

Exercise 2 (R)

Analyse the esoph dataset. Can you derive some useful statements from it? Use `data()` to see all available datasets.

In [1]:

```
1 #Zuerst muss das esoph dataset geladen werden
2 data("esoph")
3
4 #Jetzt muss ich wissen was für einen Datentyp das Dataset hat
5 class(esoph)
```

'data.frame'

In [2]:

```
1 #Das esoph-Dataset ist ein data.frame. Ich kann dann die Größe/Dimension ermitteln.
2 dim(esoph) #Das esoph-Dataset hat 88 Zeilen und 5 Spalten
```

88 5

In [3]:

```
1 #Jetzt werden die Namen der Spalten angezeigt
2 colnames(esoph) #[1] "agegp"      "alcgp"      "tobgp"      "ncases"      "ncontrols"
```

'agegp' 'alcgp' 'tobgp' 'ncases' 'ncontrols'

Um mehr Informationen über die Daten zu haben, kann man auch die Funktion: `str(esoph)`. Das Ergebnis zeigt, dass die zwei letzten Spalten numerische Werte haben, während die 3 ersten Spalten Strings enthalten.

In [14]:

```
1 str(esoph)
```

```
'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 ...
```

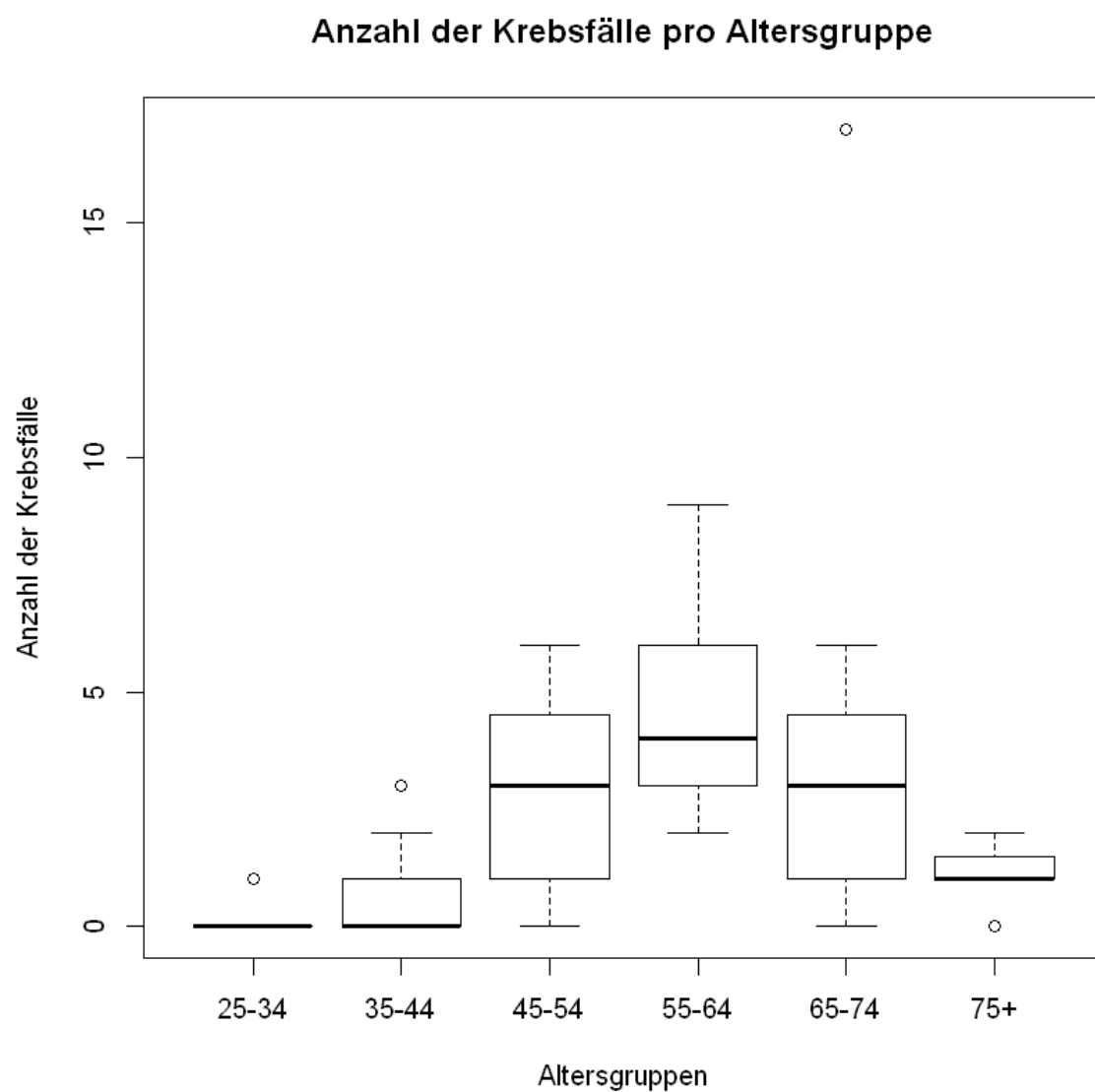
Ich kann die Anzahl der Krebsfälle pro Altersgruppe als Chart anzeigen. Das Ergebnis zeigt, dass die meisten

Fälle in dem Bereich zwischen 45 und 74 Jahren zu finden sind.

In [11]:



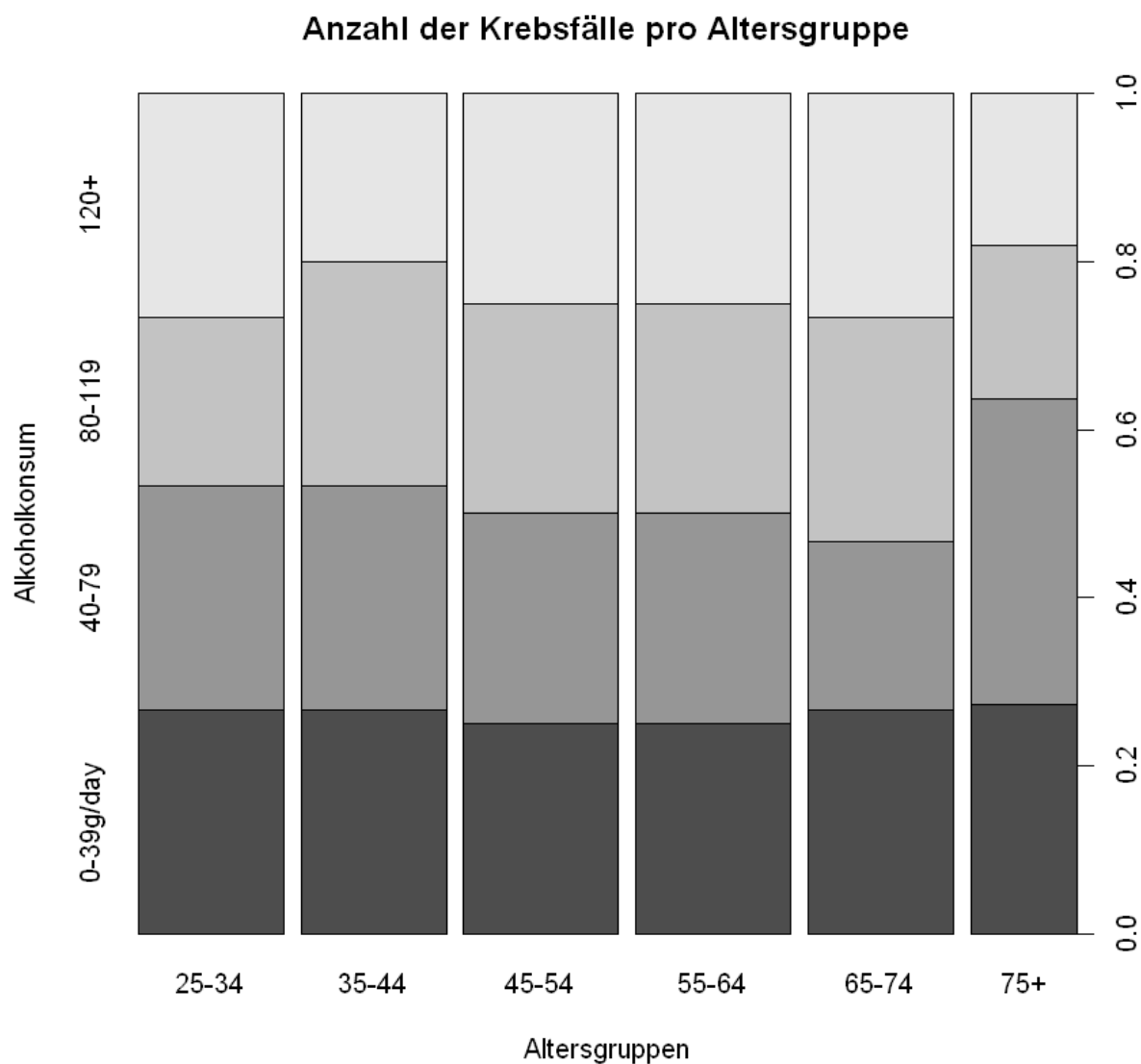
```
1 #Jetzt haben wir die Namen der Spalten und ich kann als Chart anzeigen wie viele
2 #Krebsfälle es pro Altersgruppe gibt.
3 plot(esoph$agegp, esoph$ncases, main="Anzahl der Krebsfälle pro Altersgruppe",
4       xlab="Altersgruppen", ylab="Anzahl der Krebsfälle")
```



In [10]:



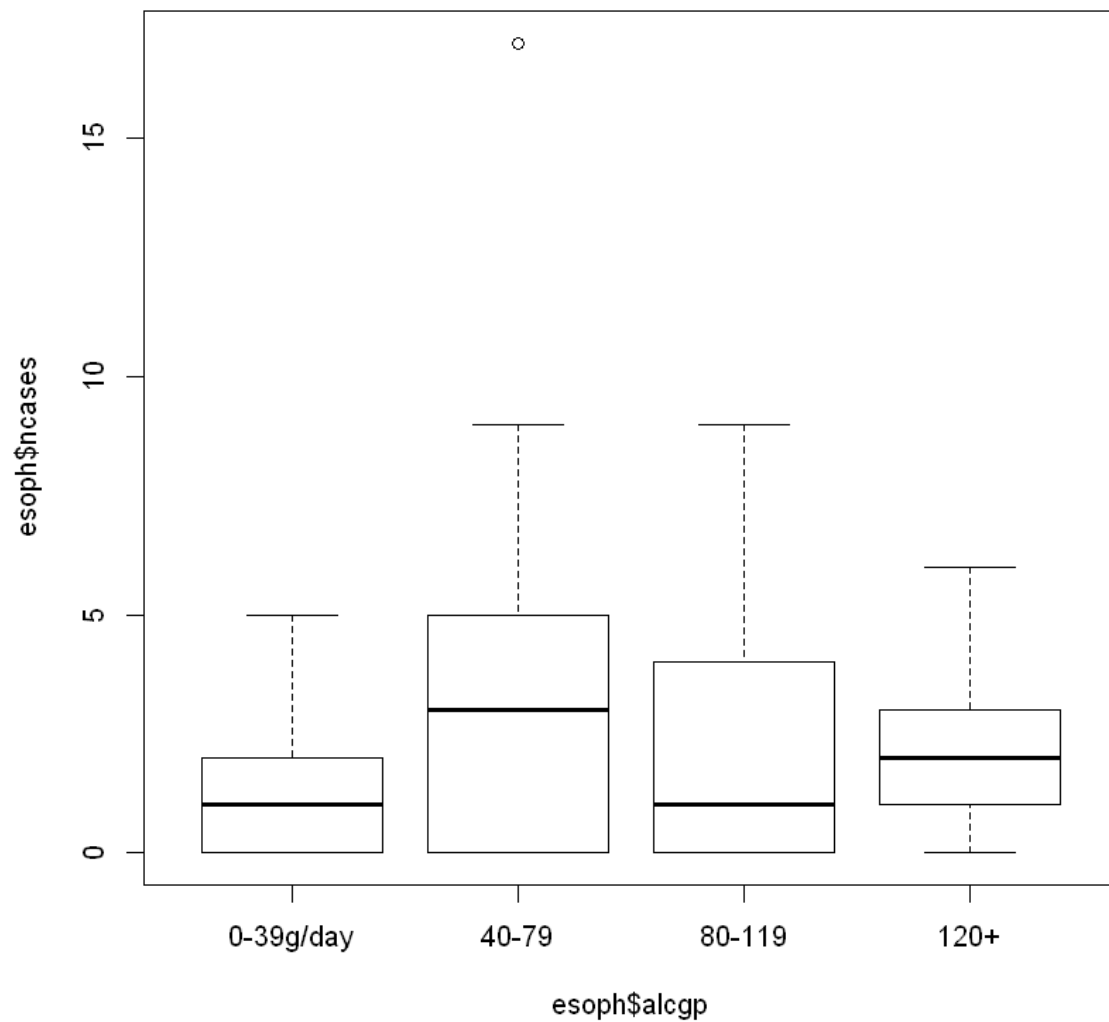
```
1 plot(esoph$agegp, esoph$alcgp, main="Anzahl der Krebsfälle pro Altersgruppe",  
2      xlab="Altersgruppen", ylab="Alkoholkonsum")
```



In [6]:



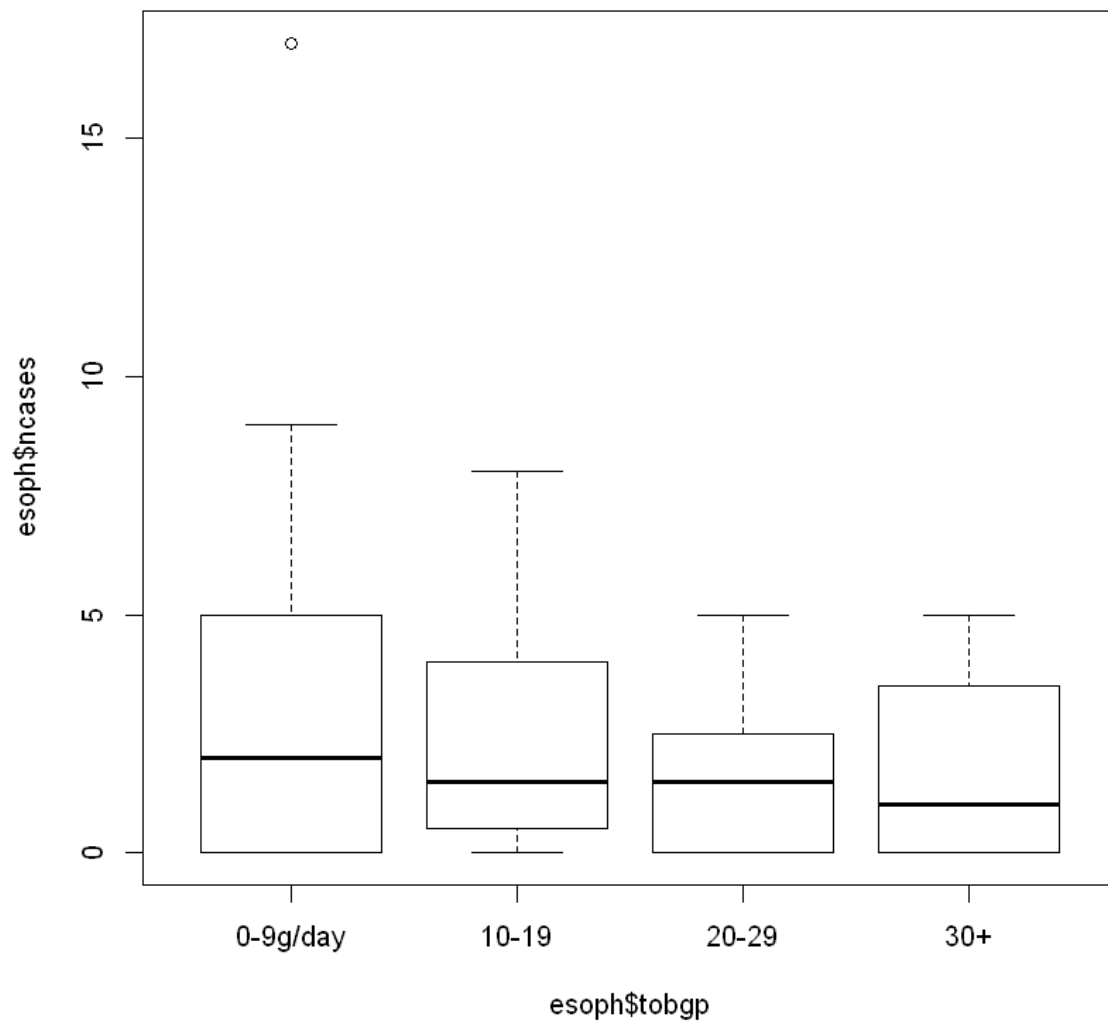
```
1 #Ich kann dann die Anzahl der Krebsfällen im Verhältnis zu dem Alkoholkonsum  
2 boxplot(esoph$ncases ~ esoph$alcgp)
```



In [7]:



```
1 boxplot(esoph$ncases ~ esoph$tobgp)
```



In R gibt es eine Funktion „summary()“, die statistischen Eigenschaften der Spalten berechnet.

In [13]:



```
1 summary(esoph)
```

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

