Let us get back on the Titanic dataset,

```
loc_fichier = "http://freakonometrics.free.fr/titanic.RData"
download.file(loc_fichier, "titanic.RData")
load("titanic.RData")
base = base[!is.na(base$Age),]
```

On consider two variables, the age x (the continuous one) and the survivor indicator y (the qualitative one)

```
X = base$Age
Y = base$Survived
```

It looks like the age might be a valid explanatory variable in the logistic regression,

```
summary(glm(Survived~Age,data=base,family=binomial))

Coefficients:
             Estimate Std. Error z value Pr(&gt;|z|)
(Intercept) -0.05672    0.17358  -0.327   0.7438
Age         -0.01096    0.00533  -2.057   0.0397 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 960.23  on 712  degrees of freedom
AIC: 964.23
```

The significance test here has a p-value just below 4%. Actually, one can relate it with the value of the deviance (the null deviance and the residual deviance). Recall that$D=2\big(\log\mathcal{L}(\boldsymbol{y})-\log\mathcal{L}(\widehat{\boldsymbol{\mu}})\big)$while$D_0=2\big(\log\mathcal{L}(\boldsymbol{y})-\log\mathcal{L}(\overline{y})\big)$Under the assumption that x is worthless, D_0-D tends to a χ^2 distribution with 1 degree of freedom. And we can compute the p-value dof that likelihood ratio test,

```
1-pchisq(964.52-960.23,1)
[1] 0.03833717
```

(which is consistent with a Gaussian test). But if we consider a nonlinear transformation

```
summary(glm(Survived~bs(Age),data=base,family=binomial))

Coefficients:
             Estimate Std. Error z value Pr(&gt;|z|)
(Intercept)    0.8648     0.3460   2.500 0.012433 *
bs(Age)1      -3.6772     1.0458  -3.516 0.000438 ***
bs(Age)2       1.7430     1.1068   1.575 0.115299
bs(Age)3      -3.9251     1.4544  -2.699 0.006961 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 948.69  on 710  degrees of freedom
```

which seems to be “more significant”

```
1-pchisq(964.52-948.69,3)
[1] 0.001228712
```

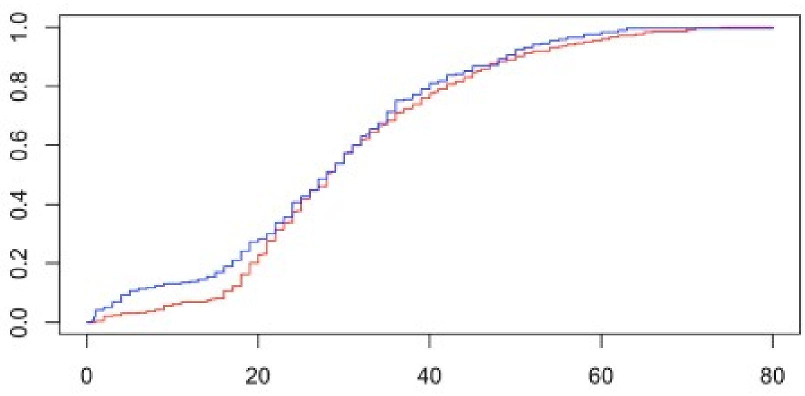
So it looks like the variable x is interesting here.

To visualize the non-null correlation, one can consider the condition distribution of x given y=1, and compare it with the condition distribution of x given y=0,

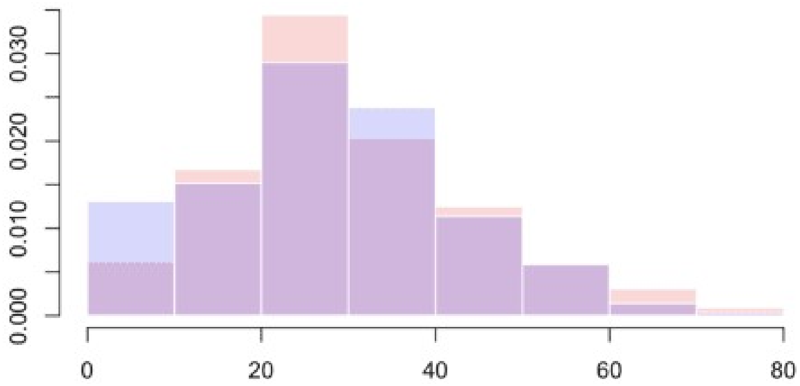
```
ks.test(X[Y==0],X[Y==1])

	Two-sample Kolmogorov-Smirnov test

data:  X[Y == 0] and X[Y == 1]
D = 0.088777, p-value = 0.1324
alternative hypothesis: two-sided
```

i.e. with a p-value above 10%, the two distributions are not significatly different.

```
F0 = function(x) mean(X[Y==0]&lt;=x)
F1 = function(x) mean(X[Y==1]&lt;=x)
vx = seq(0,80,by=.1)
vy0 = Vectorize(F0)(vx)
vy1 = Vectorize(F1)(vx)
plot(vx,vy0,col="red",type="s")
lines(vx,vy1,col="blue",type="s")
```



(we can also look at the density, but it looks like that there is not much to see)

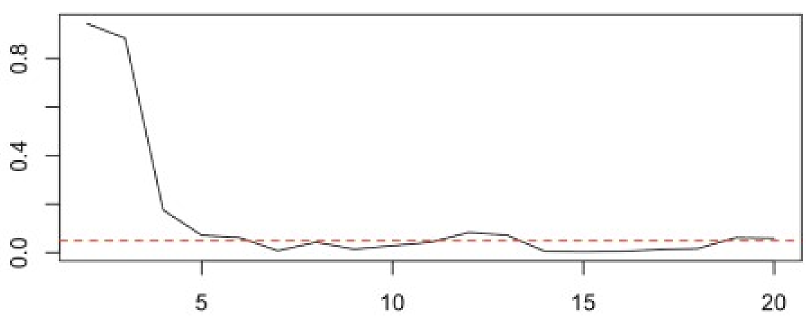


An alternative is discretize variable x and to use Pearson's independence test,

```
k=5
LV = quantile(X,(0:k)/k)
LV[1] = 0
Xc = cut(X,LV)
table(Xc,Y)
            Y
Xc            0  1
  (0,19]     85 79
  (19,25]    92 45
  (25,31.8]  77 50
  (31.8,41]  81 63
  (41,80]    89 53
chisq.test(table(Xc,Y))

        Pearson's Chi-squared test

data:  table(Xc, Y)
X-squared = 8.6155, df = 4, p-value = 0.07146
```

The p-value is here 7%, with five categories for the age. And actually, we can compare the p-value

```
pvalue = function(k=5){
LV = quantile(X,(0:k)/k)
LV[1] = 0
Xc = cut(X,LV)
chisq.test(table(Xc,Y))$p.value}
vk = 2:20
vp = Vectorize(pvalue)(vk)
plot(vk,vp,type="l")
abline(h=.05,col="red",lty=2)
```



which gives a p-value close to 5%, as soon as we have enough categories….