

TD : Compréhension des données

Exercice 1 : Test Khi-carré

On est en présence de deux variables qualitatives : les cheveux et les yeux.

Les différentes modalités de chaque variable sont :

- Pour les cheveux : Blonds, Bruns, Noirs, Roux (quatre modalités)
- Pour les yeux : Bleus, Gris ou verts, Bruns (trois modalités)

Couleurs des cheveux/Couleurs des yeux	Blonds	Bruns	Noirs	Roux	« Total »
Bleus	1768	807	189	47	2811
Gris ou verts	946	1387	746	53	3132
bruns	115	438	288	16	857
« Total »	2829	2632	1233	116	6800

Nous voulons savoir s'il existe une liaison, c'est-à-dire une relation statistique entre la couleur des cheveux et la couleur des yeux ou bien si ces deux caractères sont indépendants c'est-à-dire n'ont aucune relation ou lien entre eux.

Le seuil de signification est fixé à 5%

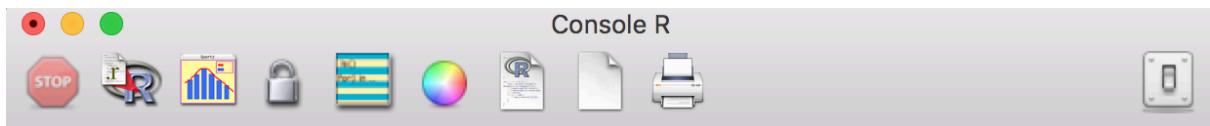
Exercice 2 : t-test

```
Console R

> test.df <- read.table("file:///Users/macbookretina/Documents/Lecture/M1/DM/Rdata/test/
testdf.txt",
+ header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
> test.df
  s.id y x
1    1 -2 0
2    2  0 2
3    3  2 2
4    4  1 5
5    5  3 5
6    6  1 9
7    7  0 9
8    8  0 9
9    9  1 9
10   10 -1 10
```

```
Console R

> vc<-rep(1,10)
> l<-seq(1,nrow(test.df))
> l
[1] 1 2 3 4 5 6 7 8 9 10
> c<-seq(3,ncol(test.df))
> c
[1] 3
> test.df[l,c]
[1] 0 2 2 5 5 9 9 9 9 10
> x.df<-cbind(vc, test.df[l,c])
> x<-as.matrix(x.df)
> x
      vc
[1,] 1 0
[2,] 1 2
[3,] 1 2
[4,] 1 5
[5,] 1 5
[6,] 1 9
[7,] 1 9
[8,] 1 9
[9,] 1 9
[10,] 1 10
> stxx<-solve(t(x)%*%x)
> stxx
      vc
vc 0.39508197 -0.049180328
   -0.04918033 0.008196721
> seb1<-stxx[2,2]
> seb1
[1] 0.008196721
```



Console R

~ Recherche dans l'aide

```
> n<-10
> p<-1
>
> #build regression model
> model <- lm(y ~ x, data = test.df)
> mod_summary <- summary(model)
> mod_summary
```

Call:
lm(formula = y ~ x, data = test.df)

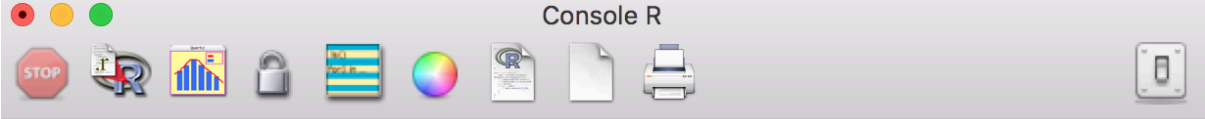
Residuals:

	Min	1Q	Median	3Q	Max
	-2.4016	-0.5492	0.0082	0.5000	2.5164

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.40164	0.95499	0.421	0.685
x	0.01639	0.13755	0.119	0.908

Residual standard error: 1.519 on 8 degrees of freedom
Multiple R-squared: 0.001772, Adjusted R-squared: -0.123
F-statistic: 0.0142 on 1 and 8 DF, p-value: 0.9081



Console R

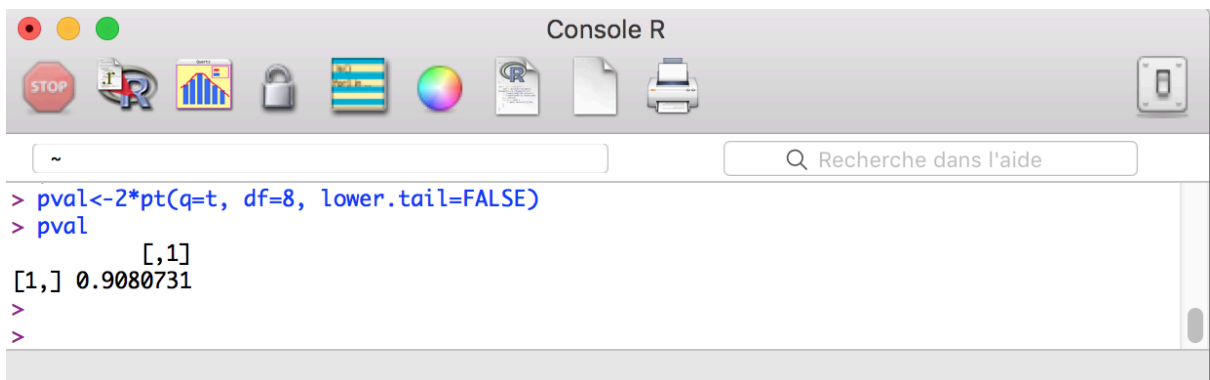
~ Recherche dans l'aide

```
> #calculate residual sum of squares (method 1)
> sse <- deviance(model)
> sse
[1] 18.46721
> #calculate residual sum of squares (method 2)
> sse <- sum(resid(model)^2)
> sse
[1] 18.46721
> se2 <- sse/(n-p-1)
> se2
[1] 2.308402
> cf <- coef(model)
> b0 <- cf[1]
> b0
(Intercept)
0.4016393
> b1 <- cf[2]
> b1
          x
0.01639344
> se <- sqrt(se2*%seb1)
> se
      [,1]
[1,] 0.1375548
> t <- b1/sqrt(se2*%seb1)
> t
      [,1]
[1,] 0.1191775
```

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Calculate the P-Value of a T-Score in R

- To find the p-value associated with a t-score in R, we can use the `pt()` function **`pt(q, df, lower.tail = TRUE)`**
 - **q**: The t-score
 - **df**: The degrees of freedom
 - **lower.tail**: If TRUE, the probability to the left of **q** in the t distribution is returned. If FALSE, the probability to the right is returned. Default is TRUE.



The screenshot shows the R Console window with the following content:

```
> pval<-2*pt(q=t, df=8, lower.tail=FALSE)
> pval
      [,1]
[1,] 0.9080731
>
>
```

Exercice 3 : t-test

Le service des ressources humaines d'une grande entreprise souhaite développer un modèle pour prédire la satisfaction au travail d'un employé à partir du nombre d'heures de travail non rémunéré par semaine, de l'âge de l'employé et du revenu de l'employé.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146

5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Au seuil de signification de 5 %,

1. tester la relation entre la variable dépendante « satisfaction au travail » et la variable indépendante « âge ».
2. tester la relation entre la variable dépendante "satisfaction au travail" et la variable indépendante "revenu"