Introduction to julia for Statistics and Data Science

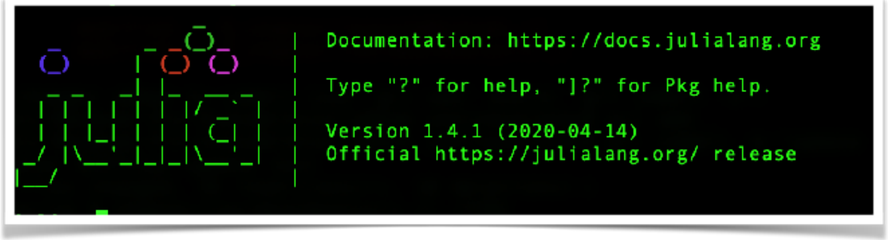
A workshop organized by the Statistical Society of Australia (VIC branch)

Fri, July 24 9:00am - 12:30pm
Mon, July 27, 9:00am - 12:30pm
2020 (AEST) via Zoom
Registration: https://www.statsoc.org.au/event-3888909

Presented by Yoni Nazarathy - @ynazarathy

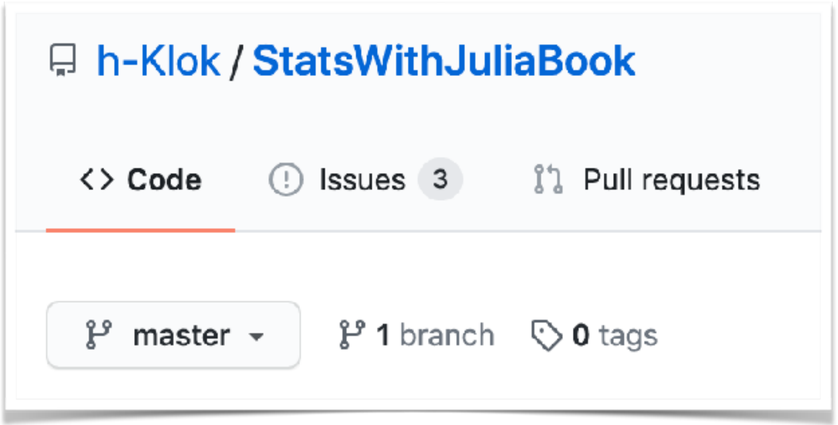
Before the workshop

- **If you don't have Julia installed, get it working on your computer**
- Recommended to use (all 3):
  - Julia REPL (read-eval-print loop)
  - Julia Pro: https://juliacomputing.com/products/juliapro (Juno)
  - Jupyter notebooks via IJulia: https://github.com/JuliaLang/IJulia.jl
- Recommended to install packages used by examples in this workshop
    - You can follow the installation instructions in this video for Julia Pro: https://www.youtube.com/watch?v=ei-xnVid3QY
    - If you use command line - add Julia to the PATH: https://julialang.org/downloads/platform/
    - You an follow the installation instructions for IJulia: https://www.youtube.com/watch?v=oyx8M1yoboY (
      —-> However you need `using Pkg` before `Pkg.add("IJulia")`
    - Run this file to install all needed packages: https://github.com/h-Klok/StatsWithJuliaBook/blob/master/rawInit.jl
- Browse tool documentation:
  - Julia REPL docs: https://docs.julialang.org/en/v1/stdlib/REPL/
  - Juno docs: http://docs.junolab.org/stable/
  - Jupyter notebook tutorial (one of many available): https://www.dataquest.io/blog/jupyter-notebook-tutorial/
- Consider registering (free) to JuliaCon 2020: https://juliacon.org/2020/



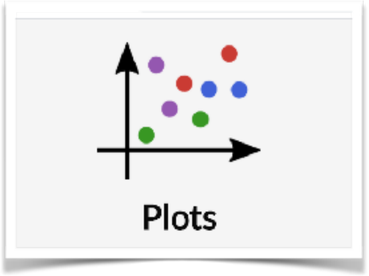
Resources

- Book DRAFT (joint work with Hayden Klok):
  Statistics with Julia: Fundamentals for Data Science, Machine Learning and Artificial Intelligence. https://statisticswithjulia.org/
- Workshop uses some code examples from book's source code GitHub:
  https://github.com/h-Klok/StatsWithJuliaBook
- Full Julia documentation: https://docs.julialang.org/
- Julia discourse: https://discourse.julialang.org/
- Julia slack: https://slackinvite.julialang.org/
- Juliacon 2020: https://juliacon.org/2020/



Resources continued: Packages

- `Plots.jl` docs: http://docs.juliaplots.org/latest/
- `DataFrames.jl` docs: https://juliadata.github.io/DataFrames.jl/stable/
- `GLM.jl` docs: https://juliastats.org/GLM.jl/stable/manual/
- `Distributions.jl` docs: https://juliastats.org/Distributions.jl/stable/
- `StatsBase.jl` docs: https://juliastats.org/StatsBase.jl/stable/
- `StatsModels.jl` docs: https://juliastats.org/StatsModels.jl/stable/
- `Flux.jl` docs: https://fluxml.ai/Flux.jl/stable/
- `DifferentialEquations.jl` docs: https://docs.sciml.ai/stable/
- `RCall.jl` docs: http://juliainterop.github.io/RCall.jl/stable/
- `HypothesisTests.jl` docs: https://juliastats.org/HypothesisTests.jl/stable/
- `MultivariateStats.jl` docs: https://multivariatestatsjl.readthedocs.io/en/stable/ind
- `Lasso.jl` docs: https://juliastats.org/Lasso.jl/stable/
- `TimeSeries.jl` docs: https://juliastats.org/TimeSeries.jl/stable/
- `Clustering.jl` docs: https://juliastats.org/Clustering.jl/stable/
- Many more julia packages: https://juliahub.com/ui/Home



Resources continued: Additional tutorials

- Julia express by Bogumił Kaminski:
  https://github.com/bkamins/The-Julia-Express
- From zero to Julia by Aurelio Amerio:
  https://techytok.com/from-zero-to-julia/
- An introductory book (free on-line), Think Julia: How to Think Like a Computer Scientist, by Ben Lauwens and Allen Downey:
  https://benlauwens.github.io/ThinkJulia.jl/latest/book.html
- Many more learning resources:
  https://julialang.org/learning/

Workshop Structure

- 10 acts
- 2 days
- 5 acts/day
- 40 min/act
- act =
  25min Yoni +
  15min self-work/break (Yoni available for support)
- Main content of acts based on “Statistics with Julia” source code
- Continue to watch some JuliaCon talks.... (next slide)

Recommendations for (pre-)JuliaCon Workshops

- Many great talks but these workshops are very relevant in particular. Times below are stated in AEST (times on website are UTC):
  - Midnight of July 24-25: Learn Julia via epidemic modelling (David Sanders): https://pretalx.com/juliacon2020/talk/LSNEWV/
  - Midnight of July 26-27: Doing scientific machine learning (SciML) with Julia (Chris Rackauckas): https://pretalx.com/juliacon2020/talk/C9FGPP/
  - Midnight of July 27-28: A deep dive into Dataframes.jl indexing (Bogumil Kamniski): https://pretalx.com/juliacon2020/schedule/#2020-07-27
  - Midnight of July 28-29: MLJ - a machine learning toolbox for Julia (Thibaut Lienart, Anthony Blaom): https://pretalx.com/juliacon2020/talk/DMHZCC/

Recommendations for Talks during JuliaCon

- Particularly relevant talks during the actual conference (Wed July 29 - Friday (early Saturday) July 31):
  - https://pretalx.com/juliacon2020/talk/RRNYRW/ (Flux)
  - https://pretalx.com/juliacon2020/talk/R7EPKS/ (Probabilistic programming)
  - https://pretalx.com/juliacon2020/talk/MASLPF/ (The state of Julia by language co-creators Stefan Karpinski and Jeff Bezanson)
  - https://pretalx.com/juliacon2020/talk/QHRQVF/ (Metaprogramming)
  - https://pretalx.com/juliacon2020/talk/ELQ8A8/ (StatsModel.jl - the @formula macro)
  - https://pretalx.com/juliacon2020/talk/NPWSWB/ (teaching Data Science)
  - https://pretalx.com/juliacon2020/talk/HBTFT7/ (VS code)
  - https://pretalx.com/juliacon2020/talk/7XARPV/ (Makie.jl)
  - https://pretalx.com/juliacon2020/talk/BMNWLJ/ (StatsMakie.jl)
  - https://pretalx.com/juliacon2020/talk/Z8WWNV/ (NLP)
  - https://pretalx.com/juliacon2020/talk/CA3SET/ (Julia and Data)

About your presenter

- Associate Professor at The University of Queensland (always looking for good PhD candidates).
- Queueing Theory, Scheduling, Control Theory, Statistics, and Data Science
- Aerospace industry prior to academia.
- Israel, US, Netherlands, Victoria, Queensland.
- Safe Blues: https://safeblues.org/
- One on Epsilon: https://oneonepsilon.com/

Acts - Day 1

Act 1: Hello World

Act 2: More Basics

Act 3: Distributions

Act 4: More on distributions

Act 5: Dataframes, linear algebra, and some GLM

Acts - Day 2

Act 6: Julia overview - a bit more in depth

Act 7: More on distributions and fitting

Act 8: Small numerical experiments

Act 9: Fitting models

Act 10: Machine learning and wrap up

One big Jupiter Notebook

- Get the notebook from:
  https://statisticswithjulia.org/presentations/SSAJuliaJuly2020Nazarathy.ipynb
  You'll need it placed in the right way to reach the data files.
- The PDF for the notebook is here:
  https://statisticswithjulia.org/presentations/SSAJuliaJuly2020Nazarathy-Jupyter-Notebook.pdf

Act 1: Hello World

Some terms from the Julia world

- Types
- Multiple dispatch
- Garbage collection
- Just in time compilation
- Metaprogramming
- Unicode
- LLVM
- Package Manager


```
julia> @code_llvm +(1,1)

define i64 @"julia_+_130862"(i64, i64) {
top:
    %2 = add i64 %1, %0
    ret i64 %2
}
```

Some more

- Subtypes, supertype, type hierarchy
- Macros
- Broadcasting
- Arrays = Vectors
- Tuples
- Dictionaries
- Short circuit evaluation
- Data frames
- Packages with package `Base`. E.g. `LinearAlgebra`, `Statistics`, `Random`, `Dates`

Act 1 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/helloWorld.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/bubbleSort.jl

Act 1 - Suggested Self Work

A) Create a script that approximates $\frac{6}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}$

B) Modify the bubble sort so that it sorts backwards

C) Search the web/docs for a way to remove all white space from a string in Julia and try it

Act 2: More Basics

Act 2: More Basics

Data Science concepts used in this act...

- Markov Chains
- Linear Algebra
- Pseudorandom number generation
- Histogram
- The R Language
- The Python Language

Act 2 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/multiWayMarkovChainStationary.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/plotSimple.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/hailstoneHistogram.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/usingR.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/usingPython.jl

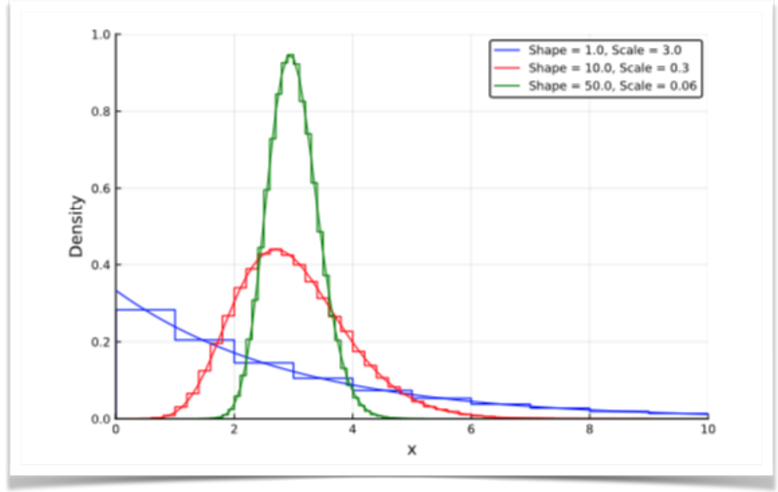
Act 2 - Suggested Self Work

A) Modify the Markov chain example to work on a chain with 4 states.

B) Plot cos(x) and sin(x) with nicely formatted labels.

C) Plot a histogram of the number of primes in each block of 100 numbers, ranging from 0 to 10,000.

Act 3: Distributions



Act 3: Distributions

Data Science concepts used in this act...

- Probability Distributions
- Mean, Variance, PDF, CDF, Quantiles...
- Families of Discrete Distributions
- Families of Continuous Distributions

Act 3 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/discreteContinuous.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/meanIntegration.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/triangularDist.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/distributionDescriptors.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/basicDistRand.jl

6) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/discreteDists.jl

7) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/continuousDists.jl

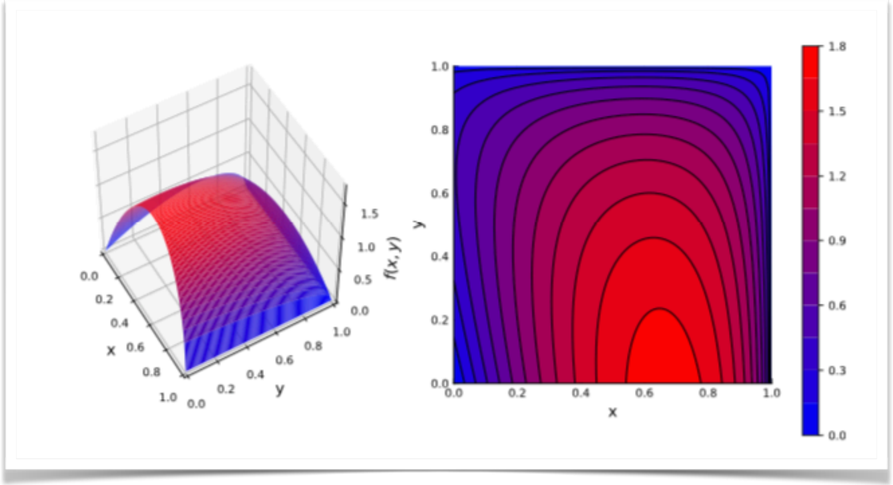
Act 3 - Suggested Self Work

A) Plot the PMF of a Poisson distribution

B) Plot the PDF of a Gamma distribution

C) Use random numbers to estimate (via Monte Carlo) the mean of a Beta distribution and compare it to the analytic value

Act 4: More on distributions



Act 4: More on distributions

Data Science concepts used in this act...

- The normal distribution
- The law of large numbers
- The Cauchy distribution
- Covariance matrices
- Multivariate normal distributions

Act 4 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/normalCalculus.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/cauchyCMA.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/multiRVgeneration.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/meanVectCovMatrixSummary.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/bivariateNormal.jl

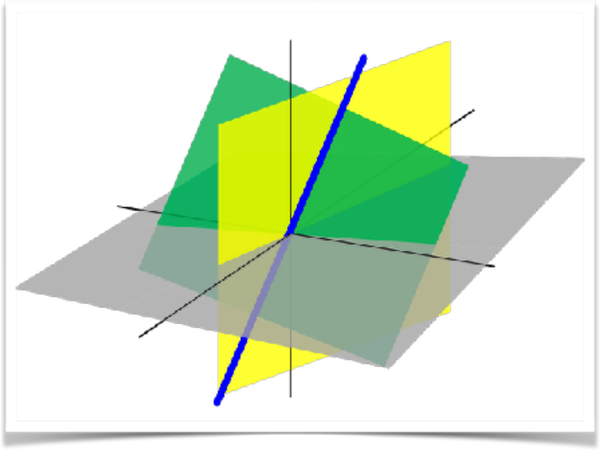
Act 4 - Suggested Self Work

A) Reproduce the example using the Cauchy distribution and see what happens if you truncate the distribution

B) Reproduce the example of multivariate random variable generation

C) Reproduce (and modify) the example plotting a bivariate normally distribution

Act 5: Dataframes, linear algebra, and some GLM



Act 5: Dataframes, linear algebra, and some GLM

Data Science concepts used in this act...

- Least Squares
- Many ways of computing least squares estimates including the pseudo inverse, QR factorization, SVD, and gradient descent
- Simple linear regression
- The distribution of the regression estimators

Act 5 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/dataframeInspection.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/dataframeReferencing.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/leastSqMethods.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/usingGLM.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/distRegEstimators.jl

Act 5 - Suggested Self Work

A) Modify the dataframe see the effect on the regression.

B) Modify the parameters for gradient descent to see if/how converges

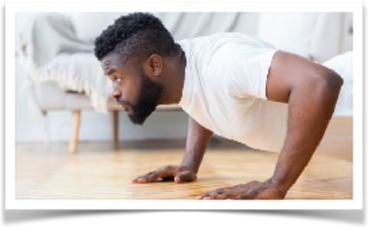
C) Modify/improve the graphics for the distribution of the regression estimators.

Day 2

Act 6: Julia overview - a bit more in depth

```
abstract type Number end
abstract type Real     <: Number end
abstract type AbstractFloat <: Real end
abstract type Integer  <: Real end
abstract type Signed   <: Integer end
abstract type Unsigned <: Integer end
```

Morning Vocabulary Workout



@which	typeof()	methods()	im	lastindex()	length(), size()	eachindex()	Linear	min() , max()	copy()	broadcast() .	NaN
begin, end	exit()	while	for	break	continue	enumerate()	::	minimum(), maximum()	deepcopy()	ntuple()	LinearAlgebra
struct	Any	eltype()	export	true, false	String, string()	undef	@inbounds	BigInt, BigFloat, big()	vec()	Int64, Int32, etc…	Float64, Float32, etc..
macro	Nothing, nothing	isprimitivetype()	isa()	global	parse()	Random	@edit	Tuple, tuple()	convert()	rand()	findmax(), findmin()
ans	mutable struct	Missing, missing	NamedTuple	local	@time, @timed	cd(), pwd()	fieldnames()	Array	dropdims()	randn()	^
const	ndims()	print(), println()	collect()	include()	findall()	@__DIR__	÷	%	try, catch, finally	[]	@.
stdout, stderr, stdin	@warn_type	varinfo()	reshape()	cat(), vcat(), hcat()	$	delete!()	//	exp()	Statistics	@isdefined	log(), log2(), log10()
map()	sum()	if, else	@code_llvm	haskey(), keys(), values(),	sqrt(),√ issqrt()	using	zero()	let	Dates	|>	π
filter()	Dict()	supertype()	resize!()	@doc raw	REPL: ?,], ;	falses(), trues()	Set()	.*, .+, ./ etc..	exit()	isempty()	argmax(), argmin()
sort(), sort!()	module	subtypes()	display(), show()	append!()	zeros(), ones()	Pkg	abstract type	count()	clipboard()	abs(), abs2()	@assert

Act 6 - Code Examples

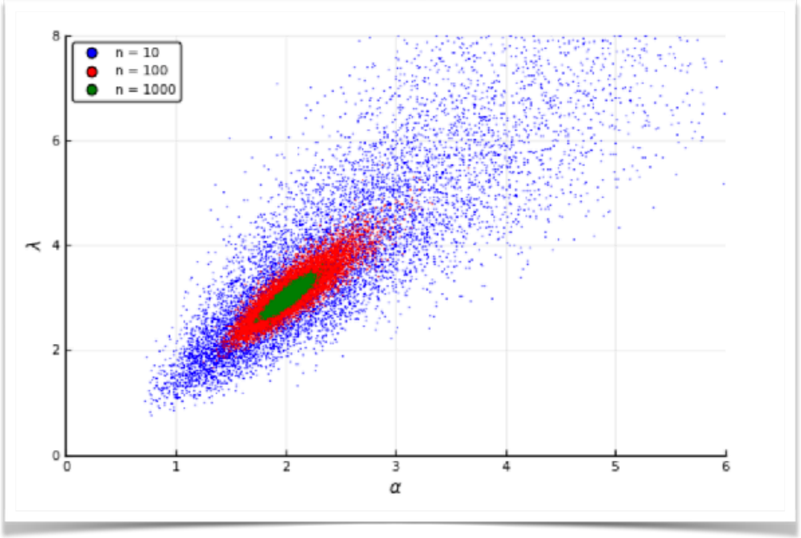
1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/1_chapter/variableScope.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/callByValueByReference.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/shallowDeepCopy.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/dataframeReshape.jl

Act 7: More distributions and fitting



Act 7: More distributions and fitting

Data Science concepts used in this act...

- The exponential and geometric memoryless distributions
- Kernel Density Estimates (KDE)
- Empirical Cumulative Distribution Functions (ECDF)
- The Gamma Distribution and Maximum Likelihood Estimates

Act 7 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/3_chapter/expGeom.jl

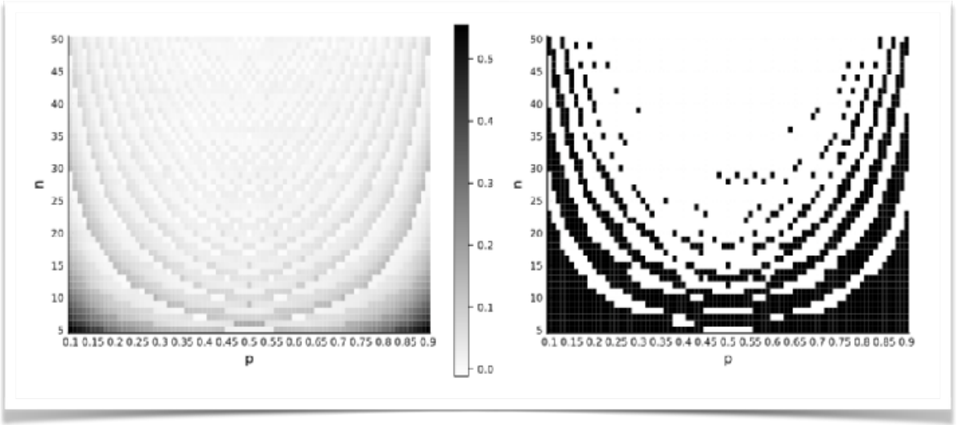
2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/statsPlotsDensity.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/KDE.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/4_chapter/ecdf.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/5_chapter/mleGamma.jl

Act 8: Small numerical experiments



Act 8: Small numerical experiments

Data Science concepts used in this act...

- Method of Moments parameter estimation
- Maximum Likelihood Estimation
- Confidence intervals for a promotion
- Confidence intervals for the variance
- T-tests
- Signed Rank tests
- The Susceptible Exposed Infected Removed (SEIR) model

Act 8 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/5_chapter/mm_vs_mle.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/5_chapter/bayesUnivariate.jl

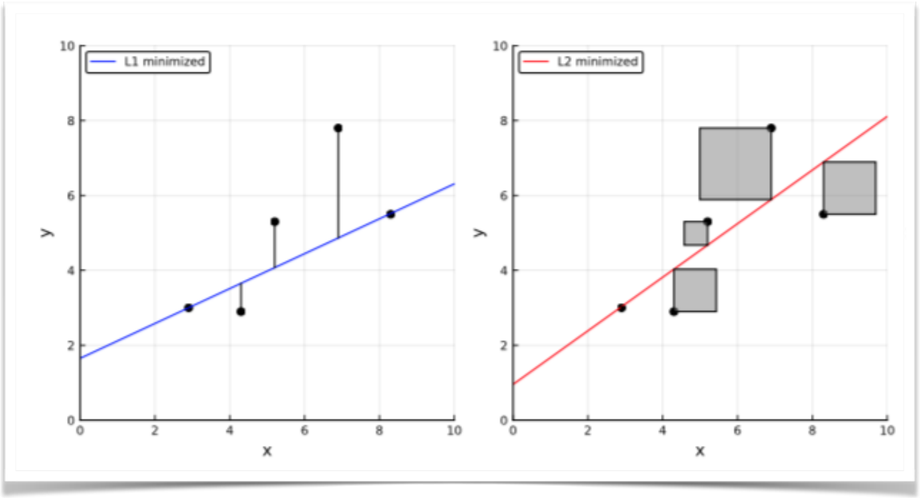
3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/6_chapter/propCIcoverageAccuracy.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/6_chapter/varianceCIalphas.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/7_chapter/TvsSign.jl

6) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/10_chapter/deterministicSEIR.jl

Act 9: Fitting models



Act 9: Fitting models

Data Science concepts used in this act...

- Residual analysis in regression
- Multiple linear regression
- Collinearity
- Interactions in regression
- Ridge regression
- K-fold cross validation
- LASSO
- Generalized Linear Models (GLM)

Act 9 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/residualAnalysis.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/multiLinReg.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/collinearity.jl

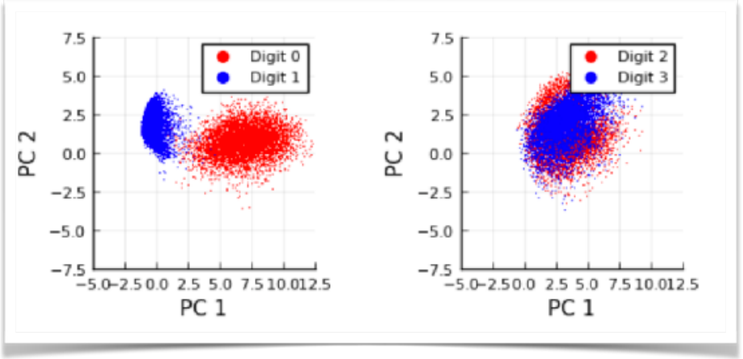
4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/interaction.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/ridgeRegressionCross.jl

6) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/lassoSelection.jl

7) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/8_chapter/linkFunctions.jl

Act 10: Machine learning and wrap up



Act 10: Machine learning and wrap up

Data Science concepts used in this act...

- Principal Component Analysis (PCA)
- Logistic Softmax Regression (Multinomial Regression)
- Neural Networks
- Q-learning

Act 10 - Code Examples

1) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/pcaMNIST.jl

2) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/logisticRegressionMNIST.jl

3) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/pretrainedMNIST.jl

4) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/NN_mnist.jl

5) https://github.com/h-Klok/StatsWithJuliaBook/blob/master/9_chapter/qLearning.jl

More for Another Day...

- Programming for type stability
- Parametric types
- Parallel and distributed computing
- Package development
- Coroutines
- ... Many packages for different tasks.
- More statistical content from the “Statistics with Julia” book includes: Probability, Hypothesis Tests, basic time series, Statistical Foundations, more aspects of machine learning, and Simulation of Stochastic Processes.

The End