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
```

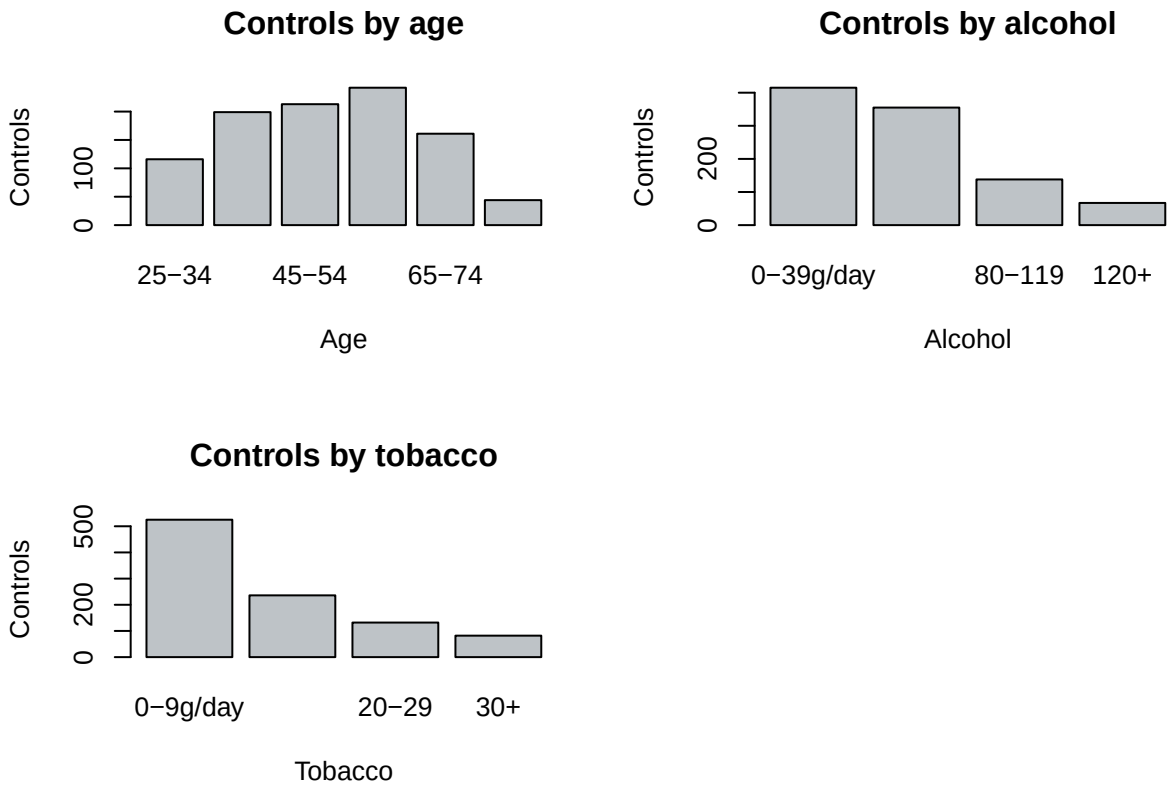
In fact, we can see using the `tapply` R function, which is a `group_by`, that for instance there are 242 controls for the age group `55-64` in the dataset, while there are only 44 for the group `75+`.

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")
```

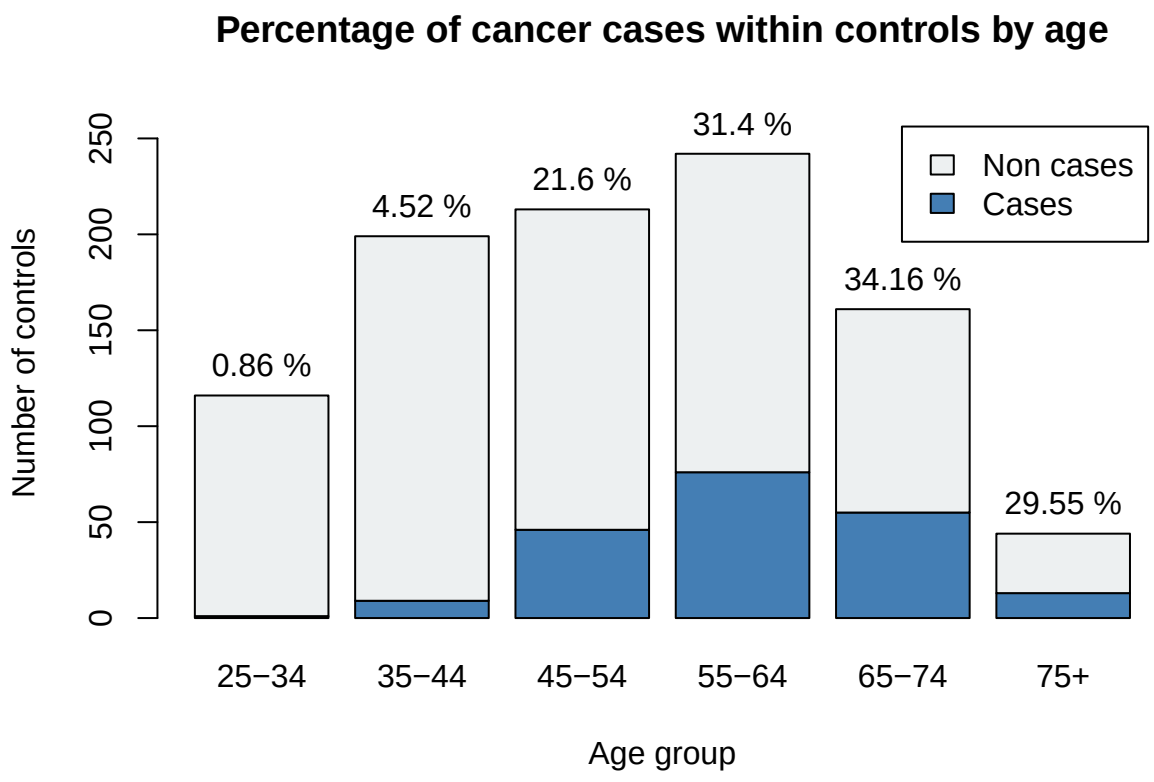


We clearly see that controls not evenly distributed. Note that for alcohol and tobacco, most controlled people present a low consumption, which shows once again that we absolutely cannot overlook the differences in the number of measurements.

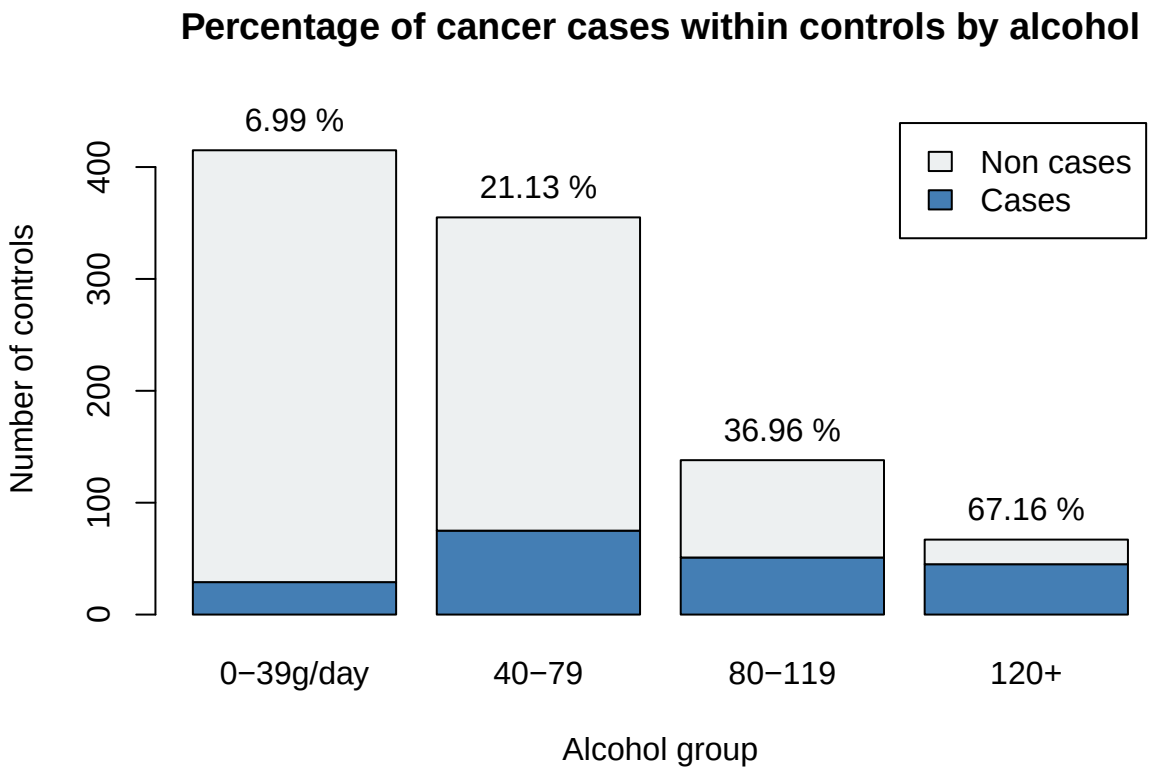
A certainly more reasonable way of analysing this dataset is to base our calculations on the proportion of cancer cases according to the number of controls for each combination of data. A first look at these proportions is given below, by age, alcohol and tobacco.

```
# 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
```

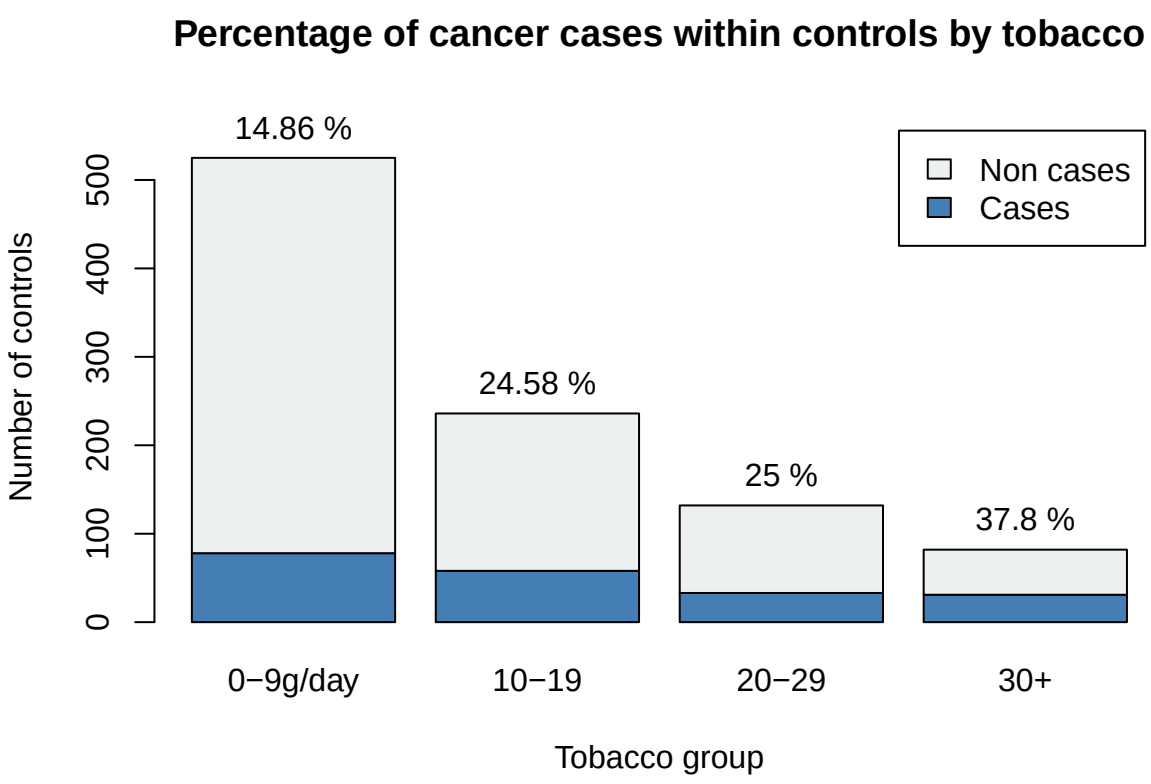
```r
age_ylim <- c(0, 1.1 * max(age_controls))

# 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
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)
```



```
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.

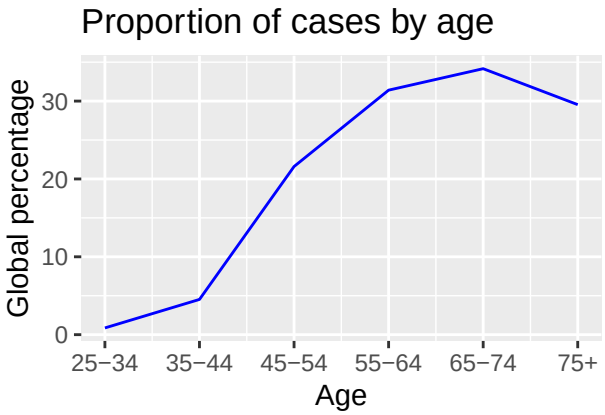
Hypothesis testing

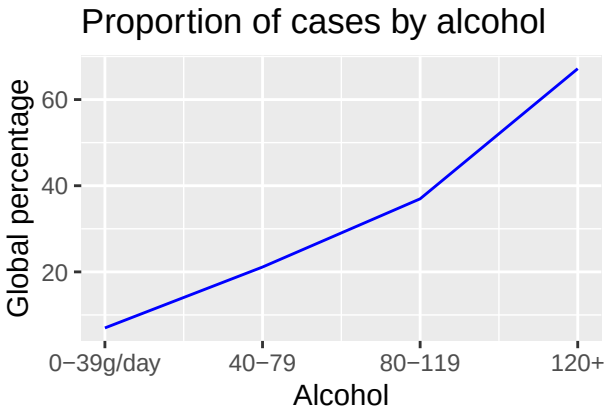
In this section, we can to confirm or not some hypothesis about the dataset. In order to do this, we will define null hypotheses, de noted H_0, and alternative hypotheses, denoted H_a. The null hypothesis is a default position that there is no relationship between two variables, and the alternative hypothesis is, on the contrary, that there is a specific relationship between those variables.

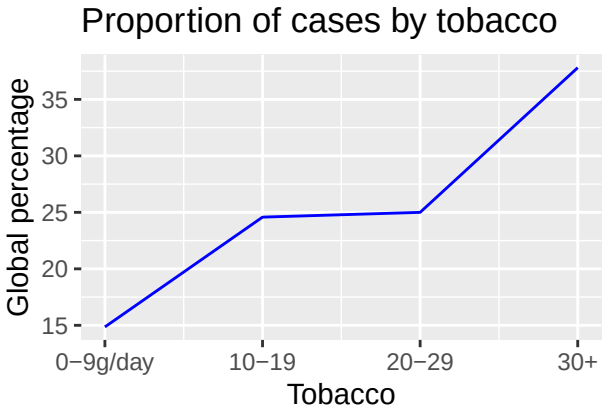
State hypothesis

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$age)) +
  labs(title="Proportion of cases by age", x="Age", y="Global percentage")
alc_plot <- ggplot(as.data.frame(alc_proportions), aes(x=c(0:3), y=alc_proportions)) +
  geom_line(col="blue") + scale_x_continuous(labels=levels(esoph$alcgp)) +
  labs(title="Proportion of cases by alcohol", 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(title="Proportion of cases by tobacco", 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 questions:

- Does age favour the development of œsophageal cancer ?
  - H_0: age and cases proportion ar enot correlated, i.e. correlation factor is below 0.5.
  - H_a: age and cases proportion correlation factor is higher or equal to 0.5.
- Do smoking habits favour the development of œsophageal cancer ?
  - H_0: alcohol and cases proportion correlation factor is below 0.5.
  - H_a: alcohol and cases proportion correlation factor is higher or equal to 0.5.
- Do alcoholic habits favour the development of œsophageal cancer ?
  - H_0: tobacco and cases proportion correlation factor is below 0.5.
  - H_a: tobacco and cases proportion correlation factor is higher or equal to 0.5.

In order to answer these questions, we need to test the null hypotheses H_0, which is the default position that there is no relationship between the variables. To do this, we are going to carry out significance tests of the correlations.

Hypothesis verification

First, let's explain briefly what a significance test is about. Eventually, the null hypothesis must be rejected if the p-value determined by the test is lower than the significance level α, generally 0.05. The p-value is the probability of finding the test-statistic value. The latest is the value expected by the null hypothesis, i.e. how closely the distribution matches the null hypothesis. This value is calculated as follows:

$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

Now that the goals are defined, we can start analysing. First of all, we will build a new cleaned dataset called `esoph_prop`, in which all data will be quantitative and where cases will be given by proportion of controls. Therefore, we create a column `cases`, which represents the positive diagnosed cases for each group combination, normalised by number of controls.

```
age <- as.numeric(esoph$agegp)
alcohol <- as.numeric(esoph$alcgp)
tobacco <- as.numeric(esoph$tobgp)
cases <- esoph$ncases / esoph$ncontrols

esoph_prop <- data.frame(age, alcohol, tobacco, cases)
summary(esoph_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, we can calculate correlations between all three group variables and cases proportion. Let's first build and display a correlation factors matrix using default pearson's correlation.

```
esoph_prop_corr <- cor(esoph_prop)
esoph_prop_corr
```

```
##                 age     alcohol     tobacco      cases
## age      1.00000000 -0.01521895 -0.06780840 0.53318708
## alcohol -0.01521895  1.00000000 -0.04896281 0.55068870
## tobacco -0.06780840 -0.04896281  1.00000000 0.08526092
## cases    0.53318708  0.55068870  0.08526092 1.00000000
```

Nous remarquons tout de suite que la diagonale de la matrice est composée essentiellement de 1, ce qui est tout à fait normal car ce sont des corrélations entre mêmes variables. Aussi, les corrélations entre les trois variables qualitatives que sont l'âge, l'alcool et le tabac n'ont aucune valeur ici. En effet, ce dataset est intéressant pour ses mesures de nombre de cas de cancer, par contre les mesures ont été faites de façon très homogène entre ces trois variables. Autrement dit, il est censé y avoir des mesures pour chaque combinaison de ces variables, c'est pour cela que leurs corrélations sont presque nulles.

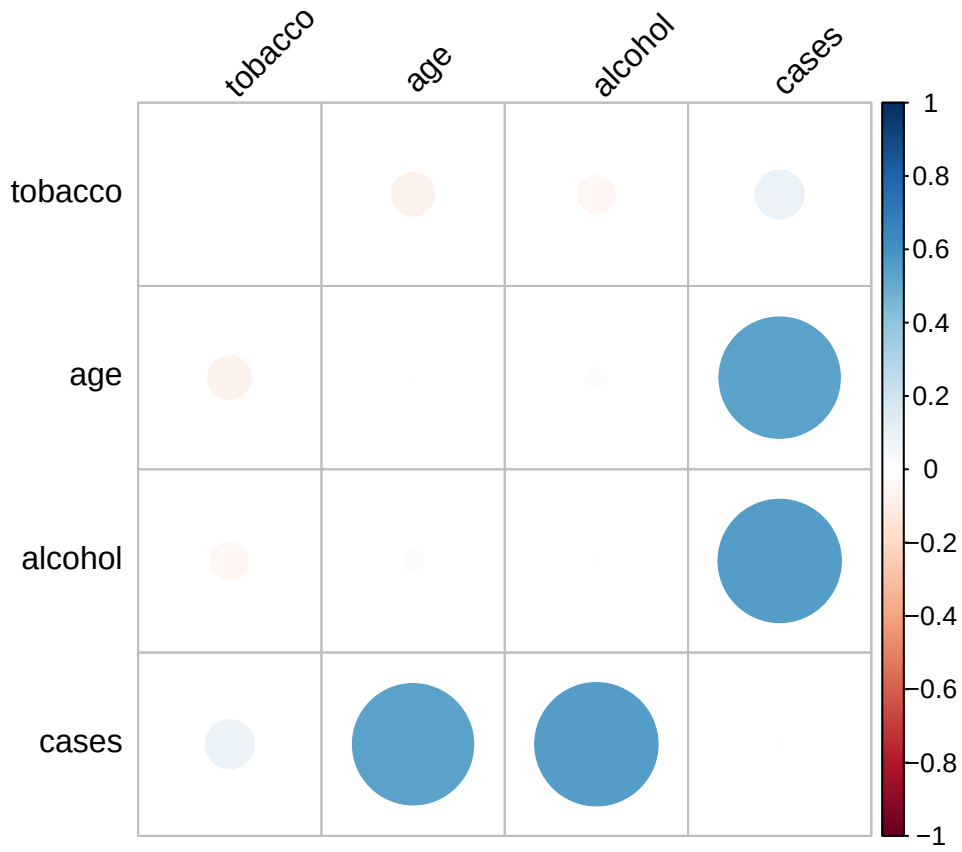
We can also display the matrix on a more clear way as below, either by masking low values or graphically.

```
cleanTable <- function(table, threshold, subst=NA, rev=FALSE) {
  for (i in seq(1, length(table[,1]))) {
        for (j in seq(1, length(table[1,]))) {
          if (i == j ||
              !is.null(threshold) &&
              xor(rev, table[i, j] > -threshold && table[i, j] < threshold)) {
```

```
                table[i, j] <- subst
            }
        }
    }
    return (table)
}

# Masking low values
print(cleanTable(esoph_prop_corr, 0.5), digits=3, na.print=".")
```

```
##           age alcohol tobacco cases
## age         .       .       . 0.533
## alcohol     .       .       . 0.551
## tobacco     .       .       .     .
## cases   0.533   0.551       .     .
```

```
# Graphically
library(corrplot)
esoph_prop_corr_plot <- cleanTable(esoph_prop_corr, NULL, 0.0)
corrplot(as.matrix(esoph_prop_corr_plot), type="full",
         order="hclust", tl.col="black", tl.srt=45)
```



We do see that age and alcohol are correlated with the proportion of cases, as expected. But surprisingly, it seems that tobacco is not correlated with the proportions of cases, contrary to what seemed to appear previously. As seen previously, to be able to reject or not the null hypothesis, we must calculate the p-values. For each couple of variables, we can perform a correlation significance test, as follows for age and cases.

```
cor.test(esoph_prop$age, esoph_prop$cases)
```

```
##
##  Pearson's product-moment correlation
##
## data:  esoph_prop$age and esoph_prop$cases
## t = 5.8447, df = 86, p-value = 8.884e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3644429 0.6680292
## sample estimates:
##       cor
## 0.5331871
```

Here, the confidence interval at 95 is $[0.364, 0.668]$. With a correlation factor of 0.533, the confidence interval is fit. In fact, the determined p-value is $8.884 \cdot 10^{-8}$, which is way lower than the significance level $\alpha = 0.05$. Hence, 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.

Let's compute the p-values in a more efficient way (i.e. not one by one), using the `rcorr` function.

```
library(Hmisc)
print(cleanTable(rcorr(as.matrix(esoph_prop))$P, 0.05, rev=TRUE), digits=3, na.print=".")
```

```
##              age  alcohol tobacco    cases
## age            .        .       . 8.88e-08
## alcohol        .        .       . 2.72e-08
## tobacco        .        .       .        .
## cases   8.88e-08 2.72e-08       .        .
```

The result is definitive: the previously identified correlations are significant, and we can reject the null hypotheses that age and alcohol are not correlated with the proportion of cases. However, the tobacco consumption seems not to be sufficiently correlated with cases proportions to reject the associated null hypothesis. We will try to explain this phenomenon later.

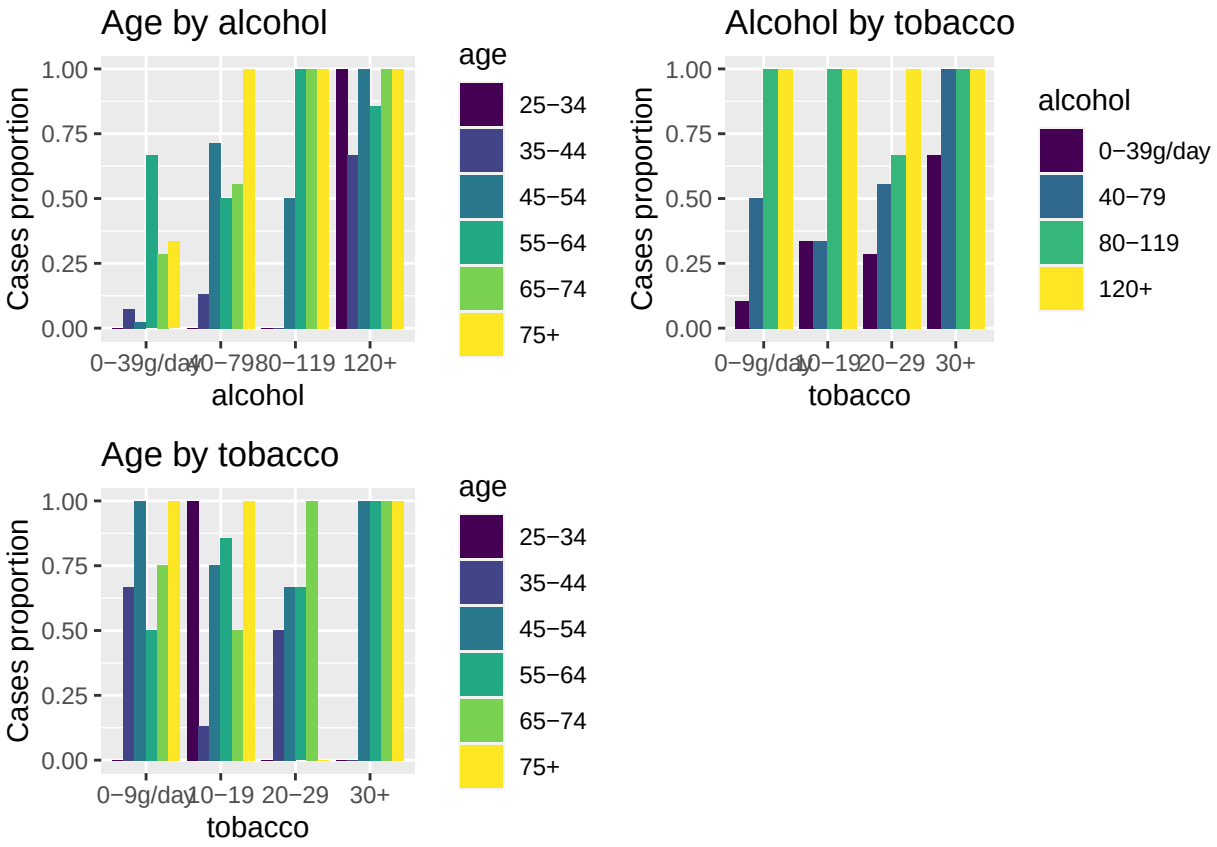
There is one thing to be aware of, however, namely that, as stated above, a correlation between two variables is not sufficient to claim causality. Many factors can influence an observation. Let us imagine, for example, that a sociological study shows that smokers become heavy drinkers with age. In this 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 œsophageal cancer.

After carrying out these last tests, we can confirm that tobacco consumption is not correlated with the proportion of cases of oesophageal cancer. To begin the investigation, let us try to understand how the three variables influence each other.

Influences of the variables on each other

To analyse the influence that these variables have on each other, we will construct three graphs combining the data from one variable according to the categories of another. We could construct all six possible graphs, but these three are sufficient.

```
esoph_prop_qual <- data.frame(esoph$agegp, esoph$alcgp, esoph$tobgp, esoph_prop$cases)
names(esoph_prop_qual) <- c("age", "alcohol", "tobacco", "cases")

age_alc <- ggplot(esoph_prop_qual, aes(alcohol, cases, fill=age)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Age by alcohol", y="Cases proportion")
alc_tob <- ggplot(esoph_prop_qual, aes(tobacco, cases, fill=alcohol)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Alcohol by tobacco", y="Cases proportion")
age_tob <- ggplot(esoph_prop_qual, aes(tobacco, cases, fill=age)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Age by tobacco", y="Cases proportion")

ggarrange(age_alc, alc_tob, age_tob)
```



Let's start with the impact of age as a function of alcohol consumption. We see that from the second alcohol category, 40-79g/day, there are already 50% or more cases of œsophageal cancer for all age categories except the first two, and that from the third category, 80-119g/day, the cancer rate is almost always 100% from the age of 45 onwards. Moreover, from an alcohol consumption of 80g per day, almost all cases develop cancer, in any age group or tobacco consumption. This shows the significant impact of heavy alcohol consumption on cancer cases, but also that the age of the person will strongly determine the resistance of his or her body, and in particular that people under 44 years of age suffer much less from it.

Secondly, for the first two age categories, there are very few cases of œsophageal cancer, apart from alcohol consumption exceeding 119g per day. Conversely, as far as tobacco consumption is concerned, cancer cases are very high from the first age category onwards. Indeed, even at low doses of tobacco, and regardless of

alcohol consumption or the age of the person, cases of cancer are already high. Moreover, certain data seem to be missing for the first age group.

We may therefore have here a first clue to understanding the non-correlation between tobacco consumption and the proportion of cases. Cases of cancer are said to be immediately high when consumption is low, and therefore to increase less following consumption of tobacco than alcohol, for example.

Building a model

Now that we have shown the correlations between the variables let's try to go further, in order to confirm our hypothesis concerning tobacco consumption. For that, we can build a model on these data, allowing us to predict a probability of cancer cases according to different criteria. Let's first try to build a linear model and see if it is appropriate.

Linear model

Let's first build a linear model on age correlation.

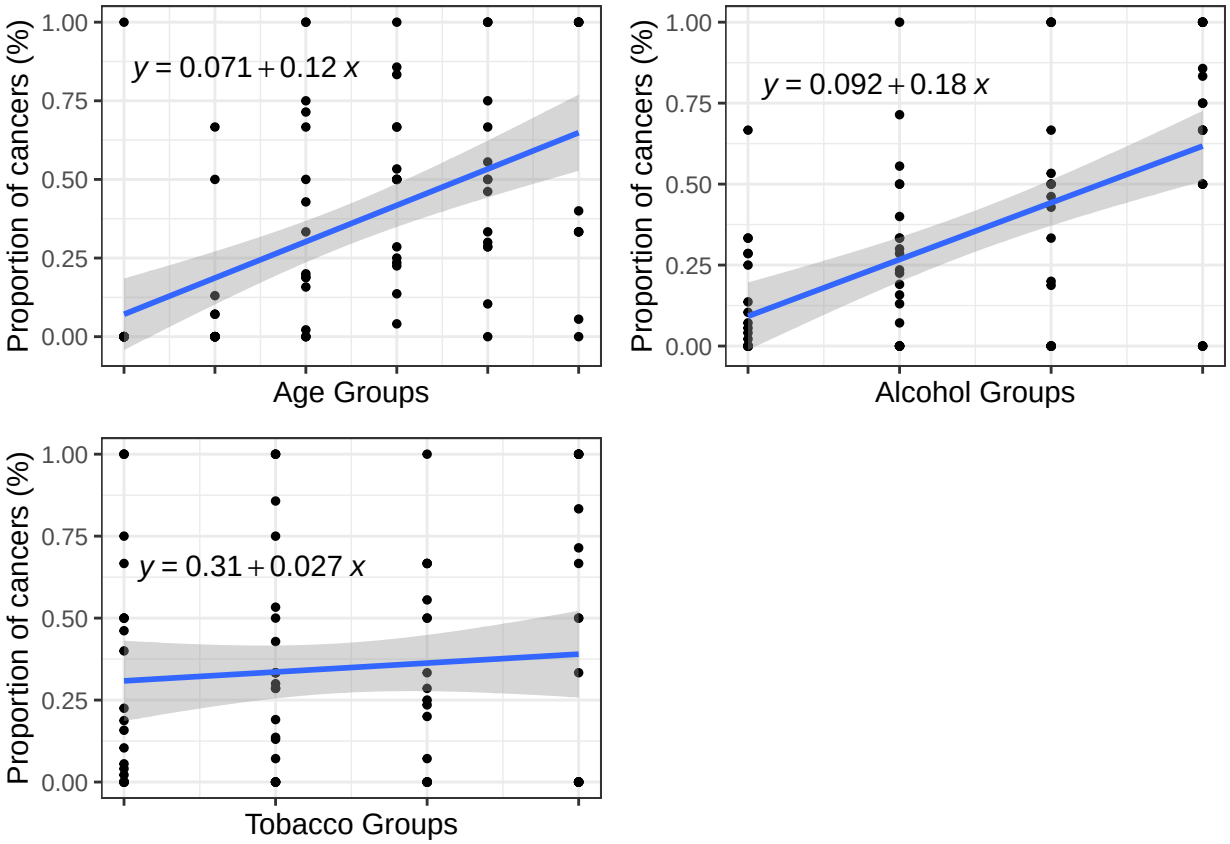
```
X <- esoph_prop$cases
y_age <- as.numeric(esoph_prop$age) - 1

age_linear_model <- lm(X ~ y_age)
summary(age_linear_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 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. Then, we can build the models for alcohol and tobacco and visualise the three linear regressions below.

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

age_linear <- ggplot(esoph_prop, aes(x=y_age, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.85) +
  scale_x_continuous(labels=levels(esoph_prop$age)) +
  labs(x="Age Groups", y="Proportion of cancers (%)") + theme_bw()
alc_linear <- ggplot(esoph_prop, aes(x=y_alc, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.8) +
  scale_x_continuous(labels=levels(esoph_prop$alcohol)) +
  labs(x="Alcohol Groups", y="Proportion of cancers (%)") + theme_bw()
```

```
tob_linear <- ggplot(esoph_prop, aes(x=y_tob, y=X)) + geom_point(size=1) +
  geom_smooth(method="lm") + stat_regline_equation(label.x=0.1, label.y=0.65) +
  scale_x_continuous(labels=levels(esoph_prop$tobacco)) +
  labs(x="Tobacco Groups", y="Proportion of cancers (%)") + theme_bw()

ggarrange(age_linear, alc_linear, tob_linear)
```



One more time, we can confirm 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 seems to be 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.

However, we can notice that the error areas are relatively large, this is because the the data is not sufficiently linear to better fit the model. Moreover, for all three models, the R^2 value is very low. For instance in the age model, it is of about 0.28, while a correct R^2 should be higher than 0.8. This shows indeed that the previous models are not accurate enough. Thus, we will need to build logistic regressions.

Logistic regression

Conclusion

Conclusion.