SM-1402/CU-0304 Basic Statistics

Introduction

Dr. Haziq Jamil

Semester 1, 2020/21

Admin

- Lecturer information

```
Dr. Haziq Jamil
Assistant Professor in Statistics
Room M1.09
haziq.jamil@ubd.edu.bn
```

- Classes are on

  - **Saturdays, 9.50am—11.40am** online

- Students have attended Basic Calculus with Dr. Norhayati Hamzah in weeks 1-7. Now proceeding to learn Basic Statistics.

- Slides and materials are available from Canvas.

Admin

- The course is assessed as follows:

  - Formative assessment

    - Exercise sheets

    - Tutorials

- Summative assessment

  - 1 × assignment, due at the end of Semester.

  - Exam.

Course contents

- **Chapter 1: Representation and Summary of Data**
  - Discrete & continuous data
  - Frequency distributions & grouping
  - Plots (histograms, frequency polygons, pie charts)
  - Data summaries (mean, median, variance, standard deviation, quartiles)
  - Cumulative frequencies
- **Chapter 2: Probability, Arrangements, Permutations, Combinations**
  - Probability notation and laws
  - Venn diagrams and probabilities
  - Conditional probabilities and independent events
  - Probability trees
  - Bayes theorem
  - Arrangements, permutations and combinations

Course contents

- **Chapter 3: Discrete Probability Distribution**
  - Probability distribution tables / functions and its features
  - Expectation and variance
  - Cumulative distribution
  - Useful results of expectation and variance
  - Bernoulli and Binomial distribution
  - Poisson distribution

Why Statistics?

Statistics is the science of learning from data.

It is the crucial process which allows us to

- make discoveries in scientific studies (*inference*)

- make *predictions* about future observations

- make informed decisions based on data

Learning Statistics allows you to make *trustworthy analyses and predictions* about the real world.
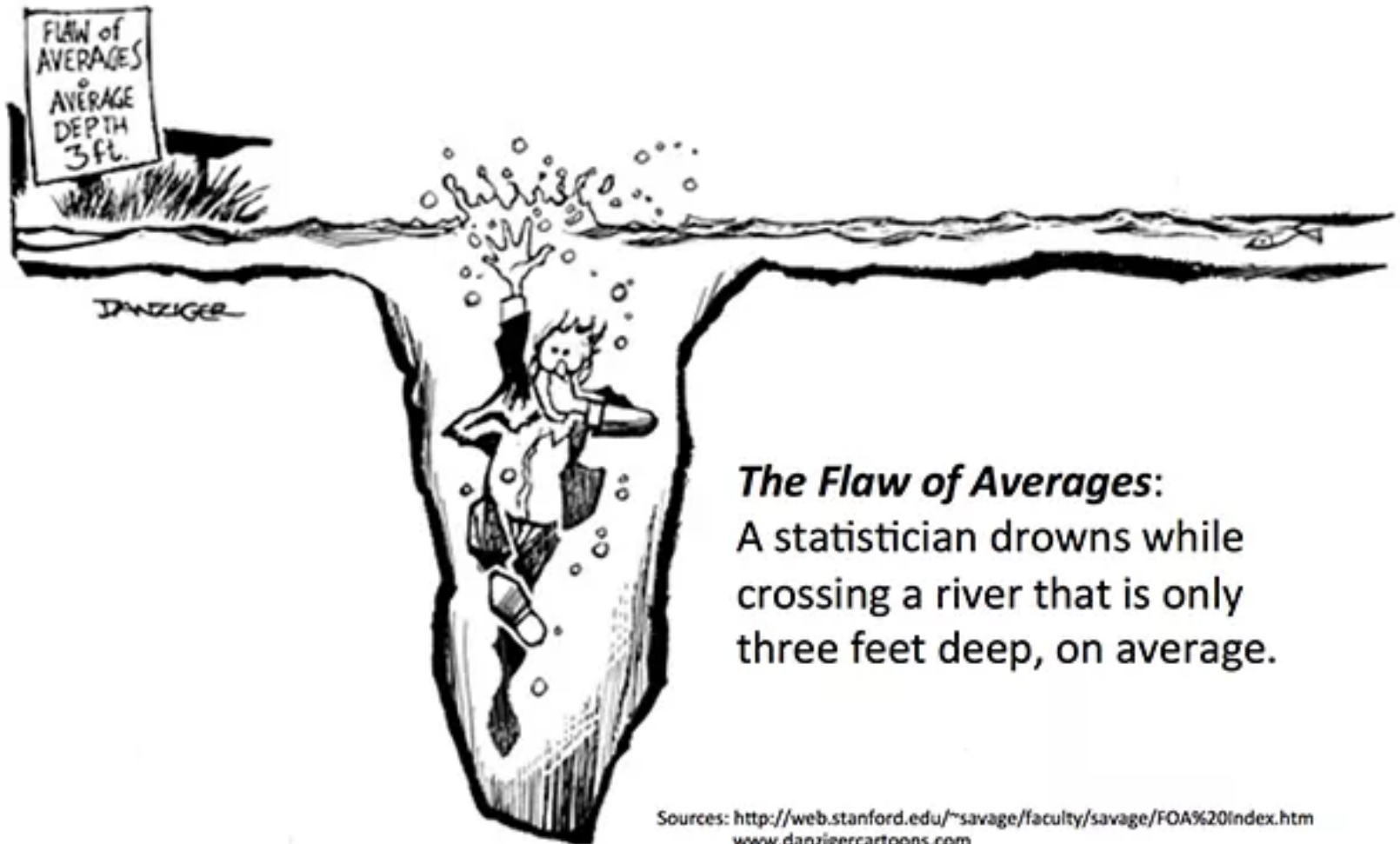
Statistical concerns

1. Data summary

- A vital part of statistics is to succintly convey and report *information*.

- Being versed in the correct statistical techniques will avoid you being duped by flawed statistics.

- One example: *The flaw of averages*

- While it is useful to condense a large amount of information to a singular 'statistic', often it can be misleading.

Statistical concerns
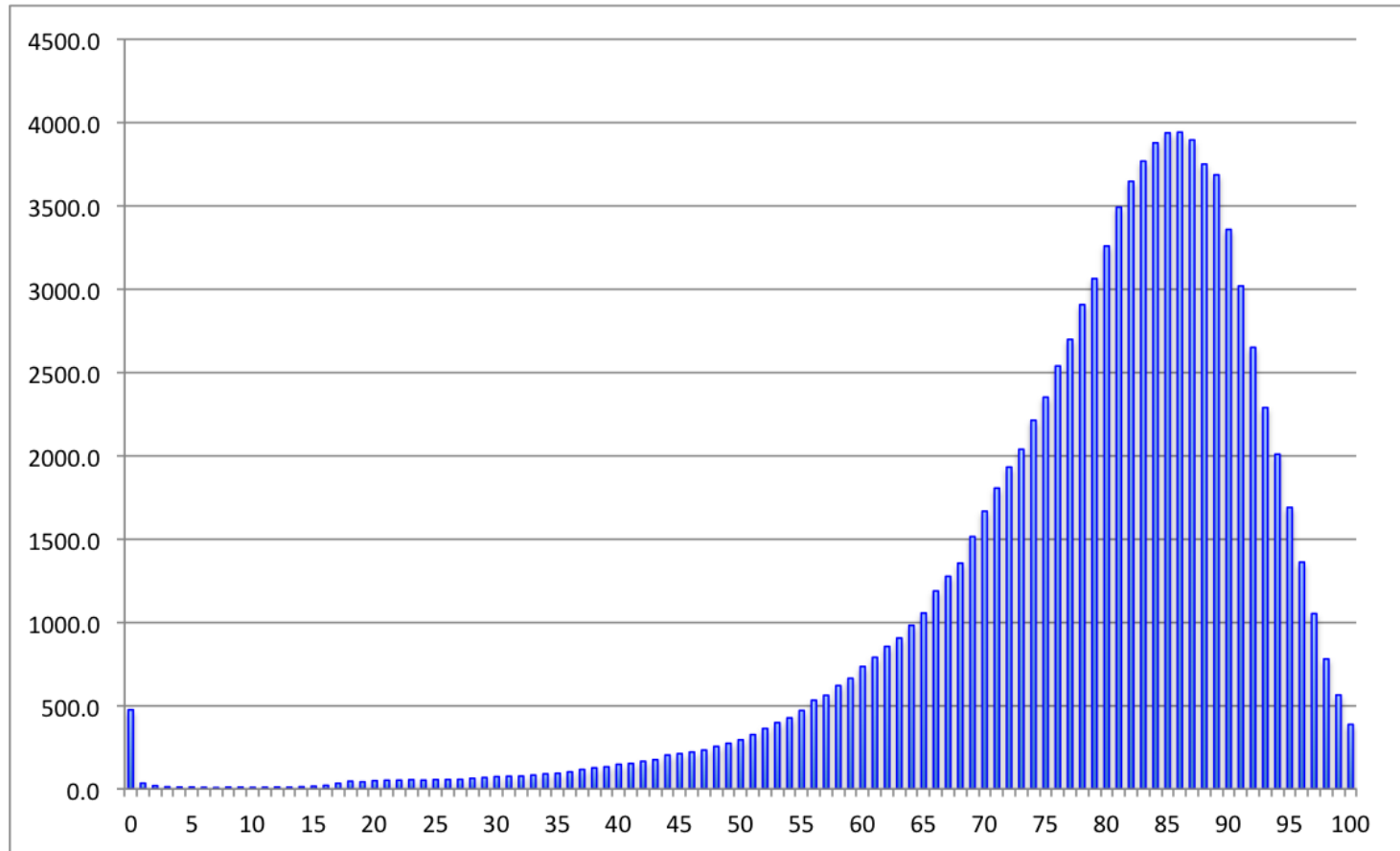
Numbers of men expected to die at each age (UK, mortality rate 2010-2012). Mean age = 79, Median = 82, Mode = 86.

Statistical concerns

The Flaw of Averages: A statistician drowns while crossing a river that is only three feet deep, on average.

Statistical concerns

2. What's your proof?

- Statistics provide the framework for discerning 'credible truth' by means of 'statistical significance'.

- *Example: Kathrina declares that by tasting her bobba drink, she is able to distinguish whether bubble tea pearls were added before or after milk tea was added to the cup.*

- How do we know she's telling the truth? We can set up an experiment of course...

- Eight cups of bobba were prepared, four of which had the pearls poured in first, and the remaining four had the tea poured in first.

- The lady tastes the bobba at random and tells us whether the tea or the pearls was poured in first for each cup.

Statistical concerns

2. What's your proof?

Milk poured first (4 cups)   Tea poured first (4 cups)

- Kathrina correctly guesses **3 out of 8**. Is she telling the truth?

- Kathrina correctly guesses **4 out of 8**. Is she telling the truth?

- Kathrina correctly guesses **5 out of 8**. Is she telling the truth?

- How large is large?

- We can use a hypothesis test to inform us the chance (p-value) of the lady correctly answering the taste tests, under the assumption that her claim is true.

- Low p-values indicate evidence against her claim.

Statistical concerns

3. Measures of association

- Statistics and statistical modelling concerns the measurement of correlations between variables.

- In a lot of scientific studies, we want to understand the behaviour of two or more variables taken together. For example,

  - Low unemployment is *associated* with high GDP output.
  - High satisfaction in the workplace is *associated* with low turnover (job quitting) rates.
  - Lung cancer is found to be *more prevalent* among people who smoke.
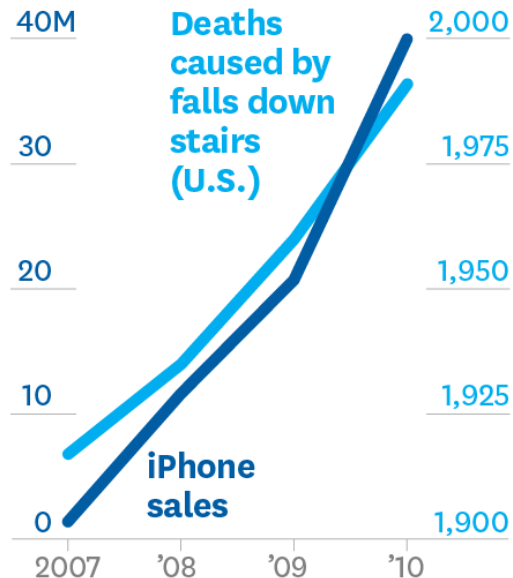
Statistical concerns

3. Measures of association

- Caution: **CORRELATION DOES NOT IMPLY CAUSATION**

- "A study found that children with small feet tend to have low IQ scores"

- Small feet causes low IQ?

- No, children with small feet are younger children, and IQ develops as they grow.
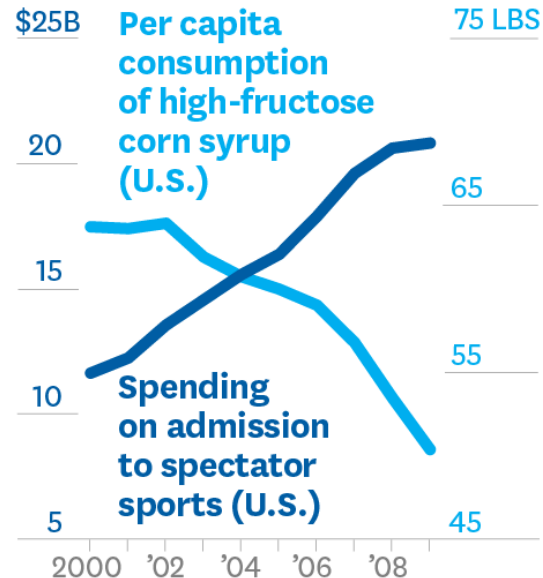
Statistical concerns
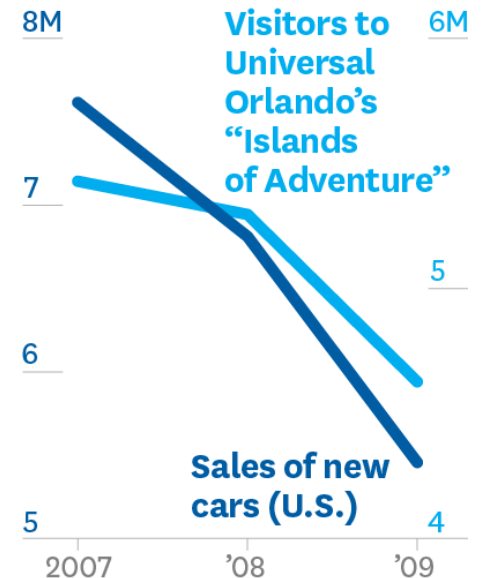
Be wary of spurious correlations

MORE IPHONES MEANS MORE PEOPLE DIE FROM FALLING DOWN STAIRS

Deaths caused by falls down stairs (U.S.)

iPhone sales

40M / 30 / 20 / 10 / 0

2,000 / 1,975 / 1,950 / 1,925 / 1,900

2007 '08 '09 '10

LET'S CHEER ON THE TEAM, AND WE'LL LOSE WEIGHT

Per capita consumption of high-fructose corn syrup (U.S.)

Spending on admission to spectator sports (U.S.)

$25B / 20 / 15 / 10 / 5

75 LBS / 65 / 55 / 45

2000 '02 '04 '06 '08

TO INCREASE AUTO SALES, MARKET TRIPS TO UNIVERSAL ORLANDO

Visitors to Universal Orlando's "Islands of Adventure"

Sales of new cars (U.S.)

8M / 7 / 6 / 5

6M / 5 / 4

2007 '08 '09

Statistical concerns

On the other hand, causation <u>requires</u> STRONG CORRELATION.

"Tea" if by sea, "cha" if by land
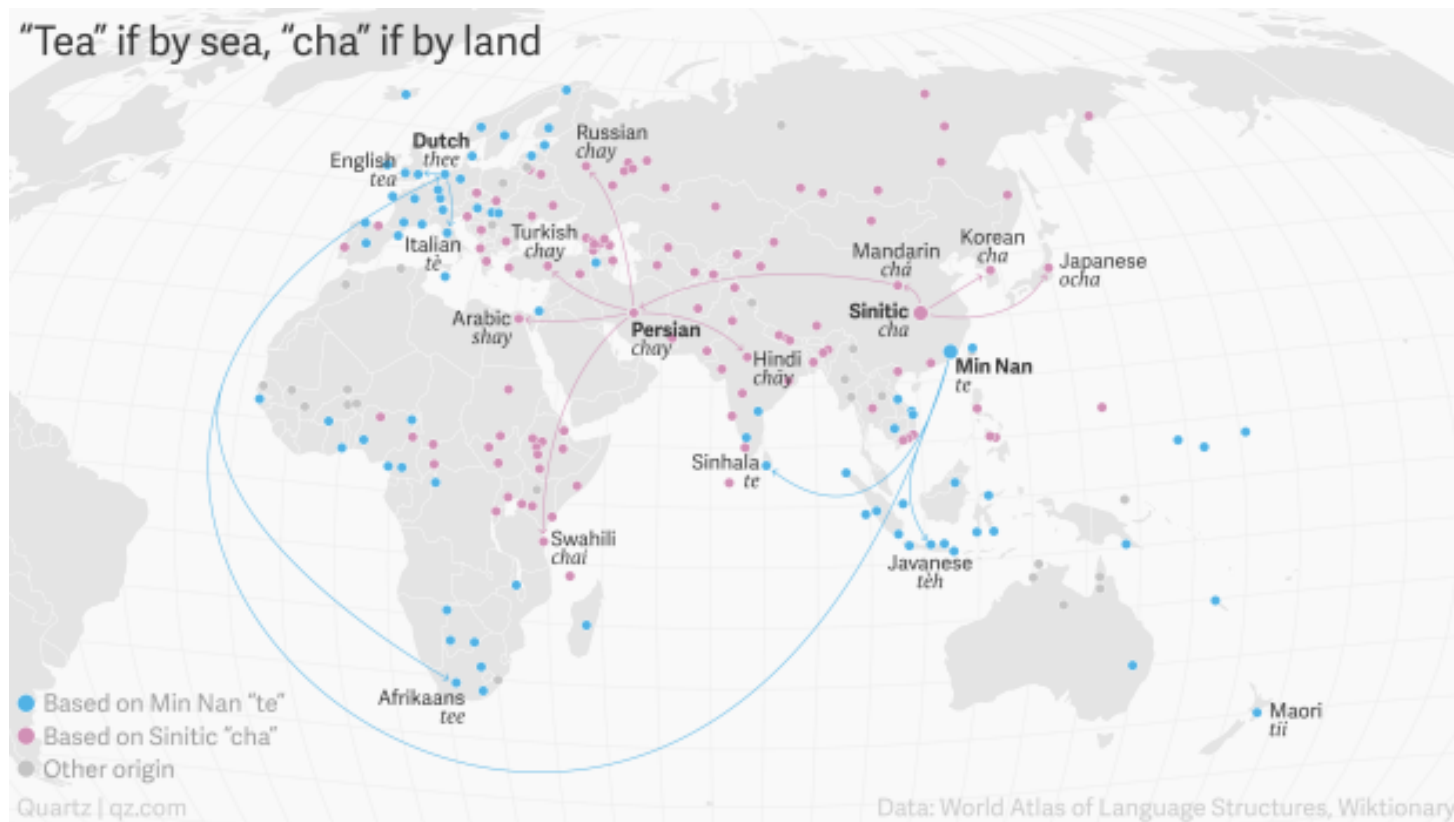
Statistical concerns

The world has only two words for tea: tea or cha. Which word you use very much depends on the region you are in (strong correlation). But there is a cause for this:

Chá (茶) is "Sinitic", common to many varieties of Chinese, eventually becoming chai (چای) in Persian. However, the character is pronounced as 'te', derived probably from tu (荼), in Min Chinese (Fujian etc.).

Statistical concerns

4. Reliability (of data and assumptions)

- Understanding *how* the data was collected, and what (model) *assumptions* were made, allows us to be vigilant when wrong conclusions are drawn.

- *Example: Survey sampling bias.*

- Data that is collected should be *representative* of the population of interest, in order for *valid* inferences to be drawn.

- Types of sampling errors:

  - Convenience sampling.
  - Substitutions.
  - Multiplicity.
  - Volunteering.

Statistical concerns

Where People Can't Live Without The Internet

Share of respondents who can't imagine life without the internet

Country		Share
India	🇮🇳	82%
United Kingdom	🇬🇧	78%
China	🇨🇳	77%
Germany	🇩🇪	73%
United States	🇺🇸	73%
Russia	🇷🇺	66%
Spain	🇪🇸	65%
France	🇫🇷	64%
Italy	🇮🇹	62%
Japan	🇯🇵	62%

Online poll of adults (Sept 09–Nov 10, 2016)

@StatistaCharts Source: Ipsos

Forbes statista

Statistical concerns

5. How likely is it?

- When done correctly, statistical modelling allows us to predict future outcomes (with a degree of confidence).

- This is the main concern of **machine learning**: Learning from data and patterns.

- Statistics also allows us to learn about trends and perform forecasting (time-series analyses).

Statistical concerns

SM-1402 Basic Statistics

- What will we learn in this course?

- Mainly, the foundations and basics of statistics. This includes data summary, probability, and distributions.

- There are several statistics modules being offered here in FOS and they advance progressively.

END