Chapitre 4 : Données catégorielles

Géraldine Marchand

Travaux pratiques - séances 2016

Comparaison de deux proportions à partir de données pairées

Questions de recherche (cf. enquête SMP) :

Pour un individu pris au hasard dans la population cible,

1. la probabilité d'avoir été l'objet d'une mesure de placement avant ses 18 ans est-elle significativement différente de celle d'avoir subi des maltraitances dans l'enfance?

2. la probabilité d'abuser d'alcool est-elle significativement différente de celle d'abuser de substance?

3. la probabilité d'être diagnostiqué dépressif est-elle significativement différente de celle d'avoir déjà tenté de suicider?

Rappels : Estimation de $\pi_2 - \pi_1$ (données pairées)

Point de départ :

Table 2x2 (+marges)

	$X_2 = 1$	$X_2 = 0$	**Tot**
$X_1 = 1$	n_{11}	n_{12}	n_{1+}
$X_1 = 0$	n_{21}	n_{22}	n_{2+}
Tot	n_{+1}	n_{+2}	n

ou alternativement :

X_2	X_1	$D = X_2 - X_1$	**Freq**
1	1	0	n_{11}
1	0	1	n_{21}
0	1	-1	n_{12}
0	0	0	n_{22}

Parallèle avec le test-t pairé

Cette présentation alternative permet de faire le parallèle entre l'estimation de $\pi_2 - \pi_1$ pairées et le test-t pairé (cf. chapitre 3).

- Distribution de probabilité pour D

Espace d'échantillionnage : -1 0 1

Probabilités : π_{12} $\pi_{11} + \pi_{22}$ π_{21}

$\mu_D = E(D) = -1 \times \pi_{12} + 0 \times (\pi_{11} + \pi_{22}) + 1 \times \pi_{21} = \pi_{21} - \pi_{12} = \pi_{21}(+\pi_{11} - \pi_{11}) - \pi_{12} = \pi_{+1} - \pi_{1+}$
$\mu_D = P(X_2 = 1) - P(X_1 = 1)$: quantité d'intérêt

- Distribution de probabilité pour μ_D

$(\mu_D|donnees) \approx N(\bar{d}, \frac{s_d^2}{n})$

avec $\bar{d} = \frac{n_{21} - n_{12}}{n}$ et $s_D^2 \approx \hat{\sigma}_D^2 = \frac{(-1)^2 \times n_{12} + 0^2 \times (n_{11} + n_{22}) + 1^2 \times n_{21}}{n} - \frac{(n_{21} - n_{12})^2}{n^2} = \frac{n_{21} + n_{12}}{n} - \frac{(n_{21} - n_{12})^2}{n^2}$

- Intervalle de crédibilité et probabilité a posteriori pour $\mu_D = \pi_2 - \pi_1$

$\bar{d} \pm 1.96 \times \sqrt{\frac{s_D^2}{n}}$

$P(\mu_D > 0|D) = P(\pi_2 - \pi_1 > 0|D) = P(\pi_2 > \pi_1|D)$

Test de McNemar

Question : est-ce que 0 est une valeur plausible pour $\pi_2 - \pi_1$

Réponse :

- oui si $\frac{(n_{21}-n_{12})^2}{n_{21}+n_{12}} < 1.96^2 = 3.84 = \chi^2_1(0.95)$
- non si $\frac{(n_{21}-n_{12})^2}{n_{21}+n_{12}} > 1.96^2 = 3.84 = \chi^2_1(0.95)$

En pratique

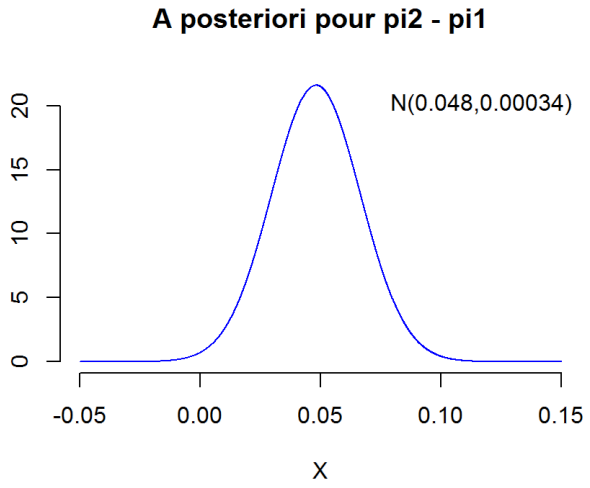
1. X_1 : Placement - X_2 : Maltraitances

- Données :

```
##          Abus=1 Abus=0 Total
## Place=1     93     88   181
## Place=0    126    482   608
## Total      219    570   789
```

```
##   place abus Freq
## 1     0    0  482
## 2     1    0   88
## 3     0    1  126
## 4     1    1   93
```

- Distribution a posteriori pour $\pi_2 - \pi_1$



- Intervalle de crédibilité à 95% pour $\pi_2 - \pi_1$: (0.012, 0.084)
- $P(\pi_2 - \pi_1 > 0|D) = 0.995$
- Test de Mc Nemar : $6.75 > 3.84(\chi^2_1(0.95)) \Rightarrow \pi_2$ sign. $\neq \pi_1$

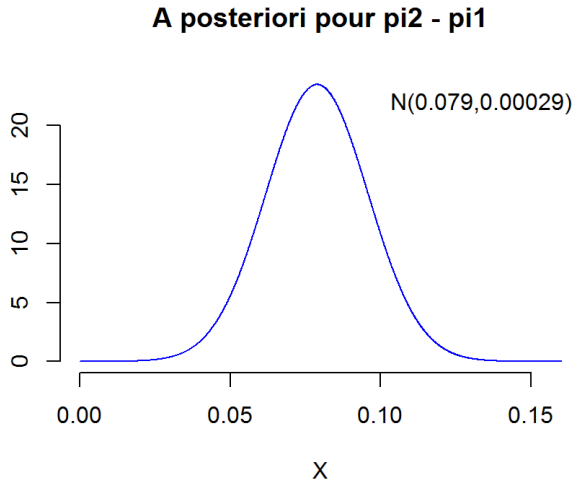
2. X_1 : Alcool - X_2 : Substances

- Données :

```
##       Subst=1 Subst=0 Total
## Alc=1      86      63   149
## Alc=0     126     524   650
## Total     212     587   799
```

```
##   alc.cons subst.cons Freq
## 1        0          0  524
## 2        1          0   63
## 3        0          1  126
## 4        1          1   86
```

- Distribution a posteriori pour $\pi_2 - \pi_1$



- Intervalle de crédibilité à 95% pour $\pi_2 - \pi_1$: (0.046, 0.112)
- Test de Mc Nemar : $21 > 3.84(\chi^2_1(0.95)) \Rightarrow \pi_2$ sign. $\neq \pi_1$

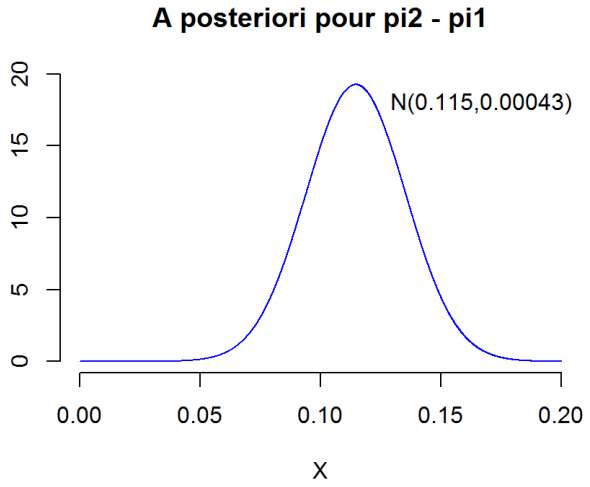
3. X_1 : Suicide - X_2 : Dépression

- Données :

```
##        Dep=1 Dep=0 Total
## TS=1     131    92   223
## TS=0     182   380   562
## Total    313   472   785
```

```
##   suicide.past dep.cons Freq
## 1            0        0  380
## 2            1        0   92
## 3            0        1  182
## 4            1        1  131
```

- Distribution a posteriori pour $\pi_2 - \pi_1$



- Intervalle de crédibilité à 95% pour $\pi_2 - \pi_1$: (0.074, 0.155)
- Test de Mc Nemar : 29.56 $> 3.84(\chi^2_1(0.95)) \Rightarrow \pi_2$ sign. $\neq \pi_1$

Association entre deux variables catégorielles

Application 1

- La durée de condamnation (cat) varie-t-elle avec l'âge (cat) des détenus?
- Si oui, quelle est la force et quel est le sens du lien? (les jeunes détenus ont-ils tendance à être condamnés plus ou moins longtemps que les détenus plus âgés?)

Application 2

- La taille de la fratrie (cat) est-elle associée au fait d'avoir ou non un diplôme?
- Si oui, quelle est la force et quel est le sens du lien? (les détenus issus de familles nombreuses sont-ils plus "à risque" d'être "sans diplôme"?)

$\Rightarrow$ **Trois étapes** :

- Analyse descriptive
- Test d'indépendance
- (Si rejet hyp d'indép.) Quantification de la force du lien via l'estimation de RR

Rappels

Test d'indépendance

- Se base sur la table de contingence :

	j=1	**...**	**j = J**	
i=1	n_{11}	n_{1j}	n_{1J}	n_{1+}
...	n_{i1}	...	n_{iJ}	n_{i+}
i=I	n_{I1}	n_{Ij}	n_{IJ}	n_{I+}
	n_{+1}	n_{+j}	n_{+J}	n

- Statistique du test : $G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{n_{i+}p_{j|i}})$
- En cas d'indépendance : $p_{j|i}$ identiques $\forall i \Rightarrow p_{j|i} = \frac{n_{+j}}{n}$ (les probabilités conditionnelles de Y ne changent pas avec X)
  $\Rightarrow$ Fréquences attendues en cas d'indépendance : $n_{i+}\frac{n_{+j}}{n} \Rightarrow G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{n_{i+}\frac{n_{+j}}{n}})$

Décision :

- $G^2 \leq X^2_{(I-1)(J-1)}(0.95)$: non rejet de l'hypothèse d'indépendance. Avec les données à disposition, des probabilités conditionnelles identitiques pour chaque modalité de X sont une configuration plausible.
- $G^2 \geq X^2_{(I-1)(J-1)}(0.95)$: rejet de l'hypothèse d'indépendance. Il existe un lien significatif entre X et Y au niveau de la population d'intérêt.

Quantification de la force du lien

$RR = \frac{\pi_1}{\pi_2}$ estimé par $\hat{RR} = \frac{p_1}{p_2}$

IC 95% pour log(RR) = $log(\frac{p_1}{p_2}) \pm 1.96\sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}} = (a, b)$

IC 95% pour RR = (e^a, e^b)

Application 1 : Durée de la peine en fonction de la tranche d'âge

Analyse descriptive : visualisation du lien éventuel

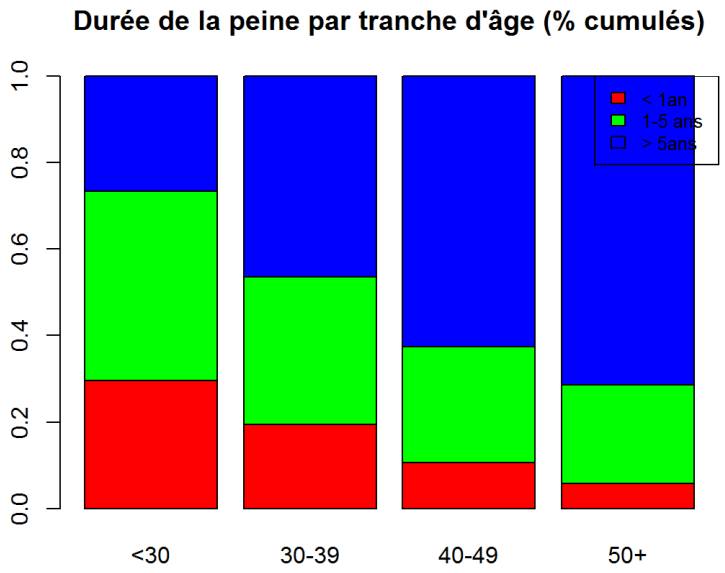
1. Point de départ : Table de contingence

```
##        duree_cat
## age_cat < 1an 1-5 ans > 5ans
##   <30      41      61     37
##   30-39    30      53     72
##   40-49    15      38     89
##   50+       8      32    100
```

2. Calcul des fréquences relatives (en %) de la variable réponse (duree) pour chaque modalité de la variable explicative (âge)

```
##          age_cat
## duree_cat  <30 30-39 40-49  50+
##   < 1an   29.5  19.4  10.6  5.7
##   1-5 ans 43.9  34.2  26.8 22.9
##   > 5ans  26.6  46.5  62.7 71.4
```

3. Graphe de la distribution conditionnelle de Y en fonction de X



4. Interprétation

Si la durée de condamnation ne dépendait pas de l'âge du détenu, la distribution conditionnelle de Y ne devrait pas changer avec X. Or, au sein de l'**échantillon**, on observe que la proportion de condamnations > 5 ans "augmente" avec la catégorie d'âge. Inversement, les détenus plus jeunes sont plus souvent condamnés moins longtemps. Il faut néanmoins vérifier si ces tendances se vérifient au niveau de la **population des détenus** $\Rightarrow$ test d'indépendance.

Test d'indépendance

1. Calcul des fréquences attendues en cas d'indépendance

```
##        duree_cat
## age_cat < 1an 1-5 ans > 5ans
##   <30    22.7    44.4   71.9
##   30-39  25.3    49.5   80.2
##   40-49  23.2    45.4   73.5
##   50+    22.8    44.7   72.4
```

2. Vérification des conditions d'utilisation du test

$A_{ij} = n_{i+}\frac{n_{+j}}{n} \geq 5 \forall ij$: ok

3. Calcul du G^2

G^2 = 73.9699

4. Seuil critique $X^2_{(I-1)(J-1)}(0.95)$

Table du Chi-carré

$X^2_6(0.95)$ = 12.59

5. Conclusion

$G^2 > X^2_6(0.95) \Rightarrow$ On rejette l'hypothèse d'indépendance. La durée de condamnation varie avec la catégorie d'âge dans la **population** des détenus.

6. Examen de la cause du rejet

Table des fréquences observées :

```
##         duree_cat
## age_cat < 1an 1-5 ans > 5ans
##   <30       41       61     37
##   30-39     30       53     72
##   40-49     15       38     89
##   50+        8       32    100
```

Table des fréquences attendues en cas d'indépendance:

```
##         duree_cat
## age_cat < 1an 1-5 ans > 5ans
##   <30    22.7    44.4   71.9
##   30-39  25.3    49.5   80.2
##   40-49  23.2    45.4   73.5
##   50+    22.8    44.7   72.4
```

Constat :

- Parmi les détenus "jeunes" (<30 et 30-39), la proportion de condamnations "courtes" (< 1 an et 1-5 ans) est sur-estimée et la proportion de condamnations "longues" (>5 ans) est sous-estimée.
- Parmi les détenus "âgés" (40-49 et 50+), la proportion de condamnations "courtes" (< 1 an et 1-5 ans) est sous-estimée et la proportion de condamnations "longues" (>5 ans) est sur-estimée.
- Ce sont les catégories extrêmes (<30 et 50+) qui s'éloignent le plus de la situation d'indépendance.

Quantification de la force du lien

Données : Duree (binaire) par age_cat (en %)

```
##        duree_bin
## age_cat >5ans <5ans
##   <30    26.6  73.4
##   30-39  46.5  53.5
##   40-49  62.7  37.3
##   50+    71.4  28.6
```

Rapports de risque (durée de condamnation > 5 ans) en fonction de la catégorie d'âge

Contrastes	$\hat{RR}$	**IC 95% pour RR**	**Conclusion**
2 Vs 1	1.75	(1.26, 2.41)	1 $\notin$ IC, V>1 : $\pi_2 > \pi_1$
3 Vs 1	2.35	(1.74, 3.19)	1 $\notin$ IC, V>1 : $\pi_3 > \pi_1$
4 Vs 1	2.68	(2.00, 3.61)	1 $\notin$ IC, V>1 : $\pi_4 > \pi_1$
3 Vs 2	1.35	(1.09, 1.67)	1 $\notin$ IC, V>1 : $\pi_3 > \pi_2$
4 Vs 2	1.54	(1.26, 1.88)	1 $\notin$ IC, V>1 : $\pi_4 > \pi_2$
4 Vs 3	1.14	(0.97, 1.34)	1 $\in$ IC : "$\pi_4 = \pi_3$"

Le RR entre les catégories "<30 ans et 50+" est le plus grand : le risque d'être condamné pour une durée > 5 ans est entre 2 fois et 3.6 fois plus élevé pour les détenus de 50 ans et plus par rapport aux détenus de moins de 30 ans.

Application 2 : Diplôme (oui/non) en fonction de la taille de la fratrie

Analyse descriptive : visualisation du lien éventuel

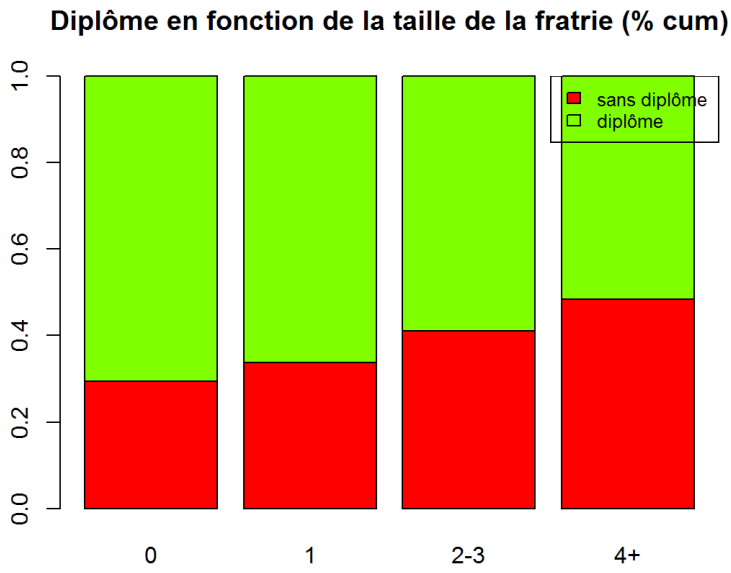
1. Point de départ : Table de contingence

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0             22      53
##           1             28      55
##           2-3           98     141
##           4+           192     205
```

2. Calcul des fréquences relatives (en %) de la variable réponse (duree) pour chaque modalité de la variable explicative (âge)

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0           29.3    70.7
##           1           33.7    66.3
##           2-3         41.0    59.0
##           4+          48.4    51.6
```

3. Graphe de la distribution conditionnelle de Y en fonction de X



4. Interprétation

Au sein de l'**échantillon**, on observe que la proportion de détenus sans diplôme "augmente" avec la taille de fratrie. Il faut néanmoins vérifier si ces tendances se vérifient au niveau de la **population** des détenus.

Test d'indépendance

1. Calcul des fréquences attendues en cas d'indépendance

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0           32.1    42.9
##           1           35.5    47.5
##           2-3        102.3   136.7
##           4+         170.0   227.0
```

2. Vérification des conditions d'utilisation du test

$A_{ij} = n_{i+} \frac{n_{+j}}{n} \geq 5 \forall ij$: ok

3. Calcul du G^2

G^2 = 13.9307

4. Seuil critique $X^2_{(I-1)(J-1)}(0.95)$

Table du Chi-carré

$X^2_3(0.95)$ = 7.81

5. Conclusion

$G^2 > X^2_3(0.95) \Rightarrow$ On rejette l'hypothèse d'indépendance. La probabilité d'être "sans diplôme" varie avec la taille de la fratrie (catégorisée) dans la **population** des détenus.

6. Examen de la cause du rejet

Table des fréquences observées :

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0             22      53
##           1             28      55
##           2-3           98     141
##           4+           192     205
```

Table des fréquences attendues en cas d'indépendance:

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0           32.1    42.9
##           1           35.5    47.5
##           2-3        102.3   136.7
##           4+         170.0   227.0
```

Constat :

- Au sein de la catégorie des détenus ayant au moins 4 frères et/ou soeurs, la proportion de "sans diplôme" est sur-estimée.
- Dans les trois autres catégories, cette même proportion est sous-estimée.
- Ce constat est non-négligeable dans la mesure où la proportion de détenus issus de familles nombreuses est particulièrement importante (estimée à 50% sur base des données récoltées).

Quantification de la force du lien

Données : Diplôme par n.fratrie_cat (en %)

```
##              ecole_cat
## n.fratrie_cat sans diplôme diplôme
##           0          29.3    70.7
##           1          33.7    66.3
##           2-3        41.0    59.0
##           4+         48.4    51.6
```

Rapports de risque (être "sans diplôme") en fonction de la taille de la fratrie (catégorisée)

Contrastes	$\hat{RR}$	IC 95% pour RR	Conclusion
2 Vs 1	1.15	(0.72, 1.83)	1 $\in$ IC : "$\pi_1 = \pi_2$"
3 Vs 1	1.40	(0.95, 2.05)	1 $\in$ IC : "$\pi_1 = \pi_3$"
4 Vs 1	1.65	(1.14, 2.38)	1 $\notin$ IC, V>1 : $\pi_4 > \pi_1$
3 Vs 2	1.22	(0.87, 1.70)	1 $\in$ IC : "$\pi_2 = \pi_3$"
4 Vs 2	1.43	(1.04, 1.97)	1 $\notin$ IC, V>1 : $\pi_4 > \pi_2$
4 Vs 3	1.18	(0.98, 1.42)	1 $\in$ IC : "$\pi_4 = \pi_3$"

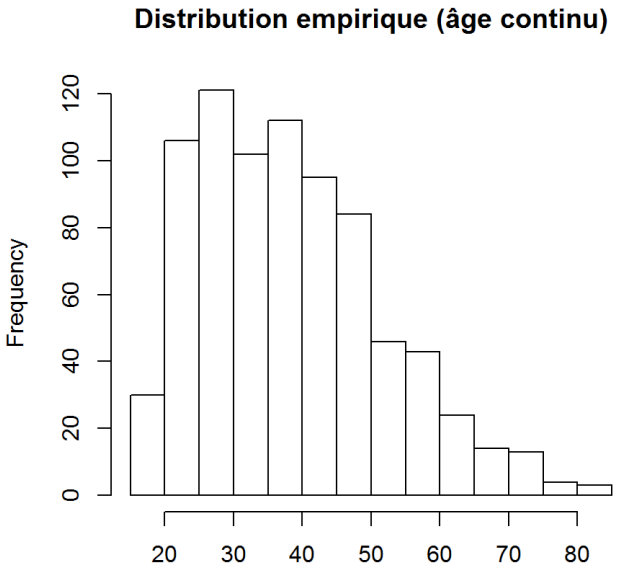
Avec les données à disposition, on a pu mettre en évidence que les détenus issus de familles nombreuses (avec minimum 4 frères et soeurs) sont significativement plus à risque d'être sans diplôme que ceux qui n'ont aucun ou maximum un frère ou une soeur.

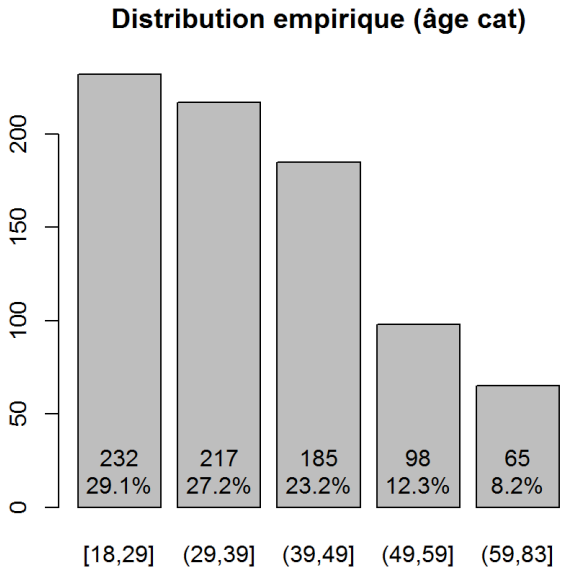
Autre utilisation de la statistique G^2 : Test d'ajustement

- Question : La répartition par tranches d'âge est-elle la même au sein de la population des détenus français qu'au sein de la population française en général (hommes uniquement)?
- Statistiques **population**: INSEE - Population France Métropolitaine 18 ans et plus - Hommes - Recensement 2004-2008:

Tranches d'âge	**Effectifs**	**%**
18-29	4679860	20.56
30-39	4220746	18.54
40-49	4247739	18.66
50-59	4023282	17.68
60+	5588852	24.56
Tot	22760479	100

- Données **échantillon** :





Test d'ajustement

1. Calcul des fréquences attendues en cas d'ajustement :

```
##                    18-29    30-39    40-49    50-59      60+
## Répartition pop   0.2056   0.1854   0.1866   0.1768   0.2456   1
## Fréq observées  232.0000 217.0000 185.0000  98.0000  65.0000 797
## Fréq attendues  163.8632 147.7638 148.7202 140.9096 195.7432 797
```

2. Vérification des conditions d'utilisation du test :

Fréquences attendues ≥ 5 : ok

3. Calcul du G^2 :

G^2 = 194.39

4. Seuil critique :

Table du Chi-carré

$X^2_{(K-1)}(0.95)$ = 9.49

5. Conclusion

$G^2 > X^2_4(0.95) \Rightarrow$ On rejette l'hypothèse d'ajustement. La répartition par catégories d'âge des détenus français n'est pas la même que dans la population française en général.

6. Examen de la cause du rejet

Les fréquences attendues sous l'hypothèse d'ajustement sous-estiment la proportion des classes 18-29, 30-39, 40-49 et sur-estiment la proportion des classes 50-59 et 60+. Autrement dit, la population carcérale est caractérisée par une sous-représentation des catégories d'âges plus avancées.