Validate Statsomat/edapy

```
# 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)
```

Dataset HolzingerSwineford1939.csv

```
# 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

```
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
## ---------------------------------------------------------------------------
```

Dataset Baitingdata.csv

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

```
# Data frame of the continuous variables
cols_continuous = c(9,10,11,12,22,23,24)
cols_continuous <- cols_continuous+1
df_num <- df[,cols_continuous]

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

	1st locate	1st attack	1st attack stop	2nd attack	CBH (cm)	DBH (cm)	height (m)
nbr.val	124.00	116.00	116.00	23.00	160.00	160.00	160.00
nbr.null	0.00	0.00	0.00	0.00	42.00	42.00	0.00

	1st locate	1st attack	1st attack stop	2nd attack	CBH (cm)	DBH (cm)	height (m)
nbr.na	36.00	44.00	44.00	137.00	0.00	0.00	0.00
min	1.00	1.00	58.00	73.00	0.00	0.00	0.40
max	595.00	595.00	600.00	595.00	29.60	9.42	9.00
range	594.00	594.00	542.00	522.00	29.60	9.42	8.60
sum	16317.00	18273.00	58509.00	7650.00	997.56	317.53	463.20
median	100.50	103.50	600.00	319.00	5.30	1.69	2.50
mean	131.59	157.53	504.39	332.61	6.23	1.98	2.90
SE.mean	11.87	13.92	15.72	31.73	0.52	0.16	0.15
CI.mean.0.95	23.51	27.57	31.14	65.81	1.02	0.32	0.30
var	17485.66	22474.43	28668.94	23160.70	42.52	4.31	3.70
std.dev	132.23	149.91	169.32	152.19	6.52	2.08	1.92
coef.var	1.00	0.95	0.34	0.46	1.05	1.05	0.66

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

```
##                 vars   n   mean     sd median trimmed    mad  min    max  range
## 1st locate         1 124 131.59 132.23 100.50  109.45 108.97  1.0 595.00 594.00
## 1st attack         2 116 157.53 149.91 103.50  136.76 122.31  1.0 595.00 594.00
## 1st attack stop    3 116 504.39 169.32 600.00  541.16   0.00 58.0 600.00 542.00
## 2nd attack         4  23 332.61 152.19 319.00  330.42 167.53 73.0 595.00 522.00
## CBH (cm)           5 160   6.23   6.52   5.30    5.11   5.34  0.0  29.60  29.60
## DBH (cm)           6 160   1.98   2.08   1.69    1.63   1.70  0.0   9.42   9.42
## height (m)         7 160   2.90   1.92   2.50    2.62   1.56  0.4   9.00   8.60
##                  skew kurtosis    se
## 1st locate       1.52     2.16 11.87
## 1st attack       1.07     0.27 13.92
## 1st attack stop -1.52     0.85 15.72
## 2nd attack       0.26    -1.08 31.73
## CBH (cm)         1.70     3.22  0.52
## DBH (cm)         1.70     3.22  0.16
## height (m)       1.40     1.99  0.15
```

```
# Data frame of the discrete variables
cols_discrete <- c(0,1,2,3,4,5,6,7,8,13,25,26,27,28,29,30,31,32)
cols_discrete <- cols_discrete+1
df_cat = df[,cols_discrete]

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

```
## df_cat
##
##  18  Variables      160  Observations
## ---------------------------------------------------------------------------
## elevation (m)
##        n  missing distinct     Info     Mean      Gmd
##      160        0        8    0.984     1050    264.2
##
## lowest :  700  800  900 1000 1100, highest: 1000 1100 1200 1300 1400
##
```

```
## Value        700   800   900  1000  1100  1200  1300  1400
## Frequency     20    20    20    20    20    20    20    20
## Proportion 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125
## ---------------------------------------------------------------------------
## date
##        n  missing distinct
##      160        0       29
##
## lowest : 14.08.13 15.08.13 16.08.13 17.08.13 19.08.13
## highest: 22.08.30 22.08.31 22.08.32 24.08.13 26.08.13
## ---------------------------------------------------------------------------
## transect
##        n  missing distinct
##      160        0        3
##
## Value            N       Y Y-not m
## Frequency       58     100       2
## Proportion   0.362   0.625   0.013
## ---------------------------------------------------------------------------
## Tree number
##        n  missing distinct
##      100       60       50
##
## lowest : 1000 - 0157 1000 - 0858 1000 - 0863 1000 - 0864 1100 - 0108
## highest: 700 - 0740  700 - 0778  800 - 0329  900 - 0627  900 - 0628
## ---------------------------------------------------------------------------
## Baiting tree no.
##        n  missing distinct
##      160        0       80
##
## lowest : 0799A 1001  1002  1018  108  , highest: B1002 B1003 B1004 B1005 B1100
## ---------------------------------------------------------------------------
## Termite/C
##        n  missing distinct
##      160        0        2
##
## Value        C   T
## Frequency   80  80
## Proportion 0.5 0.5
## ---------------------------------------------------------------------------
## Detected
##        n  missing distinct     Info      Sum     Mean      Gmd
##      160        0        2    0.523      124    0.775   0.3509
##
## ---------------------------------------------------------------------------
## Attacked
##        n  missing distinct     Info      Sum     Mean      Gmd
##      160        0        2    0.598      116    0.725   0.4013
##
## ---------------------------------------------------------------------------
## Recruited
##        n  missing distinct     Info      Sum     Mean      Gmd
##      160        0        2    0.727       94   0.5875   0.4877
##
```

```
## ------------------------------------------------------------------------------
## 2nd attack stop
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       23      137       10    0.775    520.1    119.6    267.4    319.6
##      .25      .50      .75      .90      .95
##    483.0    600.0    600.0    600.0    600.0
##
## lowest : 242 263 307 370 401, highest: 474 492 501 512 600
##
## Value        242   263   307   370   401   474   492   501   512   600
## Frequency      1     1     1     1     1     1     1     1     1    14
## Proportion 0.043 0.043 0.043 0.043 0.043 0.043 0.043 0.043 0.043 0.609
## ------------------------------------------------------------------------------
## H: 0
##        n  missing distinct     Info     Mean      Gmd
##      150       10        9     0.82     1.44    2.155
##
## lowest :  0  1  2  3  4, highest:  4  5  7  8 12
##
## Value          0     1     2     3     4     5     7     8    12
## Frequency     84    22    12    12     2     8     2     6     2
## Proportion 0.560 0.147 0.080 0.080 0.013 0.053 0.013 0.040 0.013
## ------------------------------------------------------------------------------
## H: 1-5%
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150       10       31    0.993    12.53    15.33     0.00     0.00
##      .25      .50      .75      .90      .95
##     1.00     5.00    20.75    33.00    42.10
##
## lowest :  0  1  2  3  4, highest: 41 43 47 65 71
## ------------------------------------------------------------------------------
## H: 5-33%
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150       10       40    0.998    20.57    17.59     2.45     5.00
##      .25      .50      .75      .90      .95
##     9.00    15.00    24.75    47.00    57.65
##
## lowest :  0  1  2  3  4, highest: 56 59 61 64 80
## ------------------------------------------------------------------------------
## H: 33+%
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150       10       38    0.999    20.13    16.75     4.00     5.00
##      .25      .50      .75      .90      .95
##     9.00    16.00    24.75    41.00    56.55
##
## lowest :  1  3  4  5  6, highest: 56 57 70 73 77
## ------------------------------------------------------------------------------
## ant sample
##        n  missing distinct    value
##       60      100        1        Y
##
## Value       Y
## Frequency  60
## Proportion  1
```

```
## ----------------------------------------------------------------------------
## field notes
##        n  missing distinct
##        8      152        5
##
## lowest : (Jimmy data)                                  *broken
## highest: (Jimmy data)                                  *broken
##
## (Jimmy data) (3, 0.375), *broken (2, 0.250), *broken, 5 @ resprout (1, 0.125),
## few ants (1, 0.125), few ants on lower branches but inhabited above (1, 0.125)
## ----------------------------------------------------------------------------
## species
##        n  missing distinct
##      160        0        5
##
## lowest : ANON001 ANON002 ANON009 ANON012 ANON013
## highest: ANON001 ANON002 ANON009 ANON012 ANON013
##
## Value      ANON001 ANON002 ANON009 ANON012 ANON013
## Frequency        4       6      40      68      42
## Proportion   0.025   0.038   0.250   0.425   0.262
## ----------------------------------------------------------------------------
## lab notes
##        n  missing distinct
##       15      145        8
##
## lowest : 250 misread as 230 so changed       ant ID from main data              Bigger than 802
## highest: Much bigger than B0700              Much bigger than B0805             Small compared to 8
##
## 250 misread as 230 so changed (2, 0.133), ant ID from main data (6, 0.400),
## Bigger than 802 (1, 0.067), Much bigger than B0700 (1, 0.067), Much bigger than
## B0805 (1, 0.067), Small compared to 801 (1, 0.067), Taller petiole than 801 (1,
## 0.067), Vial transferred from Jimmy samples (2, 0.133)
## ----------------------------------------------------------------------------
```