**Wee Kim Wee School of Communication and Information**

**Foundations of Information Analytics (CS2400)**

**Academic Year 2021/22 (Semester 1)**

**General Information**

Lee Chu Keong

WKWSCI #04-08

ascklee@ntu.edu.sg

67904715

**Lecture Timing**

Please refer to schedule.

**Consultation Time**

Essentially, any time before 9:00 p.m., Singapore time! I’m working from home, and if you have questions, just contact me through Microsoft Teams or Zoom!

**“Main” Textbook**

None. Statistics textbooks are typically very expensive (typically > US$100 from Amazon.com), and so I have distilled the most important parts (the examples and the questions) of twelve statistics textbooks into a study guide (a free ebook in PDF format, see below). You are encouraged to print only if necessary.

**CS2400 Resources**

These are the most important resources for CS2400:

⬩ The *Study Guide (8th Edition)* contains a large collection of questions that are relevant to this course. Working through the questions will help you to gain mastery of the concepts taught in class. The *Study Guide* focusses on *mechanique*.

⬩ The *Compilation of Articles* (a folder on NTULearn) contains articles that I refer to during the lectures. To understand how to apply statistics, real-world applications are critical, and the articles are a rich source of details on how the concepts we learn come alive in the world we live in. Many of the articles have to be read several times for the contents to sink in, but it is well worth the effort. The *Compilation of Articles* focusses on *technique*.

⬩ The *New Cambridge Statistical Tables (“NCST”, 2nd ed.)* is a compilation of statistical tables that we will be using for tests of hypothesis we conduct. As the NCST will be used during the examinations, you are encouraged to familiarise yourself with the important tables. A PDF version of the NCST has been uploaded to NTULearn under the folder “Statistical Tables”. (CMIL: QA276.25.L746)

**Reference Books**

Plenty. I will be assigning readings and questions from the following books. The assigned portions will be scanned and uploaded to NTULearn.

Agresti, A., & Franklin, C. (2013). *Statistics: The Art and Science of Learning from Data* (3rd ed.). Boston: Pearson Education. [NIE Library: QA276.12 Agr 2013]

Burleson, D.R. (1980). *Elementary Statistics*. Cambridge, MA: Winthrop Publishers. [NIE Library: QA 276.12 Bur]

Christmann, E.P. (2012). *Beyond the Numbers: Making Sense of Statistics*. Arlington, VA: NSTA Press. [NTU Library: e-book]

Crilly, T. (2007). *50 Mathematical Ideas You Need to Know*. London: Quercus Publishing. (Chapter 34 – Distributions, Chapter 35 – The Normal Curve, and Chapter 36 – Connecting Data only)

Dancey, C.P., & Reidy, J. (2002). *Statistics Without Maths for Psychology* (2nd ed.). Harlow, England, Pearson Education.

Folks, J.L. (1981). *Ideas of Statistics*. New York: John Wiley & Sons. (specifically, Chapter 1: Political Arithmetic) [NIE: HA29 Lev]

Johnson, R., & Kuby, P. (2012). *Elementary Statistics* (11th ed.). New York: BROOKS/COLE.

Kennedy, G. (1983). *Invitation to Statistics*. Oxford, England: Martin Robertson. (specifically, Chapter 2: From Counting to Statistics) [NIE Library: QA276.12 Ken]

Krieg, E.J. (2014). *Statistics and Data Analysis for Social Science*. Pearson.

Levin, J. (1977). *Elementary Statistics in Social Research* (2nd ed.). New York: Harper & Row. [NIE Library: HA29 Lev]

Levin, R.I. (1984). *Statistics for Management* (3rd ed.). New York: Prentice-Hall.

Levitin, D.J. (2016). *A Field Guide to Lies: Critical Thinking in the Information Age*. New York: Dutton. [NTU Library: BC177 L666]

Montaña, R.A., & Bantilan, M.M. (2009). *Introduction to College Algebra*. Manila, Philippines: Rex Book Store.

Montgomery, D.C., Runger, G.C., & Hubele, N.F. (2001). *Engineering Statistics* (2nd ed.). New York: John Wiley & Sons.

Rugg, G. (2007). *Using Statistics: A Gentle Introduction*. Berkshire, England: Open University Press. [NTU Library: e-book]

Salkind, N.J. (2015). *100 Questions (and Answers) About Statistics*. Los Angeles, CA: Sage.

Spiegel, M.R., & Constable, R.L. (1992). *Schaum’s Outline of Theory and Problems of Statistics* (2nd ed. in SI Units). London: McGraw-Hill. [NTU Library: HA29.S755 1992]

Steinberg, W.J. (2011). *Statistics Alive!* (2nd ed.). Thousand Oaks, CA: Sage. [NTU Library: HA29 S819 2011]

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: HarperCollins. [NTU Library: QA76.9.D343S832]

Stroud, K.A., & Booth, D.J. (2001). *Engineering Mathematics*. London: Palgrave.

Triola, M.F. (2008). *Elementary Statistics with Multimedia Study Guide* (10th ed.). New York: Pearson.

Utts, J.M., & Heckard, R.F. (2015). *Mind on Statistics* (5th ed.). Stamford, CT: Cengage Learning.

Washington, A.J. (2010). *Basic Technical Mathematics with Calculus* (9th ed.). Toronto, Canada: Pearson.



Think before you speak. Read before you think.

Fran Lebowitz



**Articles**

Brasseur, L. (2009). Florence Nightingale’s Visual Rhetoric in the Rose Diagrams. *Technical Communication Quarterly, 14*(2), 161−182.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics, 26*(4), 745-766.

Ferguson, D. (2013). How supermarkets get your data – and what they do with it. The Guardian.

Gurin, J. (2014). Opening Business Innovation with Open Data. *Business Horizon, 12*, 42−49.

Hess, A. (May 14, 2017). Open Secrets. *The New York Times Magazine*. New York: New York Times.

Khan, M.A., Uddin, M.F., & Gupta, N. (2014). Seven V’s of Big Data: Understanding Big Data to Extract Value. *Conference of the American Society for Engineering Education*.

Loukides, M. (2010). *What is Data Science*? Sebastopol, CA: O’Reilly.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10), 1−23. (Sections 1 to 3 only)

**Videos**

*The Best Stats You've Ever Seen (A TED Talk by Hans Rosling)* (https://www.youtube.com/watch?v=hVimVzgtD6w)

*The Joy of Stats. By Hans Rosling* (http://www.gapminder.org/videos/the-joy-of-stats/)

*Trash Trail: Episode 5 (Data: Finale)* (https://video.toggle.sg/en/video/series/trash-trail/ep5/478356)

**Reports on Big Data**

Hewlett-Packard Development Company. (2015). *The Disruptive Power of Big Data*.

The Networked Software and Services Initiative (NESSI). (2012). *Big Data: A New World of Opportunities*.

**Websites**

• University of Glasgow’s Statistics Glossary (http://www.stats.gla.ac.uk/steps/glossary/index.html)

• NIST/SEMATECH e-Handbook of Statistical Methods (http://www.itl.nist.gov/div898/handbook/)



Torture numbers, and they will confess to anything.

Gregg Easterbrook



**Course Description**

Today, many organisations generate, and collect (“harvest”), unimaginable quantities of data of all types. However, merely collecting lots of data is pointless. The critical step is to analyse the data so that it can be transformed into information and action. The key idea is to transform data in such a way that it can be used for business advantage. An important tool that enables this transformation is statistics. This is the subject matter of this course. Statistics will be presented in a mathematically friendly and non-threatening manner. The course emphasises conceptual understanding of the material, and not on the exact keystrokes needed to accomplish specific statistical tests.

**Course Objectives**

* To sensitise you to the fact that data is all around us
* To start you thinking about the opportunities for transforming data into action (new products / processes / …)
* To lay the statistical foundations for data analytics so that you can transform the data into actionable information

**NTULearn Website**

All course materials will be uploaded to the NTULearn website (<http://ntulearn.ntu.edu.sg/>). All the online lectures will be delivered live, and recorded for reference and revision.

**Topics Covered**

| Week | Topic |
| --- | --- |
| **1** | **Introduction to Data Analytics**  ❖ When did Data Analytics “Start”?  ❖ Data *vs* Information  ❖ Personal Data  ❖ Sensitive Personal Data  ❖ Relational Data  ❖ Big Data (The Seven Vs of Big Data)  ❖ Open Data  ❖ Data Exhaust  ❖ Data Products  **Historical Use of Big Data**  ❖ Carolus Linnaeus  ❖ James Murray  ❖ John Snow |
| **2** | **Exploratory Data Analysis**  ❖ Frequency Polygons  ❖ Stem-and-Leaf Plots  ❖ Histograms  ⬩ Lower Class Limit (LCL) and Upper Class Limit (UCL)  ⬩ Class Width  ⬩ Class Midpoint  ❖ Bar Charts  ❖ Scattergrams  ❖ Pie Charts  ❖ The Invention of the Rose / Polar Diagrams by Florence Nightingale  Specialised Diagrams or Charts  ❖ Pareto Charts  ❖ Population Pyramids  ❖ Triangular Graphs  ❖ Chernoff Faces  ❖ Digidot Plots |
| **3** | **Describing the Data You Find**  ❖ Measures of Central Tendency  ❖ Understanding the Sigma () Notation  ❖ Measures of Spread  ❖ Measures of Shape |
| **4** | **A Quick Revision of Probability**  ❖ Statistical Distributions  ❖ The Process of Normalisation  ❖ The Normal Distribution  ❖ Statistical Tables (specifically, *The New Cambridge Statistical Tables*, “NCST”)  ❖ The relationship between probability and statistics  **Applications of Distributions**  ❖ Ladislaus J. Bortkiewicz  ❖ Frank Benford  ❖ George Kingsley Zipf |
| **5** | **Statistical Hypothesis Testing**  ❖ The Logic of Hypothesis Testing  ❖ The [Five/Six/Seven] Steps Involved  The -Tests:  ❖ The One-Sample -Test  ❖ The Two-Sample -Test |
| **6** | **The -Tests**  ❖ One-Sample -Test  ❖ Two-Sample -Test  ⬩ Equal Variances: Pooled t-test  ⬩ Unequal Variances  ❖ Paired -Test |
| **7** | **Five Miscellaneous Topics**  Applications of Statistics in Research |
| **8** | **Correlation**  ❖ Calculating the Summary Statistics (, , , , , , )  ❖ Calculating , , and  ❖ Pearson Product Moment Correlation Coefficient (PMCC),  ❖ Spearman’s Rank-Order Correlation Coefficient  ❖ Kendall’s Rank Correlation Coefficient |
| **9** | **Simple Linear Regression**  ❖ Calculating the Regression Coefficients ( and )  ❖ Understanding the Difference Between Interpolation and Extrapolation  ❖ Coefficient of Determination ()  ❖ Coefficient of Non-determination ()  ❖ Calculating Leverage () |
| **10** | **Multiple Regression**  ❖ Formulating the Standard Equations  (i) in the equation format  (ii) in the matrix format  ❖ Solving the Standard Equations Using Cramer’s Rule  **Polynomial Regression**  ❖ Formulating the Standard Equations  (i) in the equation format  (ii) in the matrix format  ❖ Solving the Standard Equations Using Cramer’s Rule |
| **11** | **Chi-Square Analysis**  Characteristics of the Chi-Square Distribution  ❖ Chi-Square Goodness-of-Fit Test  ❖ Chi-Square Test of Independence |
| **12** | **Non-parametric Statistical Tests**  ❖ Mann-Whitney Test  ❖ Wilcoxon Test  ❖ Kruskal-Wallis Test  ❖ Friedman Analysis of Variance by Ranks |
| **13** | **More Non-parametric Statistical Tests**  ❖ Runs Test  ❖ Phi Coefficient |
| **14** | **ANOVA**  ❖ Limitations of the -Test  ❖ ANOVA Terminology  ❖ Null and Alternative Hypotheses for ANOVA  ❖ Source of Variations  (1) Chance Variation  (2) Variation Due to Manipulation of the Factor  ❖ The Variance Ratio,  ❖ Percentage Points of the -Distribution |

**Independent Readings**

These readings have been uploaded to NTULearn. Each student will be assigned one reading.

**Books**

1. Black, E. (2012). *IBM and the Holocaust: The Strategic Alliance Between Nazi Germany and America’s Most Powerful Corporation*. Dialog Press.

2. Johnson, S. (2007). *The Ghost Map: The Story of London’s Most Terrifying Epidemic*. Riverhead Books.

3. Winchester, S. (2005). *The Professor and the Madman*. Harper Perennial.

**Assessment Components (Covid Period)**

| Component | Deadline / Date (Day) | Weightage |
| --- | --- | --- |
| Mid-semester Test (1-hour test) | To Be Announced (TBA) | 20% |
| Group Assignment | TBA | 20% |
| Presentations | TBA | 20% |
| Final Test (1-hour test) | TBA | 20% |
| Individual Assignment | TBA | 20% |
|  | Total | 100% |

**Policy on Plagiarism**

All work presented in this class must be the product of your own effort. Your work should not be copied without appropriate citation from any source, including the Internet. Any student caught presenting work which is not his or her own will face disciplinary action, which may include award of zero marks for the assignment, receiving a failing grade for the class, or being expelled from the university. This policy applies to all work submitted, either through oral presentation, or written work, including outlines, briefings, group projects, self-evaluations, etc. You are encouraged to consult me if you have questions concerning the meaning of plagiarism or whether a particular use of sources constitutes plagiarism. Details on academic integrity can be found from <http://www.ntu.edu.sg/ai/Pages/index.aspx> (Academic Integrity at NTU).