Sophie Baillargeon 12 février 2014

Solutions, série d'exercices chapitre 2

Exercices sur les tableaux $I \times J$

1.

$$\begin{aligned} P[\mu_1 = n_1, \ldots, \mu_m = n_m | \mu_\cdot = n] &= \frac{P[\mu_1 = n_1, \ldots, \mu_m = n_m]}{P[\mu_\cdot = n]} \\ &= \frac{\prod_{j=1}^m \lambda_j^{n_j} e^{-\lambda_j}/n_j!}{\lambda_\cdot^n e^{-\lambda_\cdot}/n!} = \frac{n!}{n_1! \cdots n_m!} \prod_{j=1}^m \left(\frac{\lambda_j}{\lambda_\cdot}\right)^{n_j}, \end{aligned}$$

car la somme de v.a. Poisson indépendantes suit aussi une loi Poisson de paramètre égal à la somme des paramètres individuels, donc $\lambda_\cdot = \sum_j \lambda_j$. Comme $\sum_j n_j = n$ (sinon la proba. conditionnelle est 0) et $\sum_j \lambda_j/\lambda_\cdot = 1$, alors on a bien la fonction de probabilité conjointe d'une multinomiale$(n; \pi)$, où $\pi_j = \lambda_j/\lambda_\cdot$.

2. Commençons par l'implication la plus facile à prouver, X et Y sont indépendants $\Rightarrow RC = 1$

$$\begin{aligned} X \text{ et } Y \text{ sont indépendants} &\Rightarrow \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \\ &\Rightarrow \pi_{j|i} = \pi_{\bullet j} \\ &\Rightarrow \pi_{j|i=1} = \pi_{j|i=2} \quad \text{équivalent à} \quad \pi_1 = \pi_2 \\ &\Rightarrow \pi_1/(1-\pi_1) = \pi_2/(1-\pi_2) \\ &\Rightarrow \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = RC = 1 \end{aligned}$$

Maintenant, prouvons que $RC = 1 \Rightarrow X$ et Y sont indépendants

$$\begin{aligned} RC = 1 &\Rightarrow \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = 1 \\ &\Rightarrow \pi_1 - \pi_1\pi_2 = \pi_2 - \pi_2\pi_1 \\ &\Rightarrow \pi_1 = \pi_2 \quad \text{équivalent à} \quad \pi_{j|i=1} = \pi_{j|i=2} \\ &\Rightarrow \pi_{j|i=1} = \pi_{j|i=2} = \pi_{\bullet j} \\ &\Rightarrow X \text{ et } Y \text{ sont indépendants} \end{aligned}$$

L'avant-dernière ligne a déjà été prouvée dans les notes lorsque l'on a prouvé que l'homogénéité des sous-populations impliquait l'indépendance (p.79).

3. (a) $X|Y$

   (b) $((0.91)/(1-0.91))/((0.17)/(1-0.17)) = 49.36601$, donc les chances qu'une victime noire ait un meurtrier noir sont beaucoup, beacoup plus élevées que les chances qu'une victime blanche ait un meurtrier noir.

   (c) P(Y=blanche|X=blanc) = P(X=blanc|Y=blanche) P(Y=blanche) / P(X=blanc)
   où P(X=blanc) = P(X=blanc|Y=blanche)P(Y=blanche) + P(X=blanc|Y=noire)P(Y=noire)
   et P(Y=noire) = 1 - P(Y=blanche) et P(X=blanc|Y=noire) = 1 - P(X=noire|Y=noire).
   Donc, P(Y=blanche|X=blanc) = 0.83 P(Y=blanche) / ((0.83)*P(Y=blanche) + 0.09*(1-P(Y=blanche))).
   On n'arrive pas à éliminer P(Y=blanche) dans cette formule. Il nous manque donc la proportion de victimes blanches (ou noires car P(Y=blanche) = 1 - P(Y=noire)).

4. (a) Étude transversale

   (b) multinomial simple

   (c) Test de Wald de comparaison de deux proportions (on aurait aussi pu faire le test score)
   $\hat{\pi}_{1|i=1} = n_{11}/n_{1\cdot} = 2/3$, $\hat{\pi}_{1|i=2} = n_{21}/n_{2\cdot} = 1/5$. On a donc

$$z_w = (2/3 - 1/5)/\sqrt{(2/3)(1/3)/750 + (1/5)(4/5)/500} = 18.8,$$

   et $P(|N(0,1)| \geq 18.8) = 2 * (P(N(0,1) \geq 18.8)) \approx 0$, donc clairement, le risque d'échec est significativement différent dans les deux groupes.
   On pourrait aussi faire un test d'homogénéité de sous-populations avec la statistique du khi-deux de Pearson ou du rapport de vraisemblance. Voici le détail des calculs de la statistique du rapport de vraisemblance.

$$\begin{aligned} G^2_{obs} &= 2 * \left(500 * \ln\left(\frac{500}{(750 \times 600)/1250}\right) + 250 * \ln\left(\frac{250}{(750 \times 650)/1250}\right) \right. \\ &\quad \left. + 100 * \ln\left(\frac{100}{(500 \times 600)/1250}\right) + 400 * \ln\left(\frac{400}{(500 \times 650)/1250}\right)\right) = 275.69. \end{aligned}$$

   (d)

$$\begin{aligned} \widehat{RR} &= \hat{\pi}_{1|i=1}/\hat{\pi}_{1|i=2} = (2/3)/(1/5) = 10/3 = 3.333 \\ \text{int. pour } RR &: 3.333 \exp\left(\pm 1.96\sqrt{1/500 - 1/750 + 1/100 - 1/500)}\right) = (2.78, 4.00) \\ \widehat{RC} &= \frac{n_{11}n_{22}}{n_{12}n_{21}} = 8 \\ \text{int. pour } \ln RC &: \ln 8 \pm 1.96\sqrt{1/500 + 1/250 + 1/100 + 1/400} = (1.81, 2.35) \\ \Rightarrow \text{int. pour } RC &: (e^{1.81}, e^{2.35}) = (6.13, 10.44). \end{aligned}$$

   Le risque de subir un échec est plus de 3 fois plus élevé ($RR = 3.33333$) pour ceux qui étudient moins d'une heure que pour ceux qui étudient au moins une heure. La valeur 1 est loin d'être dans l'intervalle de confiance pour RC, ce qui confirme notre test d'hypothèse en (a).

   (e)
```
data poly;
input etude $3. echec $ freq;
datalines;
<1  oui 500
<1  non 250
>=1 oui 100
>=1 non 400
;
run;
ods exclude FishersExact;
proc freq data=poly order=data;
weight freq;
tables etude*echec / chisq riskdiff (equal) relrisk ;
run;
quit;
```

```
The FREQ Procedure

Table of etude by echec

etude     echec

Frequency|
Percent  |
```

Row Pct / Col Pct	oui	non	Total
<1	500	250	750
	40.00	20.00	60.00
	66.67	33.33	
	83.33	38.46	
>=1	100	400	500
	8.00	32.00	40.00
	20.00	80.00	
	16.67	61.54	
Total	600	650	1250
	48.00	52.00	100.00

Statistics for Table of etude by echec

Statistic	DF	Value	Prob
Chi-Square	1	261.7521	<.0001
Likelihood Ratio Chi-Square	1	275.6937	<.0001
Continuity Adj. Chi-Square	1	259.8858	<.0001
Mantel-Haenszel Chi-Square	1	261.5427	<.0001
Phi Coefficient		0.4576	
Contingency Coefficient		0.4161	
Cramer's V		0.4576	

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.6667	0.0172	0.6329	0.7004	0.6317	0.7004
Row 2	0.2000	0.0179	0.1649	0.2351	0.1658	0.2378
Total	0.4800	0.0141	0.4523	0.5077	0.4520	0.5081
Difference	0.4667	0.0248	0.4180	0.5153		

Difference is (Row 1 - Row 2)

The FREQ Procedure

Statistics for Table of etude by echec

Proportion (Risk) Difference Test
H0: P1 - P2 = 0

Proportion Difference	0.4667
ASE (Sample)	0.0248
Z	18.7980
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.3333	0.0172	0.2996	0.3671	0.2996	0.3683
Row 2	0.8000	0.0179	0.7649	0.8351	0.7622	0.8342
Total	0.5200	0.0141	0.4923	0.5477	0.4919	0.5480
Difference	-0.4667	0.0248	-0.5153	-0.4180		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	8.0000	6.1279	10.4440
Cohort (Col1 Risk)	3.3333	2.7774	4.0006
Cohort (Col2 Risk)	0.4167	0.3732	0.4653

Sample Size = 1250

5. (a) Étude transversale

   (b) multinomial (binomial) multiple

   (c) Test score de comparaison de deux proportions (on aurait aussi pu faire le test de Wald)
   $\hat{\pi}_{1|i=1} = n_{11}/n_{1\cdot} = 0.55$, $\hat{\pi}_{1|i=2} = n_{21}/n_{2\cdot} = 0.432$ et la proba commune $\hat{\pi} = (n_{11}+n_{21})/n = 0.484$.
   On a donc
   $$z_s = (0.55 - 0.432)/\sqrt{(0.484)(0.516)(1/200 + 1/250)} = 2.49,$$
   et $P[|N(0,1)| \geq 2.49] \approx 0.013$, donc le taux de succès est significativement différent entre les deux programmes.
   On aurait aussi pu tester l'homogénéité des deux sous-populations avec la statistique du khi-deux de Pearson ou du rapport de vraissemblance. Voici le détail des calculs de la statistique du khi-deux de Pearson.
   $$X^2_{obs} = \left(\frac{(110-96.89)^2}{96.89} + \frac{(90-103.11)^2}{96.89} + \frac{(108-121.11)^2}{121.11} + \frac{(142-128.89)^2}{128.89}\right) = 6.19.$$
   Ou encore avec la formule abrégée :
   $$X^2_{obs} = \frac{450(110\times 142 - 90\times 108)^2}{200\times 250\times 218\times 232} = 6.19.$$

   (d)
   $$\begin{aligned}
   \widehat{RR} &= \hat{\pi}_{1|i=1}/\hat{\pi}_{1|i=2} = (0.55)/(0.432) = 1.27 \\
   \text{int. pour } RR &: 1.27\exp\left(\pm 1.96\sqrt{1/110 - 1/200 + 1/108 - 1/250}\right) = (1.05, 1.54) \\
   \widehat{RC} &= \frac{n_{11}n_{22}}{n_{12}n_{21}} = 1.61 \\
   \text{int. pour } \ln RC &: \ln 1.61 \pm 1.96\sqrt{1/110 + 1/90 + 1/108 + 1/142} = (0.10, 0.85) \\
   \Rightarrow \text{int. pour } RC &: (e^{0.10}, e^{0.85}) = (1.11, 2.34).
   \end{aligned}$$

La probabilité de trouver un emploi est 27% plus élevé ($RR = 1.27$) pour ceux qui font le programme A que pour ceux qui font le programme B. La valeur 1 est hors de l'intervalle de confiance pour θ, ce qui confirme notre test d'hypothèse en (a).

(e)
```
data emploi;
input prog $ emploi $ freq;
datalines;
A oui 110
A non  90
B oui 108
B non 142
;
run;
ods exclude FishersExact;
proc freq data=emploi order=data;
weight freq;
tables prog*emploi / chisq riskdiff (equal var=NULL) relrisk ;
run;
quit;
```

```
The FREQ Procedure

Table of prog by emploi

prog      emploi

Frequency|
Percent  |
Row Pct  |
Col Pct  |oui     |non     |  Total
---------+--------+--------+
A        |    110 |     90 |    200
         |  24.44 |  20.00 |  44.44
         |  55.00 |  45.00 |
         |  50.46 |  38.79 |
---------+--------+--------+
B        |    108 |    142 |    250
         |  24.00 |  31.56 |  55.56
         |  43.20 |  56.80 |
         |  49.54 |  61.21 |
---------+--------+--------+
Total         218      232      450
            48.44    51.56   100.00


Statistics for Table of prog by emploi

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      6.1944    0.0128
Likelihood Ratio Chi-Square    1      6.2061    0.0127
Continuity Adj. Chi-Square     1      5.7310    0.0167
Mantel-Haenszel Chi-Square     1      6.1807    0.0129
Phi Coefficient                       0.1173
Contingency Coefficient               0.1165
```

```
Cramer's V                          0.1173


                        Column 1 Risk Estimates

                                          (Asymptotic) 95%        (Exact) 95%
                  Risk        ASE     Confidence Limits     Confidence Limits
-------------------------------------------------------------------------------
Row 1           0.5500     0.0352     0.4811     0.6189     0.4782     0.6202
Row 2           0.4320     0.0313     0.3706     0.4934     0.3697     0.4959
Total           0.4844     0.0236     0.4383     0.5306     0.4374     0.5317

Difference      0.1180     0.0471     0.0257     0.2103

                      Difference is (Row 1 - Row 2)
The FREQ Procedure

Statistics for Table of prog by emploi

Proportion (Risk) Difference Test
         H0: P1 - P2 = 0

Proportion Difference      0.1180
ASE (H0)                   0.0474
Z                          2.4889
One-sided Pr >  Z          0.0064
Two-sided Pr > |Z|         0.0128


                        Column 2 Risk Estimates

                                          (Asymptotic) 95%        (Exact) 95%
                  Risk        ASE     Confidence Limits     Confidence Limits
-------------------------------------------------------------------------------
Row 1           0.4500     0.0352     0.3811     0.5189     0.3798     0.5218
Row 2           0.5680     0.0313     0.5066     0.6294     0.5041     0.6303
Total           0.5156     0.0236     0.4694     0.5617     0.4683     0.5626

Difference     -0.1180     0.0471    -0.2103    -0.0257

                      Difference is (Row 1 - Row 2)


           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.6070        1.1051        2.3369
Cohort (Col1 Risk)             1.2731        1.0533        1.5388
Cohort (Col2 Risk)             0.7923        0.6568        0.9557

Sample Size = 450
```

6. Les données ont la forme suivante :

	Triché université	Pas triché université
Triché secondaire	20	5
Pas triché secondaire	30	45

(a) Étude cas-témoins

(b) multinomial (binomial) multiple selon les colonnes

(c) On peut faire un test d'homogénéité des sous-populations avec la statistique du khi-deux de Pearson ou du rapport de vraisemblance. Ces statsitiques sont utilisables dans une étude cas-témoin car elles restent les mêmes peu importe l'ordre des variales dans le tableau de fréquences. Étant donné que les fréquences ne sont pas très grandes ici, on va calculer la statistique du khi-deux de Pearson avec une correction pour la continuité.

$$X^2_{corr,obs} = \frac{100(|20 \times 45 - 5 \times 30| - 100/2)^2}{25 \times 75 \times 50 \times 50} = 10.45$$

Cette valeur observées est nettement supérieur à la valeur critique $\chi^2_{1,0.05} = 3.84$. On rejette donc H_0, ce qui signifie que le risque de tricher à l'université n'est pas le même pour ceux qui ont triché et ceux qui n'ont pas triché au secondaire.

(d) Si les proportions $\pi_{1|i=1}$ et $\pi_{1|i=2}$ sont de faibles valeurs, alors le risque relatif peut être approximé par le rapport de cotes. On peut supposer que c'est le cas ici.

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 6$$

Dans ce cas, on aurait que les étudiants ayant triché au secondaires ont 6 fois plus de chances de tricher à l'université que les étudiants n'ayant jamais triché au secondaire.

(e)
```
data triche;
input second $ univ $ freq;
datalines;
oui oui 20
oui non  5
non oui 30
non non 45
;
run;
ods exclude FishersExact;
proc freq data=triche order=data;
weight freq;
tables second*univ / chisk relrisk ;
run;
quit;
```

```
The FREQ Procedure

Table of second by univ

second     univ

Frequency|
Percent  |
Row Pct  |
Col Pct  |oui     |non     |  Total
---------+--------+--------+
oui      |     20 |      5 |     25
         |  20.00 |   5.00 |  25.00
```

```
         |  80.00 |  20.00 |
         |  40.00 |  10.00 |
---------+--------+--------+
non      |     30 |     45 |     75
         |  30.00 |  45.00 |  75.00
         |  40.00 |  60.00 |
         |  60.00 |  90.00 |
---------+--------+--------+
Total          50       50      100
            50.00    50.00   100.00


Statistics for Table of second by univ

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     12.0000    0.0005
Likelihood Ratio Chi-Square    1     12.6576    0.0004
Continuity Adj. Chi-Square     1     10.4533    0.0012
Mantel-Haenszel Chi-Square     1     11.8800    0.0006
Phi Coefficient                       0.3464
Contingency Coefficient               0.3273
Cramer's V                            0.3464


           Estimates of the Relative Risk (Row1/Row2)

Type of Study                  Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)     6.0000        2.0306       17.7284
Cohort (Col1 Risk)            2.0000        1.4243        2.8084
Cohort (Col2 Risk)            0.3333        0.1490        0.7459

Sample Size = 100
```

7. (a) Il s'agit d'une étude longitudinale. Si l'attribution du traitement aux sujets a été contrôlée par les chercheurs réalisant l'étude, alors c'est un essai clinique. Si on a simplement observé après coup quelle intervention médicale avait été réalisée, alors c'est plutôt une étude de cohorte.

   (b) Test d'homogénéité des populations avec le khi-deux de Pearson :

$$X^2_{obs} = \frac{n\Delta^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{41(21\times 3 - 2\times 15)^2}{23\times 18\times 36\times 5} = 0.599$$

   Sous l'hypothèse nulle d'égalité entre les deux proportions, la loi asymptotique de X^2 est une khi-deux à 1 d.d.l.. Le seuil observé du test vaut $P(\chi^2_1 \geq 0.599) = 0.439$. On accepte donc l'égalité entre les deux proportions.
   Ce test n'est pas tout à fait approprié ici puisque pour la moitié des cellules du tableau, la fréquence espérée est inférieure à 5. La loi asymptotique est mise en doute.
   Si on effectuait une correction pour la continuité (la correction de Yates), on obtiendrait :

$$X^2_{corr,obs} = \frac{n(|\Delta| - \frac{n}{2})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{41(|21\times 3 - 2\times 15| - \frac{41}{2})^2}{23\times 18\times 36\times 5} = 0.086$$

   Avec un seuil observé de $P(\chi^2_1 \geq 0.086) = 0.769$

(c) Sous H_0 en supposant les marges fixes, on a que $\mu_{11} \sim$ Hypergéométrique($n_{\cdot 1} = 36, n_{1\cdot} = 23, n = 41$). On doit donc calculer les probabilité hypergéométriques

$$P(\mu_{11} = w | H_0, n_{\cdot 1} = 36, n_{1\cdot} = 23, n = 41) = \frac{\binom{23}{w}\binom{18}{36-w}}{\binom{41}{36}}$$

pour $w = max(0, 36 + 23 - 41), \ldots, min(36, 23) = 18, \ldots, 23$. Par exemple, pour $w = 20$, la probabilité vaut :

$$\frac{\binom{23}{20}\binom{18}{36-20}}{\binom{41}{36}} = \frac{\left(\frac{23!}{20!3!}\right)\left(\frac{18!}{16!2!}\right)}{\left(\frac{41!}{36!5!}\right)} = \frac{\left(\frac{23\times22\times21}{3\times2\times1}\right)\left(\frac{18\times17}{2\times1}\right)}{\left(\frac{41\times40\times39\times38\times37}{5\times4\times3\times2\times1}\right)} = \frac{1771 \times 153}{749398} = 0.362$$

Le tableau complet de la fonction de masse est le suivant

18	19	20	21	22	23
0.045	0.213	0.362	0.275	0.094	0.011

Le seuil observé du test unilatéral est la probabitlié du résultat observé, plus la probabilité d'un résultat plus extrême : $P(\mu_{11} \geq 21 | H_0$ et marges fixes$) = 0.275 + 0.094 + 0.011 = 0.38$.
(Plus extrême équivaut ici à avoir un μ_{11} plus grand que $n_{11} = 21$ car ça donnerait une probabilité de contrôler le cancer encore plus grande pour ceux ayant eu une chirurgie.)
Avec le "mid p-value", on aurait plutôt un seuil obsevé de $0.275/2 + 0.094 + 0.011 = 0.2425$.
On accepte donc l'hypothèse nulle : la chirurgie n'a pas un meilleur taux de contrôle du cancer que la radiothérapie.
Si on avait fait un test bilatéral, le seuil observé du test aurait été :
$P(\mu_{11} \geq 21, \mu_{11} \leq 19 | H_0$ et marges fixes$) = 0.045 + 0.213 + 0.275 + 0.094 + 0.011 = 1 - 0.362 = 0.638$.

(d)
```
data cancer;
input trait $ controle $ freq @@;
datalines;
chirur oui 21 chirur non 2
radio oui 15 radio non 3
;
run;
proc freq data=cancer order=data;
weight freq;
tables trait*controle / chisq;
run;
```

```
The FREQ Procedure

Table of trait by controle

trait     controle

Frequency|
Percent  |
Row Pct  |
Col Pct  |oui     |non     |  Total
---------+--------+--------+
chirur   |     21 |      2 |     23
         |  51.22 |   4.88 |  56.10
         |  91.30 |   8.70 |
         |  58.33 |  40.00 |
---------+--------+--------+
```

```
radio    |     15 |      3 |     18
         |  36.59 |   7.32 |  43.90
         |  83.33 |  16.67 |
         |  41.67 |  60.00 |
---------+--------+--------+
Total          36        5       41
              87.80    12.20   100.00


Statistics for Table of trait by controle

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.5992    0.4389
Likelihood Ratio Chi-Square    1      0.5948    0.4406
Continuity Adj. Chi-Square     1      0.0860    0.7694
Mantel-Haenszel Chi-Square     1      0.5845    0.4445
Phi Coefficient                       0.1209
Contingency Coefficient               0.1200
Cramer's V                            0.1209

 WARNING: 50% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        21
Left-sided Pr <= F          0.8947
Right-sided Pr >= F         0.3808

Table Probability (P)       0.2755
Two-sided Pr <= P           0.6384

Sample Size = 41
```

8.
```
data projet;
input revenu $ education $ freq @@;
datalines;
faible part_sec 9 faible sec 44 faible part_post-sec 13 faible post-sec 10
moyen part_sec 11 moyen sec 52 moyen part_post-sec 23 moyen post-sec 22
eleve part_sec 9 eleve sec 41 eleve part_post-sec 12 eleve post-sec 27
;
run;
proc freq data=projet order=data;
weight freq;
tables revenu*education / chisk crosslist(stdres) scores=rank measures;
run;
```

```
The FREQ Procedure

                        Table of revenu by education

                                          Std                   Row     Column
revenu    education    Frequency    Residual    Percent    Percent    Percent
```

faible	part_sec	9	0.4061	3.30	11.84	31.03
	sec	44	1.5828	16.12	57.89	32.12
	part_pos	13	-0.1286	4.76	17.11	27.08
	post-sec	10	-2.1078	3.66	13.16	16.95
	Total	76		27.84	100.00	
moyen	part_sec	11	-0.1898	4.03	10.19	37.93
	sec	52	-0.5441	19.05	48.15	37.96
	part_pos	23	1.3042	8.42	21.30	47.92
	post-sec	22	-0.4032	8.06	20.37	37.29
	Total	108		39.56	100.00	
eleve	part_sec	9	-0.1903	3.30	10.11	31.03
	sec	41	-0.9459	15.02	46.07	29.93
	part_pos	12	-1.2374	4.40	13.48	25.00
	post-sec	27	2.4360	9.89	30.34	45.76
	Total	89		32.60	100.00	
Total	part_sec	29		10.62		100.00
	sec	137		50.18		100.00
	part_pos	48		17.58		100.00
	post-sec	59		21.61		100.00
	Total	273		100.00		

Statistics for Table of revenu by education

Statistic	DF	Value	Prob
Chi-Square	6	8.8709	0.1810
Likelihood Ratio Chi-Square	6	8.9165	0.1783
MH Chi-Square (Rank Scores)	1	3.9943	0.0457
Phi Coefficient		0.1803	
Contingency Coefficient		0.1774	
Cramer's V		0.1275	

The FREQ Procedure

Statistics for Table of revenu by education

Statistic	Value	ASE
Gamma	0.1625	0.0795
Kendall's Tau-b	0.1076	0.0530
Stuart's Tau-c	0.1064	0.0525
Somers' D C|R	0.1075	0.0530
Somers' D R|C	0.1076	0.0530

```
Pearson Correlation (Rank Scores)     0.1212    0.0597
Spearman Correlation                  0.1212    0.0600

Lambda Asymmetric C|R                 0.0000    0.0000
Lambda Asymmetric R|C                 0.0303    0.0418
Lambda Symmetric                      0.0166    0.0230

Uncertainty Coefficient C|R           0.0134    0.0088
Uncertainty Coefficient R|C           0.0150    0.0099
Uncertainty Coefficient Symmetric     0.0141    0.0094

Sample Size = 273
```

(a) $X^2 = 8.9$, d.d.l. $= 6$, seuil observé $= 0.18$; Selon ce test, on ne détecte pas d'association entre le revenu familial et le niveau d'éducation. Les tests du khi-deux de Pearson et du rapport des vraisemblances supposent que les variables à l'étude sont nomiales alors qu'elles peuvent ici être considérées ordinales. On aurait avantage à utiliser le test de Mantel et Haenszel.

(b) Deux résidus ajustés de Pearson sont supérieurs à 1.96. Ces résidus laissent croire que le niveau d'éducation projeté tend à être plus élevé pour les gens provenant d'une famille à revenu plus élevé.
Les estimations de la probabilité de projeter compléter des études post-secondaires, conditionnellement au revenu familial, sont 0.1316, 0.2037 et 0.3034 pour un ravenu familial faible, moyen et élevé respectivement. On voit encore que plus le revenu familial est élevé, plus la proportion de jeunes désirant compléter des étaudes post-secondaires est grande.
Le coefficient de Cramer est ici positif (pour un tableau 2×2, il peut prendre une valeur entre -1 et 1), il témoigne donc d'une association positive entre les variables.

(c) Nous allons utiliser une corrélation de Spearman car il est difficile ici de déterminer un score numérique représentatif de la réalité pour remplacer les modalités des variables. Le test de Mantel et Haenszel avec cette corrélation donne $M^2 = 3.9943$, d.d.l. $= 1$, seuil observé $= 0.0457$. On rejette maintenant l'hypothèse de non -association entre les variables. Le coefficient de corrélation de Spearman est $r_s = 0.1212$. Le test nous dit que cette valeur est significativement différente de zéro. Étant donné qu'elle est positive, on conlut ici que plus leur revenu familial est élevé, plus les jeunes planifient étudier longtemps.

Note concernant SAS
Ici, nous avons choisi d'utiliser la corrélation de Spearman car elle était plus pertinente. Mais si nous avions choisi la corrélation de Pearson, il est intéressant de savoir quel score aurait utilisé SAS étant donné que les variables sont entrées sous une forme non numérique.

```
proc freq data=projet order=data;
weight freq;
tables revenu*education / nopercent norow nocol chisk measures ;
run;
```

```
The FREQ Procedure

Table of revenu by education

revenu     education

Frequency|part_sec|sec     |part_pos|post-sec|  Total
---------+--------+--------+--------+--------+
faible   |      9 |     44 |     13 |     10 |     76
---------+--------+--------+--------+--------+
moyen    |     11 |     52 |     23 |     22 |    108
```

```
---------+--------+--------+--------+--------+
eleve    |      9 |     41 |     12 |     27 |     89
---------+--------+--------+--------+--------+
Total          29      137       48       59      273


Statistics for Table of revenu by education

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     6      8.8709    0.1810
Likelihood Ratio Chi-Square    6      8.9165    0.1783
Mantel-Haenszel Chi-Square     1      4.7489    0.0293
Phi Coefficient                       0.1803
Contingency Coefficient               0.1774
Cramer's V                            0.1275


Statistic                                Value       ASE
------------------------------------------------------
Gamma                                   0.1625    0.0795
Kendall's Tau-b                         0.1076    0.0530
Stuart's Tau-c                          0.1064    0.0525

Somers' D C|R                           0.1075    0.0530
Somers' D R|C                           0.1076    0.0530

Pearson Correlation                     0.1321    0.0594
Spearman Correlation                    0.1212    0.0600

Lambda Asymmetric C|R                   0.0000    0.0000
Lambda Asymmetric R|C                   0.0303    0.0418
Lambda Symmetric                        0.0166    0.0230

Uncertainty Coefficient C|R             0.0134    0.0088
Uncertainty Coefficient R|C             0.0150    0.0099
Uncertainty Coefficient Symmetric       0.0141    0.0094

Sample Size = 273
```

On constate que par défaut, SAS associe les scores 1 à k pour les k valeurs possibles d'une variable caractère, en respectant l'ordre spécifié par l'option `order` de l'énoncé `proc`. C'est la valeur par défaut de l'option `scores` dans l'énoncé `tables` qui explique ça. On obtient donc la même statitsique M^2 que si nous avions définis les scores suivants :
revenu : 1 = faible, 2=moyen, 3=élevé
education : 1 = études secondaires partielles, 2 = diplôme d'études secondaires, 3 = études post-secondaires partielles, 4 = diplôme d'études post-secondaires

```
data aspiration;
input income education count @@;
datalines;
1 1 9 1 2 44 1 3 13 1 4 10
2 1 11 2 2 52 2 3 23 2 4 22
3 1 9 3 2 41 3 3 12 3 4 27
;
```

```
run;
proc freq data=aspiration order=data;
weight count;
tables income*education / nopercent norow nocol chisk ;
run;
```

```
The FREQ Procedure

Table of income by education

income     education

Frequency|       1|       2|       3|       4|  Total
---------+--------+--------+--------+--------+
       1 |      9 |     44 |     13 |     10 |     76
---------+--------+--------+--------+--------+
       2 |     11 |     52 |     23 |     22 |    108
---------+--------+--------+--------+--------+
       3 |      9 |     41 |     12 |     27 |     89
---------+--------+--------+--------+--------+
Total          29      137       48       59      273


Statistics for Table of income by education

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     6      8.8709    0.1810
Likelihood Ratio Chi-Square    6      8.9165    0.1783
Mantel-Haenszel Chi-Square     1      4.7489    0.0293
Phi Coefficient                       0.1803
Contingency Coefficient               0.1774
Cramer's V                            0.1275

Sample Size = 273
```

9. (a) On a $\bar{x} = 2.02$ et $\bar{y} = 2.19$. A partir des données, on trouve

$$r_p = \frac{-3.95}{\sqrt{98.8 \times 104.70}} = -0.038$$

   (b) On a $\bar{x} = 44017$ et $\bar{y} = 3.156$. On trouve un coefficient de correlation de Pearson de -0.055. Le choix des scores fait donc varier la valeur de la corrélation.

   (c) Les rangs moyens pour X sont (24,84.5,147.5) et les rangs moyens pour Y sont (20,70.5,137.5). On trouve un coefficient de correlation de -0.036.

   (d) Pour répondre à cette question, on doit effectuer le test de Mantel et Haenszel. Pour la première corrélation, la statistique du test est : $M^2 = (n-1)r^2 = (173-1) \times (-0.038)^2 = 0.25974$. Sous l'hypothèse nulle d'absence d'association entre X et Y, cette statistique suit une loi du khi-deux à 1 d.d.l.. Le quantile 0.95 de cette loi vaut : $\chi^2_{1,0.95} = 3.841459$. Étant donné que la valeur de M^2 est nettement inférieure à $\chi^2_{1,0.95}$, on ne peut rejetter l'hypothèse nulle. On conlut donc que l'opinion des citoyens concernant les fusions ne dépend pas de leur revenu.
   On arrive à la même conclusion avec les deux autres corrélations (pour la corrélation de Pearson avec les nouveaux scores on obtient $M^2 = 0.5166$ et pour la corrélation de Spearman on a $M^2 = 0.22852$).

(e)
```
data fusion1;
input revenu accord freq @@;
datalines;
1 1 10 1 2 18 1 3 19
2 1 15 2 2 26 2 3 33
3 1 14 3 2 18 3 3 20
;
run;
proc freq data=fusion1 order=data;
weight freq;
tables revenu*accord / noprint chisk measures;
run;
```

The FREQ Procedure

Statistics for Table of revenu by accord

Statistic	DF	Value	Prob
Chi-Square	4	1.0549	0.9014
Likelihood Ratio Chi-Square	4	1.0337	0.9046
Mantel-Haenszel Chi-Square	1	0.2598	0.6103
Phi Coefficient		0.0781	
Contingency Coefficient		0.0778	
Cramer's V		0.0552	

Statistic	Value	ASE
Gamma	-0.0497	0.1047
Kendall's Tau-b	-0.0324	0.0683
Stuart's Tau-c	-0.0316	0.0666
Somers' D C|R	-0.0322	0.0680
Somers' D R|C	-0.0325	0.0685
Pearson Correlation	-0.0389	0.0764
Spearman Correlation	-0.0364	0.0763
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0028	0.0055
Uncertainty Coefficient R|C	0.0028	0.0055
Uncertainty Coefficient Symmetric	0.0028	0.0055

Sample Size = 173

```
data fusion2;
input revenu accord freq @@;
datalines;
20000 0 10 20000 3 18 20000 5 19
37500 0 15 37500 3 26 37500 5 33
75000 0 14 75000 3 18 75000 5 20
```

```
;
run;
proc freq data=fusion2 order=data;
weight freq;
tables revenu*accord / noprint chisq measures;
tables revenu*accord / noprint scorout chisq scores=rank ;
run;
```

The FREQ Procedure

Statistics for Table of revenu by accord

Statistic	DF	Value	Prob
Chi-Square	4	1.0549	0.9014
Likelihood Ratio Chi-Square	4	1.0337	0.9046
Mantel-Haenszel Chi-Square	1	0.5166	0.4723
Phi Coefficient		0.0781	
Contingency Coefficient		0.0778	
Cramer's V		0.0552	

Statistic	Value	ASE
Gamma	-0.0497	0.1047
Kendall's Tau-b	-0.0324	0.0683
Stuart's Tau-c	-0.0316	0.0666
Somers' D C|R	-0.0322	0.0680
Somers' D R|C	-0.0325	0.0685
Pearson Correlation	-0.0548	0.0772
Spearman Correlation	-0.0364	0.0763
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0028	0.0055
Uncertainty Coefficient R|C	0.0028	0.0055
Uncertainty Coefficient Symmetric	0.0028	0.0055

Sample Size = 173
The FREQ Procedure

Scores for Table of revenu by accord
Score Type = RANK

Row Scores

revenu	Score
20000	24
37500	84.5
75000	147.5

```
   Column Scores

accord       Score
------------------
     0          20
     3        70.5
     5       137.5
The FREQ Procedure

Statistics for Table of revenu by accord

Statistic                     DF       Value      Prob
-------------------------------------------------------
Chi-Square                     4      1.0549    0.9014
Likelihood Ratio Chi-Square    4      1.0337    0.9046
MH Chi-Square (Rank Scores)    1      0.2285    0.6326
Phi Coefficient                       0.0781
Contingency Coefficient               0.0778
Cramer's V                            0.0552

Sample Size = 173
```

10. (a) Pour répondre à cette question, on doit faire le test de McNemar. La statistique du test prend la valeur observée suivante : $(n_{12} - n_{21})^2/(n_{12} + n_{21}) = 62.88$. Le seuil observé du test est 2.22×10^{-15}, donc on conclut que la différence est très significative entre les probabilités marginles. Les adolescents jugés par le tribunal pour adultes sont beaucoup plus souvent ré-arrêtés ($(673/2097) \times 100\% = 32,09\%$ du temps comparativement à $(448/2097) \times 100\% = 21,36\%$ du temps).

    (b) Estimons d'abord la proportion d'accord observée : $(158 + 1134)/2097 = 0.616$. Cette proportion d'accord n'est pas si faible, mais les cas de désaccord ne sont pas répartis de façon aléatoire en haut et en bas de la diagonale dans le tableau de fréquences. Il y a une plus de ré-arrestation avec le tribunal pour adultes (test en (a)). On voit donc un désaccord (ou une non concordance) entre les variables.
    D'ailleurs, le kappa de Cohen détecte ici ce manque d'accord. L'estimation de cette mesure prend la valeur observée $(0.616-0.603)/(1-0.603) = 0.034$ car on estime la porporiton d'accord aléatoire par $(673 \times 448 + 1424 \times 1649)/2097^2 = 0.603$. Une valeur de kappa aussi faible est typiquement interpréter comme un mauvais accord entre les variables.

    (c)
```
data rearrest;
input adultes $ enfants $ freq @@;
datalines;
oui oui 158 oui non  515
non oui 290 non non 1134
;
proc freq data=rearrest order=data; weight freq;
tables adultes*enfants / agree;
run;
```

```
The FREQ Procedure

Table of adultes by enfants

adultes     enfants

Frequency|
```

Percent / Row Pct / Col Pct	oui	non	Total
oui	158	515	673
	7.53	24.56	32.09
	23.48	76.52	
	35.27	31.23	
non	290	1134	1424
	13.83	54.08	67.91
	20.37	79.63	
	64.73	68.77	
Total	448	1649	2097
	21.36	78.64	100.00

Statistics for Table of adultes by enfants

McNemar's Test

Statistic (S)	62.8882
DF	1
Pr > S	<.0001

Simple Kappa Coefficient

Kappa	0.0341
ASE	0.0214
95% Lower Conf Limit	-0.0078
95% Upper Conf Limit	0.0761

Sample Size = 2097