
# EMBEDDING DATA ANALYSIS INTO THE UNDERGRADUATE ACTUARIAL SCIENCE CURRICULUM

Aloke Phatak, School of Electrical Engineering, Computing and Mathematical Sciences

# Background and objectives

**Background**

- Engineer-turned statistician
- ~3 years at Curtin after ~20 years at CSIRO (DMS → Data61)
- Co-ordinator of undergraduate Data Science, UG & PG curriculum development, teaching, consulting, research, and developing CPD (including MOOCs)

**Objectives**

- Speak about recent initiatives to ‘modernize’ Curtin AS curriculum by incorporating computational tools/thinking and data analysis
- Learn about similar initiatives elsewhere
- Fill in gaps in my knowledge about how CT and DA can be used in ‘traditional’ actuarial science units

# Curtin/WA context

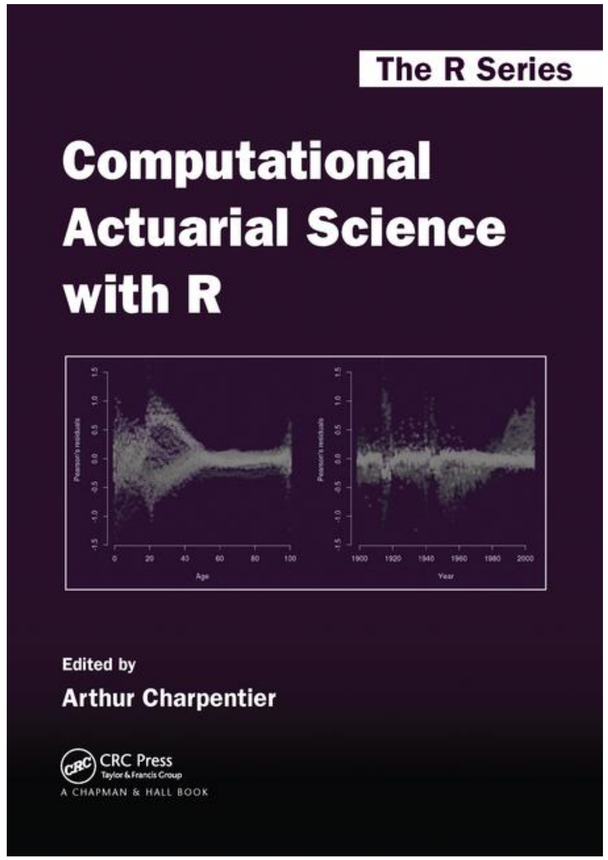
- Curtin AS program is the only AS course offered in WA
- Sits within a School that also offers degree programs in Financial Mathematics, Industrial and Applied Mathematics, and Data Science
- First-year intake of ~55 students; 10–15 articulation students join in second year; total cohort of ~200 students
- Because of the nature of the WA market, very few students who stay in Perth become practicing actuaries

# Embedding DA & CT: The vision

- Introduce computing in virtually *every* unit; emphasize reproducible analyses
  - Provide students with the means to calculate, to visualize
  - Independent of environment; we use R (and Excel)
- Emphasize computing as a means of obtaining answers to real-world problems, **but also as an aid to understanding theory**
  - Coding something helps you understand it
- Insist on the importance of modelling, data visualization and communication of results
- **Integrate case studies, hands-on data analysis, and project work into as many units as possible**
- Formalize obtaining data from different sources and then exploring and cleaning as subjects in their own right

# Why incorporate DA & CT?

- As a means of:
  - Providing our students with skills that they will need in the contemporary workplace, actuarial *and* non-actuarial
  - Revitalizing the AS curriculum
    - **Value judgments**: students taught plenty of theory, little practice; little or no computing skills; CT notes are dry, could do with much more contextual knowledge
  - Ensuring that many units can indeed do double- and triple-duty
- Happy coincidence: also consistent with broader changes occurring in AS education



# Integrating DA & CT

## Which units?

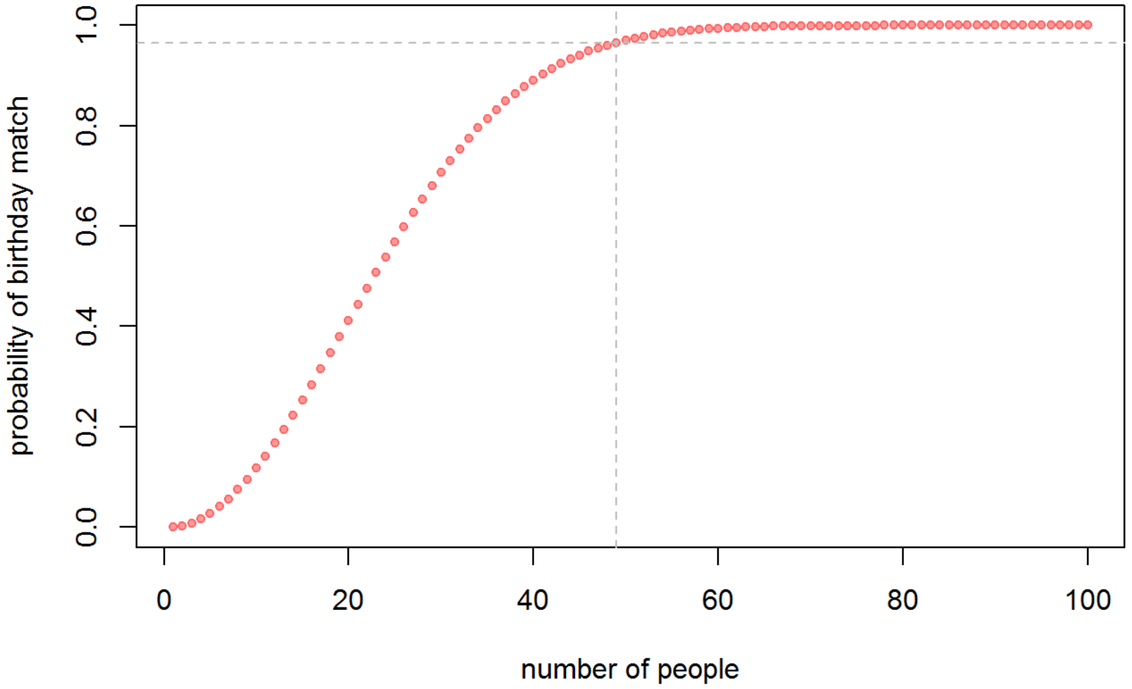
- Introductory statistics and probability
- Mathematical statistics
- Regression analysis
- Survival analysis
- Stochastic processes
- Statistical modelling (GLM, time series analysis, simulation)

## How?

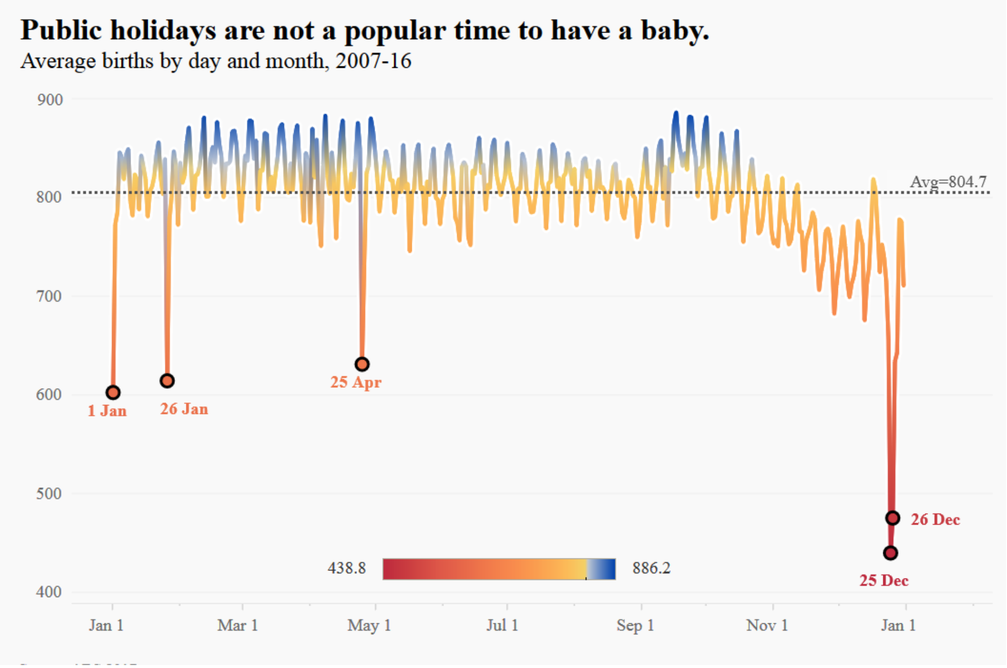
- Introduce R/RStudio/R Markdown (and Excel) early in the curriculum
- Embed code in lecture slides
- To different extents:
  - Reallocate time from lectures to existing or new computer laboratory
  - Include computer labs/tutorials in which a sheet of paper is replaced by an R Markdown file
  - Use computational tools and visualization to illustrate theory
  - Introduce projects that have a strong data analytic component
  - Incorporate computer-based tests
- Encourage students to use DA & CT in other units

# Example: the birthday problem

## Simple formulation




## More complex scenario



# Example: simple linear regression

S2_2018_STAT1000_Workshop_5.Rmd

Knit | Insert | Run

```
---
title: "STAT1000"
author: "Regression and Nonparametric Inference"
date: "**Workshop Week 5**"
output:
  html_document:
    highlight: haddock
    theme: flatly
  html_notebook:
    highlight: haddock
    theme: readable
---


```{r echo=FALSE}
knitr::opts_chunk$set(prompt=FALSE, comment=NA, tidy=TRUE, error=TRUE, warning=FALSE, message=FALSE)
```

```{r echo=FALSE}
htmltools::img(src = "https://global.curtin.edu.au/responsive-assets/img/logo-curtin-university.png",
               alt = 'logo',
               style = 'position:absolute; top:0; right:0; padding:10px;')
```


```{r}
# Install this library first if it doesn't already exist on your computer
library(rgl)
knitr::knit_hooks$set(webgl = hook_webgl, rgl=hook_rgl)
```


```{r}
# Load the workshop data here
load("S2_2018_STAT1000_Workshop_5.RData")

```
```

Curtin University

# STAT1000

*Regression and Nonparametric Inference*

*Workshop Week 5*

```
# Install this library first if it doesn't already exist on your computer
library(rgl)
knitr::knit_hooks$set(webgl = hook_webgl, rgl=hook_rgl)
```

```
# Load the workshop data here

print(load("S2_2018_STAT1000_Workshop_5.RData"))
```

```
[1] "SimulData" "wm1"
```

## Question 1

In this question, we'll use a simulated dataset (`SimulData`) to demonstrate some aspects of simple least squares (LS) regression. The model is written as

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where $(x_i, y_i)$ represent the $i$th values of the response and explanatory variables, respectively; $\beta_0$ and $\beta_1$ are the parameters we wish to estimate; and $\epsilon_i \sim N(0, \sigma^2)$ represents the iid (independently and identically) distributed random error.

Recall that in LS, we minimize the residual sum of squares ($RSS$), shown below, in order to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameter:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2$$

Doing so yields the estimates as follows:

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}; \quad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

# Project: linear regression

## STAT1000 Regression and Nonparametric Inference

**Take-Home Project**

Semester 2, 2018

### 1 Objective

This take-home project is one of four assessments (along with Tests 1 & 2 and the final exam) in this unit. It is worth 20% of the overall mark and you need to obtain 8/20 or more in order to pass the unit. **Unless you can make a convincing argument for working alone, this project is to be done in groups of two people**.

The objective of the project is to analyze one of two datasets using the methods and techniques you have learned in this unit with a view to coming up with the best predictive model you can for the response variable. The datasets, which are described below, are typical of contemporary, complex data: there are many potential predictors, many of which are highly correlated with one another; there *may* be outliers that have to be removed; and the response *may* have to be transformed before being used in a linear regression model.

Up to now, in both lectures and tutorials, we have analyzed data and fitted regression models as if the steps to do so were clear, well-laid out, and led invariably to a 'correct' answer. Reality, however, is messier. There is not a linear path from problem and data to solution, and I hope you will get some sense of that as you do this project.

- Guided analysis of one of a handful of datasets, e.g., house price data from Ames, IA
  - Some ‘messiness’ left in
- EDA, linear modelling using contemporary variable selection methods, diagnostic checking, predictive evaluation
- Brief (3–5 page), readable write-up of results
- Code to be included as appendix

# Example: accept-reject methods

## STAT3001

***Statistical Modelling***

*Solution: Workshop 7*

- Accept-Reject Methods
- Exercise
- Notes

## Accept-Reject Methods

There are many distributions for which the inverse transform method is less useful because the quantile function of the target density $f(x)$ is not easily found. In addition, we may want a method that relies only on the PDF, not the CDF, which may be difficult to compute. In such cases, we can use *indirect* methods where we generate a candidate random variate and only accept it as a draw from $f(x)$ if it passes a 'test'.

Such methods are known as *accept-reject* methods, and they only require us to know the functional form of the density $f(x)$ up to a multiplicative constant. The basic idea is simple: We use a simpler (to simulate) density $g$, called the *candidate* density, to generate random variates for which the simulation is actually done. There are, however, some constraints on this candidate density $g$:

1. $f$ and $g$ have compatible supports
2. There is a constant $C$ such that $f(x)/g(x) \leq C$ for all $x$.

In many didactic examples, we use the uniform density as the candidate density, though in real situations, this is not necessarily going to be the best or most efficient choice. If we do, however, use the uniform distribution as the candidate density, the accept-reject method amounts to the following paint-ball analogy for a target density that has finite support (thanks to Felix Chan of CBS for coming up with this):

1. Imagine that you have plotted the PDF of the target density $f(x)$ on a piece of paper and that you have drawn a box around it, i.e., the box will be as wide as the support of the target density, and it will be at least as high as the mode of the target density;
2. Hang up the piece of paper;
3. Using your paint-ball gun, randomly (in 2 dimensions) shoot at the picture $N$ times;
4. Identify all the shots that are under the PDF;
5. Use the $x$-coordinates of those shots as the random draws from $f(x)$.

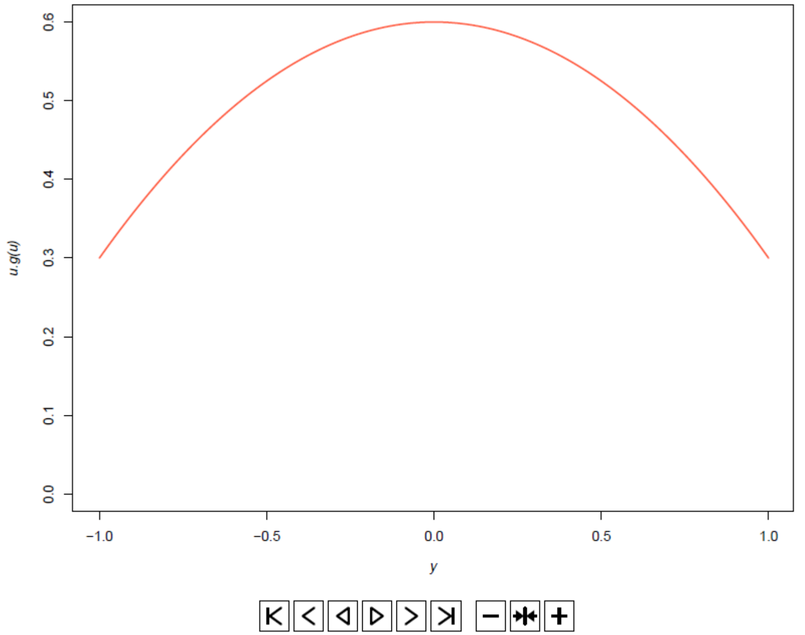



Figure 1: Accepted/rejected draws from a uniform distribution for simulating from density $f(x) = 0.6 - 0.3x^2$.

# Example: computer-based test

# STAT3001

*Statistical Modelling*

*Solution: Test 2*

## Instructions

1. Save this file to `I:\STAT3001` and rename it by adding your student ID to the beginning of the file name, i.e., `19141918_S1_2018_STAT3001_Test_2.Rmd`;
2. Carry out all of your work in this Rmarkdown file;
3. **Save your work frequently!**
4. Make sure you have downloaded the file `S1_2018_STAT3001_Test_2.RData`;
5. When you are finished, 'knit' this file to a Word document, and upload the resulting 'knitted' .docx file to the Assessments section from which you downloaded the files.
6. Total number of marks: **20**

**You only need to do 2 questions.**

```
################################################################################
# Load this library first; if you need to, install it first on your computer#
################################################################################

library(astsa)
```

# How’s it going so far?

**Barriers**

- It’s a lot of work to create correct, coherent, and compelling materials that are completely integrated with lectures
- Digital (ill)literacy: local and overseas students
- Students forget from one semester to the next, so if there are any temporal gaps, we have to start over again
- Broad integration of these ideas – even were they to be relevant in all units – depends on teaching staff becoming familiar/at ease with software environments

# How’s it going so far?

**Response**

- Generally positive: actuarial students recognize the ‘buzz’ around data science and are keen to get in on it
  - Some students also recognize the pedagogical value of computational, data analytic, and visualization skills and capabilities
- The view from out there: many of our second-year students have obtained summer internships in data analysis roles; third-year students have obtained graduate data analysis positions

# Future plans and needs

- Explore the extent to which DA and CT can be integrated into most (all?) units so that students have a seamless education
- Unified set of materials (GitHub-based?) available to all
- Needs: compelling examples from all areas of actuarial science!


# Questions?