# Mathematics for Data Scientists – esoph dataset analysis

KIESGEN de RICHTER Stanislas – VAIO Luca

06/11/2020

## Discovering the dataset

Our dataset comes from a case-control study of œsophageal cancer conducted in Ille-et-Vilaine, France. The dataset is composed of three factors: age groups, alcohol consumption and tobacco consumption. The number of people controlled and proven cases of cancer are given for each group combination. The aim of the study is to confirm the correlation between cancers and age, as well as consumption of alcohol and of tobacco.

To begin with, let's explore the esoph dataset. With R, we can display the first 8 lines, and its dimensions.

```
head(esoph, 8)
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
## 2 25-34 0-39g/day    10-19      0        10
## 3 25-34 0-39g/day    20-29      0         6
## 4 25-34 0-39g/day      30+      0         5
## 5 25-34     40-79 0-9g/day      0        27
## 6 25-34     40-79    10-19      0         7
## 7 25-34     40-79    20-29      0         4
## 8 25-34     40-79      30+      0         7
```

```
cat("Number of combinations: ", dim(esoph)[1], "\nNumber of persons controlled: ",
    sum(esoph$ncontrols))
```

```
## Number of combinations:  88
## Number of persons controlled:  975
```

As we can see, the study takes into account the age group of each individual, along with their alcohol and tobacco consumption per day. The age groups start at 25 and are divided by ranges of 10 years until 75+, the alcohol by ranges of 40g/day and the tobacco by ranges of 10g/day. In the dataset, each combination of the three groups is presented with the corresponding number of controls and proven cases.

R also provides a summary of the dataset. In this summary below, the first three columns only represent the number of lines of `agegp`, `alcgp` and `tobgp`, which are respectively age, alcohol and tobacco. We cannot get any valuable information from this, because these variables are qualitative. The two other variables, however, are quantitative, so we can analyse the given statistics.

```
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
```

The number of cases goes from 0 to 17 by group combination. The maximum seems to be quite far from the median, and even the third quartile, which is 4. This means that there is a peak in one of the combinations. As we see in the last column, the controls are not equally distributed: they vary from 1 to 60 depending on the group combination. It means that the combinations of the three groups (agegp, alcgp, tobgp) are not equally distributed.

```
tapply(esoph$ncontrols, esoph$agegp, sum)
```

```
## 25-34 35-44 45-54 55-64 65-74   75+
##   116   199   213   242   161    44
```

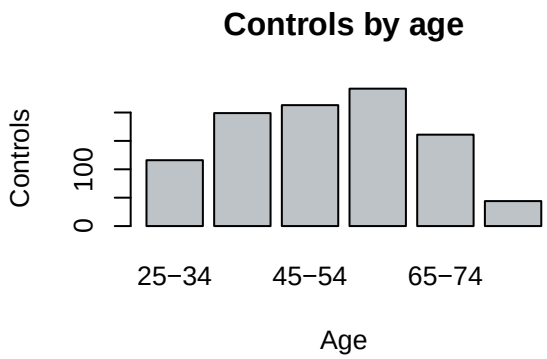
In fact, we can see using the `tapply` R function, which is a `group_by`, that for instance there are 242 controls for the age group `55-64` in the dataset, while there are only 44 for the group `75+`.

On the basis of these observations, we can conclude that we will not be able to base our calculations on the numbers of cases and controls, in particular with regard to the computation of correlation indices between the different variables.

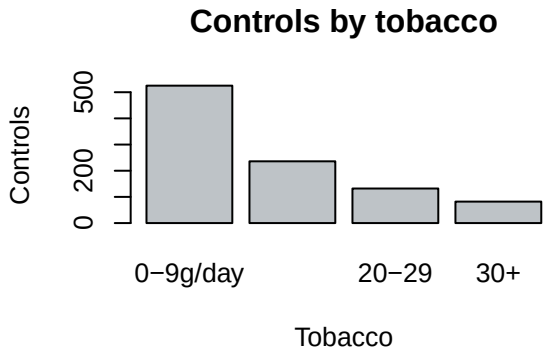
## Visualising the data

As said before, the major issue in this dataset is that the quantitative variables are not equally distributed. The following bar plots permit to visualise this well, using as previously the `tapply` function.

```
par(mfrow=c(2, 2))
barplot(tapply(esoph$ncontrols, esoph$agegp, sum), col="#BEC3C7",
        xlab="Age", ylab="Controls", main="Controls by age")
barplot(tapply(esoph$ncontrols, esoph$alcgp, sum), col="#BEC3C7",
        xlab="Alcohol", ylab="Controls", main="Controls by alcohol")
barplot(tapply(esoph$ncontrols, esoph$tobgp, sum), col="#BEC3C7",
        xlab="Tobacco", ylab="Controls", main="Controls by tobacco")
```





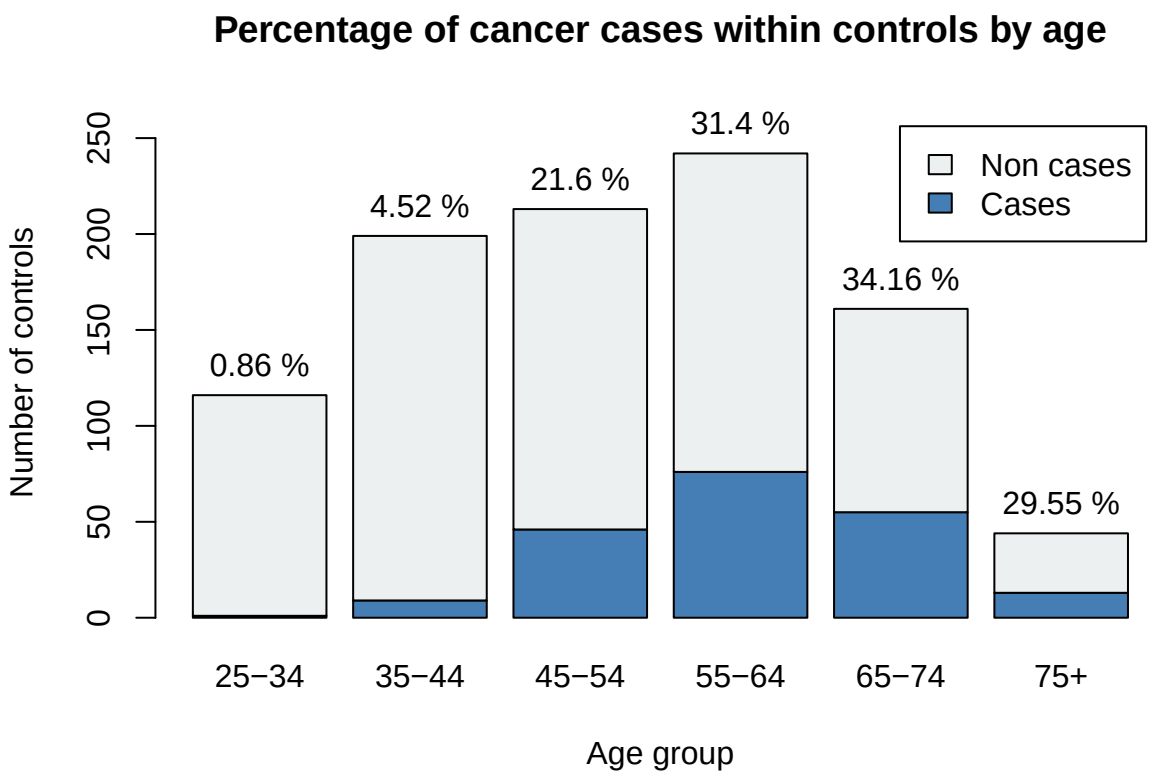



We clearly see that controls not evenly distributed. Note that for alcohol and tobacco, most controlled people present a low consumption, which shows once again that we absolutely cannot overlook the differences in the number of measurements.

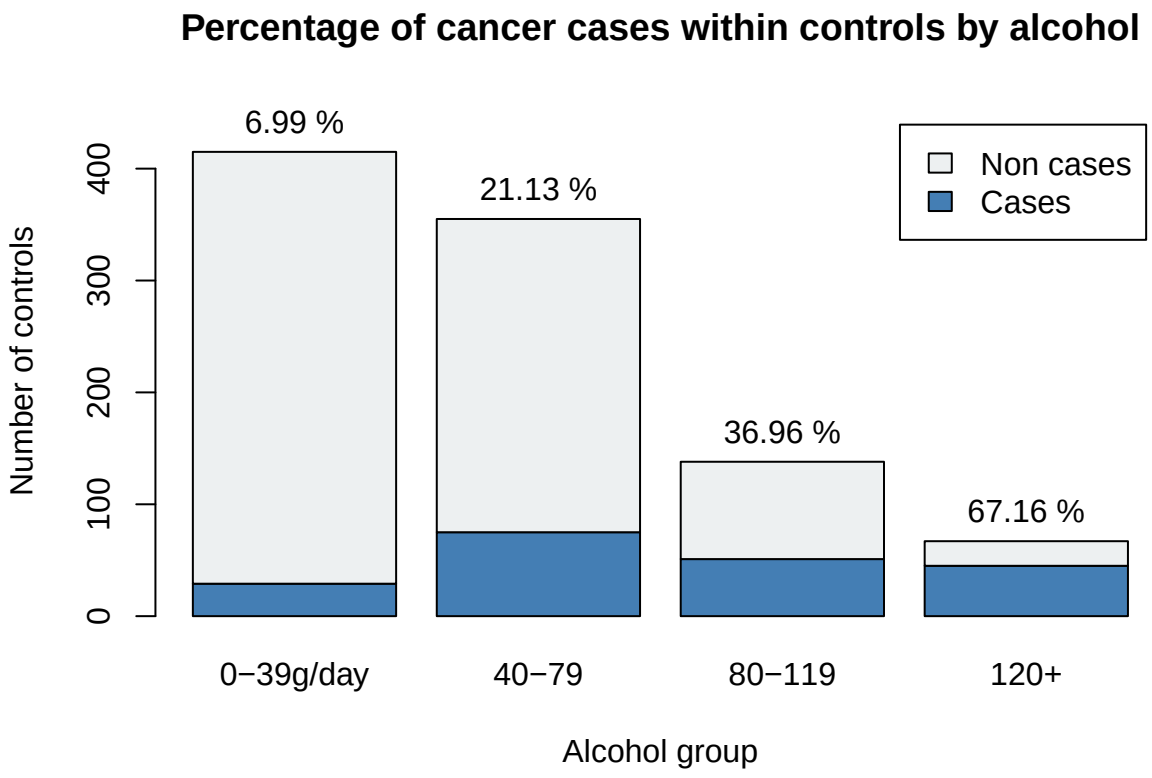
A certainly more reasonable way of analysing this dataset is to base our calculations on the proportion of cancer cases according to the number of controls for each combination of data. A first look at these proportions is given below, by age, alcohol and tobacco.

```
# Age
age_cases <- tapply(esoph$ncases, esoph$agegp, sum)
age_controls <- tapply(esoph$ncontrols, esoph$agegp, sum)
non_age_cases <- age_controls - age_cases
age_proportions <- age_cases / age_controls * 100
age_ylim <- c(0, 1.1 * max(age_controls))

# Alcohol
alc_cases <- tapply(esoph$ncases, esoph$alcgp, sum)
alc_controls <- tapply(esoph$ncontrols, esoph$alcgp, sum)
non_alc_cases <- alc_controls - alc_cases
alc_proportions <- alc_cases / alc_controls * 100
alc_ylim <- c(0, 1.1 * max(alc_controls))

# Tobacco
tob_cases <- tapply(esoph$ncases, esoph$tobgp, sum)
tob_controls <- tapply(esoph$ncontrols, esoph$tobgp, sum)
non_tob_cases <- tob_controls - tob_cases
tob_proportions <- tob_cases / tob_controls * 100
```

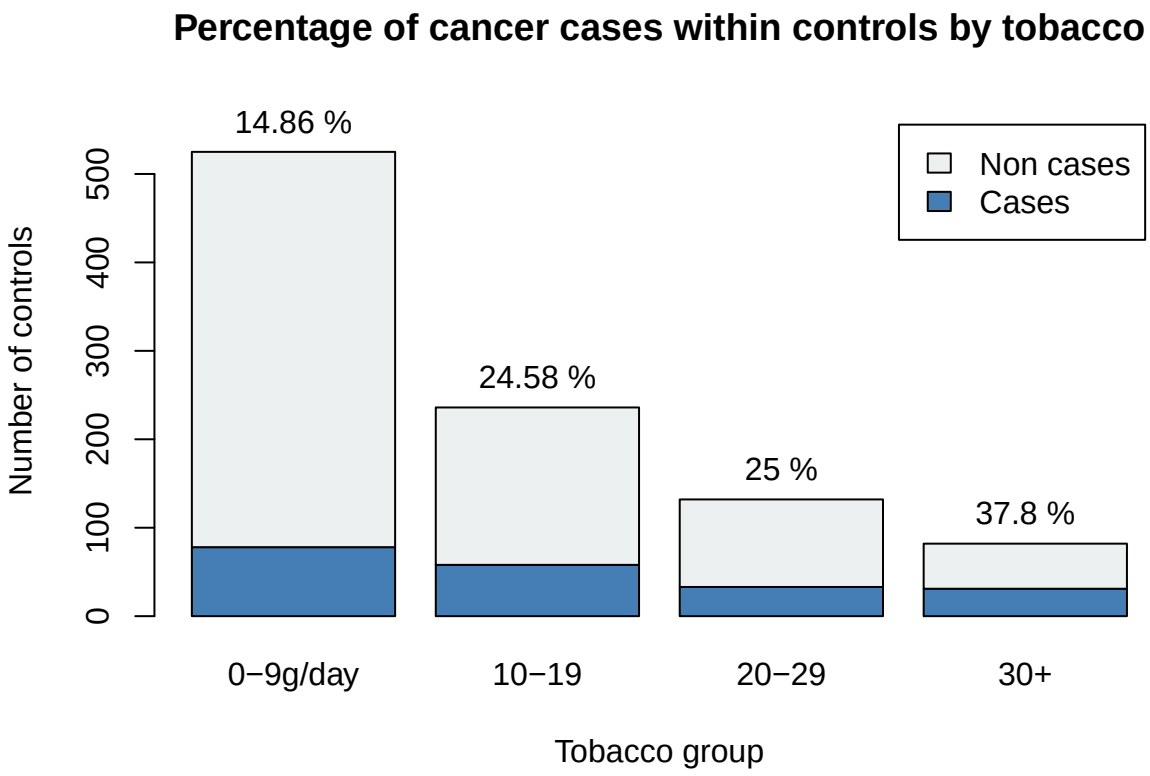
```r
tob_ylim <- c(0, 1.1 * max(tob_controls))

# Plots
legend <- c("Cases", "Non cases")
colours <- c("#447EB4", "#EDF0F1")
ylab <- "Number of controls"

age_plot <- barplot(rbind(age_cases, non_age_cases), ylim=age_ylim, legend=legend,
                col=colours, xlab="Age group", ylab=ylab,
                main="Percentage of cancer cases within controls by age")
text(x=age_plot, y=age_controls, label=paste(round(age_proportions, 2), "%"), pos=3)
```




```r
alc_plot <- barplot(rbind(alc_cases, non_alc_cases), ylim=alc_ylim, legend=legend,
                col=colours, xlab="Alcohol group", ylab=ylab,
                main="Percentage of cancer cases within controls by alcohol")
text(x=alc_plot, y=alc_controls, label=paste(round(alc_proportions, 2), "%"), pos=3)
```




```
tob_plot <- barplot(rbind(tob_cases, non_tob_cases), ylim=tob_ylim, legend=legend,
                col=colours, xlab="Tobacco group", ylab=ylab,
                main="Percentage of cancer cases within controls by tobacco")
text(x=tob_plot, y=tob_controls, label=paste(round(tob_proportions, 2), "%"), pos=3)
```




These three bar plots show, for each group of the concerned factor, the proportion of positive and negative cases within the overall controls. As we saw, and as we still see on these plots, the controls are not equitably distributed, but looking at the proportion of cases within those controls is more consistent for our data analysis. In fact, what matters is not how many controls have been done for a given age group for example, but indeed the percentage of positive cancer cases measured. For instance, the alcohol group `0-39` has 415 controls, but only 6.99 of those controlled people are positive. In parallel, the group `120+` has much less controls (67), but actually a much higher proportion of those are positive, 67.16, which means that this alcohol group is way more likely to present cancer cases that the previous group.

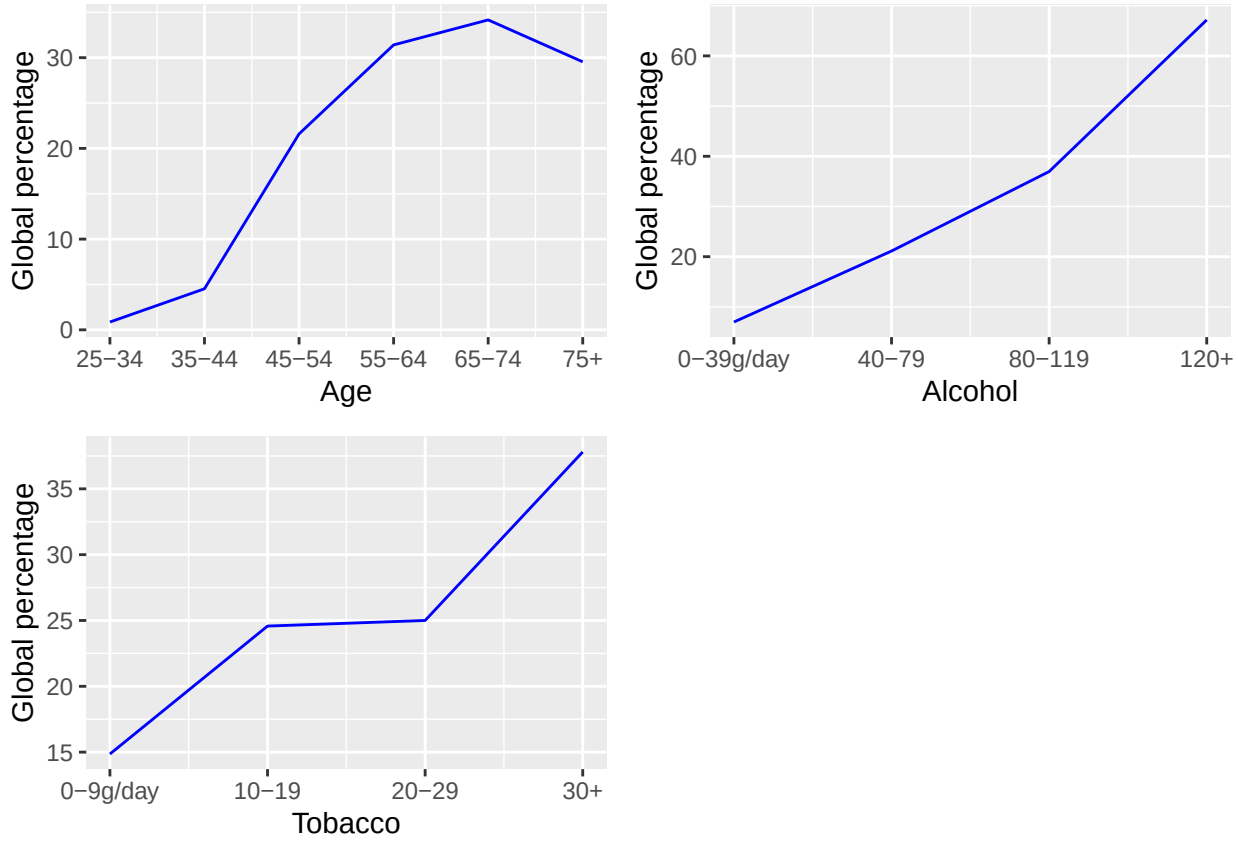
To finish with, case proportions seem to be more consistent to conduct further analysis, and we will base our computations on this model. The only drawback to this system is that with a different number of controls, the data collected, and therefore the proportions calculated at the same time, will be of very different accuracy. To come back to the previous example, the `120+` group does have 67.16 positive cases, but this proportion could well be revised downwards following further measurements. The `0-39` group has many more measurements and therefore presents more accurate data, which can therefore be relied on more confidently.

## Hypothesis testing

### State hypothesis

We have seen that, in order to carry out our analysis, it is preferable for us to work with proportions of positive cases. Now, let's try to plot these proportions graphically by age, alcohol and tobacco groups in order to begin our analysis and state our initial hypotheses. To do this, let's reuse the global percentages previously calculated for each variable (age, alcohol, tobacco)

```
library(ggplot2)
library(ggpubr)

age_plot <- ggplot(as.data.frame(age_proportions), aes(x=c(0:5), y=age_proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$age)) +
  labs(x="Age", y="Global percentage")
alc_plot <- ggplot(as.data.frame(alc_proportions), aes(x=c(0:3), y=alc_proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$alcgp)) +
  labs(x="Alcohol", y="Global percentage")
tob_plot <- ggplot(as.data.frame(tob_proportions), aes(x=c(0:3), y=tob_proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$tobgp)) +
  labs(x="Tobacco", y="Global percentage")

ggarrange(age_plot, alc_plot, tob_plot)
```




On the graphs, for instance, we clearly see that the proportion of positive cases is higher when the person is old. However, we see that the 75+ group doesn't have so much cases, which can maybe be explained by the fact that people with cancer die before. But more generally, a clear trend seems to be emerging for all three variables: cancer cases seem to increase with age or consumption. This would mean that the rate of positive cases would be correlated with the variables.

At this stage, we can therefore state three hypotheses: - Does age favour the development of œsophageal cancer ? - $H_0$: age and cases proportion are independent. - $H_a$: age and cases proportion are correlated. - Do smoking habits favour the development of œsophageal cancer ? - $H_0$: alcohol and cases proportion are independent. - $H_a$: alcohol and cases proportion are correlated. - Do alcoholic habits favour the development of œsophageal cancer ? - $H_0$: tobacco and cases proportion are independent. - $H_a$: tobacco

and cases proportion are correlated.

In order to answer these questions, we need to test the null hypotheses $H_0$. To do this, we are going to carry out what are called $\chi^2$ tests.

### Hypothesis verification

First, let's explain briefly what a $\chi^2$ test is about. Let $o_{ij}$ is the frequency observed at line $i$ of `prop_cases` and at line $j$ of `age`, and $c_{ij}$ is the calculated frequency expected for the null hypothesis $H_0$ (i.e. if the variables are independent). The null hypothesis must be rejected if the p-value, i.e. the probability of finding the expected value, determined by the following $\chi^2$ test, is lower than the significance level $\alpha$, generally 0.05.

$\chi^2 = \sum_{i,j} \frac{(o_{ij} - c_{ij})^2}{c_{ij}}$

For a $\chi^2$ test to be accurate, the given values should not be too small, otherwise R will warn it. A solution for that could be bootstrapping, i.e. simulating a certain number of samples according to the original dataset, in order to further enhance the accuracy of the test. This method is directly integrated in the $\chi^2$ test function of R, with attribute `simulate.p.value`. In our case, we will use the `xtabs` function in order to build a contingency table with frequencies of data we are interested in.

**Age test** As we want to test the independence of cases proportion and age, we should create two factors, one combining `ncases` and `ncontrols`, and the other being `agegp`. With the built-in function `chisq.test`, the independence $\chi^2$ test will be performed between these two factors.

```
age_table <- xtabs(cbind(ncases, ncontrols) ~ agegp, data=esoph)
age_table
```

```
##
## agegp   ncases ncontrols
##   25-34      1       116
##   35-44      9       199
##   45-54     46       213
##   55-64     76       242
##   65-74     55       161
##   75+       13        44
```

```
chisq.test(age_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  age_table
## X-squared = 68.382, df = 5, p-value = 2.224e-13
```

The test carried out with sufficient precision. As explained above, the most important result is the p-value. Here, the p-value $2.224 \cdot 10^{-13}$ is much lower than the 5 significance level, so we can reject the null hypothesis that the age is independent of the œsophageal cancer cases. This therefore shows a clear relationship between the age and œsophageal cancers. Before continuing, we must be careful not to interpret this conclusion in the wrong way. This correlation does not mean that there is a cause and effect relationship, and could even come about by chance.

**Alcohol test** In the same way as above, we can perform a $\chi^2$ test to reject or not the hypothesis.

```
alc_table <- xtabs(cbind(ncases, ncontrols) ~ alcgp, data=esoph)
alc_table
```

```
##
## alcgp       ncases ncontrols
##   0-39g/day     29       415
##   40-79         75       355
##   80-119        51       138
##   120+          45        67
```

```
chisq.test(alc_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  alc_table
## X-squared = 90.45, df = 3, p-value < 2.2e-16
```

As the p-value is even lower than $2.2 \cdot 10^{-16}$, it is much lower than the 5 significance level and we can reject the null hypothesis that the alcohol habits are independent of the œsophageal cancer cases. This therefore shows a clear relationship between the alcohol habits and œsophageal cancers.

```
tob_table <- xtabs(cbind(ncases, ncontrols) ~ tobgp, data=esoph)
tob_table
```

**Tobacco test**

```
##
## tobgp      ncases ncontrols
##   0-9g/day     78       525
##   10-19        58       236
##   20-29        33       132
##   30+          31        82
```

```
chisq.test(tob_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tob_table
## X-squared = 18.363, df = 3, p-value = 0.0003702
```

As the p-value 0.0003702 is much lower than the 5 significance level, we can reject the null hypothesis that the alcohol habits are independent of the œsophageal cancer cases. This therefore shows a clear relationship between the alcohol habits and œsophageal cancers.

To finish with, the three correlations has been demonstrated. Note that the p-values correspond to what we could expect from the graphs above, i.e. for example alcohol is very much correlated with positive cases, while tobacco is a bit less. However, what does this really mean? As stated above, a correlation between two variables is not sufficient to claim causality. Many factors can influence an observation. Let us imagine, for example, that a sociological study shows that smokers become heavy drinkers with age. In this case, it could very well be that tobacco as well as age influence cases of œsophageal cancer, but not alcohol at all. However, because of this sociological fact, we measure a very strong correlation between alcohol consumption and cases of oesophageal cancer. A correlation, therefore, is not sufficient. Thus, in addition to the previous tests, it would be interesting to look for links between alcoholism and smoking in cases of cancer.

### Tobacco impact depending on alchohol

We want to know the impact of tobacco on cancer frequency depending on alcohol consumption.

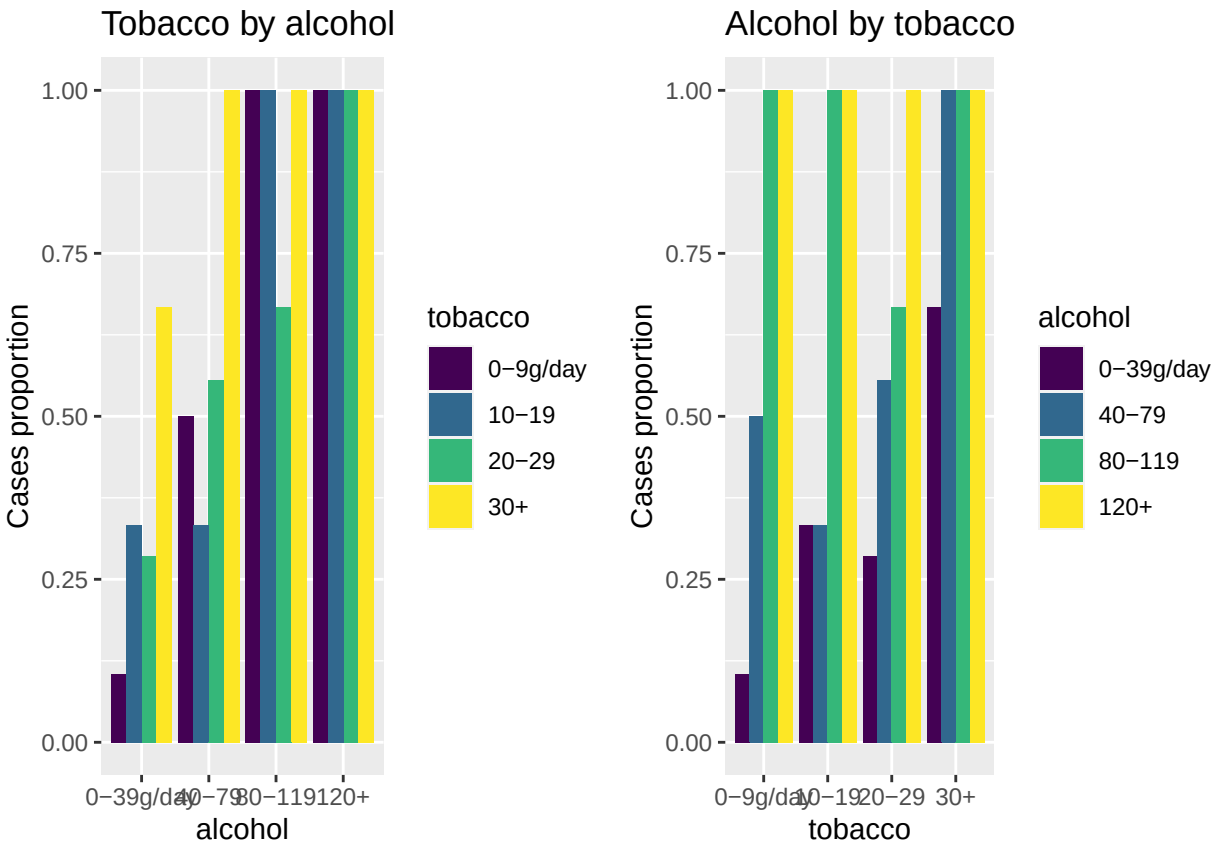
For that, we will first need to build a new dataset in which each proportion will be calculated. Therefore, we create a column `prop_cases`, which represents the positive diagnosed cases for each group combination, normalised by number of controls.

```
# Columns
age <- esoph$agegp
alcohol <- esoph$alcgp
tobacco <- esoph$tobgp
prop_cases <- esoph$ncases / esoph$ncontrols

# Dataset
esoph_prop <- data.frame(age, alcohol, tobacco, prop_cases)
summary(esoph_prop$prop_cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.2679  0.3468  0.5833  1.0000
```

In fact, we see that this column represents a proportion, from 0 to 1, with a mean of 34.68 positive cases. Now, let's build two bar plots with the proportion of cases depending on alcohol and tobacco consumption.

```
alc_tob <- ggplot(esoph_prop, aes(alcohol, prop_cases, fill=tobacco)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Tobacco by alcohol", y="Cases proportion")
tob_alc <- ggplot(esoph_prop, aes(tobacco, prop_cases, fill=alcohol)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Alcohol by tobacco", y="Cases proportion")

ggarrange(alc_tob, tob_alc)
```




First of all, it can be seen that, at low doses, alcohol consumption has less effect than tobacco on cancer cases. However, at higher doses, the effects of alcohol catch up with those of tobacco. Indeed, the two graphs show that high alcohol consumption (over 80 g/day) almost always leads to cancer of the œsophagus, while tobacco seems to have a lower impact alone, up to 60% when no alcohol is consumed. Tobacco therefore has no impact on cancer of the oesophagus above 80 g of alcohol consumed per day, as cancer certainly already exists, but still has a significant impact at lower doses compared to alcohol.

## Building a model

Now that we have shown the correlations between the variables, let's try to go further. We can build a model on these data, allowing us to predict a probability of cancer cases according to different criteria. As we saw earlier, the variables are highly correlated with cancer cases, so a linear model might be appropriate.

### Age model

Let's build a linear model on age correlation.

```
X <- esoph_prop$prop_cases
y_age <- as.numeric(esoph_prop$age) - 1

age_model <- lm(X ~ y_age)
summary(age_model)
```

```
##
```

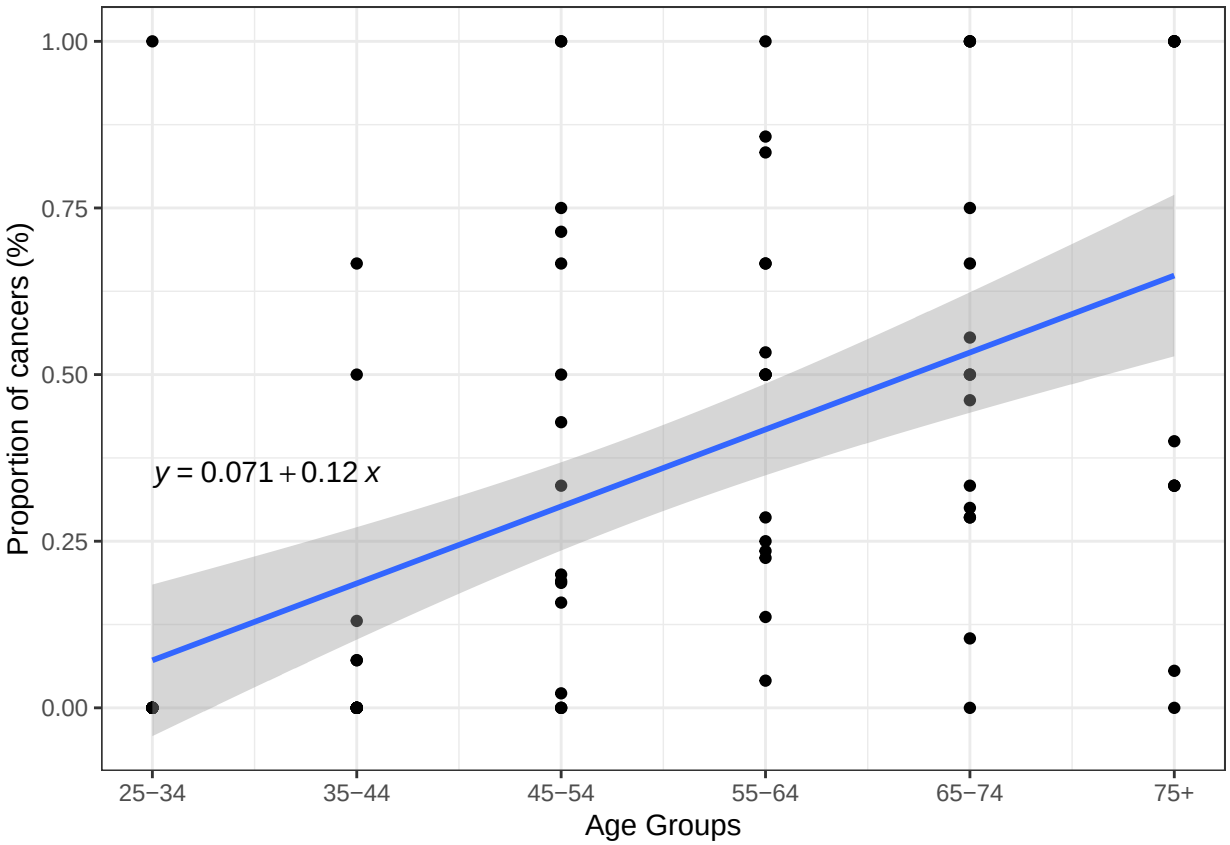
```
## Call:
## lm(formula = X ~ y_age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64861 -0.18672 -0.07125  0.20257  0.92875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07125    0.05721   1.245    0.216
## y_age        0.11547    0.01976   5.845 8.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3041 on 86 degrees of freedom
## Multiple R-squared:  0.2843, Adjusted R-squared:  0.276
## F-statistic: 34.16 on 1 and 86 DF,  p-value: 8.884e-08
```

The model shows different coefficients: - estimates: estimated effect of age on proportion of cases - standard error of the estimated values - t-statistic value: how closely the distribution matches the null hypothesis - p-value: probability of finding the t-value if $H_0$ were true

So the p-value indicates if the model fits the data well or not. Here, we can say that there is a significant positive relationship between age group and cancer proportion (p-value « 0.001), with an increase in cancer proportion of 0.11547 (11.547) for every unit increase in age group. We can visualise this linear regression below.

```
ggplot(esoph_prop, aes(x=y_age, y=X)) + geom_point() + geom_smooth(method="lm") +
  stat_regline_equation(label.x=0.01, label.y=0.35) +
  scale_x_continuous(labels=levels(esoph_prop$age)) +
  labs(x="Age Groups", y="Proportion of cancers (%)") + theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




We can notice that the error area is relatively large, this is because the the data is not sufficiently linear to better fit the model.

### Alcohol model

In the same way, we can build an alcohol model.

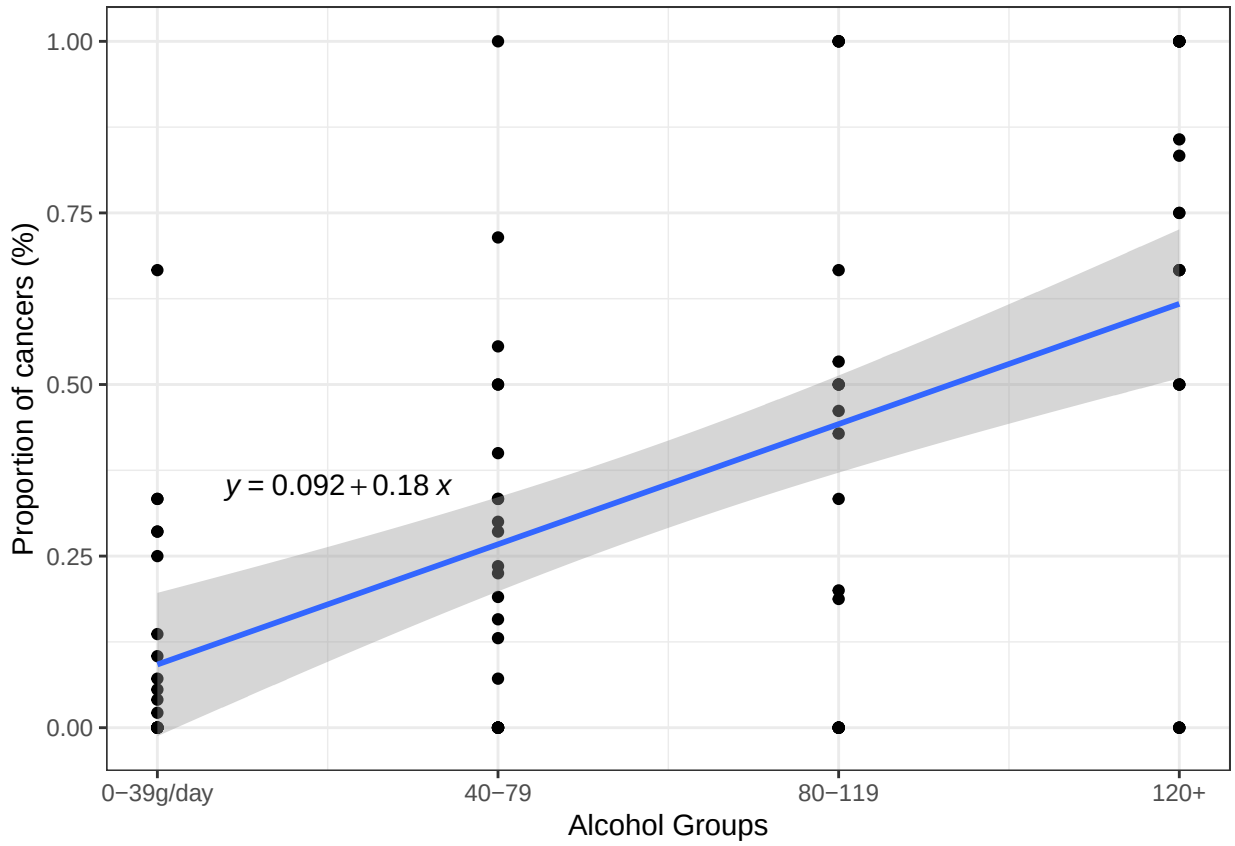
```
y_alc <- as.numeric(esoph_prop$alcohol) - 1

alc_model <- lm(X ~ y_alc)
summary(alc_model)
```

```
##
## Call:
## lm(formula = X ~ y_alc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61750 -0.12231 -0.02626  0.19921  0.73281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09204    0.05250   1.753   0.0832 .
## y_alc        0.17515    0.02863   6.118 2.72e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3 on 86 degrees of freedom
## Multiple R-squared:  0.3033, Adjusted R-squared:  0.2952
## F-statistic: 37.43 on 1 and 86 DF,  p-value: 2.716e-08
```

```
ggplot(esoph_prop, aes(x=y_alc, y=X)) + geom_point() + geom_smooth(method="lm") +
  stat_regline_equation(label.x=0.2, label.y=0.35) +
  scale_x_continuous(labels=levels(esoph_prop$alcohol)) +
  labs(x="Alcohol Groups", y="Proportion of cancers (%)") + theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




**Tobacco model**

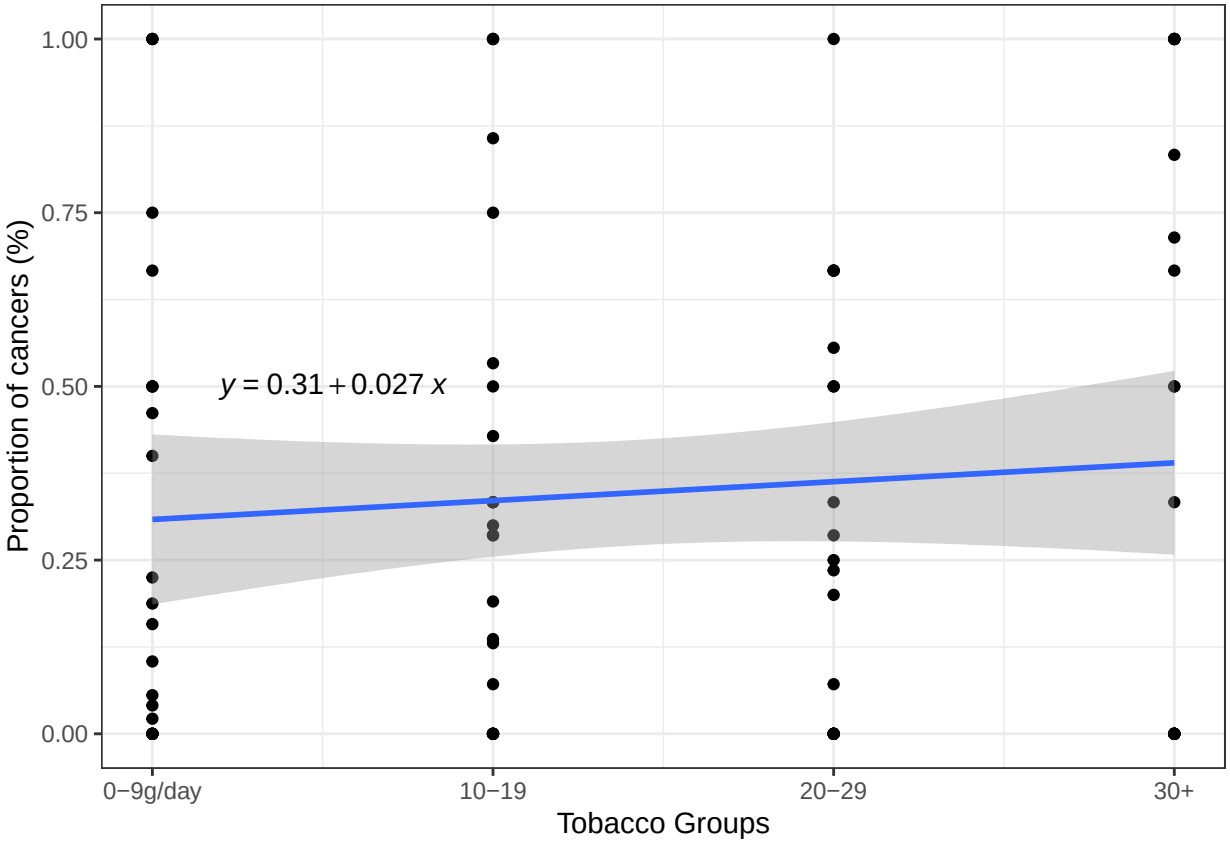
```
y_tob <- as.numeric(esoph_prop$tobacco) - 1

tob_model <- lm(X ~ y_tob)
summary(tob_model)
```

```
##
## Call:
```

```
## lm(formula = X ~ y_tob)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39006 -0.31530 -0.08033  0.21739  0.69150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.30850    0.06154   5.013 2.84e-06 ***
## y_tob        0.02719    0.03426   0.794     0.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3581 on 86 degrees of freedom
## Multiple R-squared:  0.007269,   Adjusted R-squared:  -0.004274
## F-statistic: 0.6297 on 1 and 86 DF,  p-value: 0.4296
```

```
ggplot(esoph_prop, aes(x=y_tob, y=X)) + geom_point() + geom_smooth(method="lm") +
  stat_regline_equation(label.x=0.2, label.y=0.5) +
  scale_x_continuous(labels=levels(esoph_prop$tobacco)) +
  labs(x="Tobacco Groups", y="Proportion of cancers (%)") + theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




In the end, as shown above, we see that tobacco is less correlated than age or alcohol with the increase in cancer cases, but that the value is very high from the outset. In other words, a person who smokes little

is almost as likely to develop cancer of the œsophagus as a person who smokes much more, unlike age or alcohol, for which the probability increases gradually.

## Conclusion

Conclusion.