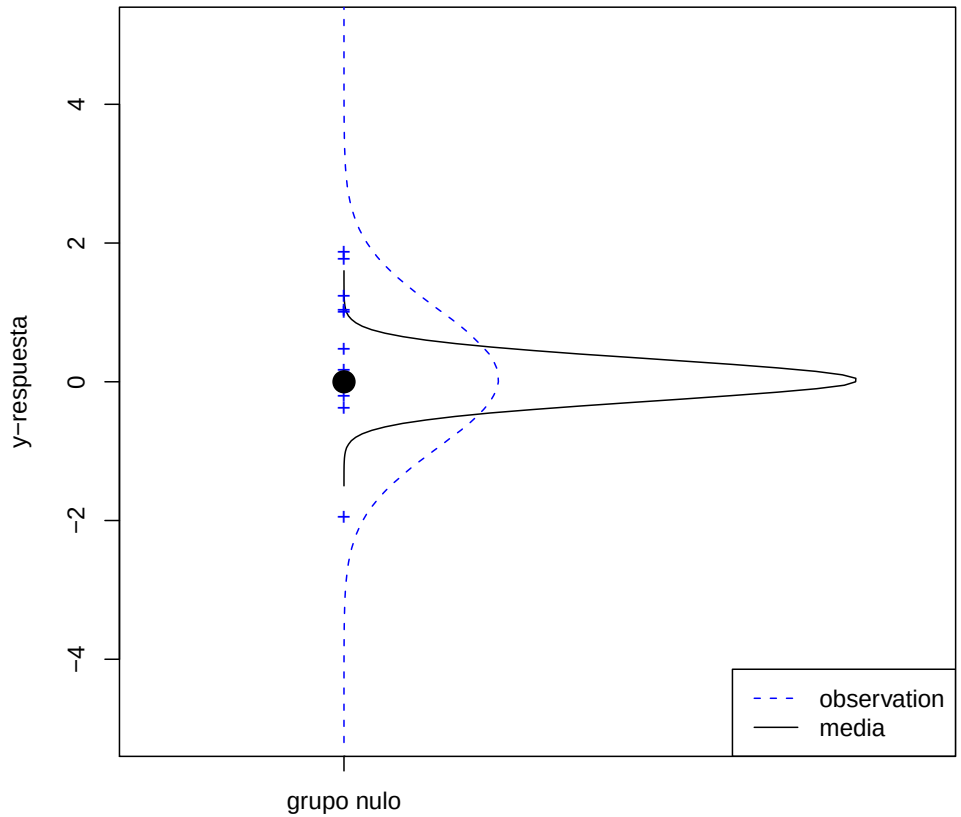
Statistics with R

Alejandro Cáceres[a]
acaceres@creal.cat

[a]Centre for Research in Environmental Epidemiology (CREAL)

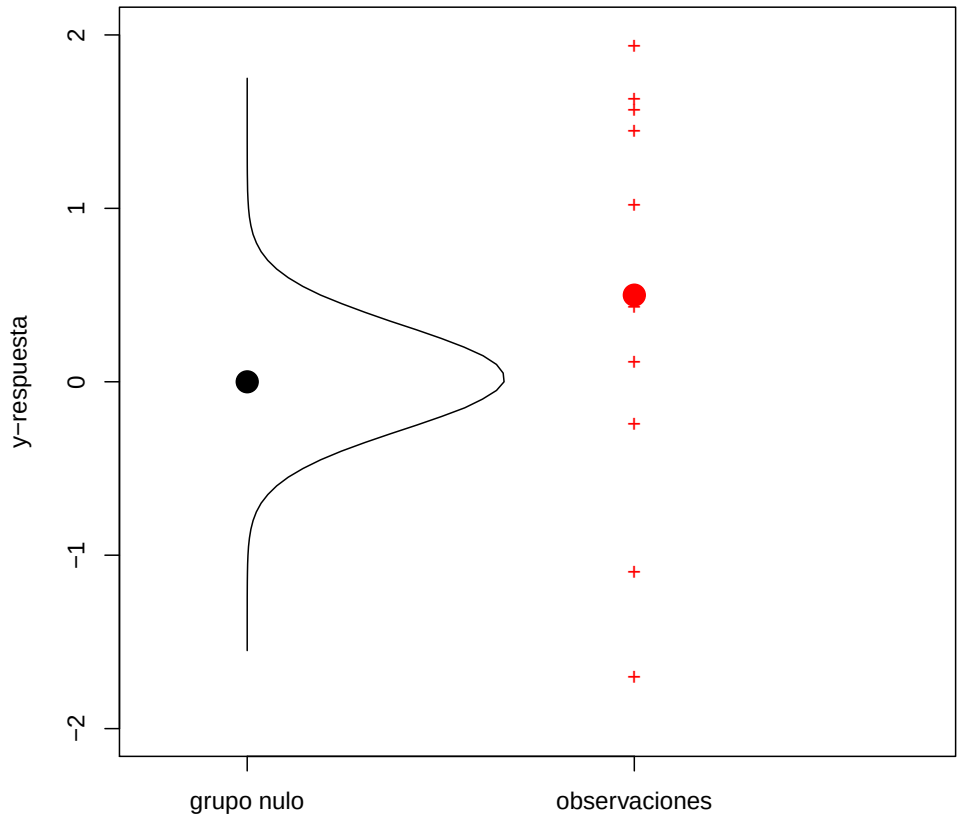
May 2016

Significance Test



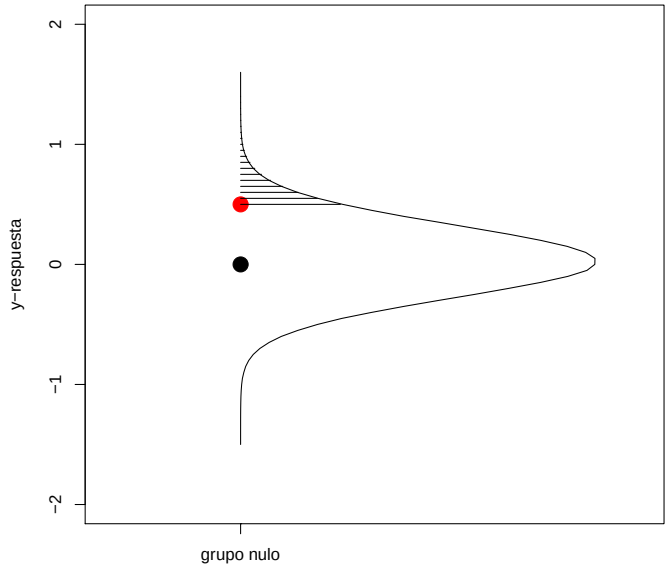
Let us suppose a system for which the mean value of 10 measurements has a Gaussian distribution (black). We call “null hypothesis” the mean of this Gaussian process where no treatment or effect is expected.

Significance Test



Now imagine we actually perform 10 observations of a system under a kind of treatment. We then want to know whether the treatment had a significant effect on the system or not.

Hypothesis testing



We then compute the mean of our 10 observations and compute the probability of having found at least this mean assuming the system of no-effect. If the probability, namely p-value, is lower than a conventional threshold (5%) then we say that our observation is probably not due to chance, that we significantly reject the null hypothesis, or that the effect is statistically significant.

t-test

Let us assume that the null hypothesis is that a typical mean of 10 observations is zero. For testing whether an observed mean of actual measurements is significantly different from zero we can use the R function `t.test`

```
>  load("obs.RData")
>  t.test(obs)

One Sample t-test

data:  obs
t = 1.3709, df = 9, p-value = 0.2036
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.3250338  1.3250338
sample estimates:
mean of x
      0.5
```

The t statistics measures how far the mean of our 10 measurements is from zero, accounting for the observed mean's variability (V) or $t = \frac{\mu}{\sqrt{V}}$. Our observed t is compared to a student's t-distribution and a p-value is thus computed.

t-test

`t.test` returns an object similar to a list

```
> t<- t.test(obs)
> names(t)
[1] "statistic"   "parameter"   "p.value"     "conf.int"    "estim
[8] "method"      "data.name

> t$statistic
       t
1.370948
> t$conf.int
[1] -0.3250338  1.3250338
attr(,"conf.level")
[1] 0.95
```

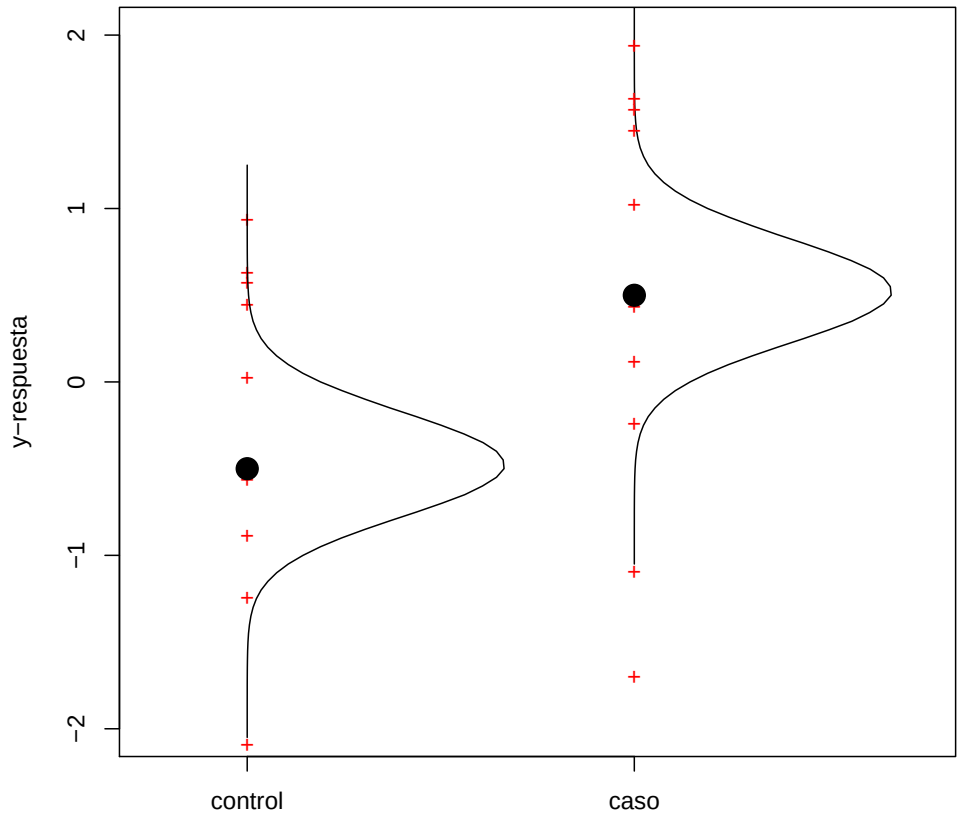
INTERLUDE:

Some of the advantages of R is its graphics and how to change other people's code

```
#simulacion
sm<-rnorm(10000000,0,1)
nm<-hist(sm,freq=FALSE,br=100)
observaciones <-sample(sm,10)
smmean<-rnorm(10000000,0,0.3)
nnmean<-hist(smmean,freq=FALSE,br=100)

#grafica
plot(nnmean$density+0.5, nnmean$breaks[-1],lty=1,col="black",
type="l",xlim=c(0,2),ylab="y-respuesta",xlab="",xaxt="n",ylim=c(-5
lines(nm$density+0.5,nm$breaks[-1],col="blue",lty=2,)
axis(1,at=c(0.5), labels=c("grupo nulo"))
points(rep(0.5, length(observaciones)), observaciones,col="blue",
pch="+" )
points(0.5,0,pch=19,cex=2,col="black")
legend("bottomright",c("observation","media"),lty=c(2,1),
col=c("blue","black"))
```

Try to paint in red the histogram of the variable obs

Differences between two group means



If we don't really know how to model a system of no-treatment, we can perform a set or group of observations on such as system and compare them with the system under treatment. Think of 2 groups of hyper-tensed individuals one under placebo and the other under medication. We want to test if the mean blood pressures for each group are significantly different.

Difference between tow group means

Load the data in phenosCont.txt, where there are a set of four observations for each individual (row) Explore the Variables: How many subjects are there? which type of variable is pop? what is the mean of X1?

```
> phenos<-read.table("phenosCont.txt",header=TRUE)
> head(phenos)
   pop caco           X1          X2
1 Pop1    1 -0.045452947  0.02677151
2 Pop1    0  0.018036264 -0.03361612
3 Pop1    1  0.001631087  0.03286500
4 Pop1    0  0.021977998 -0.05069528
5 Pop1    0 -0.003845641  0.02882352
6 Pop1    0  0.017218303 -0.01419022
```

Difference between tow group means

The differences between two group means can be computed in two ways

- with a group t-test

```
control<-phenos$X1[phenos$caco==1]
caso<-phenos$X1[phenos$caco==0]
t.test(control,caso)
```

- with a linear regression model

$$(3.2, 5.6, .., -1.1, -2.0) \sim \mu_{control} + \beta(0, 0, ..1, 1) + \epsilon \quad (1)$$

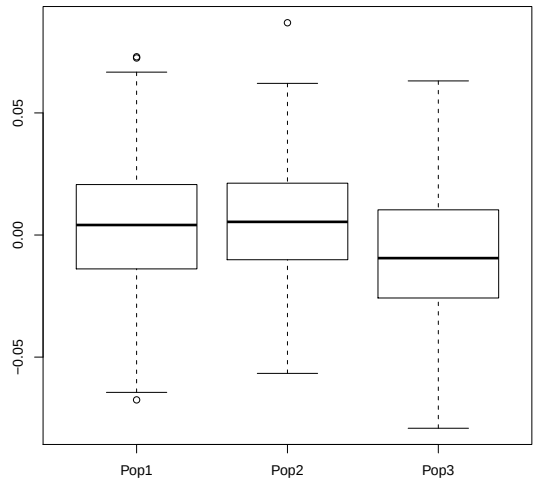
```
model<-lm(X1~caco,dat=phenos)

summary(model)
names(model)
```

where $\beta = \mu_{caso} - \mu_{control}$. Check that both models give the same result.

Group comparison of more than two groups

Let us look at the differences between three populations (three different treatments or conditions) for our variable X2. If the x axis is a factor R will plot a boxplot.

```
#pdf() saves "plotBox.pdf"
pdf("plotBox.pdf")
 plot(phenos$pop,phenos$X2)
dev.off()

# note: if the x variable is continuous then R makes a scatter plo
pdf("plotCont.pdf")
 plot(phenos$X1,phenos$X2)
dev.off()
```



Group comparison of more than two groups

Like the case for two groups, we can use a linear model to test the differences between each pair of groups. The independent variable is then a factor with three categories. The first one is taken as the reference.

```
> model<-lm(X2~pop,dat=phenos)
>
> model

Call:
lm(formula = X2 ~ pop, data = phenos)

Coefficients:
(Intercept)      popPop2      popPop3
   0.003409     0.001626    -0.011852
```

Group comparison of more than two groups

```
> summary(model)

Call:
lm(formula = X2 ~ pop, data = phenos)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0709260 -0.0164655 -0.0000323  0.0169957  0.0818985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003409   0.001124   3.032  0.00247 **
popPop2      0.001626   0.001590   1.023  0.30660
popPop3     -0.011852   0.001590  -7.455 1.51e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.02514 on 1497 degrees of freedom
Multiple R-squared: 0.05413,Adjusted R-squared: 0.05287
F-statistic: 42.84 on 2 and 1497 DF,  p-value: < 2.2e-16
```

Group comparison of more than two groups

Another hypothesis test, instead of pair-wise comparisons, is whether all the means are different between themselves. This is performed with an ANOVA test, which is a simple a function on the linear model

```
> anova(model)
Analysis of Variance Table

Response: X2
            Df  Sum Sq   Mean Sq F value    Pr(>F)
pop          2 0.05413 0.0270657  42.836 < 2.2e-16 ***
Residuals 1497 0.94587 0.0006318
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

The statistic $F = MS_{pop}/MS_{residuos}$ measures the separability between populations that is if the differences between subjects in one population are lower than the differences between populations.

Continuous number of groups

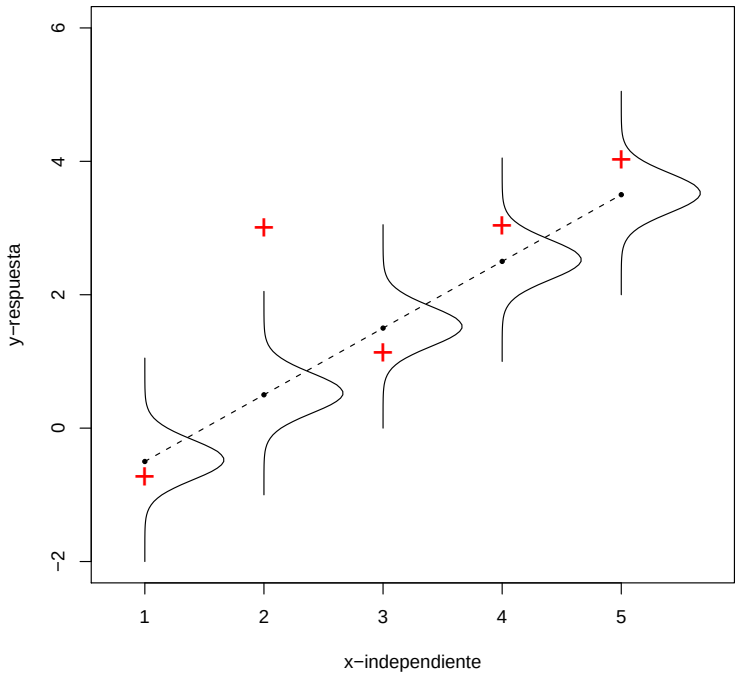
In a regression analysis each group is characterized by a continuous variable x.

$$\mu_x = \alpha + \beta x \tag{2}$$

Typically there is only one observation per group

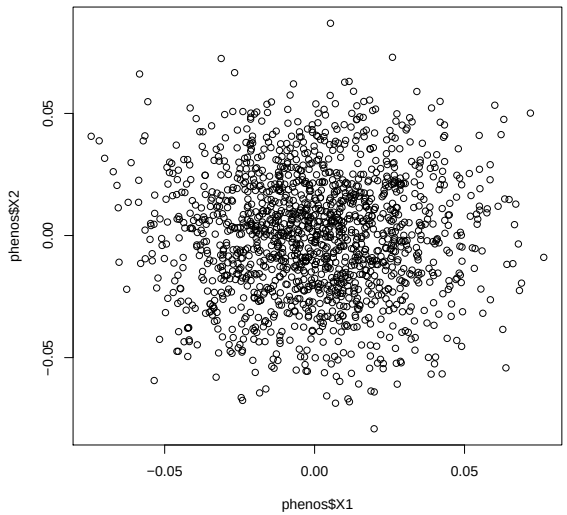
$$y_i \sim \alpha + \beta x_i + \epsilon_i \tag{3}$$

We want to test if the means are significantly described by β, or that β is not 0.



Continuous number of groups

Given that the ratio β between *y* and *x* is fixed, the regression analysis tests the relationship between the variables. We had seen the scatter plot between variables X1 and X2 from our dataset

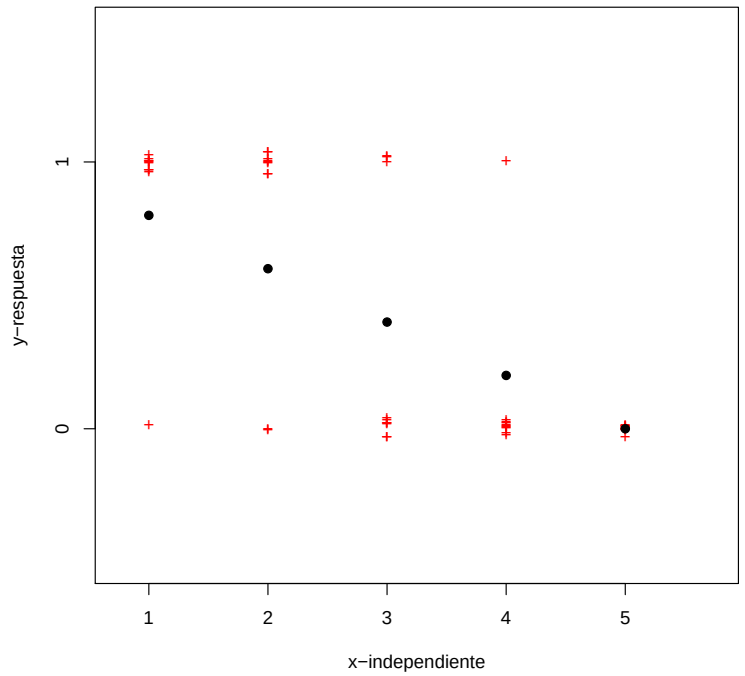


What would be the β in a regression model between these two variables? If I fix x what is its mean? how does the mean varies as I take different values of x?

```
model<-lm(X1~X2,dat=phenos)
```

Generalized linear models

So far we have assumed that the dependent variable is continuous and distributes like a Gaussian. There is an important set of experiments in which the response variable is binary (case/control).



Logistic Regression

In these cases we suppose that the variable *y* follows a binomial distribution (think of tossing a coin) then the model over the mean (proportion of heads/tails) may depend on a variable *x*

$$f(\mu) \sim \alpha + \beta x \tag{4}$$

where *f* is the *logit* function

$$logit(\mu) \sim \log(\frac{\mu}{1-\mu}) \tag{5}$$

the *logit* is a link function for binary variables that casts the regression problem into a linear regression.

Logistic Regression

In R the function *glm* generalizes *lm* for the binomial case like:

```
model<-glm(caco~X2,dat=phenos, family="binomial")
```

Question:

- How would the expression be for the regression between X1 and X2 (continuous variables)?

Correlation Measurements

Let us assume that the have a pair of observations X and Y for the same individual. The correlation between the variables is defined as

$$cor = \frac{cov(XY)}{var(X)var(Y)} \tag{6}$$

$$\tag{7}$$

where

$$cov(XY) = E[(X - \mu_x) * (Y - \mu_y)] \tag{8}$$

$$var(X) = cov(XX) \tag{9}$$

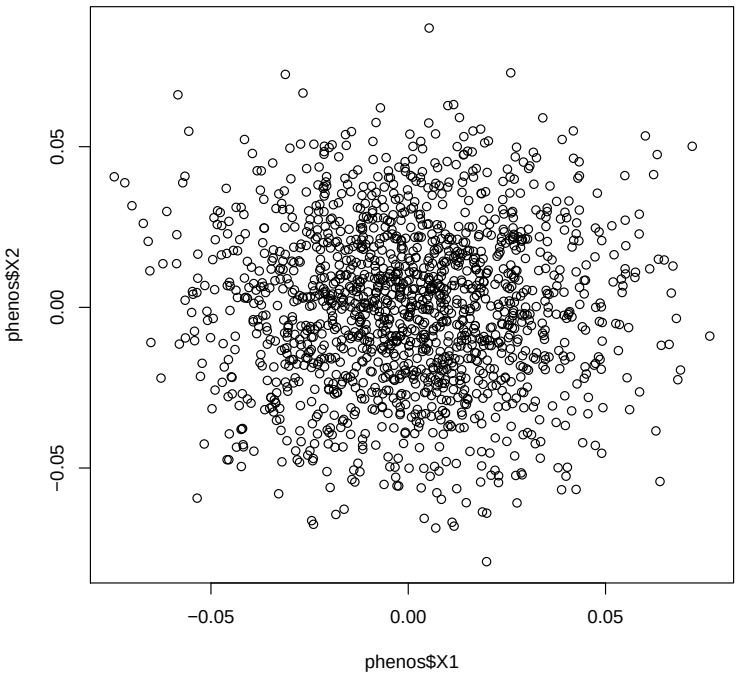
So

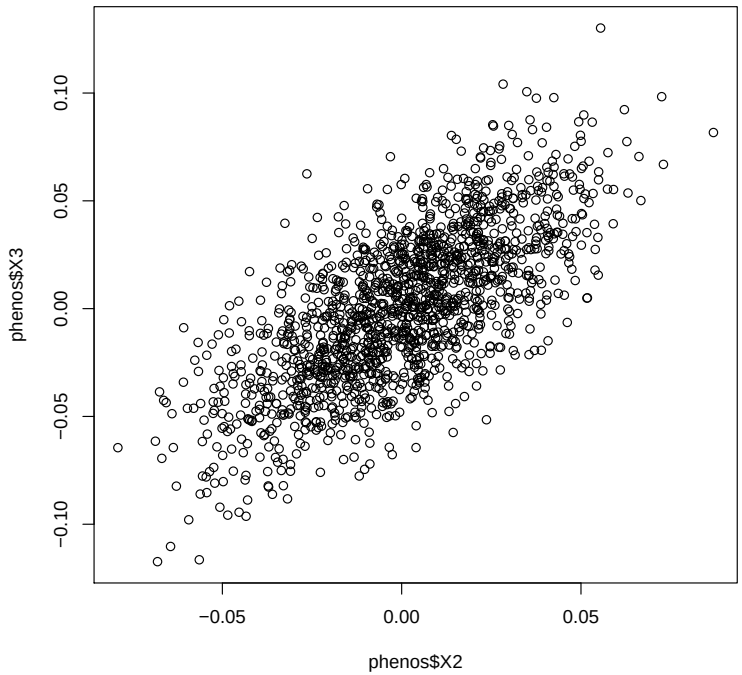
- if $X = Y$ then $cor = 1$
- if X and Y are independent then $cor = 0$ because

$$cov(XY) = E[YX] - \mu_x \mu_y = 0 \tag{10}$$

Correlation measurements

If X1 y X2 are independent the mean of X2 for a fixed X1 is independent of the value of X1. In the case of dependence the mean of X2 given X1 varies with different values of X1.

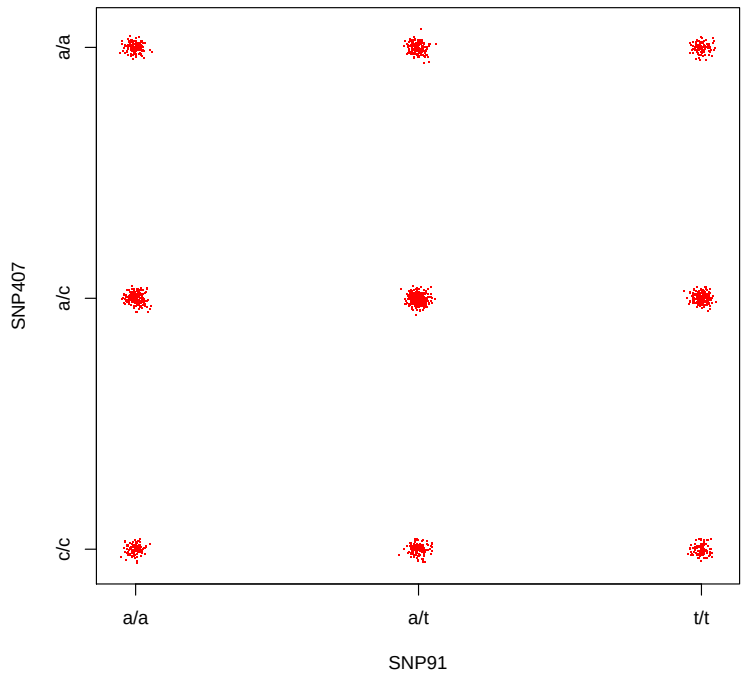


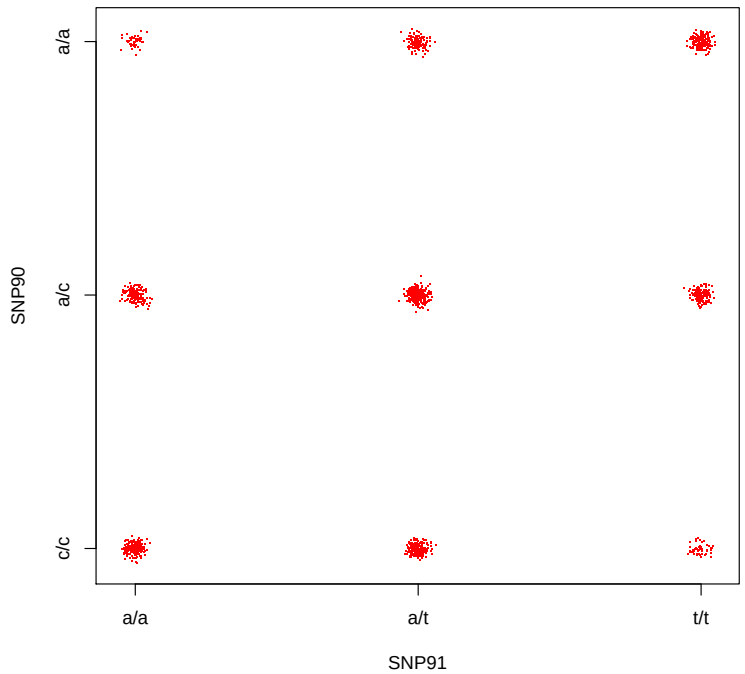


In R we can test the magnitude and significance of a correlation with `cor.test()`. What are the correlations and p-values for the plots above?

Correlation measurements

Let us suppose that the pair of measurements for each subject are from categorical variables (discreet) X and Y, with three levels (possible outcomes) each. En the case of independence, for any fixed value of $SNP91$ the mean of $SNP407$ does not change, which is not the case for $SNP90$.





Chi-squared

A measurement of statistical dependence between categorical variables is given by the chi-squared test.The statistics measures the deviation of our data from the case of no-independence.

$$chi = \Sigma_{i=celdas} \frac{(O_i - E_i)^2}{E_i} \tag{11}$$

Practice: Read the data “datGenoNum.txt”

```
genos<-read.table("datGenoNum.txt",header=TRUE)
```

Compute `chisq.test()` between `genos$SNP91` and `genos$SNP90`; and between `genos$SNP91` and `genos$SNP407`.

Chi-squared

The chi-squared test can be used if two variables have different number of categories (possible outcomes)

See for instance

```
> table(genos$SNP91,phenos$caco)

     0   1
0 166 286
1 316 340
2 172 220

> prop.table(table(genos$SNP91,phenos$caco),2)

           0         1
0 0.2538226 0.3380615
1 0.4831804 0.4018913
2 0.2629969 0.2600473
```

Chi-squared

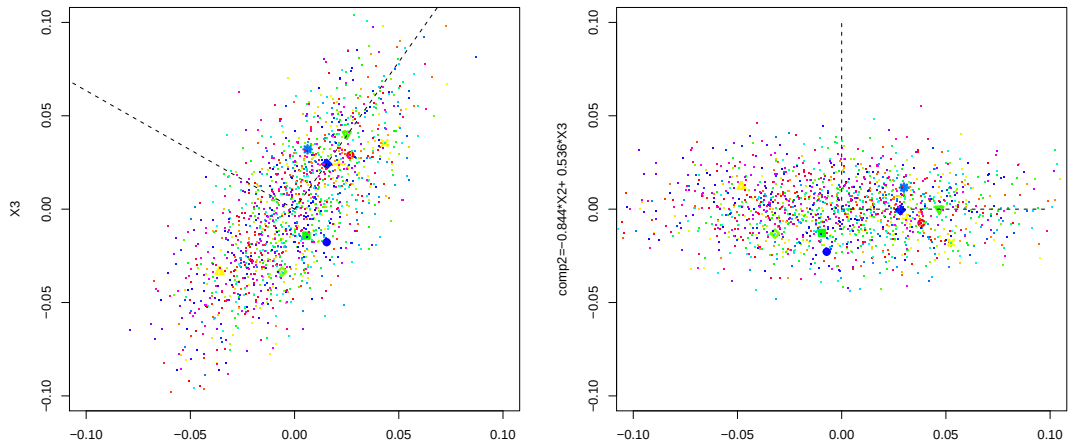
While the proportion of 0 and 1 in category 2 are similar, the full distribution of 0 and 1s for all 0,1 and 2 is not. This is captured by the chi.squared test between `caco` and `SNP91`

```
> chisq.test(genos$SNP91,phenos$caco)

Pearson's Chi-squared test

data:  genos$SNP91 and phenos$caco
X-squared = 14.2718, df = 2, p-value = 0.000796
```

Principal component analysis

PCA finds a liner combination of the variables that explains more variability

```
base<-read.table("phenosCont.txt",header=TRUE)
base2var<-base[,c(4,5)]
plot(base2var,xlim=c(-0.1,0.1),ylim=c(-0.1,0.1))
pc<-princomp(base2var)
plot(pc$scores,xlim=c(-0.1,0.1),ylim=c(-0.1,0.1))
```



If we use the first component, we can reduce the problem from 2 to 1 dimension.

knitr

- A library that processes files .Rnw that mixes LateX and R code.
- It runs R code and in a LateX document with the appropriate sintax
- Plots are inserted
- It has several options to control display
- It saves in cache the code chunks with large amount of computation

knitr

In `minimal.Rnw` is an example of a file that combines R with latex

Lets try to compile it

```
library(knitr)
```

knitr

Open `minimal.Rnw`

```
\begin{document}

\title{TITULO}
\author{A Caceres}
\maketitle

TEXTO LATEX



<< nobre chunk >>=

#CODIGO R

@


\end{document}
```

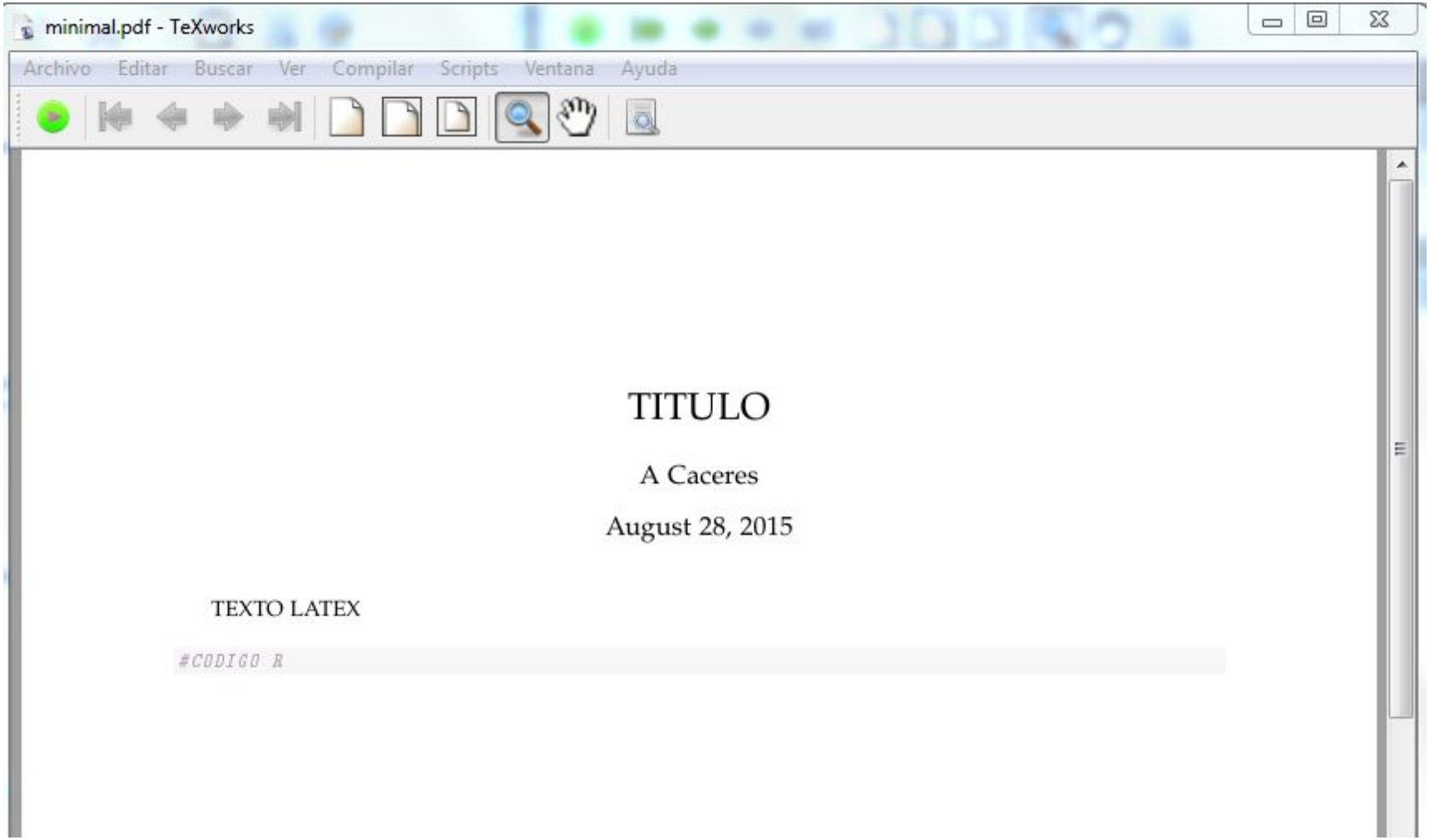
knitr

Edit title, author, text and R code. In R compile with

```
> knit("minimal.Rnw")

##
  |
  |
  |
  |.......................
##   ordinary text without R code
##
##
  |
  |.......................................
## label: unnamed-chunk-6
##
  |
  |.................................................................
##   ordinary text without R code
```

you can compile the .tex file produced with your usual pdflatex compiler. You can use the pdf button in RStudio.

knitr


minimal.pdf - TeXworks
Archivo Editar Buscar Ver Compilar Scripts Ventana Ayuda
TITULO
A Caceres
August 28, 2015
TEXTO LATEX
#CODIGO R

knitr

Chunk options are given as `<< OPCION1, OPCION2, OPCION3, ... >>=`

The most common

- `<< name>>=`
- `<< echo=FALSE>>=` omit the result of the chunk R
- `<< eval=FALSE>>=` do not run the code chunk
- `<< cahce=TRUE>>=` save results of chunk in cache
- `<< size="small">>=` controls font size

knitr

Practice Add a new chunk to $plot((1:10)^2)$