MIE237 - Statistics

Neil Montgomery

2016-01-05

preliminaries

Administrative Details

date format	ISO8601
no class	2016-02-12 (Friday before reading week)
instructor	Neil Montgomery
email	neilm@mie.utoronto.ca
office	BA8132
website	portal (announcements, grades, etc.)
github	https://github.com/mie237 (lecture material, code, etc.)
office hours	any time (except M5-6 and R3-5)

Grading

what	when	type	weight
midterm 1	2016-02-09	B type	20%
assignment 1	2016-02-26		5%
midterm 2	2016-03-22	B type	20%
assignment 2	2016-04-08		5%
exam	TBA	B type	50%

Tutorial time: one hour group meeting plus one TA office hour (depending on demand) starting next week Labs start next week

R Software

- R is the name of a programming language. R code needs to be interpreted by the computer. The name of the program that does this is also called R which you get at https://www.r-project.org
- So I suppose R $\neq$ R. This is not unusual (c.f. Matlab)
- You'll also need a nice development environment such as RStudio which you get at https://www.rstudio.com
- Suggestion: a $\LaTeX$ installation (this lets you make nice PDF versions of assignments).
- "Hold on a second don't we just do our assignments in Word, typing numbers right into the text, copying and pasting results and graphs into the document, and then if anything changes go back and edit the text and change the graphs and hope I remembered to change everything and hope Word leaves all the pictures where I want?"

R markdown

- Sadly that is a typical workflow in industry. I will expose you to a better way of doing things. Then you will go into industry and convince everyone to do things better all because of me.
- A better way is called *markdown* in which text and computer results are interwoven using a simple format and the final document is *rendered* into something suitable (PDF, Word, HTML)
- Lecture slides are written in markdown right in RStudio and will have lots of R code embedded.
- You will complete your assignments in markdown.

Suggested install sequence and notes

Works on any platform. All free software.

1. (Optional but nice) LaTeX https://latex-project.org
2. R https://r-project.org
3. RStudio https://www.rstudio.com/products/rstudio/download

You'll normally run R via RStudio. You won't usually run R directly itself.

Note that RStudio is not (yet?) on the ECF machines and might not be for 3-6 weeks or so, if ever. It apparently a huge pain to get software on ECF machines.

R, RStudio, and markdown - What to learn and when?

R

- It's probably a good idea to learn a few of the very basics about R, and you are welcome to become as expert as you like, but in this course we'll only need a tiny subset of its full capabilitities.
- The first few links here: https://support.rstudio.com/hc/en-us/articles/201141096-Getting-Started-with-R seem OK.
- Slides and other lecture material will contain lots of examples.

RStudio and markdown

- markdown is really easy
- Some nice thoughts are here: https://stat545-ubc.github.io/block002_hello-r-workspace-wd-project.html which gives some nice thoughts about RStudio after the little R intro at the top.
- Always use RStudio projects
- Every new project in a clean directory
- Code is real. Results are not. Never save the workspace.
- In particular, I suggest right away that in RStudio you do the following with Tools->Global Options:
    - Uncheck "Restore .RData into workspace at startup"
    - Change "Save workspace .RData on exit:" to "Never"

review

Guess what's going on here

```
rnorm(n = 1, mean= 20, sd = 5)
```

```
## [1] 15.60259
```

Concepts from Probability

A "random variable" is a *function*

Fundemantal property of a random variable: *distribution*

Its *distribution* is (informally): the possible outcomes and their probabilities.

Calculus poison

Functions are graphical things. The point is to draw pictures, find tangents, compute areas under curves etc. And they are always called $f(x)$, or $g(x)$ is necessary or $f_1(x), f_2(x), \ldots$ in desperate cases.

Nerds might be totally into things like $f : \mathbb{R} \longrightarrow \mathbb{R}$ and just calling functions by their actual names f and so on.

We're all nerds now

Random variables have a different naming convention.

$$X, Y, X_1, \ldots, X_n$$

with Z reserved for the "standard normal distribution", i.e.

$$Z \sim N(0, 1)$$

Calculus tools (the graphical things) are not usually directly applied to random variables. They are applied to the things that describe distributions (cdf, pdf, pmf) and result in probabilities and expected values etc.

Key points about expected value

If X has density f and g is some other function, generally:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

Example with $g(x) = x$:

$$E(g(X)) = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

Common names: "mean", "average", "population mean", μ, or just $E(X)$. We'll treat the mean of a random variable as a *fixed constant* although you might not know what the number actually is.

Another example

$$\mathrm{Var}(X) = E((X - E(X))^2) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx$$

Common names: variance, σ^2, Var(X). Again, we'll treat the variance of a random variable as a , although you might not know what the number actually is.

And $\mathrm{SD}(X) = \sigma = \sqrt{\mathrm{Var}(X)}$ is the standard deviation.

Key properties of mean and variance

X and Y are random variables and a and b are constants.

$$E(aX + b) = aE(X) + b$$
$$E(a) = a$$
$$E(X + Y) = E(X) + E(Y)$$
$$\mathrm{Var}(X) = \mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

But:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

where $\mathrm{Cov}(X, Y)$ is the *covariance* of X and Y.

Covariance is 0 when X and Y are *uncorrelated* (no linear relationship). In particular, random variables that are *independent* are also uncorrelated.