Computer Lab Practical n. 1

Today's task

Use Rstudio to produce some simple graphs and verify simple hypotheses.

How to start

- Access the course web page on Moodle (/Biology). Download and save the xenope.dat and schizophrenic.rda data files.
- Double click on "Rstudio".

How to load data available in R

- Load one of R's packages called MASS.

```
install.packages("MASS")
library(MASS)
```

- Read data set from an attached package

```
data("Pima.tr")
```

- To inspect the data we just loaded click on Pima.tr.

```
View(Pima.tr)
```

- To retrieve a description of the dataset (if available):

```
help(Pima.tr)
```

or

```
?Pima.tr
```

- To produce a numerical summary of the data:

```
summary(Pima.tr)
```

Does the prevalence of diabetes in the Pima population agree with the one of the general population?

- The prevalence of diabetes in the general population is estimated to be around 9%. Let's see whether this agrees with the prevalence of diabetes among women of the Pima tribe.
- To estimate the prevalence in the given sample:

```
# frequency table
.Table <- xtabs(~ type , data= Pima.tr )
print(.Table)
# proportion table
prop.table(.Table)
```

- To run a test and compute confidence intervals:

Proportion test

Suggestion: Use prop.test function

Question: why do we have to specify p = 0.91 and not p = 0.09?
Perform the "Exact binomial test".
Suggestion: Use binom.test function
- Your comments?

Diabetes and BMI

- To plot the histogram of BMI in the two population groups:

```
# Create two new variables: bmi for the two groups
bmi_typeYES <- Pima.tr$bmi[Pima.tr$type=="Yes"]
bmi_typeNO <- Pima.tr$bmi[Pima.tr$type=="No"]

# First distribution
hist(bmi_typeYES, xlim=c(11,55), col=rgb(1,0,0,0.5), xlab="bmi", prob=T,
      breaks=15, main="Distribution of bmi in the two populations")

# Second with add=T to plot on top
hist(bmi_typeNO, col=rgb(0,0,1,0.5), prob=T, breaks=15, add=T)

# Add legend
legend("topright", legend=c("Yes","No"), title="Diabetes",
       col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex=2, pch=15 )
```

- Is this associated with diabetes?

```
# Summary of bmi
summary(Pima.tr$bmi)

# Summary by group
# We use the function numSummary from the RcmdrMisc package.
install.packages("RcmdrMisc")
library(RcmdrMisc)
numSummary(Pima.tr$bmi, groups=Pima.tr$type)
```

- Much better to use a graphical summary.

```
boxplot(bmi ~ type, data=Pima.tr, xlab='Diabetes', ylab='bmi')
```

- A further frequently used (though less informative) graphical summary

```
# We use the function plotMeans from the RcmdrMisc package (already loaded).
plotMeans(Pima.tr$bmi, Pima.tr$type, error.bars="se", connect=TRUE,
          xlab="Diabetes", ylab='bmi')
```

Question: what do the bars represent?

- Is the difference significant?

```
t.test(bmi~type, data=Pima.tr, alternative="two.sided",
       conf.level=.95, var.equal=FALSE)
```

- You may also use a non-parametric test: Wilcoxon test

```
wilcox.test(bmi ~ type, alternative="two.sided", data=Pima.tr)
```

- We haven't discussed this in beforehand, but we can also verify whether the variability of the two groups is the same using suitable test statistics. For instance:

```
leveneTest(bmi ~ type, data=Pima.tr, center="median")
```

Question: which option of the two-sample t test should be avoided?

```
?t.test
```

Obesity and diabetes

- A person is considered to be overweight if her/his BMI exceeds 30.
- Define a new variable overweight.

```
Pima.tr$overweight <- as.factor(ifelse(Pima.tr$bmi<30,"No","Yes"))
```

Inspect the new variable by using

```
View(Pima.tr)
```

or

```
head(Pima.tr)
```

- Use Two-sample proportions test to reproduce what we did with the "alcoholic mice" data set.

Suggestion: Create two way frequency table of two variables: overweight and type by xtabs function

```
.Table <- xtabs(~overweight+type,data=Pima.tr)
```

then use prop.test function for 2-sample proportions test.

Question: what does the confidence interval refer to?

- Or, try with both, Pearson's χ^2 test and Fisher's exact test for two way frequency table of two variables: overweight and type that defined as above using chisq.test and fisher.test

Comment. The χ^2 test agrees with the two-sample test for proportions. Fisher's exact test can be used to verify whether two proportions are equal. A particularity of this test is that it doesn't require a minimal sample size, that is, it can also be used for small samples.

- We may want to test whether the prevalence of diabetes among the normal weight Pima agrees with the one of the general population. To proceed as this:

We first select a subset of data in which variables overweight=="No", and rename the new data set as normalweight

```
normalweight <- Pima.tr[which(Pima.tr$overweight=="No"),]
```

Then, repeat the previous analyses on normalweight data.

Diabetes and blood pressure

Study, as done so far, whether also blood pressure (bp variable) is associated with diabetes. Do it first using the original values of the (diastolic) blood pressure. Then, re-code assuming that bp > 80 qualifies hypertension, i.e. create a new variable:

```
# create a new variable "hypertension"
Pima.tr$hypertension <- as.factor(ifelse(Pima.tr$bp > 80, "Yes", "No"))

# check
table(Pima.tr$bp>80)
table(Pima.tr$hypertension)
```

Xenope (new)

• Load the xenope.dat file into Rstudio using

```
xenope <- read.table("~/Lab1/xenope.dat",header=TRUE)
```

Where ~ refers to the direction that you stored the data.

• View(xenope) allows you to inspect the loaded data.

• To highlight a possible dependence between the two variables, use, for instance Scatterplot.

• Perform the "paired t test" using t.test function.

MAO activity levels and schizophrenia analysis (new)

• Load data into RStudio

```
load(file = file.choose())
```

then select schizophrenic.rda from the directory to which you downloaded it.

• To plot the boxplots and inspect the different groups:

```
boxplot(mao ~ type, data=schizophrenic)
```

• To numerically summarise the data (this essentially tells the same story):

```
# Summary by group
# We use the function numSummary from the RcmdrMisc package.
library(RcmdrMisc)
numSummary(schizophrenic$mao, groups=schizophrenic$type)
```

Try and convince yourself that these summaries, particularly the mean, the median and the quantiles, tell us the same story as the previous graph.

• The same "suggestion" is provided by

```
# We use the function plotMeans from the RcmdrMisc package.
plotMeans(schizophrenic$mao, schizophrenic$type, error.bars="se",
          connect=TRUE, xlab="type", ylab='mao')
```

Comment the picture!

- Are the differences significant? Let's try with one-way analysis of variance (ANOVA) and the Kruskal-Wallis test.

For the former:

```
AnovaModel.1 <- aov(mao ~ type, data=schizophrenic)
summary(AnovaModel.1)
```

For the latter:

```
kruskal.test(mao ~ type, data=schizophrenic)
```

Your comments? Can you furthermore figure out how one-way ANOVA works?

- Student's t test is also useful for pairwise comparison.

```
pairwise.t.test(schizophrenic$mao, schizophrenic$type)
```

Pairwise comparison using the Wilcoxon test can be carried out similarly using the pairwise.wilcox.test function.
Question: what do we do exactly?