11. Connection between Module 1 (Intro to R) and Module 2 (Intro to Probability)

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

We did:

- Discrete and Continuous Random Variables and their distributions.
- PMF/PDF, plotting pmf/pdf, cdf, probability calculations by hand and using R for Discrete and Continuous uniform, Binomial and Normal distributions.

binomial distribution $\mathrm{Binom}(\text{size}, \text{prob})$

- `dbinom(x, size, prob)`
- `pbinom(q, size, prob)`
- `rbinom(n, size, prob)`
- `qbinom(p, size, prob)`

uniform distribution $\mathrm{Unif}(\text{min}, \text{max})$

- `dunif(x, min, max)`
- `punif(q, min, max)`
- `runif(n, min, max)`
- `qunif(p, min, max)`

normal distribution $N(\text{mean}, \text{sd})$

- `dnorm(x, mean, sd)`
- `pnorm(q, mean, sd)`
- `rnorm(n, mean, sd)`
- `qnorm(p, mean, sd)`

Next we will see. . .

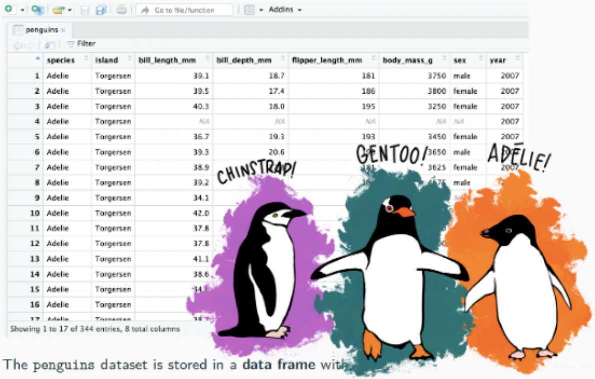
- Connection modules and other courses in PSTAT department
- Some extras
    - Creating your own .Rmd file
    - Environment
    - Working directory

Overall Connection

Module 1(Intro to R) and Module 2(Introduction to Probability)

Module 1(Intro to R) -> PStat 10

Sample



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	[illegible]	3650	male	2007
7	Adelie	Torgersen	38.9	[illegible]	181	3625	female	2007
8	Adelie	Torgersen	39.2		[illegible]	[illegible]	male	
9	Adelie	Torgersen	34.1				[illegible]	
10	Adelie	Torgersen	42.0					
11	Adelie	Torgersen	37.8					
12	Adelie	Torgersen	37.8					
13	Adelie	Torgersen	41.1					
14	Adelie	Torgersen	38.6					
15	Adelie	Torgersen	34.6					
16	Adelie	Torgersen						
17	Adelie	Torgersen	38.7					

Showing 1 to 17 of 344 entries, 8 total columns

The penguins dataset is stored in a **data frame** with

- **344 observations/samples/cases/subjects** (rows)
  - each case represents a penguin
- **8 variables** (columns)
  - species, island, bill_length_mm, bill_depth_mm etc
  - each corresponds to some measurement of the penguin

Where do random varaibles come from?

Recall, for data(a sample) we said a variable can be

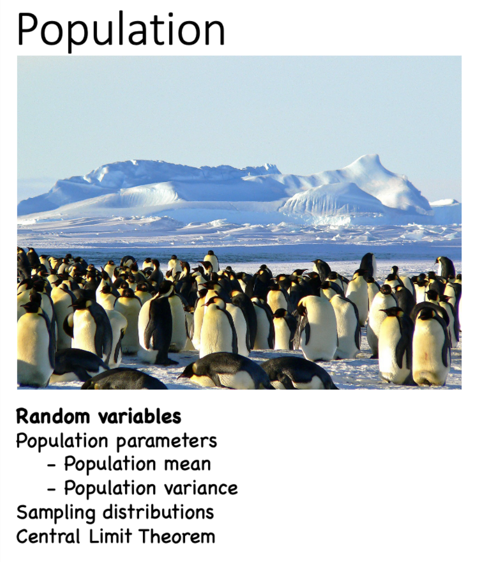
- Numerical - discrete or continuous
- Categorical - ordinal or nominal

Random variables encode all possible data we may ever see!

Why learn probability?

- Used plots and summary statistics to explore distributions and relationships of different variables in our (observed) data/sample.
- Now, Statistics aims to generalize these findings to the entire population.

Module 2(Introduction to Probability) -> PStat 120A -> Pstat 120B



Population

Random variables

Population parameters

- Population mean
- Population variance

Sampling distributions

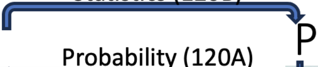
Central Limit Theorem

Statistics: Generalizing from a sample to the population

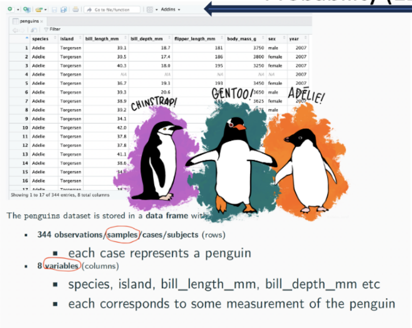
- There's always some uncertainty about the true distributions and relationships in the population
- Probability is the mathematical tool used to measure and express this uncertainty. (PSTAT 120A)
- We should clearly specify the extent of our uncertainty. (PSTAT 120B)

Connection between modules & other courses in PSTAT department

Sample



Population



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	[illegible]	3650	male	[illegible]
7	Adelie	Torgersen	38.9	[illegible]	181	3625	female	2007
8	Adelie	Torgersen	39.2	[illegible]	[illegible]	4675	male	[illegible]
9	Adelie	Torgersen	34.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	Adelie	Torgersen	42.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	Adelie	Torgersen	37.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	Adelie	Torgersen	37.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	Adelie	Torgersen	41.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	Adelie	Torgersen	38.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15	Adelie	Torgersen	34.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
16	Adelie	Torgersen	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
17	Adelie	Torgersen	38.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Showing 1 to 17 of 344 entries, 8 total columns

The penguins dataset is stored in a **data frame** with

- **344 observations/samples/cases/subjects** (rows)
  - each case represents a penguin
- **8 variables** (columns)
  - species, island, bill_length_mm, bill_depth_mm etc
  - each corresponds to some measurement of the penguin

Variables
Summary statistics
- sample mean
- sample variance

Visualizations

Random variables
Population parameters
- Population mean
- Population variance

Sampling distributions
Central Limit Theorem

Courses that build on probability fundamentals

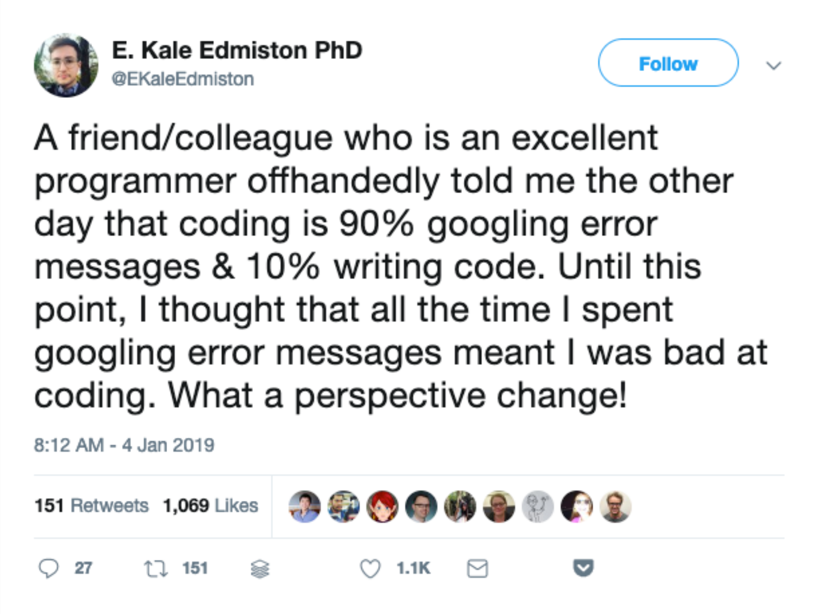
- Measure and express uncertainty in going from sample to population (PSTAT 120B)
- Hypothesis testing (PSTAT 120B)
- Bayesian statistics (PSTAT 115)
- Linear Regression (PSTAT 126)
- Statistical Machine Learning (PSTAT 131)
- Computational statistics (PSTAT 194CS)
  - Monte Carlo methods, Social Network Analysis, AI

Dangers

- Theory not used correctly

Some extras [OPTIONAL]

Learning Programming is HARD!



E. Kale Edmiston PhD
@EKaleEdmiston

A friend/colleague who is an excellent programmer offhandedly told me the other day that coding is 90% googling error messages & 10% writing code. Until this point, I thought that all the time I spent googling error messages meant I was bad at coding. What a perspective change!

8:12 AM - 4 Jan 2019

151 Retweets 1,069 Likes

Debugging

Debugging is the process of getting rid of errors in your code.

3 types of errors:

1. Syntax Errors: code does not follow `R`'s rules
2. Runtime Errors: errors that occur during knitting
3. Logic Errors: code runs but produces unexpected results.

Know Your RStudio Environment

There are a *lot* of keyboard shortcuts in RStudio. To view all the options, you must engage the keyboard shortcut that rules them all:

- Windows: `Alt` + `Shift` + `K`
- macOS: `Option` + `Shift` + `K`

Some favorites

1. Autocomplete command.
    - Both: `Tab`
2. Run the current line, selection from the editor.
    - Windows: `Ctrl` + `Enter`
    - macOS: `Cmd` + `Enter`
3. Run the current code chunk from the editor.
    - Windows: `Ctrl` + `Shift` + `Enter`
    - macOS: `Cmd` + `Shift` + `Enter`

Downloading R

Go to: https://cran.r-project.org/

Chose from:

- Download R for (Mac) OS X
- Download R for Windows

Mac users choose Mac download

Windows users choose Windowns download

Downloading RStudio

1. Download and install R first.
2. Go to https://rstudio.com/products/rstudio/download/

We did

- Connection of modules and other courses in PSTAT department
- Some extras
    - Creating your own .Rmd file
    - Environment
    - Working directory