R-Helper-Functions

Marcus Vollmer

2018-05-14

Contents

Data import and R in-build summary

Standard build-in summary functions are demonstrated using a dataset on chronic ganulotomous disease (CGD) from the `survival` *package.* `summary` *produces result summaries of the input data as given below:*

```
library(survival)
cgd = get("cgd")
summary(cgd)
```

```
##        id                            center       random
##  Min.   :  1.00   NIH                  :41   Min.   :1989-06-07
##  1st Qu.: 24.50   Scripps Institute    :36   1st Qu.:1989-08-19
##  Median : 54.00   Amsterdam            :28   Median :1989-09-15
##  Mean   : 58.09   Univ. of Zurich      :21   Mean   :1989-09-22
##  3rd Qu.: 89.50   Mott Children's Hosp:20   3rd Qu.:1989-11-03
##  Max.   :135.00   L.A. Children's Hosp:13   Max.   :1989-12-29
##                   (Other)              :44
##      treat          sex           age           height
##  placebo:120   male  :168   Min.   : 1.0   Min.   : 76.3
##  rIFN-g : 83   female: 35   1st Qu.: 6.0   1st Qu.:114.5
##                             Median :12.0   Median :140.0
##                             Mean   :13.7   Mean   :138.1
##                             3rd Qu.:20.0   3rd Qu.:169.2
##                             Max.   :44.0   Max.   :189.0
##
##      weight            inherit       steroids          propylac
##  Min.   : 10.40   X-linked :131   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 20.25   autosomal: 72   1st Qu.:0.00000   1st Qu.:1.0000
##  Median : 33.40                   Median :0.00000   Median :1.0000
##  Mean   : 39.34                   Mean   :0.03448   Mean   :0.8473
##  3rd Qu.: 58.70                   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :101.50                   Max.   :1.00000   Max.   :1.0000
##
##              hos.cat        tstart           enum           tstop
##  US:NIH          : 41   Min.   :  0.0   Min.   :1.000   Min.   :  4.0
##  US:other        :108   1st Qu.:  0.0   1st Qu.:1.000   1st Qu.:204.5
##  Europe:Amsterdam: 28   Median :  0.0   Median :1.000   Median :273.0
##  Europe:other    : 26   Mean   : 69.5   Mean   :1.665   Mean   :254.1
##                         3rd Qu.:121.0   3rd Qu.:2.000   3rd Qu.:320.0
##                         Max.   :373.0   Max.   :8.000   Max.   :439.0
##
##      status
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3744
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Required libraries and own functions for summary tables

In order to summarize the data stored as a table one can use the `strtable` *for analyzing variable names, variable classes, number of missing values, and basic stats. For factor and logical variable the level names along with their counts are given. For integer and numerical variable mean and standard deviation and quantiles (0%,25%,50%,75%,100%) including upper and lower limits are given. The* `stargazer` *package and the* `stargazer_long` *modification will plot the content as LaTeX tables.* `stargazer_long` *will convert the normal* `stargazer` *latex output to a long table which automatically splits the table whenever the table doesn't fit the paper size. Moreover, it is now possible to rotate the table heading by setting the rotation angle and to fix the table width by specifying the output paramater in standard latex notation, e.g.* `output="cccp{9cm}`.

```
  library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
  library(stringr)
  source("stargazer_long.R")
  source("strtable.R")

# generate summary table
  s = strtable(cgd[,-1], n=10, width=300)

# for LaTeX output as longtable use:
  stargazer_long(s[,1:4], summary=FALSE, rownames=FALSE, output="cccp{9cm}", rotate=60)
```

variable	NAs	class	stats
center	0	Factor w/ 13 levels	"Harvard Medical Sch" (4), "Scripps Institute" (36), "Copenhagen" (5), "NIH" (41), "L.A. Children's Hosp" (13), "Mott Children's Hosp" (20), "Univ. of Utah" (5), "Univ. of Washington" (4), "Univ. of Minnesota" (10), "Univ. of Zurich" (21), ...
random	0	Date	
treat	0	Factor w/ 2 levels	"placebo" (120), "rIFN-g" (83)
sex	0	Factor w/ 2 levels	"male" (168), "female" (35)
age	0	integer	m=13.70, sd=9.34, q=[1.00, 6.00, 12.00, 20.00, 44.00]
height	0	numeric	m=138.12, sd=31.41, q=[76.30, 114.50, 140.00, 169.50, 189.00]
weight	0	numeric	m=39.34, sd=21.83, q=[10.40, 20.10, 33.40, 59.00, 101.50]
inherit	0	Factor w/ 2 levels	"X-linked" (131), "autosomal" (72)
steroids	0	numeric	m=0.03, sd=0.18, q=[0.00, 0.00, 0.00, 0.00, 1.00]
propylac	0	numeric	m=0.85, sd=0.36, q=[0.00, 1.00, 1.00, 1.00, 1.00]
hos.cat	0	Factor w/ 4 levels	"US:NIH" (41), "US:other" (108), "Europe:Amsterdam" (28), "Europe:other" (26)
tstart	0	integer	m=69.50, sd=111.62, q=[0.00, 0.00, 0.00, 121.00, 373.00]
enum	0	integer	m=1.67, sd=1.16, q=[1.00, 1.00, 1.00, 2.00, 8.00]
tstop	0	integer	m=254.11, sd=96.38, q=[4.00, 203.00, 273.00, 322.00, 439.00]
status	0	integer	m=0.37, sd=0.49, q=[0.00, 0.00, 0.00, 1.00, 1.00]

Create stratified table of data with statistical testing

Many studies starts with a characteristics table/ study population with separate columns for each cohort, e.g. treatment vs. control. In our example data set it is the column `treat` *with has two treatment levels: placebo and gamma interferon (rIFN-g).* `characteristics_table.R` *will do the job of printing characteristics seperated by a binary response (treat). P-values are the results of statistical testing comparing both groups: T test and Wilcoxon ranksum test for continous variables with mean and SD or median and quartiles respectively. Statistical testing with categorical data is conducted with Fishers exact test or* X^2*-Test (categorical with more than 3 levels). Missing values (NA's) was omitted for this analysis. Precision of numerials can be set individually for continuous values and p values. Default is 1 digit after decimal place* `prec="%.1f"`, `prec_continuous="%.0f"` *and 4 digits for p values* `prec_p="%.4f"`.

```
source("characteristics_table.R")
s = characteristics_table(-2, "treat", cgd[,-c(1:2)], "col", prec="%.1f", prec_continuous="%.1f")
stargazer_long(s, summary=FALSE, rownames=FALSE, output="p{1.5cm}p{4cm}rrrr")
```

Variable	Level	placebo	rIFN-g	P	NAs
sex	male	100 (83.3)	68 (81.9)	0.8510	0
	female	20 (16.7)	15 (18.1)		
inherit	X-linked	74 (61.7)	57 (68.7)	0.3709	0
	autosomal	46 (38.3)	26 (31.3)		
hos.cat	US:NIH	20 (16.7)	21 (25.3)	0.4469	0
	US:other	67 (55.8)	41 (49.4)		
	Europe:Amsterdam	16 (13.3)	12 (14.5)		
	Europe:other	17 (14.2)	9 (10.8)		
age	Median (Quartiles)	11.5 (5.0,21.2)	12.0 (7.0,18.5)	0.7509	0
age	Mean (SD)	13.6 (9.4)	13.9 (9.3)	0.8008	0
height	Median (Quartiles)	140.1 (107.8,169.8)	140.0 (120.0,166.5)	0.6329	0
height	Mean (SD)	136.7 (34.9)	140.2 (25.7)	0.4108	0
weight	Median (Quartiles)	33.4 (18.1,63.5)	34.4 (22.2,52.0)	0.7373	0
weight	Mean (SD)	39.7 (23.7)	38.9 (19.0)	0.7879	0
steroids	Median (Quartiles)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.1472	0
steroids	Mean (SD)	0.1 (0.2)	0.0 (0.1)	0.1055	0
propylac	Median (Quartiles)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	0.5085	0
propylac	Mean (SD)	0.8 (0.4)	0.9 (0.3)	0.5015	0
tstart	Median (Quartiles)	0.0 (0.0,170.5)	0.0 (0.0,0.0)	0.0049	0
tstart	Mean (SD)	82.1 (116.1)	51.3 (102.8)	0.0486	0
enum	Median (Quartiles)	1.0 (1.0,2.0)	1.0 (1.0,1.0)	0.0007	0
enum	Mean (SD)	1.9 (1.4)	1.3 (0.6)	0.0001	0
tstop	Median (Quartiles)	267.0 (194.2,306.2)	279.0 (241.5,338.0)	0.0051	0
tstop	Mean (SD)	236.4 (103.9)	279.7 (78.0)	0.0009	0
status	Median (Quartiles)	0.0 (0.0,1.0)	0.0 (0.0,0.0)	0.0011	0
status	Mean (SD)	0.5 (0.5)	0.2 (0.4)	0.0007	0

Alternatively `xtable` *with longtable replacement will do the printing job.*

```
library(xtable)
# Longtable LaTeX output
  out = capture.output(xtable(s, align="rp{1.5cm}p{4cm}rrrr"))
  out = out[6:NROW(out)-1]
  out = sub("\\{tabular\\}", "\\{longtable\\}", out)
  cat(out)
```

	Variable	Level	placebo	rIFN-g	P	NAs
1	sex	male	100 (83.3)	68 (81.9)	0.8510	0
2		female	20 (16.7)	15 (18.1)		

3	inherit	X-linked	74 (61.7)	57 (68.7)	0.3709	0
4		autosomal	46 (38.3)	26 (31.3)		
5	hos.cat	US:NIH	20 (16.7)	21 (25.3)	0.4469	0
6		US:other	67 (55.8)	41 (49.4)		
7		Europe:Amsterdam	16 (13.3)	12 (14.5)		
8		Europe:other	17 (14.2)	9 (10.8)		
9	age	Median (Quartiles)	11.5 (5.0,21.2)	12.0 (7.0,18.5)	0.7509	0
10	age	Mean (SD)	13.6 (9.4)	13.9 (9.3)	0.8008	0
11	height	Median (Quartiles)	140.1 (107.8,169.8)	140.0 (120.0,166.5)	0.6329	0
12	height	Mean (SD)	136.7 (34.9)	140.2 (25.7)	0.4108	0
13	weight	Median (Quartiles)	33.4 (18.1,63.5)	34.4 (22.2,52.0)	0.7373	0
14	weight	Mean (SD)	39.7 (23.7)	38.9 (19.0)	0.7879	0
15	steroids	Median (Quartiles)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.1472	0
16	steroids	Mean (SD)	0.1 (0.2)	0.0 (0.1)	0.1055	0
17	propylac	Median (Quartiles)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	0.5085	0
18	propylac	Mean (SD)	0.8 (0.4)	0.9 (0.3)	0.5015	0
19	tstart	Median (Quartiles)	0.0 (0.0,170.5)	0.0 (0.0,0.0)	0.0049	0
20	tstart	Mean (SD)	82.1 (116.1)	51.3 (102.8)	0.0486	0
21	enum	Median (Quartiles)	1.0 (1.0,2.0)	1.0 (1.0,1.0)	0.0007	0
22	enum	Mean (SD)	1.9 (1.4)	1.3 (0.6)	0.0001	0
23	tstop	Median (Quartiles)	267.0 (194.2,306.2)	279.0 (241.5,338.0)	0.0051	0
24	tstop	Mean (SD)	236.4 (103.9)	279.7 (78.0)	0.0009	0
25	status	Median (Quartiles)	0.0 (0.0,1.0)	0.0 (0.0,0.0)	0.0011	0
26	status	Mean (SD)	0.5 (0.5)	0.2 (0.4)	0.0007	0

Cross table for count data with confidence intervals

Printing a cross table with p value and percentages (rows sums up to 100%).

```
library(PropCIs)
source("mytable.R")
mytable(cgd$sex, cgd$treat, ci=FALSE, prec="%.2f", latex=TRUE)
```

	placebo	rIFN-g
male	100 (59.52)	68 (40.48)
female	20 (57.14)	15 (42.86)

Printing the same table with 95% exact Clopper-Pearson confidence intervals (`PropCIs` package required) and more less digits with % sign and a fixed column width:

```
mytable(cgd$sex, cgd$treat, ci=0.95, prec="%.1f", latex="rp{4cm}p{4cm}", pct_sign="%")
```

	placebo	rIFN-g
male	100 (59.5%, 51.7% to 67.0%)	68 (40.5%, 33.0% to 48.3%)
female	20 (57.1%, 39.4% to 73.7%)	15 (42.9%, 26.3% to 60.6%)