

Chapter 1 – What is Statistics

Chapter 1 – What is Statistics

- The term *statistics* can refer to numerical facts such as averages, medians, percents, and index numbers that help us understand a variety of business and economic situations.
- *Statistics* can also refer to the art and science of collecting, analyzing, presenting, and interpreting data.

"There are three kinds of lies: lies, damned lies, and statistics." ~Benjamin Disraeli (debatable)

"Without data, you're just another person with an opinion." ~W. Edwards Deming

1

1

Descriptive & Inferential Statistics

Descriptive Statistics

- Organize
- Summarize
- Simplify
- Presentation of data



Describing data

Inferential Statistics

- Generalize from samples to population
- Hypothesis testing
- Relationships among variables



Make predictions

2

2

Key Statistical Concepts

Population – the group of all items of interest to a statistics practitioner.

- Frequently very large; sometimes infinite.
- A *parameter* is a descriptive measure of a population.

Sample – a set of data drawn from the population.

- Can also be very large, but always much less than the population.
- A *statistic* is a descriptive measure of a sample.

Statistical inference once again – the process of making an estimate, prediction, or decision about a large population based on sample data.

- Because conclusions and estimates are not always going to be correct, we build into the statistical inference "measures of reliability", namely *confidence level* and *significance level*.

Keller, Gerald, Statistics for Management and Economics, 12th Edition. © 2023 Cengage. All Rights Reserved

3

3

Confidence and Significance Levels

The **confidence level** is the proportion of times that the estimate of a population characteristic is expected to be correct.

Example: a confidence level of 95% means that the estimate of the number of soft drinks consumed by the students at a university may be correct 95% of the time.

The **significance level** measures how frequently the conclusion about a population will be wrong in the long run.

Example: a 5% significance level means that, in the long run, the conclusion on the winner of an election will be wrong 5% of the time.

4

4

Chapter 1 – What is Statistics

Case 1- Background and Data Pepsi's Exclusivity Agreement with a University

Case Background

A large university has offered Pepsi-Cola an agreement that would give Pepsi exclusive rights to sell its products at all university facilities for the next year with an option for future years.

In return, the university would receive 35% of the on-campus revenues and an additional lump sum of \$200,000 per year.

Pepsi's Data

The market for soft drinks is measured in terms of 12-ounce cans.

Pepsi currently sells an average of 22,000 cans per week over the 40 weeks of yearly university operations.

The cans sell for an average of \$1 each. The costs, including labor, amount to 30 cents per can. Pepsi suspects its market share to be considerably less than 50%.

5

5

Case 1- Survey Approach Pepsi's Exclusivity Agreement with a University

Survey

Pepsi assigned a recent university graduate to survey the university's students to supply the missing information.

The survey consists of tracking the number of soft drinks a random sample of 500 students purchases over 7 days.

A sample is representative of a Population through random sampling, i.e., each potential sample has an equal probability of being selected.

Inferential statistics:

The population in question is the soft drink consumption of the university's 50,000 students.

The cost of interviewing each student would be prohibitive and extremely time consuming.

Instead, we can use inferential statistics to sample a random sample of 500 students and infer from the data the number of soft drinks consumed by all 50,000 students.

We can then estimate annual profits for Pepsi.

6

6

Case 1- Analysis and Potential Gain Pepsi's Exclusivity Agreement with a University

Analysis

Assuming a 25% market share, and with an exclusivity agreement in place:

$88,000 \text{ cans/wk} \times 40 \text{ wks} = 3,520,000 \text{ cans/yr}$

After deducting the 35% university take:

Gross revenue = $65\% \times \$1/\text{can} \times 3,520,000 \text{ cans/yr} = \$2,288,000$

After subtracting \$0.30 cost/can (\$1,056,000 total) and \$200,000 payment to university:

Net profit = $\$2,288,000 - \$1,056,000 - \$200,000 = \$1,032,000$

Potential Gain:

Pepsi's current annual profit:

$40 \text{ wks} \times 22,000 \text{ cans/wk} \times \$0.70 = \$616,000$

Pepsi's potential gain at a 25% market share:

$\$1,032,000 - \$616,000 = \$416,000$

Issue:

Pepsi does not know how many soft drinks are actually sold at the university.

7

7

Example 2 - Political Elections and Exit Polls

Political Elections

When an election for political office takes place, the television networks cancel regular programming and provide election coverage instead.

When the ballots are counted the results are reported.

However, for important offices such as president or senator in large states, the networks actively compete to see which will be the first to predict a winner.

Exit Polls

Data for election predictions are collected through **exit polls**, wherein a random sample of voters who exit the polling booth is asked for whom they voted.

From the data, the sample proportion of voters supporting the candidates is computed.

A statistical technique is applied to determine whether there is enough evidence to infer that a given candidate will win the election.

8

8

Chapter 1 – What is Statistics

Example 2 – Data and Inference

Bush vs. Gore, Year 2000 Presidential Election

The exit poll results from the state of Florida during the year 2000 elections were recorded (only the votes of the Republican candidate George W. Bush and the Democrat Albert Gore).

Suppose that the results of 765 people who voted for either Bush or Gore were stored: 1 = Gore, and 2 = Bush.

The network analysts would like to know whether they can conclude that George W. Bush will win the state of Florida.

Sample vs. Population

This is a very common application of statistical inference.

The population:

The television networks wanted to make inferences about the approximately 5 million Floridians who voted for Bush or Gore for president.

The sample:

765 people randomly selected by the polling company who voted for either of the two main candidates.

9

9

Example 2 – Estimate of a Population Characteristic

Population proportion

Population characteristic:

The proportion of the total electorate that voted for Bush.

Specific question:

Did more than 50% of the electorate vote for Bush (counting only those who voted for either the Republican or Democratic candidate)?

Limitations of an estimate

Because we cannot ask every one of the 5 million voters for whom they voted, we cannot predict the outcome with 100% certainty.

A sample that is only a small fraction of the size of the population can lead to correct inferences only a certain percentage of the time.

You will find that statistics practitioners can control that fraction and usually set it between 90% and 99%.

10

10

Applications in Business and Economics

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients.

Economics

Economists use statistical information in making forecasts about the future of the economy or some aspect of it.

Finance

Financial advisors use price-earnings ratios and dividend yields to guide their investment advice.

What about Marketing and Production?

11

11

Analytics

Scientific process of transforming data into insight for making better decisions.

Types

- Descriptive analysis – Analytical techniques that describe what happened in the past.
- Predictive analysis
 - Analytical techniques that use models constructed from past data to predict future.
 - Helps assess the impact the impact of one variable on another
- Prescriptive analysis – Analytical techniques that yield a best course of action to take.

12

12

Chapter 1 – What is Statistics

Data Warehousing

- Organizations obtain large amounts of data on a daily basis by means of magnetic card readers, bar code scanners, point of sale terminals, and touch screen monitors.
- Wal-Mart captures data on 20-30 million transactions per day.
- Visa processes 6,800 payment transactions per second.
- Capturing, storing, and maintaining the data, referred to as data warehousing, is a significant undertaking.

13

Data Mining

- Analysis of the data in the warehouse might aid in decisions that will lead to new strategies and higher profits for the organization.
- Using a combination of procedures from statistics, mathematics, and computer science, analysts “mine the data” to convert it into useful information.
- The most effective data mining systems use automated procedures to discover relationships in the data and predict future outcomes, ... prompted by only general, even vague, queries by the user.
- The major applications of data mining have been made by companies with a strong consumer focus such as retail, financial, and communication firms.

14

13

14

Large Real Data Sets

Throughout the course, subsets from the two following data sources are used:

General Social Survey (GSS)

With the exception of the U.S. Census, the GSS is the most frequently used source of information about American society. The surveys are conducted every second year and feature hundreds of variables and thousands of observations about American attitudes on a wide variety of topics.

Survey of Consumer Finances (SCF)

The SCF is conducted every 3 years to provide detailed information on the finances of U.S. households. The study is sponsored by the Federal Reserve Board in cooperation with the Department of the Treasury.

For this course, we will use EXCEL and STATA software to perform statistical computations on the data in the belief that virtually all university graduates use it now and will in the future.

15

15

Ethical Guidelines for Statistical Practice

- In a statistical study, unethical behavior can take a variety of forms including:
 - Improper sampling
 - Inappropriate analysis of the data
 - Development of misleading graphs
 - Use of inappropriate summary statistics
 - Biased interpretation of the statistical results
- You should strive to be fair, thorough, objective, and neutral as you collect, analyze, and present data.
- As a consumer of statistics, you should also be aware of the possibility of unethical behavior by others.

16

16

Chapter 1 – What is Statistics

Correlation vs. Causation

A study showed that truck drivers weigh more than adults who do not drive trucks. Conclusion: Trucks cause people to gain weight.*

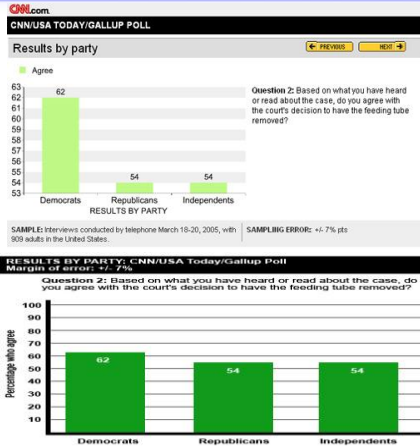
There is a strong relationship between a child's knowledge of nursery rhymes at age 3 and their reading skills over the next few years. Does this mean that nursery rhymes cause children to be able to read better?**

<https://www.proprofs.com/quiz-school/story.php?title=mt40tgwoai1ow>

*Taken from Triola, Essentials of Statistics. Addison Wesley.

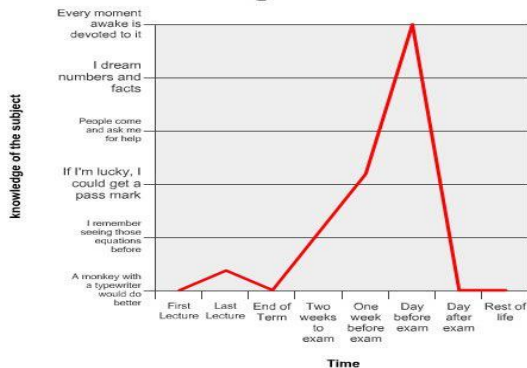
** <http://www.ncbi.nlm.nih.gov/pubmed/2760133>

17



18

Knowledge vs Time



19

Data Acquisition Considerations

Time Requirement

- Searching for information can be time consuming.
- Information may no longer be useful by the time it is available.

Cost of Acquisition

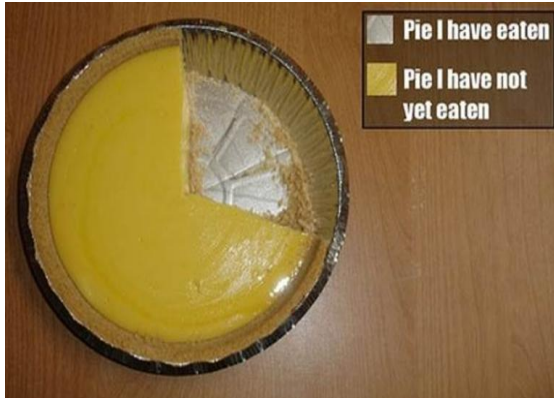
- Organizations often charge for information even when it is not their primary business activity.

Data Errors

- Using any data that happen to be available or were acquired with little care can lead to misleading information.

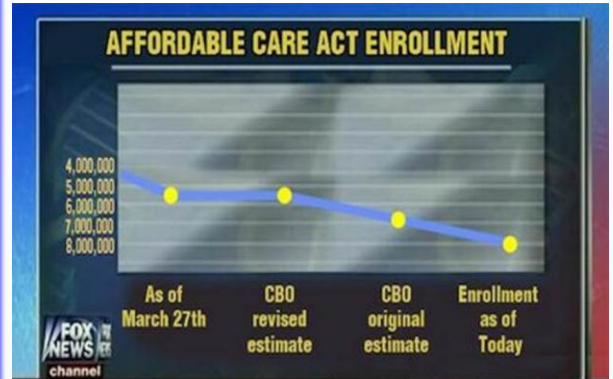
20

Chapter 1 – What is Statistics



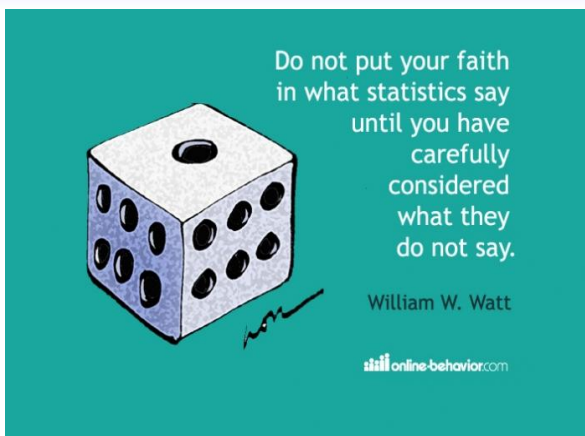
21

21



22

22



23

23



https://tstat.kuleuven.be/images/cartoon-statistics.jpg/image_view_fullscreen

24

24