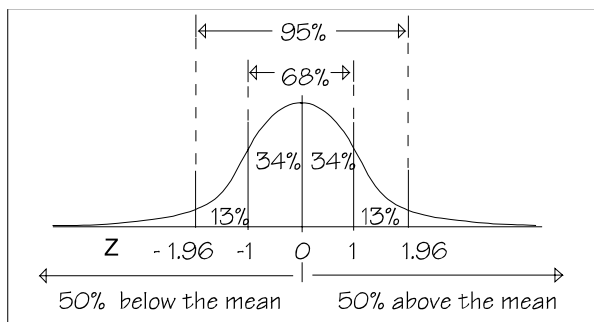*Chapter 8*

THE NORMAL CURVE AND SAMPLES: SAMPLING DISTRIBUTIONS

A picture of an ideal normal distribution is shown below. The horizontal axis is calibrated in *z*-scores— in terms of standard deviation units. The *z*-score of the mean is zero. The point one standard deviation above the mean is marked "+1".

The height of the curve over the point labeled "+1" shows how many cases have scores one standard deviation above the mean (how many people have *z*-scores of 1.0).

The area between **-1** and **+1** is the proportion of cases within one standard deviation of the mean. For normally-distributed variables, 68% of all scores are within one standard deviation of the mean; 95% are within 1.96 standard deviations, and more than 99%

(99.7%, to be exact) are within three standard deviations.

The relation between standard deviation and the areas under the normal curve turns out to be of great value when you want to generalize from a sample to a population.

This takes us into the area of **inferential statistics**, where we use information about a *sample* to make inferences about the *population*. It's almost like magic, where we use information about a small number of people to learn about a much larger population. This is probably the most amazing and valuable analytic tool you will ever encounter.

There are two kinds of estimates you can make about a population on the basis of a sample drawn from the population:

● First, you can use sample statistics, like the mean and standard deviation, to estimate the corresponding population parameters.

● Second, you can use knowledge about some of the properties of random samples to say how accurate your sample mean is as an estimate of the population's mean, and to tell how much confidence you can place in that assessment of accuracy.

It's easy to estimate the mean and standard deviation of a population:

- Your best estimate of the population's mean is simply the sample's mean.

- To estimate the population's standard deviation, all you have to do is calculate the standard deviation of the sample, using "(n-1)" in the denominator instead of "n".

Assessing the accuracy of those estimates and how much confidence you can place in them is a bit more complicated. To understand how this procedure works, you have to understand **sampling distributions**.
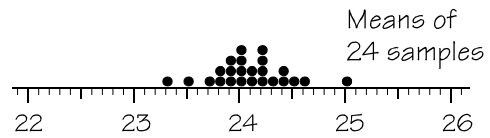
Sampling Distributions

Imagine you are going to do a study of Australian undergraduates (this is an international text). You can't afford to collect data from all of them, though, so you do a sample. Because of the higher cost of tuition and books, you are short on money, so your sample includes only two dozen (24) students, chosen *randomly* from the millions of undergraduates in the country down under. Your results show that the mean age of students in your sample is 24. You conclude that, since the sample mean is the best estimator of the population mean, the mean age of Australian undergraduates is 24.
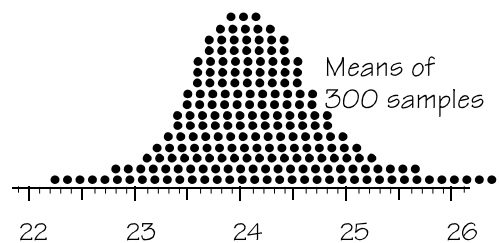

But then you learn that someone else did the same study, and they got a mean age of 25.

The difference between that result and yours makes you curious. You ask some friends, and you find two dozen more studies that used the same procedures and same types of samples. When you plot the results along with yours, this is what you see:

Means of 24 samples

Upon further investigation, you learn that three hundred people did the same study. Even more strange than that, they all followed the same procedures and asked the same questions and constructed their samples in the same way. You manage to get the results of all three hundred studies. Now you have a distribution of three hundred sample means. This is the beginning of a *sampling distribution* of sample means. If you managed to somehow take *all possible* samples of 24 Australian undergraduates, selected at random, you would have the complete sampling distribution for this variable and this sample size.
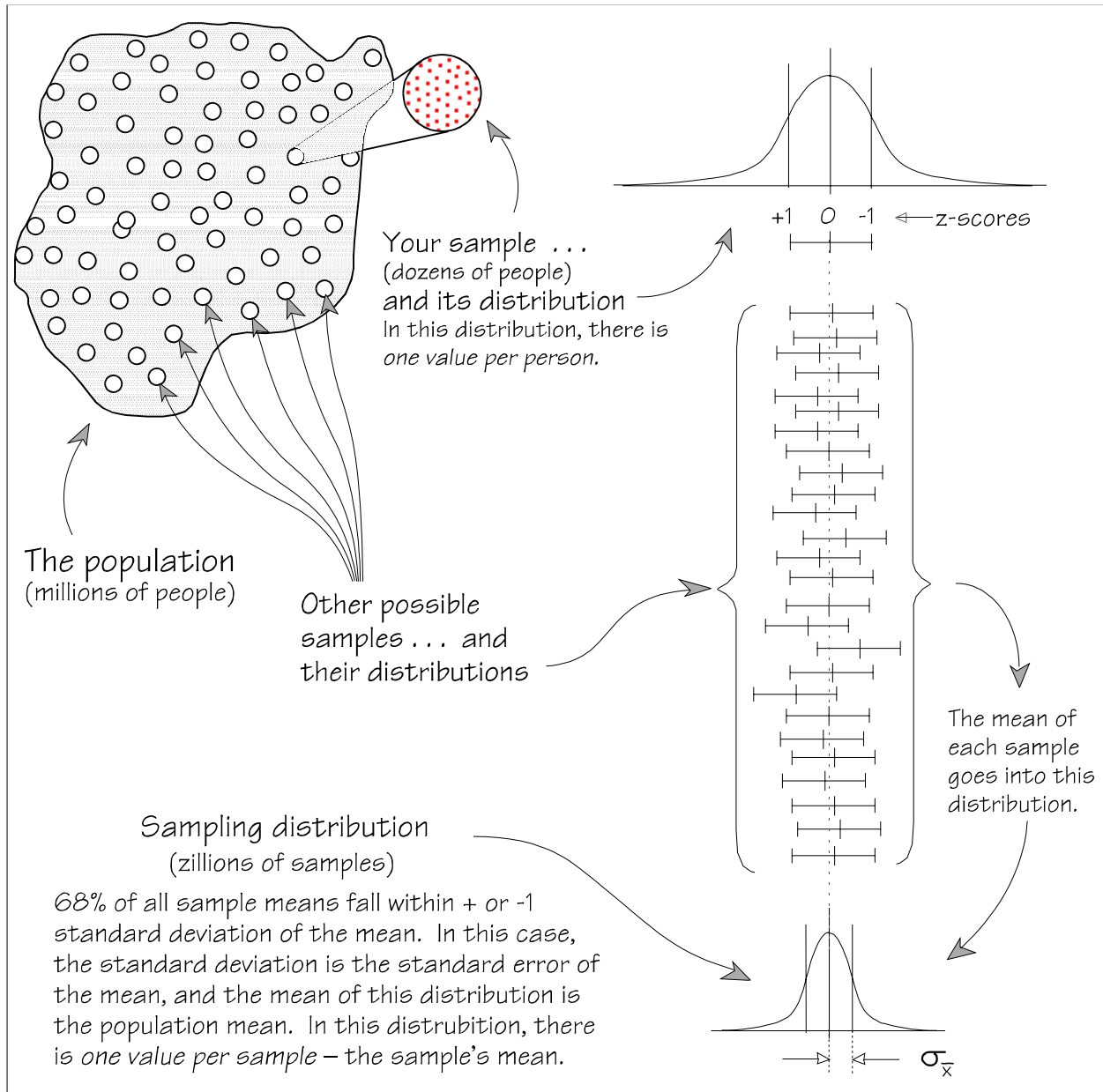

Means of 300 samples

If you were able to do this, you would notice two things that turn out to be extremely important:

- First, *the mean of the sampling distribution is the same as the population's mean*.

- Second, about 68% of all sample means fall within one standard deviation of the population mean; 95% fall within two standard deviations; and about 99% fall within three standard deviations.

You would probably suspect (correctly!) that the means of these random samples are *normally distributed*. Therefore, if you can find a way to calculate the standard deviation of the sampling distribution, you will know how certain you can be that your sample mean is within a given distance of the population's mean.

There are three important things to know about sampling distributions of sample means. If this were a statistics course taught in the Math department, you

Your sample ...
(dozens of people)
and its distribution
In this distribution, there is
one value per person.

The population
(millions of people)

Other possible
samples ... and
their distributions

Sampling distribution
(zillions of samples)

68% of all sample means fall within + or -1
standard deviation of the mean. In this case,
the standard deviation is the standard error of
the mean, and the mean of this distribution is
the population mean. In this distrubition, there
is one value per sample – the sample's mean.

+1 O -1 ←—z-scores

The mean of
each sample
goes into this
distribution.

$\sigma_{\bar{x}}$

might have to be able to prove that these things are true, but we'll just have faith in the honesty of the mathematicians and believe it when they tell us:

1. The mean of the sampling distribution is the same as the mean of the parent population.

2. The standard deviation of the sampling distribution of sample means is related to the standard deviation of the population the samples come from. It can be calculated by dividing the standard deviation of the parent population by the square root of your sample size:

$$\sigma_{\bar{x}} \;=\; \frac{\sigma}{\sqrt{n}}$$

"But," you say, "we don't know the standard deviation of the population! How can we do this?" I say "Ah hah! You know the standard deviation of your sample, and you might remember that you can use this as an estimate of the population's standard deviation."

3. The sampling distribution is normally distributed when the parent population is normally distributed, and is approximately normal *even when the parent population isn't normally distributed*—as long as the samples are large (n = 30 or more).

You can be about 68% certain that your sample mean lies within one standard error of the population mean, and 95% certain that your sample mean lies within two standard errors of the population mean. (Does "these results are accurate to within 4% nineteen times out of twenty … " sound familiar?)

Standard Errors: Standard Deviations of Sampling Distributions

The drawing on the next page shows several samples. You will notice that their means all differ slightly from one another and from the population mean. On the bottom of the drawing is a small distribution. This one shows the means of all the samples. If it had the means from *all possible samples* of this size, it would be a "*sampling distribution*." The mean of the sample means is obtained by adding all the sample means together and dividing the result by the number of samples, which in the diagram is 29, like this:

$$\frac{Mean_1 + Mean_2 + Mean_3 + \ldots + Mean_{29}}{29}$$

The sampling distribution has a *mean* and a *standard deviation*.

● The mean of the sampling distribution is the same as the population mean.

● The standard deviation of the sampling distribution is called a *standard error*. The standard deviation of the sampling distribution in the example above is the *standard error of the mean* because the sampling distribution in the example is the distribution of means of samples.

The symbol for standard error of the mean tells you a lot about what it is:

$$\sigma_{\bar{x}}$$

The Greek "σ" indicates that it is the standard deviation of a population; the subscript $\bar{x}$ tells that it is the standard deviation of the sampling distribution of sample means.

There are also the *standard error of the difference between two means*:
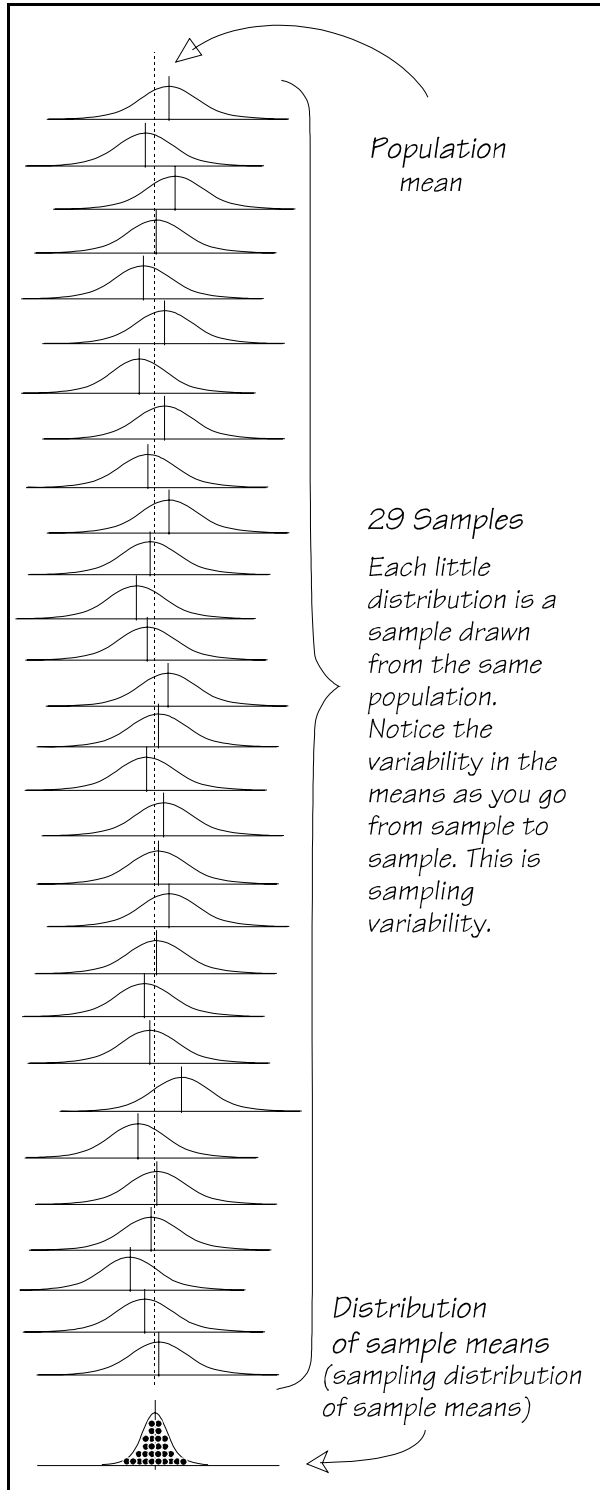
$$\sigma_{\bar{x}_1 - \bar{x}_2}$$

and the *standard error of proportions*:

$$\sigma_{\bar{p}}$$

Here are some important things to remember about standard errors:

● All standard errors are standard deviations of sampling distributions.

● Standard errors are *population parameters* and must therefore be estimated.

● The populations described by standard errors are sampling distributions—*theoretical* populations comprised of all possible samples of a given size from your sample's parent population.

● Standard errors are measures of the *sampling variability* associated with sample statistics. The standard error of the mean, for example, is a measure of the amount of sampling variability of the mean of a sample the size you are working with. In other words, they are measures of how reliable the sample statistic is as a measure of a population parameter.

A common misunderstanding of the standard error of the mean is that it is the "standard deviation of the mean." This is only partly correct. The standard error of the mean is the standard deviation of the sampling distribution of sample means. In other words, it is the standard deviation of the list of means from all possible samples of a given size drawn from a single population.

Population mean

29 Samples

Each little distribution is a sample drawn from the same population. Notice the variability in the means as you go from sample to sample. This is sampling variability.

Distribution of sample means (sampling distribution of sample means)

Well, if sample means are normally distributed about the population mean, then 68% of sample means are within one standard deviation of the population mean, therefore there is a 68% chance that your sample's mean is within one standard deviation of the population mean. In other words, you can be 68% certain that the difference between your sample's mean and the population's mean is no more than one standard deviation. *Remember that the "standard deviation" here is the standard deviation of the sampling distribution—the standard error of the mean.*

Here is an example. Say you have a sample of 49 randomly-chosen university students, and you know each student's height, accurate to within a tenth of a millimeter. You could easily calculate the mean and standard deviation of the 49 numbers in your sample. Let's say the mean turns out to be 190cm and the standard deviation (calculated with "n-1" in the denominator!) is 35cm. Your best guess for the mean height of all students would be 190cm. But how close to the true population mean would your sample's mean be?

The standard error of the mean (SEM) is

$$\frac{s}{\sqrt{n}}$$

which for you is

$$\frac{35}{\sqrt{49}} \quad \text{or } 5 \text{ cm.}$$

This tells you that the means of about 95% of all random samples of 49 students will be within two standard errors (10 cm) of the true population mean. In other words, there is a 95% probability that your sample mean is off by no more than 10cm. In still other words, you can be 95% certain that the population's mean height is between 180cm and 200cm.

An application

How would you know how certain you could be that your sample mean is a certain distance from the population mean?

Important Terms and Concepts

estimating population parameters

normal distribution

population parameters vs. sample statistics

random samples

sampling distribution

sampling distribution of sample means

sampling variability

standard errors

   standard error of the mean

   standard error of the proportion

   standard error of the difference between means