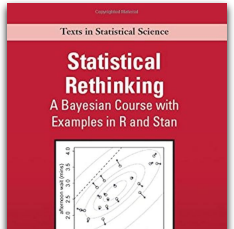
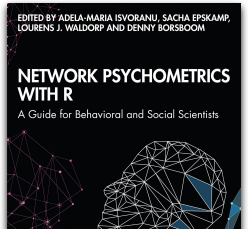
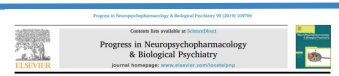
Best Practices in Writing Reproducible Code in R

Ms.C. Diego Angeles-Valdez
University of Groningen | Instituto de Neurobiología, UNAM

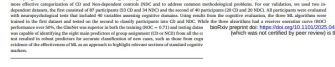
About me







Identifying cognitive deficits in cocaine dependence using standard tests and machine learning



Outline

R Packages

- Overview of essential R packages used in data analysis

Glimpse of the Data

- Initial exploration and summary of data

Statistical Analysis

- Applying statistical methods to interpret data

Visualization

- Creating visual representations to illustrate findings

Course Material

Tidyverse

Tidyverse

🇫🇷 *Tidy* for "bien rangé" and *verse* for "univers"

A collection of R 📦 developed by H. Wickham and others at Rstudio

Initial release: Sep, 2016

Hadley Wickham

Tidyverse

Tidyverse is most importantly a philosophy for data analysis.

- More efficient code
- Easier to remember syntax
- Easier to read syntax

Remember install and load the package

```
#install.packages("tidyverse")
library(tidyverse)
```

Hadley Wickham

Tidyverse

```r
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✔ ggplot2 3.3.3     ✔ purrr   0.3.4
## ✔ tibble  3.1.0     ✔ dplyr   1.0.4
## ✔ tidyr   1.1.2     ✔ forcats 0.5.1
## ✔ readr   1.4.0

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::collapse()   masks IRanges::collapse()
## ✖ dplyr::combine()    masks Biobase::combine(), BiocGenerics::combine()
## ✖ dplyr::desc()       masks IRanges::desc()
## ✖ tidyr::expand()     masks S4Vectors::expand()
## ✖ tidyr::extract()    masks magrittr::extract()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::first()      masks S4Vectors::first()
## ✖ dplyr::lag()        masks stats::lag()
## ✖ ggplot2::Position() masks BiocGenerics::Position(), base::Position()
## ✖ purrr::reduce()     masks GenomicRanges::reduce(), IRanges::reduce()
## ✖ dplyr::rename()     masks S4Vectors::rename()
```

Tidyverse

Tidyverse is a collection of R 📦

ggplot2 - visualising stuff

dplyr, tidyr - data manipulation

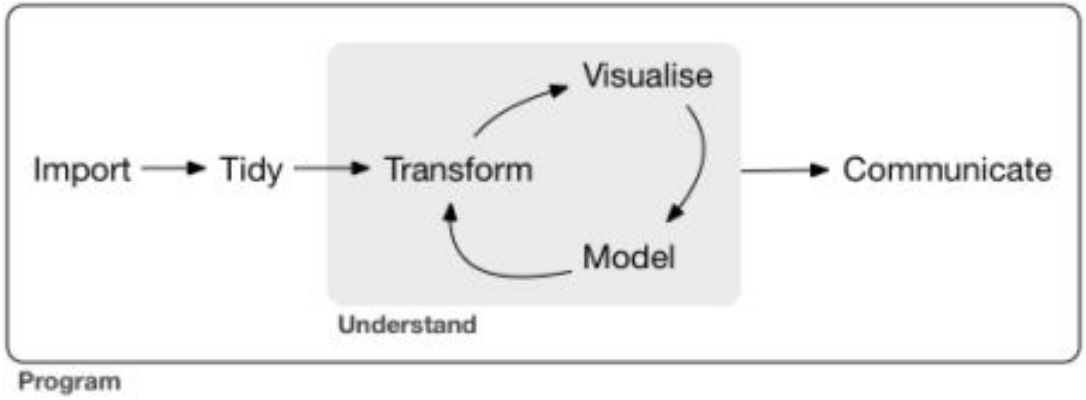
purrr - advanced programming

readr - import data

tibble - improved data.frame format

forcats - working w/ factors

stringr - working w/ chain of characters



Introduction to dplyr

All of the dplyr functions take a data frame (or tibble) as the first argument.

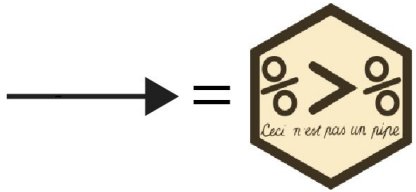
dplyr provides the **%>%**

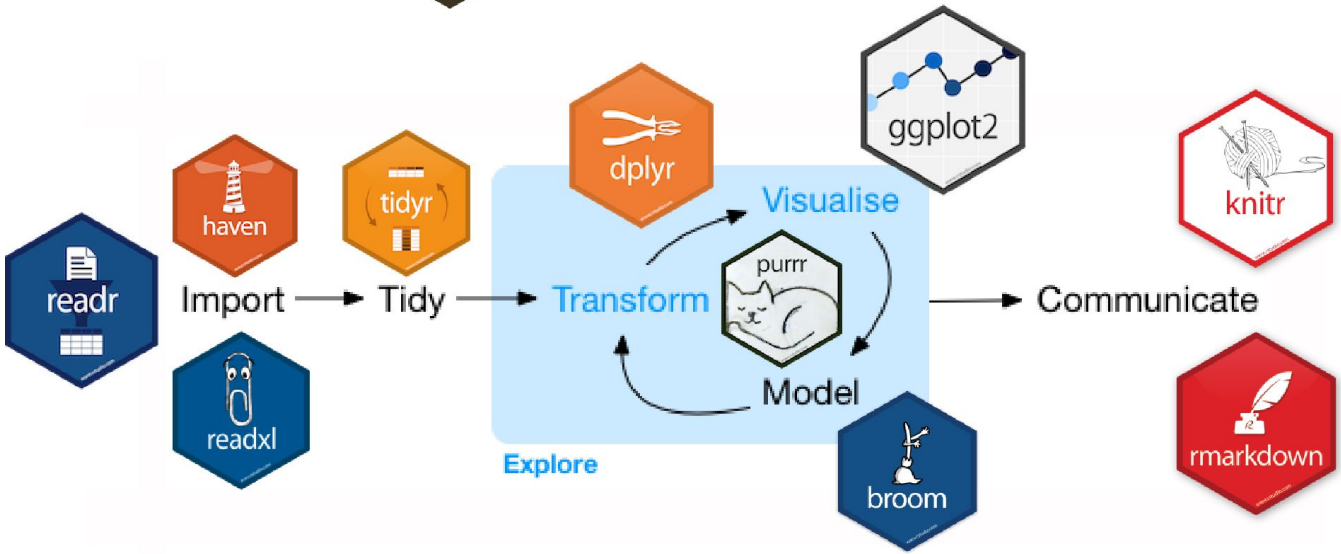
You can use the pipe to rewrite multiple operations that you can read left-to-right, top-to-bottom .

Your data frame

Your function of choice

```
dataframe %>% function(parameters, ... ) %>%  ...
```





data.frames

factors

strings

Graphics in R

The ggplot2 package was created by Hadley Wickham to provide an intuitive plotting system.

In order to produce a ggplot2 graph we need a minimum of:

- Data to be used in graph
- Mappings of data to the graph (**aesthetic mapping**)
- What type of graph we want to use (The **geom** to use).

Table

Descriptive statistics are an essential part of data analysis as they give you an initial overview of your data.

To create descriptive tables in R, you can use:

- moonBook
- table1
- tableone
- tables

Examples

	Stratified by trt 1	2	p	test
n	158	154		
time (mean (SD))	2,015.62 (1,094.12)	1,996.86 (1,155.93)	0.883	
status (%)			0.894	
0	83 (52.5)	85 (55.2)		
1	10 (6.3)	9 (5.8)		
2	65 (41.1)	60 (39.0)		
trt = 2 (%)	0 (0.0)	154 (100.0)	<0.001	
age (mean (SD))	51.42 (11.01)	48.58 (9.96)	0.018	
sex = f (%)	137 (86.7)	139 (90.3)	0.421	
ascites = 1 (%)	14 (8.9)	10 (6.5)	0.567	
hepato = 1 (%)	73 (46.2)	87 (56.5)	0.088	
spiders = 1 (%)	45 (28.5)	45 (29.2)	0.985	
edema (%)			0.877	
0	132 (83.5)	131 (85.1)		
0.5	16 (10.1)	13 (8.4)		
1	10 (6.3)	10 (6.5)		
bili (median [IQR])	1.40 [0.80, 3.20]	1.30 [0.72, 3.60]	0.842	nonnorm
chol (median [IQR])	315.50 [247.75, 417.00]	303.50 [254.25, 377.00]	0.544	nonnorm
albumin (mean (SD))	3.52 (0.44)	3.52 (0.40)	0.874	
copper (median [IQR])	73.00 [40.00, 121.00]	73.00 [43.00, 139.00]	0.717	nonnorm
alk.phos (median [IQR])	1,214.50 [840.75, 2,028.00]	1,283.00 [922.50, 1,949.75]	0.812	nonnorm
ast (median [IQR])	111.60 [76.73, 151.51]	117.40 [83.78, 151.90]	0.459	nonnorm
trig (median [IQR])	106.00 [84.50, 146.00]	113.00 [84.50, 155.00]	0.370	nonnorm
platelet (mean (SD))	258.75 (100.32)	265.20 (90.73)	0.555	
protime (median [IQR])	10.60 [10.03, 11.00]	10.60 [10.00, 11.40]	0.588	nonnorm
stage (%)			0.201	
1	12 (7.6)	4 (2.6)		
2	35 (22.2)	32 (20.8)		
3	56 (35.4)	64 (41.6)		
4	55 (34.8)	54 (35.1)		

Descriptive Statistics by 'Dx'

	NSTEMI (N=153)	STEMI (N=304)	Unstable Angina (N=400)	p
age	64.3 ± 12.3	62.1 ± 12.1	63.8 ± 11.0	0.073
sex				0.012
- Female	50 (32.7%)	84 (27.6%)	153 (38.2%)	
- Male	103 (67.3%)	220 (72.4%)	247 (61.8%)	
cardiogenicShock				0.000
- No	149 (97.4%)	256 (84.2%)	400 (100.0%)	
- Yes	4 (2.6%)	48 (15.8%)	0 (0.0%)	
entry				0.001
- Femoral	58 (37.9%)	133 (43.8%)	121 (30.2%)	
- Radial	95 (62.1%)	171 (56.2%)	279 (69.8%)	
EF	55.0 ± 9.3	52.4 ± 9.5	59.2 ± 8.7	0.000
height	163.3 ± 8.2	165.1 ± 8.2	161.7 ± 9.7	0.000
weight	64.3 ± 10.2	65.7 ± 11.6	64.5 ± 11.6	0.361
BMI	24.1 ± 3.2	24.0 ± 3.3	24.6 ± 3.4	0.064
obesity				0.186
- No	106 (69.3%)	209 (68.8%)	252 (63.0%)	
- Yes	47 (30.7%)	95 (31.2%)	148 (37.0%)	
TC	193.7 ± 53.6	183.2 ± 43.4	183.5 ± 48.3	0.057
LDLC	126.1 ± 44.7	116.7 ± 39.5	112.9 ± 40.4	0.004
HDLC	38.9 ± 11.9	38.5 ± 11.0	37.8 ± 10.9	0.501
TG	130.1 ± 88.5	106.5 ± 72.0	137.4 ± 101.6	0.000
DM				0.209
- No	96 (62.7%)	208 (68.4%)	249 (62.2%)	
- Yes	57 (37.3%)	96 (31.6%)	151 (37.8%)	
HBP				0.002
- No	62 (40.5%)	150 (49.3%)	144 (36.0%)	
- Yes	91 (59.5%)	154 (50.7%)	256 (64.0%)	
smoking				0.000
- Ex-smoker	42 (27.5%)	66 (21.7%)	96 (24.0%)	
- Never	50 (32.7%)	97 (31.9%)	185 (46.2%)	
- Smoker	61 (39.9%)	141 (46.4%)	119 (29.8%)	

Basic stats

	Total (N=205)	Alive (N=134)	Death: Melanoma (N=57)	Death: Non-melanoma (N=14)
Sex				
Male	79 (39 %)	43 (32 %)	29 (51 %)	7 (50 %)
Female	126 (61 %)	91 (68 %)	28 (49 %)	7 (50 %)
Age (years)				
Mean (SD)	52 (± 17)	50 (± 16)	55 (± 18)	65 (± 11)
Ulceration				
Absent	115 (56 %)	92 (69 %)	16 (28 %)	7 (50 %)
Present	90 (44 %)	42 (31 %)	41 (72 %)	7 (50 %)
Thickness[a] (mm)				
Mean (SD)	2.9 (± 3.0)	2.2 (± 2.3)	4.3 (± 3.6)	3.7 (± 3.6)

[a] Also known as Breslow thickness

		Sepal.Length		Sepal.Width	
Species	n	mean	sd	mean	sd
setosa	50	5.01	0.35	3.43	0.38
versicolor	50	5.94	0.52	2.77	0.31
virginica	50	6.59	0.64	2.97	0.32
Overall, we see the following:					
All	150	5.84	0.83	3.06	0.44

Our first ggplot2 graph

```
p <- ggplot(mtcars, aes(factor(cyl), mpg))
p + geom_violin()
```

Data

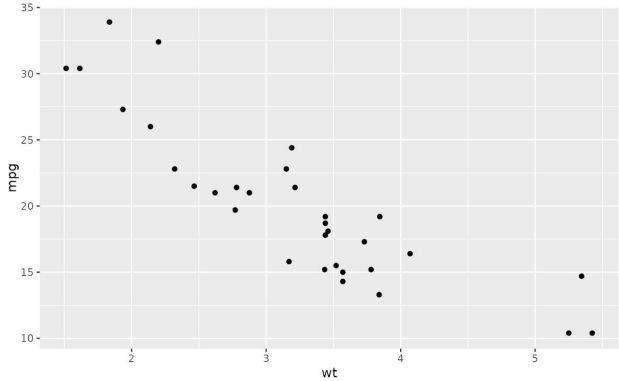
Geom

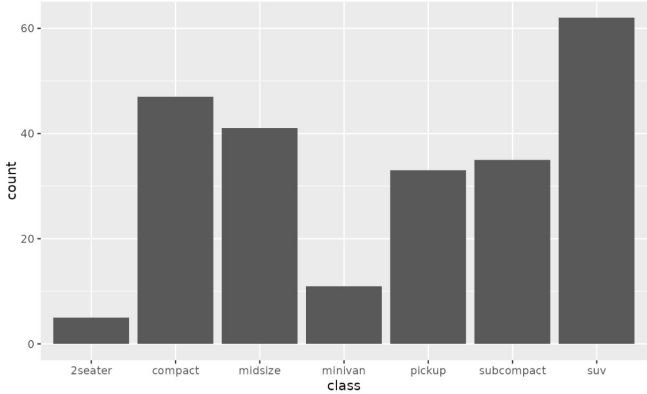
Aesthetic mapping

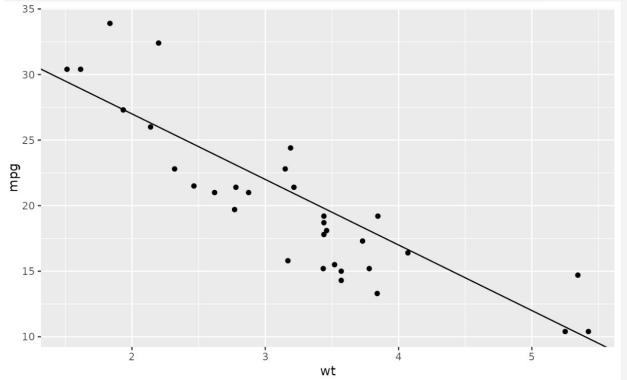
Our first ggplot2 graph

For example:

- `geom_point()` for scatter plots,
- `geom_line()` for line charts,
- `geom_bar()` for bar plots,
- `geom_histogram()` for histograms,
- `geom_smooth()` to add trend lines.







ggplot2

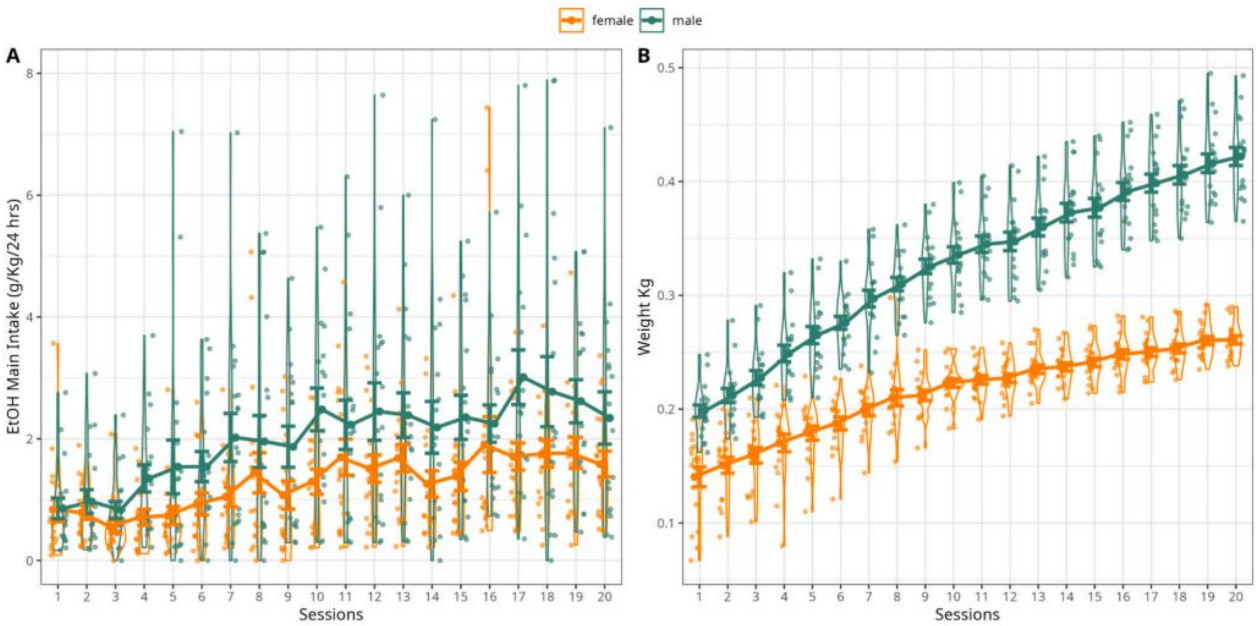
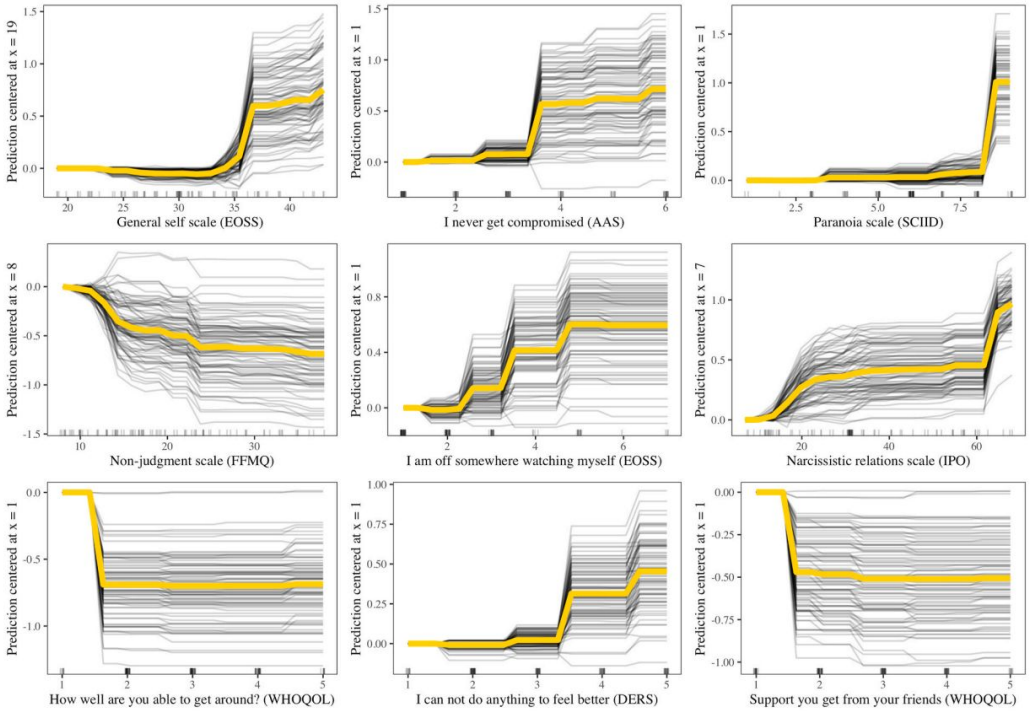


Fig. 1. Ethanol intake and weight trajectories.
Overview of **(A)** ethanol main intake, **(B)** and the weight of all rats over IA2BC model by sex. The values are expressed as mean EtOH Main intake (g/kg/24 h), and weight (Kg) ± SEM at each drinking session.

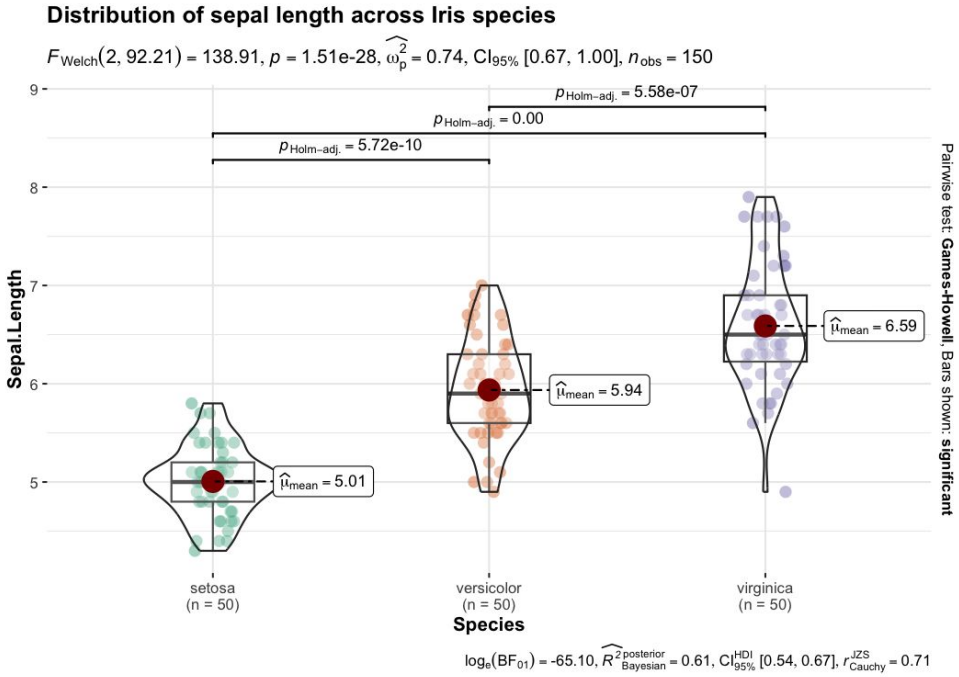
ggpubr

The ‘ggpubr’ package provides some easy-to-use functions for creating and customizing ‘ggplot2’-based publication ready plots.

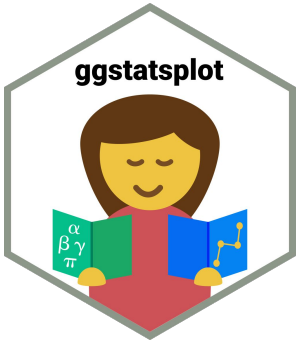


Other R packages for plotting

Based Plots with Statistical Details



Other R packages for plotting



Based Plots with Statistical Details

Results from Welch's t-test with {statsExpressions}

Template for Frequentist analysis

test | parameter | statistic | significance | effect size type + estimate + confidence intervals | number of observations

$$t_{\text{Welch}}(281.95) = -10.75,\ p = 8.31\text{e-}23,\ \hat{g}_{\text{Hedges}} = -1.27,\ \text{CI}_{99\%}[-1.61, -0.94],\ n_{\text{obs}} = 284$$

Template for Bayesian analysis

evidence in favor of null over alternative hypothesis | natural logarithm of Bayes Factor | posterior type + estimate + credible intervals | prior type and value

$$\log_e(\text{BF}_{01}) = -6.20,\ \delta^{\text{posterior}}_{\text{difference}} = -5.06,\ \text{CI}^{\text{HDI}}_{95\%}\ [-6.75, -3.53],\ r^{\text{JZS}}_{\text{cauchy}} = 0.71$$

<https://indrajeetpatil.github.io/statsExpressions/>

src: @patilindrajeets

Extra information

Linear Mixed models

- lmer
- broom

Network analysis

- bootnet
- psychonetrics
- qgraph

SEM or path model

- lavaan

Machine Learning

- caret
  - randomForest
  - xgboost
  - e1071
  - …
- tidymodels

Thank you

Contact me

M. Sc. Diego Angeles-Valdez,
PhD candidate,
University Medical Center Groningen, Cognitive Neuroscience Center,
Biomedical Sciences

Universidad Nacional Autónoma de México (UNAM)
Institute of Neurobiology (INB)

Personal Website: https://diegoangls.github.io/
Twitter: @diegoangls
Bluesky: @dangeles.bsky.social

Phone: +31 6 39 88 71 29
Email: d.angeles.valdez@rug.nl | d.angeles.valdez@umcg.nl

Work address:P.O. Box 196, 9700 AD GRONINGEN, The Netherlands