R-Helper-Functions

Marcus Vollmer

2018-08-07

Contents

Data import and R in-build summary

Standard build-in summary functions are informative but the output looks messy. This is demonstrated by plotting the summary of a standard dataset on chronic ganulotomous disease (CGD) from the `survival` package. Given the input data, `summary` produces the following output:

```
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 5.2.2. 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]

```
# as HTML or text:
stargazer(s, summary = FALSE, rownames = FALSE, type = "html")
```

```
##
## <table style="text-align:center"><tr><td colspan="9" style="border-bottom: 1px solid black"></td></t
```

```
## <tr><td colspan="9" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td style="text-align:left">random</td><td>0</td><td>Date</td><td></td><td></td><td></td><td></t
## <tr><td style="text-align:left">treat</td><td>0</td><td>Factor w/ 2 levels</td><td>"placebo" (120),
## <tr><td style="text-align:left">sex</td><td>0</td><td>Factor w/ 2 levels</td><td>"male" (168), "fema
## <tr><td style="text-align:left">age</td><td>0</td><td>integer</td><td>m=13.70, sd=9.34, q=[1.00, 6.0
## <tr><td style="text-align:left">height</td><td>0</td><td>numeric</td><td>m=138.12, sd=31.41, q=[76.3
## <tr><td style="text-align:left">weight</td><td>0</td><td>numeric</td><td>m=39.34, sd=21.83, q=[10.40
## <tr><td style="text-align:left">inherit</td><td>0</td><td>Factor w/ 2 levels</td><td>"X-linked" (131
## <tr><td style="text-align:left">steroids</td><td>0</td><td>numeric</td><td>m=0.03, sd=0.18, q=[0.00,
## <tr><td style="text-align:left">propylac</td><td>0</td><td>numeric</td><td>m=0.85, sd=0.36, q=[0.00,
## <tr><td style="text-align:left">hos.cat</td><td>0</td><td>Factor w/ 4 levels</td><td>"US:NIH" (41),
## <tr><td style="text-align:left">tstart</td><td>0</td><td>integer</td><td>m=69.50, sd=111.62, q=[0.00
## <tr><td style="text-align:left">enum</td><td>0</td><td>integer</td><td>m=1.67, sd=1.16, q=[1.00, 1.0
## <tr><td style="text-align:left">tstop</td><td>0</td><td>integer</td><td>m=254.11, sd=96.38, q=[4.00,
## <tr><td style="text-align:left">status</td><td>0</td><td>integer</td><td>m=0.37, sd=0.49, q=[0.00, 0
## <tr><td colspan="9" style="border-bottom: 1px solid black"></td></tr></table>
```

```
stargazer(s, summary = FALSE, rownames = FALSE, type = "text")
```

```
##
## ================================================================================================
## variable NAs        class
## ------------------------------------------------------------------------------------------------
## center    0  Factor w/ 13 levels "Harvard Medical Sch" (4), "Scripps Institute" (36), "Copenhagen" (
## random    0         Date
## treat     0  Factor w/ 2 levels
## sex       0  Factor w/ 2 levels
## age       0        integer
## height    0        numeric
## weight    0        numeric
## inherit   0  Factor w/ 2 levels
## steroids  0        numeric
## propylac  0        numeric
## hos.cat   0  Factor w/ 4 levels
## tstart    0        integer
## enum      0        integer
## tstop     0        integer
## status    0        integer
## ------------------------------------------------------------------------------------------------
```

Create stratified table of data with statistical testing

Many studies starts with a characteristics table of the study population with separate columns for each cohort, e.g. treatment vs. control. In our example data set we stratify by `treat` which 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 one digit after decimal place `prec="%.1f",prec_continuous="%.0f"` and 4 digits for p values `prec_p="%.4f"`.

```
source("characteristics_table.R")
characteristics_table(-2, "treat", cgd[, -c(1:2)], "col", prec = "%.1f",
    prec_continuous = "%.1f", latex = "p{1.5cm}p{4cm}rrrr", tablefootnote = TRUE)
```

Variable	Level	placebo	rIFN-g	P	NAs
sex	male	100 (83.3)	68 (81.9)	0.8510[2]	0
	female	20 (16.7)	15 (18.1)		
inherit	X-linked	74 (61.7)	57 (68.7)	0.3709[2]	0
	autosomal	46 (38.3)	26 (31.3)		
hos.cat	US:NIH	20 (16.7)	21 (25.3)	0.4469[1]	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[4]	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[4]	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[4]	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[4]	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[4]	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[4]	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[4]	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[4]	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[4]	0
status	Mean (SD)	0.5 (0.5)	0.2 (0.4)	0.0007□	0

[1] Chi-squared test
[2] Fisher's exact test
[4] Wilcoxon rank sum test
□ Student's t-test

It works also for categorical variables with more than 2 levels. Statistical testing will change to X^2-Test, Kruskal-Wallis rank sum test and One-way analysis of variance (ANOVA). Footnotes will tell you the statistical test behind the p values. You may change the footnote labeling as follows:

```
characteristics_table(-2, "hos.cat", cgd[, -c(1:2)], "col", prec = "%.1f",
    prec_continuous = "%.1f", latex = "p{1cm}p{2cm}rrrrrr", tablefootnote = TRUE,
    fn = c("'", "_", "''", "_", "'''", "_"))
```

Variable	Level	US:NIH	US:other	Europe:Amsterdam	Europe:other	P	NAs
treat	placebo	20 (48.8)	67 (62.0)	16 (57.1)	17 (65.4)	0.4469'	0
	rIFN-g	21 (51.2)	41 (38.0)	12 (42.9)	9 (34.6)		
sex	male	34 (82.9)	92 (85.2)	20 (71.4)	22 (84.6)	0.3873'*	0
	female	7 (17.1)	16 (14.8)	8 (28.6)	4 (15.4)		
inherit	X-linked	26 (63.4)	74 (68.5)	14 (50.0)	17 (65.4)	0.3388'	0
	autosomal	15 (36.6)	34 (31.5)	14 (50.0)	9 (34.6)		
age	Median (Quartiles)	14.0 (9.0,25.0)	8.5 (5.0,15.5)	19.5 (12.5,25.0)	11.0 (6.5,21.5)	0.0003"	0
age	Mean (SD)	15.8 (8.6)	11.5 (9.0)	18.3 (8.9)	14.6 (9.9)	0.0015"'	0
height	Median (Quartiles)	145.2 (135.0,168.0)	129.5 (107.4,159.0)	154.5 (137.6,170.6)	141.1 (120.3,169.6)	0.0017"	0
height	Mean (SD)	147.6 (24.5)	130.3 (32.2)	151.3 (28.4)	141.2 (32.8)	0.0011"'	0
weight	Median (Quartiles)	42.9 (33.4,63.7)	27.9 (18.0,47.9)	49.0 (30.5,65.0)	35.5 (20.7,62.8)	0.0013"	0
weight	Mean (SD)	46.0 (20.6)	34.7 (21.4)	47.0 (21.8)	39.8 (21.5)	0.0062"'	0
steroids	Median (Quartiles)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.0000"	0
steroids	Mean (SD)	0.0 (0.0)	0.0 (0.1)	0.2 (0.4)	0.0 (0.0)	0.0000"'	0
propylac	Median (Quartiles)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	1.0 (0.0,1.0)	1.0 (1.0,1.0)	0.0000"	0
propylac	Mean (SD)	1.0 (0.0)	0.8 (0.4)	0.6 (0.5)	0.9 (0.3)	0.0000"'	0
tstart	Median (Quartiles)	0.0 (0.0,118.0)	0.0 (0.0,166.0)	0.0 (0.0,127.2)	0.0 (0.0,0.0)	0.1957"	0
tstart	Mean (SD)	75.6 (122.5)	79.8 (118.0)	62.9 (100.2)	24.3 (59.4)	0.1444"'	0
enum	Median (Quartiles)	1.0 (1.0,2.0)	1.0 (1.0,2.0)	1.0 (1.0,2.0)	1.0 (1.0,1.0)	0.1873"	0
enum	Mean (SD)	1.4 (0.6)	1.9 (1.4)	1.5 (0.8)	1.3 (0.7)	0.0445"'	0
tstop	Median (Quartiles)	294.0 (246.0,365.0)	268.0 (198.5,331.2)	286.0 (257.0,304.0)	269.0 (198.5,286.2)	0.0948"	0
tstop	Mean (SD)	277.0 (102.9)	248.7 (102.1)	263.4 (62.4)	230.6 (87.6)	0.2160"'	0
status	Median (Quartiles)	0.0 (0.0,1.0)	0.0 (0.0,1.0)	0.0 (0.0,1.0)	0.0 (0.0,0.0)	0.3239"	0
status	Mean (SD)	0.4 (0.5)	0.4 (0.5)	0.3 (0.5)	0.2 (0.4)	0.3256"'	0

* Chi-squared approximation may be incorrect

' Chi-squared test

" Kruskal-Wallis rank sum test

"' One-way analysis of variance (ANOVA)

Or you just store the output table and show the results in a different way, e.g. using `stargazer_long` to set more options.

```
s = characteristics_table(-2, "treat", cgd[, -c(1:2)], "col",
    prec = "%.1f", prec_continuous = "%.1f", tablefootnote = FALSE)
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 too.

```
s = characteristics_table(-2, "treat", cgd[, -c(1:2)], "col", prec = "%.1f",
    prec_continuous = "%.1f", tablefootnote = FALSE)
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
NA	sex	male	100 (83.3)	68 (81.9)	0.8510	0
NA1		female	20 (16.7)	15 (18.1)		
NA2	inherit	X-linked	74 (61.7)	57 (68.7)	0.3709	0
NA3		autosomal	46 (38.3)	26 (31.3)		
NA4	hos.cat	US:NIH	20 (16.7)	21 (25.3)	0.4469	0
NA5		US:other	67 (55.8)	41 (49.4)		
NA6		Europe:Amsterdam	16 (13.3)	12 (14.5)		
NA7		Europe:other	17 (14.2)	9 (10.8)		
NA8	age	Median (Quartiles)	11.5 (5.0,21.2)	12.0 (7.0,18.5)	0.7509	0
NA9	age	Mean (SD)	13.6 (9.4)	13.9 (9.3)	0.8008	0
NA10	height	Median (Quartiles)	140.1 (107.8,169.8)	140.0 (120.0,166.5)	0.6329	0
NA11	height	Mean (SD)	136.7 (34.9)	140.2 (25.7)	0.4108	0
NA12	weight	Median (Quartiles)	33.4 (18.1,63.5)	34.4 (22.2,52.0)	0.7373	0
NA13	weight	Mean (SD)	39.7 (23.7)	38.9 (19.0)	0.7879	0
NA14	steroids	Median (Quartiles)	0.0 (0.0,0.0)	0.0 (0.0,0.0)	0.1472	0
NA15	steroids	Mean (SD)	0.1 (0.2)	0.0 (0.1)	0.1055	0
NA16	propylac	Median (Quartiles)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	0.5085	0
NA17	propylac	Mean (SD)	0.8 (0.4)	0.9 (0.3)	0.5015	0
NA18	tstart	Median (Quartiles)	0.0 (0.0,170.5)	0.0 (0.0,0.0)	0.0049	0
NA19	tstart	Mean (SD)	82.1 (116.1)	51.3 (102.8)	0.0486	0
NA20	enum	Median (Quartiles)	1.0 (1.0,2.0)	1.0 (1.0,1.0)	0.0007	0
NA21	enum	Mean (SD)	1.9 (1.4)	1.3 (0.6)	0.0001	0
NA22	tstop	Median (Quartiles)	267.0 (194.2,306.2)	279.0 (241.5,338.0)	0.0051	0
NA23	tstop	Mean (SD)	236.4 (103.9)	279.7 (78.0)	0.0009	0
NA24	status	Median (Quartiles)	0.0 (0.0,1.0)	0.0 (0.0,0.0)	0.0011	0
NA25	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)

Fisher's Exact Test for Count Data: p-Value=8.51e-01

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%)

Fisher's Exact Test for Count Data: p-Value=8.51e-01