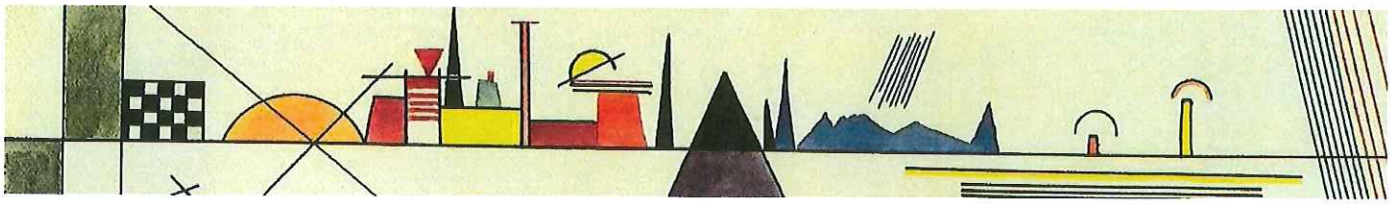

# STA303/STA1002: Methods of Data Analysis II (Summer 2016)

## REGISTERING FOR STA1002 (FOR GRAD STUDENTS): SEE HERE

**About STA303/STA1002**

*Overview:* The aim of this course is to introduce the most common data analysis techniques used for analyzing real-world data that do not conform to the assumptions of the Linear Model. We will be analyzing data that displays non-linear patterns, frequency data, count data, and longitudinal data. Students will get practice with exploratory data analysis (data visualization, model selection, formulating a hypothesis) and with statistical inference for regression models. Data analysis will be done in R and reproducible assignment reports will be authored using R Markdown.

*Prerequisites:* I will assume that students are familiar with linear regression, have used of a statistical package such as R for linear regression, and have a a reasonable degree of facility with mathematical reasoning about statistical models (at the level of STA302).

**Teaching team**

Instructor: Michael Guerzhoy. **Office:** BA5244, **Email:** `guerzhoy at cs.toronto.edu` (please include STA303/STA1002 in the subject, and please ask questions on Piazza if they are relevant to everyone.)

TAs: Tiffany Fitzpatrick, Luhui (Luke) Gan

**Getting help**

Michael's office hours: Thursday 6-7PM, Friday 3-4PM. Or email for an appointment (Thursday and Friday afternoon/evening strongly preferred). Or drop by to see if I'm in. Feel free to chat with me after lecture.

Course forum on Piazza

*Piazza is a third-party discussion forum with many features that are designed specifically for use with university courses. We encourage you to post questions (and answers!) on Piazza, and read what other questions your classmates have posted. However, since Piazza is run by company separate from the university, we also encourage you to read the privacy policy carefully and only sign up if you are comfortable with it. If you are not comfortable with singing up for Piazza, please contact me by email to discuss alternative arrangements.*

**References**

There is no perfect textbook that fits the syllabus of STA303/STA1002. The following are good starting points:

- Michael Kutner, Christopher Nachtsheim, John Neter, Applied Linear Regression Models
- Howard J. Seltman, Experimental Design and Analysis — a more elementary book than what we need (just discusses the techniques while sometimes omitting the intuition/rationale/theory), but covers t-tests and ANOVA.
- Alan Agresti, Introduction to Categorical Data Analysis — covers most of what we need, but unfortunately not t-tests, ANOVA, and multiple comparisons (available on the web via the UofT library)
- Fred Ramsey and Daniel Shafer, The Statistical Sleuth: A Course in Methods of Data Analysis (see also The Statistical Sleuth (3rd Edition) In R) — an excellent book that can sometimes be sparse on details.
- Cosma Rohilla Shalizi, Advanced Data Analysis from an Elementary Point of View — a wonderful book about modern data analysis techniques. Some chapters are very relevant (although not directly covered), and others are too advanced.

**Software**

We will be using RStudio to author reproducible data analysis reports using R and R Markdown.

**Projects**

Project 1 (10%): ANOVA, multiple comparisons, and simulation. Due: Thursday Jul. 14 11PM. Some R tips for P1: Part 1, Part 2.
Project 2 (15%). Due: Monday Aug. 1 11PM

*Lateness policy: 10% per 24 hours, rounded up. Late projects are only accepted for 48 hours after the deadline.*

**Project submission**

Projects are to be submitted on MarkUs. You can log in using your UTORid.

**Midterm**

Monday Jul. 18. Worth: 25%

**Exam**

TBA. Worth: 50%

**Practice problems**

**Conceptual problems**: Study Guide (more to come). You can add your solutions, and read other people's solutions, here.

**One-Way ANOVA and t-tests**: Problems. Supplemntary data and analysis: drug trial analysis from Kleibaum (source), Spock dataset (source). Solutions.

**Two-Way ANOVA**: Problems. Solutions.

**Logistic Regression**: Problems. Supplementary data and analysis: Donner Party (source), counterfeit banknotes (source), new cars (source). Solutions.

**Old tests and exams**: here.

Unadapted practice problems are available here.

**Lecture notes**

**Lecture 1**: Intro, t-Tests (R code, source). Video tutorials on simulation: Part 1, Part 2.

*At students' request, I am posting relevant reading. You are only responsible for what's in the lectures, but of course it's always good to read a textbook as well. I do not expect that everyone consults all the readings I post, only that people make sure that they thoroughly understand the lectures.*

**Reading**: Seltman Ch. 6 ("t-test"). Ramsey Ch. 2, 3 ("Inference Using t-Distributions", "A Closer Look at Assumptions")

Just for fun: the American Statistical Association's statement on p-values; more advanced (and *slightly* sarcastic) post from Andrew Gelman: "I've never in my professional life made a Type I error *or* a Type II error"

**Lecture 2**: t-Tests continued (R code, source). One-Way ANOVA (R code, source).

**Reading**: Seltman Ch. 7 ("One-way ANOVA"). Ramsey Ch. 3 ("A Close Look at Assumptions"), Ramsey Ch. 5 ("Comparisons Among Several Means").

**Lecture 3**: Degrees of Freedom, More on P-values, two-way anova (R code, source)

**Reading**: the appropriate chapters from Kutner (different depending on the edition); the Two-Way ANOVA chapter in Seltman

Testing hypotheses about sigma, and more simulation: R code, source.

**Lecture 4**: An overview of F-tests (R code, source), Binary response variables (R code, source), Logistic Regression (R code, source, data)

**Reading**: Agresti Chapters 4-5 (not all sections).

Also: more fish! Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon. More Brains! Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates.

**Just for fun**: the Titanic was actually a typical. Typically, more men than women survive: M. Elinder and O. Erixson, Gender, social norms, and survival in maritime disasters, PNAS vol. 109 no. 33, 2012.

**Lecture 5**: More on multiple comparisons, more on fixed intercepts (R code source), Goodness of Fit: Logistic Regression (R code, source).

**Just for fun:** FiveThirtyEight's p-value clip.

**Just for fun:** the Dunning-Kruger effect study.

**Simulation reading**: Shalizi Chapter 5

**Lecture 6**: The midterm, cross validation (R code, source), Issues in logistic regression (R code -- perfect separation, source, R code -- extrabinomial, source).

Reading: Shalizi Ch. 3 on cross-validation.