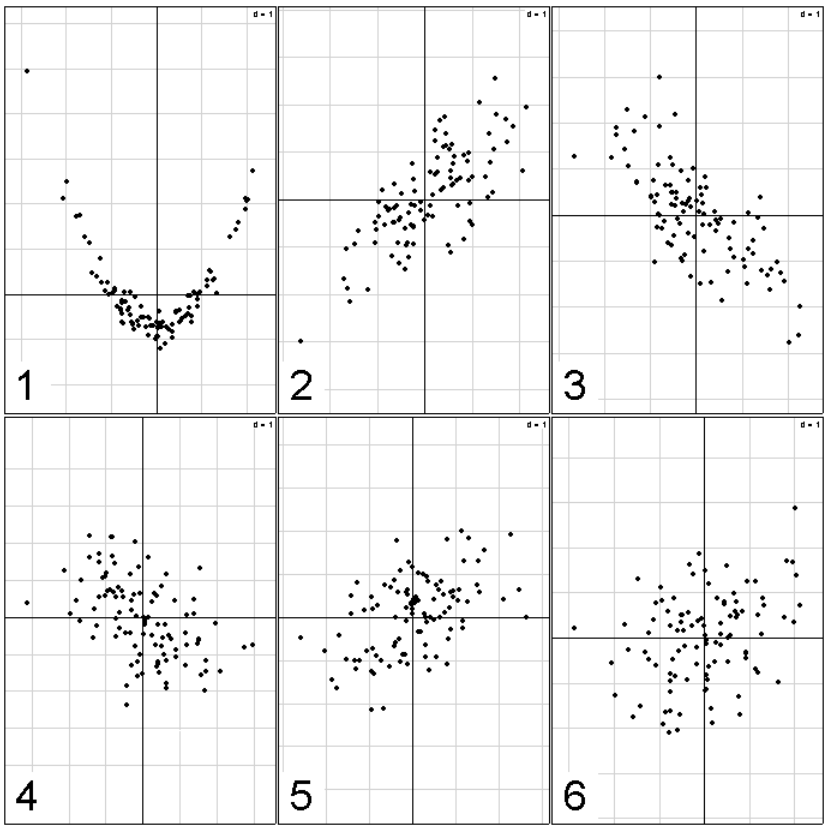
LICENCE BO - UE BMS - 06/2002 (2 HEURES)

Répondre aux questions *strictement* dans la place impartie et *justifier* simplement vos réponses par un argument qui vous paraît approprié.

Question 1. Six simulations de statistique bivariée donnent les figures :



Attribuer à chacune d'entre elles son coefficient de corrélation sachant que ces valeurs figurent parmi l'ensemble {-1, -0.73, -0.49, -0.04, 0.33, 0.50, 0.74, 1}

Question 2. Dans une grande population, on a déterminé le sexe de 19 individus et trouvé 4 mâles et 15 femelles. La valeur 0.5 appartient-elle à l'intervalle de confiance de la fréquence des mâles au seuil de confiance de 95% ?

Question 3. Six histogrammes ont été tracés sur la figure suivante en utilisant les ordres :

A `x <- rexp(n=1000,rate=2.5) # pour mémoire  mean  = 1/rate`

B `x <- rexp(n=1000,rate=0.5)`

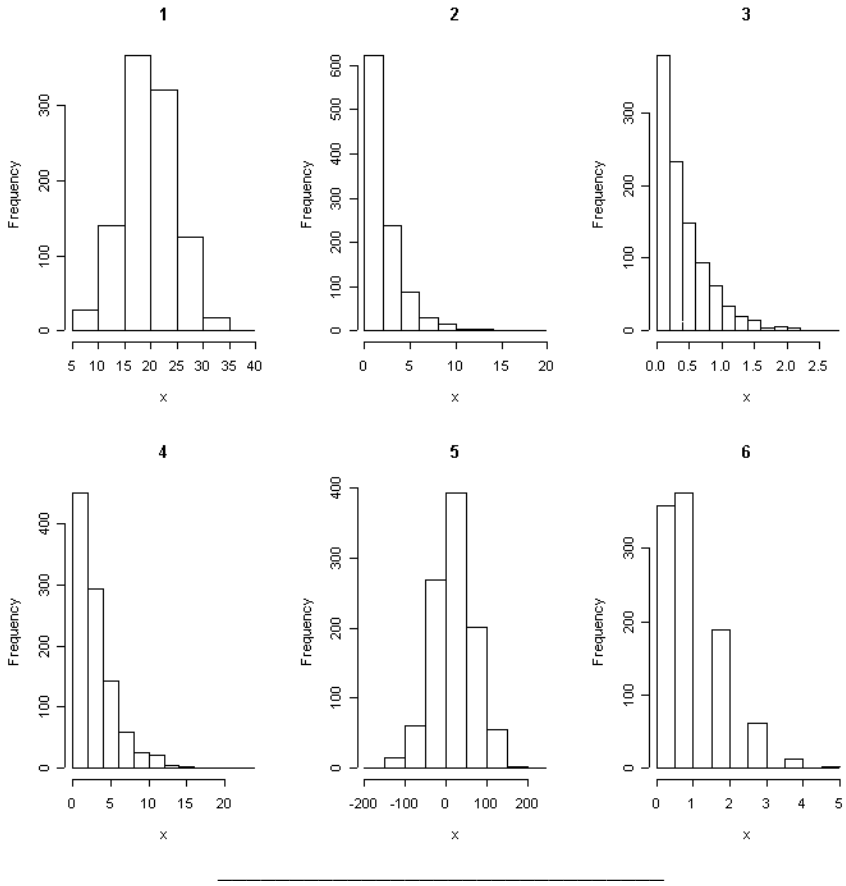
C `x <- rnorm(n=1000,mean=20, sd=5)`

D `x <- rchisq(n=1000,df=3)`

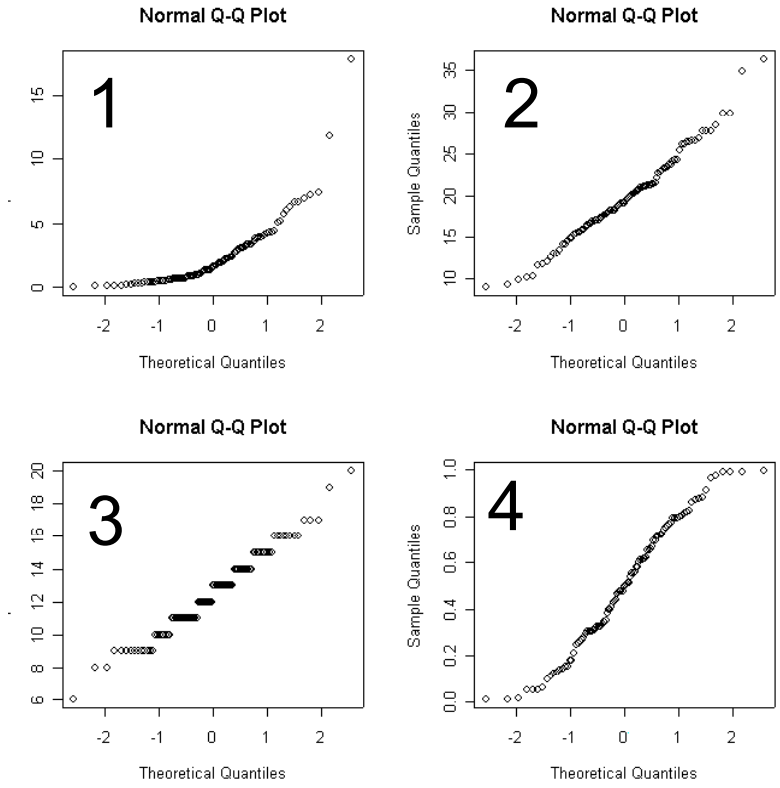
E `x <- rpois(n=1000,lambda=1)`

F `x <- rnorm(n=1000,mean=20, sd=50)`

Indiquer pour chaque histogramme lequel des ordres a été utilisé.



Question 4.



4 graphiques quantiles-quantiles ont été tracés en utilisant

A `x <- runif(100) ; qqnorm(x)`

B `x <- rbinom(100,25,0.5) ; qqnorm(x)`

C `x <- rnorm(100,20,5) ;qqnorm(x)`

D `x <- rexp(100,0.5) ; qqnorm(x)`

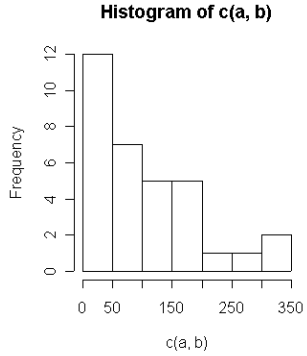
Indiquer pour chaque graphe lequel des ordres a été utilisé.

Question 5

Quelle est la fonction génératrice des probabilités $x \mapsto f(x)$ d'une loi binomiale de paramètres n (nombre d'essais) et p (probabilités de succès) ? Quelle est la valeur de la dérivée de cette fonction pour $x = 1$?

Question 6.

```
> a
 [1] 348 114   8  66 342  24  35   5  19  43 161  14 170  30  18 111  30   1 135
[20]  63
> b
 [1] 230 296  46  83 108 161 200 141  66  52 173  93  83
```




```
> t.test(a,b)

        Welch Two Sample t-test

data:  a and b
t = -1.491, df = 30.4, p-value = 0.1463
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -109.88   17.12
sample estimates:
mean of x mean of y
    86.85    133.23

> wilcox.test(a,b)

        Wilcoxon rank sum test with continuity correction

data:  a and b
W = 69, p-value = 0.02576
alternative hypothesis: true mu is not equal to 0
```

Les deux tests donnent des résultats incohérents. Lequel des deux vous semble le plus approprié ?

Question 7.

```
x_c(6,3)
y_c(21,8,7,14,9)
```

Quelle est la probabilité critique (p-value) attendue de `wilcox.test(x,y)` ?

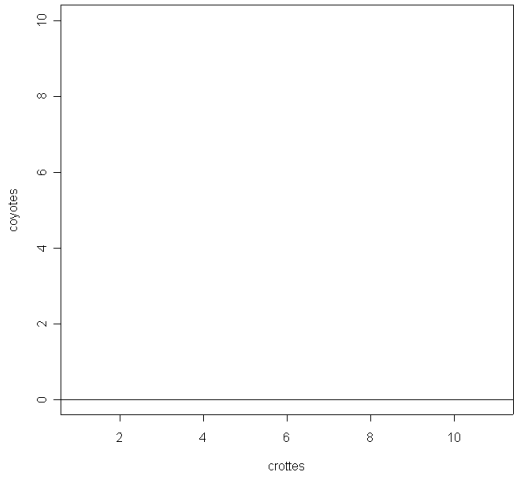
Question 8.

Par séquençage de l'ADN fécal, on a attribué 111 fèces de coyotes *Canis latrans* à 30 individus[1]. 8 individus sont représentés par une déjection, 6 individus sont représentés par 2 déjections, …, 1 individu est représenté par 11 déjections. La distribution complète est :

Crottes	1	2	3	5	6	7	8	9	11
Coyotes	8	6	5	4	2	1	1	2	1

```
dpois(1:20,111/30)*30
 [1] 2.744e+00 5.077e+00 6.262e+00 5.792e+00 4.286e+00 2.643e+00 1.397e+00
 [8] 6.461e-01 2.656e-01 9.828e-02 3.306e-02 1.019e-02 2.901e-03 7.667e-04
[15] 1.891e-04 4.374e-05 9.519e-06 1.957e-06 3.810e-07 7.049e-08
```

Représenter ces informations dans le cadre ci-dessous et commenter le résultat.



Question 9.

Trois populations de chats *Felis catus* en milieu rural échantillonnées pendant plusieurs années ont permis[2] d'examiner 324 chats classés suivant le sexe (F-M), le génotype (Orange-Non Orange), l'âge (3 classes dans l'ordre croissant) et la présence d'anticorps spécifiques du virus FIV (*feline imunodeficiency virus*). Pour chaque combinaison sexe-génotype-âge on a le nombre d'individus séropositifs (fivposi) et le nombre d'individus négatifs (fivnega).

```
> fiv
   fivnega fivposi sex gen age
1       29       1   F  NO  a1
2       32       4   M  NO  a1
3       10       2   F   O  a1
4        7       2   M   O  a1
5       43       0   F  NO  a2
6       36       7   M  NO  a2
7       18       5   F   O  a2
8       16       7   M   O  a2
9       50       4   F  NO  a3
10      16       8   M  NO  a3
11      21       2   F   O  a3
```

---

[1] Kohn, M.H., York, E.C., Kamradt, D.A., Haught, G., Sauvajot, R.M. & Wayne, R.K. (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London B* : 266, 657-663

[2] Pontier, D., Fromont, E., Courchamp, F., Artois, M. & Yoccoz, N.G. (1998) Retroviruses and sexual size dimorphism in domestic cats (Felis catus L.). *Proceedings of the Royal Society of London* B : 265, 167-173.

```
12       2       2   M   O  a3
```

Donner le taux de séropositivité par classe de sexe, d'âge et de génotype. Quel est le facteur qui vous semble le plus important ?

Question 10.

Dans les données originales de Mendel (Mendel, G. (1956) Mathematics of heredity. In : The word of mathematics. Part V. Newman, J.R. (Ed.) Tempus Books of Microsoft Press. 923-934.), on trouve :

Expt. 1. Form of seed. From 253 hybrids, 7324 seeds were obtained in the second trial year. Among them were 5474 round or roundish ones and 1850 angular wrinkled ones. Therefrom the ratio 2.96 to 1 is deduced.

Expt. 2. Colour of albumen. 258 plants yielded 8023 seeds, 6022 yellow and 2001 green; their ratio, therefore is as 3.01 to 1 ...

Expt. 3. Colour of the seed-coats. Among 929 plants 705 bore violet-red flowers and grey-brown seed-coats, giving the proportion 3.15 to 1.

Sachant que :

```
> chisq.test(c(5474,1850),p=c(3/4,1/4))$statistic
X-squared
   0.2629

> chisq.test(c(6022,2001),p=c(3/4,1/4))$statistic
X-squared
    0.015

> chisq.test(c(705,224),p=c(3/4,1/4))$statistic
X-squared
   0.3907

> pchisq(c(seq(0,0.1,by=0.01),1:10),1)

 [1] 0.00000 0.07966 0.11246 0.13751 0.15852 0.17694 0.19350 0.20866 0.22270

[10] 0.23582 0.24817 0.68269 0.84270 0.91674 0.95450 0.97465 0.98569 0.99185

[19] 0.99532 0.99730 0.99843
```

donner une approximation de la p-value (probabilité critique) d'un test de l'hypothèse nulle d'un tirage binomial de probabilité $p = 3/4$ contre l'alternative " les données sont trop proches du modèle (3/4 1/4)" ?

Question 11.

Quelle est la loi de $X + Y$ sachant que X suit une loi Khi2 à p degrés de liberté et que Y suit une loi Khi2 à q degrés de liberté. Utiliser ce résultat pour répondre à la question 10 en groupant les trois expériences.

Information :

```
> pchisq(c(seq(0,1,by=0.1),1:10),3)

 [1] 0.000000 0.008163 0.022411 0.039972 0.059758 0.081109 0.103568 0.126796

 [9] 0.150533 0.174572 0.198748 0.198748 0.427593 0.608375 0.738536 0.828203

[17] 0.888390 0.928102 0.953988 0.970709 0.981434
```

NOM : **Prénom :**

LICENCE BO - UE BMS - 06/2002 (2 HEURES)

Question 1.

Nuage	1	2	3	4	5	6
Corrélation						

Question 2. La valeur 0.5 appartient-elle à l'intervalle de confiance …

Question 3.

Histogramme	1	2	3	4	5	6
Ordre utilisé						

Question 4

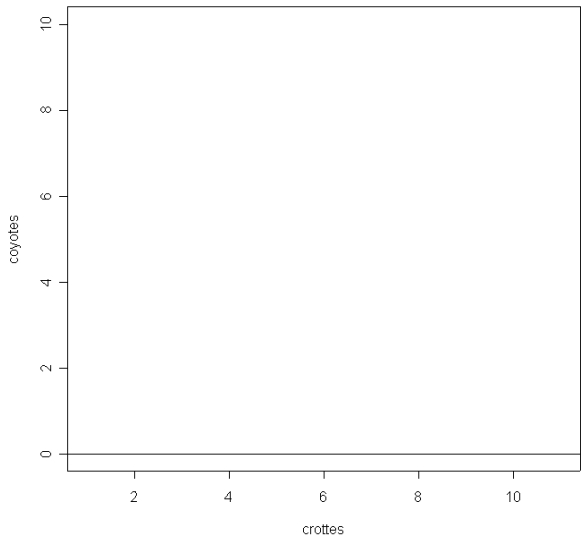
QQ-plot	1	2	3	4
Ordre utilisé				

Question 5 Quelle est la fonction génératrice des probabilités $x \mapsto f(x)$ d'une loi binomiale …

Question 6. Lequel des deux vous semble le plus approprié ?

Question 7. Quelle est la probabilité critique (p-value) attendue de wilcox.test(x,y) ?

Question 8.



Question 9. Donner le taux de séropositivité …

Question 10. Donner une approximation de la p-value (probabilité critique) d'un test …

Question 11. Utiliser ce résultat pour répondre à la question 10 …

LICENCE BO - UE BMS - 06/2002 (2 HEURES)

Question 1.

Nuage	1	2	3	4	5	6
Corrélation	-0.04	0.74	-0.73	-0.49	0.5	0.33

Je choisis les deux plus grandes corrélations en valeur absolue, ensuite les deux moyennes, enfin -0.04 pour la 1 et 0.33 pour la 6 car elle ne semble pas nulle. Pour refaire l'expérience :

```
sigma_matrix(c(0,1,1,0),2,2)

toto_rnorm(100)
x1_scale.wt(data.frame(x=toto,y=toto*toto+rnorm(100,0,0.3)))
scatter.label(x1,clab=0,cpoi=2,sub="1",csub=5)
print (cor(x1)[1,2])

x2_mvrnorm(100,rep(0,2),diag(1,2)+0.75*sigma)
scatter.label(x2,clab=0,cpoi=2,sub="2",csub=5)
print (cor(x2)[1,2])

x3_mvrnorm(100,rep(0,2),diag(1,2)-0.75*sigma)
scatter.label(x3,clab=0,cpoi=2,sub="3",csub=5)
print (cor(x3)[1,2])

x4_mvrnorm(100,rep(0,2),diag(1,2)-0.5*sigma)
scatter.label(x4,clab=0,cpoi=2,sub="4",csub=5)
print (cor(x4)[1,2])

x5_mvrnorm(100,rep(0,2),diag(1,2)+0.5*sigma)
scatter.label(x5,clab=0,cpoi=2,sub="5",csub=5)
print (cor(x5)[1,2])

x6_scale.wt(data.frame(x=toto,y=exp(toto)+rnorm(100,0,5)))
scatter.label(x6,clab=0,cpoi=2,sub="6",csub=5)
print (cor(x6)[1,2])

[1] -0.03920
[1] 0.7378
[1] -0.7254
[1] -0.4855
[1] 0.4957
[1] 0.3342
```

Question 2. La valeur 0.5 appartient-elle à l'intervalle de confiance …

NON. $P(X \leq 4)$ pour une binomiale (19, 1/2) vaut 1% (table une chance sur deux dans le polycope). Donc on rejette l'hypothèse nulle $p = 0.5$ au seuil de % dans un test bilateral, donc, par définition $p = 0.5$ n'appartient pas à l'intervalle de confiance.

Question 3.

Histogramme	1	2	3	4	5	6
Ordre utilisé	C	B	A	D	F	E

Deux histogrammes sont symétriques et il y a deux lois normales. On les sépare par la valeur de l'écart-type. On élimine la seule loi discrète. Restent 3 histogrammes dissymétriques qu'on repère avec les moyennes respectivement 1/2.5, 1/0.5 et 3.

Pour refaire l'expérience :

```
x <- rnorm(1000,20,5) ;hist(x,main="1")
x <- rexp(1000,0.5) ; hist(x,main="2")
x <- rexp(1000,2.5) ; hist(x,main="3")
x <- rchisq(1000,3) ; hist(x,main="4")
x <- rnorm(1000,20,50) ; hist(x,main="5")
x <- rpois(1000,1) ; hist(x,main="6")
```

Question 4

QQ-plot	1	2	3	4
Ordre utilisé	D	C	B	A

Le qq-plot d'une loi normale est linéaire et c'est le 2. Le 3 est discret et c'est la binomiale. Le 1 est dissymétrique et c'est l'exponentielle. Le 4 est sur-dispersé et c'est la loi uniforme.

Question 5 Quelle est la fonction génératrice des probabilités $x \mapsto f(x)$ d'une loi binomiale …

$$f(x) = \sum_{j=1}^{n} P(X = j) x^j = \sum_{j=1}^{n} \binom{n}{j} p^j (1-p)^{n-j} x^j = (px + 1 - p)^n$$

$$f'(1) = \sum_{j=1}^{n} jP(X = j) = E(X) = np$$

On retrouve toujours la moyenne.

Question 6. Lequel des deux vous semblent le plus approprié ?

La distribution n'est franchement pas normale et présente une dssymétrie très forte. Le test t est invalide et le test non paramétrique (Wilcoxon) s'impose.

Question 7. Quelle est la probabilité critique (p-value) attendue de wilcox.test(x,y) ?

Il y a $\binom{7}{2}$ manières de tirer 2 entiers parmi les 7 premiers. Dans un échantillon, on a les rangs 1 et 2. La probabilité de la configuration est de $1 \Big/ \binom{7}{2} = \frac{1}{21}$. La probabilité du test bilatéral est donc 2/21 = 0.095.

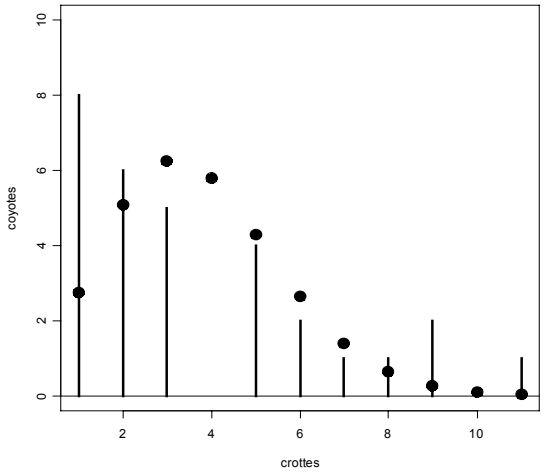
Vérification :

```
x_c(6,3)
y_c(21,8,7,14,9)
> wilcox.test(x,y)

        Wilcoxon rank sum test

data:  x and y
W = 0, p-value = 0.09524
alternative hypothesis: true mu is not equal to 0
```

Question 8.



La distribution est très éloignée d'une loi poissonienne. Certains individus dominants laissent plus de marques de leur propre territoire, alors que certains individus dominés sont fort discrets. C'est la structure sociale des canidés qui est en jeu. Pour refaire la figure :

```
coyotes_c(8,6,5,4,2,1,1,2,1)
crottes_c(1,2,3,5,6,7,8,9,11)
dpois(1:20,111/30)*30
 [1] 2.744e+00 5.077e+00 6.262e+00 5.792e+00 4.286e+00 2.643e+00 1.397e+00
 [8] 6.461e-01 2.656e-01 9.828e-02 3.306e-02 1.019e-02 2.901e-03 7.667e-04
[15] 1.891e-04 4.374e-05 9.519e-06 1.957e-06 3.810e-07 7.049e-08
plot(crottes,coyotes,type="h",lwd=3,ylim=c(0,10))
points(1:11, dpois(1:11,111/30)*30,pch=20,cex=3)
abline(h=0)
```

Question 9. Donner le taux de séropositivité …

```
> xtabs(cbind(fivposi,fivnega)~age,data=fiv)

age  fivposi fivnega
  a1      78       9       10.3 %
  a2     113      19       14.4 %
  a3      89      16       15.2 %
> cbind(fivposi, fivnega) ~ sex
cbind(fivposi, fivnega) ~ sex
> xtabs(cbind(fivposi,fivnega)~gen,data=fiv)

gen  fivposi fivnega
  NO     206      24       10.4 %
  O       74      20       21.5 %

> xtabs(cbind(fivposi,fivnega)~sex,data=fiv)

sex fivposi fivnega
  F     171      14        7.5 %
  M     109      30       21.5 %
```

Le taux augment un peu avec l'âge, beaucoup avec le gène orange et encore plus avec le sexe mâle. Le sexe est le facteur le plus important (risque de contamination dans les contacts agressifs). Pour vérification :

```
> chisq.test(xtabs(cbind(fivposi,fivnega)~age,data=fiv))

        Pearson's Chi-squared test

data:  xtabs(cbind(fivposi, fivnega) ~ age, data = fiv)
X-squared = 1.096, df = 2, p-value = 0.578

> chisq.test(xtabs(cbind(fivposi,fivnega)~sex,data=fiv))

        Pearson's Chi-squared test with Yates' continuity correction

data:  xtabs(cbind(fivposi, fivnega) ~ sex, data = fiv)
X-squared = 12.12, df = 1, p-value = 0.0004998
```

```
> chisq.test(xtabs(cbind(fivposi,fivnega)~gen,data=fiv))

        Pearson's Chi-squared test with Yates' continuity correction

data:  xtabs(cbind(fivposi, fivnega) ~ gen, data = fiv)
X-squared = 5.792, df = 1, p-value = 0.01610
```

Question 10. Donner une approximation de la p-value (probabilité critique) d'un test …

On doit prendre la probabilité d'être inférieure ou égale à l'observation, contrairement au cas ordinaire, pour un test contre une valeur trop faible du Khi2 d'ajustement.

0.2629 pour un Khi2 à un ddl, environ $p = 0.4$ (par interpolation **0.24** pour 0.1 et **0.68** pour 1)

0.015 pour un Khi2 à un ddl, environ $p = 0.11$ (par interpolation **0.08** pour 0.01 et **0.14** pour 0.02)

0.3907 pour un Khi2 à un ddl, environ $p = 0.5$ (par interpolation **0.24** pour 0.1 et **0.68** pour 1)

Vérification :

```
> pchisq(0.2669,1)
[1] 0.3946
> pchisq(0.015,1)
[1] 0.09748
> pchisq(0.3907,1)
[1] 0.4681
```

Question 11. Utiliser ce résultat pour répondre à la question 10 …

Si $X \rightarrow \chi^2_p$, c'est la somme de p carrés de gaussiennes indépendantes. Si $Y \rightarrow \chi^2_q$, c'est la somme de q carrés de gaussiennes indépendantes. Alors $X + Y$ est la somme de p + q carrés de gaussiennes indépendantes et suit une loi χ^2_{p+q}.

On peut donc additionner les Khi2 d'ajustement indépendants :

```
> 0.2629+0.015+0.3907
[1] 0.6686
```

A comparer avec un Khi2 à 3 degrés de liberté, soit par interpolation (**0.1035** pour 0.6 et **0.126796** pour 0.7) une p-value autour de 0.11.

```
> pchisq(0.6686,3)
[1] 0.1194
```

Rien n'indique que les données sont trop proches du modèle.