Part One

Estimators and Tests

1

Estimating Population Parameters

I don't know how long I stand there. I don't believe I've ever stood there mourning faithfully in a downpour, but statistically speaking it must have been spitting now and then, there must have been a bit of a drizzle once or twice.

(from *The Misfortunates*, Dimitri Verhulst, pp. 125–126)

A major goal in statistics is to make statements about populations or processes. Often, the interest is in specific parameters of the distributions or densities of the populations or processes under study. For instance, researchers in political science want to make statements about the proportion of a population that votes for a certain political party. Industrial engineers want to make statements about the proportion of defective smartphones produced by a production process. Bioscience engineers are interested in comparing the mean amounts of growth resulting from applying two or more different fertilizers. Economists are interested in income inequality and may want to compare the variance in income across different groups.

To be able to make such statements, the proportions, means, and variances under study need to be quantified. In statistical jargon, we say that these parameters need to be estimated. It is also important to quantify how reliable each of the estimates is, in order to judge the confidence we can have in any statement we make. This chapter discusses the properties of the most important sample statistics that are used to make statements about population and process means, proportions, and variances.

1.1 Introduction: Estimators Versus Estimates

In practice, population parameters such as μ, σ^2, π, and λ (see our book *Statistics with JMP: Graphs, Descriptive Statistics and Probability*) are rarely known. For example, if we study the arrival times of the customers of a bank, we know that the number of arrivals per unit of time often follows a Poisson[1] distribution. However, we do not know the exact value of the

[1] The Poisson distribution is commonly used for random variables representing a certain number of events per unit of time, per unit of length, per unit of volume, and so on. The Poisson distribution has one parameter λ, which is the average number of events per unit of time, per unit of length, per unit of volume, and so on. For more details, see *Statistics with JMP: Graphs, Descriptive Statistics and Probability*.

Statistics with JMP: Hypothesis Tests, ANOVA and Regression, First Edition. Peter Goos and David Meintrup.

Companion Website: http://www.wiley.com/go/goosandmeintrup/JMP

distribution's parameter λ. One way or another, we therefore need to estimate this parameter. This estimate will be based on a number of measurements or observations, $x_1, x_2, \ldots, x_n$, that we perform in the bank; in other words, on the sample data we collect.

The **estimate** for the unknown λ will be a function of the sample values $x_1, x_2, \ldots, x_n$; for example, the sample mean $\overline{x}$. Every researcher who faces the same problem, studying the arrival pattern of customers, will obtain different sample values, and thus a different sample mean and another estimate. The reason for this is that the number of arrivals in the bank in a given time interval is a random variable. We can express this explicitly by using uppercase letters X_1, $X_2, \ldots, X_n$ for the sample observations. The fact that each researcher obtains another estimate for λ can also be made more explicit by using a capital letter to denote the sample mean: $\overline{X}$. The sample mean is interpreted as a random variable, and then it is called an **estimator** instead of an estimate. In short, an estimate is always a real number, while an estimator is a random variable the value of which is not yet known.

The sample mean is, of course, only one of many possible functions of the sample observations $X_1, X_2, \ldots, X_n$, and thus only one of many possible estimators. Obviously, a researcher is not interested in an arbitrary function of the sample observations, but he wants to get a good idea of the unknown parameter. In other words, the researcher wishes to obtain an estimate that, on average, is equal to the unknown parameter, and that, ideally, is guaranteed to be close to the unknown parameter. Statisticians translate these requirements into "the estimator should be unbiased" and "the estimator should have a small variance". These requirements will be clarified in the next section.

1.2 Estimating a Mean Value

The requirements for a good estimator can best be illustrated by means of two simulation studies. The first study simulates data from a normally distributed population, while the second one simulates data from an exponentially distributed population.

1.2.1 The Mean of a Normally Distributed Population

We first assume that a normally distributed population with mean $\mu = 3000$ and standard deviation $\sigma = 100$ is studied by 1000 (fictitious) students. The students are unaware of the μ value and wish to estimate it. To this end, each of these students performs five measurements. A first option to estimate the unknown value μ is to calculate the sample mean. In this way, we obtain 1000 sample means, shown in the histogram in Figure 1.1, at the top left. The mean of these 1000 sample means is 2998.33, while the standard deviation is 43.38.

Another possibility to estimate the unknown μ is to calculate the median. For a normally distributed population, both the median and the expected value are equal to the parameter μ, so that this makes sense. Based on the samples that the students have gathered, the 1000 medians can also be calculated and displayed in a histogram. The resulting histogram is shown in Figure 1.1, at the top right[2]. The attentive reader will notice immediately that the second histogram is

[2] Outputs as in Figures 1.1 and 1.2 can be created in JMP with the "`Distribution`" option in the "`Analyze`" menu.

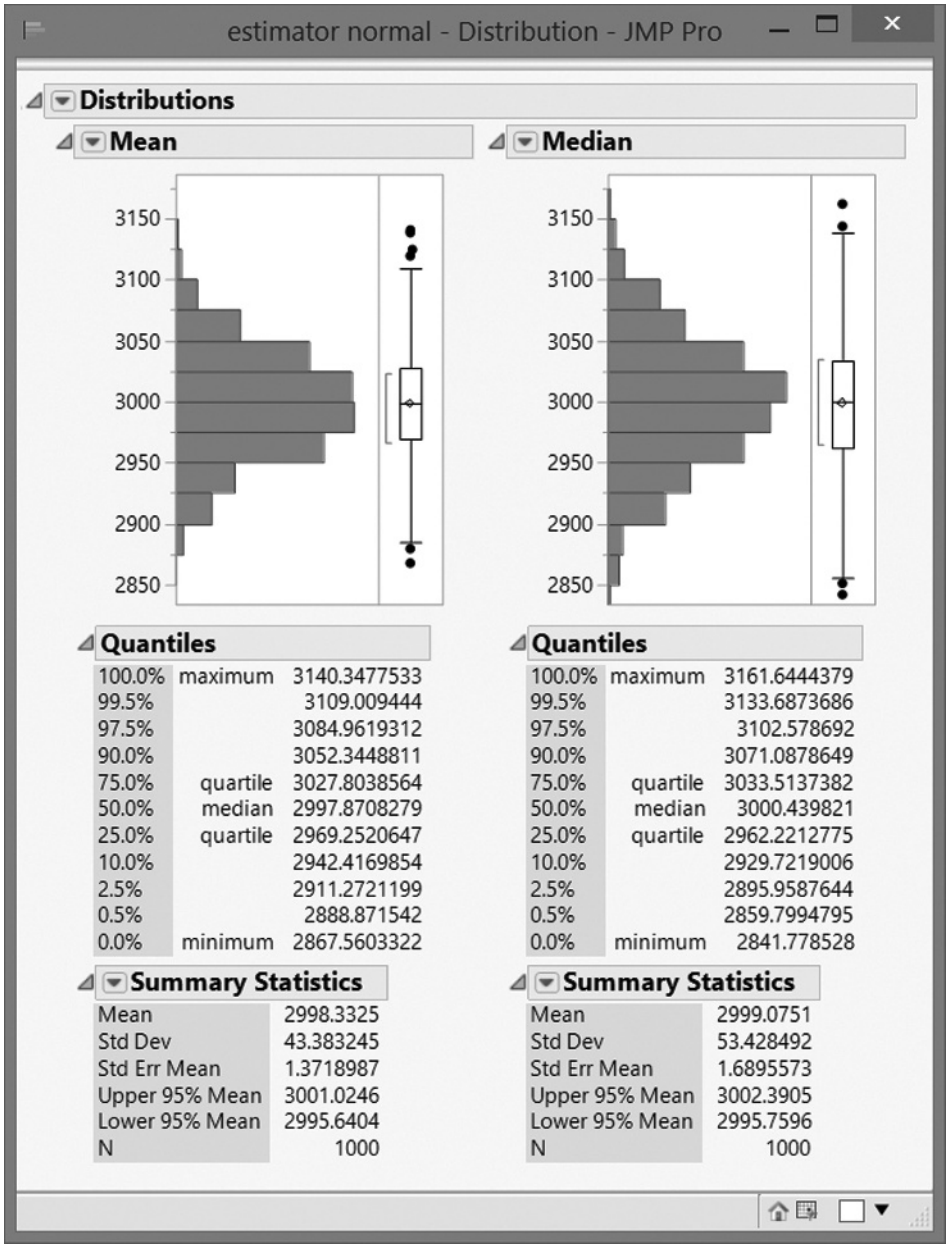


Figure 1.1 Histograms and descriptive statistics for 1000 sample means and medians calculated based on samples of five observations from a normally distributed population with mean 3000 and standard deviation 100.

just a bit wider than the first. Among other things, this is reflected by the fact that the standard deviation of the 1000 medians is 53.43. The mean of the 1000 medians is equal to 2999.08. In Figure 1.1, it can also be seen that the minimum (2841.78) and the first quartile (2962.22) of the sample medians are smaller than the minimum (2867.56) and the first quartile (2969.25) of the sample means. Also, the maximum (3161.64) and the third quartile (3033.51) of the sample medians are greater than the maximum (3140.35) and the third quartile (3027.80) of the sample means. This suggests that the sample medians are, in general, further away from the population mean $\mu = 3000$ than the sample means.

It is striking that both the mean of the 1000 sample means (2998.33) and that of the 1000 medians (2999.08) are very close to 3000. If the number of samples is raised significantly (theoretically, an infinite number of samples could be taken), the mean of the sample means and that of the sample medians will converge to the unknown $\mu = 3000$. Therefore, both the sample mean and the sample median are called **unbiased estimators** of the mean of a normally distributed population.

The fact that the range, the interquartile range, the standard deviation, and the variance of the 1000 sample means are smaller than those of the 1000 sample medians means that the sample mean is a more reliable estimator of the unknown population mean than the sample median. The larger variance of the medians indicates that the medians are generally further away from $\mu = 3000$ than the sample means. In short, a researcher should have more confidence in the sample mean because it is usually closer to the unknown μ. In such a case, we say that one estimator (here, the sample mean) is more **efficient** or **precise** than the other (here, the median).

1.2.2 The Mean of an Exponentially Distributed Population

We now investigate an exponentially distributed population with parameter $\lambda = 1/100$. The "unknown" population mean is therefore $\mu = 1/\lambda = 100$ (see *Statistics with JMP: Graphs, Descriptive Statistics and Probability*). Each of the 1000 fictitious students performs five measurements. A first option to estimate the unknown value μ is again to calculate the sample mean. A histogram of the 1000 sample means is shown in Figure 1.2, at the top left. The mean of these 1000 sample means is 99.2417, while the standard deviation is 44.10.

Based on the samples that the students have gathered, the 1000 medians can also be calculated and displayed in a histogram. This histogram is shown in Figure 1.2, at the top right. The mean of the 1000 medians is only 77.0114.

These calculations indicate that the population mean $\mu = 1/\lambda = 100$ can be approximated fairly well by using the sample means, with a mean of 99.2417. This is not the case for the medians, the mean value of which is far away from μ. This remains the case if the number of samples is increased. In this example, for an exponentially distributed population, the median is not an unbiased but a **biased** estimator of the population mean.

In addition, Figure 1.2 also shows that the standard deviation of the sample medians (46.13) is greater than that of the sample means (44.10).

1.3 Criteria for Estimators

Key properties of estimators are their expected values and their variances. These statistics are related to the concepts of bias and efficiency, respectively.

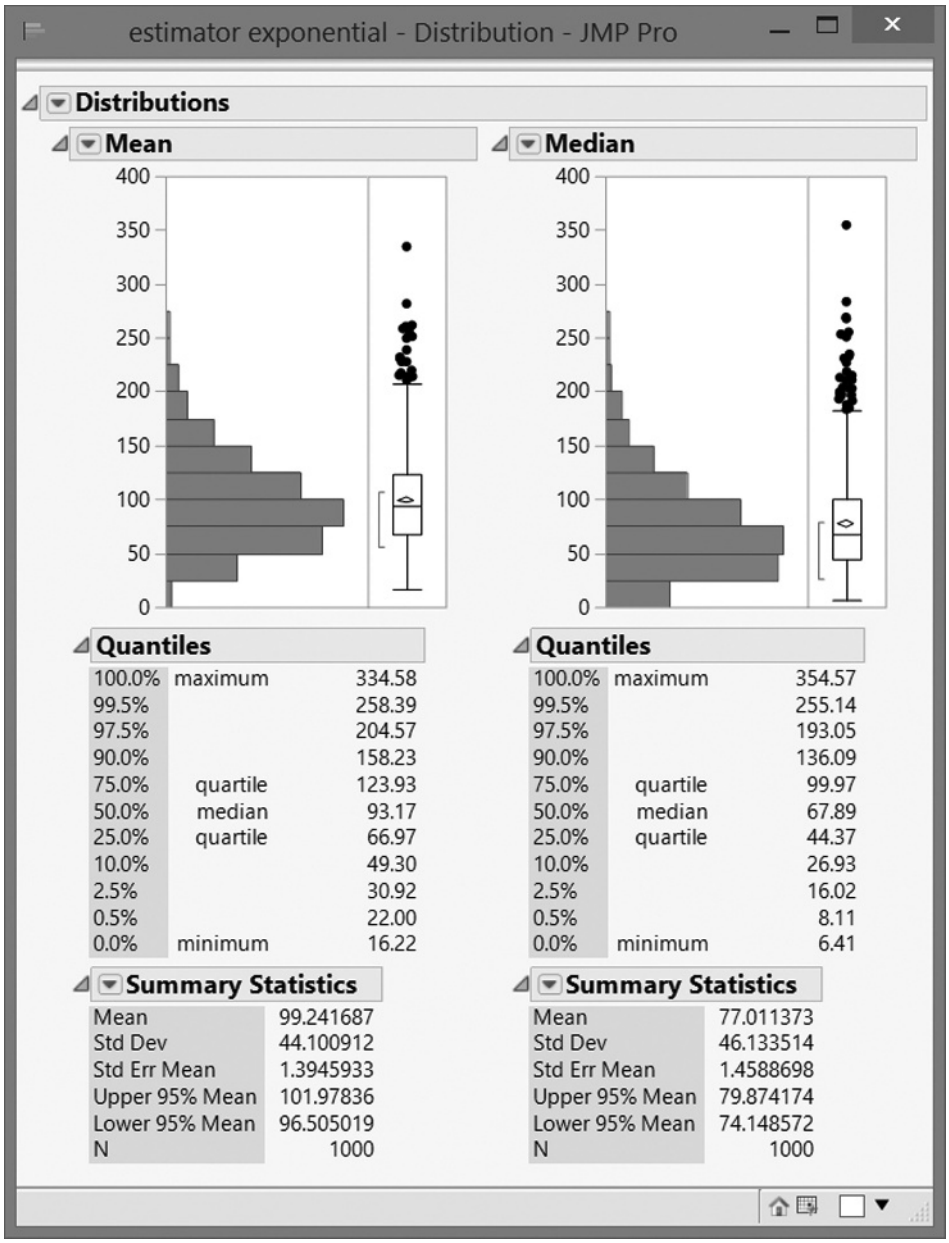


Figure 1.2 Histograms and descriptive statistics for 1000 sample means and sample medians calculated based on samples of five observations from an exponentially distributed population with parameter $\lambda = 1/100$.

1.3.1 Unbiased Estimators

An ideal estimator that always produces the exact value of an unknown population parameter does not exist. As illustrated in the above example, some estimators, namely unbiased estimators, are on average equal to the unknown population parameter, while others systematically under- or overestimate the parameter. The latter is an undesirable result for a researcher. Formally, the definition of an unbiased estimator $\hat{\theta}$ for an unknown population parameter θ is as follows:

Definition 1.3.1 *An estimator $\hat{\theta}$ of a population parameter θ is **unbiased** if*

$$E(\hat{\theta}) = \theta.$$

The **bias** of an estimator is the absolute difference $V(\hat{\theta}) = |E(\hat{\theta}) - \theta|$. An unbiased estimator has a bias of zero. For an unbiased estimator, the expected value is exactly equal to the population parameter. The histograms for the sample means on the left-hand sides of Figures 1.1 and 1.2 show that, once sample data is being used, the estimate will be close to the unknown population parameter, but not exactly equal to it. So, for any particular sample, even unbiased estimators result in estimates that differ from the population parameter that is being estimated.

Note that here the symbol $\hat{\theta}$ is used to denote an estimator of the unknown population parameter θ. As usual in statistics, we use Greek letters to denote unknown population parameters such as population means, population proportions, or population variances. If we want to estimate an unknown population parameter, we use an estimator, which is a synonym for an estimation method. In general in statistics, we indicate this using the symbol $\hat{\theta}$ (pronounced "theta hat"). We will mainly focus on three specific estimators, namely the sample mean, the sample proportion, and the sample variance. For historical reasons, the symbols $\overline{X}$, $\hat{P}$, and S^2 are used for these three estimators instead of $\hat{\mu}$, $\hat{\pi}$, and $\hat{\sigma}^2$.

The sample mean $\overline{X}$ is always an unbiased estimator of the population mean (this is proven in Theorem 1.5.1). Actually, this applies to all linear functions $Y = \sum_{i=1}^{n} \alpha_i X_i$ of sample observations for which $\sum_{i=1}^{n} \alpha_i = 1$, and the sample mean is a special case of such a linear combination, where each $\alpha_i = 1/n$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n.$$

It can be shown that, of all linear functions of $X_1, X_2, \ldots, X_n$, for which $\sum_{i=1}^{n} \alpha_i = 1$ the sample mean has the smallest variance[3]. In other words, the sample mean will usually provide an estimate that is closer to the population mean than any other linear function Y of $X_1, X_2, \ldots, X_n$.

In Theorem 1.7.1, we prove that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

[3] Therefore, the sample mean is called the "best linear unbiased estimator", abbreviated as "BLUE".

is an unbiased estimator of a population variance σ^2. This theorem also explains why we divide by $n - 1$ when computing the sample variance, and not by n. It is important to note that the sample standard deviation S is a biased estimator of the population standard deviation σ.

Finally, in Section 1.6, we will see that a sample proportion $\hat{P}$ is a special case of a sample mean. Its expected value is equal to the population proportion π, so that $\hat{P}$ is an unbiased estimator of π.

1.3.2 The Efficiency of an Estimator

It is desirable that an estimator is as reliable as possible and yields estimates that are close to the unknown population parameter under investigation. In short, the estimator should have a small variance or standard deviation. An estimator with a small variance is called an efficient estimator.

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same unknown population parameter θ, the relative efficiency of $\hat{\theta}_2$ compared to $\hat{\theta}_1$ is computed as $\text{var}(\hat{\theta}_1)/\text{var}(\hat{\theta}_2)$.

Sometimes, we have the choice between an estimator that is unbiased but has a large variance, and an estimator that is biased but has a small variance. In this case, it is not immediately clear which estimator should be used. To make a decision in such situations, one can pick the estimator that has the smaller mean squared error, $\text{MSE}(\hat{\theta})$:

Definition 1.3.2 *The **mean squared error** of an estimator $\hat{\theta}$ is the sum of its variance and the square of its bias:*

$$MSE(\hat{\theta}) = \text{var}(\hat{\theta}) + [V(\hat{\theta})]^2.$$

Finally, it is also desirable that the precision of an estimator increases with the number of observations. More observations provide more information, so that better estimates can be expected. For example, Theorem 1.5.2 shows that the variance of the sample mean is equal to σ^2/n. The variance decreases as the sample size n increases. The precision of the sample mean is thus improved when more data is used.

1.4 Methods for the Calculation of Estimators

Finding estimators with good properties is not always easy. In the statistical literature[4], three methods are frequently used:

(1) the method of moments;
(2) the method of least squares; and
(3) the maximum likelihood method.

These general methods are beyond the scope of this book. This book primarily focuses on the following estimators: sample means, sample proportions, and sample variances. In the remainder of this chapter, each of these estimators is shown to be unbiased, and the probability density of each estimator is discussed.

[4] See *Statistics with JMP: Linear and Generalized Linear Models*.

1.5 The Sample Mean

1.5.1 The Expected Value and the Variance

If the sample mean is considered as an estimator and thus as a random variable, we can determine the expected value, the variance, and even the probability density. The sample mean is then written using a capital letter,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

to indicate this explicitly. We consider the sample mean as an estimator or a random variable as long as we have no data; that is, the individual observations $X_1, X_2, \ldots, X_n$ are not known. Once the data has been collected, we use lowercase letters for the individual observations: $x_1, x_2, \ldots, x_n$. For the sample mean that we compute based on the observed values $x_1, x_2, \ldots, x_n$, we also use a lowercase letter:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Theorem 1.5.1 *For a random sample from a population with expected value μ, we have*

$$E(\overline{X}) = \mu.$$

Proof.

$$\begin{aligned} E(\overline{X}) &= E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big), \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i), \\ &= \frac{1}{n}(\mu + \mu + \cdots + \mu), \\ &= \frac{n\mu}{n}, \\ &= \mu. \end{aligned}$$

■

This theorem states that before sample data is obtained, the expected value of the sample mean is equal to the population mean. In other words, the theorem states that the sample mean is an unbiased estimator of the population mean.

Once we have sample observations $x_1, x_2, \ldots, x_n$, the sample mean is $\overline{x}$. Of course, this sample mean will not be exactly equal to μ. This was already illustrated in Section 1.2, where each student obtained a different sample mean. To get an idea of the size of the possible deviation between the sample mean $\overline{X}$ and the population mean μ, one should study the variance and standard deviation of $\overline{X}$. Figure 1.1 showed that the standard deviation of the sample means of

1000 (fictitious) students was equal to 43.38, while the original population of the individual values had a standard deviation of 100. In general, the standard deviation of a sample mean is lower than the standard deviation of the population studied. The same is true for the variance. The next theorem tells us how much smaller the variance of a sample mean is.

Theorem 1.5.2 *For a random sample of n observations from a population with variance σ^2, we have*

$$\sigma^2_{\overline{X}} = \text{var}(\overline{X}) = \frac{\sigma^2}{n}$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Proof.

$$\begin{aligned}
\sigma^2_{\overline{X}} = \text{var}(\overline{X}) &= \text{var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big), \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i), \\
&= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2), \\
&= \frac{n\sigma^2}{n^2}, \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

In the second step of this proof, it is assumed that the covariance between two different sample observations, X_i and X_j, is equal to zero. In that case, the variance of a linear combination of random variables is equal to a linear combination of the variances, with squared coefficients[5]. ■

[5] In *Statistics with JMP: Graphs, Descriptive Statistics and Probability*, we show that

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X, Y),$$

which can be simplified to

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y)$$

if the random variables X and Y are independent or uncorrelated (and thus $\text{cov}(X, Y) = 0$). This result can be generalized to scenarios involving more than two random variables.

The proof shows that the variance of the sample mean decreases linearly when the sample size n increases. This means that as the sample size increases, the probability that the sample mean $\overline{x}$ is close to (the unknown) μ increases as well.

The square root of the variance, namely $\sigma_{\overline{X}}$, is called the **standard error**. The estimated version of this statistic, namely $s/\sqrt{n}$, can be found in the reports of statistical packages. Figures 1.1 and 1.2 illustrate that the standard error[6] is also reported in JMP, namely as "Std Err Mean". It is not difficult to verify that the standard error in these figures is a factor of $\sqrt{n} = \sqrt{1000} = 31.62$ smaller than the corresponding standard deviation (abbreviated as "Std Dev").

1.5.2 The Probability Density of the Sample Mean for a Normally Distributed Population

If the sample is drawn from a normally distributed population, we can use the following theorem.

Theorem 1.5.3 *Let $X_1, X_2, \ldots, X_k$ be independent normally distributed random variables with expected values $E(X_1) = \mu_1, E(X_2) = \mu_2, \ldots, E(X_k) = \mu_k$ and variances $\text{var}(X_1) = \sigma_1^2, \text{var}(X_2) = \sigma_2^2, \ldots, \text{var}(X_k) = \sigma_k^2$. Then, the linear function $Y = \alpha_0 + \sum_{i=1}^{k} \alpha_i X_i$ is also normally distributed, with expected value $E(Y) = \alpha_0 + \sum_{i=1}^{k} \alpha_i \mu_i$ and variance $\text{var}(Y) = \sum_{i=1}^{k} \alpha_i^2 \sigma_i^2$.*

It follows from this theorem that the mean of a number of normally distributed random variables with the same mean μ and the same variance σ^2 is also normally distributed. Indeed, the mean of the variables $X_1, X_2, \ldots, X_n$ is a linear function with $\alpha_0 = 0$ and $\alpha_i = 1/n$ for $i \geq 1$. In this case, the sample mean $\overline{X}$ is normally distributed with mean μ and variance σ^2/n. We denote this by

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

This result is valid for any sample size – also for a small one – and is illustrated in Section 1.5.4.

1.5.3 The Probability Density of the Sample Mean for a Nonnormally Distributed Population

If the population under investigation has an unknown probability density or probability distribution, the probability distribution of the sample mean typically cannot be determined exactly.

[6] The standard error can be calculated in JMP using the menu "`Analyze`" and the option "`Distribution`". An alternative way to obtain the standard error for a particular variable in a data table is to use the "`Summary`" option in the "`Tables`" menu, and to choose "`Std Err`" in the "`Statistics`" drop-down menu that then becomes available.

In this case, however, a large sample size may help, because the central limit theorem can be used for large samples. One version of this theorem, namely Theorem 1.5.6, indeed indicates that the sample mean is approximately normally distributed for large n, with mean μ and variance σ^2/n.

The **central limit theorem** is one of the main theorems of statistics. This theorem also explains to a large extent why the normal probability density is so crucial in statistics. There are different versions of this theorem.

Theorem 1.5.4 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with expected values $E(X_i) = \mu_i$ and variances $\text{var}(X_i) = \sigma_i^2$. Then, under very general conditions and for a sufficiently large value of n:*

(1) the random variable $Y = \sum_{i=1}^{n} X_i$ is approximately normally distributed with mean $\mu_Y = \sum_{i=1}^{n} \mu_i$ and variance $\sigma_Y^2 = \text{var}(Y) = \sum_{i=1}^{n} \sigma_i^2$;
(2) and, consequently, the random variable

$$\frac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

approximately follows a standard normal distribution.

The general conditions mentioned in the theorem refer to the fact that none of the individual variances σ_i^2 makes a dominant contribution to the total variance of Y. In many practical applications of the central limit theorem, all random variables $X_1, X_2, \ldots, X_n$ have the same distribution or density, and therefore the same variance. In that case, this condition is automatically met. If all random variables $X_1, X_2, \ldots, X_n$ have the same distribution or density, the central limit theorem can be rewritten as follows:

Theorem 1.5.5 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with expected value $E(X_i) = \mu$ and variance $\text{var}(X_i) = \sigma^2$. Then, for a sufficiently large value of n:*

(1) the random variable $Y = \sum_{i=1}^{n} X_i$ is approximately normally distributed with mean $\mu_Y = n\mu$ and variance $\sigma_Y^2 = \text{var}(Y) = n\sigma^2$;
(2) and, consequently, the random variable

$$\frac{Y - n\mu}{\sigma\sqrt{n}}$$

approximately follows a standard normal distribution.

The central limit theorem can also be formulated in terms of the sample mean $\overline{X} = Y/n$:

Theorem 1.5.6 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with expected value $E(X_i) = \mu$ and variance $\text{var}(X_i) = \sigma^2$. Then, for a sufficiently large value of n:*

(1) the random variable $\overline{X} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$;
(2) and, consequently, the random variable

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approximately follows a standard normal distribution.

An important practical question is how large the sample size n must be before one can apply the central limit theorem. There is no general answer to this question. The required size of n depends on the distribution or density of the individual random variables X_i:

- If the probability density of X_i is similar to the normal density, $n = 5$ is sufficient.
- If the probability density of X_i does not show any pronounced peaks – such as, for example, the uniform density – then $n = 12$ should be sufficient.
- If the probability distribution or density of X_i shows pronounced peaks, it is difficult to specify a value of n. A value of $n = 100$ will usually suffice. An example of a distribution with a peak is $P(X = 1) = 0.06$ and $P(X = 10) = 1 - P(X = 1) = 0.94$.
- For continuous variables that appear in practice, typically $n = 25$ or $n = 30$ is sufficient.

In the next section, the third version of the central limit theorem (Theorem 1.5.6) is illustrated in detail, using simulations.

1.5.4 An Illustration of the Central Limit Theorem

Suppose that some students are interested in the value of the Euro Stoxx 50 index, which summarizes the performance of the 50 most important stocks inside the eurozone. Student 1 will take a sample of n observations of the Euro Stoxx 50 index and calculate the mean, namely $\overline{X}_1$. Student 2 will also take a sample of n observations. Since the Euro Stoxx 50 index changes from minute to minute, Student 2 will obviously observe different values of the Euro Stoxx 50 index (unless, by coincidence, the two students make their observations at exactly the same times). Student 2 also calculates the mean of his sample: $\overline{X}_2$. In the same way, all students collect n observations and calculate their sample means. If there are 200 students, we finally obtain 200 sample means: $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_{200}$.

The third version of the central limit theorem now states that these 200 means have a distribution that is very similar to that of the normal density. With a histogram of these 200 means, this is easy to verify.

This is exactly what we will do in this section. We will not use real students, but we will simulate the scenario sketched above in JMP. To this end, we will create 200 samples

of n observations (one for each hypothetical student), calculate the mean for each sample, and create a histogram of the 200 sample means. This simulation requires that we specify a probability distribution or probability density in JMP for the generation of the observations.

We start with a normal distribution. Hence, we first assume that the Euro Stoxx 50 index behaves like a normally distributed random variable. We use $\mu = 3000$ as the mean of this normal distribution (more or less the value of the index when work on this book was started in March 2014), and we choose $\sigma = 100$ as the standard deviation. We assume that all of the students' observations are independent of each other.

Normally Distributed *X*

First, suppose that each student collects a sample of five observations; in other words, that $n = 5$. In this scenario, we need to simulate 200 sets of five observations using JMP. To this end, we create a JMP data table with 200 rows and five columns, filled with pseudo-random numbers from a normal probability density with parameters $\mu = 3000$ and $\sigma = 100$. The formula we use for each of the five columns is "`Random Normal(3000, 100)`". We calculate the mean of the five observations in each row, and then display all means in a histogram. If we create a second data table in the same way, this corresponds to a second group of 200 hypothetical students who also collect samples of five observations. Two possible histograms obtained in this way are shown in Figure 1.3. The resulting histograms are quite bell-shaped, indicating that the sample means are normally distributed, as predicted by Theorem 1.5.6 (and also Theorem 1.5.3, because, here, we assume that the observations are normally distributed).

The fastest way to generate 200 new samples is to ask JMP to recalculate the formula "`Random Normal(3000, 100)`". This is done using the command "`Rerun Formulas`", which appears when you click on the hotspot (red triangle) menu next to the name of the data table. This is illustrated in Figure 1.4.

If each student takes samples of 20 instead of five observations, the histograms have a different shape: they are still bell-shaped but they are significantly narrower. Two histograms

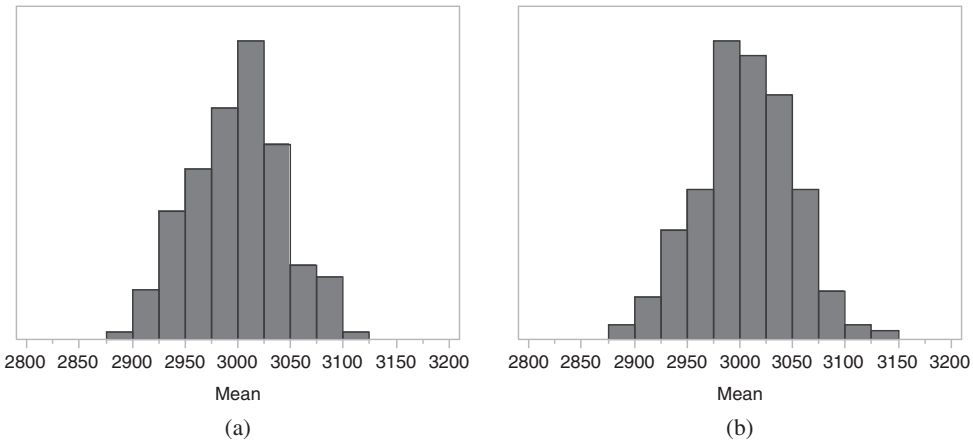


Figure 1.3 Two histograms of 200 sample means for normally distributed data and samples of five observations.

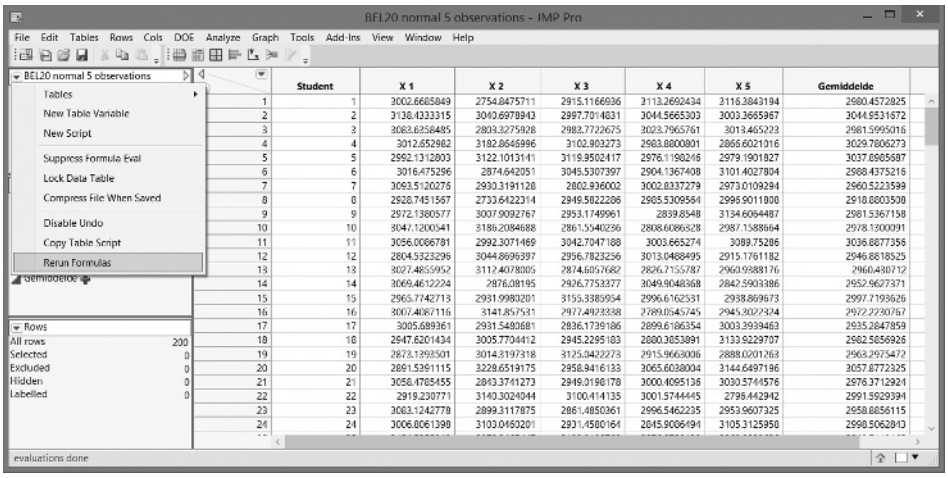


Figure 1.4 Generating new pseudo-random observations in JMP with the option "`Rerun Formulas`".

for 200 sample means of samples with 20 observations are shown in Figure 1.5. The bell shape tells us that the sample means are still normally distributed. The fact that the histograms are narrower should not come as a surprise, since the central limit theorem and Theorem 1.5.3 imply that the variance of the sample mean is equal to σ^2/n. As a consequence, sample means of 20 observations have a variance that is four times smaller than the variance of sample means of five observations.

Uniformly Distributed X

Suppose that the value of the Euro Stoxx 50 index is not normally distributed, but is uniformly distributed between 2800 and 3200. First, suppose again that each student takes a sample of

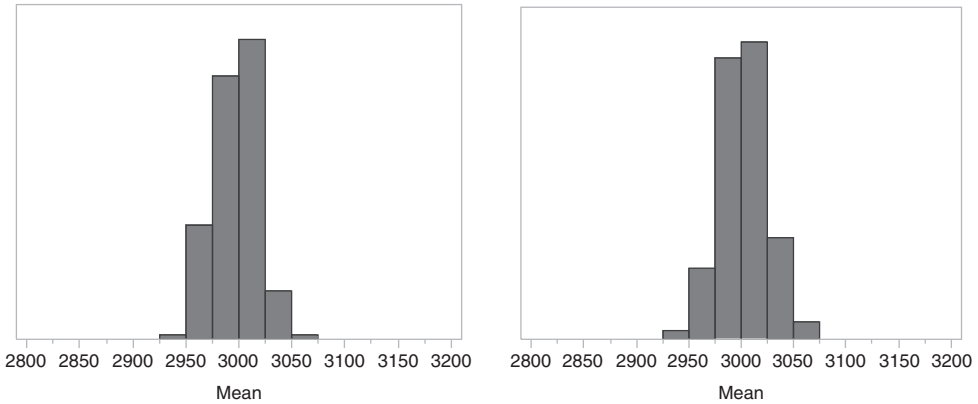


Figure 1.5 Two histograms of 200 sample means for normally distributed data and samples of 20 observations.

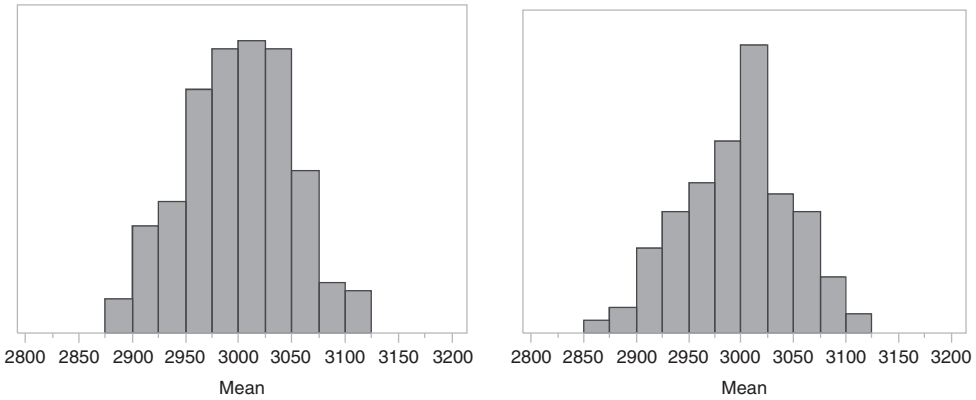


Figure 1.6 Two histograms of 200 sample means for uniformly distributed data and samples of five observations.

five observations. For this new scenario involving the uniform density, we again simulate 200 samples of five observations using JMP. To this end, we need to enter the formula "`Random Uniform(2800, 3200)`" in five columns of a data table with 200 rows. For each sample of five observations, we calculate the mean, and then we display all means in a histogram. Two possible histograms obtained in this way are shown in Figure 1.6. It is striking that, again, the histograms are quite bell-shaped, indicating that the sample means are still approximately normally distributed, even though the original data is uniformly distributed.

When the students take samples of 20 instead of five observations, the corresponding bell-shaped histograms are significantly narrower. Two histograms for 200 means of samples of 20 observations are shown in Figure 1.7.

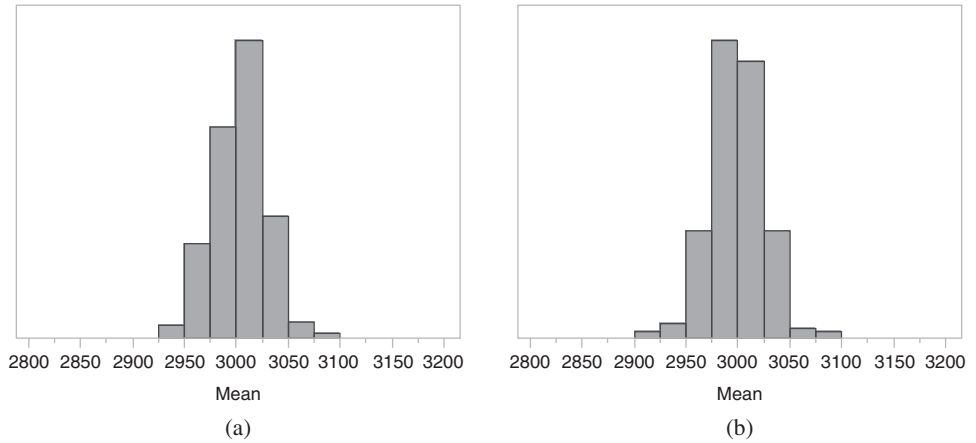


Figure 1.7 Two histograms of 200 sample means for uniformly distributed data and samples of 20 observations.

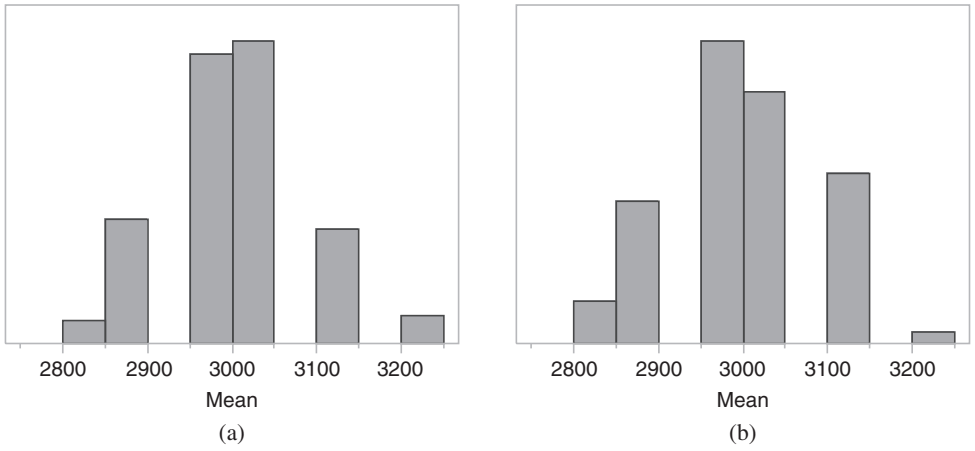


Figure 1.8 Two histograms of 200 sample means for Bernoulli distributed data and samples of five observations.

Bernoulli Distributed *X*

Now, suppose that the value of the Euro Stoxx 50 index is Bernoulli distributed, with a 50% chance that its value is 2800 and a 50% chance that its value is 3200. First, suppose again that each student takes a sample of five observations. We again need to simulate 200 samples of five observations using JMP. This time, we need to enter the formula "2800 + 400 * `Random Binomial(1, 0.5)`" in five columns of a data table with 200 rows. For each sample of five observations, we calculate the mean, and display the resulting 200 means in a histogram. Two possible histograms obtained in this way are shown in Figure 1.8. This time, the histograms are not bell-shaped. It is clearly visible that the original data comes from a discrete distribution, namely the Bernoulli distribution. The central limit theorem does not seem to work for the Bernoulli distribution and a sample of five observations.

When, however, the students take samples of 20 instead of five observations, the histograms look totally different. Although the histograms still do not exhibit a perfect bell shape, it is no longer obvious that the original data had a discrete probability distribution. Two possible histograms for 200 sample means of samples with 20 observations are shown in Figure 1.9. In order to obtain an even better bell shape, a slightly larger sample size is required.

This last example demonstrates that the central limit theorem is very powerful. Even probability distributions or probability densities that are quite different from the normal density still lead to distributions of sample means that are approximately normal, provided that the number of observations is sufficiently large.

1.6 The Sample Proportion

A sample proportion is a special case of the sample mean. Oftentimes, a variable under study can only take the values 0 or 1. Examples of such variables are gender (male/female) or

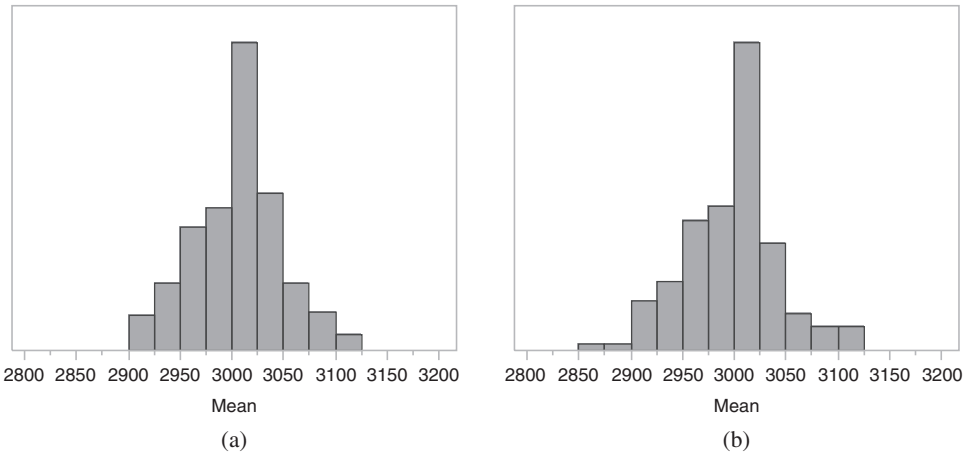


Figure 1.9 Two histograms of 200 sample means for Bernoulli distributed data and samples of 20 observations.

quality (defective/not defective). In general, the terms "success" and "failure" are used. The unknown population proportion is denoted by the letter π. This unknown population proportion indicates the proportion of successes in the entire population and is estimated using the sample proportion, which is simply the relative frequency of successes in a sample:

Definition 1.6.1 *The **sample proportion** is the number of successes in a sample divided by the number of observations.*

We can consider the sample proportion as a random variable or as a computed real number. We consider the sample proportion as a random variable as long as no sample data has been obtained. In that case, we use the symbol $\hat{P}$ to denote the sample proportion. If data is available and the sample proportion has been calculated, we use the symbol $\hat{p}$. For the sample proportion as random variable, we also use the symbol $\hat{P}$ to avoid confusion with a probability, which is typically denoted by the letter P.

If we adopt the convention of assigning the value "1" to the random variable X_i when the ith observation is a success, and the value "0" in the event of a failure, then we have

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X},$$

which shows that a sample proportion is actually a sample mean. Hence, the central limit theorem can be used: for large samples, the sample proportion is approximately normally distributed. The expected value and variance of the sample proportion are easily determined. Indeed, the sample proportion is a sum (and thus a linear combination) of n independent Bernoulli distributed random variables X_i with parameter π. As the expected value of a linear

combination of random variables is equal to the linear combination of the expected values, we obtain

$$\begin{aligned} E(\hat{P}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \\ &= \frac{1}{n}E(X_1 + X_2 + \cdots + X_n), \\ &= \frac{1}{n}\big(E(X_1) + E(X_2) + \cdots + E(X_n)\big), \\ &= \frac{1}{n}(\pi + \pi + \cdots + \pi), \\ &= \frac{1}{n}n\pi, \\ &= \pi. \end{aligned}$$

Since the variance of a linear combination of independent random variables is the linear combination of the variances with squared coefficients, we have

$$\begin{aligned} \text{var}(\hat{P}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \\ &= \frac{1}{n^2}\text{var}(X_1 + X_2 + \cdots + X_n), \\ &= \frac{1}{n^2}\big(\text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n)\big), \\ &= \frac{1}{n^2}(\pi(1-\pi) + \pi(1-\pi) + \cdots + \pi(1-\pi)), \\ &= \frac{1}{n^2}n\pi(1-\pi), \\ &= \frac{\pi(1-\pi)}{n}. \end{aligned}$$

The expected value of the sample proportion thus proves to be equal to the population proportion π, so that the sample proportion is an unbiased estimator of the population proportion. The variance of the sample proportion is equal to $\pi(1-\pi)/n$. This variance decreases linearly with the number of observations, n. In other words, if you collect more data, the sample proportion will give you a more precise estimate of the population proportion. All this is summarized in the following theorem:

Theorem 1.6.2 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with only two possible outcomes,* 0 *(failure) or* 1 *(success), and with a probability of success equal to π. Then, for a sufficiently large value of n:*

(1) the sample proportion $\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n}$ is approximately normally distributed, with expected value π and variance $\pi(1-\pi)/n$;

(2) and, consequently, the random variable

$$\frac{\hat{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

approximately follows a standard normal distribution.

As a rule of thumb for using the normal density as an approximation, the conditions $n\pi > 5$ and $n(1-\pi) > 5$ must be met. If one of these conditions is not valid, the normal distribution cannot be used for the sample proportion. In that case, the binomial distribution needs to be used instead of the normal probability density. Indeed, the number of successes in a sample can be described by a binomially distributed random variable with parameters n and π, if the probability of success for each individual observation is equal to π.

Figures 1.10 and 1.11 illustrate why the normal density can be used if $n\pi > 5$ and $n(1-\pi) > 5$. Figure 1.10 shows three binomial distributions for $n = 25$. Figure 1.10a shows the binomial distribution for $\pi = 0.5$. This distribution is nicely symmetrical and almost perfectly bell-shaped. In this case, the binomial distribution seems similar to a normal density. Figure 1.10b shows the binomial distribution for $\pi = 0.25$. While this distribution is not perfectly symmetrical, it still looks bell-shaped and resembles a normal probability density. The same applies to the binomial distribution with $\pi = 0.75$ in Figure 1.10c. In all these cases, $n\pi > 5$ and $n(1-\pi) > 5$.

Figure 1.11 shows two other binomial distributions, again with $n = 25$. Figure 1.11a shows the binomial distribution for $\pi = 0.05$. This distribution is far from symmetrical. In this case, the binomial probability distribution does not look like a normal probability density. Figure 1.11b shows the binomial distribution for $\pi = 0.95$. Again, the binomial distribution is neither symmetrical nor bell-shaped. In other words, the binomial probability distribution does not look like a normal probability density. For $\pi = 0.05$, the value of $n\pi$ is smaller than 5. For $\pi = 0.95$, the value of $n(1-\pi)$ is smaller than 5.

1.7 The Sample Variance

The sample variance can be viewed as an estimator for the population variance. Thus, just like the sample mean and the sample proportion, it can also be treated as a random variable. In this case, we use capital letters to define the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{1.1}$$

If we consider the sample variance as an estimate and therefore as a real number, we denote it by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

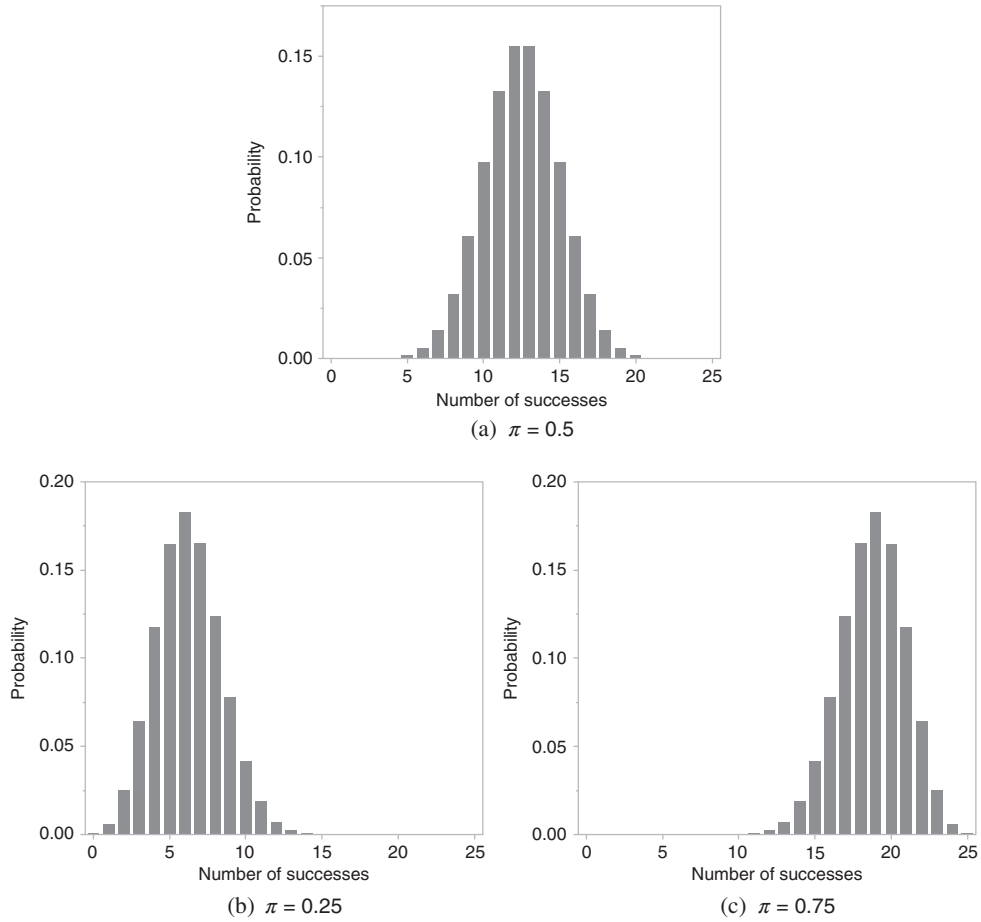


Figure 1.10 Binomial distributions with $n = 25$, with both $n\pi > 5$ and $n(1 - \pi) > 5$.

This second notation is used in descriptive statistics (see *Statistics with JMP: Graphs, Descriptive Statistics and Probability*). Once again, we use capital letters as long as there is no data, and lowercase letters when sample data has been used.

Like any other random variable, the random variable S^2 has an expected value, a variance, and a probability density.

1.7.1 The Expected Value

Theorem 1.7.1 *For a random sample from a population with variance σ^2, we have*

$$E(S^2) = \sigma^2.$$

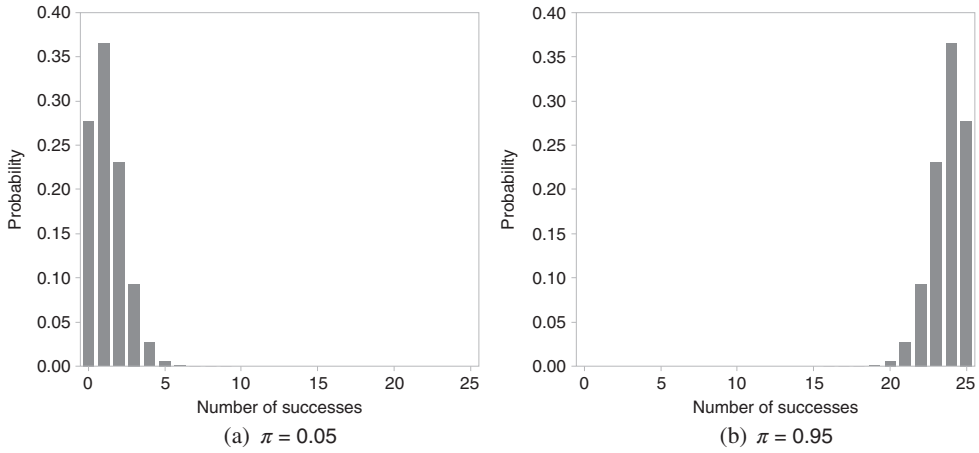


Figure 1.11 Binomial distributions with $n = 25$, with either $n\pi$ or $n(1 - \pi)$ less than 5.

Proof.

$$
\begin{aligned}
E(S^2) &= E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu)^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \overline{X}) + \sum_{i=1}^{n}(\mu - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) + n(\mu - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})(n\overline{X} - n\mu) + n(\mu - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\mu - \overline{X})^2 + n(\mu - \overline{X})^2\right\}, \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \overline{X})^2\right\},
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\{(X_i-\mu)^2\} - nE\{(\mu-\overline{X})^2\}\right],\\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\{(X_i-\mu)^2\} - nE\{(\overline{X}-\mu)^2\}\right],\\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n} \mathrm{var}(X_i) - n\ \mathrm{var}(\overline{X})\right\},\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} \sigma^2 - n\frac{\sigma^2}{n}\right),\\
&= \frac{1}{n-1}(n\sigma^2-\sigma^2),\\
&= \sigma^2.
\end{aligned}
$$

■

The theorem proves that the sample variance S^2 is an unbiased estimator of the population variance σ^2, regardless of the probability distribution or probability density of the population. In addition, the theorem explains why it is important that we divide by $n-1$ instead of n in the computation of the sample variance. It is important to note that the unbiasedness is a property of the sample variance, but not of the sample standard deviation S. In general,

$$E(S) \neq \sigma.$$

This implies that the sample standard deviation S is a biased estimator of the population standard deviation. We briefly discuss the sample standard deviation in Section 1.8. Before we discuss the distribution of the sample variance S^2, we introduce a new probability density, namely the χ^2-distribution (pronounced "chi-square").

1.7.2 The χ^2-Distribution

Apart from the normal distribution, the χ^2-distribution[7], which is derived from normal distributions, is a very important family of probability densities. This family has one parameter k, which is called the **degrees of freedom**. The probability density of this distribution is

$$f_X(x;k) = \frac{x^{\frac{k}{2}-1}\mathrm{e}^{-x/2}}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}},\ \text{for } x>0.$$

[7] The name of the χ^2-distribution derives from the fact that the square of a standard normally distributed random variable has a χ^2-distribution: a random variable is typically denoted by an X and its square by X^2, and X is the capital form of the Greek letter χ.

In this expression, $\Gamma()$ is the so-called gamma function[8]. The χ^2-distribution is a special case of a gamma distribution (see *Statistics with JMP: Graphs, Descriptive Statistics and Probability*). The expected value and variance of a χ^2-distributed random variable X with k degrees of freedom are

$$\mu_X = E(X) = k$$

and

$$\sigma_X^2 = \text{var}(X) = 2k,$$

respectively. The median is approximately equal to

$$k\left(1 - \frac{2}{9k}\right)^3.$$

To indicate that a random variable has a χ^2-distribution with k degrees of freedom – that is, with parameter k – we use the notation

$$X \sim \chi_k^2.$$

χ^2-distributions with two, four, eight, and 12 degrees of freedom are shown in Figure 1.12.

Probabilities based on the χ^2-distribution can, of course, be computed using JMP. The probability

$$P(\chi_k^2 \leq x),$$

where χ_k^2 is a χ^2-distributed random variable with k degrees of freedom, can be calculated using the formula "`ChiSquare Distribution(x, k)`". If you want to find a quantile or percentile of a χ^2-distributed random variable with k degrees of freedom, the function "`ChiSquare Quantile(p, k)`" is available. Some specific quantiles can also be found in the table in Appendix C.

1.7.3 The Relation Between the Standard Normal and the χ^2-Distribution

A sum of k squared independent standard normally distributed random variables $X_1, X_2, \ldots, X_k$ is χ^2-distributed with k degrees of freedom. In other words, the random variable

$$Y = \sum_{i=1}^{k} X_i^2,$$

[8] The gamma function is an extension of the factorial function for integers n: $\Gamma(n+1) = n! = n \times (n-1) \times \cdots \times 2 \times 1$. If z is a positive real number, then $\Gamma(z) = (z-1)\Gamma(z-1) = \int_0^{+\infty} \mathrm{e}^{-t} t^{z-1} dt$. A special case is $\Gamma(1/2) = \sqrt{\pi}$. In addition, $\Gamma(2) = \Gamma(1) = 1! = 0! = 1$, and $\Gamma(n) = (n-1)!$ for all integers n.

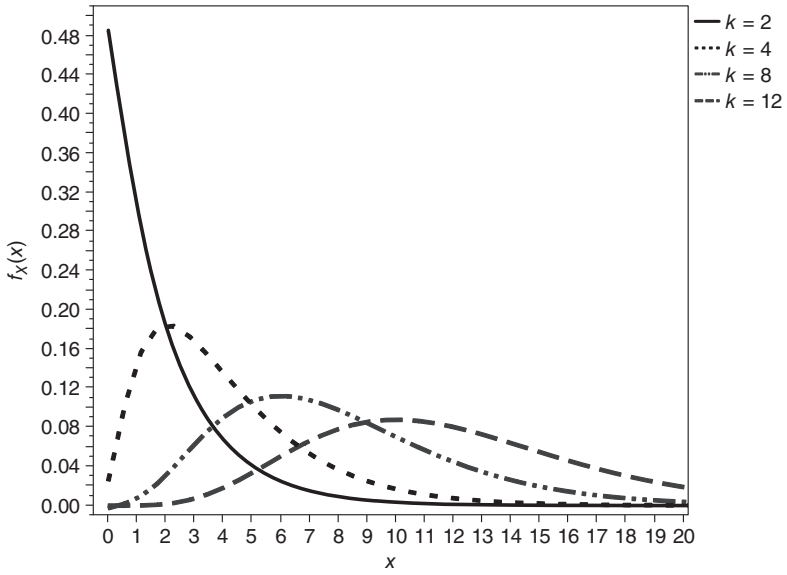


Figure 1.12 χ^2-distributions with two, four, eight, and 12 degrees of freedom.

where the X_i are standard normally distributed, is χ^2-distributed. The number of degrees of freedom of the resulting χ^2-distribution is equal to the number of independent variables involved in the sum of squares.

This assertion can easily be verified using JMP. To do so, generate 10 000 sets of eight pseudo-random numbers from the standard normal density. Next, square all numbers in each set of eight and sum the squares. Finally, create a histogram of the 10 000 sums of squares. You can then compare the shape of the histogram with that of the χ^2-distribution with eight degrees of freedom in Figure 1.12. Alternatively, you can compute the mean and the variance of the 10 000 sums of squares that you obtained. The mean should lie close to 8, while the variance should be close to 16. A histogram that was obtained in this way is shown in Figure 1.13. As the JMP output in Figure 1.14 shows, the mean of the 10 000 values is 7.9755, while the variance is 15.8786. Obviously, if you repeat this exercise in JMP by yourself, you will obtain other pseudo-random numbers and therefore a (slightly) different mean and a (slightly) different variance.

As the number of degrees of freedom of the χ^2-distribution increases, the distribution looks more and more like a normal probability density. This is a consequence of the central limit theorem: a χ^2-distributed random variable with a large number of degrees of freedom is a sum of a large number of independent random variables (the square of a standard normally distributed random variable is, of course, also a random variable), and the central limit theorem tells us that such a sum is approximately normally distributed. Figure 1.15 compares the χ^2-distributions with 25 and 30 degrees of freedom with the normal probability density with $\mu = 25$ and $\sigma^2 = 50$ and the normal probability density with $\mu = 30$ and $\sigma^2 = 60$, respectively. The figure clearly shows that, for a large number of degrees of freedom, there is a strong resemblance between the χ^2-distribution and the normal distribution. Figure 1.12 shows that this resemblance is absent if the number of degrees of freedom is small.

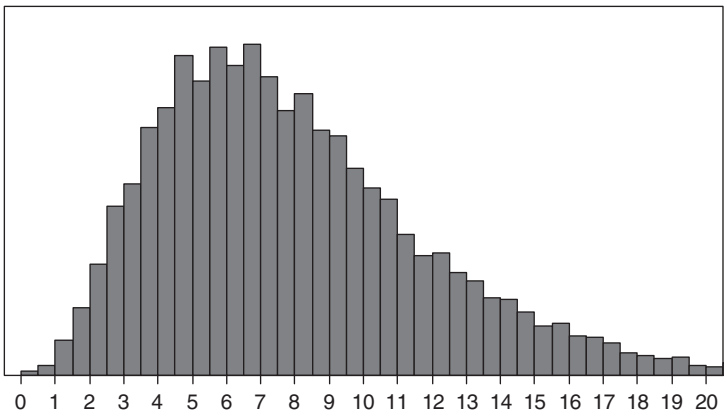
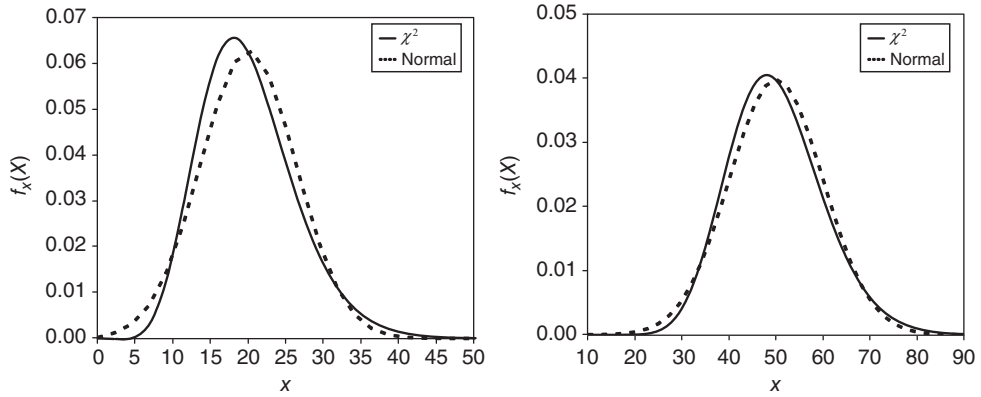


Figure 1.13 The histogram obtained by generating 10 000 sets of eight (pseudo-random) draws from a standard normal density, squaring them, and summing the squares.

Summary Statistics	
Mean	7.9754589
Std Dev	3.984791
Std Err Mean	0.0398479
Upper 95% Mean	8.0535688
Lower 95% Mean	7.897349
N	10000
Variance	15.878559

Figure 1.14 The descriptive statistics of the 10 000 (pseudo-random) numbers shown in Figure 1.13.



(a) $k = 25$ (b) $k = 50$

Figure 1.15 A comparison of a χ^2-distribution with k degrees of freedom and the corresponding normal probability density with $\mu = k$ and $\sigma^2 = 2k$.

1.7.4 The Probability Density of the Sample Variance

Starting with the definition of a sample variance in Equation (1.1), it is not difficult to see that there is a connection between the sample variance and the χ^2-distribution. Dividing both sides of Equation (1.1) by σ^2 and multiplying by $(n-1)$ yields

$$\begin{aligned}\frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2},\\ &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2.\end{aligned} \tag{1.2}$$

Replacing the sample mean $\overline{X}$ in the right-hand side of this expression by the population mean μ, we obtain

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

If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables with expected value μ and variance σ^2, then this expression is a sum of squared independent standard normally distributed random variables. Accordingly, this expression has a χ^2-distribution with n degrees of freedom.

In Equation (1.2), however, the population mean μ is estimated by the sample mean $\overline{X}$. As a consequence, the n terms in the sum

$$\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}$$

are not independent. This is due to the fact that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i, \tag{1.3}$$

which implies that the X_i are subject to the following constraint:

$$\sum_{i=1}^{n}(X_i-\overline{X}) = 0.$$

As a result of this one constraint, the sum in Equation (1.3) contains only $n-1$ independent terms. Therefore, the χ^2-distribution of $(n-1)S^2/\sigma^2$ has only $n-1$ degrees of freedom.

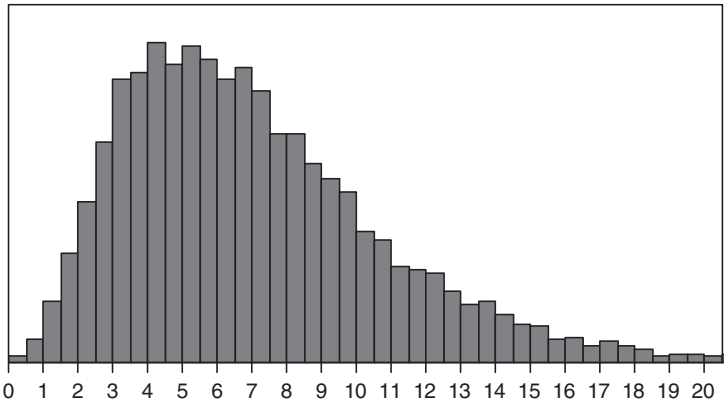


Figure 1.16 The histogram of 10 000 (pseudo-random) values of the random variable $(n-1)S^2/\sigma^2$.

Theorem 1.7.2 *Let $X_1, X_2, \ldots, X_n$ be independent normally distributed random variables with variance σ^2, and $n \geq 2$. Then, the random variable*

$$\frac{(n-1)S^2}{\sigma^2}$$

has a χ^2-distribution with $n-1$ degrees of freedom.

These theoretical results can also be verified using JMP. To do so, we can generate 10 000 sets of eight numbers from a normal density. We determine the sample variance for each of these 10 000 sets, multiply it by $n-1=7$, and divide the result by σ^2. A histogram of 10 000 values obtained in this way is shown in Figure 1.16. The mean of all these numbers is 7.0280 (see Figure 1.17), which indicates that the underlying χ^2-distribution has only seven degrees of freedom instead of eight. The histogram in Figure 1.16 looks a lot like the one in Figure 1.13 (generated from a χ^2-distribution with eight degrees of freedom). However, the difference between the two histograms is that the maximum is reached slightly earlier in Figure 1.16 than in the histogram in Figure 1.13, as a result of the smaller number of degrees of freedom of the distribution in Figure 1.16.

Summary Statistics	
Mean	7.0280156
Std Dev	3.7366969
Std Err Mean	0.037367
Upper 95% Mean	7.1012624
Lower 95% Mean	6.9547688
N	10000
Variance	13.962903

Figure 1.17 The descriptive statistics of the 10 000 (pseudo-random) values of the random variable $(n-1)S^2/\sigma^2$ shown in Figure 1.16.

Now that the probability density of the sample variance is known, we can determine the variance of the sample variance. We know that the variance of a χ^2-distributed random variable with k degrees of freedom is equal to $2k$. Thus a χ^2-distributed random variable with $n-1$ degrees of freedom has a variance of $2(n-1)$. Therefore,

$$\text{var}\left\{\frac{(n-1)S^2}{\sigma^2}\right\} = 2(n-1).$$

Hence,

$$\left(\frac{n-1}{\sigma^2}\right)^2 \text{var}(S^2) = 2(n-1),$$

and

$$\text{var}(S^2) = 2(n-1)\left(\frac{\sigma^2}{n-1}\right)^2 = \frac{2\sigma^4}{n-1}.$$

The variance of the sample variance S^2 thus decreases with the number of observations in the sample, n. In other words, the sample variance is a more precise and efficient estimator when a larger amount of data is used. The same applies to the sample mean and the sample proportion.

Finally, it should be emphasized that the use of the χ^2-distribution in the context of sample variances is only valid for normally distributed variables $X_1, X_2, \ldots, X_k$. In other words, if you study a nonnormally distributed population, you should not assume that $(n-1)S^2/\sigma^2$ has a χ^2-distribution. Therefore, the above expression for the variance of the sample variance, $\text{var}(S^2)$, is also invalid for nonnormally distributed populations.

1.8 The Sample Standard Deviation

The sample standard deviation S is the square root of the sample variance S^2:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

In Section 1.7.1, we showed that the sample variance is an unbiased estimator of the population variance. However, the sample standard deviation turns out to be a biased estimator of the population standard deviation. It can be shown that

$$E(S) < \sigma,$$

so that the sample standard deviation S generally underestimates the population standard deviation σ.

If the population under investigation is normally distributed, a bias correction can be applied to eliminate the systematic underestimation. This correction is based on the fact that

$$E(S) = \sigma\sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}, \tag{1.4}$$

where $\Gamma()$ again represents the gamma function. The bias correction factor

$$\sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

is referred to as c_4 in the statistical literature.

It follows from the expression for the expected value of S in Equation (1.4) that the corrected sample standard deviation

$$S^* = \frac{S}{c_4} = S\sqrt{\frac{n-1}{2}} \cdot \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

Table 1.1 Values of the factor c_4 and the extent to which the population standard deviation is underestimated by the sample standard deviation for a normally distributed population.

	c_4		
n	Exact	Approximation	Underestimation (%)
2	$\sqrt{\frac{2}{\pi}}$	0.7979	20.21
3	$\frac{\sqrt{\pi}}{2}$	0.8862	11.38
4	$2\sqrt{\frac{2}{3\pi}}$	0.9213	7.87
5	$\frac{3}{4}\sqrt{\frac{\pi}{2}}$	0.9400	6.00
6	$\frac{8}{3}\sqrt{\frac{2}{5\pi}}$	0.9515	4.85
7	$\frac{5\sqrt{3\pi}}{16}$	0.9594	4.06
8	$\frac{16}{5}\sqrt{\frac{2}{7\pi}}$	0.9650	3.50
9	$\frac{35\sqrt{\pi}}{64}$	0.9693	3.07
10	$\frac{128}{105}\sqrt{\frac{2}{\pi}}$	0.9727	2.73
100	$\frac{6511077190}{1570338389}\sqrt{\frac{2}{11\pi}}$	0.9975	0.25

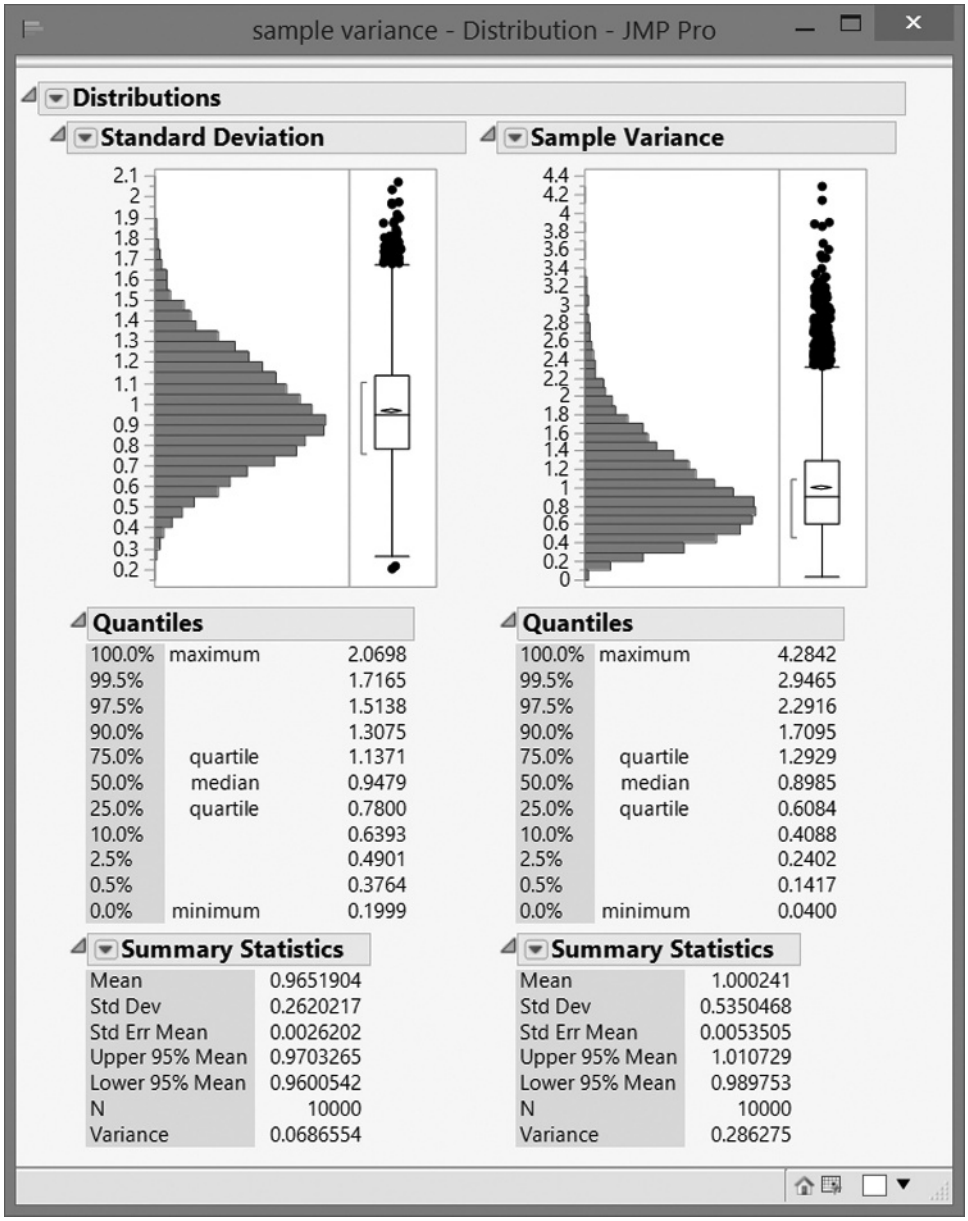


Figure 1.18 Histograms, box plots, and descriptive statistics for the sample standard deviations and variances of 10 000 samples of eight (pseudo-random) values drawn from a standard normally distributed population.

is an unbiased estimator of the population standard deviation σ. The extent to which σ is underestimated is larger for a small number of observations. The factor c_4 is approximately 0.80 or 80% if $n = 2$. This means that the sample standard deviation yields estimates that are generally 20% smaller than the population standard deviation. If n is equal to 10, the factor c_4 is approximately 0.97. In that case, the sample standard deviation provides estimates that are, on average, 3% too low. If the number of observations increases up to 100, then c_4 is nearly equal to 1, and the sample standard deviation is almost an unbiased estimator. Table 1.1 contains a list of values of the factor c_4 and the degree of underestimation of the population standard deviation by the sample standard deviation. Note that in this table, π is not a success probability, but the circle constant 3.1415 … .

To illustrate all this, we have used JMP to generate 10 000 samples consisting of eight pseudo-random numbers drawn from the standard normal probability distribution. As a result, both the population variance and the population standard deviation are equal to 1 because the data is generated from the standard normal probability density. For each sample, the sample standard deviation and the sample variance are computed. Histograms, box plots, and descriptive statistics for the 10 000 sample standard deviations and sample variances are shown in Figure 1.18. The mean sample standard deviation is 0.9652, so that the population standard deviation is underestimated by about 3.5%. This corresponds to the value of c_4 and the underestimation displayed in Table 1.1. The mean sample variance is 1.0002, which is very close to the population variance of 1. This illustrates that the sample variance is an unbiased estimator of the population variance. The variance of the 10 000 sample variances is 0.2863, which is close to the theoretical value of $2\sigma^4/(n-1) = 2 \times 1^4/(8-1) = 2/7 = 0.2857$.

1.9 Applications

The derived probability densities for the sample mean, the sample proportion, and the sample variance are the cornerstones for the remaining chapters of the book. The normal (and therefore also the standard normal) distribution, and the χ^2-distribution will be used to establish so-called confidence intervals for the population mean, the population proportion, and the population variance. In addition, two new probability densities, namely the t-distribution and the F-distribution, will be derived. These new distributions will be needed for other confidence intervals and hypothesis tests.