An Analysis

Jane Doe
Department of Biostatistics
Vanderbilt University School of Medicine

June 28, 2009

Contents

1 Descriptive Statistics

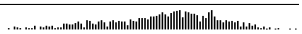
```
getHdata(support)    # Use Hmisc/getHdata to get dataset from VU DataSets wiki
d ← subset(support, select=c(age,sex,race,edu,income,hospdead,slos,dzgroup,
                             meanbp,hrt))
latex(describe(d), file='')
```

d

10 Variables **1000 Observations**

---

age : Age



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	970	62.47	33.76	38.91	51.81	64.90	74.50	81.87	86.00

```
lowest :  18.04  18.41  19.76  20.30  20.31
highest:  95.51  96.02  96.71 100.13 101.85
```

---

sex

n	missing	unique
1000	0	2

```
female (438, 44%), male (562, 56%)
```

---

race

n	missing	unique
995	5	5

	white	black	asian	other	hispanic
Frequency	781	157	9	12	36
%	78	16	1	1	4

---

edu : Years of Education

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
798	202	25	11.78	6	8	10	12	14	16	18

```
lowest :  0  1  2  3  4, highest: 20 21 22 24 30
```

---

income

n	missing	unique
651	349	4

```
under $11k (309, 47%), $11-$25k (161, 25%), $25-$50k (106, 16%)
>$50k (75, 12%)
```

hospdead : Death in Hospital

n	missing	unique	Sum	Mean
1000	0	2	253	0.253

slos : Days from Study Entry to Discharge

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	88	17.86	4	4	6	11	20	37	53

```
lowest :   3   4   5   6   7, highest: 145 164 202 236 241
```

dzgroup

n	missing	unique
1000	0	8

	ARF/MOSF w/Sepsis	COPD	CHF	Cirrhosis	Coma	Colon Cancer	Lung Cancer
Frequency	391	116	143	55	60	49	100
%	39	12	14	6	6	5	10

	MOSF w/Malig
Frequency	86
%	9

meanbp : Mean Arterial Blood Pressure Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	122	84.98	47.00	55.00	64.75	78.00	107.00	120.00	128.05

```
lowest :   0  20  27  30  32, highest: 155 158 161 162 180
```

hrt : Heart Rate Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	124	97.87	54.0	60.0	72.0	100.0	120.0	135.0	146.1

```
lowest :   0  11  30  35  36, highest: 189 193 199 232 300
```

Race is reduced to three levels (white, black, OTHER) because of low frequencies in other levels (minimum relative frequency set to 0.05).

```
d ← upData(d,
           race = combine.levels(race, minlev = 0.05))
```

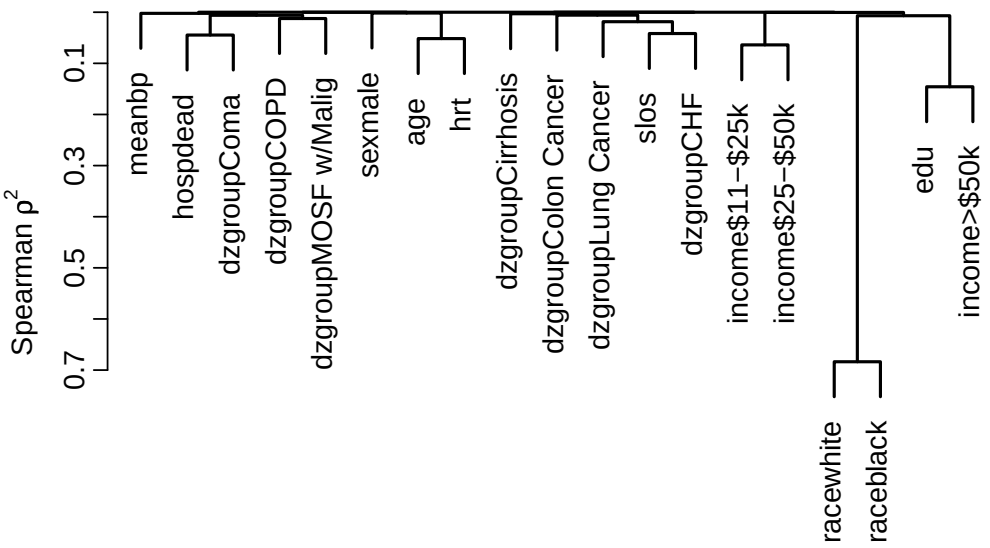
```
Input object size:         84200 bytes;         10 variables
Modified variable         race
New object size:         84112 bytes;         10 variables
```

2 Redundancy Analysis and Variable Interrelationships

```
v ← varclus( ~., data=d)
plot(v)
redun(~age+sex+race+edu+income+dzgroup+meanbp+hrt, data=d)
```

```
Redundancy Analysis

redun(formula = ~age + sex + race + edu + income + dzgroup +
    meanbp + hrt, data = d)

n: 617         p: 8         nk: 3

Number of NAs:         383
Frequencies of Missing Values Due to Each Variable
    age     sex    race     edu  income dzgroup  meanbp     hrt
      0       0       5     202     349       0       0       0
```

```
Transformation of target variables forced to be linear

R-squared cutoff: 0.9         Type: ordinary

R^2 with which each variable can be predicted from all other variables:

    age     sex    race     edu  income dzgroup  meanbp     hrt
  0.196   0.088   0.120   0.284   0.339   0.253   0.067   0.163

No redundant variables
```



Note that the clustering of black with white is not interesting; this just means that these are mutually exclusive higher frequency categories, causing them to be negatively correlated.

3 Logistic Regression Model

Here we fit a tentative binary logistic regression model. The coefficients are not very useful so they are not printed.

```
dd ← datadist(d); options(datadist='dd')
f ← lrm(hospdead ~ rcs(age,4) + sex + race + dzgroup + rcs(meanbp,5),
        data=d)
f
```

```
Logistic Regression Model

lrm(formula = hospdead ~ rcs(age, 4) + sex + race + dzgroup +
    rcs(meanbp, 5), data = d)


Frequencies of Responses
  0   1
```

```
744 251

Frequencies of Missing Values Due to Each Variable
hospdead      age      sex     race  dzgroup   meanbp
       0        0        0        5        0        0

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy
       995      1e-09     245.83         17          0        0.8      0.601
     Gamma      Tau-a         R2      Brier
     0.602      0.227      0.323      0.144

...
```

```
latex(anova(f), where='h', file='')    # can also try where='htbp'
```

Table 1: Wald Statistics for `hospdead`

	χ^2	$d.f.$	P
age	7.12	3	0.0683
Nonlinear	2.91	2	0.2338
sex	2.16	1	0.1413
race	1.38	2	0.5005
dzgroup	78.77	7	< 0.0001
meanbp	65.62	4	< 0.0001
Nonlinear	48.11	3	< 0.0001
TOTAL NONLINEAR	50.15	5	< 0.0001
TOTAL	151.71	17	< 0.0001

4 Test Calculations

```
x ← 3; y ← 2
if(x <= y) 'this' else 'that'
```

```
[1] "that"
```

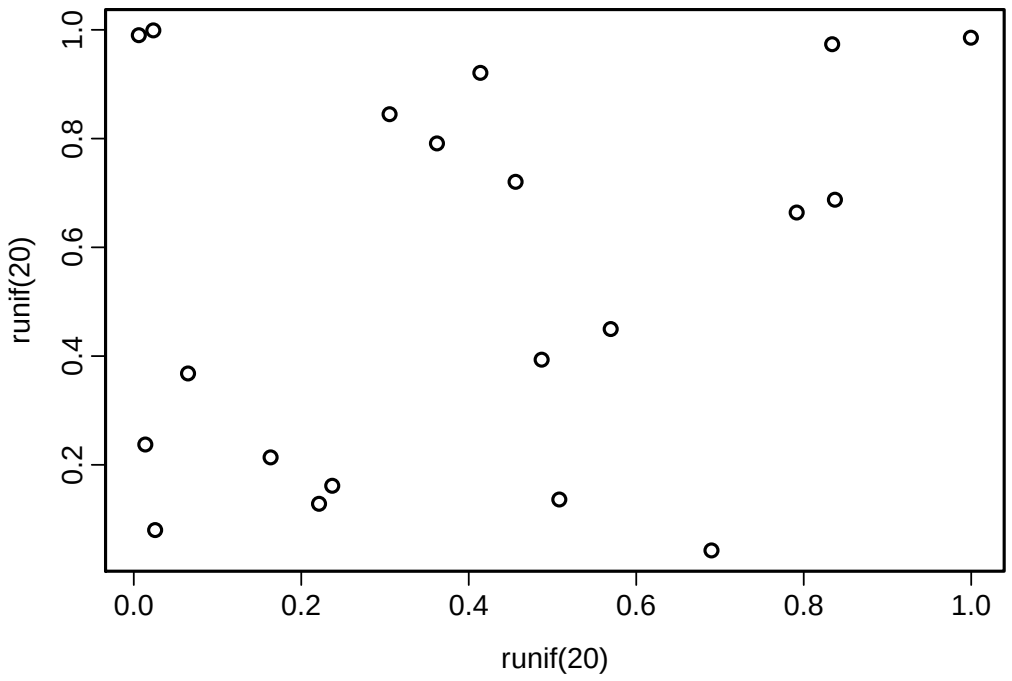
```
if(y >= x) 'that' else 'this'
```

```
[1] "this"
```

```
x^y
```

```
[1] 9
```

```
plot(runif(20),runif(20))
```



5 Computing Environment

These analyses were done using the following versions of R[1], the operating system, and add-on packages `Hmisc`[2], `Design`[3], and others:

- R version 2.9.0 (2009-04-17), `i486-pc-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: Design 2.1-2, Hmisc 3.4-4, survival 2.35-3
- Loaded via a namespace (and not attached): cluster 1.11.13, grid 2.9.0, lattice 0.17-22

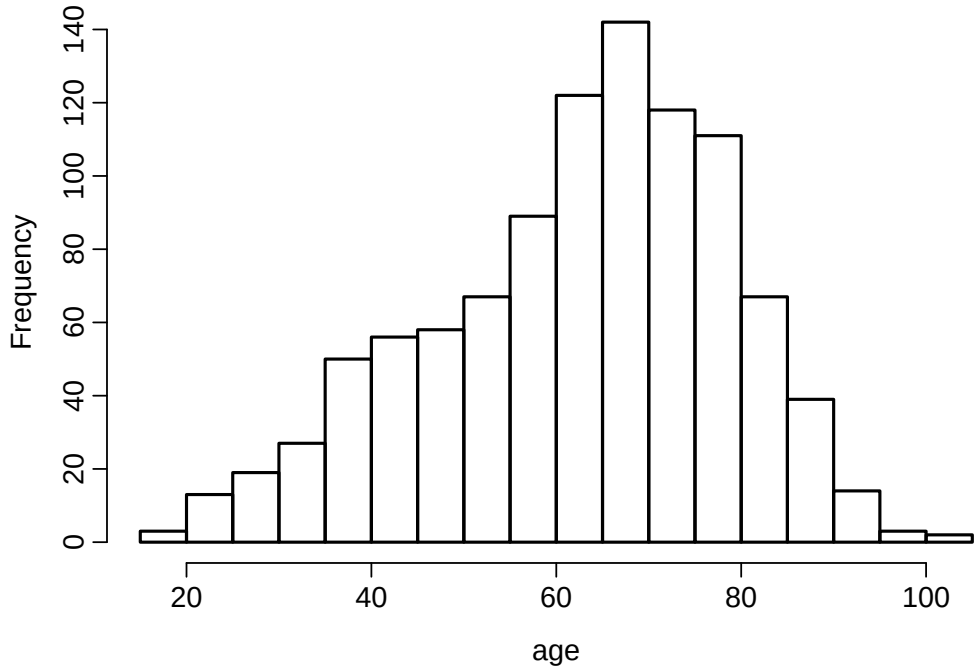
6 Descriptive Statistics Again

```
getHdata(support)
d ← subset(support, select = c(age, sex, race, edu, income,
    hospdead, slos, dzgroup, meanbp, hrt))
summary(d)
```

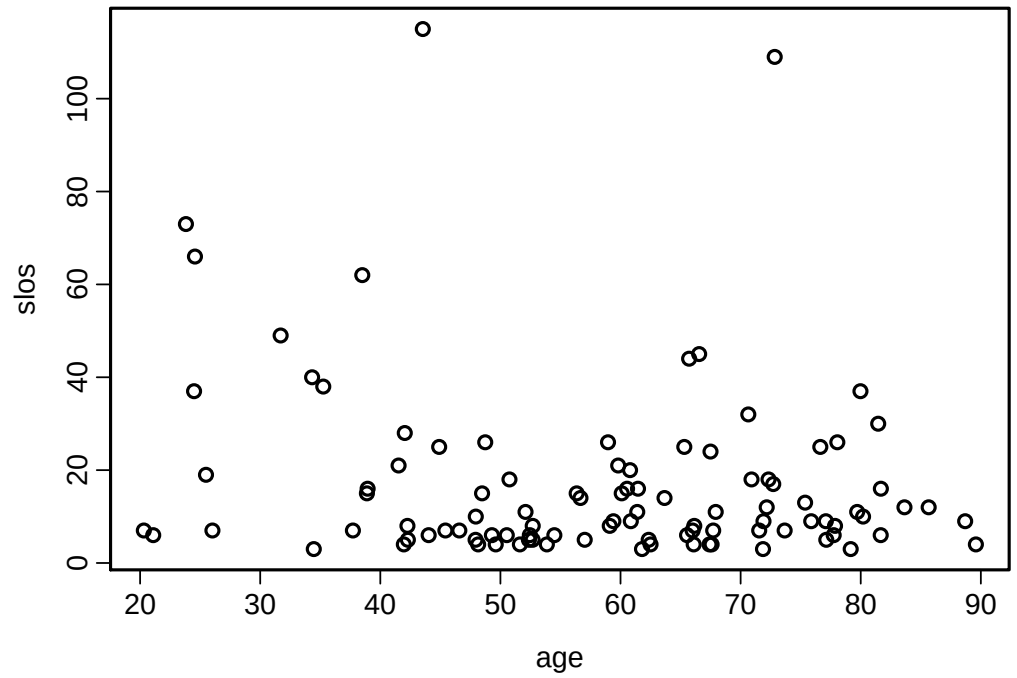
```
      age              sex            race          edu               income
 Min.   : 18.04   female:438   white   :781   Min.   :  0.00   under $11k:309
 1st Qu.: 51.81   male  :562   black   :157   1st Qu.: 10.00   $11-$25k  :161
 Median : 64.90                asian   :  9   Median : 12.00   $25-$50k  :106
 Mean   : 62.47                other   : 12   Mean   : 11.78   >$50k     : 75
 3rd Qu.: 74.50                hispanic: 36   3rd Qu.: 14.00   NA's      :349
 Max.   :101.85                NA's    :  5   Max.   : 30.00
                                              NA's   :202.00

   hospdead          slos                      dzgroup        meanbp
 Min.   :0.000   Min.   :  3.00   ARF/MOSF w/Sepsis:391   Min.   :  0.00
 1st Qu.:0.000   1st Qu.:  6.00   CHF              :143   1st Qu.: 64.75
 Median :0.000   Median : 11.00   COPD             :116   Median : 78.00
 Mean   :0.253   Mean   : 17.86   Lung Cancer      :100   Mean   : 84.98
 3rd Qu.:1.000   3rd Qu.: 20.00   MOSF w/Malig     : 86   3rd Qu.:107.00
 Max.   :1.000   Max.   :241.00   Coma             : 60   Max.   :180.00
                                  (Other)          :104

     hrt
```

```
Min.    :  0.00
1st Qu.: 72.00
Median :100.00
Mean   : 97.87
3rd Qu.:120.00
Max.   :300.00
```

```
with(d, hist(age, nclass = 25, main = ""))
```




```
with(d[1:100, ], plot(age, slos))
```



That was a very concise set of descriptive statistics.

References

[1] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0, available from `www.R-project.org`.

[2] Frank E. Harrell. `Hmisc`: A library of miscellaneous S functions. Available from `biostat.mc.vanderbilt.edu/s/Hmisc`, 2009.

[3] Frank E. Harrell. `Design`: S functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Available from `biostat.mc.vanderbilt.edu/s/Design`, 2009.