag{9.116}$$

$$= \int_{0}^{\infty} \frac{\ell}{\pi(1+\ell^2)} \, \mathrm{d}\ell - \int_{0}^{\infty} \frac{a}{\pi(1+a^2)} \, \mathrm{d}a, \tag{9.117}$$

by the change of variable $a = -\ell$. We obtain a difference of two identical terms. You may be tempted to cancel out these terms and declare that the mean is zero, but this is incorrect! The reason is that both terms diverge. By the change of variables $t = \ell^2$,

$$\int_{0}^{\infty} \frac{\ell}{\pi(1+\ell^2)} \, \mathrm{d}\ell = \int_{0}^{\infty} \frac{1}{2\pi(1+t)} \mathrm{d}t = \lim_{t\to\infty} \frac{\log(1+t)}{2\pi} = \infty. \tag{9.118}$$

The tail of the pdf $f_{\tilde{\ell}}$ decays fast enough to ensure that it integrates to one, but not fast enough for $f_{\tilde{\ell}}(\ell)\ell$ to have a finite integral. From basic calculus, the difference of two terms that tend to infinity is not well defined, because we can manipulate the expression to yield any value we want.

The random variable $\tilde{\ell}$ therefore does not have a mean, so the law of large numbers does not apply to it. When we sample from its distribution, there is a small but non-negligible probability of observing enormous values, which dominate the sample mean, preventing it from converging to a fixed value. These values, which can be positive or negative, cause the fluctuations observed in Figure 9.8.

....

The distribution of the random variable $\tilde{\ell}$ in Example 9.28 is called a Cauchy distribution.

Definition 9.29 (Cauchy distribution)**.** *A Cauchy random variable $\tilde{a}$ has a pdf of the form*

$$f_{\tilde{a}}(a) = \frac{1}{\pi(1+a^2)}. \tag{9.119}$$

Figure 9.9 compares the pdfs of a Gaussian and a Cauchy distribution. The tail of the Cauchy distribution decays much more slowly, which means that it can take extreme values with much higher probability than the Gaussian. As a result, its mean is not well defined, as we establish in Example 9.28. At first glance, this might seem a mathematical curiosity, but Figure 9.8 shows that it has a very practical implication: a running average of i.i.d. samples from this distribution will never converge.

The following example shows that the sample mean can be unstable in the presence of extreme values, even if the samples are measured from a finite population.

Example 9.30 (Local economic activity)**.** The G-Econ research project at Yale

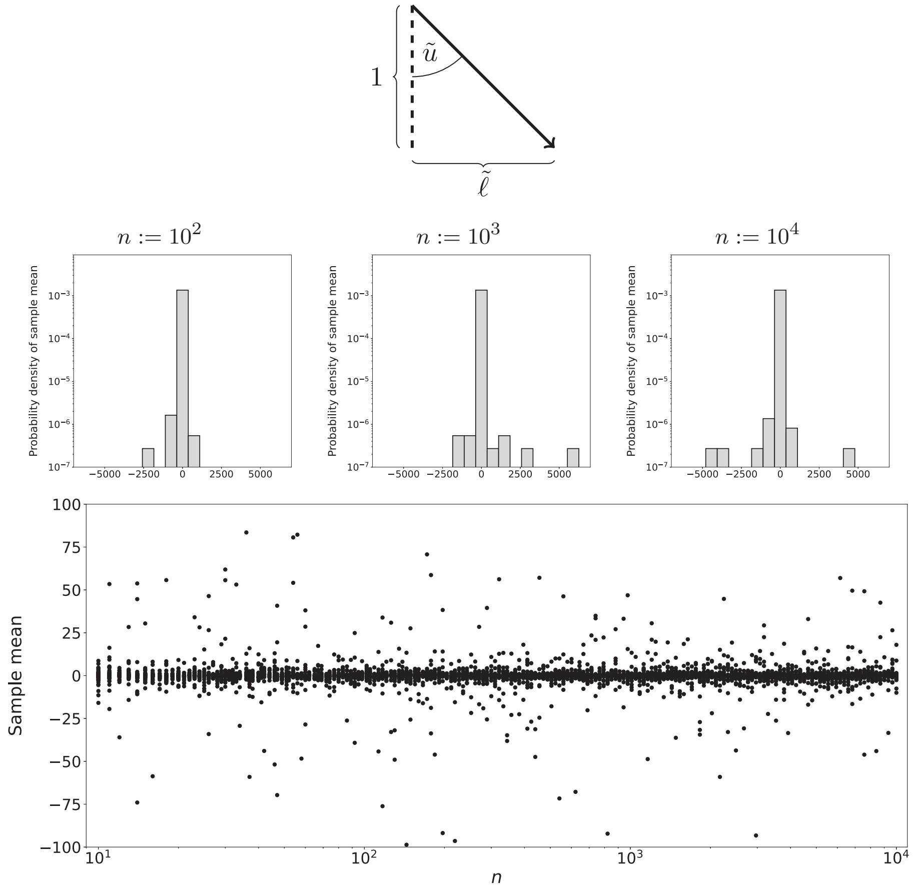


Figure 9.8 Locating a radioactive sample. The diagram at the top shows the probabilistic model for the trajectory of the radioactive particle emitted by the source in Example 9.28. The random variables $\tilde{u}$ and $\tilde{\ell}$ represent the angle of the trajectory and the location of the detector hit by the particle, respectively. The middle row shows normalized histograms of the sample mean of n i.i.d. samples of $\tilde{\ell}$. Each histogram is computed using 10^4 independent instances of the sample mean. The scatterplot below shows that as n increases, the distribution of the sample mean does not concentrate around a fixed value (compare to Figures 9.3 and 9.4), because the probability of encountering values with extremely large magnitudes does not become negligible.

has developed a metric to measure local economic activity all over the world. The gross cell product (GCP) quantifies the economic output of small regions called cells, which partition the globe. In total, there are $N := 20{,}100$ cells. We consider the problem of estimating the mean GCP using the sample mean of 100 cells selected uniformly at random with replacement from Dataset 16. Figure 9.10 shows the results. The value of the sample mean is extremely volatile.

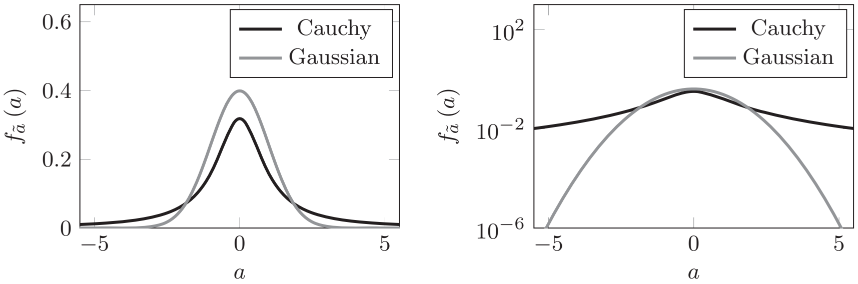


Figure 9.9 Comparison between the Cauchy and Gaussian pdfs. The plots show the pdfs of Cauchy and Gaussian distributions centered at the origin on a linear (left) and logarithmic (right) scale. The tail of the Cauchy distribution decays much more slowly than that of the Gaussian. As a result, extreme values have much higher probability.

The population mean is 2 million dollars, but the sample mean can be as large as 20 million, an order of magnitude larger.

The reason for the instability of the sample mean is that some of the cells have very large values: the median GCP is 0.03 million, but multiple cells have a GCP of more than 200 million! When one of these outliers is selected, this produces an increase in the sample mean greater than the actual value of the population mean (since $200/n = 2$), as can be observed in the bottom left graph of Figure 9.10. Note that this does not contradict the law of large numbers and the consistency of the sample mean. The catch is that the proof of Theorem 9.22 relies on bounding the variance of the i.i.d. random samples, which is very large in this case; the population standard deviation is 17.7 million. Consequently, we require substantially more than 100 samples for the sample mean to concentrate close to the population mean.

The examples in this section show that averaging is not to be trusted in the presence of extreme values. In such situations, it is advisable to not even attempt to estimate the underlying mean in the first place. The median is a more reasonable description of a typical value, as discussed in Section 7.5.

9.7 The Central Limit Theorem

The central limit theorem is a fundamental result in probability, which has crucial implications in statistics. It states that sums of independent quantities tend to have a Gaussian distribution. In Section 9.7.1 we study the distribution of sums of independent random variables. In Section 9.7.2 we present the central limit theorem and use it to characterize the behavior of sample means computed from random samples. Section 9.7.3 provides a cautionary example inspired by the 2008 Financial Crisis, which illustrates situations where the central limit theorem does not hold.

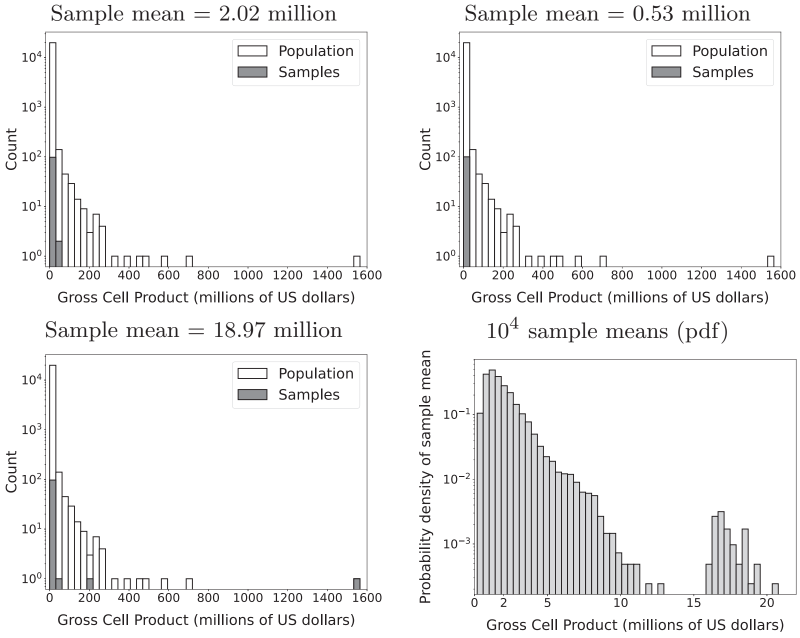


Figure 9.10 Sample mean of data with outliers. In the top row and the bottom left, the white histogram represents the local economic activity of a population of 20,100 locations, quantified by gross cell product (GCP) as described in Example 9.30. The gray histograms represent the values of 100 random samples measured uniformly at random with replacement from the population. The sample mean of the random samples is indicated above each graph (for comparison, the population mean equals 2 million). The plot on the bottom right shows a histogram of 10^4 independent instances of the sample mean obtained in the same way. The histogram is skewed, and does not concentrate tightly around the population mean (compare to Figures 9.1 and 9.2).

9.7.1 Sums Of Independent Random Variables

In this section we derive the distribution of the sum of two independent random variables, showing that it can be obtained applying a convolution to the corresponding pmfs or pdfs. The convolution operation is used in signal processing to represent linear translation-invariant systems. We begin by studying discrete variables.

Theorem 9.31 (Sum of two independent discrete random variables)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two independent discrete random variables with ranges A and B belonging to the same probability space. The pmf of their sum $\tilde{s} = \tilde{a} + \tilde{b}$ equals*

$$p_{\tilde{s}}\left(s\right) = \sum_{a \in A} p_{\tilde{a}}\left(a\right) p_{\tilde{b}}\left(s - a\right), \tag{9.120}$$

where $p_{\tilde{a}}$ and $p_{\tilde{b}}$ denote the pmfs of $\tilde{a}$ and $\tilde{b}$, respectively. Note that $p_{\tilde{s}}$ is nonzero only for values of s such that there exists some $a \in A$ for which $s - b \in B$.

If $\tilde{a}$ and $\tilde{b}$ are integer valued (A and B are subsets of the integers), then the pmf of their sum $\tilde{s} = \tilde{a} + \tilde{b}$ is equal to the convolution of their respective pmfs,

$$p_{\tilde{s}}(s) = p_{\tilde{a}} * p_{\tilde{b}}(s) = \sum_{a=-\infty}^{\infty} p_{\tilde{a}}(a)\, p_{\tilde{b}}(s-a). \tag{9.121}$$

In words, $p_{\tilde{s}}(s)$ is equal to the inner product between $p_{\tilde{a}}$ and a flipped copy of $p_{\tilde{b}}$ shifted by s.

Proof We express the event $\tilde{a} + \tilde{b} = s$ as the union of the intersections between the events $\tilde{a} = a$ and $\tilde{b} = s - a$ for all $a \in A$. The union is over disjoint events, so by the independence assumption,

$$\begin{aligned} p_{\tilde{s}}(s) &= \mathrm{P}(\tilde{a} + \tilde{b} = s) && (9.122)\\ &= \sum_{a \in A} \mathrm{P}(\tilde{a} = a, \tilde{b} = s - a) && (9.123)\\ &= \sum_{a \in A} \mathrm{P}(\tilde{a} = a)\, \mathrm{P}(\tilde{b} = s - a) && (9.124)\\ &= \sum_{a \in A} p_{\tilde{a}}(a)\, p_{\tilde{b}}(s-a). && (9.125) \end{aligned}$$

If A and B are subsets of the integers, then (9.125) can be rewritten as (9.121) because $p_{\tilde{a}}$ and $p_{\tilde{b}}$ are nonzero only on certain integers and zero elsewhere. ■

For random variables with integer values, it follows immediately from Theorem 9.31 that the pmf of the sum of multiple discrete random variables is equal to the convolution of their individual pmfs.

Corollary 9.32 (Sum of multiple independent discrete random variables)**.** *Let $\tilde{a}_1$, $\tilde{a}_2$, …, $\tilde{a}_n$ be n independent discrete random variables with integer values belonging to the same probability space. The pmf of their sum $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ equals the convolution of their pmfs,*

$$p_{\tilde{s}_n}(s) = p_{\tilde{a}_1} * p_{\tilde{a}_2} * \cdots * p_{\tilde{a}_n}(s). \tag{9.126}$$

Example 9.33 (Soccer league)**.** In Example 2.9 we derive the distribution of the points earned by Barcelona in a game against Atletico de Madrid. Here, we consider the problem of modeling the total points earned by Barcelona over several similar games. We assume that for each individual game, the distribution is the same as in Example 2.9 and the games are independent.

Let n be the number of games and let $\tilde{x}_1$, $\tilde{x}_2$, …, $\tilde{x}_n$ denote i.i.d. random variables representing the points obtained in each game. We are interested in the pmf of the sum $\tilde{s}_n := \sum_{i=1}^{n} \tilde{x}_i$ for different values of n. From Example 2.9, the marginal pmf of $\tilde{x}_i$ is

$$p_{\tilde{x}_i}(0) = 0.3, \quad p_{\tilde{x}_i}(1) = 0.3, \quad p_{\tilde{x}_i}(3) = 0.4, \tag{9.127}$$

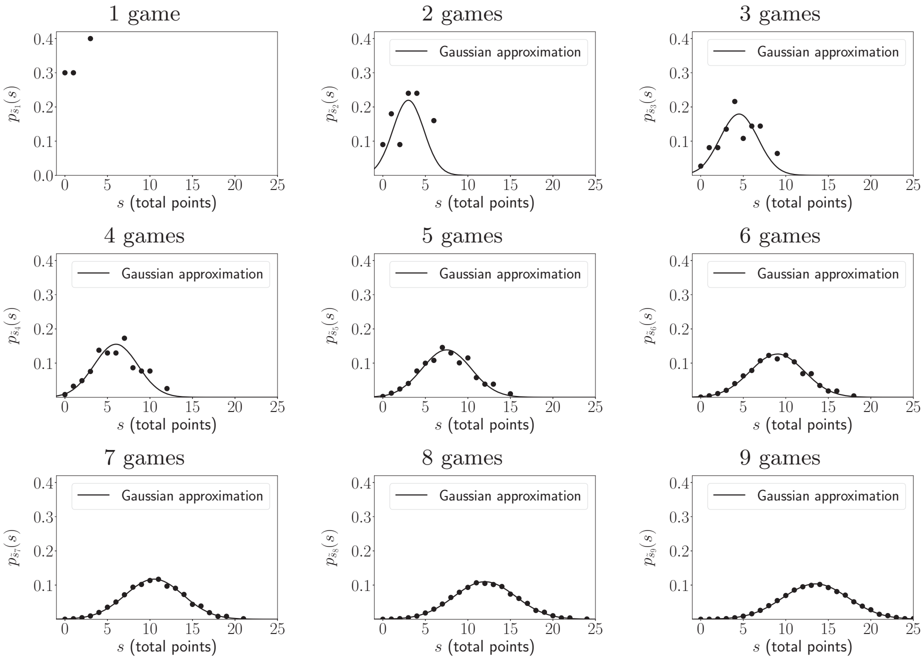


Figure 9.11 Points earned by soccer team. The circular markers represent the pmf of the sum of points earned by the soccer team in Example 9.33 for different numbers of games. The pmfs converge to a Gaussian pdf (black line), as predicted by the central limit theorem (Theorem 9.37).

and $p_{\tilde{x}_i}(x) = 0$ for any other value of x. By Theorem 9.31 the pmf of $\tilde{s}_2$ is equal to the convolution of $p_{\tilde{x}_1}$ and $p_{\tilde{x}_2}$,

$$p_{\tilde{s}_2}(s) = p_{\tilde{x}_1} * p_{\tilde{x}_2}(s) \tag{9.128}$$

$$= \sum_{x=-\infty}^{\infty} p_{\tilde{x}_1}(x)\, p_{\tilde{x}_2}(s-x). \tag{9.129}$$

Very few terms in the infinite sum are nonzero. For example,

$$p_{\tilde{s}_2}(1) = p_{\tilde{x}_1}(0)\, p_{\tilde{x}_2}(1) + p_{\tilde{x}_1}(1)\, p_{\tilde{x}_2}(0) \tag{9.130}$$

$$= 0.18. \tag{9.131}$$

Similar calculations yield the remaining values of the pmf,

$$p_{\tilde{s}_2}(0) = 0.09, \quad p_{\tilde{s}_2}(1) = 0.18, \quad p_{\tilde{s}_2}(2) = 0.09, \tag{9.132}$$

$$p_{\tilde{s}_2}(3) = 0.24, \quad p_{\tilde{s}_2}(4) = 0.24, \quad p_{\tilde{s}_2}(6) = 0.16. \tag{9.133}$$

For all other values of s, $p_{\tilde{s}_2}(s) = 0$.

By Corollary 9.32 the pmf of $\tilde{s}_n$ is

$$p_{\tilde{s}_n}(s) = p_{\tilde{x}_1} * p_{\tilde{x}_2} * \cdots * p_{\tilde{x}_n}(s). \tag{9.134}$$

The random variables are identically distributed, so all the pmfs are the same. Consequently, $p_{\tilde{s}_n}$ is obtained by convolving the same pmf with itself $n-1$ times. Figure 9.11 shows plots of $p_{\tilde{s}_n}$ for $1 \leq n \leq 9$. As n increases, the nonzero support of the pmf grows, and its shape becomes increasingly smooth and similar to a Gaussian pdf. To verify this, we compare the pmf of the sum to a Gaussian pdf with the same mean and variance, which equal

$$\mathrm{E}\left[\tilde{s}_n\right] = \sum_{i=1}^{n} \mathrm{E}\left[\tilde{x}_i\right] = 1.5n, \qquad \mathrm{Var}\left[\tilde{s}_n\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\tilde{x}_i\right] = 1.65n, \tag{9.135}$$

by linearity of expectation and Theorem 9.11, respectively. Figure 9.11 shows that the pmf of the sum and the pdf of the Gaussian are remarkably similar, even for small values of n.

. . . .

The following theorem derives the distribution of the sum of two independent continuous random variables. In this case, the pdf of the sum is equal to the continuous convolution of the individual pdfs.

Theorem 9.34 (Sum of independent continuous random variables)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two independent continuous random variables, belonging to the same probability space, with pdfs $f_{\tilde{a}}$ and $f_{\tilde{b}}$. The pdf of their sum $\tilde{s} = \tilde{a} + \tilde{b}$ equals the convolution of their pdfs,*

$$f_{\tilde{s}}\left(s\right) = f_{\tilde{a}} * f_{\tilde{b}}(s) = \int_{a=-\infty}^{\infty} f_{\tilde{a}}\left(a\right) f_{\tilde{b}}\left(s-a\right) \, \mathrm{d}a. \tag{9.136}$$

In words, $f_{\tilde{s}}\left(s\right)$ is equal to the inner product between $f_{\tilde{a}}$ and a flipped copy of $f_{\tilde{b}}$ shifted by s.

Let $\tilde{a}_1$, $\tilde{a}_2$, . . . , $\tilde{a}_n$ be n independent continuous random variables with pdfs denoted by $f_{\tilde{a}_1}$, . . . , $f_{\tilde{a}_n}$ belonging to the same probability space. The pdf of their sum $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ equals the convolution of their pdfs,

$$f_{\tilde{s}_n}\left(s\right) = f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n}\left(s\right). \tag{9.137}$$

Proof First we derive the cdf of $\tilde{s}$, integrating the joint pdf of $\tilde{a}$ and $\tilde{b}$ over the region corresponding to the event $\tilde{a} + \tilde{b} \leq s$. By the independence assumption, the joint pdf is the product of the individual pdfs, so

$$F_{\tilde{s}}\left(s\right) = \mathrm{P}\left(\tilde{a} + \tilde{b} \leq s\right) \tag{9.138}$$

$$= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{s-a} f_{\tilde{a}}(a) f_{\tilde{b}}\left(b\right) \, \mathrm{d}a \, \mathrm{d}b \tag{9.139}$$

$$= \int_{a=-\infty}^{\infty} f_{\tilde{a}}\left(a\right) F_{\tilde{b}}\left(s-a\right) \, \mathrm{d}a. \tag{9.140}$$

We now differentiate the cdf to obtain the pdf. This requires an interchange of a limit operator with a differentiation operator and another interchange of an integral operator with a differentiation operator, which are justified because the functions involved are bounded and integrable:

$$f_{\tilde{s}}(s) = \frac{\mathrm{d}}{\mathrm{d}s} \lim_{t\to\infty} \int_{a=-t}^{t} f_{\tilde{a}}(a)\, F_{\tilde{b}}(s-a)\, \mathrm{d}a \tag{9.141}$$

$$= \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}s} \int_{a=-t}^{t} f_{\tilde{a}}(a)\, F_{\tilde{b}}(s-a)\, \mathrm{d}a \tag{9.142}$$

$$= \lim_{t\to\infty} \int_{a=-t}^{t} \frac{\mathrm{d}}{\mathrm{d}s} f_{\tilde{a}}(a)\, F_{\tilde{b}}(s-a)\, \mathrm{d}a \tag{9.143}$$

$$= \lim_{t\to\infty} \int_{a=-t}^{t} f_{\tilde{a}}(a)\, f_{\tilde{b}}(s-a)\, \mathrm{d}a. \tag{9.144}$$

The result for n random variables follows immediately from (9.136). ■

Example 9.35 (Coffee supply). A coffee shop in Manhattan is considering how to source their coffee. They have access to many suppliers around the world, but the supply from each supplier is quite volatile, because it depends on local demand and weather conditions. It can be approximately modeled as a uniform random variable between 0 and 1 ton. In order to protect themselves from this volatility, the cafe makes a deal with n suppliers: they will buy a quantity of coffee equal to the each supplier's total supply divided by n.

We model the coffee available from the ith supplier as a random variable $\tilde{c}_i$, $1 \le i \le n$. The suppliers are from very different locations, so we assume that the variables are independent. The total coffee offered by the suppliers is therefore equal to the sum of n i.i.d. uniform random variables,

$$\tilde{s}_n := \sum_{i=1}^{n} \tilde{c}_i, \tag{9.145}$$

and the coffee purchased by the company is equal to the average $\widetilde{m}_n := \tilde{s}_n/n$.

If there are only two suppliers, by Theorem 9.34 the pdf of the sum is equal to the convolution between two uniform pdfs,

$$f_{\tilde{s}_2}(s) = \int_{c=-\infty}^{\infty} f_{\tilde{c}_1}(c)\, f_{\tilde{c}_2}(s-c)\, \mathrm{d}c \tag{9.146}$$

$$= \int_{c=0}^{1} f_{\tilde{c}_2}(s-c)\, \mathrm{d}c. \tag{9.147}$$

The pdf $f_{\tilde{c}_2}(s-c)$ is equal to one if $0 \le s-c \le 1$, i.e. if $s-1 \le c \le s$, and zero otherwise. Consequently, if $0 \le s \le 1$,

$$f_{\tilde{s}_2}(s) = \int_{c=0}^{s} \mathrm{d}c = s, \tag{9.148}$$

if $1 \le s \le 2$,

$$f_{\tilde{s}_2}(s) = \int_{c=s-1}^{1} \mathrm{d}c = 2-s, \tag{9.149}$$

and if $s < 0$ or $s > 2$, then $f_{\tilde{s}_2}(s) = 0$. The pdf of the sum is triangular between 0 and 2. By Theorem 3.20, the pdf of the average is also triangular. It equals

$$f_{\widetilde{m}_2}(m) = 2 f_{\tilde{s}_2}(2m) = \begin{cases} 4m & \text{for } 0 \leq s \leq \frac{1}{2}, \\ 4(1-m) & \text{for } \frac{1}{2} \leq s \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.150}$$

By Theorem 9.34, the pdf of the total coffee from n suppliers is

$$f_{\tilde{s}_n}(s) = f_{\tilde{c}_1} * f_{\tilde{c}_2} * \cdots * f_{\tilde{c}_n}(s), \tag{9.151}$$

and, by Theorem 3.20, the pdf of the average is

$$f_{\widetilde{m}_n}(m) = n f_{\tilde{s}_n}(nm) \tag{9.152}$$

$$= n f_{\tilde{c}_1} * f_{\tilde{c}_2} * \cdots * f_{\tilde{c}_n}(nm). \tag{9.153}$$

In words, the pdf is obtained by convolving the uniform pdf with itself $n-1$ times, and then scaling it. Figure 9.12 shows the pdf for different values of n. The convolutions gradually smooth the shape of the pdf until it resembles a Gaussian. We compare it to a Gaussian with the same mean and variance, which equal

$$\mathrm{E}[\widetilde{m}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\tilde{c}_i] = 0.5, \qquad \mathrm{Var}[\widetilde{m}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[\tilde{c}_i] = \frac{1}{12n}, \tag{9.154}$$

by linearity of expectation, Lemma 7.8, Theorem 9.11 and Lemma 7.34. The Gaussian approximation, depicted in Figure 9.12, is extremely precise.

. . . .

In Examples 9.33 and 9.35, we observe that repeatedly summing independent random variables has a smoothing effect on their pmf or pdf that results in a Gaussian-like distribution. The following theorem provides an additional connection between the convolution operation and the Gaussian distribution. The convolution of two Gaussians pdfs is Gaussian. This means that if we sum independent Gaussian random variables together, the result is Gaussian.

Theorem 9.36 (Sum of independent Gaussian random variables)**.** *Let $\tilde{a}_1$ and $\tilde{a}_2$ be two independent continuous random variables with Gaussian distributions belonging to the same probability space. Their sum $\tilde{s} := \tilde{a}_1 + \tilde{a}_2$ is Gaussian with mean $\mu_{\tilde{s}} := \mu_1 + \mu_2$ and variance $\sigma_{\tilde{s}}^2 := \sigma_1^2 + \sigma_2^2$, where μ_1 and σ_1^2 denote the mean and variance of $\tilde{a}_1$ and μ_2 and σ_2^2 denote the mean and variance of $\tilde{a}_2$.*

Proof By Theorem 9.34, the pdf of $\tilde{s}$ equals

$$f_{\tilde{s}}(s) = \int_{a=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(a-\mu_1)^2}{2\sigma_1^2}\right) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(s-a-\mu_2)^2}{2\sigma_2^2}\right) \mathrm{d}a$$

$$= \int_{a=-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{(a-\mu_1)^2}{\sigma_1^2} + \frac{(s-a-\mu_2)^2}{\sigma_2^2}\right)\right) \mathrm{d}a. \tag{9.155}$$

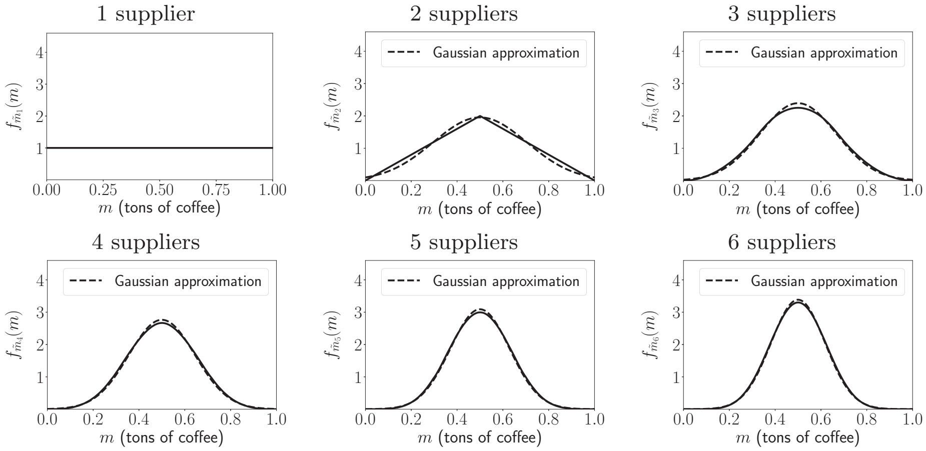


Figure 9.12 Coffee supply. The black curves in each graph represent the pdfs of the quantity of coffee purchased by the coffee shop in Example 9.35 for different numbers of suppliers. The dashed line depicts a Gaussian approximation that is increasingly precise, as predicted by the central limit theorem (Theorem 9.37).

Manipulating the expression in the exponential, we obtain

$$\frac{(a-\mu_1)^2}{\sigma_1^2} + \frac{(s-a-\mu_2)^2}{\sigma_2^2} \tag{9.156}$$

$$= \frac{\sigma_2^2(a^2+\mu_1^2-2\mu_1 a) + \sigma_1^2(s^2+a^2+\mu_2^2-2sa-2s\mu_2+2a\mu_2)}{\sigma_1^2\sigma_2^2} \tag{9.157}$$

$$= \frac{\sigma_{\tilde{s}}^2 a^2 - 2(\sigma_2^2\mu_1 + \sigma_1^2(s-\mu_2))a + \sigma_2^2\mu_1^2 + \sigma_1^2(s^2+\mu_2^2-2s\mu_2)}{\sigma_1^2\sigma_2^2} \tag{9.158}$$

$$= \frac{\sigma_{\tilde{s}}^2}{\sigma_1^2\sigma_2^2}\left((a-b)^2 - b^2 + c\right), \tag{9.159}$$

where

$$b := \frac{\sigma_2^2\mu_1 + \sigma_1^2(s-\mu_2)}{\sigma_{\tilde{s}}^2}, \tag{9.160}$$

$$c := \frac{\sigma_2^2\mu_1^2 + \sigma_1^2(s-\mu_2)^2}{\sigma_{\tilde{s}}^2}. \tag{9.161}$$

These operations are called *completing the square*. We realize that

$$c - b^2 = \frac{\sigma_1^2\sigma_2^2(s-\mu_{\tilde{s}})^2}{\sigma_{\tilde{s}}^4}. \tag{9.162}$$

Combining (9.159), (9.155) and (9.162) we obtain

$$\begin{aligned} f_{\tilde{s}}(s) &= \int_{a=-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{\sigma_{\tilde{s}}^2}{2\sigma_1^2\sigma_2^2}\left((a-b)^2 + \frac{\sigma_1^2\sigma_2^2(s-\mu_{\tilde{s}})^2}{\sigma_{\tilde{s}}^4}\right)\right) \mathrm{d}a \\ &= \frac{1}{\sqrt{2\pi}\sigma_{\tilde{s}}} \exp\left(-\frac{(s-\mu_{\tilde{s}})^2}{2\sigma_{\tilde{s}}^2}\right) \int_{a=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\frac{\sigma_1\sigma_2}{\sigma_{\tilde{s}}}} \exp\left(-\frac{(a-b)^2}{\frac{2\sigma_1^2\sigma_2^2}{\sigma_{\tilde{s}}^2}}\right) \mathrm{d}a \\ &= \frac{1}{\sqrt{2\pi}\sigma_{\tilde{s}}} \exp\left(-\frac{(s-\mu_{\tilde{s}})^2}{2\sigma_{\tilde{s}}^2}\right). \end{aligned} \tag{9.163}$$

The integral is equal to one, because the term inside is a Gaussian pdf with mean b and standard deviation $\sigma_1\sigma_2/\sigma_{\tilde{s}}$. ■

9.7.2 Convergence To The Gaussian Distribution

In this section, we study the distribution of the average of sequences of independent identically distributed (i.i.d.) random variables. Let $\tilde{x}_1$, $\tilde{x}_2$, … denote a sequence of i.i.d. random variables with mean μ and variance σ^2. The sample mean of the first n variables is

$$\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j. \tag{9.164}$$

As derived in the proof of Theorem 9.22, the mean and variance of the sample mean equal

$$\mathrm{E}\left[\widetilde{m}_n\right] = \mu, \qquad \mathrm{Var}\left[\widetilde{m}_n\right] = \frac{\sigma^2}{n}, \tag{9.165}$$

so the sample mean is an unbiased estimator of the true mean μ, and its variance or standard error decays linearly with n. Consequently, the sample mean converges to μ with high probability, as stated in the law of large numbers. However this does not provide a precise characterization of the *distribution of the sample mean*. For example, we may be interested in the probability $\mathrm{P}\left(|\widetilde{m}_n - \mu| > \epsilon\right)$ that $\widetilde{m}_n$ deviates from μ by some constant ϵ for a fixed value of n. As illustrated in Example 9.26, our bounds based on Chebyshev's inequality provide a terrible estimate of this probability.

Empirically, in Examples 9.33 and 9.35, we observe that the distribution of sums of independent random variables is close to being Gaussian. This suggests approximating $\widetilde{m}_n$ using a Gaussian random variable with mean μ and variance σ^2/n. By Theorem 3.32, this is equivalent to approximating the standardized sample mean

$$s(\widetilde{m}_n) := \frac{\widetilde{m}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{9.166}$$

using a standard Gaussian with zero mean and unit variance. The central limit theorem establishes that the approximation becomes arbitrarily accurate as n tends to infinity.

Theorem 9.37 (Central limit theorem)**.** *Let $\tilde{x}_1$, $\tilde{x}_2$, … be a countably infinite sequence of independent identically distributed random variables with mean μ and variance σ^2 belonging to the same probability space. The standardized running average*

$$s(\widetilde{m}_n) := \frac{\widetilde{m}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{9.167}$$

converges in distribution to a standard Gaussian random variable with zero mean and unit variance, in the sense that the cdf of $F_{s(\widetilde{m}_n)}$ converges to the cdf of a standard Gaussian as $n \to \infty$.

Proof Unfortunately, the proof of the central limit theorem is beyond the scope of this book, as it requires introducing advanced concepts from probability theory and functional analysis. We refer to Chapter 3 in (Durrett, 2019) for the proof and the necessary mathematical background. ■

It is important to emphasize that the central limit theorem states that the sample mean *converges in distribution*. In contrast to convergence in mean square or convergence in probability (see Corollary 9.15 and Theorem 9.22), this does not mean that the sample mean becomes arbitrarily close to a certain value, but rather that its distribution is increasingly well approximated as Gaussian. Even though the theorem is asymptotic, the Gaussian approximation is often very accurate even for small n, as illustrated in Examples 9.33 and 9.35. Motivated by this, we define a Gaussian approximation to the sample mean inspired by the central limit theorem.

Definition 9.38 (Gaussian approximation to the sample mean)**.** *Given n independent identically distributed random variables $\tilde{x}_1$, $\tilde{x}_2$, …, $\tilde{x}_n$ with mean μ and σ^2, the Gaussian approximation of the sample mean*

$$\tilde{m} := \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \tag{9.168}$$

is a Gaussian random variable with mean μ and variance $\frac{\sigma^2}{n}$.

A common application of the Gaussian approximation to the sample mean is to approximate the binomial distribution.

Definition 9.39 (Gaussian approximation to the binomial distribution)**.** *Recall that a binomial random variable with parameters n and θ can be represented as the sum of n independent Bernoulli random variables $\tilde{b}_1$, …, $\tilde{b}_n$ with parameter θ (see Example 2.16),*

$$\tilde{a} = \sum_{i=1}^{n} \tilde{b}_i. \tag{9.169}$$

By the Gaussian approximation of the sample mean in Definition 9.38, we can approximate $\tilde{a}/n$ as a Gaussian random variable with mean θ and variance $\theta(1-$

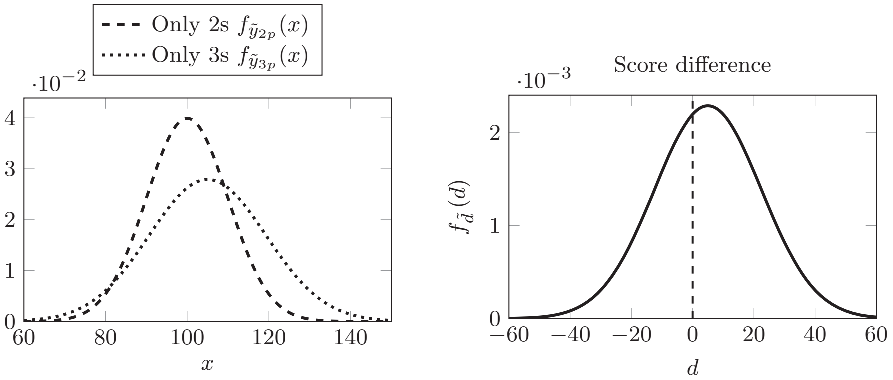


Figure 9.13 Basketball strategy. The graph on the left shows the Gaussian approximations derived in Example 9.40 for the distribution of points scored scored by a team that only attempts 2-point ($\tilde{y}_{2p}$) or 3-point ($\tilde{y}_{2p}$) shots. The graph on the right shows the Gaussian approximation to the score difference between the teams.

$\theta)/n$. Equivalently, by Theorem 3.32, we can approximate $\tilde{a}$ as a Gaussian random variable with mean $n\theta$ and variance $n\theta(1-\theta)$.

Example 9.40 (Basketball strategy)**.** A basketball team hires you as an analyst. Your first task is to compare two strategies: only taking 2-point shots (*Strategy 2p*) or only taking 3-point shots (*Strategy 3p*). To compare the strategies, you decide to model a game between two identical teams, where one employs Strategy 2p and the other Strategy 3p.

From past data, you determine that for the current team roster, the probability of making a 2-point and a 3-point shot are $\theta_2 := 0.5$ and $\theta_3 := 0.35$, respectively, and that the different shots in a game can be assumed to be independent. There are typically around 100 possessions in a game, so you model the number of shots made when following Strategy 2p and Strategy 3p as two binomial random variables $\tilde{x}_{2p}$ and $\tilde{x}_{3p}$ with parameters $n := 100$ and $\theta_2 := 0.5$ for $\tilde{x}_{2p}$ and $n := 100$ and $\theta_3 := 0.35$ for $\tilde{x}_{3p}$. We are interested in the distribution of the score difference $\tilde{d} := 3\tilde{x}_{3p} - 2\tilde{x}_{2p}$, and, in particular, in the probability that it is positive or negative, as this determines the outcome of the game. Deriving the exact pmf of $\tilde{d}$ is quite complicated. Instead, we approximate it based on the central limit theorem.

By the Gaussian approximation to the binomial distribution (Definition 9.39), we can approximate $\tilde{x}_{2p}$ as a Gaussian with mean $100\,\theta_2$ and variance $100\,\theta_2(1-\theta_2)$. By Theorem 3.32, if $\tilde{x}_{2p}$ is Gaussian, then the random variable $\tilde{y}_{2p} := 2\tilde{x}_{2p}$, which represents the corresponding score, is Gaussian with mean $200\,\theta_2 = 100$ and variance $400\,\theta_2(1-\theta_2) = 100$. By the same reasoning, $\tilde{x}_{3p}$ is approximately Gaussian with mean $100\,\theta_3$ and variance $100\,\theta_3(1-\theta_3)$. Consequently, the random variable $\tilde{y}_{3p} := 3\tilde{x}_{3p}$, representing the corresponding score, is approximately Gaussian with mean $300\,\theta_3 = 105$ and variance $900\,\theta_3(1-\theta_3) = 204.75$.

The left graph in Figure 9.13 compares the approximate distributions of $\tilde{y}_{2p}$ and $\tilde{y}_{3p}$. Attempting 3-point shots results in more points on average (the mean is larger), but is also more volatile (the variance is also larger). By Theorem 9.36, if $\tilde{y}_{2p}$ and $\tilde{y}_{3p}$ followed their approximate Gaussian distributions exactly, then the score difference $\tilde{d} := \tilde{y}_{3p} - \tilde{y}_{2p}$ would be Gaussian with mean $105 - 100 = 5$ and variance $204.75 + 100 = 304.75$ (depicted on the right in Figure 9.13). By Theorem 3.32 we can express such a Gaussian as $\sqrt{304.75}\tilde{z} + 5$, where $\tilde{z}$ is a standard Gaussian with zero mean and unit variance. The probability that Strategy 3p beats Strategy 2p can therefore be approximated as,

$$\mathrm{P}(\text{Strategy 3p beats Strategy 2p}) \approx \mathrm{P}\left(\sqrt{304.75}\tilde{z} + 5 > 0\right) \tag{9.170}$$

$$= \mathrm{P}\left(\tilde{z} > -0.2864\right) = 0.613. \tag{9.171}$$

Just shooting 3 pointers wins approximately 60% of the time. To evaluate our approximation, we leverage the Monte Carlo method described in Section 1.7. We simulate one million games between a team following Strategy 2p and a team following Strategy 3p. Strategy 3p indeed wins close to 60% of the games (599,790). This is obviously a cartoon example, but it illustrates why the proportion of 3-point shots taken by NBA teams has risen steadily since the introduction of the 3-point shot in 1979.

. . . .

We now consider the implications of the central limit theorem for the behavior of a sample mean computed from random samples. The law of large numbers establishes that the sample mean converges to the population mean, if the samples are measured independently and uniformly at random with replacement (see Corollary 9.15 and Theorem 9.23). The central limit theorem states that the sample mean converges in distribution to a Gaussian random variable centered at the population mean with a standard deviation equal to the standard error (see Definition 9.9).

Corollary 9.41 (Distribution of the sample mean)**.** *Let $a_1, a_2, \ldots, a_N$ denote a dataset of size N with population mean μ_{pop} and population variance σ^2_{pop}, defined as in* (9.9) *and* (9.19)*, respectively. Let $\tilde{x}_1, \tilde{x}_2, \ldots,$ be a sequence of independent, uniform random samples following Definition 9.3. We define the standardized sample mean as*

$$s(\widetilde{m}_n) := \frac{\widetilde{m}_n - \mu_{\text{pop}}}{\text{se}\left[\widetilde{m}_n\right]}, \tag{9.172}$$

where $\widetilde{m}_n := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j$ and $\text{se}\left[\widetilde{m}\right] = \sigma_{\text{pop}}/\sqrt{n}$ denotes the standard error (see Theorem 9.13). As $n \to \infty$, the standardized sample mean converges in distribution to a standard Gaussian random variable with zero mean and unit variance.

Proof The random variables $\tilde{x}_1, \tilde{x}_2, \ldots$ are i.i.d. with mean μ_{pop} and variance

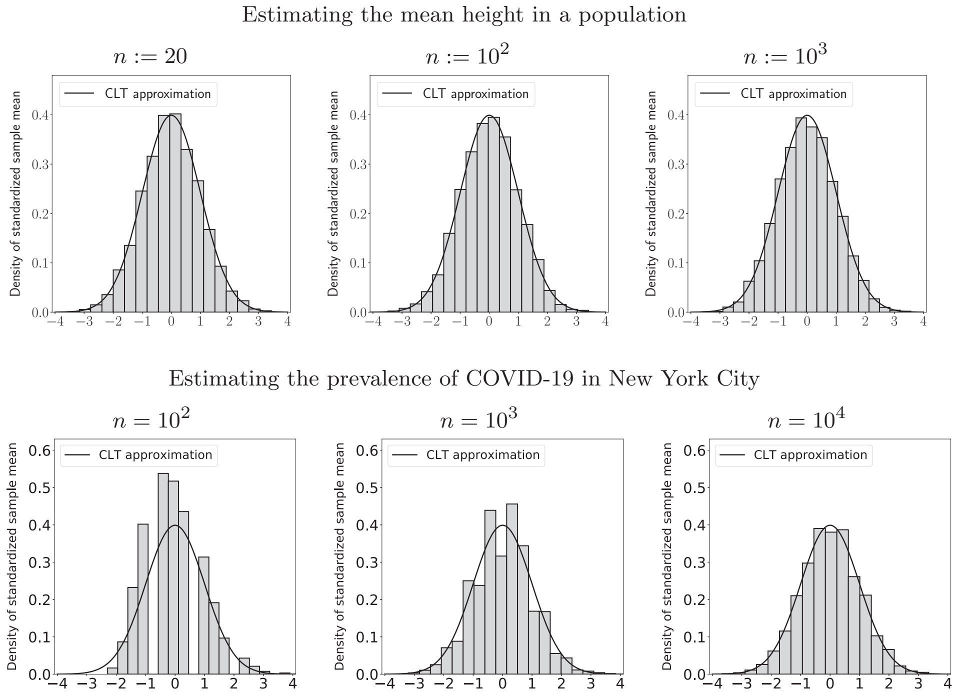


Figure 9.14 Distribution of the sample mean. The top row shows histograms of the standardized sample mean, defined in Corollary 9.41, of n independent, uniform samples from the dataset in Example 9.1 for different values of n. The bottom row shows histograms of the standardized sample proportion computed from n independent measurements in the scenario described in Example 9.2, also for different values of n. Each histogram is computed using 10^4 sample means or proportions. A Gaussian approximation based on the central limit theorem is shown superposed on each histogram. As n increases, the approximation becomes increasingly accurate, as predicted by Corollary 9.41.

σ^2_{pop} by (9.14) and (9.36), so the result follows directly from the central limit theorem. ■

Figure 9.14 shows histograms of the standardized sample mean $s(\tilde{m})$ of n independent, uniform samples from the dataset in Example 9.1 for different values of n. The Gaussian approximation from Corollary 9.41 is very accurate even for small values of n. The figure also shows histograms of the standardized sample proportion of positive tests among n individuals chosen independently and uniformly at random in Example 9.2. Since the sample proportion can be interpreted as a sample mean (see Example 9.4), the central limit theorem applies: the Gaussian approximation improves as we increase n, and is already very accurate for $n := 10{,}000$.

9.7.3 The Financial Crisis And The Central Limit Theorem: How Not To Estimate Risk

The Gaussian approximation provided by the central limit theorem is widely used in probabilistic modeling. However, it is critical not to forget that it relies on the assumption that the quantity of interest is an average of *independent* quantities. In this section we illustrate the dangers of blindly relying on Gaussian approximations when their underlying assumptions do not hold.

We focus on a toy example inspired by the 2008 financial crisis. Our goal is to evaluate the risk of a collateralized debt obligation (CDO), a financial instrument that gained notorious fame due to the 2008 financial crisis. Our CDO of interest consists of a pool of n subprime mortgages from borrowers with a poor credit history (so, to be more precise, it is a collateralized mortgage obligation). The probability that each of the borrowers defaults and does not pay back their debt is 2/3.

In order to alleviate the risk, the CDO is divided into ten different sections called *tranches*. When borrowers default, the tranches suffer losses sequentially. If 10% or less borrowers default, then only the first tranche is affected. If between 10% and 20% default, then the first and second tranches lose money. The last tranche, which is known as the *senior* tranche, is the most protected; more than 90% of the borrowers must default for it to suffer losses.

In the remainder of this section, we evaluate the risk of the senior tranche under two different modeling assumptions. Our analysis illustrates the dramatic results of inadequate independence assumptions in risk estimation.

Assumption 1: Borrowers Default Independently

Let us assume that the borrowers default independently. In that case, the number of defaults is distributed as a binomial random variable $\tilde{d}_1$ with parameters n and $\theta := 2/3$. By Definition 9.39, the probability that more than 90% of the borrowers default can be approximated as

$$\mathrm{P}\left(\tilde{d}_1 > 0.9n\right) = \mathrm{P}\left(\frac{\tilde{d}_1 - \theta n}{\sqrt{\theta(1-\theta)n}} > \frac{0.9n - \theta n}{\sqrt{\theta(1-\theta)n}}\right) \tag{9.173}$$

$$\approx \mathrm{P}\left(\tilde{z} > 0.49\sqrt{n}\right), \tag{9.174}$$

where $\tilde{z}$ is a Gaussian random variable with zero mean and unit variance. The probability that the senior tranche suffers losses decreases as we increase n. By selecting n large enough, we can make it into a low-risk investment. For instance, if there are $n := 100$ mortgages in the CDO, the risk of losing money from the senior tranche is essentially zero (less than 10^{-6})! Figure 9.15 shows the pmf of $\tilde{d}_1$ and the corresponding Gaussian approximation for $n := 100$.

Assumption 2: Borrowers Default Depending On Economic Context

We now assume that the probability of default depends on the state of the economy. We define a random variable $\tilde{r}$ that represents to what extent the economy

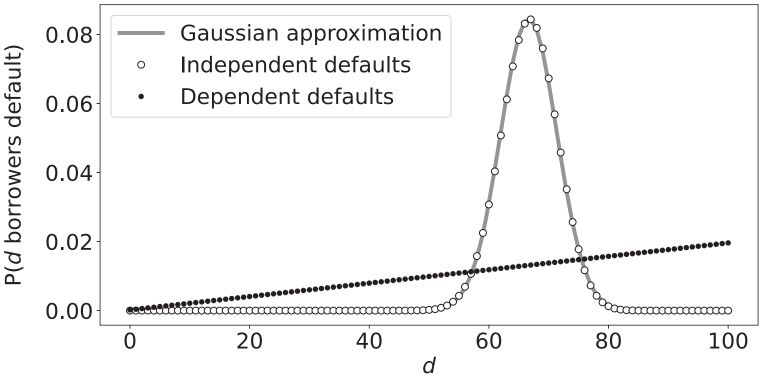


Figure 9.15 Effect of independence assumptions on risk estimates. The figure compares the estimated distributions of the number of defaults in a pool of 100 subprime mortgages for the two scenarios described in Section 9.7.3. The white markers depict the pmf of the number of borrowers that default under Assumption 1 (the borrowers default independently). The gray curve shows a Gaussian approximation to the pmf. Under this scenario, it is extremely unlikely that more than 80% of the borrowers will default. The black markers depict the pmf of the number of borrowers that default under Assumption 2 (the default risk depends on a common latent variable representing economic context). Due to the dependence between borrowers under Assumption 2, the probability of many of them defaulting simultaneously is much higher.

is in recession: $\tilde{r} = 0$ indicates a strong economy, whereas $\tilde{r} = 1$ corresponds to economic disaster. The pdf of $\tilde{r}$ is equal to

$$f_{\tilde{r}}(r) := 2r, \qquad 0 \leq r \leq 1, \tag{9.175}$$

and zero otherwise. Conditioned on $\tilde{r} = r$, each borrower is modeled as a Bernoulli random variable $\tilde{b}_i$ with parameter r, where $\tilde{b}_i = 1$ indicates that the ith borrower defaults. In other words, the probability of default is equal to r, so higher values of $\tilde{r}$ indeed correspond to a worse economic context. By Theorem 6.5, the probability that each individual borrower defaults equals

$$p_{\tilde{b}_i}(1) = \int_{r=-\infty}^{\infty} p_{\tilde{b}_i \,|\, \tilde{r}}(1 \,|\, r) f_{\tilde{r}}(r) \, \mathrm{d}r \tag{9.176}$$

$$= \int_{r=0}^{1} 2r^2 \, \mathrm{d}r = \frac{2}{3}, \tag{9.177}$$

which is *exactly the same* as under Assumption 1.

We assume that there is no additional dependence between the borrowers, so that $\tilde{b}_1, \ldots, \tilde{b}_n$ are conditionally independent given $\tilde{r}$. As a result, conditioned on $\tilde{r} = r$, the number of defaults $\tilde{d}_2$ has a binomial distribution with parameters n

and r. By Theorem 6.5, its pmf equals

$$p_{\tilde{d}_2}(d) = \int_{r=-\infty}^{\infty} f_{\tilde{r}}(r) p_{\tilde{d}_2 \,|\, \tilde{r}}(d \,|\, r) \; \mathrm{d}r \tag{9.178}$$

$$= \int_{r=0}^{1} 2r \binom{n}{d} r^{d} (1-r)^{n-d} \; \mathrm{d}r \tag{9.179}$$

$$= 2 \binom{n}{d} \int_{r=0}^{1} r^{d+1} (1-r)^{n-d} \; \mathrm{d}r \tag{9.180}$$

$$= \frac{2n!}{d!(n-d)!} \frac{(d+1)!(n-d)!}{(n+2)!} \tag{9.181}$$

$$= \frac{2(d+1)}{(n+1)(n+2)}. \tag{9.182}$$

Figure 9.15 shows the pmf of $\tilde{d}_2$ for $n := 100$. It does not resemble a Gaussian at all! Note that this does *not* contradict the central limit theorem, because the borrowers do not default independently.

The probability that more than 90% of the borrowers default in this scenario is equal to

$$\mathrm{P}\left(\tilde{d}_2 > 0.9n\right) = \sum_{d=0.9n+1}^{n} p_{\tilde{d}_2}(d) \tag{9.183}$$

$$= \frac{2}{(n+1)(n+2)} \left(0.1n + \sum_{d=0.9n+1}^{n} d \right) \tag{9.184}$$

$$= \frac{0.2n + 0.1n(1.9n+1)}{(n+1)(n+2)} \tag{9.185}$$

where we apply the arithmetic series formula $\sum_{d=a}^{b} d = \frac{(a+b)(b-a+1)}{2}$. For $n := 100$, the probability is 0.187. This is dramatically higher than under Assumption 1. In this scenario, when the economy is in recession, borrowers are likely to default *simultaneously.*

Under Assumption 1, we can reduce the risk of the senior tranche by including more mortgages in the CDO. Under Assumption 2, this is not possible: the limit of the probability in (9.185) does not converge to zero as n tends to infinity, it equals 0.19. No matter how many mortgages we pool together, the senior tranche remains a risky investment. It is important to clarify that this is a cartoon example: the actual models used to evaluate the risk of CDOs prior to the 2008 financial crisis did incorporate dependence. However, they definitely underestimated the risk of CDOs based on subprime mortgages, which suffered enormous losses despite being considered low-risk investments.

9.8 Confidence Intervals

When we perform an estimate using random samples, the estimate is *uncertain*, as illustrated in Figures 9.1 and 9.2. It is therefore often useful to report a *confidence interval* containing the parameter with high probability, as opposed to a single point estimate. In Sections 9.8.1 and 9.8.2, we explain how to build confidence intervals for the mean, and for proportions and probabilities. Section 9.8.3 discusses how to interpret confidence intervals and cautions against applying them when the available data are not sampled independently.

9.8.1 Confidence Interval For The Mean

As explained in Section 9.1, estimates obtained from random samples are uncertain: if we were to repeat our measurements, we would obtain a different estimate. Confidence intervals allow us to quantify this uncertainty by providing a range of values that contain the population parameter with high probability. In this section we show how to construct confidence intervals for the mean of a quantity of interest.

We consider the problem of estimating the population mean μ_{pop} of a population a_1, a_2, ..., a_N from random samples. The available measurements are modeled as a n i.i.d. random variables $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ following Definition 9.3. Our estimator is the sample mean $\widetilde{m} := \sum_{j=1}^{n} \tilde{x}_j$. In order to quantify its uncertainty, we would like to build an interval around $\widetilde{m}$ that contains μ_{pop} with high probability.

By the central limit theorem, $\widetilde{m}$ is approximately Gaussian with mean μ_{pop}. Consequently, we can select a constant c such that $\widetilde{m}$ belongs to the interval $[\mu_{\text{pop}} - c, \mu_{\text{pop}} + c]$ with high probability. Unfortunately, to compute this interval, we need to know μ_{pop}, which is the parameter that we are trying to estimate in the first place! However, if $\widetilde{m}$ is in $[\mu_{\text{pop}} - c, \mu_{\text{pop}} + c]$, then μ_{pop} is in $[\widetilde{m} - c, \widetilde{m} + c]$. Therefore, $[\widetilde{m} - c, \widetilde{m} + c]$ is a confidence interval for μ_{pop}! This interval *can* be computed from the available data, as long as we are able to determine c.

The interval $[\mu_{\text{pop}} - c, \mu_{\text{pop}} + c]$ is fixed and deterministic, but the confidence interval $[\widetilde{m} - c, \widetilde{m} + c]$ is random and depends on the available measurements, as illustrated in Figure 9.16. The following lemma derives the value of the constant c for any estimator that has a Gaussian distribution.

Lemma 9.42 (Confidence interval for the mean of a Gaussian random variable)**.** *Let $\tilde{a}$ be a Gaussian random variable with mean μ and variance σ^2 and let $F_{\tilde{z}}$ denote the cdf of a standard Gaussian random variable $\tilde{z}$ with zero mean and unit variance. For any $\alpha \in (0, 1)$, the random interval*

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{a} - c_\alpha \sigma, \tilde{a} + c_\alpha \sigma\right], \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{9.186}$$

is a 1-α confidence interval for μ, in the sense that

$$\mathrm{P}\left(\mu \in \widetilde{\mathcal{I}}_{1-\alpha}\right) = 1 - \alpha. \tag{9.187}$$

In particular,

$$\widetilde{\mathcal{I}}_{0.95} := [\tilde{a} - 1.96\sigma, \tilde{a} + 1.96\sigma] \tag{9.188}$$

is a 0.95 confidence interval for μ.

Proof The event $\mu \in \widetilde{\mathcal{I}}_{1-\alpha}$ is the complement of the union of the events $\tilde{a} - c_\alpha \sigma > \mu$ and $\tilde{a} + c_\alpha \sigma < \mu$, which are disjoint, so

$$\begin{aligned}
\mathrm{P}\left(\mu \in \widetilde{\mathcal{I}}_{1-\alpha}\right) &= 1 - \mathrm{P}\left(\tilde{a} - c_\alpha \sigma > \mu\right) - \mathrm{P}\left(\tilde{a} + c_\alpha \sigma < \mu\right) && (9.189)\\
&= 1 - \mathrm{P}\left(\frac{\tilde{a} - \mu}{\sigma} > c_\alpha\right) - \mathrm{P}\left(\frac{\tilde{a} - \mu}{\sigma} < -c_\alpha\right) && (9.190)\\
&= 1 - \mathrm{P}\left(\tilde{z} > c_\alpha\right) - \mathrm{P}\left(\tilde{z} < -c_\alpha\right) && (9.191)\\
&= 1 - 2\mathrm{P}\left(\tilde{z} > c_\alpha\right), && (9.192)
\end{aligned}$$

where $\tilde{z} := (\tilde{a} - \mu)/\sigma$ is a standard Gaussian by Theorem 3.32, and the last step follows from the symmetry of the Gaussian pdf. By the definition of c_α,

$$\begin{aligned}
\mathrm{P}\left(\tilde{z} > c_\alpha\right) &= 1 - \mathrm{P}\left(\tilde{z} \leq c_\alpha\right) && (9.193)\\
&= 1 - F_{\tilde{z}}(c_\alpha) && (9.194)\\
&= \frac{\alpha}{2}. && (9.195)
\end{aligned}$$

Plugging this into (9.192) yields $\mathrm{P}(\mu \in \widetilde{\mathcal{I}}_{1-\alpha}) = 1 - \alpha$. The expression for $\widetilde{\mathcal{I}}_{0.95}$ follows from the fact that $c_{0.05} = 1.96$ because $F_{\tilde{z}}(1.96) = 1 - 0.05/2$. ■

Armed with Lemma 9.42, we can now build confidence intervals for the mean based on the Gaussian approximation of the sample mean in Definition 9.38.

Definition 9.43 (Approximate confidence interval for the mean)**.** *Let $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ be n independent identically distributed random variables with mean μ and variance σ^2, and let $\tilde{m} := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j$ be their sample mean. For any $\alpha \in (0, 1)$, the random interval*

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - \frac{c_\alpha \sigma}{\sqrt{n}}, \tilde{m} + \frac{c_\alpha \sigma}{\sqrt{n}}\right], \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{9.196}$$

is a 1-α confidence interval for μ, where $F_{\tilde{z}}$ denotes the cdf of a standard Gaussian with zero mean and unit variance.

Specifically, if $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$ are independent, uniform random samples from a dataset $\{a_1, a_2, \ldots, a_N\}$ with population mean μ_{pop} and population variance σ_{pop}^2 (defined as in (9.9) and (9.19), respectively), then the random interval

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - \frac{c_\alpha \sigma_{\text{pop}}}{\sqrt{n}}, \tilde{m} + \frac{c_\alpha \sigma_{\text{pop}}}{\sqrt{n}}\right] \tag{9.197}$$

is a 1-α confidence interval for μ_{pop}.

The careful reader may have realized that the length of the confidence intervals in Definition 9.43 depends on the population variance, which is probably

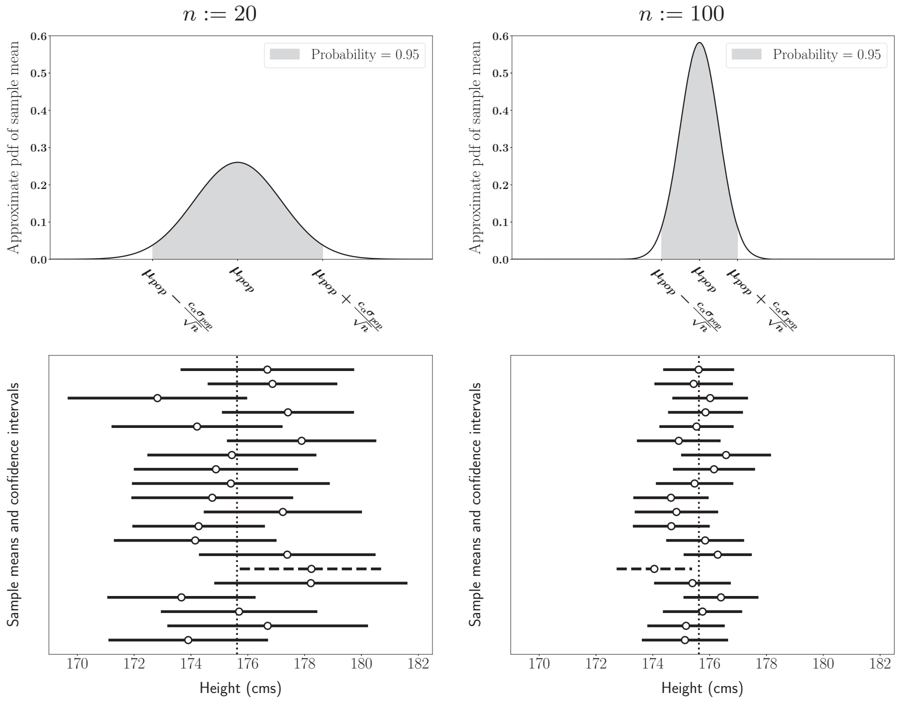


Figure 9.16 Confidence intervals based on the central limit theorem. The top row depicts the Gaussian approximation of the distribution of the sample mean in Definition 9.38 for the dataset in Example 9.1 for two different values of n. The sample mean belongs to the shaded deterministic interval with probability 0.95. The bottom row shows 0.95 confidence intervals based on Definition 9.43. These random intervals are centered at the sample mean (white markers) and have approximately the same length as the deterministic interval (but not exactly the same because they are computed based on the sample variance, not the true population variance). 95% (38/40) of the intervals contain the population mean (dotted line). Those that do not are depicted by dashed lines.

unknown to us! To build confidence intervals in practice, we replace the population variance by the best available estimate or upper bound. A reasonable choice is the sample variance, which is a consistent estimator of the population variance by Theorem 9.25 and typically results in an accurate approximation unless the sample size is very small. Figure 9.16 shows approximate 0.95 confidence intervals for the population mean in Example 9.1 computed from different batches of random samples using the corresponding sample variance.

9.8.2 Confidence Intervals For Probabilities And Proportions

In this section we explain how to build confidence intervals for probabilities and proportions estimated from independent measurements.

Definition 9.44 (Approximate confidence interval for a probability)**.** *Let $\tilde{b}_1$, $\tilde{b}_2$, ..., $\tilde{b}_n$ be n independent Bernoulli random variables with parameter θ, and let $\widetilde{m} := \frac{1}{n}\sum_{j=1}^{n}\tilde{b}_j$ be the proportion of these variables that are equal to one. For any $\alpha \in (0,1)$, the random interval*

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - c_\alpha\sqrt{\frac{\theta(1-\theta)}{n}}, \tilde{m} + c_\alpha\sqrt{\frac{\theta(1-\theta)}{n}}\right], \tag{9.198}$$

is a 1-α confidence interval for θ, where $c_\alpha := F_{\tilde{z}}^{-1}\left(1-\frac{\alpha}{2}\right)$ and $F_{\tilde{z}}$ denotes the cdf of a standard Gaussian with zero mean and unit variance. This interval is included in the following wider confidence interval, which does not depend on θ:

$$\widetilde{\mathcal{I}}_{1-\alpha} \subset \left[\tilde{m} - \frac{0.5c_\alpha}{\sqrt{n}}, \tilde{m} + \frac{0.5c_\alpha}{\sqrt{n}}\right]. \tag{9.199}$$

In particular,

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right]. \tag{9.200}$$

Derivation The mean of each Bernoulli random variable $\tilde{b}_i$ is θ by Lemma 7.21 and the variance equals $\theta(1-\theta)$ by Lemma 7.38. The confidence interval $\widetilde{\mathcal{I}}_{1-\alpha}$ is obtained by applying Definition 9.43 with $\tilde{x}_j := \tilde{b}_j$, $1 \leq j \leq n$. The wider interval (9.199) is obtained from the bound $\theta(1-\theta) \leq 0.25$, which holds because the maximum of $h(\theta) := \theta(1-\theta)$ equals 0.25.* ■

Example 9.45 (Prevalence of COVID-19: Sample size)**.** We consider the setting in Example 9.2. Our goal is to determine how many tests we need to perform in order to obtain an accurate estimate of the prevalence θ_{pop} of COVID-19 in New York. More specifically, we would like to ensure that we make an error of less that 1% in our estimate with probability at least 0.95. Equivalently, we would like the half-length of a 0.95 confidence interval for the fraction of positive tests to be 0.01 or less.

Assuming that we test n individual selected independently and uniformly at random from the population, the test results are independent Bernoulli random variables with parameter θ_{pop}, as explained in Example 9.4. Therefore, by (9.200), the half-length of the approximate 0.95 confidence interval is smaller than 0.01 if

$$\frac{c_\alpha\sigma_{\text{pop}}}{\sqrt{n}} = \frac{0.98}{\sqrt{n}} \leq 0.01, \tag{9.201}$$

*The second derivative of $h(\theta)$ is -2, so the function is strictly concave. We can therefore set the first derivative $1-2\theta$ equal to zero to find the maximum.

which yields the following bound on n:

$$n \geq 9604 = \frac{0.98^2}{0.01^2}. \tag{9.202}$$

We need to test less that 10,000 people in a population of 8 million, which demonstrates the power of random sampling!

Our result is actually an overestimate, unless the prevalence is close to 50%. Otherwise, the bound $\theta_{\text{pop}}(1-\theta_{\text{pop}}) \leq 0.25$ in Definition 9.44 is quite conservative. If $\theta_{\text{pop}} := 0.05$, plugging the true prevalence into (9.198) reduces the bound to 1,825 tests. Let us verify that this is precise. If we carry out 1,825 tests, we make an error of less than 1% if the fraction of positive tests is between 4% (73 tests) and 6% (110 tests). The number of positive tests $\tilde{t}$ is binomial with parameters $n = 1825$ and $\theta_{\text{pop}} = 0.05$. The exact probability that we make an error of less than 1% is very close to 0.95; it equals

$$\begin{aligned} \mathrm{P}(73 \leq \tilde{t} \leq 110) &= \sum_{i=73}^{110} \binom{1825}{i} 0.05^i \, 0.95^{1825-i} && (9.203) \\ &= 0.959\%. && (9.204) \end{aligned}$$

. . . .

An important practical application of confidence intervals is to quantify the uncertainty in probability estimates obtained via the Monte Carlo method described in Section 1.7.

Example 9.46 (Confidence intervals for the Monte Carlo method)**.** Confidence intervals allow us to perform uncertainty quantification when applying the Monte Carlo method in Definition 1.35. Let $\theta := \mathrm{P}(A)$ be the true probability of an event of interest A. If we perform n independent Monte Carlo simulations, we can model whether the corresponding outcome is in A as a Bernoulli random variable with parameter θ. Applying Definition 9.44 to these Bernoulli random variables yields the approximate $1-\alpha$ confidence interval

$$\left[\mathrm{P}_{\text{MC}}(A) - \frac{0.5 c_\alpha}{\sqrt{n}}, \mathrm{P}_{\text{MC}}(A) + \frac{0.5 c_\alpha}{\sqrt{n}} \right], \tag{9.205}$$

where the probability estimate $\mathrm{P}_{\text{MC}}(A)$ is the proportion of simulations for which the outcome is in A.

Figure 9.17 shows 0.95 confidence intervals for the probability that Serbia and Latvia win the gold medal in the 3x3 basketball tournament at the Tokyo Olympics, according to the model described in Example 1.36. For $n := 10^3$ the probability estimate for Latvia is higher than that of Serbia, but the confidence intervals overlap, so we cannot be sure that this is correct. In fact, it is wrong! Increasing the number of simulations to $n := 10^5$ results in much shorter confidence intervals, which no longer overlap and reveal that the probability that Serbia wins is actually higher. This illustrates how confidence intervals can help

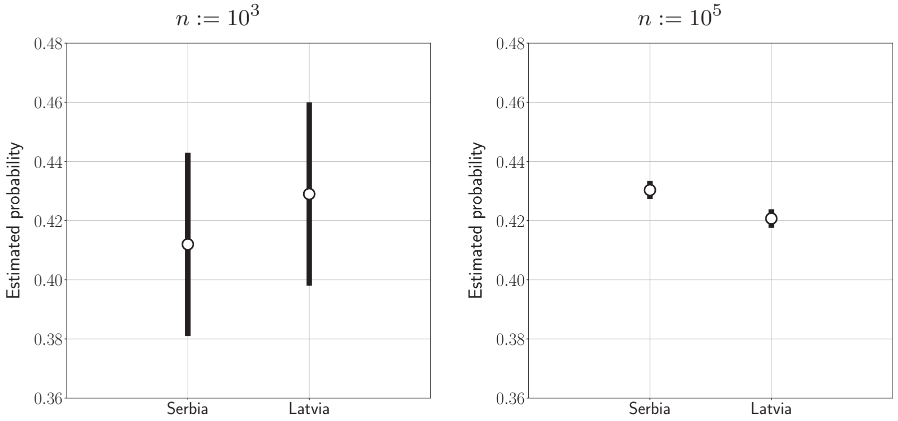


Figure 9.17 Confidence intervals for Monte Carlo simulations. Probability that Serbia and Latvia win the gold medal in the 3x3 basketball tournament at the Tokyo olympics, estimated via Monte Carlo simulation as described in Example 1.36. The graph shows the estimated probabilities and associated 0.95 confidence intervals for 10^3 (left) and 10^5 (right) simulations. On the left, the confidence intervals overlap, so even though the probability of Latvia winning is higher (compare the white markers), the result is inconclusive. On the right, Serbia's probability is higher and the confidence intervals do not overlap, indicating that we have performed enough simulations to differentiate between the two probabilities.

us decide whether we have performed enough Monte Carlo simulations to be confident of our conclusions.

9.8.3 Interpretation And Applicability Of Confidence Intervals

Interpreting the meaning of confidence intervals is somewhat tricky. Imagine that we have computed the 0.95 confidence interval $[174.6, 177.4]$ for the population mean of the height data in Figure 9.16. We might be tempted to state:

The probability that the mean height in the population is between 174.6 cms and 177.4 cms is 0.95.

However, the population mean is a deterministic quantity, so there are no random quantities in this statement! In our height example, we actually know the value of the population mean: it equals 175.6 cms. It clearly does not make any sense to state that 175.6 is in $[174.6, 177.4]$ with probability 0.95 (it always is!). The correct probabilistic statement describing the confidence interval should be:

If I were to repeat the sampling process many times, then the population mean *would belong to a confidence interval like this one 95% of the time.*

Equivalently, if we compute many 0.95 confidence intervals, we can expect 95% of them to actually contain the corresponding population parameter. Figure 9.16 illustrates this. Out of 40 intervals, 38 (95%) contain the population mean.

Confidence intervals are challenging to interpret because they involve deterministic population parameters. In statistics, this is known as the frequentist framework, in contrast to the Bayesian framework described in Section 6.7, where the parameters are modeled as random variables. The following example compares the two perspectives.

Example 9.47 (Poll in Pennsylvania: Confidence interval)**.** We consider the data reported in Example 6.19 from a poll in Pennsylvania for the 2020 US election, where 281 people intend to vote for Trump and 300 for Biden. The corresponding estimate for the fraction of people θ_{pop} that intend to vote for Trump is 281/581 = 48.4%. We assume that the people in the poll were selected independently and uniformly at random from the general population. Consequently, we can interpret the estimate as the sample proportion of i.i.d. Bernoulli random variables with parameter θ_{pop}. Applying Definition 9.44 to these variables yields the 0.95 confidence interval

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[0.484 - \frac{0.98}{\sqrt{581}}, 0.484 + \frac{0.98}{\sqrt{581}}\right] = [0.443, 0.524]. \tag{9.206}$$

Unfortunately, within the frequentist framework, we cannot convert the confidence interval to a statement about the probability that either of the candidates wins, because we are modeling θ_{pop} as being deterministic. All we can say is that we cannot rule out that Trump will win. In particular, as explained above, we *cannot* conclude that the confidence interval contains θ_{pop} with probability 0.95, because θ_{pop} is a constant, so such a statement doesn't make sense. From the frequentist viewpoint, we can only state that if we repeat the polling process over and over, the resulting confidence interval will contain the parameter of interest 95% of the time.

In contrast, the Bayesian analysis in Example 6.19 interprets the fraction of Trump voters as a random variable $\tilde{\theta}$, and is therefore able to generate probabilistic statements such as: *the probability that Biden wins in Pennsylvania is 0.75.* Remember, however, that there is a price to pay: these statements completely depend on the prior distribution chosen for $\tilde{\theta}$, and can change substantially for different priors, as demonstrated in Figure 6.13.

....

When computing confidence intervals on real data, it is easy to forget that we are implicitly modeling the measurements as *independent*. In some situations, such as the Monte Carlo simulations in Example 9.46, it is straightforward to ensure that the independence assumption holds. However, in many others it is not. In Example 9.47 we assume that the participants in the election poll are chosen independently at random from the whole population of Pennsylvania, but in practice this is very difficult to achieve. For instance, young urban voters are typically much easier to reach than old rural voters. Correcting for this is a fundamental challenge when predicting elections. The following example computes confidence intervals on data that are not independent, with catastrophic results.

Example 9.48 (Confidence intervals require independent sampling)**.** We consider the problem of estimating the probability of precipitation in Coos Bay (Oregon) in 2015 using 500 hourly measurements from Dataset 9, each indicating the presence or absence of precipitation (this is the same data used in Example 4.34). We compare two strategies: (1) using 500 successive measurements and (2) using 500 random independent measurements sampled with replacement. We compute 0.95 confidence intervals for all estimates based on Definition 9.44. We consider the *true* probability of precipitation to be 0.113, because precipitation occurs in 11.3% of the total hourly measurements over the whole year.

Figure 9.18 shows multiple estimates, along with the associated 0.95 confidence intervals, obtained with the two strategies. The estimates computed from the successive samples are not very accurate. More worryingly, very few of the corresponding confidence intervals actually contain the true probability. They are not valid confidence intervals! The reason is that the measurements are not independent, which completely violates the assumptions of Definition 9.44. In contrast, the independent samples result in valid confidence intervals that contain the true probability of precipitation with high probability.

In this case, it is quite obvious why lack of independence is problematic: precipitation tends to concentrate at certain times of the year. Over the summer it rains very little, so the probability estimated from successive samples is very small. Conversely, in the winter there are periods of intense precipitation when the estimated probability is very large. Independent random sampling is unaffected by this temporal structure, because it is unlikely to select measurements that are concentrated in time. Unfortunately, in practice, it is often difficult to implement truly independent sampling, and dependence can be much more challenging to spot than in this example.

9.9 The Bootstrap

The bootstrap is a computational approach to quantify the uncertainty of parameter estimates obtained from random samples. The key idea is to simulate additional random samples by resampling the available data. Section 9.9.1 explains how to use the bootstrap to estimate the standard error of an estimate, and then build confidence intervals for estimators that are approximately Gaussian. Section 9.9.2 describes bootstrap percentile confidence intervals, which can be applied to estimators that are approximately Gaussian after a monotonic transformation.

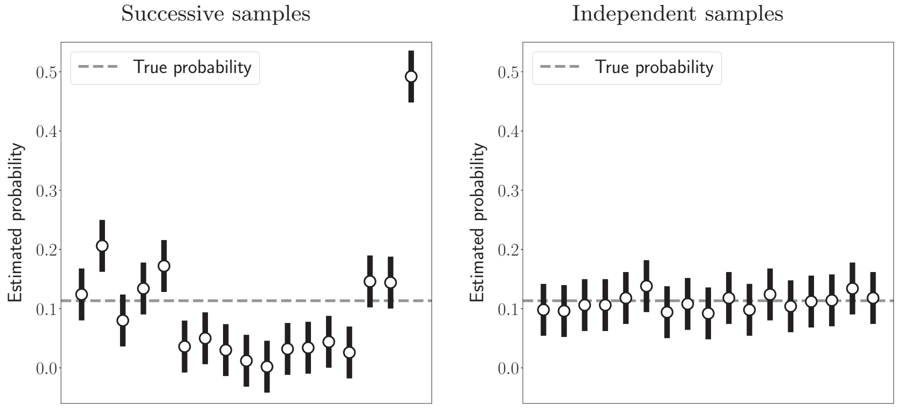


Figure 9.18 Confidence intervals require independent sampling. Estimates of the probability of precipitation in Coos Bay (Oregon) in 2015. Each estimate is equal to the fraction of 500 hourly measurements that report precipitation. The associated confidence intervals are computed based on Definition 9.44. On the left, the 500 measurements used to compute the estimate are sampled successively, which violates the independence assumption used to derive the confidence intervals. As a result, very few of the intervals contain the true probability (gray dashed line). On the right, the 500 measurements are sampled independently with replacement, which results in correct confidence intervals that do contain the true probability.

9.9.1 The Bootstrap Standard Error

As explained in Section 9.3, the standard error quantifies the fluctuations of estimators around their mean. In some cases, we can derive expressions for the standard error analytically; Theorem 9.13 provides one for the sample mean. However, this may not be possible for more complicated estimators. Here we present a *computational* approach to obtain standard errors.

Let $X := \{x_1, \ldots, x_n\}$ denote n samples selected independently and uniformly at random from a certain population, and let $h(x_1, \ldots, x_n)$ denote an estimator of a certain population parameter that we want to approximate. The standard error is the standard deviation of the estimator *with respect to the sampling process*, as captured by Definition 9.9. More formally, if we represent the data by the random variables $\tilde{x}_1$, ..., $\tilde{x}_n$ following Definition 9.3, and define the random variable $\tilde{w} := h(\tilde{x}_1, \ldots, \tilde{x}_n)$ to represent the estimator, then the standard error is the standard deviation of $\tilde{w}$.

In order to estimate the standard error computationally, a tempting idea is to use the sample standard deviation of several independent realizations of $\tilde{w}$. By Theorem 9.25 this estimate converges to the true standard error as the number of realizations tends to infinity (as long the variance and fourth central moment of $\tilde{w}$ are finite). The problem is that in order to obtain independent realizations of $\tilde{w}$, we need additional independent samples from the population, which we don't have! We need to estimate the standard error from the n available data points, which correspond to a single realization of $\tilde{w}$. Bootstrapping takes a pragmatic approach to this problem: since we cannot obtain more samples from the population, *how about we simulate more samples by resampling from the available data?*

Definition 9.49 (Bootstrap samples)**.** *To obtain bootstrap samples from a real-valued dataset $X := \{x_1, \ldots, x_n\}$, we first produce bootstrap indices $\tilde{k}_1$, $\tilde{k}_2$, ..., $\tilde{k}_n$ by sampling independently and uniformly at random with replacement from the set of possible indices $\{1, ..., n\}$. These indices are independent and satisfy*

$$\mathrm{P}\left(\tilde{k}_j = i\right) = \frac{1}{n}, \qquad 1 \leq i, j \leq n. \tag{9.207}$$

The bootstrap indices are then used to select bootstrap samples $\tilde{b}_1$, ..., $\tilde{b}_n$ by setting

$$\tilde{b}_j = x_{\tilde{k}_j}, \qquad 1 \leq j \leq n. \tag{9.208}$$

Asymptotically, as $n \to \infty$, bootstrapping is equivalent to sampling from the whole population. For finite n, the hope is that the available dataset X is still somewhat representative of the population. In that case, we can use the bootstrap estimator

$$\tilde{w}_{\mathrm{bs}}(X) := h(\tilde{b}_1, \ldots, \tilde{b}_n) \tag{9.209}$$

as a proxy for $\tilde{w}$ and approximate the standard error by computing the standard deviation of $\tilde{w}_{\mathrm{bs}}(X)$.

Definition 9.50 (Bootstrap standard error)**.** *Let $X := \{x_1, \ldots, x_n\}$ be a real-valued dataset and let $h : \mathbb{R}^n \to \mathbb{R}$ be an estimator of a parameter of interest. The bootstrap standard error of h equals*

$$\mathrm{se}_{\mathrm{bs}} = \sqrt{\mathrm{Var}\left[h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n)\right]}, \tag{9.210}$$

where $\tilde{b}_1$, ..., $\tilde{b}_n$ are bootstrap samples of X following Definition 9.49.

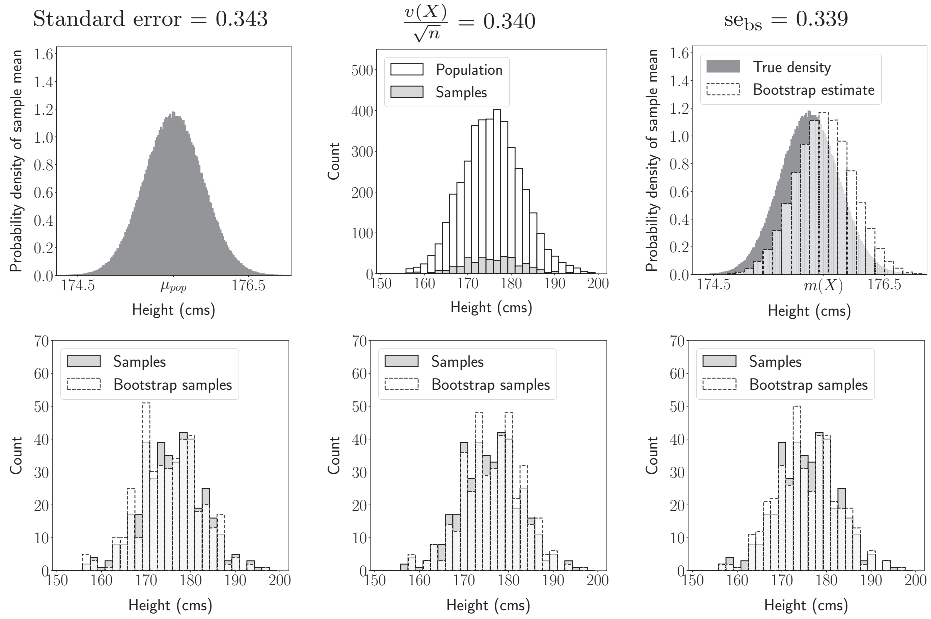
In practice, the bootstrap standard error is approximated via the Monte Carlo method, by generating K batches of n bootstrap samples, $b_j^{[k]}$, $1 \leq j \leq n$, $1 \leq k \leq K$, and compute the sample standard deviation $\sqrt{v(W)}$ of the parameter estimates

$$W := \{w_1, w_2, \ldots, w_K\}, \qquad w_k := h(b_1^{[k]}, b_2^{[k]}, \ldots, b_n^{[k]}). \tag{9.211}$$

As $K \to \infty$, $\sqrt{v(W)}$ converges to $\mathrm{se}_{\mathrm{bs}}$ by Theorem 9.25.

Figure 9.19 illustrates the computation of the bootstrap standard error for the sample mean

$$\widetilde{m}_{\mathrm{bs}} := \frac{1}{n}\sum_{k=1}^{n} \tilde{b}_k. \tag{9.212}$$



Standard error = 0.343

$\frac{v(X)}{\sqrt{n}} = 0.340$

$\mathrm{se}_{\mathrm{bs}} = 0.339$

Figure 9.19 Bootstrap standard error of the sample mean. The top left graph shows the normalized histogram of 10^6 independent instances of the sample mean of the height data in Example 9.1 computed from 400 independent, uniform samples. The corresponding standard deviation provides a ground-truth value for the standard error of the sample mean. The top middle graph shows a dataset X of 400 samples, which can be used to estimate the standard error by plugging in the sample standard deviation into the formula provided by Theorem 9.13. Alternatively, we can generate bootstrap samples from the 400 samples, as illustrated in the bottom row, and compute the bootstrap standard error following Definition 9.50. The top right graph shows a normalized histogram of 10^5 bootstrap sample means, and compares it to the true density of sample means (which cannot be computed from a single dataset of 400 samples). The bootstrap standard error is the standard deviation of the histogram. As predicted by Theorem 9.51, the two estimates of the standard error are almost the same.

Notice that the bootstrap distribution of the sample mean is centered at the sample mean $m(X)$ of the data, not at the population mean like the true distribution. This follows from Theorem 9.6, because $m(X)$ is the population mean of the dataset X.

The standard error of the sample mean can be estimated via the theoretical formula in Theorem 9.13, by plugging in the sample variance as an estimate of the population variance:

$$\frac{\sigma_{\mathrm{pop}}}{\sqrt{n}} \approx \sqrt{\frac{v(X)}{n}}. \tag{9.213}$$

The following theorem shows that the bootstrap standard error produces a very similar estimate, without using the analytical formula. In fact, the two estimates of the standard error are exactly the same if we instead use the biased estimate $\frac{1}{n}\sum_{j=1}^{n}(x_j - m(X))^2$ to approximate the population variance in (9.213).

Theorem 9.51 (Bootstrap standard error of the sample mean)**.** *Let* $X := \{x_1, \ldots, x_n\}$ *be a real-valued dataset and let* $\tilde{b}_1, \ldots, \tilde{b}_n$ *be bootstrap samples of* X *following Definition 9.49. The bootstrap standard error of the sample mean of* X *equals*

$$\text{se}_{\text{bs}} = \sqrt{\frac{n-1}{n^2} v(X)}, \tag{9.214}$$

where $v(X)$ *is the sample variance of* X.

Proof Interpreting X as a population, by Theorem 9.13 the variance of $\widetilde{m}_{\text{bs}}$ is equal to the population variance of X, which equals $\frac{1}{n}\sum_{j=1}^{n}(x_j - m(X))^2$, divided by n:

$$\text{se}_{\text{bs}}^2 := \text{Var}\left[\widetilde{m}_{\text{bs}}\right] = \frac{1}{n^2}\sum_{j=1}^{n}(x_j - m(X))^2 \tag{9.215}$$

$$= \frac{n-1}{n^2} v(X). \tag{9.216}$$

■

The bootstrap standard error enables us to build confidence intervals for any parameter estimate with a distribution that is approximately Gaussian. By Lemma 9.42, if the distribution of the parameter estimate $\tilde{w}$ is well approximated by a Gaussian with a mean equal to the population parameter γ, then $[\tilde{w} - c_\alpha \sigma, \tilde{w} + c_\alpha \sigma]$ is a 1-α confidence interval for γ, where σ is the standard error of $\tilde{w}$ and c_α is an appropriately-chosen constant. Plugging in the bootstrap standard error as an estimate for σ yields the bootstrap Gaussian confidence interval.

Definition 9.52 (Bootstrap Gaussian confidence interval)**.** *Let* $X := \{x_1, \ldots, x_n\}$ *be a real-valued dataset, and let* $h(X)$ *denote an estimator of a parameter* γ *computed from the elements of* X. *For any* $\alpha \in (0, 1)$, *the 1-*α *bootstrap Gaussian confidence interval for* γ *is*

$$\mathcal{I}_{1-\alpha}^{\text{BSG}} := [h(X) - c_\alpha \,\text{se}_{\text{bs}}, h(X) + c_\alpha \,\text{se}_{\text{bs}}], \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where se$_{\text{bs}}$ *is the bootstrap standard error of* h *and* $F_{\tilde{z}}$ *denotes the cdf of a standard Gaussian with zero mean and unit variance.*

Definition 9.52 can be used to build confidence intervals for any estimator, without the need to derive the corresponding standard error analytically. However, the resulting confidence intervals may not be very accurate if the distribution of the parameter estimate is not approximately Gaussian, as illustrated by the following example.

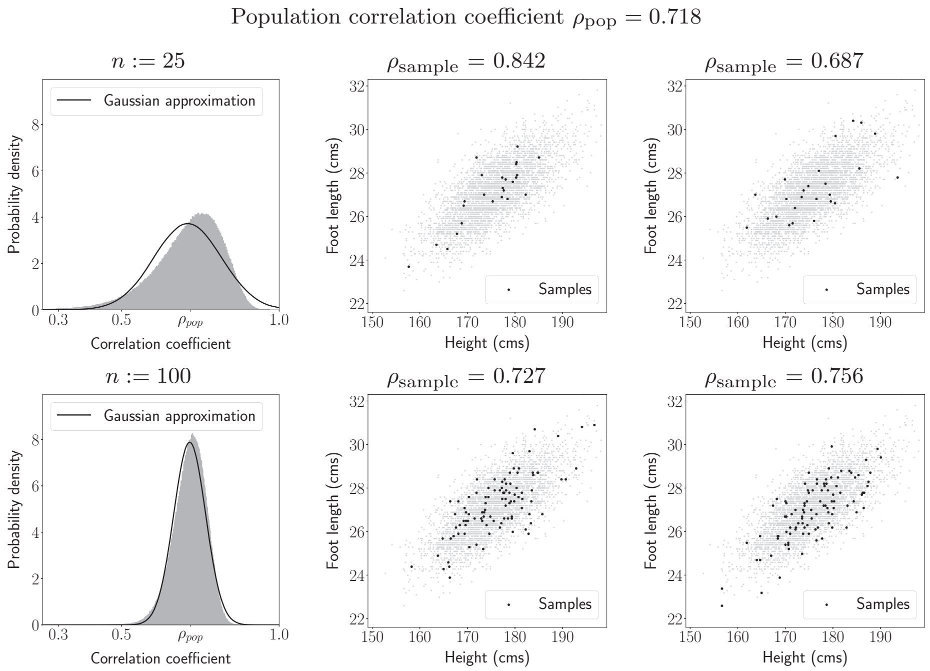


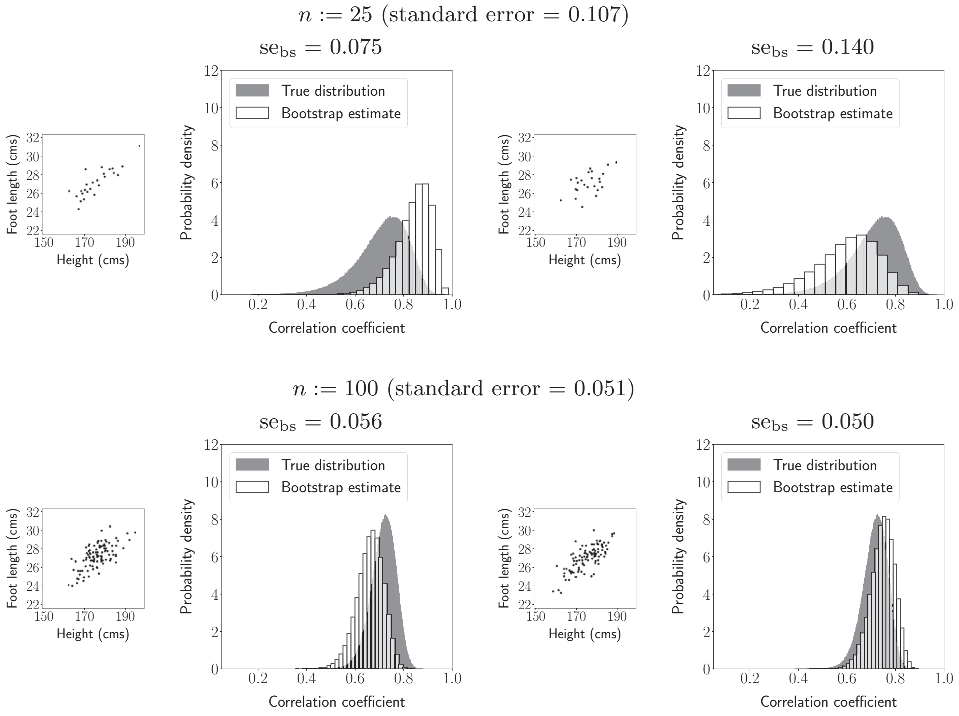
Figure 9.20 Distribution of the sample correlation coefficient. The left column shows the normalized histogram of the sample correlation coefficient between height and foot length for the dataset in Example 9.53. The histogram is computed using 10^6 instances of the sample correlation coefficient computed from 25 (top) and 100 (bottom) independent, uniform samples with replacement. The black line shows a Gaussian fit to the histogram, which approximates the distribution well for $n := 100$, but not for $n := 25$. The two right columns show example scatterplots of some of the samples (black), as well as the underlying population (light gray). The corresponding sample correlation coefficient is reported above each plot.

Example 9.53 (Confidence interval for the correlation coefficient)**.** We consider the problem of estimating the correlation coefficient between height and foot length within a population. We use the data in Figure 5.18, extracted from Dataset 5, as the complete ground-truth population. The population correlation coefficient equals

$$\rho_{\text{pop}} := \frac{\text{Cov}_{\text{pop}}(\text{height}, \text{foot})}{\sigma_{\text{pop}}(\text{height})\sigma_{\text{pop}}(\text{foot})} = 0.718. \tag{9.217}$$

Here, $\sigma_{\text{pop}}(\text{height})^2$ and $\sigma_{\text{pop}}(\text{foot})^2$ denote the population variance of height and foot length, as defined by (9.19), and

$$\text{Cov}_{\text{pop}}(\text{height}, \text{foot}) := \frac{1}{N}\sum_{k=1}^{N}(h_i - \mu_{\text{pop}}(\text{height}))(l_i - \mu_{\text{pop}}(\text{foot})), \tag{9.218}$$



$n := 25$ (standard error $= 0.107$)

$\mathrm{se}_{\mathrm{bs}} = 0.075$

$\mathrm{se}_{\mathrm{bs}} = 0.140$

$n := 100$ (standard error $= 0.051$)

$\mathrm{se}_{\mathrm{bs}} = 0.056$

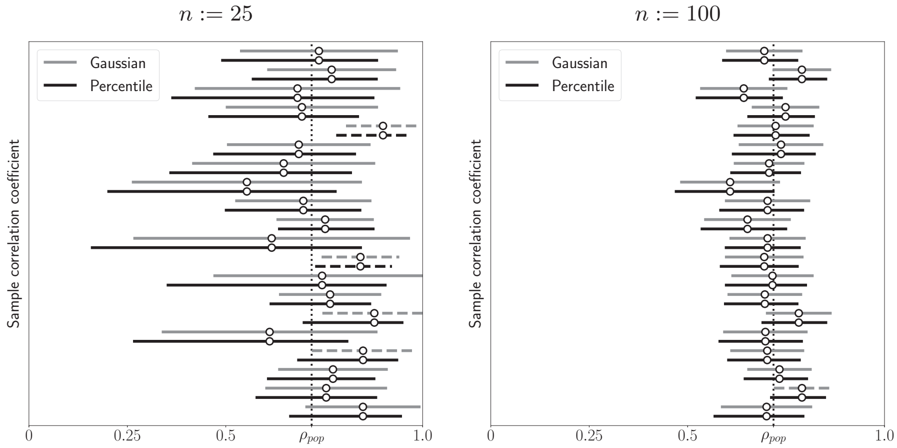
$\mathrm{se}_{\mathrm{bs}} = 0.050$

Figure 9.21 Bootstrap standard error of the sample correlation coefficient. The scatterplots depict n samples of height and foot length from the dataset in Example 9.53 for $n := 25$ (top row) and $n := 100$ (bottom row). The graph to the right of each scatterplot compares the probability density of the sample correlation coefficient (approximated as explained in the caption of Figure 9.20), with a normalized histogram of 10^5 sample correlation coefficients computed from n bootstrap samples following Definition 9.49. The standard error is the standard deviation of the true distribution; the bootstrap standard error is the standard deviation of the bootstrap distribution.

is the population covariance, where $(h_1, l_1), (h_2, l_2), \ldots, (h_N, l_N)$ denote the height and foot length of the $N := 4{,}082$ individuals in the population, and $\mu_{\mathrm{pop}}(\mathrm{height})$ and $\mu_{\mathrm{pop}}(\mathrm{foot})$ denote the population mean of height and foot length respectively.

We estimate the correlation coefficient by computing the sample correlation coefficient (see Definition 8.10) from the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ corresponding to n individuals selected independently and uniformly at random with replacement from the population. Figure 9.20 shows scatterplots of the samples, and plots of the distribution of the sample correlation coefficient. For $n := 100$, the distribution is reasonably well approximated as Gaussian (although it is somewhat skewed), but for $n := 25$ the Gaussian approximation is not very accurate.

In order to quantify the uncertainty of our estimate from just n samples, we build bootstrap Gaussian confidence intervals following Definition 9.52. Fig-



Interval	Coverage % (out of 10^4)	Average length
Gaussian	90.7	0.403
Percentile	92.8	0.399

Interval	Coverage % (out of 10^4)	Average length
Gaussian	93.7	0.196
Percentile	94.3	0.195

Figure 9.22 Bootstrap confidence intervals. Bootstrap Gaussian and percentile confidence intervals computed following Definitions 9.52 and 9.55, respectively, for the correlation coefficient of the data in Example 9.53. The graph shows 20 examples of each type of interval for $n := 25$ and $n := 100$. The table reports the coverage probability (how often the intervals contain the population correlation coefficient) and the average length of 10^4 intervals. The percentile intervals have a higher coverage probability and smaller average length, because they are able to automatically adapt to the skewness of the distribution of the sample correlation coefficient.

ure 9.21 shows the distribution of the bootstrap samples for different sets of n data. By Definition 9.50, the bootstrap standard error is equal to the standard deviation of the distribution, approximated using a large number of bootstrap samples. Figure 9.22 shows 0.95 bootstrap Gaussian confidence intervals ($\alpha := 0.05$) based on this estimate of the standard error. Since we have access to the whole population, we can compute the true coverage probability of the intervals, which is the fraction that actually contain the population correlation coefficient. For $n := 100$, it equals 93.7%, but for $n := 25$, it is only 90.7%. The problem is that the Gaussian interval does not account for the skew of the distribution of the sample correlation coefficient. In the following section, we discuss how to address this using a different type of confidence intervals based on the bootstrap.

. . . .

9.9.2 The Bootstrap Percentile Confidence Interval

In this section, we show that the bootstrap can be used to build confidence intervals directly, without having to approximate the standard error of the estimator of interest. This alternative approach has the important advantage that it provides valid confidence intervals for estimators that do not have Gaussian distributions, as long as they can be rendered approximately Gaussian by a monotonic transformation. This is the case, for instance, for the sample correlation coefficient. To motivate the approach, we first focus on the sample mean.

A sample mean computed from random samples has a distribution that is approximately Gaussian (see Definition 9.38) centered at the population mean (see Theorem 9.6). For a fixed dataset of samples X, the bootstrap distribution of the sample mean (i.e. the distribution of sample means computed from bootstrap samples of X following Definition 9.49) is also approximately Gaussian, but is instead centered at the sample mean of X. This follows again from Definition 9.38 and Theorem 9.6, interpreting X as the population. Crucially, these two Gaussian distributions have approximately the same standard deviation, because by Theorem 9.51 the standard deviation of the bootstrap distribution is very close to the sample standard error, which converges to the true standard deviation of the sample mean by Theorem 9.25. This has a very important consequence: an interval containing the bootstrap samples with probability $1-\alpha$ *contains the population mean with probability* $1-\alpha$, as established in the following theorem. The result applies to any unbiased estimator $\tilde{g}$ with a Gaussian distribution centered at the corresponding population parameter γ, as long as the conditional distribution of the bootstrap samples given $\tilde{g}=g$ is Gaussian and centered at g.

Theorem 9.54 (Percentile confidence interval under Gaussian assumptions)**.** *Let $\tilde{g}$ be a Gaussian random variable with mean γ and variance σ^2, representing an estimator of γ. Let $\tilde{w}$ be a random variable representing a bootstrap approximation of the estimator. Conditioned on $\tilde{g}=g$, $\tilde{w}$ is Gaussian with mean g and variance σ^2. Let the $\alpha/2$ and $1-\alpha/2$ percentiles of $\tilde{w}$ conditioned on $\tilde{g}=g$ be the values $q_{\alpha/2}(g)$ and $q_{1-\alpha/2}(g)$ satisfying*

$$\mathrm{P}\left(\tilde{w} \leq q_{\alpha/2}(g) \,\middle|\, \tilde{g}=g\right) = \frac{\alpha}{2}, \qquad \mathrm{P}\left(\tilde{w} \leq q_{1-\alpha/2}(g) \,\middle|\, \tilde{g}=g\right) = 1-\frac{\alpha}{2}. \quad (9.219)$$

Then, the interval $[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})]$ is a 1-α confidence interval for γ, i.e.

$$\mathrm{P}\left(\gamma \in \left[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})\right]\right) = 1-\alpha. \quad (9.220)$$

Proof Let $c_\alpha := F_{\tilde{z}}^{-1}\left(1-\frac{\alpha}{2}\right)$ for a standard Gaussian random variable $\tilde{z}$ with zero mean and unit variance. By Theorem 3.32 and the assumed distribution of $\tilde{w}$ given $\tilde{g}=g$,

$$\mathrm{P}\left(\tilde{w} \leq g + c_\alpha \sigma \,|\, \tilde{g}=g\right) = \mathrm{P}\left(\frac{\tilde{w}-g}{\sigma} \leq c_\alpha \,\middle|\, \tilde{g}=g\right) \quad (9.221)$$

$$= F_{\tilde{z}}(c_\alpha) = 1-\frac{\alpha}{2} \quad (9.222)$$

and by symmetry of the Gaussian pdf, $\mathrm{P}(\tilde{w} \leq g - c_\alpha\sigma \,|\, \tilde{g}=g) = \alpha/2$. Therefore,

$$q_{\alpha/2}(g) = g - c_\alpha\sigma, \qquad q_{1-\alpha/2}(g) = g + c_\alpha\sigma, \quad (9.223)$$

so

$$[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})] = [\tilde{g} - c_\alpha \sigma, \tilde{g} + c_\alpha \sigma]\,, \tag{9.224}$$

which is a 1-α confidence interval for γ by Lemma 9.42. ■

Theorem 9.54 suggests a very simple procedure to build $1-\alpha$ confidence intervals: pick an interval such that the fraction of bootstrap samples it contains is $1-\alpha$. This is typically achieved by determining the appropriate percentiles of the bootstrap distribution (as in the proof of Theorem 9.54), so the resulting interval is known as a bootstrap percentile confidence interval.

Definition 9.55 (Bootstrap percentile confidence interval)**.** *Let $X := \{x_1, \ldots, x_n\}$ be a real-valued dataset, and let $h(X)$ denote an estimator of a parameter γ computed from the elements of X. For any $\alpha \in (0,1)$, we define the $\alpha/2$ and $1-\alpha/2$ bootstrap percentiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ as the values satisfying*

$$\mathrm{P}\left(h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n) \leq q_{\alpha/2}\right) = \frac{\alpha}{2}, \tag{9.225}$$

$$\mathrm{P}\left(h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n) \leq q_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}, \tag{9.226}$$

where $\tilde{b}_1$, …, $\tilde{b}_n$ are bootstrap samples of X following Definition 9.49. The 1-α bootstrap percentile confidence interval for γ is

$$\mathcal{I}_{1-\alpha}^{\text{BSP}} := [q_{\alpha/2}, q_{1-\alpha/2}]. \tag{9.227}$$

In practice, the bootstrap percentiles are approximated via the Monte Carlo method, by generating K batches of n bootstrap samples, $b_i^{[k]}$, $1 \leq k \leq K$, and setting $q_{\alpha/2}$ and $q_{1-\alpha/2}$ to be the $\alpha/2$ and $1-\alpha/2$ percentiles of the parameter estimates

$$W := \{w_1, w_2, \ldots, w_K\}\,, \qquad w_k := h(b_1^{[k]}, b_2^{[k]}, \ldots, b_n^{[k]}). \tag{9.228}$$

Bootstrap percentile confidence intervals can be applied to estimators that are not approximately Gaussian, but are rendered approximately Gaussian by a monotonic transformation. If such a transformation exists, then the percentile confidence interval is a valid confidence interval, as long as (1) the transformed Gaussian distribution is centered at the transformed population parameter of interest, and (2) the transformed bootstrap samples have a Gaussian distribution centered at the transformed parameter estimate. Notice that the transformation must exist, but we *don't actually need to know it* in order to build the interval!

Theorem 9.56 (Percentile confidence interval and monotonic transformations)**.** *Let $\tilde{g}$ be a random variable such that $M(\tilde{g})$ is Gaussian with mean $M(\gamma)$ and variance σ^2, for a monotonic transformation $M : \mathbb{R} \to \mathbb{R}$. Let $\tilde{w}$ be a random variable such that conditioned on $\tilde{g} = g$, $M(\tilde{w})$ is Gaussian with mean $M(g)$ and variance σ^2. Let the $\alpha/2$ and $1-\alpha/2$ percentiles of $\tilde{w}$ conditioned on $\tilde{g} = g$ be the values $q_{\alpha/2}(g)$ and $q_{1-\alpha/2}(g)$ satisfying*

$$\mathrm{P}\left(\tilde{w} \leq q_{\alpha/2}(g) \,|\, \tilde{g} = g\right) = \frac{\alpha}{2}, \qquad \mathrm{P}\left(\tilde{w} \leq q_{1-\alpha/2}(g) \,|\, \tilde{g} = g\right) = 1 - \frac{\alpha}{2}. \tag{9.229}$$

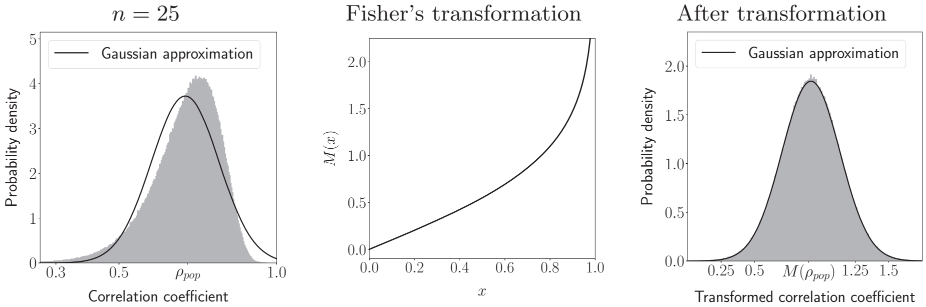


Figure 9.23 Fisher's transformation of the sample correlation coefficient. The plot on the left shows the normalized histogram of the sample correlation coefficient between height and foot length computed using 25 independent, uniform random samples from the dataset in Example 9.53. The distribution is skewed, and not well approximated as Gaussian. The plot in the middle shows Fisher's transformation M given by (9.235). The plot on the right shows the normalized histogram of sample correlation coefficients after applying the transformation: their distribution is very well approximated as a Gaussian centered at $M(\rho_{\text{pop}})$, where ρ_{pop} is the population correlation coefficient.

Then the interval $[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})]$ *is a 1-*α *confidence interval for* γ*, i.e.*

$$\mathrm{P}\left(\gamma \in \left[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})\right]\right) = 1 - \alpha. \tag{9.230}$$

Proof By the assumptions of the theorem, we can apply Theorem 9.54 to the transformed population parameter $\gamma' := M(\gamma)$ and the corresponding estimator $\tilde{g}' := M(\tilde{g})$, to obtain

$$\mathrm{P}\left(\gamma' \in \left[q_{\alpha/2}(\tilde{g}'), q_{1-\alpha/2}(\tilde{g}')\right]\right) = 1 - \alpha, \tag{9.231}$$

where $q_{\alpha/2}(\tilde{g}')$ and $q_{1-\alpha/2}(\tilde{g}')$ are the $\alpha/2$ and the $1 - \alpha/2$ percentiles of $M(\tilde{w})$ given $\tilde{g} = g$.

Here we leverage a key insight: quantiles are invariant to monotonic transformations. To simplify the exposition, let us assume that M is nondecreasing (the same argument can be easily adapted when M is nonincreasing). Then the events $\tilde{w} \leq q$ and $M(\tilde{w}) \leq M(q)$ are equivalent for any q by monotonicity of M, even if we condition on $\tilde{g} = g$ for any g. This implies that the $\alpha/2$ percentile of $M(\tilde{w})$ given $\tilde{g} = g$ is equal to $M\left(q_{\alpha/2}(g)\right)$, because

$$\mathrm{P}\left(M(\tilde{w}) \leq M(q_{\alpha/2}(g)) \,|\, \tilde{g} = g\right) = \mathrm{P}\left(\tilde{w} \leq q_{\alpha/2}(g) \,|\, \tilde{g} = g\right) = \frac{\alpha}{2}. \tag{9.232}$$

Consequently, $q_{\alpha/2}(\tilde{g}') = M\left(q_{\alpha/2}(\tilde{g})\right)$, and by the same argument $q_{1-\alpha/2}(\tilde{g}') = M\left(q_{1-\alpha/2}(\tilde{g})\right)$. Combined with (9.231) this yields

$$\mathrm{P}\left(M(\gamma) \in \left[M\left(q_{\alpha/2}(\tilde{g})\right), M\left(q_{1-\alpha/2}(\tilde{g})\right)\right]\right) = 1 - \alpha. \tag{9.233}$$

The monotonicity of M is again crucial, since it implies that the events

$$M\left(q_{\alpha/2}(\tilde{g})\right) \leq M(\gamma) \leq M\left(q_{1-\alpha/2}(\tilde{g})\right) \quad \text{and} \quad q_{\alpha/2}(\tilde{g}) \leq \gamma \leq q_{1-\alpha/2}(\tilde{g})$$

are equivalent. We conclude that

$$\mathrm{P}\left(\gamma \in \left[q_{\alpha/2}(\tilde{g}), q_{1-\alpha/2}(\tilde{g})\right]\right) = 1 - \alpha. \tag{9.234}$$

■

An important example of a monotonic transformation that renders the distribution of an estimator approximately Gaussian is Fisher's transformation,

$$M(\rho) := \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right), \tag{9.235}$$

which achieves this for the sample correlation coefficient (Fisher, 1915), as depicted in Figure 9.23. Figure 9.22 compares bootstrap percentile and Gaussian confidence intervals for the correlation coefficient of the height and foot length data in Example 9.53. The coverage probability of the percentile intervals is closer to $1-\alpha := 0.95$, even though their average length is smaller. The improvement is more noticeable for $n := 25$, when the Gaussian approximation of the estimator is less accurate (see Figure 9.20). This demonstrates that the percentile confidence intervals are able to automatically adapt to the skewed distribution of the sample correlation coefficient, as suggested by Theorem 9.56.