Statistics

Week 1: Introduction (Chapter 1)

ESD, SUTD

Term 5, 2017

Outline

Introduction

Lecturer: James Wan, james_wan@sutd.edu.sg

TA: Qifang Bao, qfbao@mit.edu, 1.715-S10

Office hour (with either lecturer or TA): by appointment

Read the *Course Description* and *Project Options* on eDimension.

Course structure

Lectures: 2×2 hours weekly, focus on theory, contain some short activities. **Start on time**, with 10 minute break in the middle.

Recitations: 1 hour weekly, focus on applications, contain longer activities.

In weeks 2, 4, 6, 8, 10, 12, there will be an **additional recitation** for the project (Thursdays, 2–3pm, TT21).

Homework sets: contain problems that require software.

We will mainly use *Excel*, and occasionally *R*.

Exams: hand computations and proofs (no software allowed).

Makeup class: Wednesday 1 February, 11am–1pm, TT21.

Course notes

These lecture slides will provide a sufficient outline of the course material. They are *not* meant to be complete or comprehensive, nor will they be uploaded far ahead of time.

To *improve* your study skills, you will need to take your own notes to complement the slides, based on what is covered in class.

For a fuller understanding of the course, you should refer to the text book, *Statistics and Data Analysis: From Elementary to Intermediate*, by Tamhane and Dunlop.

Feel free to use online resources, but stay away from ones that *spoon-feed* you. We are aiming for conceptual understanding, not memorization.

Outline

Data

In everyday life we both generate and receive data, for instance when we access websites, use social media, fill out forms, . . .

We have the need to analyze and interpret data: in daily life, we need to understand what the media gives us; managers and governments need data to *identify patterns* and *make decisions*.

Some people, such as engineers, analysts, scientists, design *models* using data. Models aid in decision-making by trying to *predict* future events, and finding *relationships* that may not be readily apparent.

In statistics, we typically collect data $\rightarrow$ organize and summarize data $\rightarrow$ analyze and interpret data.

Too much data...

- The number of photos taken is growing exponentially. There are around 300 million photo uploads on Facebook per day.
- Google receives around 4 million search queries per minute.
- Twitter users tweet about 450 000 times every minute.
- Email users send over 200 million messages every minute.
- About 300 hours of video are uploaded to YouTube every minute (compare with 100 hours in 2014).

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom. – Clifford Stoll

Data sources

- Department of Statistics – http://www.singstat.gov.sg/
- Singapore Government – http://data.gov.sg/
- World Bank – http://data.worldbank.org/
- Gapminder – http://www.gapminder.org/data/
- World Factbook – https://www.cia.gov/library/publications/the-world-factbook/
- Interactive visualizations – http://www.informationisbeautiful.net/
- Correlation vs causation: http://www.tylervigen.com/

Outline

Relationship between probability and statistics

Probability: You know how the population behaves; from that information, you use probability to infer how a sample behaves.

Statistics: You know how a sample behaves; from that information, you use statistics to infer how the population behaves.

Learning statistics

Consistently, many university graduates find statistics to be among the most useful subjects offered.

To be successful in statistics, you need a good understanding of *probability*, and also *common sense*.

Therefore you need to retain everything covered in the Probability course, as well as some calculus and linear algebra.

In particular,

$\mathrm{E}(c_1\,X_1 + c_2\,X_2) = c_1\,\mathrm{E}(X_1) + c_2\,\mathrm{E}(X_2)$, for any constants c_1, c_2 and *any random variables* X_1, X_2.

$\mathrm{Var}(c_1\,X_1 + c_2\,X_2) = c_1^2\,\mathrm{Var}(X_1) + c_2^2\,\mathrm{Var}(X_2)$, for any constants c_1, c_2 and *uncorrelated random variables* X_1, X_2.

Also, binomial distribution, normal distribution, CLT, ...

Examples

Example 1: Among 1 million items, 100 of them are defective. In a random sample of 10000, how many do we expect to be defective?

In a sample of 10000 items, we observe 5 defective ones. Estimate the number of defective items in the population of 1 million.

Example 2: Suppose adult female height in Singapore is normally distributed with mean 160cm and standard deviation 5.0cm. Find the probability that 10 randomly selected females are all between 150cm and 170cm tall.

Given the heights of 10 randomly selected females, estimate the average height of the female population.

Exercise – intro to hypothesis testing

A man claims to have extra-sensory perception. He demonstrates this by correctly predicting the outcomes of 4 out of 5 coin tosses.

Your task: find the probability that one can do *at least as well as* him by random guessing, and decide if his claim is likely to be true.

Outline

Survey questions

A statistical study may be *observational* or *experimental*.

- Observations studies often require a sample *survey*.

  Unless the sample equals the population (in which case it is called a *census*), there will be errors in any conclusion we draw about the population (sampling errors).

  Errors can also come from the survey design (such as *selection bias*).

- In experiments, errors can come from faulty equipment, measurement, design, etc.

Poorly designed surveys

What is wrong with . . .

- Conducting a survey on how many telephones the average household has, by dialing people selected randomly from a phone book?

  Truman was famously predicted to lose the 1948 US presidential election. The mistake came from reliance on the results of a phone survey.

- Conducting a survey on the average family size, by asking people how many children their parents had?

- Conducting a survey about what jobs people have, by choosing a random suburb, and knocking on doors whose street number is divisible by 5, between 11am and 4pm?

Poorly designed surveys

- What is wrong with the following survey question?

Which Chinese dialect do you speak at home?

- Mandarin
- Hokkien
- Teochew
- Cantonese

- What is wrong with asking the respondents to rate a new type of coffee on a scale from 0 to 100?

Surveys with sensitive questions

The *wording* of a survey may prompt respondents to answer in a particular way. For example, an early version of the Scottish independence referendum read,

'Do you agree that Scotland should be an independent country?'

In some surveys, the questions asked are sensitive no matter how you phrased them:

- Have you taken illicit drugs?
- Are you a racist?

People often do not respond or give false answers. Techniques such as *unmatched count* and *randomized response* can be used to reduce this.

Unmatched count

In the 1991 survey, white Americans were questioned to measure racial hatred against black Americans. Respondents were randomly divided into two groups, and asked the following questions:

Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

(1) "the federal government increasing the tax on gasoline;"
(2) "professional athletes getting million-dollar-plus salaries;"
(3) "large corporations polluting the environment."

Now I'm going to read you four things that sometimes make people angry or upset. After I read all four, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

(1) "the federal government increasing the tax on gasoline;"
(2) "professional athletes getting million-dollar-plus salaries;"
(3) "large corporations polluting the environment;"
(4) "a black family moving next door to you."

Suppose there are n people in each group, and the total number of things that upset people are respectively u_1 and u_2. Then the proportion of people with racial hatred can be estimated by $\frac{u_2 - u_1}{n}$.

Randomized response

To estimate the proportion of the people who have consumed marijuana, a survey contains two questions:

Q1: "I have consumed marijuana." Answer YES or NO.

Q2: "I have never consumed marijuana." Answer YES or NO.

The interviewee is asked to secretly throw a dice, and answer Q1 if they throw a 6, otherwise answer Q2.

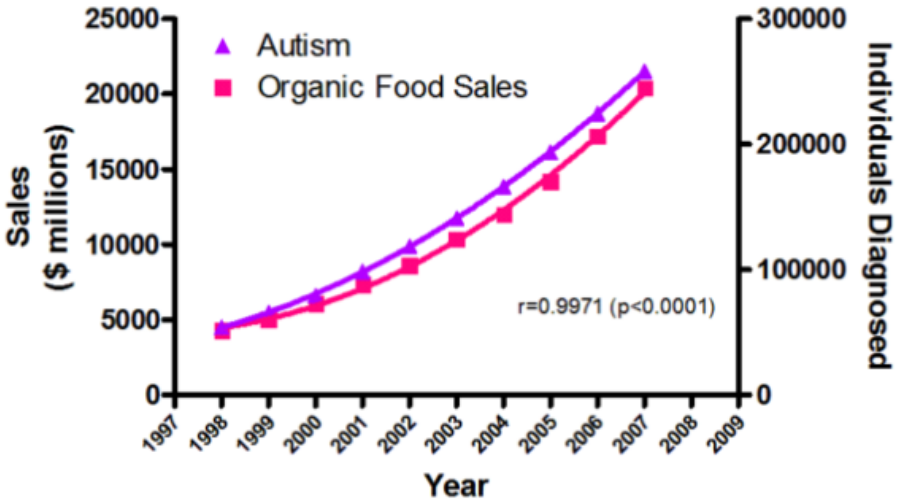
The interviewer does *not* know which question is answered.

Exercise:

If 70% of the responses are YES, estimate the proportion of the population that has consumed marijuana.

Outline

Correlation does not mean causation!



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

A strong relationship (*correlation*) between two sets of data does not imply *anything* about causation, even if the study is error-free.

Both sets of data may be influenced by another variable (a *confounding* variable), or the correlation could occur by chance.

Research (of non-statistical nature) needs to be carried out to prove causation. E. g. from the strong correlation between smoking and cancer *alone*, we cannot conclude that 'smoking causes cancer' or 'cancer causes smoking'.

Excel exercise

- Open the Excel file on stork population vs human birth rate.
- Construct a *scatter plot* and comment on the correlation. Offer an explanation.
- Do you notice something suspicious about the data?
- Go to File $\rightarrow$ Options $\rightarrow$ Add-Ins $\rightarrow$ Go, and select the Analysis ToolPak.