

2

General Notions

2.1 DATA

What do you want to know? The answer when doing surveys begins first with the question, then moves to appropriate variables, and finally rests with data. People are examined, interviewed or observed to learn more about them. The items of interest are termed *variables*. Findings based on a set of variables are recorded as *data*, to be processed and analyzed so that questions can be answered. In this book on rapid surveys we will consider only two types of data – equal interval and binomial – that account for much of what people want to know.

2.1.1 Types of Data

This chapter describes equal interval and binomial data and their average values as means or proportions. In addition, it shows how a combination of two equal interval or binomial variables becomes a *ratio estimator*, used in rapid surveys to estimate means or proportions in the population.

Equal interval variables are those that are measured with a scale consisting of equal-sized units. There are many outcomes for equal interval variables, depending on the number and size of units in the measuring scale. Conversely, binomial variables are those with only two possible outcomes, such as "yes" or "no" or 0 or 1. Bicycles and binoculars share a reference to two parts. Instead of wheels or ocular pieces, however, binomial variables feature two names or categories.

Figure 2-1

An example of the two types of data is shown in Figure 2-1. Describing this ample figure are two variables, *height* and *obesity*. We measure height with a ruler, separated into units of equal length. Thus *height* is an equal interval variable and the resulting data are equal interval data. Obesity is based on a combination of skinfold measurements, height, and weight – all of which use scales of equal interval. The information is summarized in an anthropometric index with a cutpoint for obesity. Persons above the cutpoint are termed "obese," while those below the cutpoint are classified as "not obese." The variable *obesity* with its two outcomes is a binomial variable and the data are binomial data.

Figure 2-2

A second example is given in Figure 2-2. A woman is asked her opinion of a proposed adult education program. The variable *opinion* has two outcomes, favorable – coded 1 – and unfavorable – coded 0. It is therefore a *binomial* variable. She is also asked the number of years she attended school. Years of education is a numeric scale with each year counting one unit. This is an equal

interval variable.

2.1.2 Average Value of Data

While we could present the values of measured variables for each person, it is also more useful to summarize individual data as a single average value for the group. For equal interval data, the term *mean* signifies the average value. The mean is calculated by adding the values for all persons being sampled and dividing by the total number of sampled persons:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (2.1)$$

where Σ is the sum of the calculations for all persons in the sample (n), y_i is the value of the variable of interest for person i , and \bar{y} is the mean or average value. For example, the mean years of education for a sample of five people with 10, 12, 12, 16, and 18 years of education, respectively is calculated as

$$\bar{y} = \frac{10 + 12 + 12 + 16 + 18}{5} = 13.6$$

For binomial data the average value is termed the *proportion*. It is calculated in the same way as the mean of an equal interval variable. Here, however, the binomial variable has an outcome of either 0 or 1 rather than a range of numbers. The formula is

$$p = \frac{\sum_{i=1}^n a_i}{n} \quad (2.2)$$

where Σ is the sum of the calculations for all persons in the sample, a_i is the value of the attribute of interest for person i (either 0 or 1), p is the proportion, and n is the number of sampled persons. If we want to derive the proportion who are immunized among five children coded as 0, 0, 1, 0, 1, respectively, the calculation is

$$p = \frac{0 + 0 + 1 + 0 + 1}{5} = 0.4$$

Notice that the mean of an equal interval variable is calculated in the same way as the proportion of a binomial variable. This is because a proportion is a mean, but a mean of a binomial variable.

Figure 2-3

The average value of a binomial variable is often presented as a percentage rather than a proportion. A proportion has values between 0 and 1. A percentage, being a proportion multiplied

by 100, has values from 0 to 100 (see Figure 2-3).

Figure 2-4

Equal interval or binomial data can be used to derive ratios of two variables that resemble means or proportions. These ratios are termed *ratio estimators*. As an example of such an estimator, consider a sample of three intravenous drug addicts who have injected themselves various times during the past two weeks (see Figure 2-4). One variable is the total number of injections, an equal interval variable. A second variable, also equal interval, is the number of shared injections. The ratio of the number of shared injections to the total number of injections in the group is the *proportion* of total injections that are shared. This proportion is a *ratio estimator*, slightly different from a regular proportion presented in most statistics texts. Why so? Notice that we sampled addicts, not injections. That is, for each of the three sampled addicts we counted the number of total and shared intravenous drug injections. The sampled units are drug addicts while the random variables in the sampled units are total and shared drug injections.

Figure 2-5

Another example features a sample of three households selected from a large population of households (see Figure 2-5). In this survey of households, information was collected on three variables: the number of preschool children, the number of children who had been vaccinated at least once (shown in black), and the number of vaccinations. All three are random variables because the counts vary from household to household. A ratio estimator is created with combinations of these variables to derive both mean and proportion.

If we divide the total number of immunizations (8) by the total number of children (4) in the three households, we are using a ratio of two random variables to estimate the *mean* number of immunizations per child (2.0). If we divide the number of vaccinated children (3) by the total number of children in the three households (4), the ratio of the two variables is used to estimate the *proportion* who are immunized (0.75).

The formula for a ratio estimator is

$$r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \quad (2.3)$$

where y_i and x_i are both random variables and Σ is the sum of all values in the n sampled units. Notice that the sampling units, counted from 1 to n , are different from the random variables y_i or x_i . That is, households (n) are different from children (x_i), or immunizations (y_i).

For another example, assume we did a survey of five homeless shelters and found three drug addicts in residence. These three addicts collectively injected themselves with drugs 30 times during the past two weeks (10, 8, and 12, respectively). In addition, the three addicts shared syringes in 13 of the 30 intravenous injections (6, 2, and 5, respectively). The sampling units are homeless shelters, counted from 1 to 5, and the three random variables are numbers of addicts, injections, and shared injections, respectively. Derived as a *mean*, the ratio estimator for the average number of injections

per addict in the five sampled homeless shelters is calculated with Formula 2.3 as

$$r = \frac{0 + 10 + 8 + 0 + 12}{0 + 1 + 1 + 0 + 1} = \frac{30}{3} = 10$$

Derived as a *proportion*, the ratio estimator for the proportion of injections that was shared in the five shelters is calculated with Formula 2.3 as

$$r = \frac{0 + 6 + 2 + 0 + 5}{0 + 10 + 8 + 0 + 12} = \frac{13}{30} = 0.43$$

Since the sampling units are homeless shelters, not drug addicts (a random variable) or injections (another random variable), we must use Formula 2.3 for a ratio estimator to derive the mean or proportion, rather than Formula 2.1 for a mean or Formula 2.2 for a proportion.

2.1.3 Analysis of Data

People who conduct surveys do so because they want to know something about a population but have neither time nor money to measure everyone. Data from surveys can be gathered and analyzed quickly, as long as there are not too many variables and the analysis is not too complicated. In this text the analysis will be limited to means and proportions, primarily using ratio estimators. In addition we will derive confidence intervals for the respective means and proportions.

Often, all that is needed is the average value of a variable. For example, if a rapid survey is being done to assemble knowledge of acquired immune deficiency syndrome (AIDS), the outcome may be the proportion (or percentage) of a sample who know how the disease is transmitted. If the survey is of smoking habits, the outcome might be the proportion who currently smoke. If blood pressure is the topic of interest, the outcome may be the mean systolic or diastolic pressure (if analyzed as equal interval data) or the percentage who are hypertensive.

More advanced statistical tests can be conducted on rapid survey data but require sophisticated formulas beyond the scope of this text. As you will see, data from rapid surveys of people in households, schools, census tracts or villages have a greater variance than expected by statistical tests featured in introductory textbooks. These statistical tests assume that people counted in surveys are independent of one another with respect to their characteristics, practices, attitudes, knowledge. Clearly this may not always be the case, occur especially in households, neighborhoods, schools or small villages where people tend to think and act in similar ways. In many instances, standard variance formulas featured in most introductory statistics texts tend to *under-estimate* the variability of data derived from a rapid survey. Thus, we will be using a different set of variance formulas; those that calculate the variability of survey data measured as ratio estimators.

Figure 2-6

Rapid surveys do not measure everyone in the population. Instead, they sample persons selected to represent the population. With this sample, the intent is to *estimate* the true mean or proportion in the surveyed population for the variable of interest. In the surveyed population the average value is designated as Y if an equal interval variable or P if a binomial variable. There is only one true value of Y or P in a population and therefore at any moment the average value is fixed (see Figure 2-6, left).

When drawing a sample from a population, the average value of a given variable is not fixed. Instead it can have many values, depending on the combination of persons included in the sample. Rather than using capital letters, the mean and proportion in a sample are cited as \bar{y} and p , respectively. In doing our calculations, we cannot state with complete certainty that the sample mean or proportion is equal to the true mean or proportion in the population. To show this uncertainty, we present for samples an interval that brackets the true value with a given level of confidence – usually 95 percent (see Figure 2.6, right). The interval surrounding \bar{y} and p is termed the *confidence interval*.

Figure 2-7

When analyzing study findings, we start with the data and derive an estimate of the mean, variance of the mean, and standard error (see Figure 2-7). Then we use the mean and standard error to compute the confidence interval for the variable of interest. The mean is easy to calculate – you sum the individual values and divide by the number of people as shown in Formulas 2.1, 2.2, or 2.3. You do not need to know much about statistics to derive a mean. What is more difficult, however, is to calculate the variance of the mean.

The formulas necessary to calculate the variance of rapid survey data are not taught in introductory statistics classes or presented in most statistics texts. Instead you must consult statistical sampling books that contain often-complicated formulas for a variety of survey designs. For rapid surveys the variance formulas will be presented and explained in the coming chapters. Only a few formulas are needed to do rapid surveys, but many more are useful to understand the logic of rapid surveys. Mastering the statistical logic of rapid surveys should make it easier to follow the mathematical formulas and logic of more complicated survey designs.

Figure 2-8

Formulas cannot easily be described with words. Instead they are typically represented with symbols. I have already noted that the mean of equal interval data is \bar{y} . The terms for the variance of the mean, standard error and confidence interval are shown in Figure 2-8. Observe that the confidence interval is the mean plus or minus z times the standard error. That is, the lower limit of the confidence interval is \bar{y} minus z times the standard error while the upper limit is \bar{y} plus z times the standard error. The term z is a number derived from the standard normal distribution and will be explained later in this chapter and in Chapter 3.

Figure 2-9

For binomial data, the process of creating a confidence interval is similar except that we use the proportion p and the standard error of the proportion $se(p)$ to derive the confidence interval (see Figure 2-9). Finally, the concept of analysis is also the same for ratio estimators, as shown in Figure 2-10, although the formulas are different from those presented in most introductory statistics textbooks.

Figure 2-10

Surveys are samples of households or people drawn from a population. If the sample is

drawn in an unbiased manner, we can use it to estimate the mean or proportion in the population (see Figures 2-11). In a sample, the mean or proportion is influenced by its variance – a parameter that is estimated from the sample data. The mean or proportion in a population is fixed. That is, Y or P has only one value in the population, typically called the *true* value in the study population. If the sample is selected in an unbiased manner, \bar{y} or p in the sample will on average equal Y or P in the population.

Figure 2-11

If the analysis uses the ratio estimator r to derive the mean or proportion, there may be a small bias, as shown in Figure 2-12, but it is often not large enough to effect the accuracy of the findings. If care is taken in the sampling procedure (as explained in Chapter X), r provides an acceptable estimate on average of R , the ratio estimator of the mean or proportion in the population.

Figure 2-12

Notice that I stated that the sample findings *on average* will estimate the true value in the study population. How should we interpret *on average* since our survey is done only once? This point will be further discussed in the section on **Variability and Bias** below. For now *on average* means that if the population at some moment in time had been sampled over and over again, the average value of all samples would be the same as the true value in the population. Of course the values of \bar{y} , p or r would not be the same from one sample to the next. Sometimes the values would be too high, other times too low, and still other times very close to the true value. We will use the sample data to estimate how much \bar{y} , p or r would vary from one sample to the next. That is, we use the variability of the data to estimate $v(\bar{y})$, $v(p)$ or $v(r)$, the variance of \bar{y} , p or r , respectively. By taking the square root of the variance, we derive the *standard error*, shown in Figures 2-11 and 2-12 as $se(\bar{y})$, $se(p)$ or $se(r)$. The standard error is then combined with the mean, proportion or ratio estimator to calculate the *confidence interval* for our estimates.

2.1.4 Data and Action

After all the mathematical manipulations have taken place, we are left with a mean or proportion and a confidence interval that may be wide or narrow, depending on the size of the sample and the characteristics of people selected for the sample. So what do we do with the information?

There are two important questions that should be asked about potential information before actually doing a survey. First, is the anticipated information of value for planning or improving a program or activity, or for understanding a research problem? Second, will the information be worth more than the cost of doing the survey?

These two issues – utility and cost – are central to planning in a variety of fields and for many activities. The same principles apply when doing surveys. Money spent gathering data cannot be spent delivering services or helping people in need. On the other hand, poor allocation of service resources may waste money – funds that could otherwise be spent more efficiently if only more had been known about the needs of the population. Data from rapid surveys are very useful for action-oriented people.

Figure 2-13

The link between survey data and action is shown in Figure 2-13. Data are first gathered in a survey. As part of the analysis, the raw data are converted to means or proportions and serve as information. If collected in an unbiased manner and clearly presented to those in power, the information is converted to knowledge of the population. Text, tables and graphs help convert information to knowledge. Once knowledge is in the mind of the administrator or policymaker, it may cause action. I say "may," because the information may not be germane to the decisionmaking process. That is, the survey specialist may have gone off on a tangent that holds little interest for those charged with action. If so, information becomes very costly since money is spent gathering data with no cost savings arising from use of the data. To be cost-effective, rapid surveys must respond to the needs or interests of those persons holding the purse-strings. No statistical theory or mathematical wizardry can overcome such a flaw in focus.

An easy check on the eventual use of data is to ask the person making decisions to describe for key variables the different actions that might be taken (see Figure 2-14, left side). Will action be different if the value in the population is high versus middle or low? If the answer is "yes," the data will be well used. If the answer is "no," the survey findings will have less value.

Figure 2-14

The set of possible actions also helps determine the number of persons to be surveyed. Most people planning a survey are concerned with the size of the sample. While the answer may appear to be entirely a statistical matter, it is not. Instead, the answer depends on the set of actions to be taken based on the study findings (see Figure 2-14). If there are only a few actions and the range of values for decisionmaking is wide, then the value in the population does not have to be determined with great certainty. That is, the confidence interval could be quite wide, as occurs with smaller surveys of 200 to 300 people. Conversely, if there are many potential actions and knowing the exact mean or proportion is critical to choosing the best action, more people would need to be sampled.

Figure 2-15

An example of action levels for a family planning program is shown in Figure 2-15. Assume that program managers in a developing country are interested in delivering family planning services to women in need. They reason that if more than 60 percent of the eligible women in a region are currently using family planning services, adequate saturation of the community has occurred. Thus, if this is the study finding, no further action is necessary. They also have guidelines that state that immediate action is necessary if less than 20 percent of the eligible women are using a family planning method. These actions include community education programs and efforts to improve local family planning services. For the middle range between 20 and 60 percent, the administrators would conduct a series of smaller studies of non-users to find out what the problems are and try to improve the management of the local family planning program. With these guidelines, the program manager is interested in knowing only if family planning use is in the high, middle or low range, not in the exact percentage. A small survey with wide confidence intervals would be adequate to address this issue. By knowing the

action ranges, the survey specialist can plan rapid surveys of modest size that will satisfy the needs of the administrator but not cost more than the program can afford.

2.2 VARIABILITY AND BIAS

Rapid surveys tell us about peoples' characteristics, thoughts, illnesses, practices and much more. Yet the information from rapid surveys will not be accepted by decisionmakers unless they have faith in the *unbiased* nature of the sample estimate. Repeated sample surveys of the same population will come up with slightly different answers, yet the sample may on average still be unbiased. How do we describe this variability among repeat sample surveys for decisionmakers or policymakers, given that only one survey was done? More important, how do we know if the estimate from our *one* survey is too high or too low? To answer these questions we need to understand the terms *precision* and *accuracy* and the role of *confidence intervals*.

2.2.1 Accuracy and Precision

The concepts of sampling become much easier to understand if it's assumed that we have a lot of money and time – so much money and time that the same sample survey can be done over and over again. The sample means, proportions or ratio estimators for these repeated surveys can then be used to find the variability that exists from one sample of a population to the next.

Figure 2-16

Assume that we are interested in knowing the percentage of women who currently use a contraceptive method for family planning purposes. The true value in the population is 50 percent. This true value would not be known, but is mentioned here so that we can see the effects of sampling. We draw a small sample of 20 women, the results of which are shown in Figure 2-16. Seven of the 20 women report using a family planning method – or 35 percent of the sampled women. These women comprise *one* sample drawn from the population of interest.

Figure 2-17

Now I will stretch your imagination somewhat. Assume that the sampling process is repeated over and over again. Each sample survey is of the same number of women, and is done at the same time (easy to imagine, but hard to do). The results of 64 such repeated surveys are shown in Figure 2-17. Our single survey – with 35 percent using a family planning method – sits among the 64 sample surveys. The means of the various repeated surveys range from 25 percent to 75 percent, with most percentages falling near 50 percent, the true value.

The terms *precision* and *accuracy* in Figure 2-17 refer to the variability among the 64 sample surveys. *Precision* is the variation of the survey means in relation to the average value for all surveys combined, often termed the "expected value." Here the average value is 50 percent, so precision is the variation of the percentages in the 64 individual surveys from 50 percent. *Accuracy* also refers to variation of individual surveys, but in relation to the true value in the population rather the average value of all samples. If the true value in the population is the same as the average value of repeated surveys – as is so in our example – then precision equals accuracy.

Figure 2-18

2.2.2 Bias

A third term, *bias*, helps clarify the distinction between precision and accuracy. *Bias* is the deviation of the average value of all possible samples from the true value in the population (see Figure 2-18). If there is a difference between the two – as shown on the left side of Figure 2-18 – then the sample, on average, is biased. If the average value and true value are the same – as shown in the right side of Figure 2-18 – the sample, on the average, is *unbiased*. Notice that just because a sample is unbiased, the value of the one survey actually done (shown in black) will not necessarily be the true value. Instead, being unbiased implies that if our survey was done repeatedly, the average value of the sample mean for the different surveys would equal the true value in the population.

Figure 2-19

Next, we see in Figure 2-19 how bias effects the relationship between precision and accuracy. If a sampling method is biased (seen in the left side of Figure 2-19), the level of precision will be less than the level of accuracy, since accuracy reflects both precision and bias. If the sample is unbiased (the right side of Figure 2-19), accuracy and precision will be the same.

On the surface, *precision* is not a very useful concept. After all, why should we care about the deviation of our sample from the average value of all sample means or proportions? Instead, our real interest should be in *accuracy* since it is the true value of a variable in a population that we are after. After all, the highest compliment for a sample survey is that it is accurate, not that it is precise.

The problem is that we usually cannot measure accuracy. What is missing is knowledge of the true value in the population. Of course, if the true value were known, why would we do a survey? Fortunately, even lacking truth we can estimate the *accuracy* of our survey, assuming the sampling method is unbiased. The path to accuracy, however, leads first to precision. With some statistical manipulations of the survey data, we derive a variance of the sample mean that allows us to estimate precision. When the sample is selected and analyzed in an unbiased manner, our measure of precision will equal accuracy, thereby giving us what we want. So how do we measure precision?

2.2.3 Standard Error

Precision is defined in a general way as the inverse of the variance of the sample mean (or proportion) in the population. That is, the smaller the variance of the mean, the greater the level of precision. A more useful term for understanding precision, however, is the standard error – the square root of the variance of the mean (or proportion). It has the same units as the mean or proportion and therefore is easier for most people to understand. If the mean is measured in centimeters, the standard error is also measured in centimeters. If the outcome is a proportion or percentage, the standard error is also stated as a proportion or percentage.

The formula presented in most statistics books for the standard error of the mean, $se(\bar{y})$, is

$$se(\bar{y}) = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n(n-1)}} \quad (2.4)$$

where the bracket is the square root of the formula, Σ is the sum of the calculations in the parentheses for all persons in the sample, y_i is the value of the variable of interest for person i , \bar{y} is the mean for all values of y_i , and n is the number of sampled persons. The standard error of the proportion, $se(p)$, is

$$se(p) = \sqrt{\frac{p q}{n-1}} \quad (2.5)$$

where the bracket is the square root of the formula, p is the proportion with the attribute, q is the proportion without the attribute, and n is as previously defined. Formulas 2.4 and 2.5 are correct for a simple random sample of a population but not for the more involved sampling scheme used with rapid surveys. Nevertheless, the formulas serve to introduce the concept of a standard error.

Figure 2-17 showed that when women were repeatedly sampled in a survey of family planning methods, most sample findings were close to 50 percent, the true value. Some were ten percentage points from the true value (that is, 40% and 60%), while a few were 20 or more percentage points from the true value (30% or less and 70% or more).

Statisticians for centuries have noted that the frequency distribution of the means of samples repeatedly selected from the same population resembles the bell-shaped curve of the well-known normal distribution (see Figure 2-20). Their observation is applicable to both sample means and proportions. The horizontal axis of the normal distribution is generally labeled as standard error units rather than *scale (Y or P)*, as shown in Figure 2-20. The units on the horizontal axis measure the deviation of each sample mean or proportion from the average value of all possible samples, termed the *expected value*. The deviations from the expected value are measured in multiples of the *standard error*, a statistic that has the same units as the mean or proportion. This use of the normal distribution – taught in all introductory statistics courses – is central to the theory and practice of sampling statistics.

 Figure 2-20

Now we have two ways to describe the variability of means or proportions from replicate samples: first with the term *precision* and second by the position in the normal distribution, using standard error units. But why confuse matters with standard error units when other units of measurement such as percentages, centimeters or kilograms are easier to understand?

While the size of the deviation from the expected value is important to know, the units we use to describe the deviation are less important, as long as everyone understands what the units are. For example, when measuring the width of a highway, it does not matter if the units are yards, meters, or lengths of an automobile. While the numbers may be different, the distance is always the same. If the width of a 22-meter highway is measured with a 5.5 meter automobile,

the width would be four car lengths. Measured with a metric ruler, it would be 22 meters. Thus both the metric ruler and the automobile are measuring the same thing, but with scales of different units. The same holds true when measuring the deviation of sample means or proportions from the expected value in a population. The measuring units could be centimeters, kilograms, percentage points, or standard errors. So how do we convert the scale of measured units (for example, percentage points) to standard errors?

Consider again the example of the family planning survey mentioned previously. As shown in Figure 2-16, our small survey of 20 women found that 35 percent were currently using a family planning method. The true percentage of users in the population from which the sample was derived was 50 percent. When the same sample was drawn repeatedly, Figure 2-17 shows that some sample values were well above 50 percent and others well below 50 percent. Each of these sample values – shown as percentages – can be converted to standard error units, using knowledge of the variability of the individual samples to derive the standard errors. This conversion process is illustrated in Figure 2-21.

Figure 2-21

Figure 2-21A starts with the bottom row of the sample distribution shown in Figure 2-17 for repeated family planning surveys. The value of our one survey is 35 and the expected value is 50. Figure 2-21B shows the deviation of each sample value from the expected value; minus 15 percentage points for our one survey. The standard error is derived for each survey in Figure 2-21C, using Formula 2.9. Observe that the standard error – in units of a proportion – is multiplied times 100 to derive units of a percentage point. For our example, using Formula 2.5 the standard error is calculated as

$$se(p) = \sqrt{\frac{.35 (.65)}{20 - 1}} \times 100 = 10.9\%$$

Since our single sample survey is 15 percentage points below the expected value and a standard error unit is 10.9 percentage points, the survey value is -15 divided by 10.9 or -1.4 standard error units from the expected value (see Figure 2-21D). When the same calculations are done for all 64 repeated sample surveys, Figure 2-17 can be redrawn with a new horizontal axis, *Standard Errors*, as shown in Figure 2-22.

Figure 2-22

The principle that was illustrated with the family planning surveys is central to sampling statistics. That is, means or proportions of repeated sample surveys are distributed in a manner similar to the normal distribution. Sometimes this does not hold true, as when sampling rare events or persons with mainly high or low values. Yet most of the time the theory is valid and is very helpful for analyzing the variability of rapid surveys.

Figure 2-23

If the same survey is done repeatedly and in an unbiased manner, the mean or proportion of most results will lie within a few standard error units of the expected value (see Figure 2-23).

Some will be further than one and a half to two standard error units from the expected value. Only few of the repeated proportions or means will be more than two and a half to three standard error units from the expected value. Although we would only do *one* sample survey, there is an underlying distribution of all possible samples that could have been done, similar to the distributions shown in Figure 2-23. What is not known is where our one sample survey lies in the distribution.

Precision, as mentioned previously, is related to the inverse of the variability of the means (or proportions) of the repeated surveys. The more precise a measurement, the smaller the degree of variability, as measured by the variance of the sample mean (or proportion). Since the standard error is the square root of the variance, it also is inversely related to precision. If two surveys measuring the same variable are done with the same number of subjects, the more precise survey is the one with the smaller standard error.

Precision is not an absolute term; there is no cutpoint separating precise from imprecise. Yet in common usage, we would say that a survey is precise if the standard error is small in relation to the mean or proportion. This relative measure, termed the *coefficient of variation*, is defined for a mean as

$$cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}} \quad (2.6)$$

where $se(\bar{y})$ is the standard error and \bar{y} is the mean. The coefficient of variation for a proportion is

$$cv(p) = \frac{se(p)}{p} \quad (2.7)$$

where $se(p)$ is the standard error and p is the proportion. If cited as a percentage, both $se(p)$ and p are multiplied by 100.

Two surveys illustrate how we can use the *coefficient of variation* to describe in general terms the *precision* of a variable. The first survey is a sample survey of young children, 90 percent of whom are vaccinated for measles with a standard error of 3 percent. We would regard the survey findings to be very precise since the standard error is only one-thirtieth the size of the sample mean (a percentage). Specifically, using Formula 2.7, the coefficient of variation is calculated as

$$cv(p) = \frac{0.03}{0.90} = 0.033$$

In our second survey, we are measuring HIV infection in a low risk population. Here, the same standard error of 3 percent would be considered very imprecise. Assume the prevalence of HIV infection is measured as 0.3 percent. The standard error of 3 percent would then be ten times the size of the sample mean (a percentage). Again, using Formula 2.7, the calculation of the coefficient of variation is

$$cv(p) = \frac{0.03}{0.003} = 10.0$$

This example shows that it is helpful to relate the standard error to the mean (or proportion)

before describing a measured variable to others as precise or imprecise.

Figure 2-24

Another view of the distribution of proportions from repeated samples is shown in Figure 2-24. Here we assume the samples were large – say 500 to 1,000 persons in each – and were drawn by repeated random sampling of the underlying population. Note that 90 percent of the repeated samples have proportions within 1.64 standard error units of the expected value. As previously observed in Figure 2-23, the expected value for the distribution lies at 0 standard error units. Ninety percent of the samples are within plus or minus 1.64 standard error units of the expected value, 95 percent are within plus or minus 1.96 standard error units of the expected value, while 99 percent lie within plus or minus 2.58 standard error units. So how does this knowledge help us to describe the variability of our one rapid survey? The answer lies with the *confidence interval*.

2.2.4 Confidence Interval

For every proportion in Figure 2-24, we draw a horizontal line on both sides, each side being 1.96 standard errors in length (see Figure 2-25). Instead of a distribution of proportions, we now have a distribution of horizontal lines or intervals, as shown in Figure 2-25 for four of the many possible samples. If repeated for all possible samples in Figure 2-25, most of the intervals will bracket the *expected value* of all possible samples. If the sampling method is unbiased, they would also bracket the *true value* in the population. Which intervals will not bracket the expected value? The answer is those few sample surveys with proportions far from the expected value. One such survey is shown in the bottom left of Figure 2-25 with p more than 1.96 units on the negative side of 0.

Figure 2-25

If the sample means or proportions are normally distributed, 2.5 percent of them would lie more than 1.96 standard error units below the expected value and 2.5 percent will be more than 1.96 standard error units above the expected value (see Figure 2-24). Therefore intervals of 1.96 standard errors would not enclose the expected value for five percent of all possible samples. Conversely, 95 percent of all intervals of plus or minus 1.96 standard error units would bracket the expected value. The interval of plus or minus 1.96 standard error units is termed the *95 percent confidence interval*.

Figure 2-26

If we had drawn 100 samples from the same population, calculated a 95 percent confidence interval for each and plotted the confidence intervals in ascending order, the values would look like those in Figure 2-26. On average, five of the 100 confidence intervals would not bracket the expected value of all possible samples. In Figure 2-26, three of the five intervals are for proportions below the expected value (labeled as 1,2,3) while two are for proportions above the expected value (labeled as 4,5).

A more typical view of the confidence intervals for 100 repeated samples is seen in Figure 2-27. Here the order is random. For some sample surveys, the interval lies above the

expected value and for others it's below. If we did just one survey, we could not say where it would lie. It might fall below the expected value, right on the expected value or well above the expected value. What we can say, however, is that *in advance of sampling* if the selection is unbiased there is a 95 percent probability that the interval we create with 1.96 standard error units will enclose the expected value of all possible samples in the population.

Figure 2-27

Some might want to use the term *probability interval* instead of *confidence interval*. This would not be correct. *Probability* and *confidence* are related but different concepts. Probability refers to events that have not yet happened. Accordingly you might talk about the probability of it raining tomorrow, or the probability of a measles epidemic in the coming months. Once the event has occurred, the probability of it occurring is one, or certainty. If the event had not occurred, the probability is zero. Confidence has a different meaning, both in common and statistical usage. In common usage, it refers to a personal feeling of certainty or being free from doubt. In the world of sampling, the level of confidence is a reflection of how convinced a person is that an event has happened.

Figure 2-28

Coins may help to explain the difference between probability and confidence (see Figure 2-28). Assume that someone has three coins labeled heads on one side and tails on the other. All of the coins will be flipped in the air and when they come down, hidden in a covered box. Before the flips, we could say that if the coin and flipping process are unbiased the probability is 0.5^3 or 0.125 that all three will be heads. Once the flips have occurred, however, the statement no longer holds. Of course the coins are placed in a box so that we cannot see them. Yet the flips have occurred. Either three have landed with heads up or they have not. Thus, the probability of three heads *after* the flips is either 0 or 1. Since the coins are hidden in a box after the flips, we need another word to describe our conviction about the outcome – that word is *confidence*. After the coins have been flipped but before we know the outcome, we could say that we are 12.5 percent confident that all three of the hidden coins show heads. Of course we are assuming the coin was unbiased and that our calculation of the outcome probability is correct.

Sampling presents a similar situation. *Before* sampling takes place, we can say that the probability is 95 percent that a created confidence interval will enclose the true value in the sampled population (assuming of course that there is no bias). *After* sampling has taken place and we have derived our confidence interval, the outcome has already occurred. The true value in the sampled population is either inside or outside the calculated interval. Since we do not know the true value, we cannot be certain that it is bracketed by the interval. Yet from our understanding of statistics, we are 95 percent *confident* that an interval created with 1.96 standard error units will surround the true value. Conversely, this also implies we are 5 percent confident that the interval does not bracket the true value. This interval is correctly termed the *confidence interval* and not the *probability interval*.

A greater sense of certainty requires a larger confidence interval. Thus if we had used 2.58 standard error units to construct the interval, we could be 99 percent confident that our single interval encloses the true value. Conversely, a narrower interval of 1.64 standard error units corresponds to a confidence level of only 90 percent. In general the more confident we

want to be, the wider we must make the confidence interval. Unfortunately, if the confidence interval is too wide, the information may no longer be useful for decisionmaking. To reduce the size of the confidence interval we can either reduce the level of confidence, say from 99% to 95% to 90%, or reduce the size of the standard error of the sample survey. Methods for reducing $se(\bar{y})$ or $se(p)$ will be presented in the coming chapters.

2.2.5 Summary

If the sample has been drawn in an unbiased manner, the mean or proportion of the sample will on average be the same as the true value. Sometimes the value will be higher, other times lower. For all possible samples, however, the average value and the true value will be the same when the sampling procedure is unbiased.

Variability of the proportion or mean among all possible samples is termed the *precision* of the sample. It is estimated by the *standard error* and represented by the *confidence interval*. If precision is high, the confidence interval will be narrow. If the precision is low, the confidence interval will be wide. The standard error derived for one survey is used to construct the confidence interval. If the sample is unbiased and the interval is constructed to be 1.96 standard errors in length on either side of the sample proportion or mean, then we can be 95 percent confident that the interval brackets the true value in the study population.

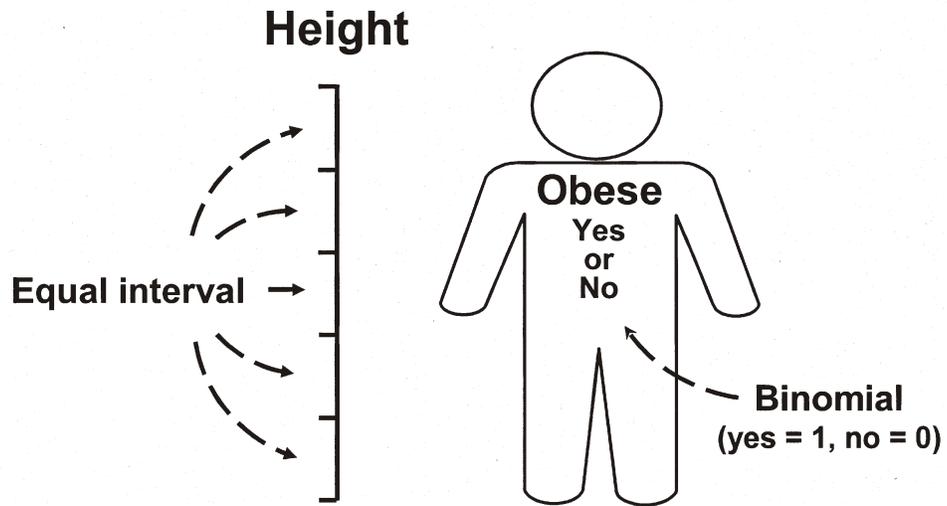


Figure 2-1. Height and obesity as equal interval and binomial variables.

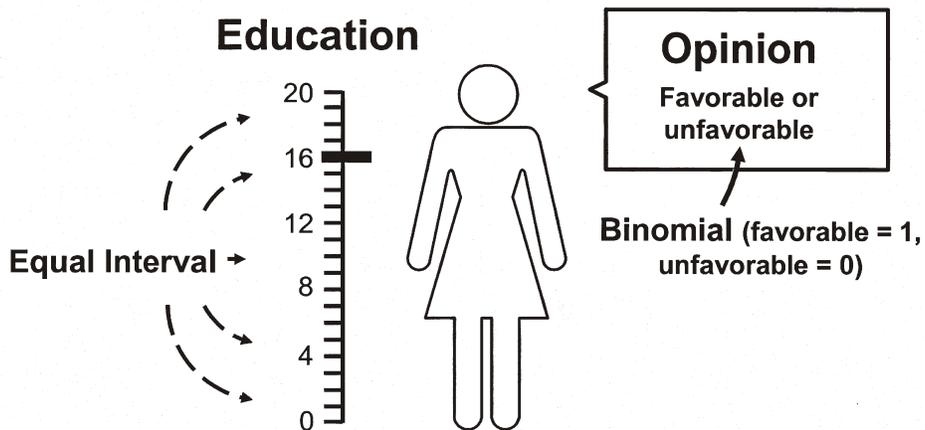


Figure 2-2. Opinion and education as equal interval and binomial variables.

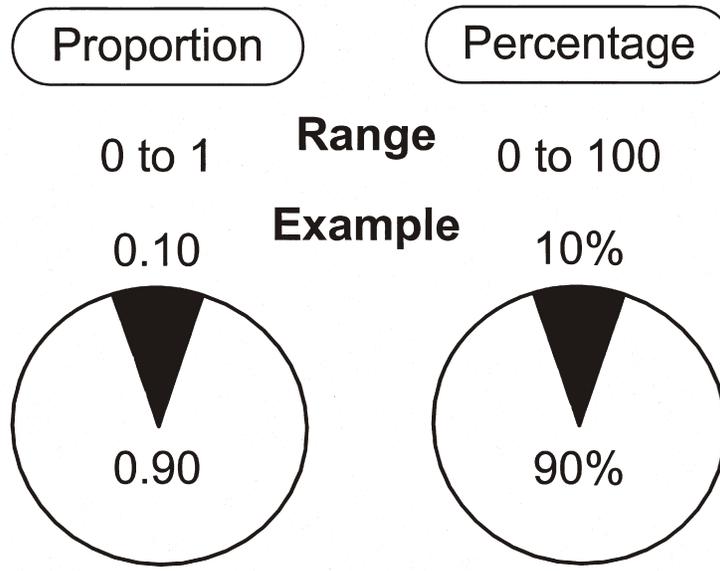


Figure 2-3. Range and example of proportions and percentages.

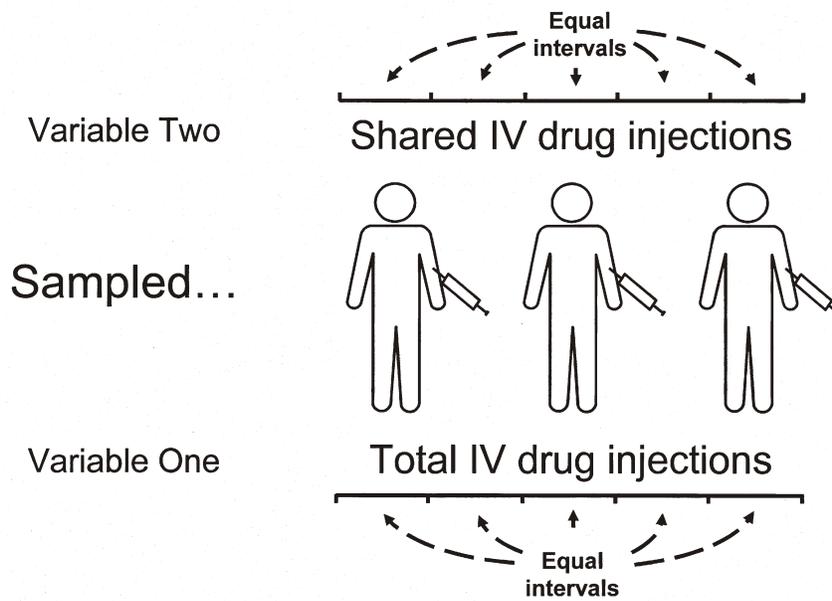


Figure 2-4. Total and shared IV drug injections among addicts as equal interval variables and a ratio estimator.

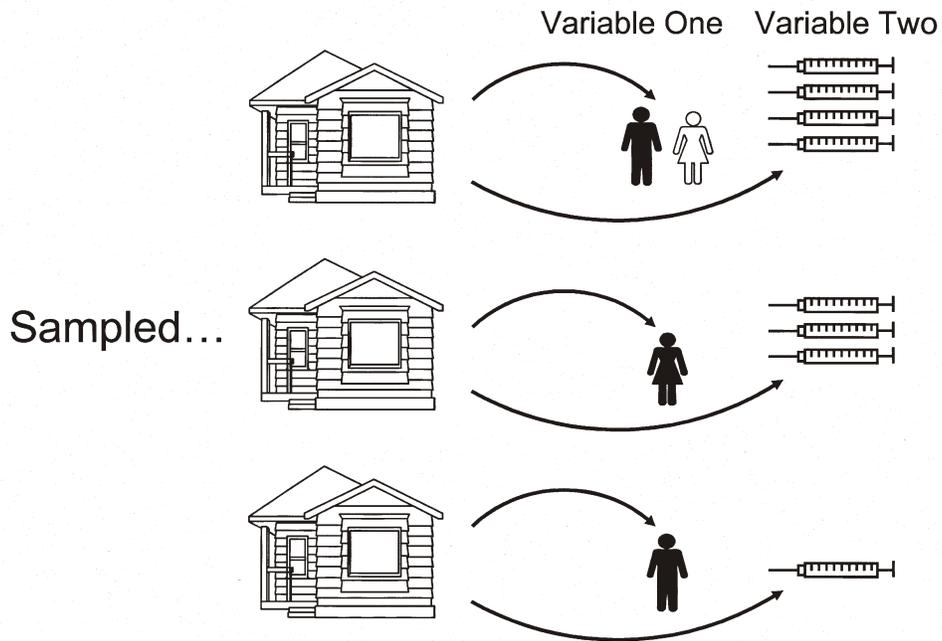


Figure 2-5. Household immunization survey and ratio estimators of a mean and a proportion.

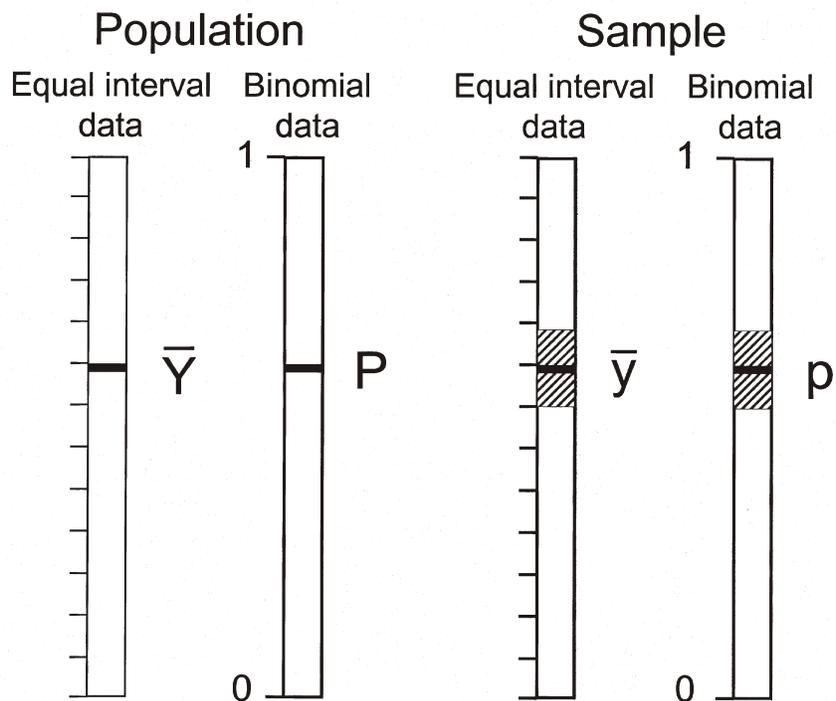


Figure 2-6. Mean and proportion in population and sample.

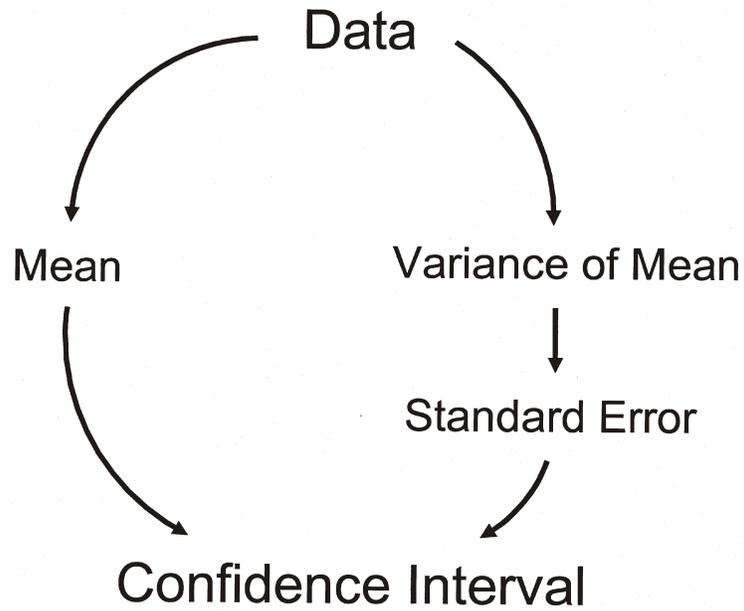


Figure 2-7. Changing data into a mean and confidence.

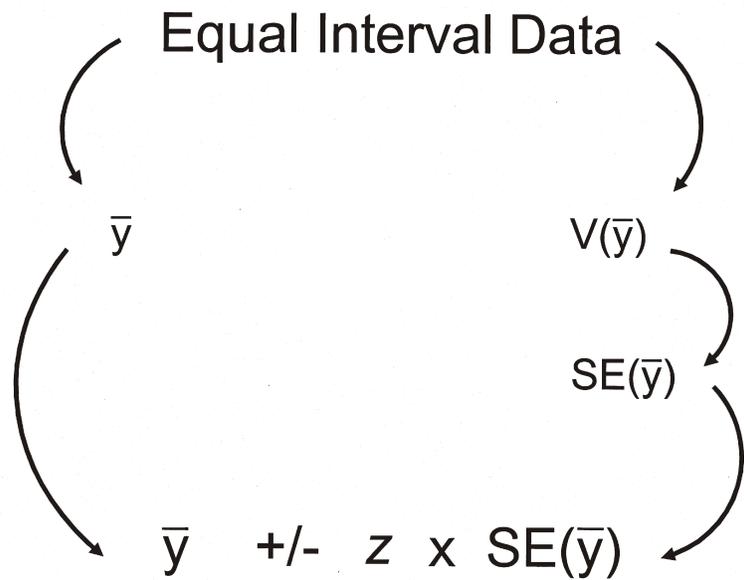


Figure 2-8. Changing equal interval data into a mean and confidence interval.

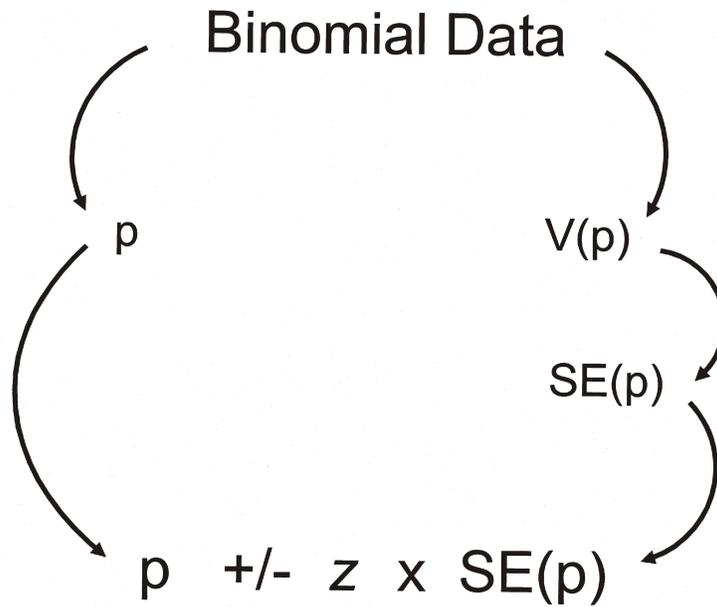


Figure 2-9. Changing binomial data into a proportion and confidence interval.

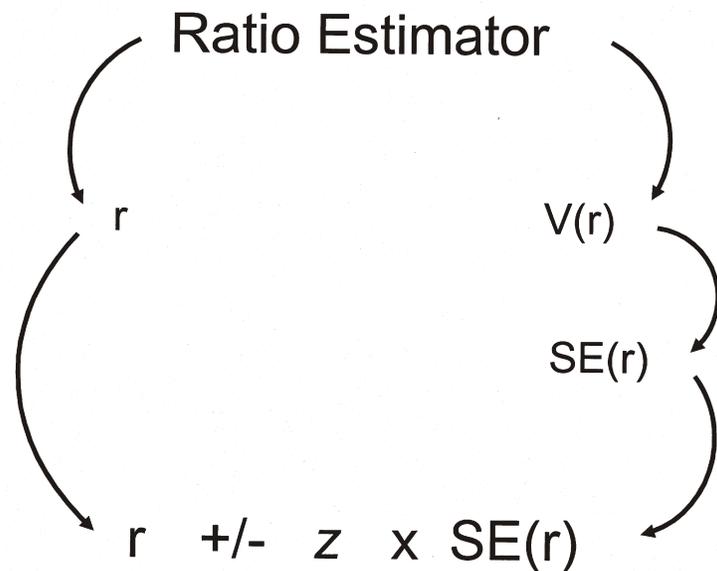


Figure 2-10. Changing ratio estimator data into a proportion or mean and confidence interval.

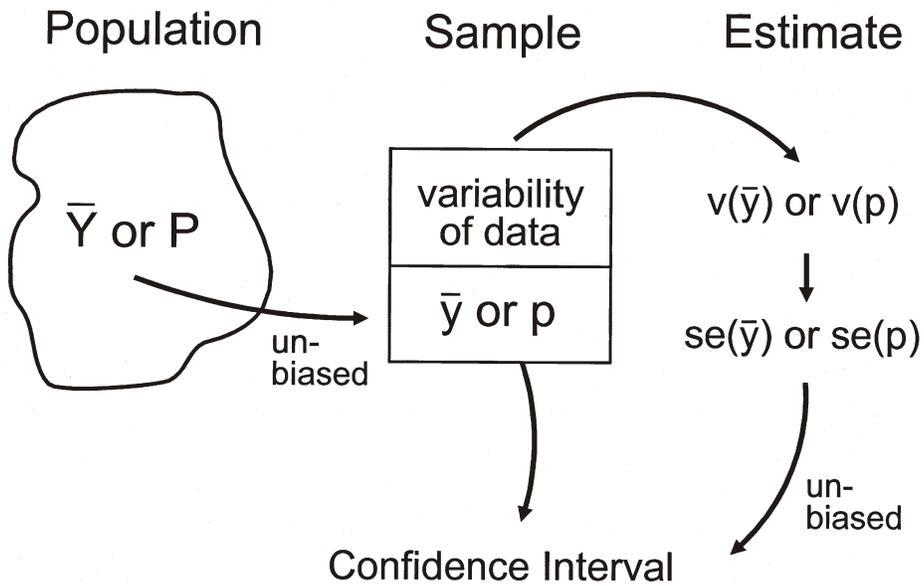


Figure 2-11. Derivation of the mean, proportion and confidence interval after sampling a population.

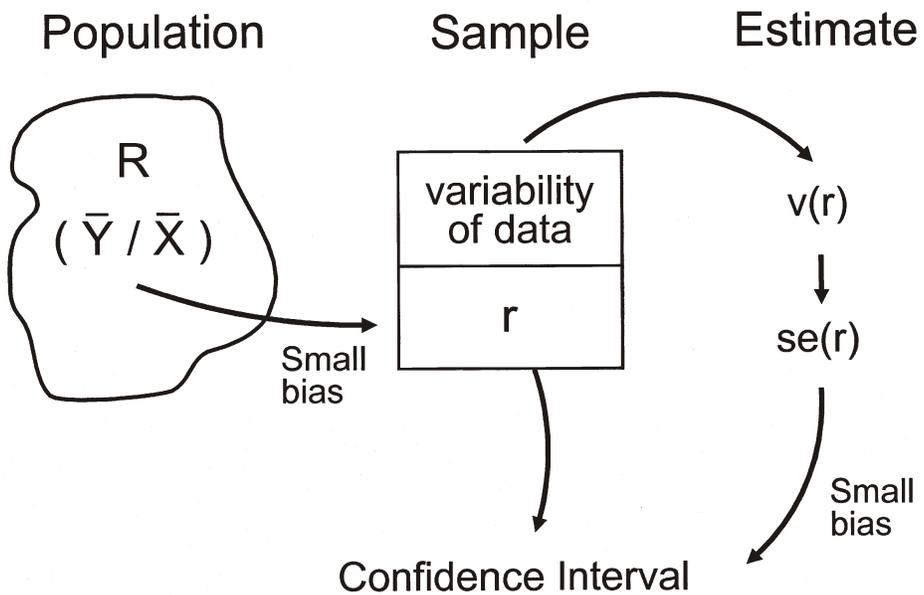


Figure 2-12. Derivation of the ratio estimator and confidence interval after sampling a population.

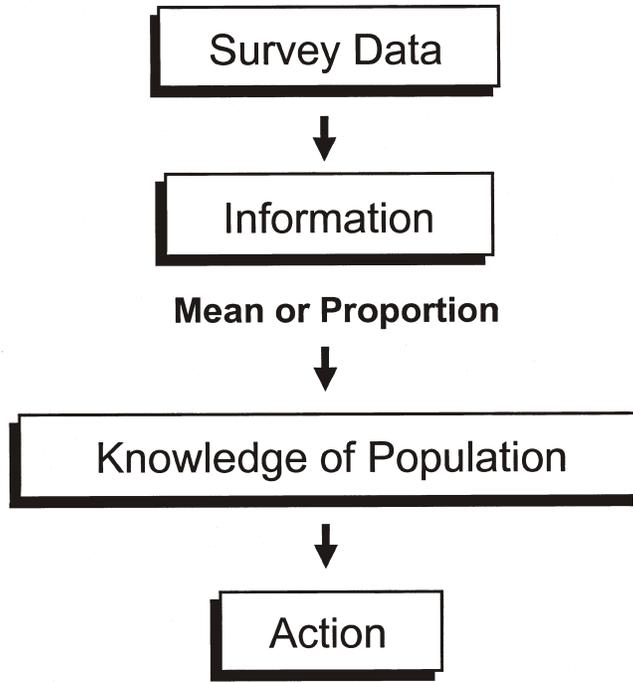


Figure 2-13. Flow of survey findings from data to action.

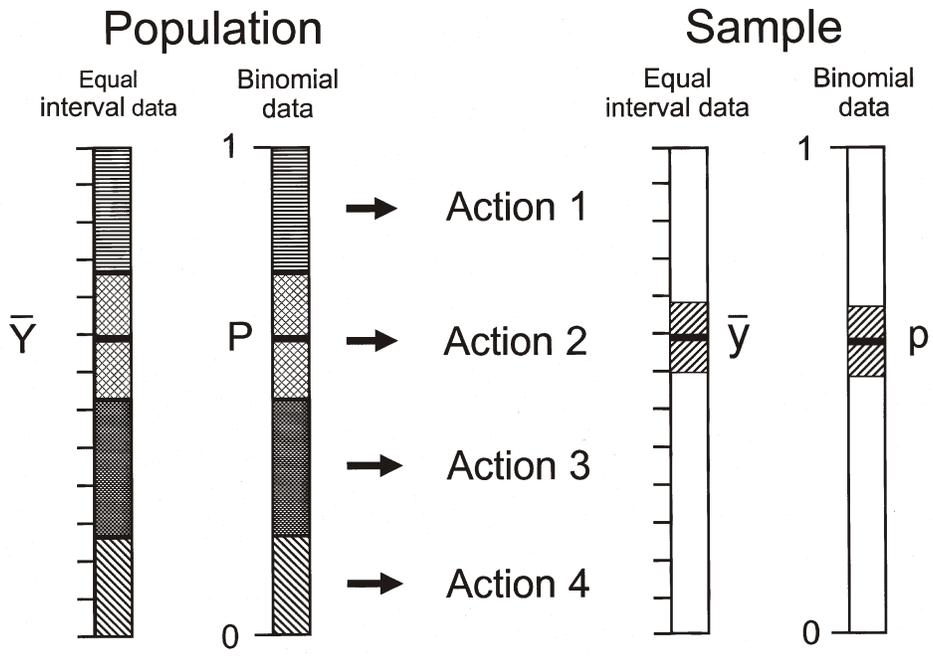


Figure 2-14. Actions based on mean or proportion in population.

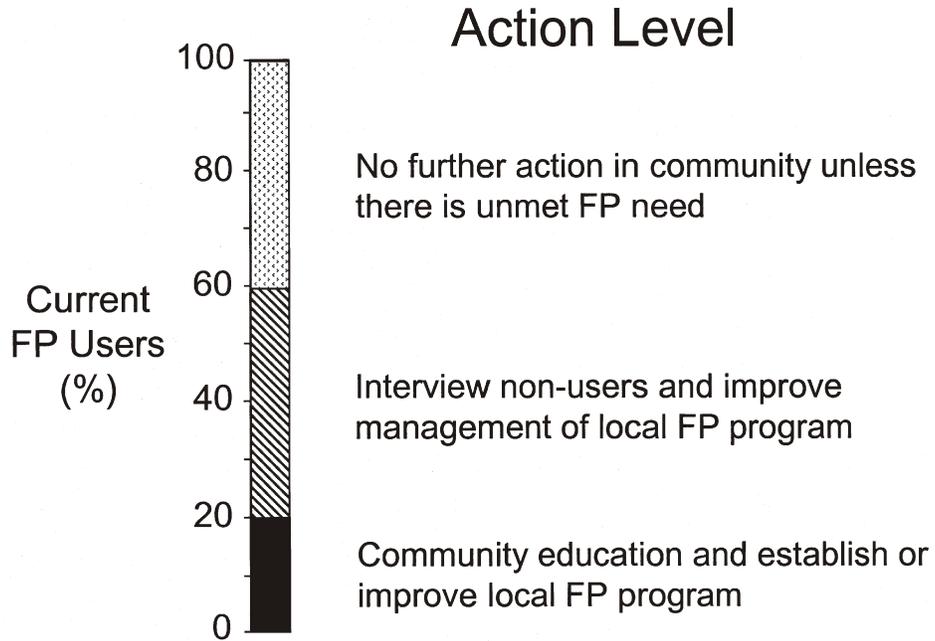


Figure 2-15. Action levels for a family planning survey.

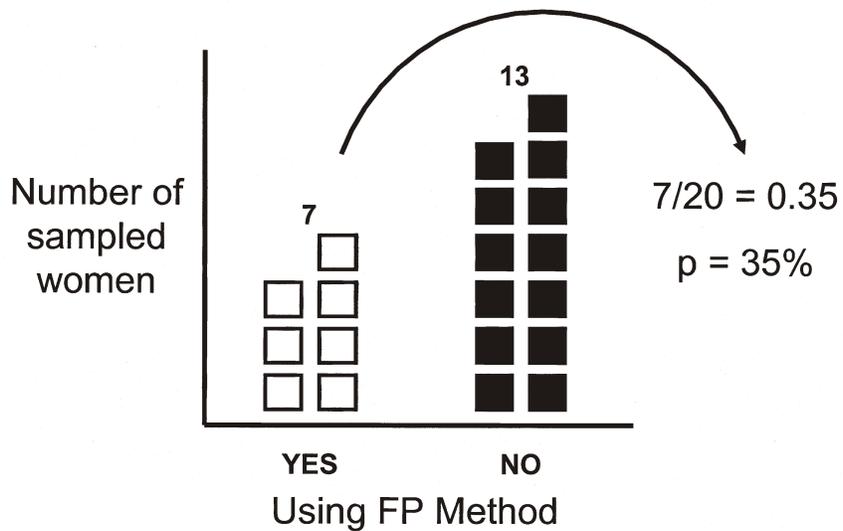


Figure 2-16. Survey of women currently using a family planning method.

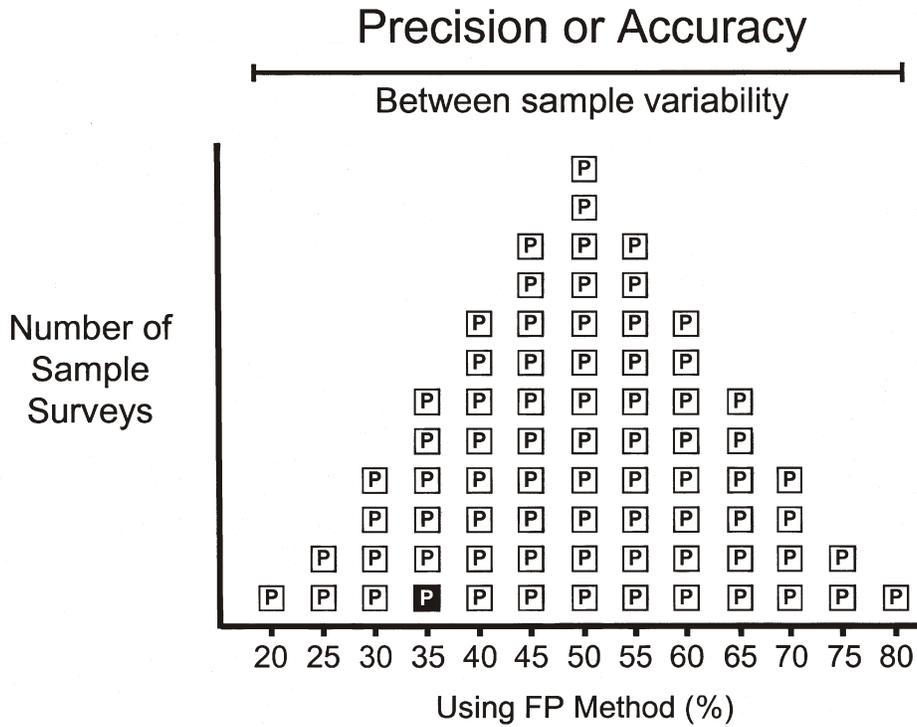


Figure 2-17. Repeated samples of use of family planning methods.

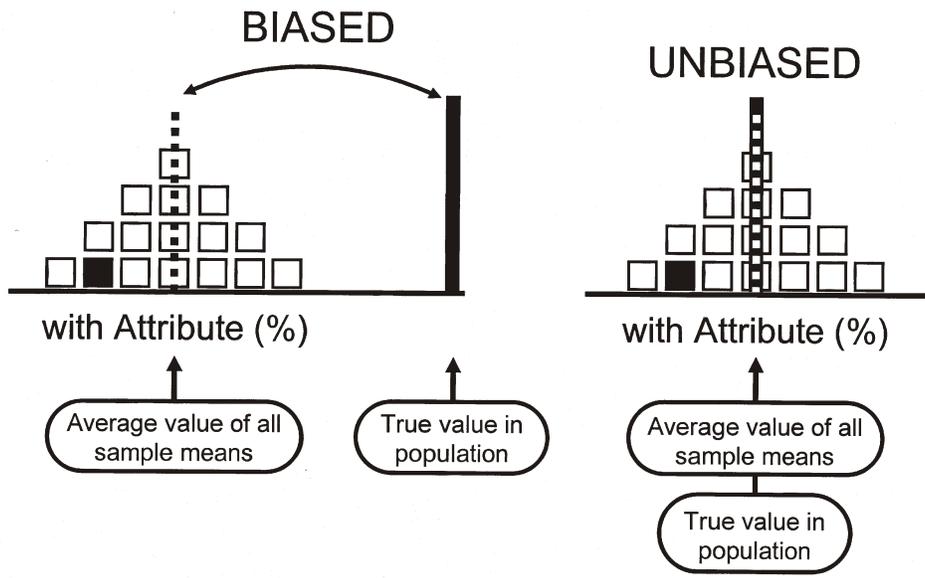


Figure 2-18. Biased and unbiased samples.

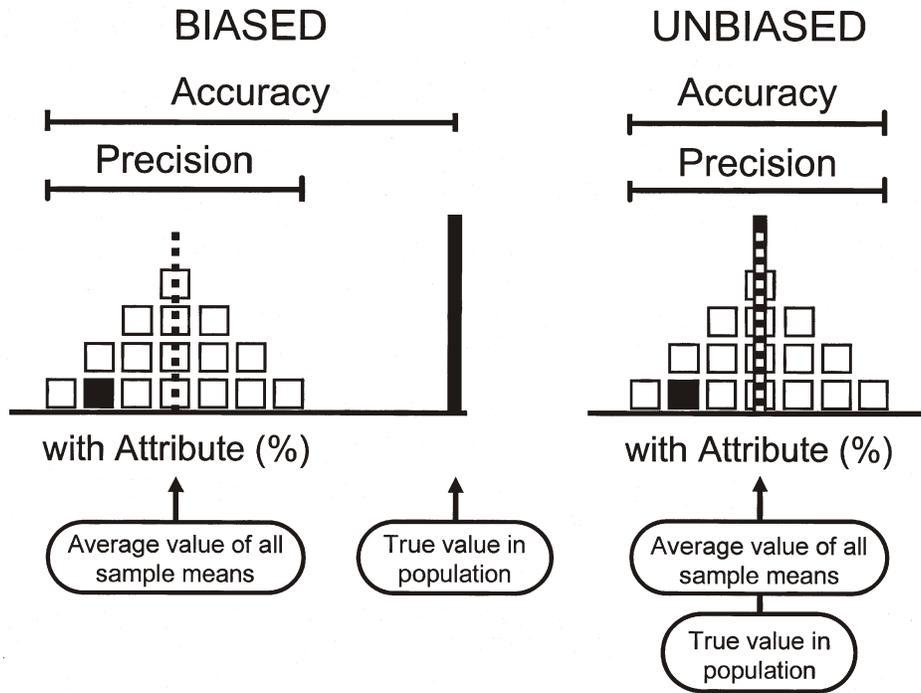


Figure 2-19. Bias, accuracy and precision of samples.

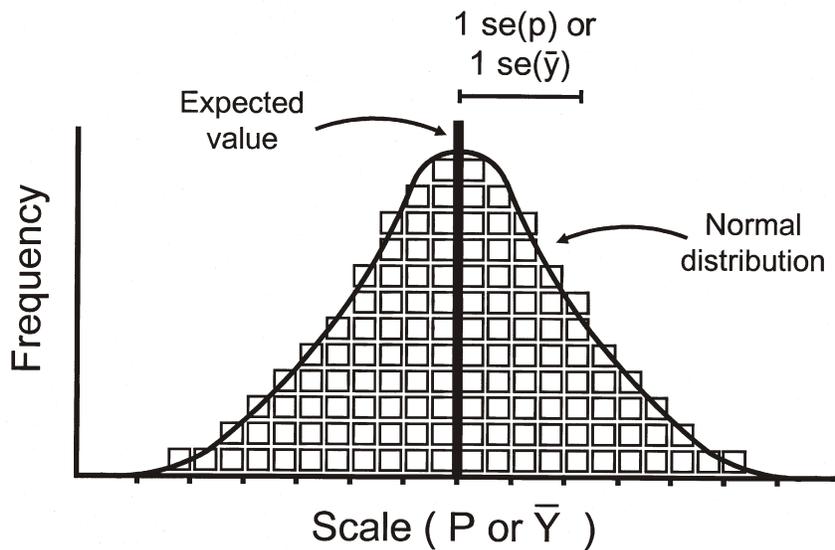


Figure 2-20. Repeated samples and the normal distribution.

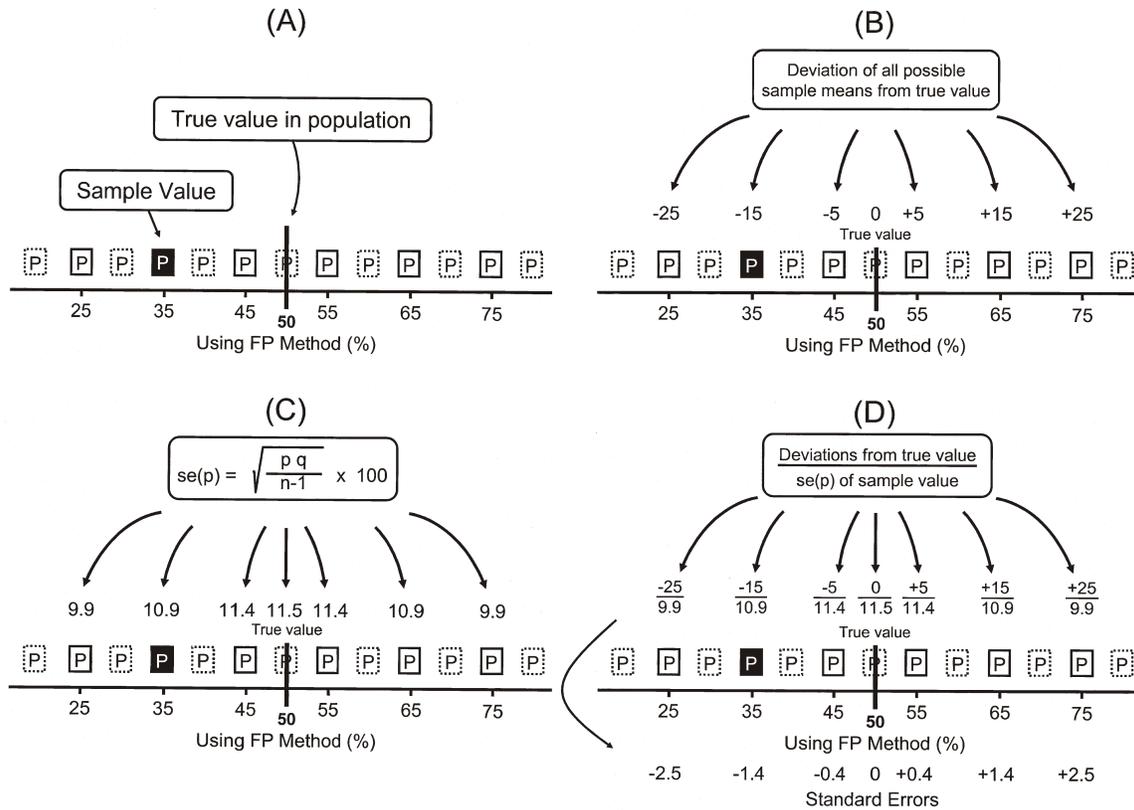


Figure 2-21. Conversion of percentage units to standard error units.

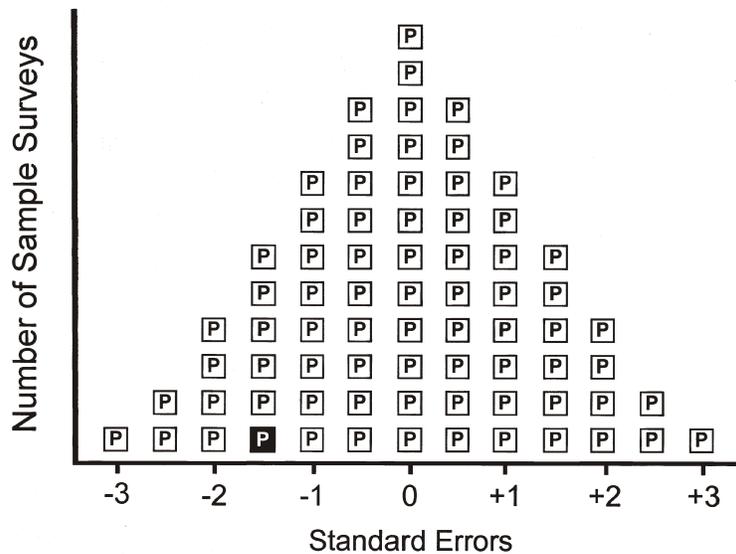


Figure 2-22. Repeated samples of use of family planning methods, with standard error units.

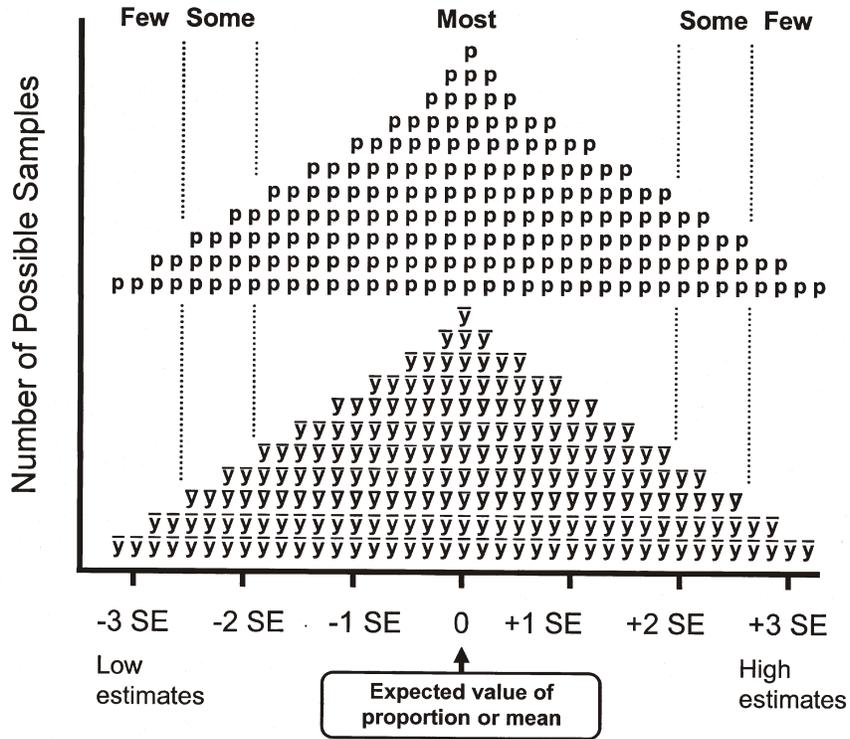


Figure 2-23. Distribution of proportions and means from repeated samples.

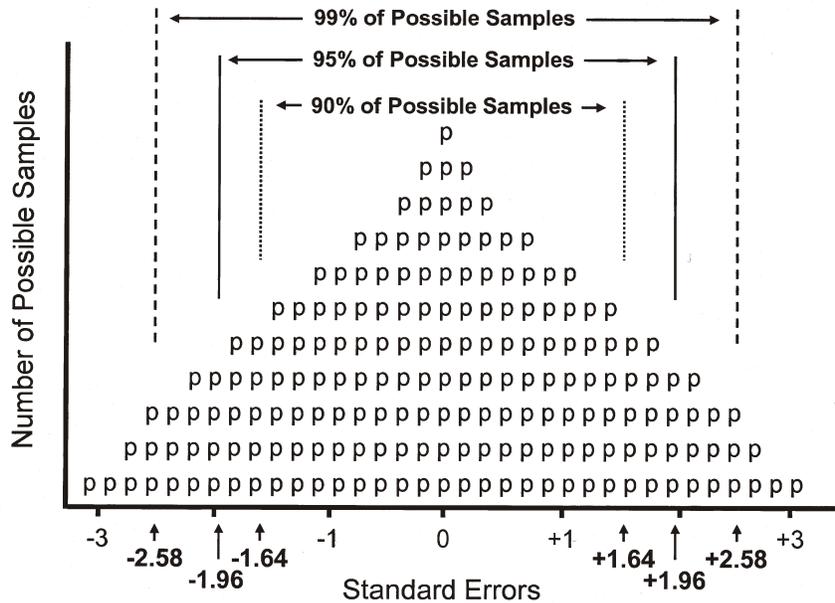


Figure 2-24. Standard errors for 90, 95 and 99 percent of all possible sample.

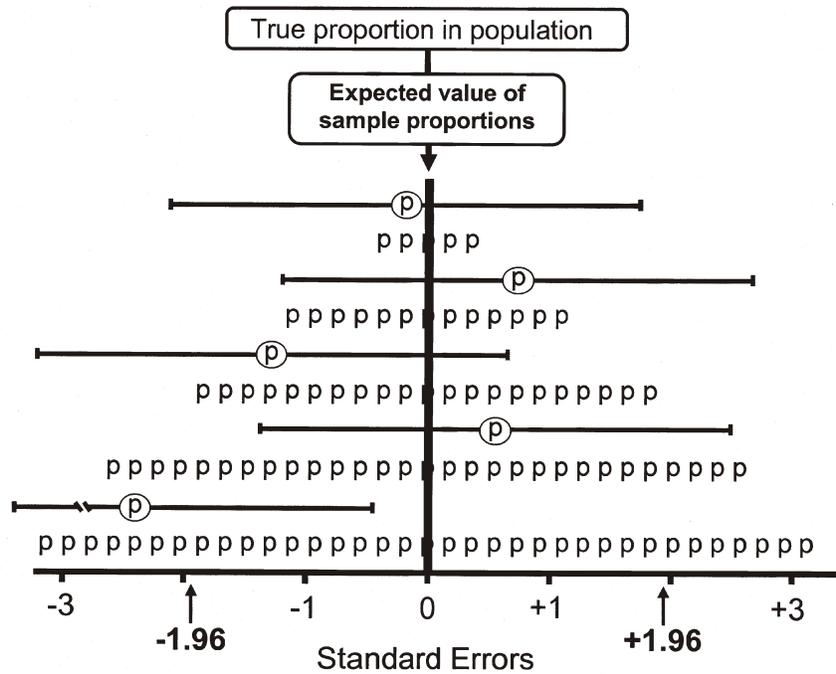


Figure 2-25. Intervals of plus or minus 1.96 standard error units bracketing sample proportions.

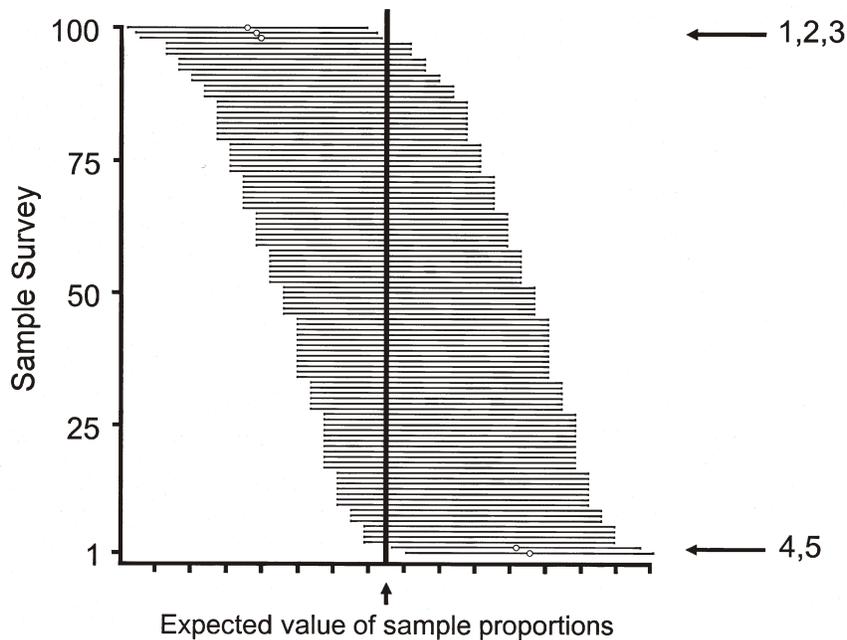


Figure 2-26. One hundred 95 percent confidence intervals for repeated samples, arranged in ascending order.

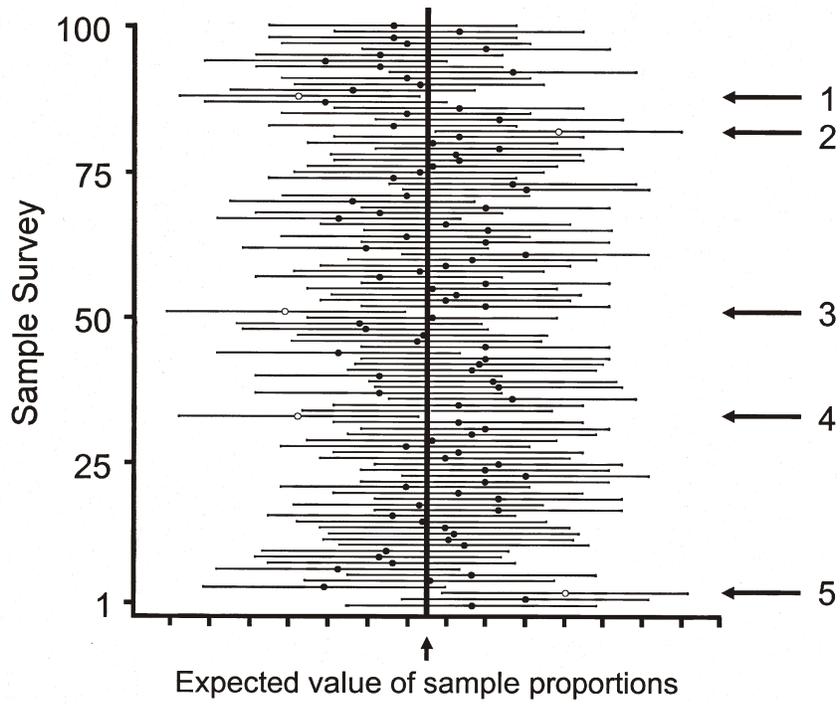


Figure 2-27. One hundred 95 percent confidence intervals for repeated samples, arranged in random order.

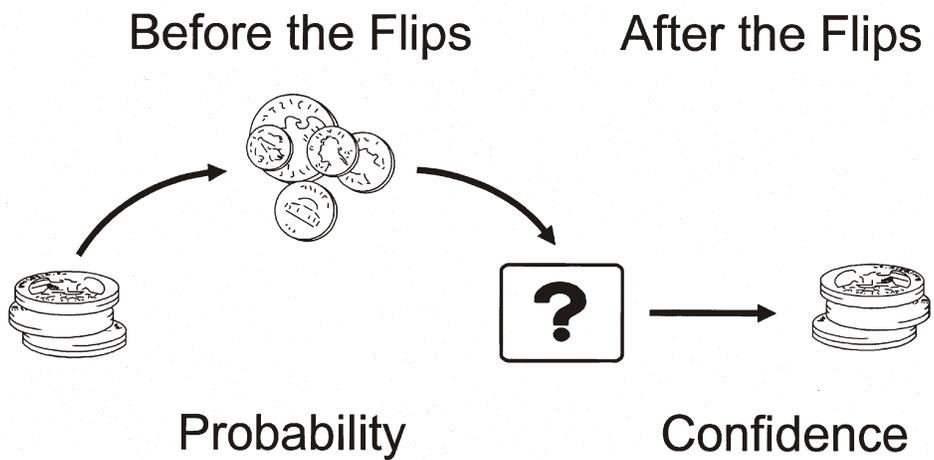


Figure 2-28. Coin flips, probability and confidence.