Validate Statsomat/edapy

```r
# Import
library(pastecs)
library(Hmisc)
library(knitr)
library(data.table)
library(psych)

# Upload and prepare dfs
filepath = "HolzingerSwineford1939.csv"
df <- fread(filepath, data.table=FALSE)
```

```r
# Data frame of the continuous variables
cols_continuous = c(0,1,7,8,9,10,11,12,13,14,15)
cols_continuous <- cols_continuous+1
df_num <- df[,cols_continuous]

# Validate table for continuous variables
kable(stat.desc(df_num),digits=2)
```

	V1	id	x1	x2	x3	x4	x5	x6	x7	x8	x9
nbr.val	301.00	301.00	301.00	301.00	301.00	301.00	301.00	301.00	301.00	301.00	301.00
nbr.null	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	1.00	1.00	0.67	2.25	0.25	0.00	1.00	0.14	1.30	3.05	2.78
max	301.00	351.00	8.50	9.25	4.50	6.33	7.00	6.14	7.43	10.00	9.25
range	300.00	350.00	7.83	7.00	4.25	6.33	6.00	6.00	6.13	6.95	6.47
sum	45451.00	53143.00	1485.67	1832.50	677.38	921.33	1306.50	657.86	1259.96	1663.65	1617.61
median	151.00	163.00	5.00	6.00	2.12	3.00	4.50	2.00	4.09	5.50	5.42
mean	151.00	176.55	4.94	6.09	2.25	3.06	4.34	2.19	4.19	5.53	5.37
SE.mean	5.02	6.11	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06
CI.mean.0.95	9.87	12.02	0.13	0.13	0.13	0.13	0.15	0.12	0.12	0.11	0.11
var	7575.17	11222.96	1.36	1.39	1.28	1.36	1.67	1.20	1.19	1.03	1.02
std.dev	87.04	105.94	1.17	1.18	1.13	1.16	1.29	1.10	1.09	1.01	1.01
coef.var	0.58	0.60	0.24	0.19	0.50	0.38	0.30	0.50	0.26	0.18	0.19

```r
psych::describe(df_num)
```

```
##      vars   n   mean      sd median trimmed    mad  min    max  range  skew
## V1      1 301 151.00  87.04 151.00  151.00 111.19 1.00 301.00 300.00  0.00
## id      2 301 176.55 105.94 163.00  176.78 140.85 1.00 351.00 350.00 -0.01
## x1      3 301   4.94   1.17   5.00    4.96   1.24 0.67   8.50   7.83 -0.25
## x2      4 301   6.09   1.18   6.00    6.02   1.11 2.25   9.25   7.00  0.47
## x3      5 301   2.25   1.13   2.12    2.20   1.30 0.25   4.50   4.25  0.38
```

```
## x4      6 301   3.06   1.16   3.00   3.02   0.99 0.00   6.33   6.33  0.27
## x5      7 301   4.34   1.29   4.50   4.40   1.48 1.00   7.00   6.00 -0.35
## x6      8 301   2.19   1.10   2.00   2.09   1.06 0.14   6.14   6.00  0.86
## x7      9 301   4.19   1.09   4.09   4.16   1.10 1.30   7.43   6.13  0.25
## x8     10 301   5.53   1.01   5.50   5.49   0.96 3.05  10.00   6.95  0.53
## x9     11 301   5.37   1.01   5.42   5.37   0.99 2.78   9.25   6.47  0.20
##    kurtosis   se
## V1    -1.21 5.02
## id    -1.36 6.11
## x1     0.31 0.07
## x2     0.33 0.07
## x3    -0.91 0.07
## x4     0.08 0.07
## x5    -0.55 0.07
## x6     0.82 0.06
## x7    -0.31 0.06
## x8     1.17 0.06
## x9     0.29 0.06
```

```r
# Data frame of the discrete variables
cols_discrete <- c(2,3,4,5,6)
cols_discrete <- cols_discrete+1
df_cat = df[,cols_discrete]

# Validate tables for discrete variables
Hmisc::describe(df_cat)
```

```
## df_cat
##
##  5  Variables      301  Observations
## --------------------------------------------------------------------------------
## sex
##        n  missing distinct     Info     Mean      Gmd
##      301        0        2    0.749    1.515   0.5012
##
## Value          1     2
## Frequency    146   155
## Proportion 0.485 0.515
## --------------------------------------------------------------------------------
## ageyr
##        n  missing distinct     Info     Mean      Gmd
##      301        0        6    0.907       13    1.123
##
## lowest : 11 12 13 14 15, highest: 12 13 14 15 16
##
## Value         11    12    13    14    15    16
## Frequency      8   101   110    55    20     7
## Proportion 0.027 0.336 0.365 0.183 0.066 0.023
## --------------------------------------------------------------------------------
## agemo
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      301        0       12    0.993    5.375    3.976        0        1
##      .25      .50      .75      .90      .95
##        2        5        8       10       11
```

```
##
## lowest :  0  1  2  3  4, highest:  7  8  9 10 11
##
## Value           0     1     2     3     4     5     6     7     8     9    10
## Frequency      22    31    26    26    27    27    21    25    26    23    19
## Proportion 0.073 0.103 0.086 0.086 0.090 0.090 0.070 0.083 0.086 0.076 0.063
##
## Value          11
## Frequency      28
## Proportion 0.093
## -------------------------------------------------------------------------------
## school
##        n  missing distinct
##      301        0        2
##
## Value      Grant-White     Pasteur
## Frequency          145         156
## Proportion       0.482       0.518
## -------------------------------------------------------------------------------
## grade
##        n  missing distinct     Info     Mean      Gmd
##      300        1        2    0.748    7.477   0.5006
##
## Value           7     8
## Frequency     157   143
## Proportion 0.523 0.477
## -------------------------------------------------------------------------------
```