1 Paradoxe de Simpson

Ref: sciencetonnante

1.1 Fumer est bon pour la santé

Je vous invite à voir (ou revoir) l'excellent film ***merci de fumer*** réalisé par *jason reitman*.

Ref: http://www.biostat.envt.fr/wp-content/uploads/Faouzi/Enseignement/A1ENVT-2015-16-ver01.pdf
1314 femmes ont été suivies pendant 20 ans, et l'objectif était de comparer le taux de mortalité des fumeuses et des non-fumeuses *(données issues du fichier smoking du package SMPractical pour R)*.

âge	vivante	morte
18-24	53	2
25-34	121	3
35-44	95	14
45-54	103	27
55-64	64	51
65-74	7	29
75+	0	13
Total	443	139

TABLE 1: Fumeuse

âge	vivante	morte
18-24	61	1
25-34	152	5
35-44	114	7
45-54	66	12
55-64	81	40
65-74	28	101
75+	0	64
Total	502	230

TABLE 2: Non fumeuse

1. Calculer le pourcentage des personnes décédées parmi les personnes fumeuses puis parmi les personnes non fumeuses.
2. Est il préférable de fumer ou de ne pas fumer ?

1.2 Berkeley

Ref: apprendre-en-ligne
En 1973, l'université américaine de Berkeley (Californie), fut poursuivie pour discrimination envers les filles.
L'étude a été précisée sur les six départements les plus importants, notés ici de A à F :

Département	Admis	Non admis
A	512	313
B	353	207
C	120	205
D	138	279
E	53	138
F	15	256
Total	1192	1398

TABLE 3: Garçon

Département	Admis	Non admis
A	89	19
B	17	8
C	202	391
D	131	244
E	94	299
F	24	317
Total	557	1278

TABLE 4: Fille

1. Calculer le pourcentage d'admis parmi les garçon puis parmi les filles.
2. L'université américaine est elle discriminatoire envers les filles ?

Explication :

➔ l'université américaine de Berkeley : L'explication de ce paradoxe apparent vient quand on regarde le nombre de candidatures dans ces départements. Les femmes semblent avoir tendance à postuler en masse à des départements très sélectifs. Dans ceux-ci, leur taux d'admission est à peine plus faible que celui des hommes. Dans les autres, elles sont plus largement sélectionnées que les hommes.
Quand on fait la moyenne globale, ce sont les départements sélectifs qui ont plus de poids, puisqu'elles y postulent en masse.

2 Paradoxe de Simpson

Ref: sciencetonnante

2.1 Berkeley

Ref: apprendre-en-ligne
En 1973, l'université américaine de Berkeley (Californie), fut poursuivie pour discrimination envers les filles. L'affaire semblait claire : parmi les candidates, seules 30% étaient retenues, alors que 46% des candidatures masculines l'étaient. L'étude a été précisée sur les six départements les plus importants, notés ici de A à F :

Département	Garçons	Admis	Filles	Admises
A	825	62%	108	82%
B	560	63%	25	68 %
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6 %	341	7 %
Total	2590	46%	1835	30%

Ce tableau, si l'on excepte la dernière ligne, ne montre aucune discrimination envers les femmes. Au contraire, le taux d'admission des filles dans le principal département (A) est nettement supérieur à celui des garçons.
L'explication de ce paradoxe apparent vient quand on regarde le nombre de candidatures dans ces départements. Les femmes semblent avoir tendance à postuler en masse à des départements très sélectifs. Dans ceux-ci, leur taux d'admission est à peine plus faible que celui des hommes. Dans les autres, elles sont plus largement sélectionnées que les hommes.
Quand on fait la moyenne globale, ce sont les départements sélectifs qui ont plus de poids, puisqu'elles y postulent en masse.

2.2 Fumer est bon pour la santé

Je vous invite à voir (ou revoir) l'excellent film ***merci de fumer*** réalisé par *jason reitman*.

Ref: http://www.biostat.envt.fr/wp-content/uploads/Faouzi/Enseignement/A1ENVT-2015-16-ver01.pdf
1314 femmes ont été suivies pendant 20 ans, et l'objectif était de comparer le taux de mortalité des fumeuses et des non-fumeuses *(données issues du fichier smoking du package SMPractical pour R)*.

	vivantes	mortes	Total	% mortes
non fumeuses	502	230	732	31,42%
fumeuses	443	139	582	23,88%
Total	945	369	1314	28,08%

Et si on regardait les données par classe d'âge :

âge	fume ?	vivante	morte	% mortes
18-24	fumeuse	53	2	3.64%
18-24	non fumeuse	61	1	1.61%
25-34	fumeuse	121	3	2.42%
25-34	non fumeuse	152	5	3.18%
35-44	fumeuse	95	14	12.84%
35-44	non fumeuse	114	7	05.79%
45-54	fumeuse	103	27	20.77%
45-54	non fumeuse	66	12	15.38%
55-64	fumeuse	64	51	44.35%
55-64	non fumeuse	81	40	33.06%
65-74	fumeuse	7	29	80.56%
65-74	non fumeuse	28	101	78.29%
75+	fumeuse	0	13	100%
75+	non fumeuse	0	64	100%

2.3 Pharmacie

On mène des tests en double aveugle sur un nouveau médicament traitant la maladie grave MG.
On a traité 160 patients, dont 80 ont reçu le médicament, et les 80 autres un placebo.
Le taux de guérison varie selon que l'on considère les malades ayant pris le médicament ou ceux ayant pris le placebo (voir le tableau 1 ci-contre). Parmi les 80 patients ayant pris le médicament, 40 ont été guéris (50 %). Parmi les 80 patients ayant reçu le placebo, seuls 32 ont été guéris (40 %).
Ces résultats suggèrent que le médicament est efficace. Il faut donc le prescrire pour soigner les patients atteints de la maladie MG.
Mais en analysant plus en détail les données et en considérant le sexe des personnes

Total	Guéri	Non guéri	Taux de guérison
Médicament	40	40	50 %
Placebo	32	48	40 %

Hommes	Guéri	Non guéri	Taux de guérison
Médicament	36	24	60%
Placebo	14	6	70%

Femmes	Guéri	Non guéri	Taux de guérison
Médicament	4	16	20 %
Placebo	18	42	30 %

a) Comment raisonner ?

Tout n'est cependant pas réglé pour autant et, dans la réalité, un médecin face aux données des trois tableaux indiqués doit prendre une décision : oui ou non, faut-il prescrire le médicament qui semble efficace (d'après le tableau 1) et qui semble moins bon que le placebo (d'après les tableaux 2 et 3) ?
Imaginez-vous à la place du médecin.
Comment allez-vous raisonner ? Plusieurs attitudes sont possibles.

➔ **Point de vue 1**. Quand j'ignore si c'est un homme ou une femme (car par exemple je suis en train de traiter le cas d'un malade anonyme), je ne tiens compte que de la statistique générale qui est celle qui s'applique dans ce cas. Je donne donc le médicament, car je sais qu'il conduit à la guérison dans 50 % des cas d'après le tableau 1, alors que le placebo ne conduit à la guérison que dans 40 % des cas. En revanche, lorsque j'ai plus de précisions sur la personne à traiter et que je sais s'il s'agit d'un homme ou d'une femme, je regarde la statistique correspondante (donc le tableau 2 ou le tableau 3).

Pour un homme, je donne le placebo, car la statistique du tableau 2 pour les hommes me conseille de donner le placebo. Pour une femme, je donne aussi le placebo, car la statistique du tableau 3 pour les femmes me dit que c'est préférable. L'information sur le sexe du patient me conduit à changer de prescription. Ce n'est pas absurde : l'apport de nouvelles informations justifie souvent de changer ses choix.

Mais à y regarder de près, ce point de vue met mal à l'aise et ne semble pas rationnel.

C'est ce qu'exprime le second point de vue.

➔ **Point de vue 2**. La conclusion du point de vue 1 est absurde. Un patient est soit un homme, soit une femme, et que ce soit l'un ou l'autre quand je connais le sexe du patient la consigne déduite est la même : donner le placebo. Ma décision ne dépend pas du résultat de la question : « S'agit-il d'un homme ou d'une femme ? » Il en résulte donc, de toute évidence, que même quand j'ignore le sexe du patient, je dois donner le placebo et ne pas tenir compte de la statistique générale qui ne sert plus à rien dès l'instant où je dispose des deux statistiques particulières.

On se trouve dans un cas assez usuel : la connaissance de nouvelles données (ici les statistiques particulières concernant séparément les hommes et les femmes) change la conclusion que je faisais avant d'en disposer. Ce changement est général et n'est pas dû à ma connaissance du sexe du patient, mais à ma connaissance des tableaux 2 et 3. Il n'y a rien d'irrationnel à changer d'avis quand on est mieux informé, certes, mais ici le changement doit conduire à oublier le tableau 1 qui ne sert plus à rien quand on a les deux autres.

Notons que, pour tirer une règle de données statistiques, on exige en général d'avoir des effectifs totaux plus élevés que ceux qui apparaissent dans nos tableaux. Mais cela est sans importance pour tout ce que nous venons de dire et pour tout ce que nous dirons plus loin, car on peut très bien imaginer que tous les nombres de nos tableaux sont multipliés par 100 ou même 10 000 et que les déductions que nous faisons des statistiques des tableaux sont de ce fait parfaitement sûres.

➔ **Point de vue 3**.

3 Avec Python

a) Lire un fichier de data R

Par exemple le fichier ***smokingR*** se trouve dans le dossier *dataR*.

```
In [1]: from  rpy2.robjects  import r
In [2]: r['load']('dataR/smoking.rda')
In [3]: dataR = r["smoking"]
In [4]: print(type(dataR))
Out[4]: <class 'rpy2.robjects.vectors.DataFrame'>
In [5]: print(dataR)
Out[5]:       age smoker alive dead
      : 1  18-24      1    53    2
      : 2  18-24      0    61    1
      : 3  25-34      1   121    3
      : 4  25-34      0   152    5
      : 5  35-44      1    95   14
      : 6  35-44      0   114    7
      : 7  45-54      1   103   27
      : 8  45-54      0    66   12
      : 9  55-64      1    64   51
      : 10 55-64      0    81   40
      : 11 65-74      1     7   29
      : 12 65-74      0    28  101
      : 13   75+      1     0   13
      : 14   75+      0     0   64
```

Si on ne dispose pas du fichier ***smoking.rda*** il faut alors le télécharger.
https://github.com/cran/SMPracticals/raw/master/data/smoking.rda

```
In [1]: from  rpy2.robjects  import r
In [2]: from  urllib.request import urlopen
In [3]: url='https://github.com/cran/SMPracticals/raw/master/data/smoking.rda'
In [4]: with urlopen(url) as response:
      :     smokingR = response.read()   #.decode("utf8")
In [5]: fichier = open("smoking.rda", "wb")
In [6]: fichier.write(smokingR)
In [7]: fichier.close()
In [8]: r['load']('smoking.rda')
In [9]: dataR = r["smoking"]
In [10]: print(dataR)
Out[10]:       age smoker alive dead
       : 1  18-24      1    53    2
       : 2  18-24      0    61    1
       : 3  25-34      1   121    3
.....................................
```

Les données peuvent être converties pour être utilisé avec panda :

```
In [11]: from rpy2.robjects import (default_converter, pandas2ri)
In [12]: from rpy2.robjects.conversion import localconverter
In [13]: with localconverter(default_converter + pandas2ri.converter) as cv:
       :     dataP = r["smoking"]
```

```
In [14]: print(type(dataP))
Out[14]: <class 'pandas.core.frame.DataFrame'>
In [15]: print(dataP)
       :       age  smoker  alive   dead
       : 0   18-24     1.0   53.0    2.0
       : 1   18-24     0.0   61.0    1.0
       : 2   25-34     1.0  121.0    3.0
       : 3   25-34     0.0  152.0    5.0
       : ..................................
```

Il ne reste plus qu'à travailler avec ces données avec ***panda***.

```
In [16]: smokeP = dataP[dataP['smoker']==1]
In [17]: nosmokeP = dataP[dataP['smoker']==0]
In [18]: print(smokeP)
Out[18]:       age  smoker  alive  dead
       : 0   18-24     1.0   53.0   2.0
       : 2   25-34     1.0  121.0   3.0
       : 4   35-44     1.0   95.0  14.0
       : 6   45-54     1.0  103.0  27.0
       : 8   55-64     1.0   64.0  51.0
       : 10  65-74     1.0    7.0  29.0
       : 12    75+     1.0    0.0  13.0
In [19]: smokeP[['alive','dead']].sum()
Out[19]: alive    443.0
       : dead     139.0
       : dtype: float64
In [20]: nosmokeP[['alive','dead']].sum()
Out[20]: alive    502.0
       : dead     230.0
       : dtype: float64
```

4 Avec le logiciel R

Ouvrez R-cran

```
install.packages("SMPracticals")
library("ellipse")
library("SMPracticals")
data("smoking")
```

Et enfin pour voir les données :

```
> smoking
     age smoker alive dead
1  18-24      1    53    2
2  18-24      0    61    1
3  25-34      1   121    3
4  25-34      0   152    5
5  35-44      1    95   14
6  35-44      0   114    7
7  45-54      1   103   27
8  45-54      0    66   12
```

```
9  55-64      1    64   51
10 55-64      0    81   40
11 65-74      1     7   29
12 65-74      0    28  101
13   75+      1     0   13
14   75+      0     0   64
> Smoke<-subset(smoking,smoker==1)
> Smoke
     age smoker alive dead
1  18-24      1    53    2
3  25-34      1   121    3
5  35-44      1    95   14
7  45-54      1   103   27
9  55-64      1    64   51
11 65-74      1     7   29
13   75+      1     0   13
> NoSmoke<-subset(smoking,smoker==0)
> NoSmoke
     age smoker alive dead
2  18-24      0    61    1
4  25-34      0   152    5
6  35-44      0   114    7
8  45-54      0    66   12
10 55-64      0    81   40
12 65-74      0    28  101
14   75+      0     0   64
> Smoke[c('alive','dead')]  #  ou Smoke[3:4]
   alive dead
1     53    2
3    121    3
5     95   14
7    103   27
9     64   51
11     7   29
13     0   13
> colSums(Smoke[c('alive','dead')])
alive  dead
  443   139
> colSums(NoSmoke[c('alive','dead')])
alive  dead
  502   230
```