**Data Collection and Interpretation**

Data interpretation is part of daily life for most people. Interpretation is the process of making sense of numerical data that has been collected, analyzed, and presented. People interpret data when they turn on the television and hear the news anchor reporting on a poll, when they read advertisements claiming that one product is better than another, or when they choose grocery store items that claim they are more effective than other leading brands.

A common method of assessing numerical data is known as **statistical analysis** , and the activity of analyzing and interpreting data in order to make predictions is known as **inferential statistics** . Informed consumers recognize the importance of judging the reasonableness of data interpretations and predictions by considering sources of bias such as sampling procedures or misleading questions, **margins of error** , **confidence intervals** , and incomplete interpretations.

**Why Is Accurate Data Collection Important?**

The repercussions of inaccurate or improperly interpreted data are wide-ranging. For example, every 10 years a major census is done in India. The results are used to help determine where new roads will be built; where new schools and libraries are needed; where new nursing homes, hospitals will be required; where new parks and recreational centers will be built; and the sizes of police and fire departments.

**Gathering Reliable Data**

The process of data interpretation begins by gathering data. Because it is often difficult, or even impossible, to look at all the data (for example, to poll every high school student in the India), data are generally obtained from a smaller unit, a subset of the population known as a **sample**. Then data from the sample are used to predict (or infer) what the characteristics of the population as a whole may be. For example, a telephone survey of one thousand car owners in Punjab might be conducted to predict the popularity of various cars among all car owners. The one thousand car owners who are surveyed are the sample and all car owners in Punjab are the population.

**But there’s both an art and science to collecting high-quality data. Several key elements must be considered: bias, sample size, question design, margin of error, and interpretation.**

**Avoiding Bias.** In order for data interpretation to be reliable, a number of factors must be considered. First and perhaps foremost, an **unbiased sample** must be used. In other words, every person (or item) in the population should have an equal chance of being in the sample.

For example, what if only Ford owners were surveyed in the telephone survey? The survey would be quite likely to show that Fords were more popular. A biased sample is likely to **deviate** the data in favor of one particular individual or item, thus making data interpretation unreliable. If we want to know what sorts of cars are preferred by Punjab car owners, we need to be sure that our sample of car owners is representative of the entire car owner population.

One way of ensuring an unbiased sample is to choose randomly from the population. However, it is often difficult to design a study that will produce a truly unbiased sample. For example, suppose a surveyor decides to choose car owners at **random** to participate in a phone interview about car preferences. This may sound like a good plan, but car owners who do not have telephones or whose telephone numbers are unavailable will not have a chance to participate in the survey. Maybe car owners with unlisted telephone numbers have very different car preferences than the broader population, but we will never know if they are not included in the sample.

**Achieving a Large Enough Sample.** A second important factor in data collection is whether the chosen sample is large enough. For example, are one thousand car owners a sufficient number of car owners from which to infer the opinion of all car owners? In order to answer this question, the margin of error needs to be calculated.

The margin of error is a statistic that represents a range in which the surveyor feels confident that the population as a whole will fall. A sufficient sample size needs to have a small margin of error, usually around 5 percent. To determine the margin of error (*m* ), divide one by the square root of the sample size (*n* ): . Therefore, the sample of one thousand car owners gives us a margin of error of about 3 percent, an allowable margin of error.

Margin of error (m) = 1/√n

**Asking the Proper Questions.** Informed citizens who are assessing survey results must consider the type of questions that are asked when a survey is conducted. Were the questions leading? Were they easy or difficult to understand? For example, suppose a study carried out by a local ice cream manufacturer states that 75 percent of students prefer ice cream. It seems self-evident that an ice cream company would not report a study that showed students do not like ice cream. So perhaps the question in the study was leading: for example, "Do you prefer ice cream or spinach?" It is therefore important to find out exactly what questions were asked and of whom.

**Giving a Proper Interpretation.** Data are often interpreted with a bias, and the results can therefore be misleading or incomplete. For example, a bath soap company claims that its soap is 99 percent pure. This statement is misleading because the soap manufacturer does not explain what "pure" is. When reading an unclarified percentage such as in the previous example, one needs to ask such questions. An example of another incomplete or misleading interpretation is that the average child watches approximately 5 hours of television a day. The reader should question what an "average child" is.

**Considering Margin of Error.** Margin of error is important to consider when statistics are reported. For example, we might read that the high school dropout rate declined from 18 percent to 16 percent with a margin of error of 3 percent. Because the 2-percentage point decline is smaller than the margin of error (3 percent), the new dropout rate may fall between 13 percent to 19 percent. We cannot be entirely sure that the high school dropout rate actually declined at all.

Confidence intervals, a term usually employed by statisticians, and related to margins of error, is reported by a percentage and is constructed to relay how confident one can be that the sample is representative of the population. The producers of this survey may only be 95 percent confident that their sample is representative of the population. If this is the case then there is a 5 percent chance that this sample data does not typify or carry over to the population of the United States. The margin of error represents the range of this 95-percent confidence interval (the range that represents plus or minus two **standard deviations** from the **mean** ).

**Understanding and Interpreting Data**

Figuring out what data means is just as important as collecting it. Even if the data collection process is sound, data can be misinterpreted. When interpreting data, the data user must not only attempt to discern the differences between causality and coincidence, but also must consider all possible factors that may have led to a result.

After considering the design of a survey, consumers should look at the reported data interpretation. Suppose a report states that 52 percent of all Punjab residents prefer Maruti-Suzuki to other car manufacturers. The surveyors want you to think that more than half of all Punjabis prefer maruti-suzuki, but is this really the case? Perhaps not all those surveyed were Punjab residents. Also, the 52 percent comes from the sample, so it is important to ask if the sample was large enough, unbiased, and randomly chosen. One also needs to be aware of margins of error and confidence intervals. If the margin of error for this survey is 5 percent than this means that the percentage of car owners in the Punjab who prefer Chevrolet could actually be between 47 and 57 percent (5 percent higher or lower than the 52 percent).

Similar questions are important to consider when we try to understand **polls**. During the 2000 presidential race, the evening news and newspapers were often filled with poll reports. For example, one poll stated 51 percent of Americans preferred George W. Bush, 46 percent preferred Al Gore, and 3 percent were undecided, with a margin of error of plus or minus 5 percent.

The news anchor then went on to report that *most* Americans prefer George W. Bush. However, given the data outlined above, this conclusion is questionable. Because the difference between George W. Bush and Al Gore is the same as the margin of error, it is impossible to know which candidate was actually preferred. In addition, if we do not know any of the circumstances behind the poll, we should be skeptical about its findings.

**A Data Checklist.** When reading any survey, listening to an advertisement, or hearing about poll results, informed consumers should ask questions about the soundness of the data interpretation. A recap of key points follows.

1. Was the sample unbiased (representative of the whole population)?
2. Was the sample large enough for the purpose of the survey (margin of error of the sample)?
3. What type of questions did the surveyor ask? Were they simple and unambiguous? Were they leading (constructed in such a way to get the desired response)?
4. Can the conclusions drawn be justified based on the information gathered?
5. How was the survey done (mail, phone, interview)? Does the survey report mention margins of error or confidence intervals, and, if so, are these such that the conclusions drawn are warranted?

By using these checkpoints and learning to think critically about data collection and interpretation, individuals can become more savvy consumers of information.