**Descriptive Data Analysis[[1]](#footnote-1) of the juul data set**

**Judy Cushing**

The juul data set is used as an exemplar dataset by P. Dalgaard in his *Introduction to Statistics with R* (ISwR), to teach the R statistical language. The juul data set can be extracted from the ISwR R library, and contains data collected by Anders Juul (Dept. for Growth and Reproduction, Rigshospitalet, xxxwhere) for a study concerning growth factor hormones in a group of 1339 healthy humans (primarily schoolchildren)[[2]](#footnote-2). For this descriptive analysis of the dataset,basic descriptive statistics were run using R (Ver 2.12.2). Table 1 thus summarizes the data.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variable | Type | Range/  values | units | Mean | Median | 1st Q | 3RD Q | Std  dev | var | Missing  (N/A) | comment |
| age | numeric | 0.17..83.0 | years? | 15.1 | 12.6 | 9.1 | 16.9 | 11.3 | 0 | 5, 0% |  |
| Igf1 | numeric | 25..321 | unknown | 340 | 314 | 202 | 463 | 171 | .18 | 321  24% | are these values for subjects >20? |
| testvol | numeric? | 1..30 | unknown | 7.9 | 3.0 | 1.0 | 15.0 | 8.2 | .23 | 859  64% | Noted many missing values. Plot of testvol by sex *plot(igf1~sex)* showed testvol missing for all 713 females; sp 146 (18%) males have missing tesvol. note to self: to explore whether testvol missing for subjects >age20. |
| sex | categorical | 1, 2 recoded  M, F | n/a | M F  621 713  46% 53% | | | |  |  | 5, 0% |  |
| menarch | categorical | 1, 2 recoded  No, Yes | n/a | No Yes  369 335  52% 47% | | | |  |  | 635 47%  Of sex:F:  14 2% | Again, many missing values; most likely for males and females >20, though this should be confirmed. |
| tanner | categorical | 1..5  recoded  I,II,III,IV,V | n/a | I II III IV V  515 103 72 81 328  47% 9% 7% 7% 30% | | | |  |  | 240  18% | Again, are these values for subjects >20? |

**Table 1:** Summary of data in the jull dataset. Range is given for numeric variables and values for categorical variables. Since mean, median, and quartiles are not meaningful for categorical data, counts are given for each category. Summary statistics (mean, median, ec.) were rounded to one decimal place. No units were given for these data, but the author made some assumptions or marked these “unknown”.

Semantics of variables age and sex are self explanatory, but some readers might need clarification for the following:

* menarch: whether females have started menstruating
* igf1: serum IGF-1, an insulin like growth factor
* tanner: Tanner stage of puberty, a classification according to primary and secondary sexual characteristics
* testvol: Testicular volume.

Topic of interest: human growth factors, e.g., IGF-1, an insulin like growth factor. This author’s potential scientific question as she looked at the data set: What is the relationship of igf1 to secondary sexual characteristics (tanner group) and to menarche (in females)? How do these data vary with age ? My initial perusal of the summary data suggested age was unevenly distributed across the data; see Fig.1. Subsequent analysis might separate the data by age class, e.g., <puberty, “puberty years”, >puberty, > postpubescent (i.e., 20). Perhaps variables tanner and igf1 will contribute to setting these cutoffs.

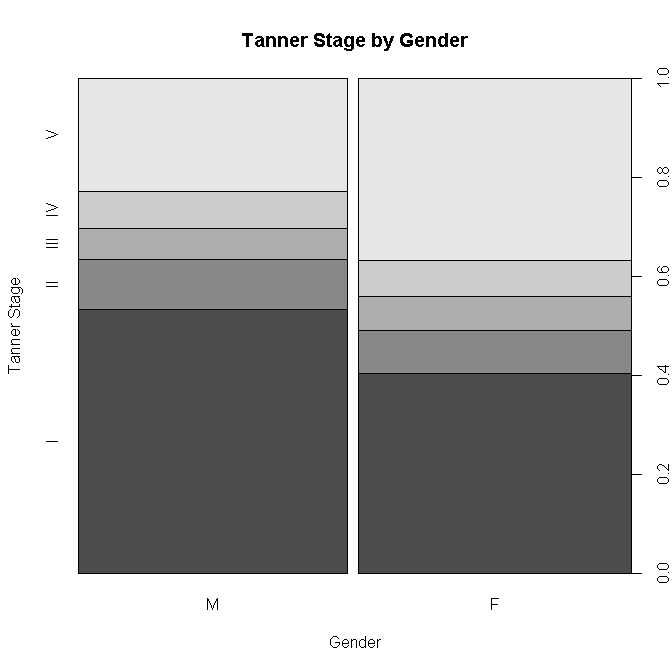
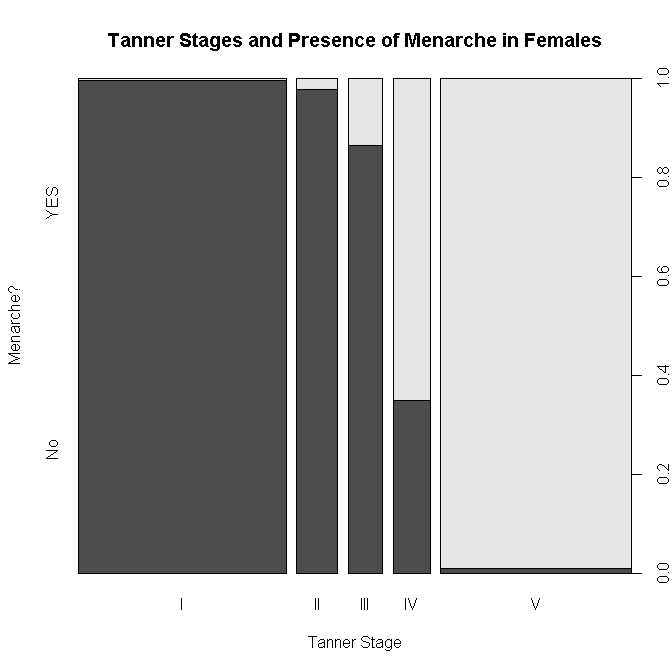
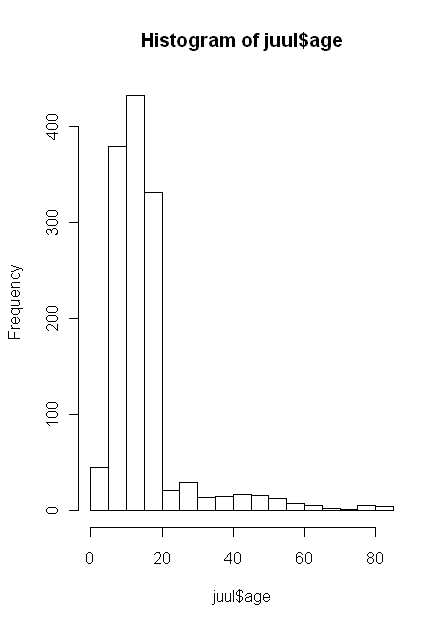


Fig. 1 (left above) Histogram of Age. Shows that most subjects in the study were younger than age 20.

Fig. 2 (center above) Tanner Stage by Menarche. Thanks to Hayduk and Kruse. Shows a more females reach menarche at a higher tanner stage. These two variables are correlated and co-predict. At tanner stage 1 almost no females have reached puberty; at stage V almost all have.

Fig 3. (right above) Tanner Stage by Gender. Thanks to Hayduk and Kruse. This histogram suggests that females mature (reach higher levels of tanner) earlier than males. However, this assumes? that the age distribution for the two groups were similar.

Table 2. Mean igf1 value, reported by tanner group. This suggests that tanner level and igf1 are related, with those subjects having a higher tanner level also having higher igf1 levels. tapply(igf1, tanner, mean, na.rm=T) was the R command used to generate this table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tanner | I | II | III | IV | V |
| IGF1 | 207.5 | 352.7 | 483.2 | 513.0 | 465.3 |

Table 3. Mean age and IGf1 by sex. This suggests that females and males differ by age wrt igf1 levels. perhaps females

aggregate(juul[c("age","igf1")], list(sex=juul$sex), mean, na.rm=T)was the R command used to generate this table.

|  |  |  |
| --- | --- | --- |
| sex | age | igf1 |
| M | 15.38436 | 310.8866 |
| F | 14.84363 | 368.1006 |

One take home lesson from this exercises is that for the data analysis project, a second step (after physically acquiring the data and perusing it) should be to read the literature associated with the data set!

1. Information for this data analysis is gleaned primarily from P. Dalgaard’s *Introductory Statistics with R*, Ch. 4. The data set itself was obtained from the R library ISwR, available for download from any R CRAN site. The one used in this report came from the CRAN site at USA(WA), and was retrieved on April 7, 2011 . [↑](#footnote-ref-1)
2. reference for juul pub here. [↑](#footnote-ref-2)