UN GROUPE D'INDIVIDUS
DÉCRIT PAR UNE VARIABLE QUALITATIVE BINAIRE

ANALYSER LA RÉPARTITION EN DEUX SOUS-GROUPES

ANALYSER L'UN DES DEUX GROUPES

COMPARER UN POURCENTAGE À UN POURCENTAGE DE RÉFÉRENCE

Mots-clés : Variable binaire ; Codage disjonctif ; Variable indicatrice ; Groupe d'intérêt ; Groupe de référence ; Mode ; Valeur modale ; Proportion ; Pourcentage ; Rapport des chances ; Odds ; Pourcentage de référence ; Inférence sur un pourcentage ; Test du *Khi*² de conformité à une distribution de référence ; Intervalle de confiance.

Ce document a été établi en indiquant comment obtenir les différents résultats avec le logiciel SES-Pegase (version 7). Cependant, il peut être utilisée comme guide méthodologique et d'interprétation, quel que soit le logiciel utilisé.

TYPE DES DONNÉES ANALYSÉES

Nous présenterons l'analyse d'un dossier particulier, le dossier ETUDIANTS. Mais cette analyse s'applique à toutes données de la forme suivante (Tableau 1) :

Tableau 1 : Structure du tableau des données individuelles

INDIV	B
i1	1
i2	2
i3	2
i4	1
i5	2
i6	2
i7	1
i8	1
(…)	(…)

On a recueilli des informations sur un ensemble d'individus (INDIV : i1, i2…). Ces individus peuvent être, des personnes, des pays, des animaux, des voitures…
Parmi les données recueillies sur ces individus, on a une variable qualitative (B) à deux modalités (1 et 2). Cette variable qualitative binaire peut être, leur sexe (H/F), leur état de santé (malade/pas malade), leur état-civil (célibataire/non célibataire), leur résultat à une épreuve (réussite/échec)…

QUESTIONS

- Quelle est la répartition dans les deux groupes ?
- Quelles sont les caractéristiques du groupe d'intérêt ?
- Comparer le pourcentage d'un sous-groupe à un pourcentage de référence.

UN EXEMPLE : LE DOSSIER ETUDIANTS

On s'intéresse aux réponses à une question posée à des étudiants en 3ème année de psychologie à l'université Paris Descartes (Paris 5) : - « Quel est votre sexe ? »
Les étudiants étaient invités à se choisir un pseudonyme pour pouvoir, s'ils le souhaitaient, se situer dans les tableaux et graphiques.

Le questionnaire a été rempli par l'ensemble des étudiants. Les données ci-dessous sont celles d'un échantillon de 36 questionnaires, extraits au hasard parmi cette population plus vaste.

On considèrera qu'une différence de pourcentage est faible si elle est inférieure à 5 points de pourcentage et importante si elle est supérieure à 10 points de pourcentage.

Tableau 2: Données (partielles) du dossier ETUDIANTS

ETUDIANT	SEX
SRP	2
EMA	2
SABX	2
EMIC	2
AMELIEM	2
CHEBLI	1
FIOR	2
FUERTES	2
SCARLETT	2
CENIE	2
(…)	(…)
TIGRIS	2
MIKY	2
LOULOU	2
ZORRO	2
MELIS	2

La colonne SEXE rapporte le sexe des étudiants interrogés (1 = H ; 2 = F).

Questions

- Quelle est la répartition dans les deux groupes, filles et garçons ?
- On s'intéresse au groupe des filles, les plus nombreuses. Quel est le pourcentage de filles ?
- Le pourcentage de filles chez les étudiants en psychologie est-il différent du pourcentage de filles chez l'ensemble des étudiants durant cette période, soit 55% ?[1]

Source

Denis Corroyer – Université Paris Descartes (Paris 5) – denis.corroyer@parisdescartes.fr
http://www.inegalites.fr/spip.php?article1096

[1] *De la même manière, on pourrait se demander si ce pourcentage est supérieur au pourcentage de filles chez l'ensemble des étudiants, toutes disciplines confondues, ou s'il est supérieur au pourcentage de filles chez les étudiants en Lettres et Sciences Humaines.*

Codage disjonctif et variables indicatrices

Le sexe peut être codé sous deux formes différentes (cf. Tableau 3) :
- soit sous forme d'une seule variable (SEX) à deux modalités (1=G, 2=F),
- soit sous forme de deux variables (GARCON et FILLE) chacune étant codée 1 ou 0. Si l'individu est une fille, la variable GARCON est codée 0 et la variable FILLE est codée 1. Si l'individu est un garçon, la variable GARCON est codée 1 et la variable FILLE est codée 0.

Ce type de codage est appelé *codage disjonctif*. Les deux variables ainsi crées, codées 1 ou 0, sont appelées *variables indicatrices* du sous-groupe (modalité) correspondant.
Le codage disjonctif est nécessaire en préalable à certaines procédures (analyses factorielles, régression multiple…)[2].

Tableau 3 : Données (partielles) du dossier ETUDIANTS avec codage disjonctif (deux variables indicatrices 0/1)

ETUDIANT	SEX	GARCON	FILLE
SRP	2	0	1
EMA	2	0	1
SABX	2	0	1
EMIC	2	0	1
AMELIEM	2	0	1
CHEBLI	1	1	0
FIOR	2	0	1
(…)	(…)		1
TIGRIS	2	0	1
MIKY	2	0	1
LOULOU	2	0	1
ZORRO	2	0	1
MELIS	2	0	1

Ouvrir le dossier et sélectionner la variable

```
SES-Pegase
Lancer SESAnalyse
Menu Fichier – Ouvrir un dossier SES (*.SES)
- Sélectionner le dossier ETUDIANT.SES Ouvrir
Menu Nouvelle analyse
- Sélectionner SEX comme Variable à analyser
```

RÉPARTITION EN DEUX SOUS GROUPES ?

On va chercher à caractériser les deux sous-groupes du point de vue de leurs effectifs et de leurs pourcentages respectifs.

Groupe le plus nombreux ?

```
Menu Statistiques
- Distribution
– Mode.
```

Tableau 4: Mode de la distribution par sexe dans l'échantillon

	SEX
Mod	F
Effectif	33

Le Tableau 4 nous indique que, pour la variable SEX, le groupe le plus nombreux est F (Fille). Ce groupe comprend 33 individus, soit 92% de l'effectif total.

[2] *Ce codage n'est pas indispensable pour l'analyse de cet exemple.*

Terminologie : La modalité Fille est, pour la variable Sexe, le *mode* ou *valeur modale.*

Attention : Ce que l'on désigne par le *mode* ou la *valeur modale* est la modalité de la variable (ici Fille) et non pas l'effectif de ce groupe (33).

Répartition en pourcentages ?

```
Menu Statistiques
- Distribution
- Diagramme à secteurs (camembert)
```

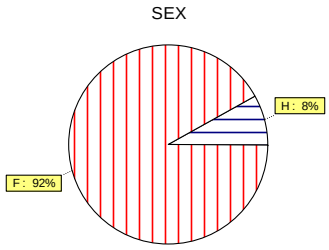


Figure 1 : Répartition des deux sous-groupes en pourcentages (Diagramme à secteurs)

Pour voir les valeurs (effectifs et pourcentages) :

```
Au dessus du graphique précédent, cliquer sur l'icône "Tableau"
ou
Menu Statistiques
- Distribution
- Distribution en effectifs et %
```

Tableau 5 : Distributions des effectifs des deux groupes

SEX	H	F	Total
n	3	33	36
%	8%	92%	100%

On constate, dans cet échantillon de 36 étudiants, une majorité de filles :
on a 92% (33/36) de filles et 8% (3/36) de garçons.

ANALYSER L'UN DES DEUX GROUPES

La première étape consiste à choisir un des deux groupes, le *groupe d'intérêt*, auquel on va plus particulièrement s'intéresser.

Choisir le groupe d'intérêt

Lorsqu'on analyse une variable binaire, on peut s'intéresser à l'une ou l'autre des deux modalités de la variable. Si l'on connaît le pourcentage d'une des modalités, on en déduit immédiatement l'autre. Par exemple, si on connaît le pourcentage des garçons dans un groupe, on peut en déduire, par différence, le pourcentage des filles. Dans le cas d'une variable comme le sexe, le choix de s'intéresser à l'une des modalités plutôt que l'autre est arbitraire. Dans ce cas, on analyse la modalité la plus fréquente. Les résultats sont ainsi, en général, plus lisibles. Ici l'analyse va donc se centrer plutôt sur le groupe des filles. Mais, dans certains cas, ce choix s'impose de par la nature des données. Ainsi, si on analyse une variable du type État de santé (malade/pas malade), on va plutôt s'intéresser au pourcentage de malades.

Terminologie : Le groupe auquel on va s'intéresser plus particulièrement est… le *groupe d'intérêt*. L'autre groupe est le *groupe de référence*.

On choisit ici le sous-groupe de plus fort effectif, le sous-groupe des filles (F).

```
Menu Statistiques
- Indiquer la modalité d'intérêt
Cliquer sur F
```

Puis ce groupe va être étudié du point de vue de deux indices :

- l'indice le plus classique : la fréquence (exprimée sous forme d'un pourcentage),
- un autre indice, certes moins courant, mais ayant des propriétés intéressantes : l'odds.

Fréquence du groupe d'intérêt ?

a/ Dans l'échantillon ?

```
Menu Statistiques
- Analyser la modalité d'intérêt
- % de la modalité d'intérêt
```

Tableau 6: Pourcentage de filles dans l'échantillon

n	36
F	33
%obs	91.70%

Pour cet échantillon de 36 étudiants, la proportion de filles (cf. Tableau 6) est d'environ 92% (33 / 36 = 0.917).
Nous le savions déjà, mais ce tableau va nous permettre de prolonger ce résultat par l'inférence.

b/ Dans la population ?

Il arrive fréquemment que l'on analyse, du point de vue d'une variable, un groupe limité d'individus dans la perspective de connaître, du point de vue de cette même variable, une population plus vaste dont fait partie ce groupe. C'est le cas ici : le groupe de 36 étudiants est un *échantillon* de la *population* des étudiants qui ont répondu au questionnaire. On souhaite connaître la répartition par sexe chez ces étudiants. Deux solutions : analyser tous les questionnaires et on connaîtra cette répartition avec certitude ou, si on en n'a, ni le temps, ni le courage, ni les financements nécessaires, on pourra avoir recours aux procédures statistiques qui nous donneront une estimation de cette répartition.

```
Au-dessus du tableau précédent, cliquer sur Inférence
ou :
Menu Statistiques
- Analyser la modalité d'intérêt
- IC sur le % de la modalité d'intérêt.
```

Tableau 7: Intervalle de confiance sur le pourcentage de filles dans la population

SEX	**F**
%obs	91.7%
p	.05
Garantie	95%
IC_inf.	76.4 %
IC_sup.	99.0 %

Ce tableau nous indique que, au seuil .05 (on dit aussi à la garantie 95%), le pourcentage de filles dans la population est compris entre 76 % et 99 % (*IC* 95% = [76 % ; 99 %]).
On constate que l'incertitude reste relativement grande. Ceci est dû au faible effectif de l'échantillon recueilli ici.[3]

[3] *Plus l'échantillon est de taille importante (se rapproche de la taille de la population) plus l'intervalle de confiance est étroit. A la limite, si l'échantillon est égal à la population, l'intervalle de confiance ne comprendra qu'une seule valeur. Cette valeur sera, avec certitude, le pourcentage des filles dans la population.*

Il importe de définir le plus précisément la population dont il est question. Ici, ce n'est pas la population des étudiants en psychologie, mais celle des étudiants de psychologie…
… de cette année universitaire (1998-1999). Il est probable que la répartition par sexe ait changé depuis,
… de cette année d'études (3ème). Il semble que la proportion des garçons augmente au fil des années d'études,
… de cette université (Paris Descartes). Il est possible que cette répartition soit différente d'une université à l'autre.

On constate que
sur un échantillon de 36 étudiants de psychologie en 3ème année de psychologie de l'université Paris Descartes,
le pourcentage de filles est de 92%.
Il semble que,
pour l'ensemble des étudiants d'où a été extrait cet échantillon,
le pourcentage des filles est compris entre 76% et 99% (*IC* 95% = [76 % ; 99 %]),

L'odds

Une autre manière d'exprimer la part des filles dans l'échantillon est de rapporter le nombre de filles (33) au nombre de garçons (3). Le rapport de ces deux effectifs (33/3 = 11) est parfois désigné par le terme *cote*, mais plus fréquemment par le terme anglais correspondant *odds*.[4]

```
Menu Statistiques
– Analyser la modalité d'intérêt
- Odds
```

Tableau 8 : Odds ou Rapport des pourcentages

	SEX
F	33
H	3
odds	11

L'odds (11) est égal au rapport de l'effectif de la modalité d'intérêt (33) sur l'effectif de l'autre modalité (3).
Comment s'interprète un odds ? [5]
On peut utiliser plusieurs formulations :
« Dans cet échantillon de 36 étudiants…
…le nombre de filles est 11 fois plus élevé que le nombre de garçons »
…le nombre de garçons est 11 fois plus faible que celui des filles »
…pour 1 garçon, on a 11 filles »
…il y a 11 filles pour 1 garçon »

COMPARER UN POURCENTAGE À UN POURCENTAGE DE RÉFÉRENCE

Comparons la fréquence des filles chez ces étudiants en psychologie à la fréquence des filles dans l'ensemble des étudiants, fréquence égale à 55%. Le pourcentage de filles dans cette population particulière est-il différent de 55% ? Si oui dans quel sens ? Inférieur ou supérieur ? L'écart à 55% est-il faible ou important ?

Existence et sens de l'écart

a/ Dans l'échantillon ?

```
Menu Statistiques
– Comparer la modalité d'intérêt à un % de référence
– % modalité d'intérêt et % de référence
Indiquer le % de référence : 55
```

[4] *Il est essentiel de se familiariser avec cet indice, moins « naturel » qu'un pourcentage, car il constitue la base d'un indice, l'odds ratio, utilisé pour analyser des données plus complexes.*
[5] *On peut également donner une interprétation probabiliste de cet indice. On parle alors, en français, des « chances » plutôt que des odds. Si odds = 11, on dira : « si je croise par hasard un étudiant en psychologie, il y a 11 fois plus de chances que ce soit une fille, plutôt qu'un garçon ».*

Tableau 9 : Pourcentage de filles dans l'échantillon et valeur de référence (55%)

SEX	F
%obs	91.7%
%ref	55%

Ce tableau nous indique que,
dans l'échantillon de 36 étudiants,
le pourcentage de filles (92%)
est supérieur au pourcentage de filles dans la population générale (55%).

b/ Dans la population

```
Cliquer sur Inférer
ou
Menu Statistiques
– Comparer la modalité d'intérêt à un % de référence
– Test Z comparaison à un % de référence
Indiquer le % de référence : 55
```

Tableau 10 : Test Z corrigé pour la comparaison du pourcentage de filles

%obs	91.7 %
%ref	55 %
Z_corr	4.25
p	< .0001

Lorsque la valeur *p* est inférieure à .05 (seuil repère usuel), comme ici, le test est, selon la formule usuelle, déclaré *significatif*. Dans ce cas, on peut conclure :
1. Sur l'existence d'une différence : en ce qui concerne le pourcentage de filles, il existe une différence entre les deux populations d'étudiants, celle des étudiants en psychologie et l'ensemble des étudiants.
2. Sur le sens de cette différence : le pourcentage de filles est (comme dans l'échantillon) plus élevé dans la population des étudiants en psychologie que dans la population de tous les étudiants, toutes disciplines confondues.

Si la valeur *p* avait été supérieure à .05, le test aurait été déclaré *non significatif*.
Dans ce cas, on ne peut pas conclure sur l'existence, ou non, d'un écart entre le pourcentage de filles parmi les étudiants en psychologie et parmi l'ensemble des étudiants, toutes disciplines confondues.

Attention : Dans le cas de test non significatif, ne pas pouvoir conclure à l'existence d'un écart, ne permet pas de conclure à l'absence d'écart.

Ampleur et importance de l'écart

a/ Dans l'échantillon ?

```
Menu Statistiques
– Comparer le groupe d'intérêt à une référence
– Différence des % (intérêt - référence)
Indiquer le % de référence : 55
```

Tableau 11 : Écart entre les pourcentages de filles dans l'échantillon et chez l'ensemble des étudiants.

SEX	F
%obs	91.7%
%ref	55.0%
%obs-%Ref	+37pts%

Ce tableau nous indique que,
dans l'échantillon de 36 étudiants,
l'ampleur de l'écart entre ces deux pourcentages
est de 37 points (de pourcentage).
Il peut être qualifié d'important
car supérieur à 10 points de % (cf. présentation des données).

b/ Dans la population ?

```
Cliquer sur Inférer
ou
Menu Statistiques
– Comparer la modalité d'intérêt à un % de référence
- IC écart à un % de référence
Indiquer le % de référence : 55
```

Tableau 12 : Intervalle de confiance sur l'écart entre le pourcentage de filles en psychologie et chez la population générale

SEX	F
%obs	91.70%
%ref	55.00%
%obs - %ref	+37 pts %
p	.05
Garantie	95%
IC_inf	21.4 pts%
IC_sup	44.0 pts%

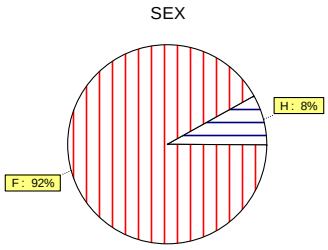
Ce tableau indique que, au seuil .05 (on dit aussi à la garantie 95%), l'écart entre le pourcentage de filles chez les étudiants en psychologie et chez l'ensemble des étudiants, toutes disciplines confondues, est compris entre 21 points de % et 44 point de % (*IC* 95% = [21.4 pts % ; 44.0 pts %]).

On situe cet intervalle par rapport aux valeurs-repères (5 pts et 10 pts de %) indiquées dans la présentation des données. Cet écart peut être jugé important car toutes les valeurs de l'intervalle sont supérieures à 10 pts de %.

RÉDIGER UN COMPTE RENDU DE L'ANALYSE

On s'intéresse au pourcentage de filles
dans la population des étudiants de psychologie en 3ème année de Licence à Paris Descartes.

Dans un échantillon de 36 étudiants,
les filles sont plus nombreuses que les garçons.
Le pourcentage de filles est de 92%,
soit 11 fois plus de filles que de garçons (*odds* = 11)



On peut estimer que,
dans la population d'où provient l'échantillon observé,
le pourcentage de filles est compris entre 76% et 99% (*IC* 95% = [76% ; 99%]).

Le pourcentage de filles dans l'échantillon de 36 étudiants (92%)
est supérieur à celui de l'ensemble des étudiants, toutes disciplines confondues (55%).
L'écart observé (37 points) peut être considéré comme important car supérieur à 10 points de pourcentage.

Il semble que,
dans la population d'où provient l'échantillon observé,
le pourcentage de filles est supérieur à celui de l'ensemble des étudiants, toutes disciplines confondues
(z_{cor}= 4.25, p < .0001).
L'écart peut être considéré comme important
(*IC* 95% = [21 pts% ; 44 pts%] > 10 pts%).

ANNEXES

EXERCICE

On peut également se demander si le pourcentage des filles en psychologie diffère du pourcentage de filles dans la population des étudiants en Lettres et Sciences Humaines. Ce pourcentage est de 77%.

Reprendre les procédures ci-dessous avec ces autres pourcentages de référence. A quelles conclusions parvient-on ?

CALCULS

Calcul d'un odds

L'*odds* peut être calculé, soit à partir des effectifs, soit à partir des pourcentages.

Dans l'exemple ETUDIANTS, parmi les 36 étudiants de l'échantillon, on a 33 filles et 3 garçons. L'*odds* d'être une fille est donc égal à 33/3 = 11

Par ailleurs, les garçons représentent 8% (plus précisément 8.333…%) et les filles 92 % (plus précisément 91.666…%) de l'échantillon.
L'*odds* d'être une fille peut être également obtenu en calculant le rapport de ces deux pourcentages : 91.666… % / 8.3333… % = 11

En effet, le pourcentage de filles est égal = 33/36 et le pourcentage de garçons est égal à 3/36. Le rapport de ces deux pourcentages…

$$\frac{\frac{33}{36}}{\frac{3}{36}} = \frac{33}{36} \times \frac{36}{3} = \frac{33}{3} = 11$$

…est bien égal au rapport des effectifs, c'est-à-dire à l'odds (11).

RÉFÉRENCES

Agresti, A. (1990). *Categorical data analysis*. Wiley.

Corroyer, D., & Wolff, M. (2003). *L'Analyse Statistique des Données en Psychologie; Concepts et Méthodes de base*. Paris: Armand Colin (Cursus).

Rouanet, H., Bernard, J.-M., & Le Roux, B. (1990). *Statistique en Sciences Humaines. Analyse Inductive des Données*. Paris: Dunod.

Bernard, J.-M. (1998). Bayesian Inference for Categorized Data. *In H. Rouanet & J.-M. Bernard & M.-C. Bert & B. Lecoutre & M.-P. Lecoutre & B. L. Roux (Eds.), New Ways in Statistical Methodology - From Significance Tests to Bayesian Inference* (pp. 159-226). Berne: Peter Lang.

LISTE DES TABLEAUX

LISTE DES FIGURES

SOMMAIRE