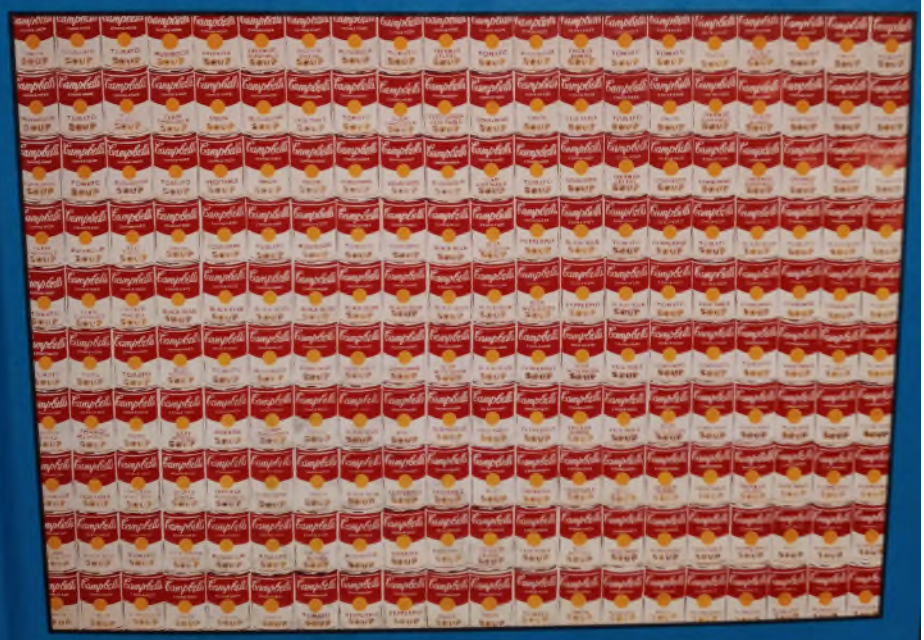
David S. Moore • George P. McCabe

INTRODUCTION to the PRACTICE of STATISTICS

THIRD EDITION



Complimentary CD-ROM with every book

INTRODUCTION: What Is Statistics?

Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data*. We are bombarded by data in our everyday lives. Most of us associate "statistics" with the bits of data that appear in news reports: baseball batting averages, imported car sales, the latest poll of the president's popularity, and the average high temperature for today's date. Advertisements often claim that data show the superiority of the advertiser's product. All sides in public debates about economics, education, and social policy argue from data. Yet the usefulness of statistics goes far beyond these everyday examples.

The study and collection of data are important in the work of many professions, so that training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisors, as well as policy makers in government and business study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals if they are to offer their patients the most effective treatments. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal consumer tastes. Farmers study data from field trials of new crop varieties. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers, and therefore also make use of the methods of statistics.

We can no more escape data than we can avoid the use of words. Just as words on a page are meaningless to the illiterate or confusing to the partially educated, so data do not interpret themselves but must be read with understanding. Just as a writer can arrange words into convincing arguments or incoherent nonsense, so data can be compelling, misleading, or simply irrelevant. Numerical literacy, the ability to follow and understand numerical arguments, is important for everyone. The ability to express yourself numerically, to be an author rather than just a reader, is a vital skill in many professions and areas of study. The study of statistics is therefore essential to a sound education. We must learn how to read data, critically and with comprehension. We must learn how to produce data that provide clear answers to important questions. And we must learn sound methods for drawing trustworthy conclusions based on data.

and standard deviation of $aX + bY$ are found as usual from the addition r
for means and variances. These facts are often used in statistical calculatic

Tom and George are playing in the club golf tournament. Their scores vary
they play the course repeatedly. Tom's score X has the $N(110, 10)$ distribution, ar
George's score Y varies from round to round according to the $N(100, 8)$ distributio
If they play independently, what is the probability that Tom will score lower tha
George and thus do better in the tournament? The difference $X - Y$ between thei
scores is normally distributed, with mean and variance

$$\mu_{X-Y} = \mu_X - \mu_Y = 110 - 100 = 10$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 10^2 + 8^2 = 164$$

Because $\sqrt{164} = 12.8$, $X - Y$ has the $N(10, 12.8)$ distribution. Figure 5.8 illustrates the probability computation:

$$\begin{aligned} P(X < Y) &= P(X - Y < 0) \\ &= P\left(\frac{(X - Y) - 10}{12.8} < \frac{0 - 10}{12.8}\right) \\ &= P(Z < -0.78) = 0.2177 \end{aligned}$$

Although George's score is 10 strokes lower on the average, Tom will have the lower score in about one of every five matches.

The central limit theorem

The sampling distribution of $\bar{x}$ is normal if the underlying population itself has a normal distribution. What happens when the population distribution is not normal? It turns out that as the sample size increases, the distribution of $\bar{x}$ gets closer to a normal distribution. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation σ. This famous fact of probability theory is called the *central limit theorem*. It is much more useful than the fact that the distribution of $\bar{x}$ is exactly normal if the population is exactly normal.* For large sample size n, we can regard $\bar{x}$ as having the $N(\mu, \sigma/\sqrt{n})$ distribution.

*The first general version of the central limit theorem was established in 1810 by the French mathematician Pierre Simon Laplace (1749–1827).

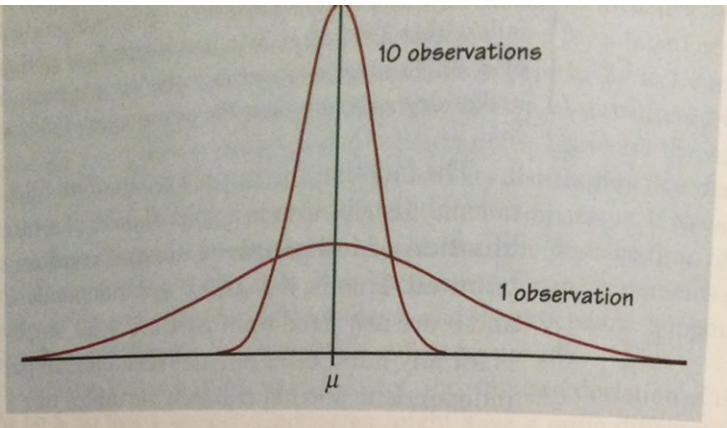


FIGURE 5.7 The sampling distribution of $\bar{x}$ for samples of size 10 compared with the distribution of a single observation.

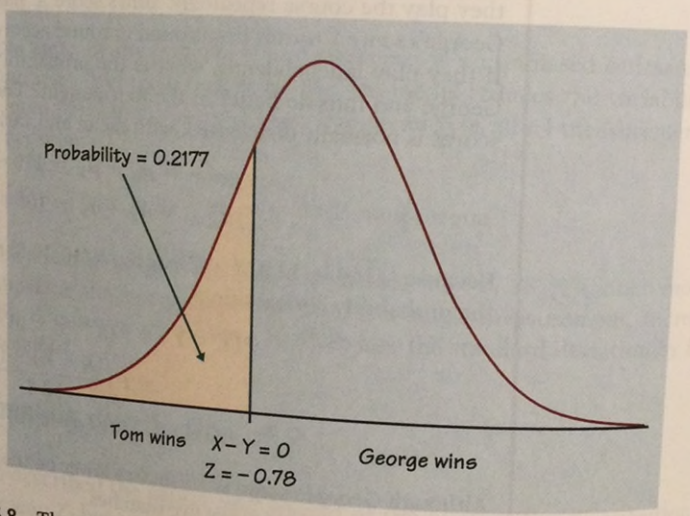


FIGURE 5.8 The normal probability calculation for Example 5.17.

Central Limit Theorem

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

More generally, the central limit theorem says that the distribution of a sum or average of many small random quantities is close to normal. This is true even if the quantities are not independent (as long as they are not too

highly correlated) and even if they have different distributions (as long as no one random quantity is so large that it dominates the others). The central limit theorem suggests why the normal distributions are common models for observed data. Any variable that is a sum of many small influences will have approximately a normal distribution.

How large a sample size n is needed for $\bar{x}$ to be close to normal depends on the population distribution. More observations are required if the shape of the population distribution is far from normal.

AMPLE 5.18 Figure 5.9 shows the central limit theorem in action for a very nonnormal population. Figure 5.9(a) displays the density curve of a single observation, that is, of the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean μ of this distribution is 1, and its standard deviation σ is also 1. This particular continuous distribution is called an **exponential distribution.** Exponential distributions are used as models for the lifetime in service of electronic components and for the time required to serve a customer or repair a machine.

Figures 5.9(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As n increases, the shape becomes more normal. The mean remains at $\mu = 1$, and the standard deviation decreases, taking the value $1/\sqrt{n}$. The density curve for 10 observations is still somewhat skewed to the right but already resembles a normal curve having $\mu = 1$ and $\sigma = 1/\sqrt{10} = 0.32$. The density curve for $n = 25$ is yet more normal. The contrast between the shapes of the population distribution and of the distribution of the mean of 10 or 25 observations is striking.

The central limit theorem allows us to use normal probability calculations to answer questions about sample means from many observations even when the population distribution is not normal.

AMPLE 5.19 The time X that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by the exponential distribution whose density curve appears in Figure 5.9(a). The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

The central limit theorem says that the sample mean time $\bar{x}$ (in hours) spent working on 70 units has approximately the normal distribution with mean equal to the population mean $\mu = 1$ hour and standard deviation

$$\frac{\sigma}{\sqrt{70}} = \frac{1}{\sqrt{70}} = 0.12 \text{ hour}$$

The distribution of $\bar{x}$ is therefore approximately $N(1, 0.12)$. Figure 5.10 shows this normal curve (*solid*) and also the actual density curve of $\bar{x}$ (*dashed*).

Because 50 minutes is 50/60 of an hour, or 0.83 hour, the probability we want is $P(\bar{x} > 0.83)$. A normal distribution calculation gives this probability as 0.9222. This is the area to the right of 0.83 under the solid normal curve in Figure 5.10. The exactly correct probability is the area under the dashed density curve in the figure. It is 0.9294. The central limit theorem normal approximation is off by only about 0.007.