Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 12: June 22, 2012

Abstract

Review session.

Quantitative methods in business

- Accounting is business's way of handling precise and accurate data. It does not admit mistakes, and uses techniques like double entry to detect and correct mistakes.
- In contrast, statistics is business's way of handling imprecise, inaccurate, and uncertain data. There is always possibility of error and the job of statistical science is to provide consistent ways of tracking, measuring, and adjusting for error.

Dealing with error

- What makes statistics difficult is the fact that there is always error, always uncertainty, and that our measurement errors are affected by that uncertainty.

- However for business purposes, typically we can formulate the costs of risk and uncertainty in terms that reflect the basic values that we are dealing with us rather than the uncertainty itself (*e.g.*, maximizing profit).
  - In economics (in consumer theory and general decision theory), we need to worry about the all aspects of uncertainty, and all the possible values of the distribution of outcomes and so on.
  - But in typical business applications we are mostly interested in the distribution of a financial outcome or similar. Then we

use our best estimate of the measure of error. We don't need to know more about the error variance beyond our best estimate.

The job of statistics

- We do have good ideas of how to calculate best estimates. If we use those we are able to make good assessments of risks that are involved in many or most business situations.
- The job of statistics is not to directly make decisions. Making decisions involve value judgments, and value judgments need to be made by the affected parties.
- In business of course that's management.
- The job of statistics then is to come up with algorithms (known methods of computation) to assess values and the errors and risks associated with them.

Measurement of value

- Human beings have relatively imprecise notions of value, even when units are well defined (such as yen and dollars).
- We deal with big and small losses differently.
- Nevertheless, statistics needs numbers. Always remember that the numbers only partly reflect what people want and think.
- Defining and getting the numbers is not the job of statistics, but of various specialties such as economics, finance, marketing, and so on.

Measurement of uncertainty

- Human beings notion of uncertainty is partly quantitative and qualitative.
  - We can usually say that one event is more or less likely than another.
  - But we usually can't assign additive probabilities to these evaluations.
  - Often we don't have a clear idea of how events relate to each other.
- This means that there is tension the human side of decision-making and quantitative side. The human is better at making judgments of what he really wants but these judgments are not useful for extended calculations.

Making extended calculations

- The purpose of statistics is the foundation of and methods for extended calculations about uncertainty.
- Statistics quantifies uncertainty and for that purpose we use probability.
- By using probability, we can develop standard methods for measuring the uncertainty and reliability of our quantitative statements.

Experiments and observation

- Sometimes we merely study past history, and infer "natural laws" from that. This is called an *observational study.*
- Sometimes we can test our hypothetical laws directly. This is called an *experimental study.*
- Experimental studies are preferred by scientists because we can directly infer cause and effect, because only our proposed cause changes; this implies that if effect changes as well, there's a direct relationship.
- When we have uncontrolled observations, we must have a model of the behavior and of the uncertainty to conduct inference.

Randomized experiments

- Often there are many important factors that we cannot control as precisely as physical scientists do, especially when dealing with human beings.
- Sometimes we can observe those factors, sometimes we cannot.
  - It is important to recognize that we don't always know what the important factors are, and unrecognized factors will often not be observed or recorded at all.
- We can ensure that those uncontrolled factors have a specific distribution by use of *random samples*.

The distribution of factors in random samples

- We usually don't know what that distribution is quantitatively.
- We can say that the observed distribution of the sample *approximates* that of the underlying population. We say that such a sample is *representative*.
- We make the meaning of "approximate" exact by using the **Law of Large Numbers**.
- We can make the probability of a "large" error as small as we like, at the cost of taking a sufficiently large sample.
- If the purpose of our study is to understand the unknown distribution, then random sampling helps a lot.

Distribution

- The fundamental idea of statistics is that of *distribution.* We assume that we do not care about the individuals, but only about certain measurable properties. (*Cf.* accounting, where the individuals do matter.)
- We count the number of individuals with each combination of values of properties.
- We can count either the number (an *absolute distribution*) or the fraction (a *relative distribution*).
- We can count for a particular combination or range of combinations, called *cell* (a *frequency distribution*) or for all cells up to some level (a *cumulative distribution*).
- Sometimes we need to account for the size of cells; in that case we use a *density.*

- Distributions are represented graphically by *histograms*.
-