Analyse des esoph Datensatzes mit RStudio

Erster Schritt

Datensatz in Data Frame einlesen.

```
df <- esoph
```

Zweiter Schritt

Struktur des Datensatzes ausgeben.

```
str(df)
```

```
## 'data.frame':    88 obs. of  5 variables:
##  $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...
```

Dritter Schritt

Zusammenfassung des Datensatzes ausgeben.

```
summary(df)
```

```
##    agegp          alcgp         tobgp        ncases         ncontrols
##  25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
##  35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
##  45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
##  55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
##  65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
##  75+  :11                                Max.   :17.000   Max.   :60.00
```

Analyse des Datensatzes

Allgemeine Infos des Datensatzes (mean, median, var) ausgeben.

```
mean(df$ncases)
```

```
## [1] 2.272727
```

```
mean(df$ncontrols)
```

```
## [1] 11.07955
```

```
median(df$ncases)
```

```
## [1] 1
```

```
median(df$ncontrols)
```

```
## [1] 6
```

```r
var(df$ncases)
```

```
## [1] 7.579937
```

```r
var(df$ncontrols)
```

```
## [1] 161.8672
```

Eine beschränkte Zusammenfassung erstellen.

```r
summary(df[c("ncases", "ncontrols", "agegp")])
```

```
##      ncases          ncontrols       agegp
##  Min.   : 0.000   Min.   : 1.00   25-34:15
##  1st Qu.: 0.000   1st Qu.: 3.00   35-44:15
##  Median : 1.000   Median : 6.00   45-54:16
##  Mean   : 2.273   Mean   :11.08   55-64:16
##  3rd Qu.: 4.000   3rd Qu.:14.00   65-74:15
##  Max.   :17.000   Max.   :60.00   75+  :11
```

Konsum von Alkohol und Tabak in Abhänigkeit von Alter und Krebs-Erkrankung (Exemplarisch an zwei Altersstufen).

```r
df.sub0 <- subset(df, agegp == "25-34" & ncases > 0)
df.sub0
```

```
##    agegp alcgp tobgp ncases ncontrols
## 13 25-34  120+ 10-19      1         1
```

```r
df.sub1 <- subset(df, agegp == "65-74" & ncases > 0)
df.sub1
```

```
##    agegp     alcgp    tobgp ncases ncontrols
## 63 65-74 0-39g/day 0-9g/day      5        48
## 64 65-74 0-39g/day    10-19      4        14
## 65 65-74 0-39g/day    20-29      2         7
## 67 65-74     40-79 0-9g/day     17        34
## 68 65-74     40-79    10-19      3        10
## 69 65-74     40-79    20-29      5         9
## 70 65-74    80-119 0-9g/day      6        13
## 71 65-74    80-119    10-19      4        12
## 72 65-74    80-119    20-29      2         3
## 73 65-74    80-119      30+      1         1
## 74 65-74      120+ 0-9g/day      3         4
## 75 65-74      120+    10-19      1         2
## 76 65-74      120+    20-29      1         1
## 77 65-74      120+      30+      1         1
```

Wie viele Erkrankungen gibt es bei der jüngsten Altersstufe in Verbindung mit gesundem Lebensstil?

```r
df.sub2 <- subset(df, agegp == "25-34" & alcgp == "0-39g/day" & tobgp == "0-9g/day")
df.sub2
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
```

Bei der ältesten Altersgruppe in Verbindung mit gesundem Lebensstil?

```r
df.sub3 <- subset(df, agegp == "75+" & alcgp == "0-39g/day" & tobgp == "0-9g/day")
df.sub3
```

```
##    agegp     alcgp    tobgp ncases ncontrols
## 78   75+ 0-39g/day 0-9g/day      1        18
```

Eingrenzung von Daten mithilfe der select Funktion.

```
df.sub4 <- subset(df, agegp == "35-44" & ncases > 0, select = c(agegp, alcgp, tobgp, ncases))
df.sub4
```

```
##    agegp     alcgp    tobgp ncases
## 17 35-44 0-39g/day    10-19      1
## 21 35-44     40-79    10-19      3
## 22 35-44     40-79    20-29      1
## 28 35-44      120+ 0-9g/day      2
## 30 35-44      120+    20-29      2
```

Krebserkrankungen bei gesundem Lebensstil über alle Altersgruppen.

```
df.sub5 <- subset(df, alcgp == "0-39g/day" & tobgp == "0-9g/day" & ncases > 0, select = c(agegp, alcgp,
df.sub5
```

```
##    agegp     alcgp    tobgp ncases
## 31 45-54 0-39g/day 0-9g/day      1
## 47 55-64 0-39g/day 0-9g/day      2
## 63 65-74 0-39g/day 0-9g/day      5
## 78   75+ 0-39g/day 0-9g/day      1
```

Krebserkrankungen bei sehr ungesundem Lebensstil über alle Altersgruppen.

```
df.sub6 <- subset(df, alcgp == "120+" & tobgp == "30+" & ncases > 0, select = c(agegp, alcgp, tobgp, nca
df.sub6
```

```
##    agegp alcgp tobgp ncases
## 46 45-54  120+   30+      4
## 62 55-64  120+   30+      5
## 77 65-74  120+   30+      1
```

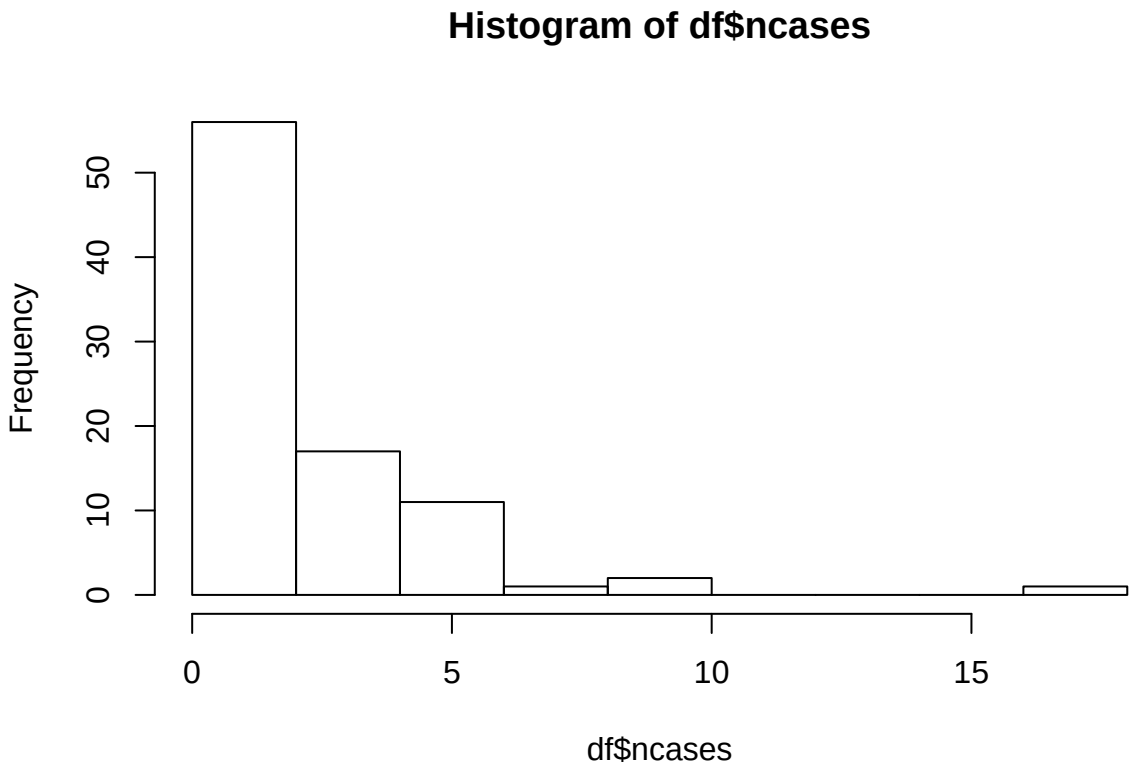
Zusammenfassung

Die Betrachtung der Kontrollen in Bezug zu den Erkrankungen müßte auch untersucht werden. Weiterhin könnte man Summen, Maximal-Werte und Minimal-Werte bilden. Der Analyse des Datensatzes sind hier kaum Grenzen gesetzt.

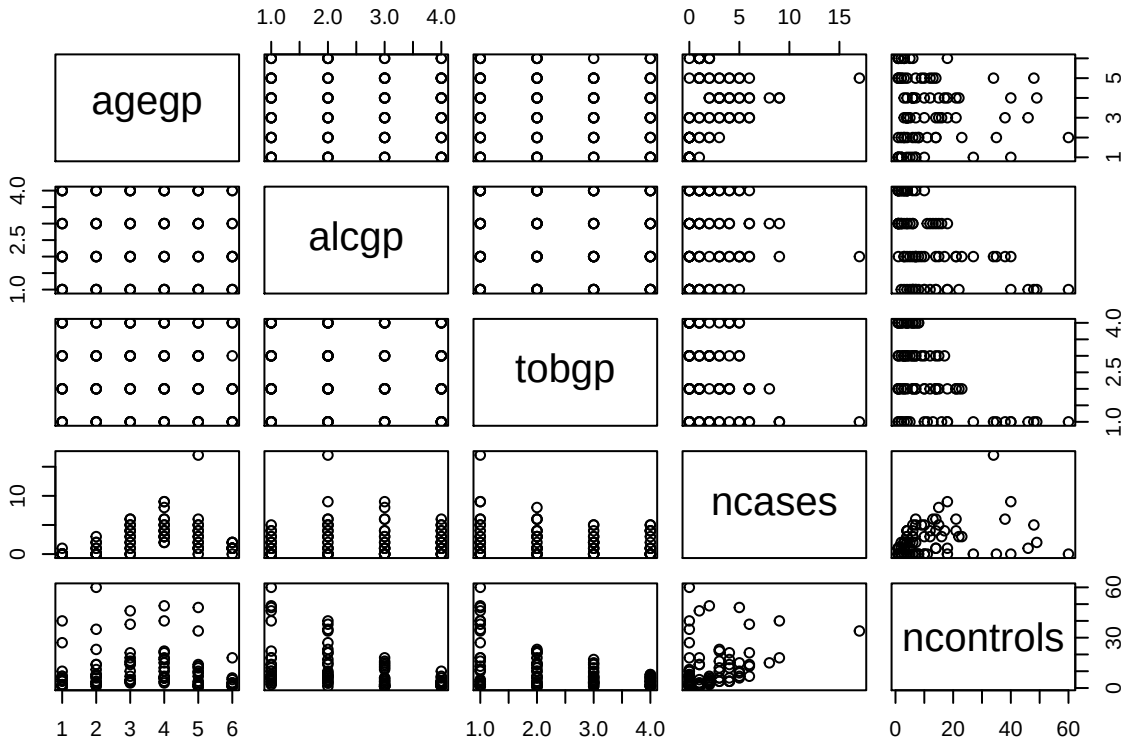
Plots

Zum Abschluss werden noch ein paar Plots gezeigt.

```
hist(df$ncases)
```


Histogram of df$ncases
Frequency
0
10
20
30
40
50
0
5
10
15
df$ncases

```
plot(df)
```


agegp
alcgp
tobgp
ncases
ncontrols