

STA2201H Methods of Applied Statistics II

Monica Alexander

Week 1: Introduction

Overview

- ▶ Introductions
- ▶ What is applied statistics
- ▶ Course outline and goals
- ▶ Motivating example
- ▶ Reproducible research
- ▶ Tools
- ▶ Lab: Intro to git, tidyverse, RMarkdown

Contact

- ▶ Email: monicaalexander@utoronto.ca.
 - ▶ I do not check/answer emails after 5pm or on weekends.
- ▶ Office hours will be via appointment.

A bit about me

Me:

- ▶ statistics \cap chemistry \rightarrow social science \cap statistics
- ▶ 50/50 Statistical Sciences and Sociology departments
- ▶ Not Canadian (Australia \rightarrow USA \rightarrow Canada)

What I work on: a mix of demography, applied stats, epidemiology and computational social science

What is demography?

Demography is the scientific study of population dynamics. We are interested the size, composition and distribution of populations over time, and study these changes with respect to the three main population processes:

- ▶ Births (Fertility)
- ▶ Deaths (Mortality)
- ▶ Migration

Statistical methods become important for estimation because:

- ▶ often have no/bad data
- ▶ often dealing with survey data/lots of measurement errors
- ▶ demographic events as stochastic processes

What is applied statistics? / how does it relate
to data science?

What is applied statistics?

Using statistical methods to answer questions and draw reasonable conclusions from data that have uncertainty and randomness.

Emphasis is on **data**

- ▶ you need to understand your data in order to make decent inferences
- ▶ data generating process, measurement errors, correlations, dependence. . .
- ▶ as statisticians we often don't collect the data so easy to forget this
- ▶ EDA, data visualization
- ▶ By definition applied to some other area we may or may not be (probably not) trained in: need to be aware of substantive topic and issues

How does it relate to data science?

Data science is:

- ▶ "... a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data" (Wikipedia)
- ▶ Being able "to understand [data], to process it, to extract value from it, to visualize it, to communicate it." (Hal Varian)
- ▶ "... the ability to extract knowledge and insights from large and complex data sets." (DJ Patil)
- ▶ Statistics + data munging + data viz (Nathan Yau)

How does it relate to data science?

My take: data science is a mix of CS, stats and ML. An emphasis on the whole pipeline:

1. data collection/extraction
2. data storage/maintenance
3. data manipulation/processing
4. **data analysis** (applied statistics, ML)
5. communicate output (often predictions)

Notes:

- ▶ The applied stats part may only be a small part and may be automated (this is doable if focus is on prediction)
- ▶ Relatively easy to re-brand as a data scientist if that's your jam
- ▶ Reproducibility is important

Course overview and goals

Course overview

- ▶ Topics will include generalized linear models, Bayesian inference, generalized linear mixed models, generalized additive models involving non-parametric smoothing, model evaluation and selection. We will also cover some core statistical computing techniques.
- ▶ A large focus of the outcomes on this course will also be on reproducible research, identifying and dealing with data and modeling issues, and model interpretation and communication.
- ▶ Throughout the course we will be using R in all examples, labs and homework assignments. Exams will also require interpretation of R output.
- ▶ Each week will be a lecture (~1-1.5hrs) then a lab
- ▶ Everything is online

We're all out here doing our best

- ▶ The current situation makes both learning and teaching challenging
- ▶ Try to be understanding of everyone's sub-optimal situation
- ▶ Communication is key
- ▶ There may be guest appearances from my toddler

Doing okay?

It is always okay to reach out for support.

- ▶ My Student Support Program – My SSP – mental health support for all U of T students. Free, confidential, immediate support. Available 24/7 in multiple languages. Download the My SSP App or call 1-844-451-9700. uoft.me/myssp
- ▶ Call Good2Talk. Free, confidential helpline with professional counseling, information and referrals for mental health, addictions and well-being, 24/7/365 1-866-925-5454

Assessment

- ▶ Lab exercises, 2% per week
 - ▶ Due 9am the following Friday
 - ▶ Hand in via git
- ▶ Two assignments, 15% each
 - ▶ Mostly data analysis
 - ▶ Hand in via Quercus
- ▶ Mid term, 15%
 - ▶ Online during class hours (2 hours)
 - ▶ Multiple choice and short answer
- ▶ Research project 35%
 - ▶ Pick a dataset, research question and statistical approach
 - ▶ Research proposal (7.5%)
 - ▶ Research paper (20 %)
 - ▶ Presentation last week of class (7.5%)

Goals

We will be doing applied statistics in the truest sense of the term

- ▶ Learn a useful suite of statistical techniques
- ▶ Be able to deal with real data
- ▶ Assess data and model issues
- ▶ Establish/streamline/improve project workflow

Expectations

- ▶ Understand main ideas behind important techniques for applied statistics
- ▶ Coding in R (and in particular, the tidyverse)
- ▶ R markdown
- ▶ Git
- ▶ Code readability
- ▶ Clear communication of methods, findings, limitations
 - ▶ Data exploration is part of this!
- ▶ Aim for reproducible research

Roadmap

Roadmap

Subject to change depending on time and priorities.

Planned lecture content:

- ▶ Generalized linear models recap
- ▶ Survival analysis
- ▶ Bayesian inference
- ▶ Visualizing the Bayesian workflow and model checks
- ▶ Multilevel models
- ▶ Temporal models
- ▶ Non-linear/ non-parametric models (splines)
- ▶ Time/interest permitting: text analysis?

Roadmap

Planned lab content:

- ▶ Rmarkdown, git
- ▶ Tidyverse
- ▶ EDA, data viz
- ▶ RShiny
- ▶ glm
- ▶ Stan, brms
- ▶ Probably: web scraping
- ▶ Maybe: Extracting data from API (e.g. Facebook or Twitter), AWS

Motivating example

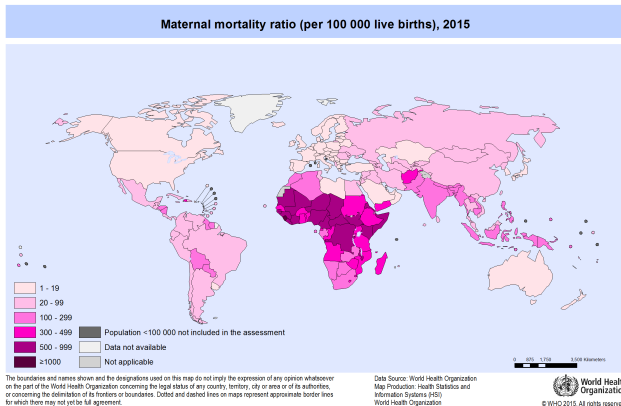
Global estimation of the causes of maternal death

- ▶ **Maternal mortality:** the death of a woman while pregnant or within 42 days of termination of pregnancy, from any cause related to or aggravated by the pregnancy.
- ▶ Very important indicator of health and development of a country
- ▶ Part of the Sustainable Development Goals (3.1)



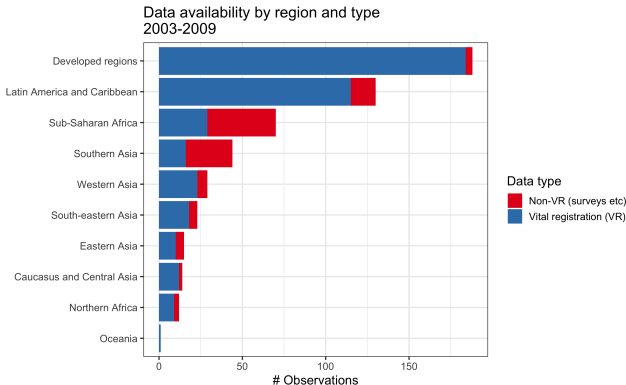
Global estimation of the causes of maternal death

- ▶ Large variation in maternal mortality ratio (deaths per 100,000 births) across the world (highest: 1150; lowest: 2)
- ▶ In order to reduce number of deaths, need to know underlying causes
- ▶ But this is difficult information to obtain/estimate



How do we get information on causes of (maternal) death?

- ▶ In high-income countries and some middle-income countries: civil registration systems
- ▶ In low-income countries: ???
 - ▶ surveys (why is this hard?)
 - ▶ facility-based administrative data
 - ▶ other specialized studies



How do we get information on causes of (maternal) death?

- ▶ If we had complete coverage of all deaths and a reliable way of classifying cause of death, then we could just count deaths and call it a day
- ▶ But in most countries (particularly high-burden countries) we have very little information, and what we do have is full of problems
- ▶ → Use statistical methods to obtain as reliable estimates as possible

Issues

To name a few:

- ▶ Years with no data
- ▶ Only some causes observed (even in high-income countries)
- ▶ Non-representative data (subnational, facility-based)
- ▶ Cause of death classification issues (death not witnessed, definition changes, differences across countries etc)
- ▶ Under/over-reporting (especially abortion)
- ▶ Not all civil registration systems are high quality
- ▶ Low death counts (~ 25 deaths in Australia)

Intro to statistical set-up

Notation:

- ▶ observations $i = 1, \dots, n$
- ▶ d_i is total number of maternal deaths for the i th observation
- ▶ observed maternal deaths $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,7})$
- ▶ $y_{i,j}$ is the number of deaths due to cause j for the i th observation
- ▶ cause groups $j = 1, \dots, 7$ corresponding to {ABO, EMB, HEM, SEP, DIR, IND, HYP}

Intro to statistical set-up

Think of deaths as a stochastic process:

- ▶ Given total number of maternal deaths d_i , the probability of a death is due to cause j is $p_{i,j}$. This is a Multinomial distribution, with 7 categories:

$$\mathbf{y}_i \sim \text{Multinomial}(d_i, \mathbf{p}_i)$$

$$\mathbf{p}_i = (p_{i,1}, \dots, p_{i,7})$$

- ▶ We observe $y_{i,j}$ and d_i
- ▶ We are interested in estimating \mathbf{p}_i . These will help us get estimates for the 'true' proportions \mathbf{p}_c for countries $c = 1, \dots, 193$ (UN member countries)

Intro to statistical set-up

$$\mathbf{y}_i \sim \text{Multinomial}(d_i, \mathbf{p}_i)$$

$$\mathbf{p}_i = (p_{i,1}, \dots, p_{i,7})$$

Put a model on \mathbf{p}_i :

- ▶ Transform to ensure probabilities sum to 1
- ▶ Model can include effects/adjustments for different things e.g. region, data quality, temporal changes, subnational adjustments. . .
- ▶ This is a (Bayesian) hierarchical model. We will learn about these!

Maternal mortality summary

- ▶ Real world problem, working with WHO and statisticians, epidemiologists, clinicians, public health officials
- ▶ So many data problems
- ▶ Data complexities lead to relatively complex models
- ▶ Substantive area knowledge helps to understand data issues
- ▶ Results have big impact (policy, \$\$\$): need to be careful, transparent with assumptions, reproducible

Reproducible research

What is reproducibility?

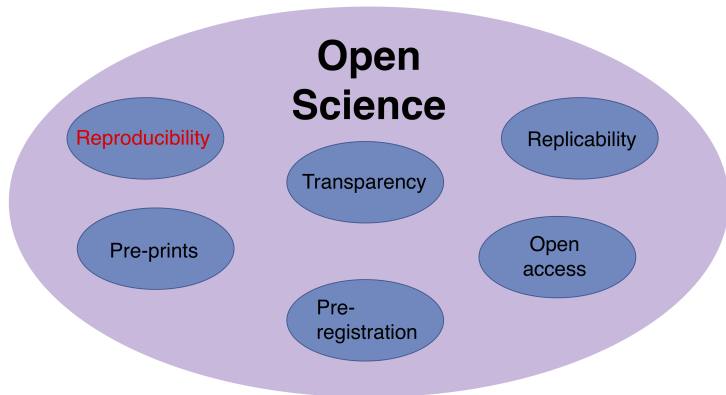
- ▶ Research is **reproducible** if it can be reproduced exactly, given all the materials used in the study
 - ▶ Note that materials need to be provided!
 - ▶ For us, 'materials' usually means data, code and software
 - ▶ Reproduce the data, methods and results (including figures, tables)
- ▶ Another person should be able to take the exact same data, run the exact same analysis, and produce the exact same results
- ▶ Different to **replicability**
 - ▶ carrying out a new study based on the description of the data and method provided in the original publication, and obtaining results that are similar enough.

Increased awareness recently

'Replication crisis', starting in Psychology, now extended to other fields

- ▶ Famous results called into question (e.g. Marshmallow test, power poses, ego depletion)
- ▶ Issues range from weaker evidence than originally thought, to fabrication of data

Just one part of doing open science



Tools

Tools

- ▶ R
- ▶ Tidyverse
- ▶ RMarkdown
- ▶ git

R

We will be using R in this course. Pros:

- ▶ Free
 - ▶ reproducibility
 - ▶ portability
- ▶ Open
 - ▶ large community
 - ▶ lots of packages
 - ▶ lots of help

RStudio:

- ▶ IDE for R that makes using R a lot nicer and easier
- ▶ If you haven't already got it, download the free version here:
<https://rstudio.com/products/rstudio/download/>

Tidyverse

- ▶ R Packages contain R functions, the documentation that describes how to use them, and sample data.
- ▶ The 'tidyverse' is "an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures."
<https://www.tidyverse.org/>
- ▶ ggplot probably the most well known
- ▶ Style of coding fundamentally different to base R.
- ▶ A lot of other packages now produce output objects in the 'tidy' form

RMarkdown

- ▶ Markdown is plain text formatting syntax that can be converted into lots of different outputs (eg HTML, PDF)
- ▶ R Markdown allows you to combine Markdown (for the report writing) and embedded R chunks, which are dynamically updated when the document is compiled
- ▶ R code can be in chunks or inline (e.g the fourth root of π is 0.7853982)
- ▶ These slides are written in RMarkdown and knitted to PDF (beamer)

```
282 - ## RMarkdown
283
284 - Markdown is plain text formatting syntax that can be converted into lots of different outputs (eg HTML, PDF)
285 - R Markdown allows you to combine Markdown (for the report writing) and embedded R chunks, which are dynamically updated when
the document is compiled
286 - R code can be in chunks or inline (e.g the fourth root of  $\pi$  is `r pi*(1/4)`)
287 - These slides are written in RMarkdown and knitted to PDF (beamer)
288
289 \begin{figure}
290 \includegraphics[width = 0.8\textwidth]{turtles.png}
291 \end{figure}
```

RMarkdown

- ▶ Good reproducibility tool
- ▶ Can do most things you can do in LaTeX (writing math is the same)
- ▶ You are expected to write up assignments in RMarkdown

git

- ▶ git is a version control system (think a more complicated Dropbox)
- ▶ Designed for software engineers, but useful for all sorts of code
- ▶ Useful for both collaborative and solo projects
- ▶ GitHub is useful place to host open source projects

Summary

Summary

- ▶ Applied statistics has a focus on **data**
 - ▶ Understanding where the data come from, generating process, issues, etc
- ▶ The implication is that we need to think carefully about model assumptions and whether they make sense
- ▶ Reproducibility is important, especially if we want our research to be useful

Lab

To-dos

- ▶ Install R/RStudio if you haven't already
- ▶ Get GitHub account if you haven't already

Git component

Git

- ▶ Git is a version control system
- ▶ The system tracks changes you make to git repositories ('repos')
- ▶ Think of repos as folders
- ▶ In order for file versions to be tracked, they need to be **committed** to the git repo
- ▶ Think of committing as like saving, but with slightly more steps

GitHub

<https://github.com/>

- ▶ A hosting service for git repos
- ▶ You can sign up for free, and host an unlimited number of public or private repos
- ▶ You will be submitting lab exercises via GitHub, so you need to set up an account!

The simplest Git/GitHub workflow

- ▶ New repo on GitHub
- ▶ Clone onto local computer
- ▶ Do work on local computer
- ▶ Save
- ▶ Add and commit to git repo
- ▶ Push to GitHub (this means your new work will appear on the GitHub website)

If you are working on your own, on one computer, this is it!

Git/GitHub

- ▶ If you are working on a couple of different computers / servers, you may also need to **pull** from GitHub to update any new work done elsewhere
- ▶ Git is designed for collaborative work. More complicated workflows have branches, pull requests, merges
- ▶ More later (time permitting)

Git

For now, you just need to learn

- ▶ clone
- ▶ status
- ▶ add
- ▶ commit
- ▶ push
- ▶ pull

Steps on GitHub

1. Create an account on GitHub
2. Click the new repository green button
3. Name it something sensible (e.g. STA2201H-applied-stats), select private, select initialize with README, click create
4. Settings -> Collaborators -> Add MJAlexander as a collaborator

Steps to clone on your computer

Disclaimer: I use the terminal and will show you these steps.

You are welcome to use the GitHub Desktop:

<https://desktop.github.com/>

Steps to clone on your computer

1. open terminal window
2. cd into place you want to save the folder
3. `git clone`
`https://github.com/yourusername/yourrepo`

Steps to add work, commit and push

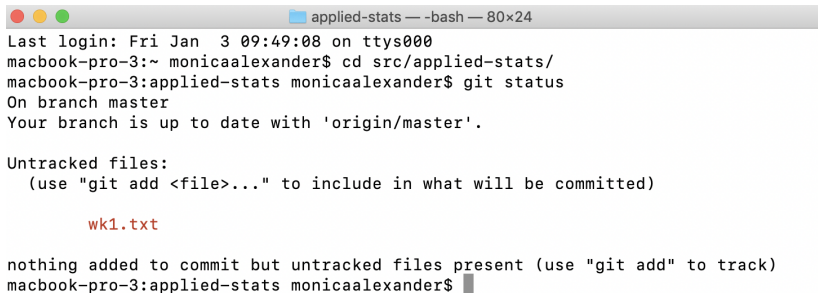
Save some work:

- ▶ Create a text file with your name, the program you are enrolled in and something you miss doing during lockdown
- ▶ Save it as something like `wk1.txt` to the git repo folder (as you would normally save something)

Steps to add work, commit and push

Commit work

- ▶ open terminal window
- ▶ cd into git repo
- ▶ git status should show uncommitted work

A screenshot of a macOS terminal window. The title bar shows three colored window control buttons (red, yellow, green) on the left and a title bar with a blue folder icon, the text 'applied-stats', and a shell prompt '-bash' followed by the window size '80x24'. The terminal text shows the user logging in, navigating to the 'src/applied-stats/' directory, and running 'git status'. The output indicates the user is on the 'master' branch and the branch is up to date. It then lists 'Untracked files:' as 'wk1.txt' and prompts the user to use 'git add' to track the file. The prompt 'nothing added to commit but untracked files present (use "git add" to track)' is shown, followed by the terminal prompt 'macbook-pro-3:applied-stats monicaalexander\$' and a cursor.

```
macbook-pro-3:~ monicaalexander$ cd src/applied-stats/
macbook-pro-3:applied-stats monicaalexander$ git status
On branch master
Your branch is up to date with 'origin/master'.

Untracked files:
  (use "git add <file>..." to include in what will be committed)

        wk1.txt

nothing added to commit but untracked files present (use "git add" to track)
macbook-pro-3:applied-stats monicaalexander$
```

Steps to add work, commit and push

- ▶ `git add wk1.txt` adds the file to staging area
- ▶ do a `git status` again to see what's going on

```
[macbook-pro-3:applied-stats monicaalexander$ git add wk1.txt ]
```

```
[macbook-pro-3:applied-stats monicaalexander$ git status ]
```

On branch master

Your branch is up to date with 'origin/master'.

Changes to be committed:

(use "git reset HEAD <file>..." to unstage)

new file: wk1.txt

```
macbook-pro-3:applied-stats monicaalexander$ █
```


Steps to add work, commit and push

- ▶ `git commit -m "adding wk1 file"` commits the file (`-m` gives the option of a message)
- ▶ do a `git status` again to see what's going on

```
[macbook-pro-3:applied-stats monicaalexander$ git commit -m "adding wk1 file"
[master 96c178f] adding wk1 file
 1 file changed, 1 insertion(+)
 create mode 100644 wk1.txt
[macbook-pro-3:applied-stats monicaalexander$ git status
On branch master
Your branch is ahead of 'origin/master' by 1 commit.
  (use "git push" to publish your local commits)

nothing to commit, working tree clean
macbook-pro-3:applied-stats monicaalexander$ █
```

Steps to add work, commit and push

Push to Github

► git push

```
macbook-pro-3:applied-stats monicaalexander$ git push
Enumerating objects: 4, done.
Counting objects: 100% (4/4), done.
Delta compression using up to 4 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (3/3), 335 bytes | 335.00 KiB/s, done.
Total 3 (delta 0), reused 0 (delta 0)
To https://github.com/MJAlexander/applied-stats
   8488fdc..96c178f  master -> master
macbook-pro-3:applied-stats monicaalexander$ █
```

- Now check your repo on GitHub to check the new work is there

Part of this week's lab assessment

- ▶ Make a repo and add me as a collaborator
- ▶ Add a text file with your name and your favorite type of food
- ▶ Push changes to GitHub

R component

Lab

- ▶ Each week, the lab will be contained in an R Markdown file in the `rmd` folder of the class repo
- ▶ You should go through it and then attempt the lab exercises
- ▶ Answers to the lab exercises should be saved in your own R Markdown file and pushed to your git repo (by Friday 9am)

Important functions to learn

- ▶ `read_*` (csv, table, rds)
- ▶ The pipe `%>%`
- ▶ `select` (columns)
- ▶ `filter` (rows)
- ▶ `arrange`
- ▶ `mutate`
- ▶ `summarize`
- ▶ `group_by`
- ▶ `pivot_*`

ggplot basics (more next week)

- ▶ `ggplot`
- ▶ `aes`
- ▶ `geom_*`
- ▶ `labs` or `ggtitle` `ylab` etc
- ▶ `scale_color_*` and `scale_fill_*`

This week's lab assessment: summary

- ▶ git exercise described above
- ▶ lab exercises at end of lab