Business Analytics

Introduction: Exploration of Data

Faton Berisha

Chapter 1

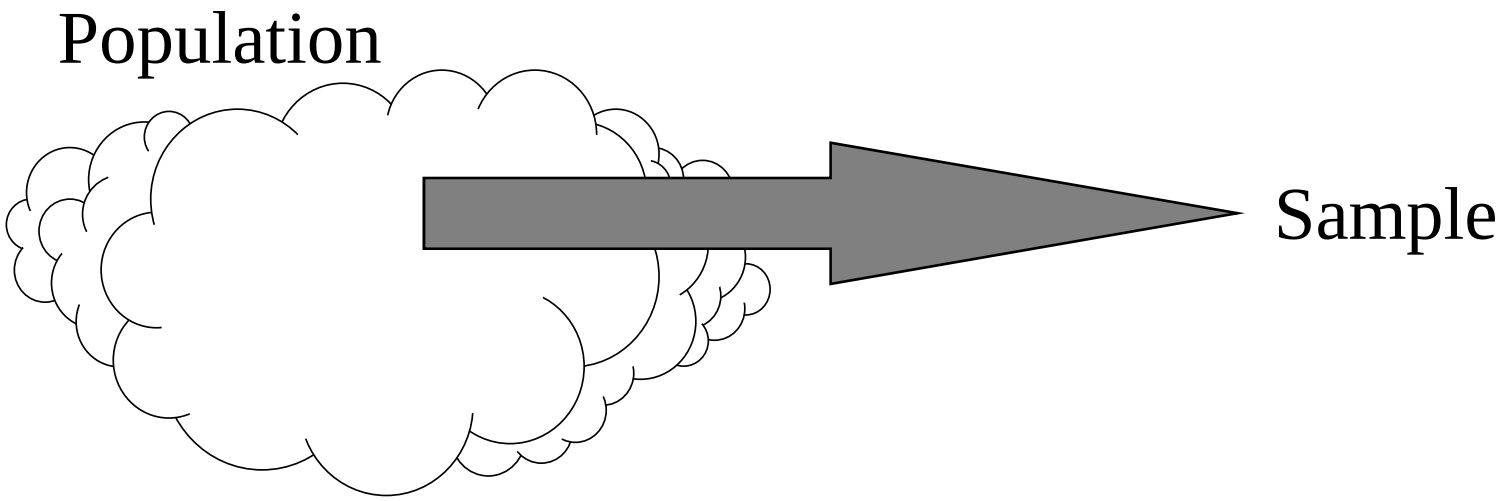
Introduction: Exploration of Data

An Introduction to Business Statistics

Populations and Samples

Population	A set of existing units (people, objects or events)
Variable	A measurable characteristic of the population
Census	An examination of the entire population of measurements
Sample	A selected subset of the units of a population

Sample from Population



Terminology 1

Measurement

The process of determining the extent, quantity, amount, etc, of the variable of interest for some a particular item of the population

- Produces data
- For example, collecting annual starting salaries of graduates from last year's MBA program

Value

The result of measurement

- The specific measurement for a particular unit in the population
- For example, the starting salaries of graduates from last year's MBA Program

Terminology 2

Quantitative

Measurements that represent quantities (for example, “how much” or “how many”)

- Annual starting salary is quantitative
- Age and number of children are also quantitative

Qualitative

A descriptive category to which a population unit belongs: a descriptive attribute of a population unit

- A person’s gender is qualitative
- A person’s hair color is also qualitative

Terminology 3

Population of Measurements

Measurement of the variable of interest for each and every population unit

- Sometimes referred to as an observation
- For example, annual starting salaries of all graduates from last year's MBA program

The process of collecting the population of all measurements is a *census*

- Census usually too expensive, too time consuming, and too much effort for a large population

Terminology 4

Sample

A subset of population units

- For example, a university graduated 8,742 students
- This is too large for a census
- So, we select a sample of these graduates and learn their annual starting salaries

Sample of measurements

- Measured values of the variable of interest for the sample units
- For example, the actual annual starting salaries of the sampled graduates

Terminology 5

Descriptive Statistics

The science of describing the important aspects of a set of measurements

- For example, for a set of annual starting salaries, want to know:
  - How much to expect
  - What is a high versus low salary
  - How much the salaries differ from each other
- If the population is small enough, could take a census and not have to sample and make any statistical inferences
- But if the population is too large, then …

Terminology 6

Statistical Inference

The science of using a sample of measurements to make generalizations about the important aspects of a population of measurements

- For example, use a sample of starting salaries to estimate the important aspects of the population of starting salaries

Sampling a Population of Existing Units

Random Sample

A random sample is a sample selected from a population so that:

- Each population unit has the same chance of being selected as every other unit
  - Each possible sample (of the same size) has the same chance of being selected
- For example, randomly pick two different people from a group of 15:
  - Number the people from 1 to 15 and write their numbers on 15 different slips of paper
  - Thoroughly mix the papers and randomly pick two of them
  - The numbers on the slips identifies the people for the sample

How to Pick?

Sample with Replacement

Replace each sampled unit before picking next unit

- The unit is placed back into the population for possible reselection
- However, the same unit in the sample does not contribute new information

Sample without Replacement

A sampled unit is withheld from possibly being selected again in the same sample

- Guarantees a sample of different units
  - Each sampled unit contributes different information
  - Sampling without replacement is the usual and customary sampling method

Example: The Accounts Receivable Case

Example 1.1: Reducing payment times

- Typical payment times with the former billing system: 39 days or more
- The consulting firm assesses the effectiveness of the newly installed billing:
  - Choosing a sample of size 65
    out of a population of 7,823 payment times.

Approximately Random Samples

In general, must make a list identifying each and every individual population unit (called a *frame*)

If the population is very large, it may not be possible to list every individual population unit

So instead draw a “systematic” sample

- Randomly enter the population and systematically sample every k^{th} unit
- This usually approximates a random sample
  - Read Example 1.2, “Marketing Research Case: Rating a New Bottle Design,” in the textbook

Example: The Marketing Research Case

Example 1.2: Rating a new package design

- E.g., Every 100th shopper (or 200th, 300th,...) is selected
- By using the systematic sampling process, a sample of 60 shoppers is selected
- Each shopper rates the package design by a composite score of 5 to 35
- From experience, the smallest acceptable composite score for a successful package design: 25

Another Sampling Method

Voluntary Response Sample

Participants select themselves to be in the sample

- Participants “self-select”
- For example, calling in to vote on *American Idol*
- Commonly referred to as a “non-scientific” sample

Usually not representative of the population

- Over-represent individuals with strong opinions
  - Usually, but not always, negative opinions

Sampling a Process

Process

A sequence of operations that takes *inputs* (labor, raw materials, methods, machines, etc) and turns them into *outputs* (products, services, and the like)

Process “Population”

The “population” from a process is all output produced in the past, present, and the future

For example, all automobiles of a particular make and model (i.e. Lincoln Towncar)

- Cars will continue to be made over time

Population Size

A population may be “finite” or “infinite”

Finite if it is of fixed and limited size

- Finite if it can be counted
  - Even if very large
  - For example, all the Lincoln Towncars actually made during just this model year is a finite population
    - Because a specific number of cars was made between the start and end of the model year

Infinite if it is unlimited

- Infinite if listing or counting every element is impossible
  - For example, all the Lincoln Towncars that could have possibly been made this model year is an infinite population

Example: The Coffee Temperature Case

Shembull 1.3: Monitoring coffee teperatures

- Measuring the temperature of the coffee at half-hour intervals from 10:00–21:30
- Range: from 152ºF to 170ºF
- Representative sample if the process is in a state of statistical control

Statistical Control

A process is in *statistical control* if it does not exhibit any unusual process variations

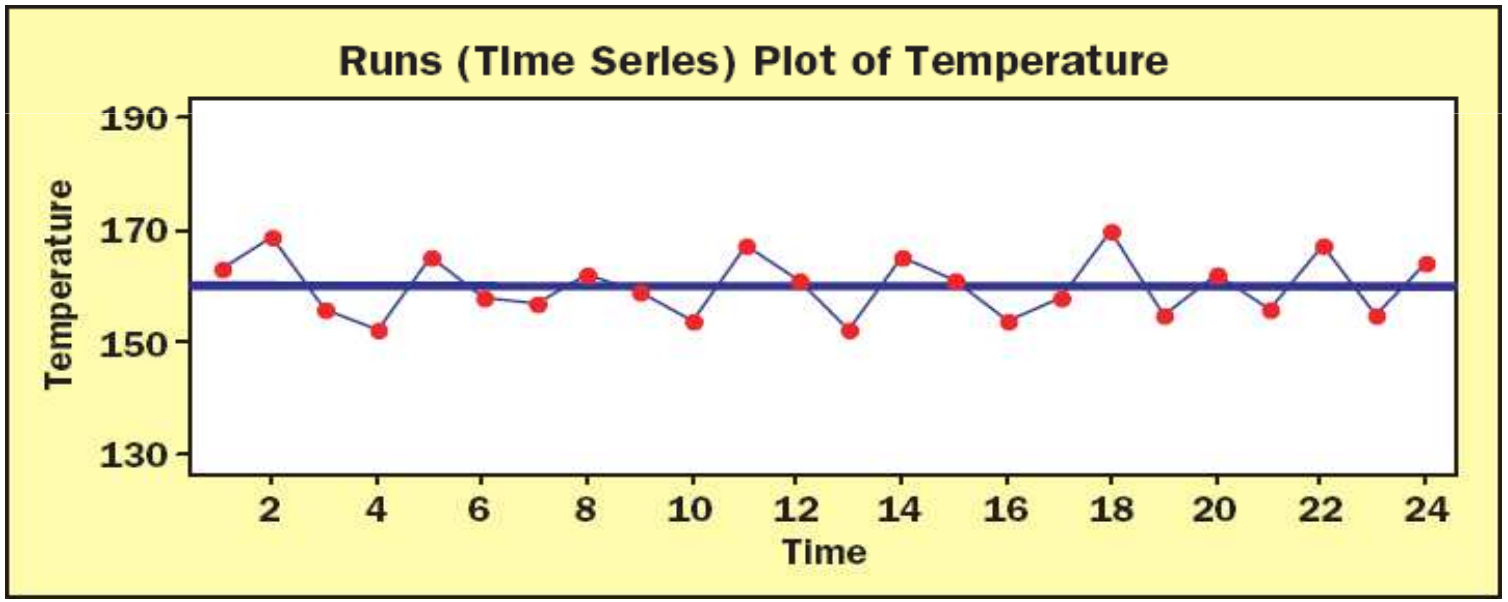
- A process in statistical control displays a constant amount of variation around a constant level
- A process not in statistical control is “out of control”

To determine if a process is in control or not, sample the process often enough to detect unusual variations

- Issue: How often to sample?
- See Example 1.3, “The Coffee Temperature Case: Monitoring Coffee Temperature,” in the textbook

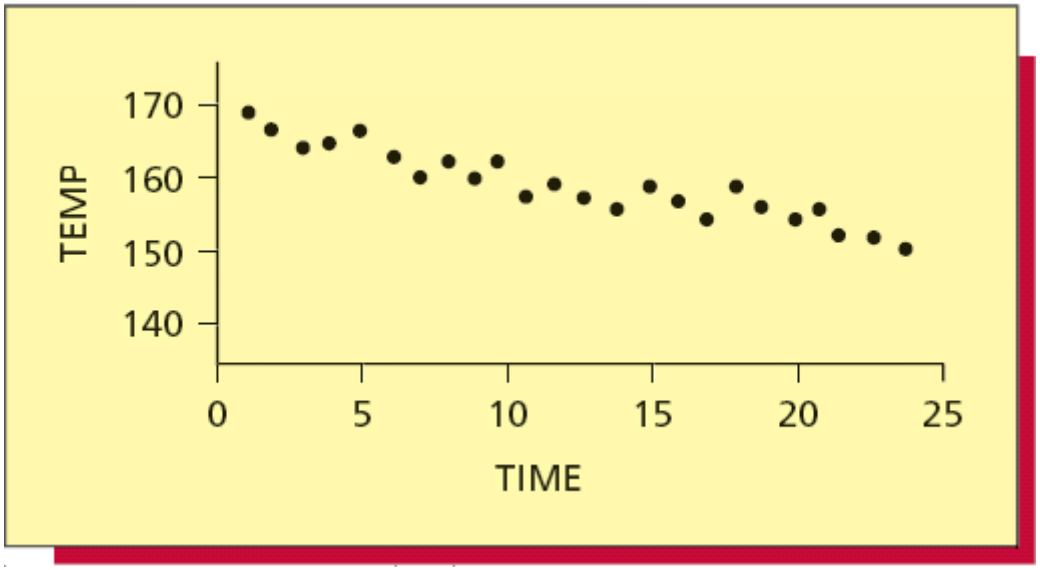
Runs Plot

A *runs plot* is a graph of individual process measurements over time



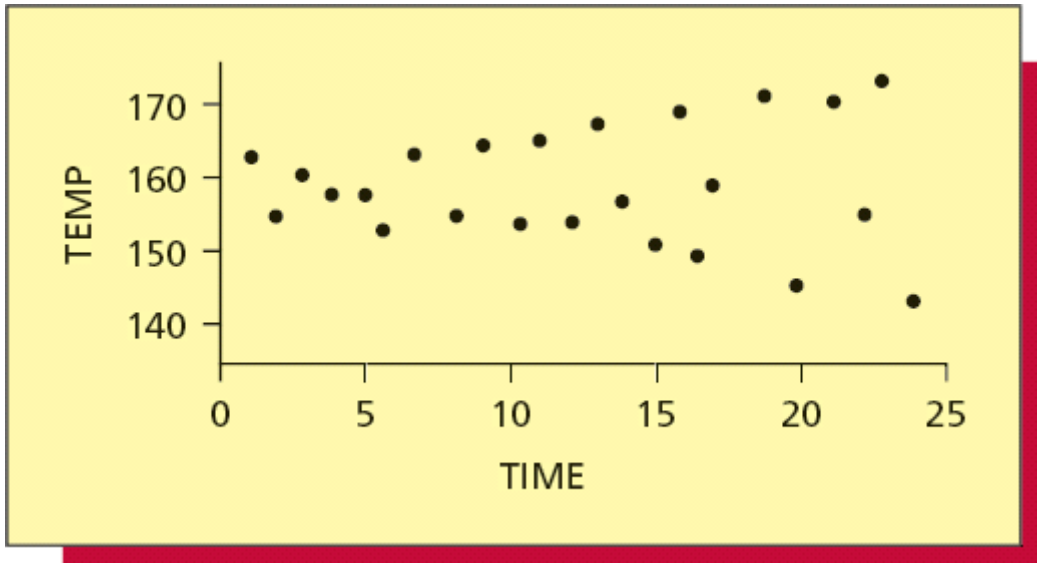
Out of Control

- If, instead of a constant level, there is a trend in the process performance
  - Following the trend, future performance of the process will be outside established limits



Out of Control

- If, there is a constant level, but the amount of the variation is varying as time goes by
  - Data points fan out from or neck down to the constant level



Example: The Car Mileage Case

Example 1.4: Estimating mileage

- The estimated midsize car mileage: 26 mpg
- Government offers the tax credit to automakers who produce a model that achieves at least 31 mpg
- Statistical process for making a statistical inference
  - Describe the practical problem
  - Describe the variable of interest and how it will be measured
  - Describe the sampling procedure:
    - A sample of size 49
  - Describe the statistical inference of interest
  - Describe how the statistical inference will be made and evaluate the reliability of the inference

Statistical Process Control

- The real purpose is to see if the process is out of control so that corrective action can be taken if necessary
- Must investigate further to find out why the process is going out of control
  - See Example 1.4, The Car Mileage Case: Estimating Mileage
- More on statistical process control in Chapter 14

Scales of Measurement

- Qualitative variables
  - Descriptive categorization of population or sample units
  - Two types:
    - Nominative
    - Ordinal
- Quantitative variables
  - Numerical values represent quantities measured with a fixed or standard unit of measure
  - Two types:
    - Interval
    - Ratio

Qualitative Variables

- Nominative:
  - Identifier or name
  - Unranked categorization
    - Example: gender, car color
- Ordinal:
  - All characteristics of nominative plus…
  - Rank-order categories
  - Ranks are relative to each other
    - Example: Low (1), moderate (2) or high (3) risk

Interval Variable

- All of the characteristics of ordinal plus…
- Measurements are on a numerical scale with an arbitrary zero point
  - The “zero” is assigned: it is nonphysical and not meaningful
  - Zero does not mean the absence of the quantity that we are trying to measure
- Can only meaningfully compare values in terms of the interval between them
  - Cannot compare values by taking their ratios
  - “Interval” is the arithmetic difference between the values
- Example: temperature
  - 0° F means “cold,” not “no heat”
  - 80° F is NOT twice as warm as 40° F
    - But 80° F is 40° warmer than 40° F

Ratio Variable

- All the characteristics of interval plus…
- Measurements are on a numerical scale with a meaningful zero point
  - Zero means “none” or “nothing”
- Values can be compared in terms of their interval and ratio
  - $30 is $20 more than $10
  - $30 is 3 times as much as $10
  - $0 means no money
- In business and finance, most quantitative variables are ratio variables, such as anything to do with money
  - Examples: Earnings, profit, loss, age, distance, height, weight

Survey Sampling

- Already know some sampling methods
  - Also called sampling designs, they are:
  - Random sampling
    - The focus of this book
  - Systematic sampling
  - Voluntary response sampling
- But there are other sample designs:
  - Stratified random sampling
  - Cluster sampling

Stratified Random Sample

- Divide the population into non-overlapping groups, called *strata*, of similar units
- Separately, select a random sample from each and every stratum
- Combine the random samples from each stratum to make the full sample
- Appropriate when the population consists of two or more different groups so that:
  - The groups differ from each other with respect to the variable of interest
  - Units within a group are similar to each other
    - For example, divide population into strata by age, gender, income, etc

Cluster Sampling

- "Cluster" or group a population into subpopulations
  - Cluster by geography, time, etc…
- Each cluster is a representative small-scale version of the population (i.e. heterogeneous group)
- A simple random sample is chosen from each cluster
- Combine the random samples from each cluster to make the full sample
- Appropriate for populations spread over a large geographic area so that…
  - There are different sections or regions in the area with respect to the variable of interest
  - A random sample of the cluster

Sampling Problems

- Random sampling should eliminate bias
- But even a random sample may not be representative because of:
  - Under-coverage
    - Too few sampled units or some of the population was excluded
  - Non-response
    - When a sampled unit cannot be contacted or refuses to participate
  - Response bias
    - Responses of selected units are not truthful

Further Discussion of Data Acquisition and Survey Sampling

- There are a number of potential sources of data
- Some of sources for existing data include…
  - Internet search
  - Company records
  - Data collection agency
- When the data is not available from these sources, you must collect it yourself

Survey Concepts

Response variable	The variable of interest
Independent variables	Variables that may be related to the variable of interest that will also be measured.

Types of Studies

- When we are able to change or manipulate variables, we have an *experimental* study
- When we are unable to change or manipulate variables, we have an *observational* study

Types of Survey Questions

Dichotomous	Questions with two answers, usually yes/no
Multiple choice	Questions where the respondent must select from a list of possible answers
Open-ended	Questions where the respondent is allowed to write their own answer

Some Different Survey Types

- In a phone survey, questions are asked during a phone conservation
  - Inexpensive
  - Fast
  - Low response rate
- Self-administered surveys are mailed to the respondent to be completed on-their-own
  - Inexpensive
  - Low response rate
  - Takes much longer than a phone survey
- Personal interviews involve face-to-face contact
  - Higher response rate
  - Less chance of misunderstanding
  - Higher cost