GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 12

- **CETTE MATIÈRE NE SERA PAS EXAMINÉE ! !**
- Tests χ^2 :
  - **CETTE MATIÈRE NE SERA PAS EXAMINÉE ! !**
  - **Adéquation :** tester la conformité ('goodness-of-fit' / 'adéquation') des données à une distribution théorique
  - **Indépendence :** étude de la distribution conjointe de deux variables qualitatives
  - **Homogénéité :** comparaison de la distribution d'une variable qualitative dans plusieurs échantillons
- Paradoxe de Simpson

Révision : Variables (I)

- En statistique, les caractéristiques qui varient pour les individus de la populations sont appelées **variables**
- Différentes sortes de variables :
  - **Variables qualitatives** : les modalités sont des mots ou 'etiquettes' que l'on appelle des *catégories*
    Exemples : couleur des yeux ('bleu', 'brun', 'vert') ; votre programme de télé préféré
  - **Variables quantitatives** : les modalités sont des valeurs numériques
    Exemples : âge, nombre de membres d'une famille, le poids

Révision : Variables (II)

- **Variables qualitatives** peuvent être classées comme :
  - Échelle *nominale* – les catégories ne sont pas naturellement ordonnées (e.g. couleur des yeux, sexe)
    - *Même si* les modalités ont des codes numériques (e.g. sexe = '0' pour 'mâle', = '1' pour 'femelle')
  - Échelle *ordinale* – les catégories peuvent être ordonnées (e.g. 'toujours', 'quelquefois', 'jamais')
- **Variables quantitatives** sont distinguées comme :
  - Variables *discrètes* – les valeurs possibles peuvent être énumérées sous la forme d'une list de chiffres (le plus souvent : les entiers naturels 0, 1, 2, ...)
  - Variables *continues* – l'ensemble des valeurs possibles est constitué par une intervalle

Analyse des données catégorielles

- Jusqu'à présent, on a examiné des variables continues (en régression) et variables factorielles (catégorielles) dans le contexte de l'ANOVA
- Une variable catégorielle pourrait être considérée comme un *classement* des observations
- Classement : une variable
  - test d'adéquation
- Classements : plusieurs variables (au moins 2)
  - représentation par tableau de contingence
  - test d'indépendance
  - test de l'homogénéité des distributions
    - p. ex. test d'équalité des deux proportions

Test d'adéquation

- Comment pouvons-nous mesurer l'adéquation des données à une théorie ?
- Situation : on observe la répartition de n objets dans k classes
- On veut tester l'*adéquation des données à une distribution théorique*

Example 12.1 Soixante lancers d'un dé ont donné les résultats :

huit fois '1'	dix fois '2'	neuf fois '3'
seize fois '4'	treize fois '5'	quatre fois '6'

On veut tester l'hypothèse selon laquelle le dé est équilibré – comment peut-on faire cela *??*

Révision : Multinomial distribution

- L'une des distributions conjointes les plus importantes (VAs discrètes) est la **distribution multinomiale**
- Il s'agit d'un analogue à plusieurs variables de la *distribution binomiale* :
  - Pour la distribution *binomiale*, il y a 2 résultats possibles pour chacun des n épreuves indépendantes
  - Pour la distribution *multinomiale*, chaque épreuve peut entraîner l'un des r résultats possibles, de probabilités respectives $p_1, p_2, \ldots, p_r$, telles que $\sum_{i=1}^{r} p_i = 1$
- La loi conjointe de la distribution multinomiale est :

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r) = \frac{n!}{n_1! n_2! \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

- *Si $r = 2$, c'est quelle distribution ? ?*

Test d'adéquation, cont.

- Pour tester l'adéquation des données à une répartition théorique, on dispose de deux éléments :
  - n *observations réparties* dans k 'cellules' :

n_1	n_2	$\cdots$	n_k

  - une fonction de *fréquence théorique* :

p_1	p_2	$\cdots$	p_k

- *Une mesure de distance* entre la répartition empirique (observée) et la loi théorique se base sur la répartition des n observations selon la loi théorique (les *nombres espérés* dans les cellules) :

$n \times p_1$	$n \times p_2$	$\cdots$	$n \times p_k$

Statistique de test

- La statistique de test :

$$X^2_{obs} = \sum_{i=1}^{k} \frac{(\text{observé} - \text{espéré})^2}{\text{espéré}}$$

$$= \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

- Pour n 'suffisamment grand' ($\forall i, = \boxed{np_i \geq 5}$), sous H

$$X^2_{obs} \sim \chi^2_{k-p-1},$$

où p = nombre de paramètres *estimés à partir des données*

- $\Rightarrow$ **Règle générale :** il faut un nombre espéré de **5** dans chaque cellule

Table de χ^2

TABLE C: Chi-Square distributions

cum probability	0.025	0.80	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
right tail	0.975	0.2	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
df									
1	0.00098	1.64	2.71	3.84	5.02	6.63	7.88	10.83	12.12
2	0.051	3.22	4.61	5.99	7.38	9.21	10.60	13.82	15.20
3	0.216	4.64	6.25	7.81	9.35	11.34	12.84	16.27	17.73
4	0.48	5.99	7.78	9.49	11.14	13.28	14.86	18.47	20.00
5	0.83	7.29	9.24	11.07	12.83	15.09	16.75	20.51	22.11
6	1.24	8.56	10.64	12.59	14.45	16.81	18.55	22.46	24.10
7	1.69	9.80	12.02	14.07	16.01	18.48	20.28	24.32	26.02
8	2.18	11.03	13.36	15.51	17.53	20.09	21.95	26.12	27.87
9	2.70	12.24	14.68	16.92	19.02	21.67	23.59	27.88	29.67
10	3.25	13.44	15.99	18.31	20.48	23.21	25.19	29.59	31.42
11	3.82	14.63	17.28	19.68	21.92	24.73	26.76	31.26	33.14
12	4.40	15.81	18.55	21.03	23.34	26.22	28.30	32.91	34.82
13	5.01	16.98	19.81	22.36	24.74	27.69	29.82	34.53	36.48
14	5.63	18.15	21.06	23.68	26.12	29.14	31.32	36.12	38.11
15	6.26	19.31	22.31	25.00	27.49	30.58	32.80	37.70	39.72
16	6.91	20.47	23.54	26.30	28.85	32.00	34.27	39.25	41.31
17	7.56	21.61	24.77	27.59	30.19	33.41	35.72	40.79	42.88
18	8.23	22.76	25.99	28.87	31.53	34.81	37.16	42.31	44.43
19	8.91	23.90	27.20	30.14	32.85	36.19	38.58	43.82	45.97
20	9.59	25.04	28.41	31.41	34.17	37.57	40.00	45.31	47.50
21	10.28	26.17	29.62	32.67	35.48	38.93	41.40	46.80	49.01
22	10.98	27.30	30.81	33.92	36.78	40.29	42.80	48.27	50.51
23	11.69	28.43	32.01	35.17	38.08	41.64	44.18	49.73	52.00
24	12.40	29.55	33.20	36.42	39.36	42.98	45.56	51.18	53.48
25	13.12	30.68	34.38	37.65	40.65	44.31	46.93	52.62	54.95
30	16.79	36.25	40.26	43.77	46.98	50.89	53.67	59.70	62.16
40	24.43	47.27	51.81	55.76	59.34	63.69	66.77	73.40	76.10
50	32.36	58.16	63.17	67.50	71.42	76.15	79.49	86.66	89.56
60	40.48	68.97	74.40	79.08	83.30	88.38	91.95	99.61	102.7
80	57.15	90.41	96.58	101.9	106.6	112.3	116.3	124.8	128.3
100	74.22	111.7	118.5	124.3	129.6	135.8	140.2	149.4	153.2

Exemple, cont.

Exemple 12.1, cont. Faisons le test ($\alpha = 0,05$) **...**

1.

2.

3.

4.

5.

Autre exemple

Exemple 12.2 Supposons que la proportion des éléments 'défectueux' dans une grande population est inconnue, et que l'on veut tester les hypothèses : $H : p = 0.1$ vs. $A : p \neq 0.1$

Activité Faire un test d'adéquation en supposant un échantillon aléatoire de 100 éléments dont 16 défectueux...

1.

2.

3.

4.

5.

Pouvez-vous penser à un autre test pour résoudre ce problème ? ?

Distributions continues

- Pour une loi *discrète*, il est facile de faire la répartion nécessaire pour le test...
- ... ce qui *ne marche plus* pour les distributions *continues* :
  - plusieurs répartitions possibles
  - comment choisir *? ?*

Exemple spécifique : loi uniforme

- Considérons une VA X qui prend les valeurs entre 0 et 1, mais la densité n'est pas connue
- Supposons en plus qu'on a un échantillon de 100 valeurs tirées aléatoirement de cette distribution, et qu'on veut tester si les observations suivent la loi uniforme
- Alors, on peut diviser l'intervalle (0, 1) en 20 sous-intervalles (0, 0.05), (0.05, 0.10), *etc.*
- Si la distribution est uniforme, alors la probabilité qu'une observation particulière est dans le sous-intervalle i est $p_i = 1/20$, $i = 1, \ldots, 20$
- On a donc que le nombre espéré dans chaque sous-intervalle est $np_i = 100/20 = 5$
- Ensuite, on peut faire le calcul de la statistique X^2 comme avant

Distributions continues, cont.

- Cette méthode pourrait être utilisée pour *toute distribution continue* (elle ne dépende pas de la distribution uniforme)
- Le choix de k et un peu arbitraire : en général, choisissez un nombre k et les sous-intervalles tel que les nombres espérés sont **à peu près égaux et pas trop petits**

1. Faire une partition de l'intervalle (qui pourrait être toutes les réelles) en k sous-intervalles disjoints
2. Déterminer la probabilités $p_i^0, i = 1, \ldots, k$, en supposant une distribution particulière (H), et donc les *nombres espérés* (np_i^0)
3. $\forall i$, trouver N_i, le nombre *d'observations* dans le sous-intervalle i
4. Calculer X^2_{obs}
5. Sous H, $X^2_{obs} \sim \chi^2_{k-1}$

Test des hypothèses composées

- On a considéré les hypothèses (nulles) **simples** : la loi est *complètement spécifiée* par l'hypothèse
- P. ex. H : $p = 0.1$ est une hypothèse *simple*
- Dans ce cas, il n'est pas nécessaire de faire l'estimation des paramètres, et la statistique $X^2_{obs} \sim \chi^2_{k-1}$ sous H
- Par contre, il se peut que l'hypothèse nulle comprend *plusieurs valeurs* pour le(s) paramètre(s)
- Une telle hypothèse est dite **composée**

Statistique de test pour les hypothèses composées

- Pour une hypothèse nulle composée, il faut *modifier* la statistique de test X^2_{obs}, car le nombre espéré d'observations de classe i n'est plus complètement spécifié par l'hypotheèse nulle H
- **Modification :** remplacer np_i *par son EMV* $n\hat{p}_i$ dans le calcul de X^2_{obs}
- Sous H, cette $X^2_{obs} \sim \chi^2_{k-p-1}$, où p est le *nombre de paramètres estimés* pour le calcul des $\hat{p}_i$

Example

Example 12.3 Dans une population d'une plante diploïde annuelle, on a observé la composition génotypique d'un échantillon de 146 individus pour un locus autosomal. Trois allèles codominants F, S et C ont été mis en évidence. On a obtenu les résultats suivants (les parents sont inconnus, alors l'ordre des allèles n'est pas important ; donc p. ex. 'FS' se réfère a FS ou SF) :

FF	FS	FC	SS	SC	CC
18	76	24	13	10	5

La population présente-t-elle des proportions conformes à celles de Hardy-Weinberg ?

[*Indication* : HWE proportions pour 3 allèles sont $(p+q+r)^2$]

Test

1

2

3

4

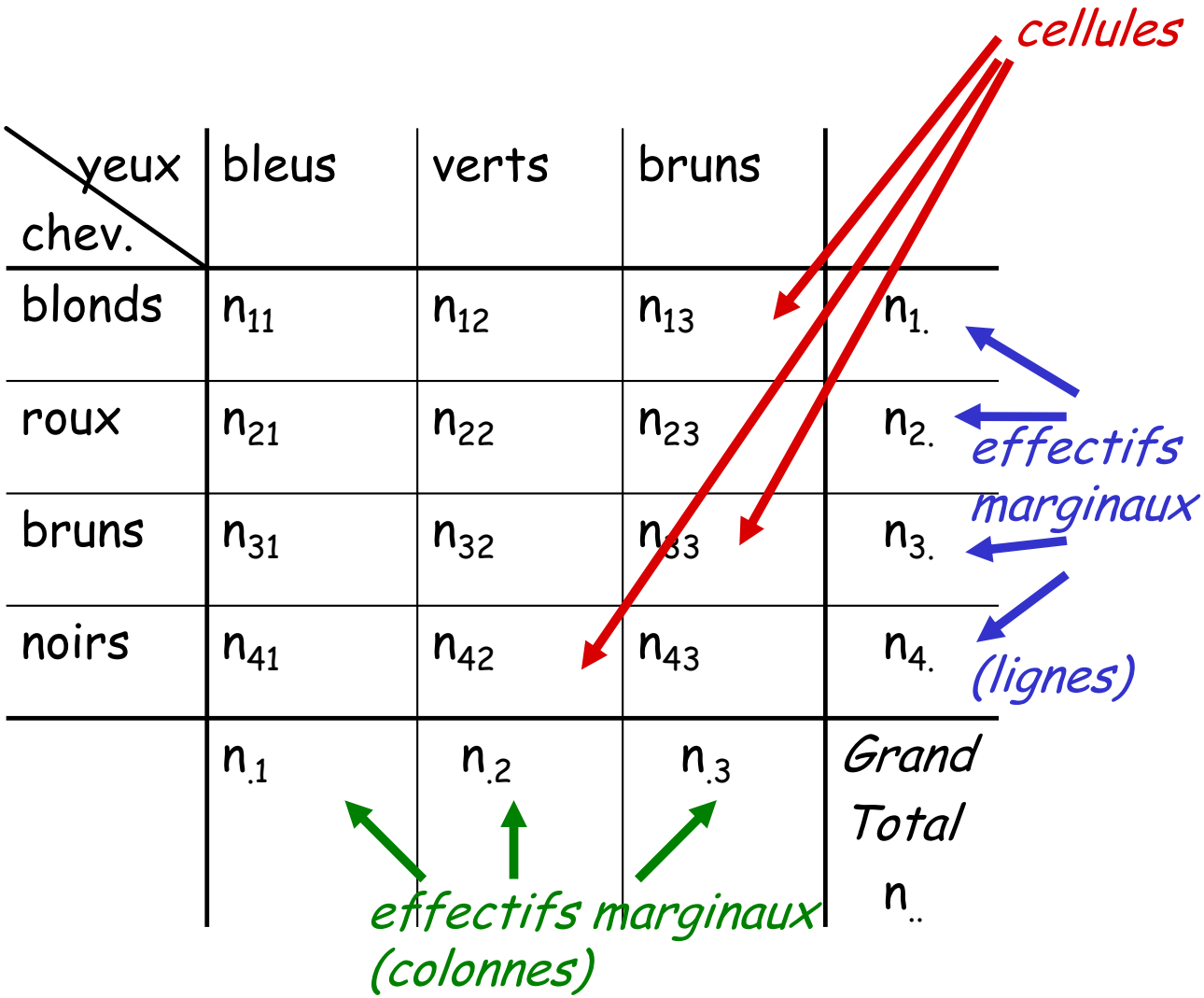
5

PAUSE

Tableau de contingence

- Le *tableau de contingence* est un moyen particulier de représenter *simultanément* deux caractères observés sur une même population (ou échantillon)
- On représente les *fréquences* des valeurs d'un couple de variables $(X,\ Y)$ de modalités $x_i = 1, \ldots, r,\ y_j = 1, \ldots, c$, dans un tableau à double entrée, le *tableau de contingence*
- r = nombre de lignes, c = nombre de colonnes
- Exemple :
  - Couleur des cheveux = blonds, roux, bruns, noirs
  - Couleur des yeux = bruns, verts, bleus
- Les valeurs dans le tableau représente *le nombre d'observations de chaque combinaison* des valeurs possibles pour le couple ('cellule')
- Les *sommes* des valeurs d'une ligne ou d'une colonne sont les *effectifs marginaux*

Tableau cheveux/yeux



Test d'indépendance : intuition

- Construire le tableau bivarié théorique sous la NULLE *i.e.* s'il n'y a *pas d'association*
- Comparer le tableau *réel, observé* au tableau théorique
- Mesurer *la différence* entre ces deux tableaux
- S'il existe *une différence suffisamment grande*, on conclure qu'il y a une *relation significative* (*i.e.* déviation significative de l'independance)
- Autrement, on conclure que nos observations varions *seulement en raison de la variabilité aléatoire*

Test d'indépendance

- Deux caractères sont *indépendants* si la valeur de l'un *n'influe pas sur les distributions des valeurs de l'autre*
- $\Longrightarrow$ les fréquences conjointes doivent être proches des *produits de fréquences marginales*
- Sous l'hypothèse que X et Y sont *indépendantes*, les probabilités conjointes sont :

$$p_{ij} = P(X = x_i \text{ et } Y = y_j) = P(X = x_i) \times P(Y = y_j)$$

- Sur la base du tableau de contingence, on estime les deux *probabilités marginales* par :

$$P(X = x_i) = \frac{\sum_{j=1}^{c} n_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}} = \frac{n_{i.}}{n_{..}} = \frac{n_{i.}}{n}$$

$$P(Y = y_j) = \frac{\sum_{i=1}^{r} n_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}} = \frac{n_{.j}}{n_{..}} = \frac{n_{.j}}{n}$$

Test d'indépendance, cont.

- Donc, on obtient $\hat{p}_{ij} = n_{i.} n_{.j} / n^2$
  ($n = n_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$, la taille de l'échantillon)

- $\Longrightarrow$ Sous l'hypothèse (nulle) d'indépendance, les nombres espérés de la cellule (i, j) sont :

$$n \times P(X = x_i) \times P(Y = y_j) = \frac{\sum_{j=1}^{c} n_{ij} \times \sum_{i=1}^{r} n_{ij}}{n} = \frac{n_{i.}\, n_{.j}}{n}$$

Statistique de test

- La statistique de test est toujours de la forme

$$X^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\text{observé} - \text{espéré})^2}{\text{espéré}}$$

$$= \frac{(n_{11} - n_{1.}n_{.1}/n)^2}{n_{1.}n_{.1}/n} + \frac{(n_{12} - n_{1.}n_{.2}/n)^2}{n_{1.}n_{.2}/n} + \cdots + \frac{(n_{rc} - n_{r.}n_{.c}/n)^2}{n_{r.}n_{.c}/n}$$

- Pour n 'suffisament grand' $(\forall i, j, \quad np_{ij} \geq 5)$, sous H

$$X^2_{obs} \sim \chi^2_{(r-1)\times(c-1)}$$

Exemple

Le tableau suivant montre les résultats d'une enquête auprès de 6672 personnes concernant le lien entre le *sexe* et le fait que la personne est *droitière ou gauchère* :

	hommes	femmes	
droitiers	2780	3281	6061
gauchers	311	300	611
	3091	3581	6672

Tester au niveau $\alpha = 0,05$ l'hypothèse que les caractères 'sexe' et 'main dominante' sont indépendants ...

Solution

Le tableau théorique :

	hommes	femmes	
droitiers	2808	3253	6061
gauchers	283	328	611
	3091	3581	6672

$2808 = 6061 \times 3091/6672$, etc.

$$X^2 = \frac{(2780-2808)^2}{2808} + \frac{(3281-3253)^2}{3253} + \frac{(311-283)^2}{283} + \frac{(300-328)^2}{328}$$

$= 5.68$

Sous H, $X^2 \sim \chi^2_1$; $\chi^2_{1,0.95} = 3.84 < 5.68$.

Donc, on *REJETTE* l'hypothèse nulle : les données indiquent *une déviation significative* de l'hypothèse d'indépendance.

Test d'indépendance : autre exemple

Exemple 12.4 Est-ce que la participation à classe influent sur la réussite à l'examen ? Tester au niveau $\alpha = 0,01$ **...**

	réussite	échec	
présent	25	6	31
absent	8	15	23
	33	21	54

Nombres espérés :

PS :

PF :

AS :

AF :

Test

1

2

3

4

5

Test d'homogénéité

- Le **test d'homogénéité** (ou de comparaison) consiste à vérifier que J échantillons (groupes) *proviennent de la même population*
- C.-à-d., la distribution de la variable d'intérêt est *la même* dans toutes les populations
- Il s'agit d'une *comparaison de la distribution* d'une variable qualitative dans plusieurs échantillons
- Considérons un facteur A qui peut prendre I valeurs différentes sur une population
- La probabilité d'apparition des différentes valeurs de A est $p_{i.}$, $i = 1, \ldots, I$
- Soit J échantillons C_j de tailles $n_{.j}$ respectives issu de la population

Test d'homogénéité, cont.

- Les fréquences d'apparition des valeurs A_i du facteur A dans l'échantillon C_j sont notées n_{ij}
- L'hypothèse nulle H pose que les distributions sont *les mêmes*
- $\Longrightarrow$ les différences observées entre tous les échantillons sont dûes à la *variabilité d'échantillonage* (variation aléatoire)

Modalité du facteur	C_1	C_2	$\cdots$	C_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$					
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot J}$	$n_{\cdot\cdot} = n$

Statistique de test

- Sous H, *les probabilités* d'apparition des différentes modalités du facteur *sont toutes égales dans les distributions marginales*
- Probabilité d'apparition de la modalité j du facteur A :

$$p_{i.} = n_{i.}/n$$

- Nombre espéré de modalité i dans l'échantillon j :

$$E_{ij} = n_{.j}\, p_{i.} = \frac{n_{i.}\, n_{.j}}{n}$$

- La statistique de test (homogénéité) est toujours de forme :

$$X^2_{obs} = \sum \frac{(\text{observé} - \text{espéré})^2}{\text{espéré}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

- À noter : c'est *la même formule* que celle du test d'indépendance
- Sous H, $X^2 \sim \chi^2_{(I-1)(J-1)}$

Exemple

Exemple 12.3, cont.

La même population a été analysée l'année suivante (échantillon de 119 individus) ; la structure génotypique est telle que :

FF	*FS*	*FC*	*SS*	*SC*	*CC*
12	60	22	11	8	6

La distribution génotypique de la population a-t-elle changé depuis l'année précédente ?

Test

1

2

3

4

5

Le paradoxe de Simpson

- Également connu sous le nom de *effet Yule-Simpson*
- Un effet qui se produit lorsque le *association marginale* entre deux variables catégorielles est *qualitativement différent* de la *association partielle* entre les deux mêmes variables **après avoir contrôlé une (ou plusieurs) autres variables**
- Pour l'historique, voir `https://www.britannica.com/topic/Simpsons-paradox`

Le paradoxe de Simpson - exemple concret (I)

- Considérons deux hôpitaux : A et B
- L'année dernière, l'hôpital A a réalisé 2100 interventions chirurgicales, dont 63 patients sont décédés
- L'hôpital B a effectué 800 interventions chirurgicales, dont 16 patients sont décédés
- *Activité :* Calculer la proportion de décès dans chaque hôpital et décider lequel vous choisiriez si vous auriez besoin d'une intervention chirurgicale.

*Choix :*_____

Le paradoxe de Simpson - exemple concret (II)

- Considérons maintenant une *variable cachée* : quel était l'état des patients lorsqu'ils sont arrivés à l'hôpital ?
- Pour ceux en *bon état* :
  - 600 est allé à A et 6 sont morts
  - 600 est allé à B et 8 sont morts
- Pour ceux en *mauvais état* :
  - 1500 est allé à A et 57 sont morts
  - 200 est allé à B et 8 sont morts
- *Activité, suite :* Encore une fois, calculer la proportion de décès dans chaque hôpital et décider lequel vous choisiriez si vous auriez besoin d'une intervention chirurgicale

*Choix :*_____

Le paradoxe de Simpson - exemple concret (III)

- Donc – vous n'avez pas fait le même choix, non *? ?*
- Ce que nous venons de voir est un exemple de *Le paradoxe de Simpson*
- Les données d'un groupe global donnent des proportions qui semblent favoriser un choix plutôt qu'un autre
- MAIS : lorsque les groupes sont **répartis en sous-groupes**, chaque groupe plus petit peut avoir des proportions qui produire le *résultat opposé*