BASIC FACTS ABOUT STATISTICS AND DATA

STATISTICS IS THE SCIENCE OF LEARNING FROM DATA

Data are numbers, but they are not "just numbers." ***Data are numbers with a context.*** *The number 2.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 2.5 kgs at birth, we congratulate her on the healthy size of the child. The context engages our background knowledge and allows us to make judgments. We know that a baby weighing 2.5 kg is quite healthy, and that a human baby is unlikely to weigh 2.5 g or 2.5 Tons. The context makes the number informative. To gain insight from data, we make graphs and do calculations. But graphs and calculations are guided by ways of thinking that amount to educated common sense.*

WHERE THE DATA COME FROM MATTERS

Data may come from observational studies, experiments, questionnaires or Interviews. Experiments are hard to design in the business setting and may raise ethical concerns, and so are rarely used. Observational studies are often useful. We can learn from observational studies how chimpanzees behave in the wild, or which popular songs sold best last week, or what percent of workers were unemployed last month. Can we trust the results? We'll this isn't a simple yes-or-no question- because sampling is often involved.

ALWAYS LOOK AT THE DATA

Yogi Berra said it: "You can observe a lot by just watching." That's a motto for learning from data. A few carefully chosen graphs are often more instructive than great piles of numbers.

BEWARE THE LURKING VARIABLE

Women who chose hormone replacement after menopause were on the average richer and better educated than those who didn't. No wonder they had fewer heart attacks. Children who play cricket or golf tend to have prosperous and well-educated parents. No wonder they do better in school (on the average) than children who don't play these sports. We can't conclude that hormone replacement reduces heart attacks or that playing cricket or golf increases school grades just because we see these relationships in data. In both examples, education and affluence are lurking variables; background factors that help explain the relationships between hormone replacement and good health and between cricket/ golf and good grades.

Almost all relationships between two variables are influenced by other variables lurking in the background. To understand the relationship between two variables, you must often look at other variables. Careful statistical studies try to think of and measure possible lurking variables in order to correct for their influence. As the hormone saga illustrates, this doesn't always work well. News reports often just ignore possible lurking variables that might ruin a good headline like "Playing cricket/ golf can improve your grades." The habit of asking "What might lie behind this relationship?" is part of thinking statistically.

VARIATION IS EVERYWHERE

Too many business people assign equal validity to all numbers printed on paper. They accept numbers as representing Truth and find it difficult to work with the concept of probability. They do not see a number as a kind of shorthand for a range that describes our actual knowledge of the underlying condition.

Business data such as sales and prices vary from month to month for reasons ranging from the weather to a customer's financial difficulties to the inevitable errors in gathering the data. The manager's challenge is to say when there is a real pattern behind the variation. We'll see that statistics provides tools for understanding variation and for seeking patterns behind the screen of variation.

CONCLUSIONS ARE NOT CERTAIN

Cervical cancer is second only to breast cancer as a cause of cancer deaths in women. Almost all cervical cancers are caused by human papillamavirus (HPV). The first vaccine to protect against the most common varieties of HPV became available in 2006.

How well does the vaccine work? Doctors rely on experiments (called "clinical trials" in medicine) that give some women the new vaccine and others a dummy vaccine. (This is ethical when it is not yet known whether or not the vaccine is safe and effective.) The conclusion of the most important trial was that an estimated 98% of women up to age 26 who are vaccinated before they are infected with HPV will avoid cervical cancers over a 3-year period. On the average women who get the vaccine are much less likely to get cervical cancer. But because variation is everywhere, the results are different for different women. Some vaccinated women will get cancer, and many who are not vaccinated will escape. Statistical conclusions are "on the average" statements only.

Well then, can we be certain that the vaccine reduces risk on the average? No. We can be very confident, but we can't be certain.

Because variation is everywhere, conclusions are uncertain. Statistics gives us a language for talking about uncertainty that is used and understood by statistically literate people everywhere. In the case of HPV vaccine, the medical journal used that language to tell us that "Vaccine efficiency . . . was 98% (95 percent confidence interval 86% to 100%)." That "98% effective" is, in Arthur Nielsen's words, "shorthand for a range that describes our actual knowledge of the underlying condition."The range is 86% to 100%, and we are 95 percent confident that the truth lies in that range. We will soon learn to understand this language. We can't escape variation and uncertainty. Learning statistics enables us to live more comfortably with these realities.

CHAPTER 1
PART 1 SUMMARY

A data set contains information on a number of ***individuals.*** *Individuals may be people, animals, or things. For each individual, the data give values for one or more* ***variables.*** *A variable describes some characteristic of an individual, such as a person's height, sex, or salary.*

Some variables are ***categorical*** *and others are* ***quantitative.*** *A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or salary in Kenya Shillings. You can perform arithmetic manipulations (add, subtract, divide, multiply, average) on quantitative variables, but you can't do it on categorical variables.*

Exploratory data analysis (descriptive statistics) *uses graphs and numerical summaries to describe the variables in a data set and the relations among them.* ***Inferential statistics*** *is aimed at arriving at conclusions that are beyond the immediate data alone.*

After you understand the background of your data (individuals, variables, units of measurement), the first thing to do is almost always ***plot your data.***

The ***distribution*** *of a variable describes what values the variable takes and how often it takes these values.* ***Pie charts*** *and* ***bar graphs*** *display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units.* ***Histograms*** *and* ***stem-plots*** *graph the distribution of a quantitative variable. Be sure you know how to plot all these.*

When examining any graph, look for an ***overall pattern*** *and for notable* ***deviations*** *from the pattern.*

Shape, center, and spread *describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as* ***symmetric*** *or* ***skewed.*** *Not all distributions have a simple overall shape, especially when there are few observations. Here you should be able to compute the mean, median, mode, variance and standard deviation, kurtosis and skewness.*

Outliers *are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.*

When observations on a variable are taken over time, make a ***time plot*** *that graphs time horizontally and the values of the variable vertically. A time plot can reveal* ***trends, cycles,*** *or other changes over time. Be sure you are able to make* ***time plots****.*