MATHEMATICAL STATISTICS. A statistical analysis of NHANES variables. Part 2

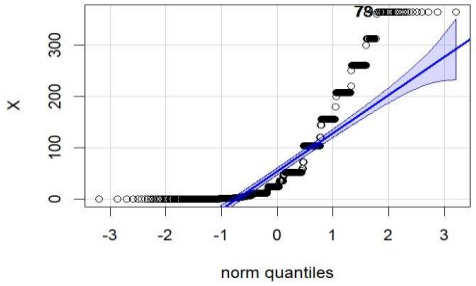
Exercise 1

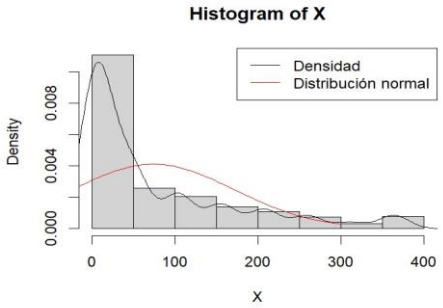
Study the normality of a variable, denoted herein X. Compute a confidence interval at level 0.90 for the mean of the variable X and another one for its median.

We have selected the following variables:

- X2: Number of days that a person drinks alcohol a year
- F1: "YES" or "NO" if the person consumes marihuana regularly
- F2: Education

In the last exercise, we study the normality of our variable X=X2, plotting the sample





We saw that it does not follow a normal distribution.

To confirm it, we use a normality test: the null hypothesis is that the population from which the sample comes is Normal, but since each of them evaluate different aspects of the normal behaviour, we can obtain different results. (Any condition is required)

Using Kolmogorov-Smirnov test we get a p-value of $\mathbf{2.2^{*}\ 10{-}16}$. Since it is lower than 0.05, we reject the hypothesis that the distribution is normal.
Using Shapiro-Wilk test we get, as before, p-value of $\mathbf{2.2^{*}\ 10{-}16}$. Since it is lower than 0.05, we reject the hypothesis that the distribution is normal.

Also, since n=800 is large we can apply the t.test although our variable is not normal. (We can t.test since, if the sample size is large, the distribution tn−1 can be approximated by a Normal distribution).
We obtain this asymptotic confidence interval for the mean: $[\mathbf{68.\ 17},\ \mathbf{79.\ 98}]$, which includes the mean of our sample, 74.075.

For the median of X (which is 24) we study the hypothesis

$$\boldsymbol{H_0: M = M_0\ \ vs\ H_1: M \neq M_0}$$

Where M denotes the median of the distribution and M_0 the median of our sample

Since we do not know the distribution of the variable, a non-parametric test should be used and we obtain the following confidence intervals:

- Using SIGN. test: $[\mathbf{24}, \mathbf{36}]$, which contains the median of X.
- Using Wilcox-test: $[\mathbf{52}, \mathbf{76}]$, which does not contain the median of X.

Exercise 2

We define two groups based on one of the factors, F1 or F2, denoted herein F. If F takes more than two values, you can define the two groups using different approaches; for example, you can consider the groups defined by one of the values, and the complementary group or you can consider the groups defined by the two values with the highest frequency. Study whether the mean of the variable X is the same in the two groups considered. Study also if the median is the same in the two groups.

We take the variables:

- F=F1 Marihuana use, Yes or No
- X =X2 Alcohol Use, number of days a year

The groups are:

- X[F=='Yes'] Use of alcohol of people who use marihuana
- X[F=='No'] Use of alcohol of people who do not use marihuana

Test for the mean

To study the difference of the means we state the test of hypothesis:

$$\boldsymbol{H_0: \mu_{YES} = \mu_{NO} \ \ vs \ H_1: \mu_{YES} \neq \mu_{NO}}$$

Where μ_{YES} and μ_{NO} denote the mean of the variables X[F=='Yes'] and X[F=='No'].

We will use a **t.test**. The variable X is not normal but since n = 800 is large we will get an asymptotic CI for the difference of the means and we can apply the test. With α = 0.05 and without the assumption that the variances are equal we get [$\mathbf{8.85}$, $\mathbf{39.67}$] and a p-value 0.002.

0 does not belong to our CI and the p-value is lower than 0.05. Thus, we have evidence to reject our initial hypothesis and we deduce that the mean of the variable X is different in each group.

Test for the median

To study the difference of the medians we state the test of hypothesis:

$$\boldsymbol{H_0: M_{YES} = M_{NO} \ \ vs \ H_1: M_{YES} \neq M_{NO}}$$

Where M_{YES} and M_{NO} denote the median of the variables X[F=='Yes'] and X[F=='No'].

We use a **Wilcox. Test.** Since the variable is not normal. We get the following CI: [$\mathbf{3}$, $\mathbf{24}$], a difference in location of 12, and a p-value of $\mathbf{3.54 * 10{-}5}$.

Again 0 does not belong to our CI, and the p-value is lower than 0.05 so we reject our hypothesis. We deduce that the median of the variable X must be different in each group.

Exercise 3

Study if the two factors (F1 and F2, the original variables) are independent, that is if the distribution of one of the factors is the same in the groups defined by the other.

We want to check if the consume of marijuana is dependent to the education. To answer this, we need to study the hypothesis of independence of $H_0: p_{ij} = p_i p_j$ vs $H_1: at\ least\ p_{ij} \neq p_i p_j$ where p_{ij} is the conditional probability of the occurrence of the event *i* given that the group is *j*.

Computing the contingency table, we get:

	8th grade	**9-11th grade**	**Highschool**	**Some college**	**College grad**	**Sum**
No consume	26	59	90	171	202	548
Consume	5	33	56	78	42	214
Sum	31	92	146	249	244	762

And doing the Chi-Squared test, we obtain a p-value $1.16 * 10^{-5}$. As it is lower than our significance level $\alpha = 0.05$ we reject the null hypothesis. We conclude then that our two variables are dependent.

Now, we are comparing the Pearson Residuals in order to check which groups provoke the dependence.

	8th grade	**9-11th grade**	**Highschool**	**Some college**	**College grad**
No consume	0.78	-0.88	-1.46	-0.60	2
Consume	-1.26	1.41	2.34	0.97	-3.2

As every residual is far away from zero, we deduce that the dependence is among the whole sample.

APPENDIX

Code to get the sample:

```
knitr::opts_chunk$set(echo = TRUE)
library(NHANES)
library(RcmdrMisc)
#group H(j=8)
set.seed(800)
n=800

index <- complete.cases(NHANES$RegularMarij)

set.seed(800)
F1 <- sample(NHANES$RegularMarij[index], 800)
set.seed(800)
X2 <- sample(NHANES$AlcoholYear[index], 800)
set.seed(800)
X1 <- sample(NHANES$HHIncomeMid[index], 800)
set.seed(800)
F2 <- sample(NHANES$Education[index], 800)
```

Exercise 1 R Code

```
X=X2
#normality
sigma<-sd(X,na.rm=TRUE)
mu<-mean(X,na.rm=TRUE)

qqPlot(X)

hist(X, probability=TRUE, breaks=11, xlim=c(0,400))
dxE <- density(X, na.rm=TRUE)
lines(dxE, col="black")
xn <- qnorm(seq(0,1,length.out=100), mu, sigma)
lines(xn, dnorm(xn, mu, sigma), col='red')

# Añadir leyenda
legend("topright", legend=c("Densidad", "Distribución normal"),
       col=c("black", "red"), lty=c(1,1), lwd=c(1,1))

ks.test(X,"pnorm",mean=mu,sd=sigma) #Kolmogorov-Smirnov test
shapiro.test(X)#shapiro.Wilk test

#The variable X is not Normal, but the sample size is large, so that we can use the asymptotic CI for μ
#(X¯−z1−α/2Sn−−√,X¯+z1−α/2Sn−−√)
#This CI is not implemented in any specific R function but we can use again t.test since,
#it the sample size is large, the distribution tn−1 can be approximated by a Normal distirbution.
#However, we do not know any CI for the variance of a non-Normal distribution.
```

```
t.test(X, conf.level=0.9)


library(BSDA)
median(X,na=TRUE)
SIGN.test(X,md=24,alternative = "two.sided",conf.level = 0.9)
help("SIGN.test")
wilcox.test(X, alternative='two.sided', mu=24, conf.int=TRUE)
```

Exercise 2 R Code

```
F=F1

t.test(X[F=='Yes'], X[F=='No'], conf.level=0.95, var.equal=FALSE)#for the difference of the mean


wilcox.test(X[F=='Yes'],X[F=='No'], alternative="two.sided", conf.int=TRUE)# test for the difference of the medians

median(X[F=='No'],na.rm=TRUE)

median(X[F=='Yes'],na.rm=TRUE)

91.17051-66.91120
```

Exercise 3 R Code

```
F11 <- factor(F1, labels = c("No consume","Consume"))

F22 <- factor(F2, labels=c("8th grade","9-11th grade", "Highschool", "SomeCollege",
"CollegeGrad"))

tabla <- xtabs(~F11 + F22, data = NHANES,)

addmargins(tabla)


tabla.test<-chisq.test(tabla, correct=TRUE)

tabla.test

tabla.test[3]>0.05 #They are dependent


round(tabla.test$expected,2)

round(tabla,2)

round(tabla.test$residuals,2)
```