Hypothesis testing on several parameters
Test d'hypothèse sur plusieurs paramètres

Advanced Statistical Analysis

Feryal WINDAL

1.Introduction

Introduction

➡Dans ce chapitre propose une extension à $k > 2$ paramètres les tests d'hypothèses présentés dans le chapitre précédent.

➡Ces tests sont également utilisables lorsque $k=2$.

➡Ils présentent ainsi une véritable généralisation du traitement des tests d'homogénéité.

➡Le principe de fonctionnement est le même qu'au chapitre précédent :

- on dispose d'un caractère dont les modalités permettent d'identifier k sous-populations
- sur lesquelles on observe les réalisations d'une même variable, quantitative ou qualitative

➡In this chapter proposes an extension to $k> 2$ parameters of the hypothesis tests presented in the previous chapter.

➡These tests can also be used when $k = 2$.

➡They thus present a true generalization of the treatment of homogeneity tests.

➡The principle of operation is the same as in last chapter :

- we have a trait whose modalities allow us to identify k sub-populations
- on which we observe the achievements of the same variable, quantitative or qualitative

Introduction

➡L'hypothèse nulle du test est l'égalité entre paramètres.

➡L'hypothèse alternative est la négation de l'hypothèse nulle.

➡L'identification de la région de non-rejet de l'hypothèse nulle nécessite l'usage de la table de deux distributions de probabilités distinctes :

- la table de Fisher pour un test portant sur des moyennes.
- la table du Khi-Deux pour un test portant sur des pourcentages.

➡The null hypothesis of the test is equality between parameters.

➡The alternative hypothesis is the negation of the null hypothesis.

➡Identifying the non-rejection region of the null hypothesis requires the use of the two distinct probability distributions table:

- Fisher's table for a test on means
- The chi-square table for a test on percentages.

2.Hypothesis Tests on Means: Analysis of Variance

2.1 Test principles

Test principles

Le test ANOVA (ANalysis Of VAriance) présente les hypothèses nulle et alternative suivantes :

The ANOVA (ANalysis Of VAriance) test presents the following null and alternative hypotheses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j \end{cases}$$

ce qui peut être réécrit :

which can be rewritten:

$$\begin{cases} H_0 : \text{ all means are equal} \\ H_1 : \text{ all means are not equal,} \end{cases}$$

où μ_i est la moyenne de la variable sur la sous-population i.

where μ_i is the mean of the variable over subpopulation i.

Test principles

➡ Le test ANOVA est réalisé sous 3 hypothèses :

1. Les échantillons tirés des populations sont aléatoires et indépendants
2. Les distributions de probabilités de la variable sur chaque population sont au moins approximativement normales
3. Les variances de la variable sont égales sur les différentes sous-populations :

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

➡ The ANOVA test is carried out under 3 assumptions:

1. Samples drawn from populations are random and independent
2. The probability distributions of the variable on each population are at least approximately normal
3. The variances of the variable are equal over the different sub-populations:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

Test principles

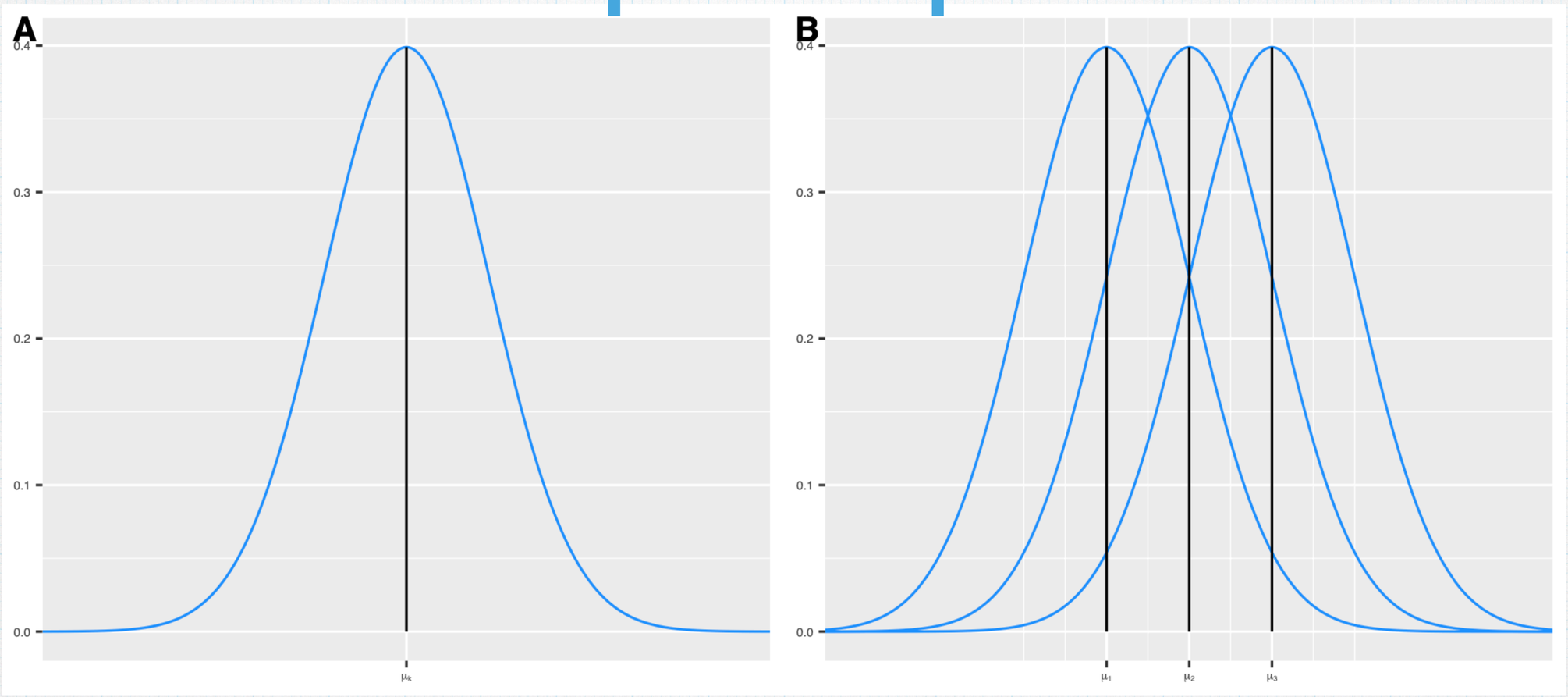


Figure A: k populations normales avec $\mu_1 = \mu_2 = \ldots = \mu_k$ et $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

Figure B : populations normales avec $\mu_1 \neq \mu_2 \neq \mu_3$ et $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

Figure A: k normal populations with $\mu_1 = \mu_2 = \ldots = \mu_k$ and $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

Figure B: normal populations with $\mu_1 \neq \mu_2 \neq \mu_3$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

2.2 Within (intra) and between (inter) estimators

Within (intra) and between (inter) estimators

Le test ANOVA consiste à calculer deux estimateurs alternatifs de la variance d'une variable.

➡Le premier estimateur se présente sous la forme d'une moyenne pondérée des écarts des moyennes calculées sur les k échantillons à la moyenne de l'échantillon total.

- Cet estimateur est appelé estimateur between.
- Cette estimation de la variance de la variable sur la population reste valide que l'hypothèse nulle soit non-rejetée ou non.

The ANOVA test consists in calculating two alternative estimators of the variance of a variable.

➡The first estimator is in the form of a weighted average of the deviations of the means calculated on the k samples from the mean of the total sample.

- This estimator is called the between estimator.
- This estimate of the variance of the variable over the population remains valid whether the null hypothesis is not rejected or not.

Sampling distribution of the two means differences

➡Le second estimateur propose une estimation sans biais de la variance de la variable sous l'hypothèse que les moyennes sur chaque sous-population sont égales, i.e., si l'hypothèse nulle n'est pas rejetée.

- Cet estimateur est appelé estimateur within.

➡ Le rapport critique (RC) utilisé fait le rapport de la valeur de ces deux estimateurs :

- si l'hypothèse nulle d'égalité des moyennes est vérifiée, alors les valeurs des deux estimateurs sont très proches et RC est proche de 1.

➡The second estimator provides an unbiased estimate of the variance of the variable under the assumption that the means over each subpopulation are equal, i.e., if the null hypothesis is not rejected.

- This estimator is called the within estimator.

➡ The critical ratio (CR) used reports the value of these two estimators:

- if the null hypothesis of equality of means is true, then the values of the two estimators are very close and CR is close to 1.

Between estimator

L'estimateur between de la variance de la variable s'écrit:

L'estimateur between de la variance de la variable s'écrit: $$\hat{\sigma}^2_{between} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1}$$

n_i the size of the sub-sample i

la taille du sous-échantillon i

$\bar{x}_i$ the mean of the variable on sub-sample i

la moyenne de la variable sur le sous-échantillon i

$\sum_{i=1}^{k} n_i = n$ the total sample size

la taille totale de l'échantillon

$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$ the mean of the sub-sample means (i.e., nor the sample mean)

la moyenne des moyennes sous-échantillonnales (i.e., ni la moyenne de l'échantillon)

k the number of parameters to be estimated (equal to the number of distinguished sub-populations

le nombre de paramètres à estimer (égal au nombre de sous-populations distinguées).

Within estimator

L'estimateur between de la variance de la variable s'écrit:

L'estimateur between de la variance de la variable s'écrit:

$$\hat{\sigma}^2_{within} = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{n - k}$$

n_i the size of the sub-sample i

la taille du sous-échantillon i

$\overline{x}_i$ the mean of the variable on sub-sample i

la moyenne de la variable sur le sous-échantillon i

s_i^2 the variance of the variable over the sub-sample i

la variance de la variable sur le sous-échantillon i

k the number of parameters to be estimated (equal to the number of distinguished sub-populations).

le nombre de paramètres à estimer (égal au nombre de sous-populations distinguées).

2.2 Statistic test

Statistic test

Le rapport critique s'écrit :

The critical report is written:

$$CR = \frac{\hat{\sigma}^2_{between}}{\hat{\sigma}^2_{within}} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1}}{\frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{n-k}} \sim \mathfrak{F}^{\alpha}_{k-1,n-k}$$

La région de non-rejet de l'hypothèse nulle est définie par :

The null hypothesis non-rejection region is defined by:

$$AR = \left[0; \mathfrak{F}^{\alpha}_{(k-1,n-k)}\right]$$

où $\mathfrak{F}^{\alpha}_{(k-1,n-k)}$ est lue dans une table de Fisher au seuil de significativité avec $k-1$ degrés de liberté au numérateur et $n-k$ degrés de liberté au dénominateur.

where $\mathfrak{F}^{\alpha}_{(k-1,n-k)}$ is read from a Fisher table at the significance level with $k-1$ degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator.

- Si $CR \in AR$, l'hypothèse nulle d'égalité des moyennes ne peut pas être rejetée
- Si $CR \notin AR$ l'hypothèse nulle d'égalité des moyennes peut être rejetée

- If $CR \in AR$ the null hypothesis of equality of means cannot be rejected
- If $CR \notin AR$ the null hypothesis of equality of means can be rejected

Example 1

D’après une revue médicale, la satisfaction des différentes professions médicales ne serait pas la même.
Quatre catégories de médecins ont été interrogées : médecins généralistes, pédiatres, dermatologues et chirurgiens. Les médecins ont exprimé leur satisfaction vis-à-vis de leur profession à l’aide d’un indice variant de 0 (aucune satisfaction) à 100 (satisfaction maximale). Les données sont récapitulées dans le tableau.
Au seuil $\alpha = 5\%$, peut-on dire si la satisfaction des médecins vis-à-vis de leur profession est la même quelle que soit leur spécialité?

According to a medical journal, the satisfaction of different medical professions is not the same.
Four categories of doctors were questioned: general practitioners, pediatricians, dermatologists and surgeons. Doctors expressed their satisfaction with their profession using an index ranging from 0 (no satisfaction) to 100 (maximum satisfaction). The data is summarized in the table.
At the $\alpha = 5\%$ threshold, can we say whether physicians' satisfaction with their profession is the same whatever their specialty?

Example 1

Médecin généraliste = general practitioner

Pédiatre = pediatrician

Dermatologue =dermatologist

Chirurgien = Surgeon

Médecins généralistes	Pédiatres	Dermatologues	Chirurgiens
44	55	54	44
42	78	65	73
74	80	79	71
42	86	69	60
53	60	79	64
50	59	64	66
45	62	59	41
48	52	78	55
64	55	84	76
38	50	60	62

On construit les indicateurs nécessaires au calcul des estimateurs between et within :

We construct the indicators necessary to calculate the between and within estimators:

	Médecins généralistes	Pédiatres	Dermatologues	Chirurgiens
$\overline{x}_i$	50	63,7	69,1	61,2
s_i^2	111,80	148,21	95,29	122,96
n_i	10	10	10	10

Example 1

➡Les hypothèses du test s'écrivent :

➡The hypotheses of the test are written:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j \end{cases}$$

On fait l'hypothèse :

- ➡que les échantillons ont été tirés aléatoirement et sont indépendants ;
- ➡que le score de satisfaction suit, au moins approximativement, une distribution de probabilité normale;
- ➡et que le score admet la même variance sur les quatre sous-populations identifiées.

We make the assumption:

- ➡that the samples were drawn at random and are independent;
- ➡that the satisfaction score follows, at least approximately, a normal probability distribution
- ➡and that the score admits the same variance over the four identified sub-populations.

Example 1

La moyenne sur l'échantillon total est égale à :

The average over the total sample is equal to:

$$\bar{x} = \frac{\sum_{i=1}^{4} n_i \bar{x}_i}{\sum_{i=1}^{4} n_i} = \frac{10 \times 50 + 10 \times 63{,}7 + 10 \times 69{,}1 + 10 \times 61{,}2}{10 + 10 + 10 + 10} = 61$$

L'estimateur between s'écrit :

The between estimator is written:

$$\hat{\sigma}^2_{between} = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{k - 1}$$

$$= \frac{10 \times (50 - 61)^2 + 10 \times (63{,}7 - 61)^2 + 10 \times (69{,}1 - 61)^2 + 10 \times (61{,}2 - 61)^2}{4 - 1}$$

$$= 646{,}5$$

Example 1

L'estimateur within s’écrit :

The within estimator is written:

$$\hat{\sigma}^2_{within} = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{n - k}$$

$$= \frac{(10-1) \times 111{,}05 + (10-1) \times 148{,}21 + (10-1) \times 95{,}29 + (10-1) \times 122{,}96}{40 - 4}$$

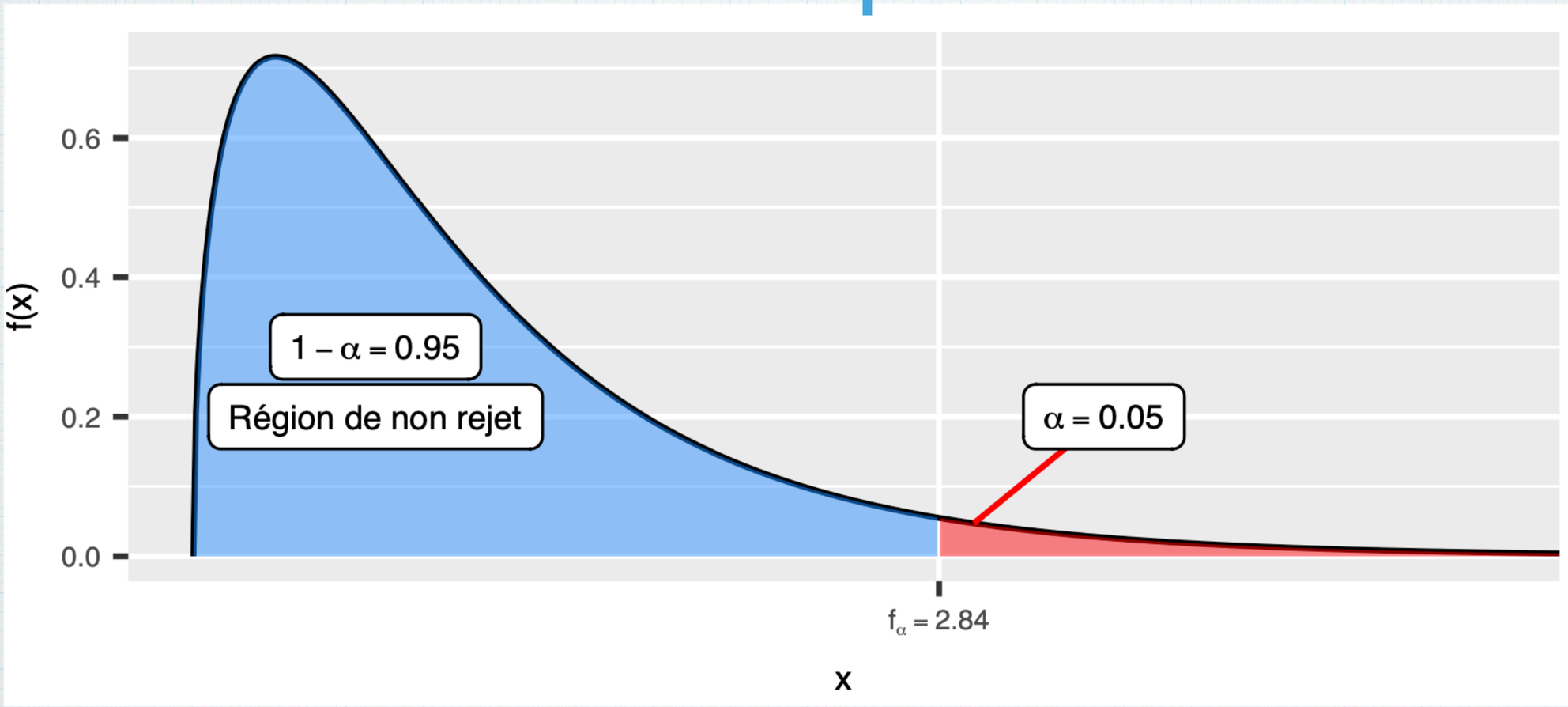
$$= 119{,}6$$

The critical ratio is equal to:

Le rapport critique est égal à :

$$CR = \frac{\hat{\sigma}^2_{between}}{\hat{\sigma}^2_{within}} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})}{k-1}}{\frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{n-k}} = \frac{646{,}5}{119{,}6} = 5{,}40$$

Example 1



On lit le quantile d'ordre $\alpha = 5\,\%$ dans une table de Fisher à (4-1; 40-4) soit (3;36) degrés de liberté :

We read the quantile of order $\alpha = 5\,\%$ in a Fisher table with (4-1; 40-4) that is (3; 36) degrees of freedom:

$$\mathfrak{F}_{3,36}^{0,05} = 2,84$$

Example 1

La région de non-rejet de l'hypothèse nulle a pour écriture :

The non-rejection region of the null hypothesis is written:

$$AR = \left[0; \mathfrak{F}^{\alpha}_{(n_2-1,n_1-1)}\right] = \left[0; \mathfrak{F}^{0,05}_{(3,36)}\right] = [0; 2{,}84]$$

Donc $CR \notin AR \Rightarrow H_0$ rejetée

Then $CR \notin AR \Rightarrow H_0$ is rejected

Au seuil $\alpha = 5\,\%$ et sur la base des informations disponibles, on ne peut pas considérer que les médecins retirent la même satisfaction de leur métier, quelle que soit leur spécialité.

At the threshold $\alpha = 5\,\%$ and on the basis of the information available, we cannot consider that doctors derive the same satisfaction from their profession, whatever their specialty.

Example 2

En 1966, Sax et Cromach étudièrent l'effet de l'ordre de présentation des items sur la performance obtenue à un test. Ayant choisi 70 items, ils formèrent 4 types de tests constitués des mêmes items mais ordonnés différemment:

- Type A : les items sont présentés du plus facile au plus difficile ;
- Type B : les items sont présentés du plus difficile au plus facile ;
- Type C : Un item facile sépare (en moyenne) 6 items difficiles ;
- Type D : L'ordre des items est aléatoire.

In 1966, Sax and Cromach studied the effect of the order in which items were presented on test performance. Having chosen 70 items, they formed 4 types of tests made up of the same items but ordered differently:

- Type A : items are presented from easiest to most difficult;
- Type B : the items are presented from the most difficult to the easiest;
- Type C : An easy item separates (on average) 6 difficult items;
- Type D :The order of the items is random.

Example 2

A chaque type de test est affecté un échantillon différent de sujets. Sax et Cromach mesurèrent la performance et obtinrent les résultats résumés dans le tableau suivant.

Au seuil $\alpha = 5\,\%$, l'ordre de présentation des items paraît-il avoir une incidence sur la performance au test ?

Each type of test is assigned a different sample of subjects. Sax and Cromach measured the performance and obtained the results summarized in the following table.

At the $\alpha = 5\,\%$ Does the order of presentation of items seem to have an impact on test performance?

Type of test	A	B	C	D
Mean	48	46	43	42
S-deviation	11	15	12	13
Effectif	50	45	49	47

Example 2

➡Les hypothèses du test s'écrivent :

➡The hypotheses of the test are written:

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D \\ H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j \end{cases}$$

On fait l'hypothèse :

- ➡que les échantillons ont été tirés aléatoirement et sont indépendants ;
- ➡que le score de satisfaction suit, au moins approximativement, une distribution de probabilité normale;
- ➡et que le score admet la même variance sur les quatre sous-populations identifiées.

We make the assumption:

- ➡that the samples were drawn at random and are independent;
- ➡that the satisfaction score follows, at least approximately, a normal probability distribution
- ➡and that the score admits the same variance over the four identified sub-populations.

Example 2

La moyenne sur l'échantillon total est égale à :

The average over the total sample is equal to:

$$\bar{x} = \frac{\sum_{i=A}^{D} n_i \bar{x}_i}{\sum_{i=A}^{D} n_i} = \frac{50 \times 48 + 45 \times 46 + 49 \times 43 + 47 \times 42}{50 + 45 + 49 + 47} = 44{,}77$$

L'estimateur between s'écrit :

The between estimator is written:

$$\hat{\sigma}^2_{between} = \frac{\sum_{i=A}^{D} n_i (\bar{x}_i - \bar{x})^2}{k - 1}$$

$$= \frac{50 \times (48 - 44{,}77)^2 + 45 \times (46 - 44{,}77)^2 + 49 \times (43 - 44{,}77)^2 + 47 \times (42 - 44{,}77)^2}{4 - 1}$$

$$= 367{,}95$$

Example 2

L'estimateur within s'écrit :

The within estimator is written:

$$\hat{\sigma}^2_{within} = \frac{\sum_{i=A}^{D}(n_i - 1)s_i^2}{\sum_{i=A}^{D} n_i - k}$$

$$= \frac{(50-1)\times 11^2 + (45-1)\times 15^2 + (49-1)\times 12^2 + (47-1)\times 13^2}{\underbrace{50+45+49+47}_{=179} - 4}$$

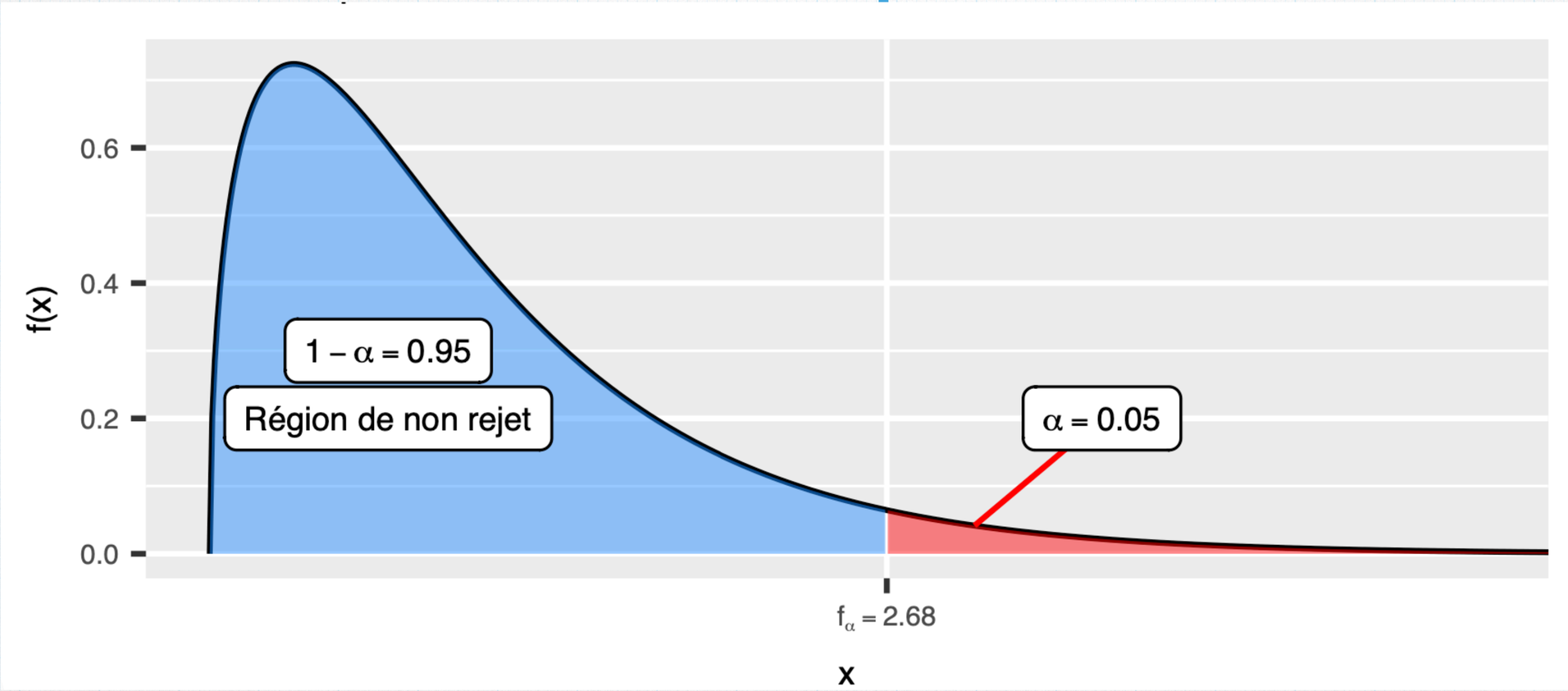
$$= 174{,}37$$

The critical ratio is equal to:

Le rapport critique est égal à :

$$CR = \frac{\hat{\sigma}^2_{between}}{\hat{\sigma}^2_{within}} = \frac{\frac{\sum_{i=A}^{D} n_i(\bar{x}_i - \bar{x})}{k-1}}{\frac{\sum_{i=A}^{D}(n_i-1)s_i^2}{n-k}} = \frac{367{,}95}{174{,}37} = 2{,}11$$

Example 2



On lit le quantile d'ordre $\alpha = 5\,\%$ dans une table de Fisher à (4-1; 179-4) soit (3;175) degrés de liberté :

We read the quantile of order $\alpha = 5\,\%$ in a Fisher table with (4-1; 179-4) that is (3; 175) degrees of freedom:

$$\mathfrak{F}_{3,175}^{0,05} = 2,68$$

Example 2

La région de non-rejet de l'hypothèse nulle a pour écriture :

The non-rejection region of the null hypothesis is written:

$$AR = \left[0; \mathfrak{F}^{\alpha}_{(n_2-1,n_1-1)}\right] = \left[0; \mathfrak{F}^{0,05}_{(3,175)}\right] = [0; 2{,}68]$$

Donc $CR \in AR \Rightarrow H_0$ n'est pas rejetée

Then $CR \in AR \Rightarrow H_0$ is not rejected

Au seuil $\alpha = 5\,\%$ on ne rejette pas l'hypothèse nulle d'égalité des performances quel que soit l'ordre de présentation des items. Autrement dit, l'ordre d'apparition des questions ne semble pas, dans l'expérience de Sax et Cromach, avoir un impact sur la qualité des réponses.

we do not reject the null hypothesis of equality of performance regardless of the order in which the items are presented. In other words, the order of appearance of the questions does not seem, in the experience of Sax and Cromach, to have an impact on the quality of the answers.

3.Hypothesis tests on percentages: Chi-square test

3.1 Test principles

Test principles

Le test du Khi-Deux permet de juger de l'homogénéité avec laquelle un phénomène se manifeste dans une population.
Bien entendu, il permet avant tout de juger de l'égalité de la fréquence d'occurrence d'un événement discret dans plusieurs sous-populations. Par exemple :

- ✓ taux de vaccination contre la grippe saisonnière dans les différents pays
- ✓part de la population ayant vu des extra-terrestres au Texas, dans le Kansas et le Nevada

The chi-square test is used to judge the homogeneity with which a phenomenon manifests itself in a population.
Of course, above all, it allows to judge the equality of the frequency of occurrence of a discrete event in several sub-populations. For example :

- ✓vaccination rate against seasonal influenza in the various countries
- ✓share of the population having seen aliens in Texas, Kansas and Nevada

Plus généralement encore, le test du Khi-Deux autorise à tester l'égalité entre distributions en fréquences de plusieurs sous-populations selon un nombre de modalités supérieur à 2. Par exemple :

- ✓ distribution en fréquences des réponses: jamais / rarement / souvent / très souvent à la question: allez-vous au cinéma ? sur trois sous-populations :
  - ➡les jeunes (<20 ans),
  - ➡les jeunes adultes (20-40 ans),
  - ➡les adultes (>40 ans) ;
- ✓distribution en fréquences des suffrages exprimés au cours du premier tour de l'élection présidentielle de 2012 sur quatre sous-populations: communes de moins de 10000 habitants, de 10001 à 100 000 habitants, de 100001 à 500000 habitants et de plus de 500000 habitants.

More generally, the chi-square test allows to test the equality between frequency distributions of several sub-populations according to a number of modalities greater than 2. For example:

- ✓frequency distribution of responses: never / rarely / often / very often to the question: do you go to the cinema? on three sub-populations:
  - ➡young people (<20 years),
  - ➡young adults (20-40 years old),
  - ➡adults (> 40 years old);
- ✓frequency distribution of the votes cast during the first round of the 2012 presidential election over four sub-populations: municipalities with less than 10,000 inhabitants, 10,001 to 100,000 inhabitants, 100,001 to 500,000 inhabitants and more than 500,000 inhabitants.

Test principles

On considère un caractère décrit à l'aide de p modalités sur q sous- populations.

On compte ainsi n_{ij} individus qui présentent la modalité i du caractère X dans la sous-population j.

We consider a characteristic described using p modalities on q subpopulations.

There are thus n_{ij} individuals who present the modality i of the characteristic X in the subpopulation j.

		Sous-populations						Total
		$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\cdots$	$\mathbf{Y}_j$	$\cdots$	$\mathbf{Y}_q$	
Modalités du caractère $\mathbf{X}$	$\mathbf{X}_1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\bullet}$
	$\mathbf{X}_2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2q}	$n_{2\bullet}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\mathbf{X}_i$	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\bullet}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\mathbf{X}_p$	n_{p1}	n_{p2}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet q}$	

Test principles

On peut exprimer à partir de ce tableau les q distributions en fréquences conditionnelles du point de vue des sous-populations..

We can express from this table the q conditional frequency distributions from the point of view of the sub-populations.

		Sous-populations						Total
		$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\cdots$	$\mathbf{Y}_j$	$\cdots$	$\mathbf{Y}_q$	
Modalités du caractère $\mathbf{X}$	$\mathbf{X}_1$	$f_{11/j=1}$	$f_{12/j=2}$	$\cdots$	$f_{1j/j}$	$\cdots$	$f_{1q/j=q}$	$f_{1\bullet}$
	$\mathbf{X}_2$	$f_{21/j=1}$	$f_{22/j=2}$	$\cdots$	$f_{2j/j}$	$\cdots$	$f_{2q/j=q}$	$f_{2\bullet}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\mathbf{X}_i$	$f_{i1/j=1}$	$f_{i2/j=2}$	$\cdots$	$f_{ij/j}$	$\cdots$	$f_{iq/j=q}$	$f_{i\bullet}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\mathbf{X}_p$	$f_{p1/j=1}$	$f_{p2/j=2}$	$\cdots$	$f_{pj/j}$	$\cdots$	$f_{pq/j=q}$	$f_{p\bullet}$
Total		100%	100%	$\cdots$	100%	$\cdots$	100%	

$$\begin{cases} H_0 : p_{i1} = p_{i2} = \ldots = p_{iq}, \forall i = 1,\ldots,p \\ H_1 : \exists i \neq j \text{ tel que } p_{ij} \neq p_{ik} \text{ for } i = 1,\ldots,p \end{cases}$$

In practice

Dans les faits :

➡on compare les effectifs n_{ij} observés à ceux théoriques qui auraient été observés si dans chaque sous-population j, les $n_{\bullet j}$ individus s'étaient répartis exactement de la même manière que dans la population totale.

In the facts :

➡we compare the numbers n_{ij} observed with the theoretical ones that would have been observed if in each subpopulation j, the $n_{\bullet j}$ individuals had been distributed in exactly the same way as in the total population.

The estimated workforce is written as follows:

L'effectif estimé s'écrit ainsi :

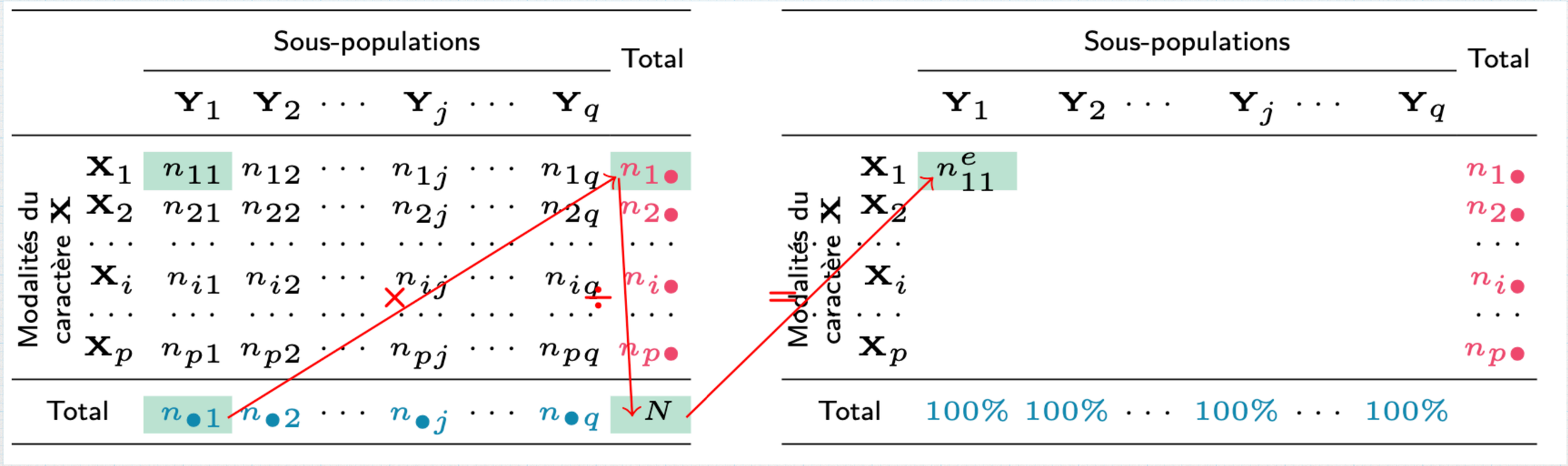
$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

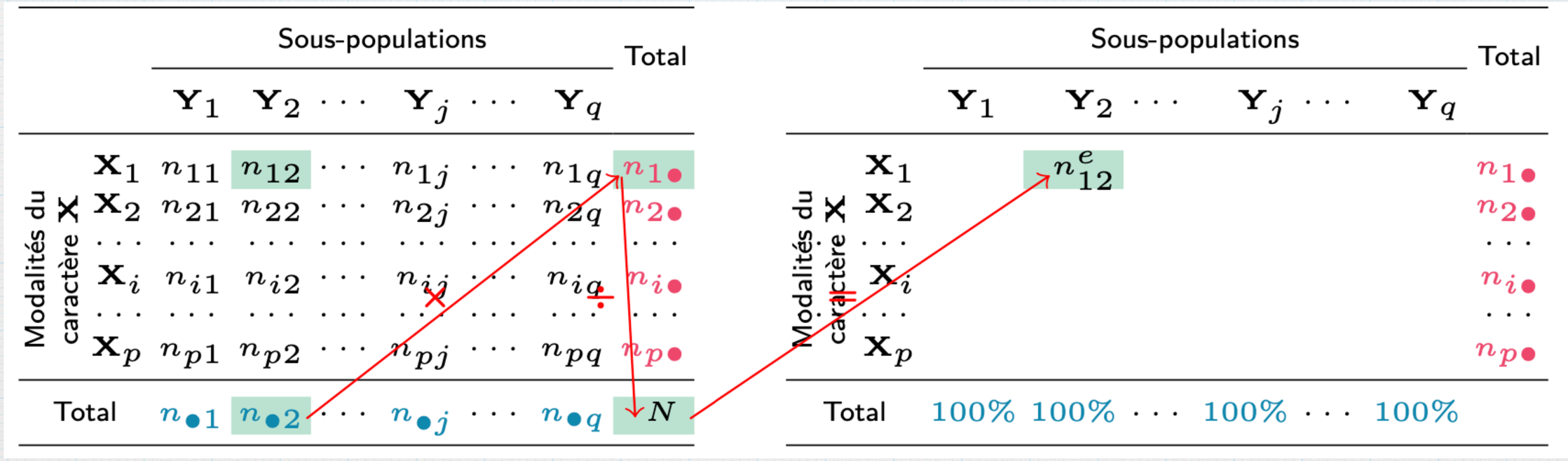


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

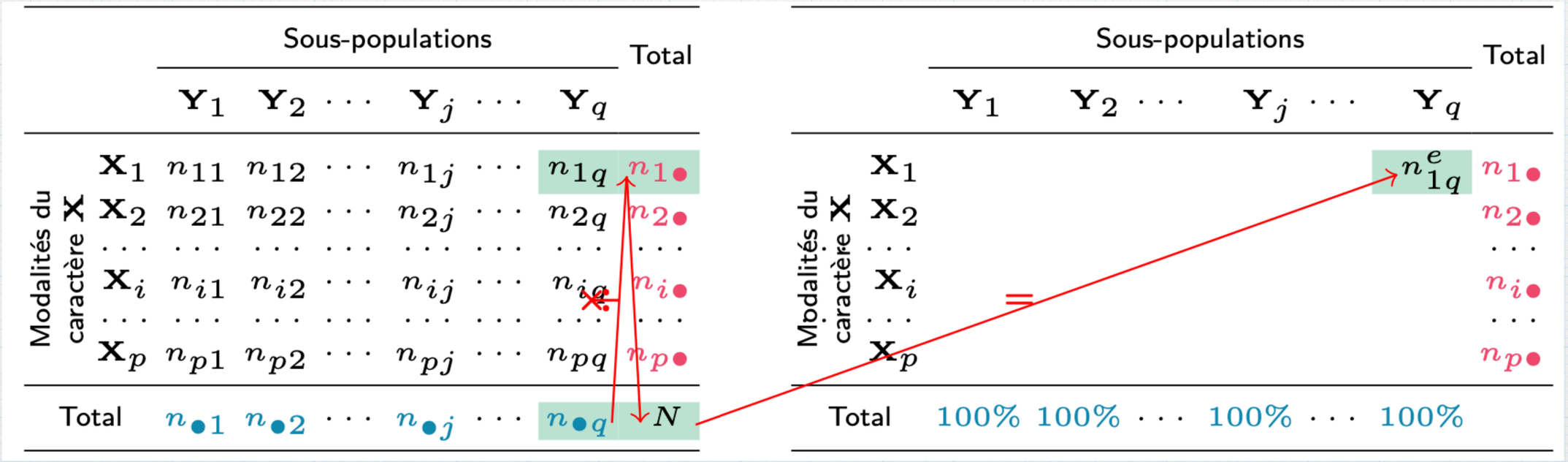


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

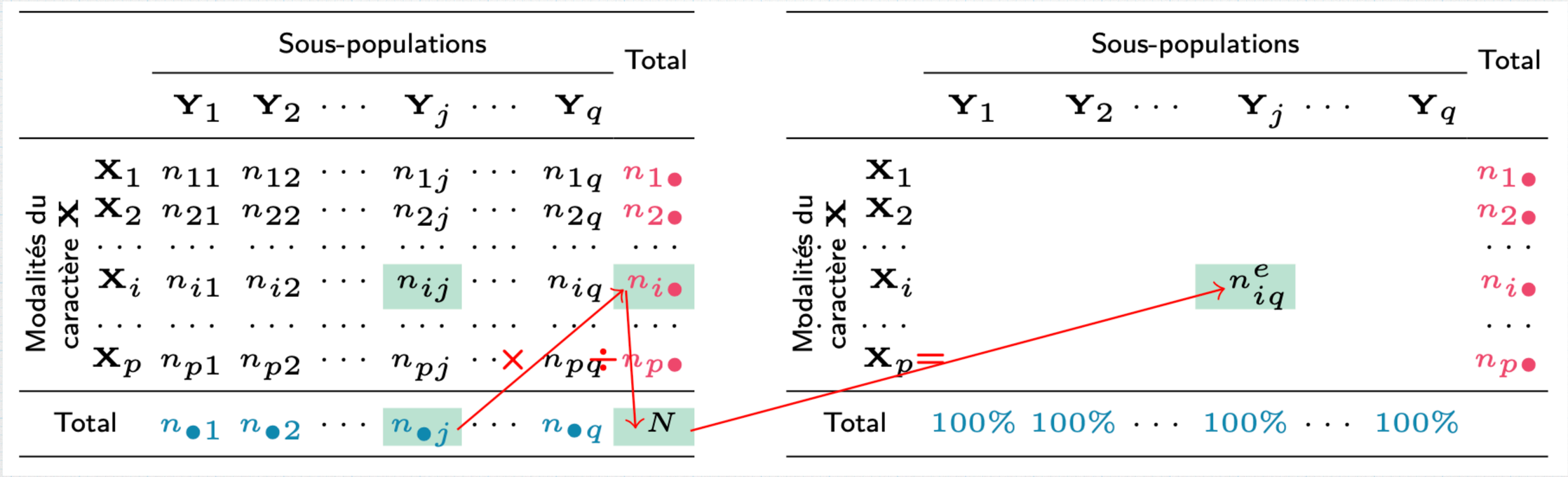


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

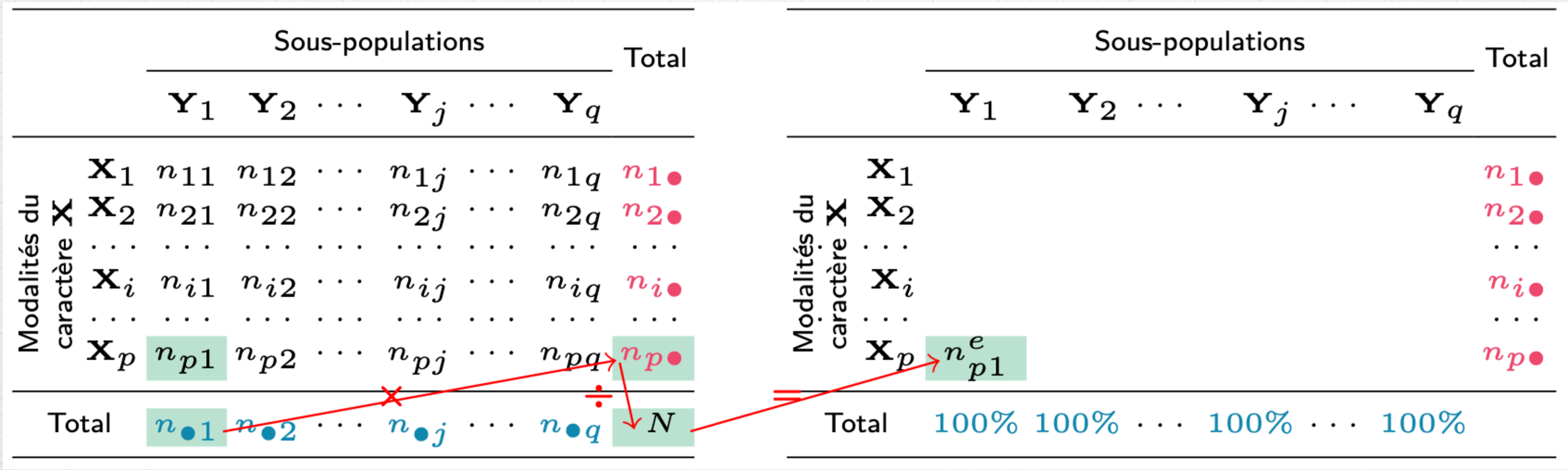


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

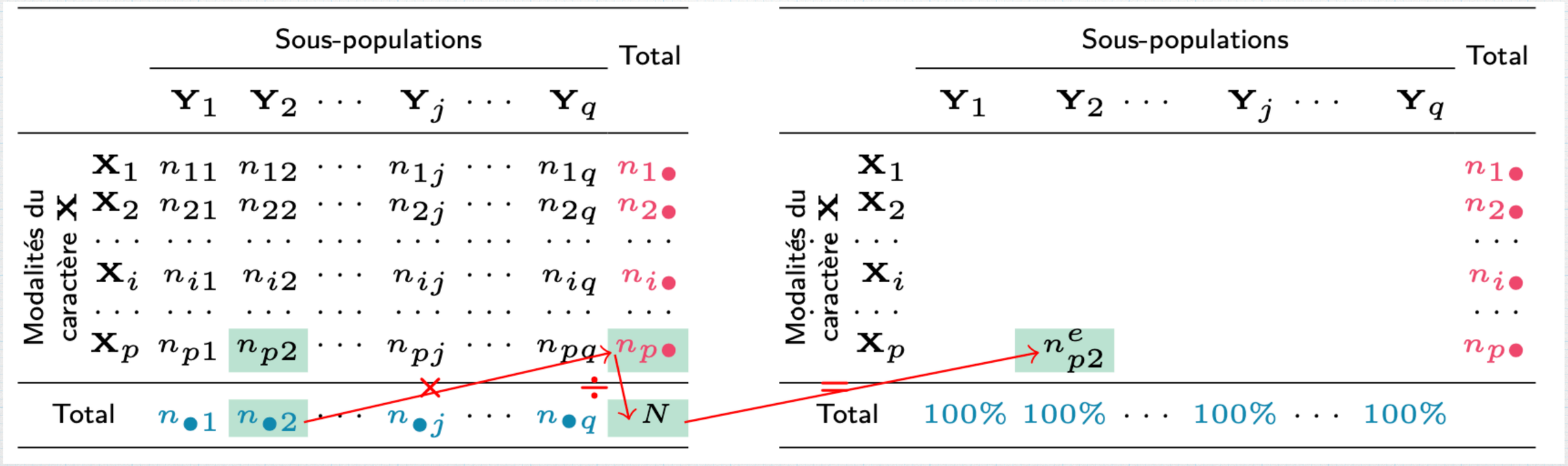


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

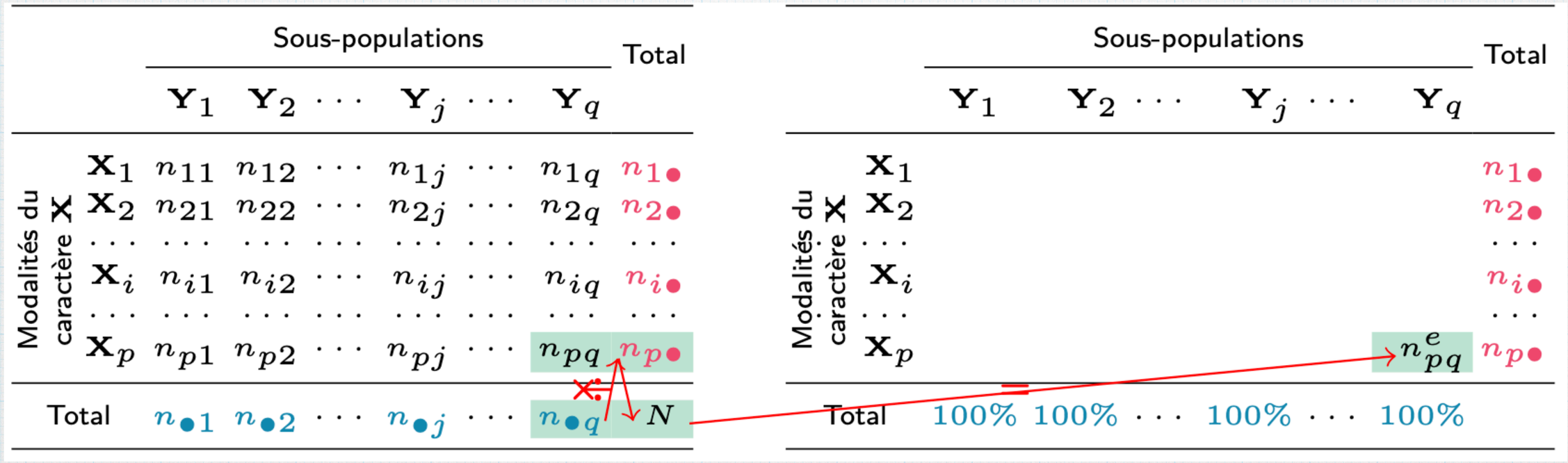


In practice

L'effectif estimé s'écrit ainsi :

The estimated workforce is written as follows:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$



3.2 Statistic test

Statistic test

La statistique du test du Khi-Deux s'écrit :

The chi-square test statistic is written:

$$\chi^2 = \sum_{j=1}^{q} \sum_{i=1}^{p} \frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e} \sim \chi^{2,\alpha}_{(p-1)(q-1)}$$

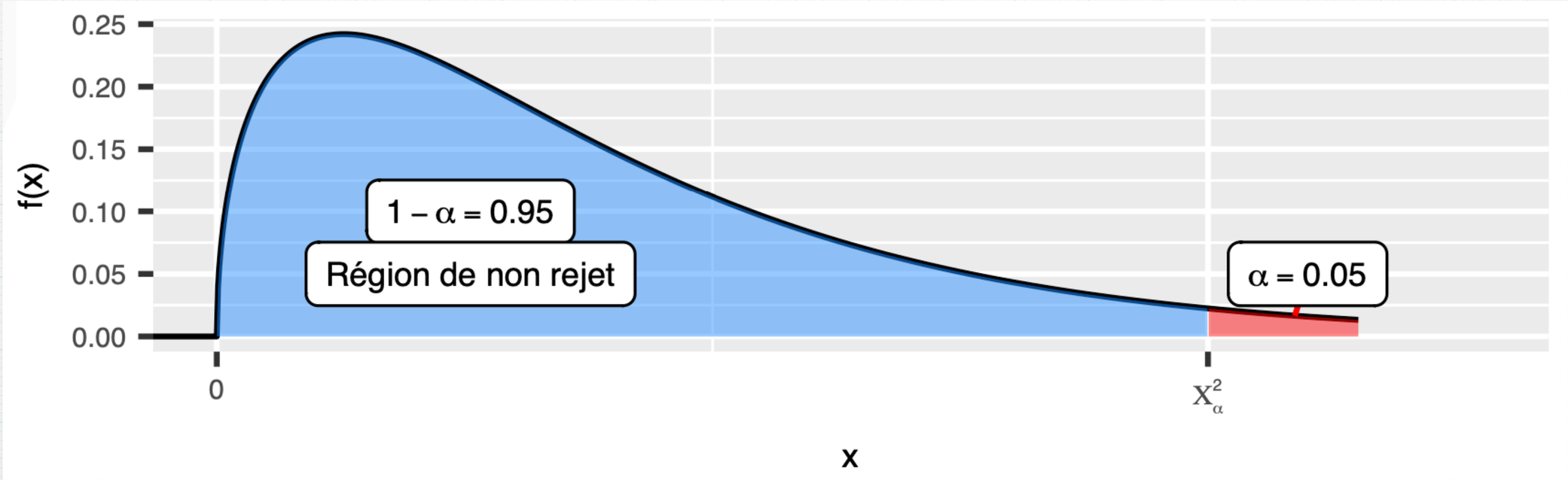
Le quantile d'ordre $\alpha, \chi^{2,\alpha}_{(p-1)(q-1)}$ est la valeur valeur lue dans dans la table du Khi-Deux à $(p-1)(q-1)$ degrés de liberté, au seuil de significativité α.

The quantile of order $\alpha, \chi^{2,\alpha}_{(p-1)(q-1)}$ is the value read in the chi-square table at $(p-1)(q-1)$ degrees of freedom, at the significance level α.

The non-rejected area for H_0

La région de non-rejet de l'hypothèse nulle s'écrit :
The non-rejection region of the null hypothesis is written:

$$AR = \left[0; \chi^{2,\alpha}_{(p-1)(q-1)}\right]$$



➡Si $CR \in AR$, l'hypothèse nulle d'égalité des proportions ne peut pas être rejetée
➡Si $CR \notin AR$, l'hypothèse nulle d'égalité des proportions peut être rejetée

➡If $CR \in AR$, the null hypothesis of equality of proportions cannot be rejected
➡If $CR \notin AR$, the null hypothesis of equality of proportions can be rejected

Example

En 2010, l'Institut National de la Prévention et de l'Education à la Santé (INPES) a enquêté sur l'état de santé de la population résidant en France auprès de 25 830 individus. Les résultats ont été publiés dans le Baromètre Santé 2010, Attitudes et comportements de santé (La Documentation Française, Paris). On reproduit ci-après un tableau concernant les expériences d'ivresse déclarées par les personnes interrogées.

In 2010, the National Institute for Prevention and Health Education (INPES) surveyed the state of health of the population residing in France with 25,830 individuals. The results were published in the Health Barometer 2010, Attitudes et behaviors of health (La Documentation Française, Paris). The table (next slide) reproduce the experiences concerning the drunkenness declared by the asked persons

Example

	Employed workers	Students	Unemployed	Retirees	other inactive	Total
Had at least one experience of drunkenness in 2010	2607	962	529	75	131	4304
Had no experience of drunkenness in 2010	12728	2588	2064	2131	2015	21526
Total	15335	3550	2593	2206	2146	25830

Sur la base des informations recueillies, peut-on dire, au seuil de signification de $\alpha = 5\,\%$, si les problèmes d'alcool considérés sont liés au statut professionnel des individus ? Préciser les différentes étapes de la démarche.

On the basis of the information collected, can we say, at the significance level of $\alpha = 5\,\%$, whether the alcohol problems considered are linked to the professional status of individuals? Specify the different stages of the process.

Example

Les problèmes d'alcool sont liés au statut socioprofessionnel des individus s'il apparaît des différences dans la prévalence des expériences d'ivresse entre les différentes catégories sociales retenues.

Alcohol problems are linked to the socio-professional status of individuals if there are differences in the prevalence of experiences of drunkenness between the different social categories selected.

Si l'on note p_{ij} la proportion d'individus au statut socioprofessionnel j, $(j = 1,2,3,4,5)$ ayant vécu la situation i, $(i = 1,2)$ en 2010, il s'agit de tester :

If we denote by p_{ij} the proportion of individuals with socio-professional status j $(j = 1,2,3,4,5)$ having experienced situation i, $(i = 1,2)$ in 2010, it is a question of testing:

$$\begin{cases} H_0 : p_{i1} = p_{i2} = p_{i3} = p_{i4} = p_{i5}, \forall i = 1,2 \\ H_1 : \exists i \neq j \text{ tel que } p_{ij} \neq p_{ik} \text{ for } i = 1,2 \end{cases}$$

Example

La région d'acceptation de l'hypothèse nulle, au seuil de signification de 5%, s'écrit :

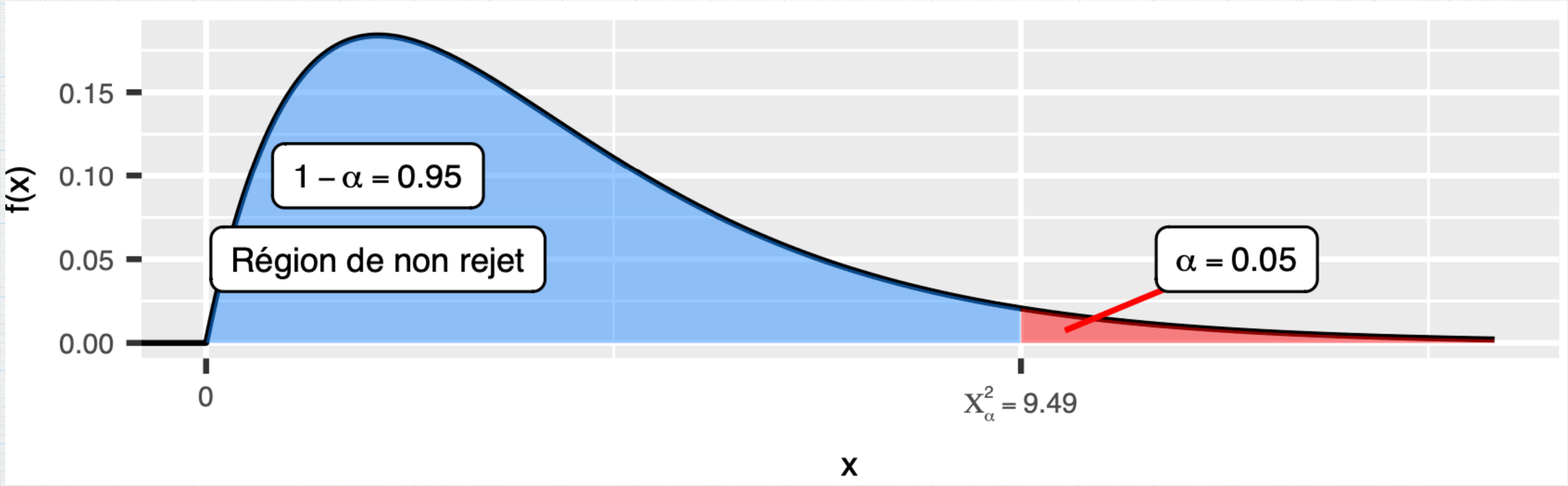
The acceptance region of the null hypothesis, at the 5% significance level, is written:

$$AR = \left[0; \chi^{2,\alpha}_{(p-1)(q-1)}\right]$$

où p et q désignent respectivement le nombre de lignes et le nombre de colonnes du tableau reproduisant les données, c'est-à-dire $p = 2$ et $q = 5$.

where p and q denote respectively the number of rows and the number of columns of the table reproducing the data, that is to say $p = 2$ and $q = 5$.

Example



Le quantile d'ordre $\alpha = 5\,\%$, $\chi_4^{2,\alpha}$ est la valeur lue dans dans la table du Khi-Deux à 4 degrés de liberté, au seuil de significativité α, soit : $\chi_4^{2,\alpha} = 9{,}49$

The quantile of order $\alpha = 5\,\%$ is the value read in table 4 of the chi-square with 4 degrees of freedom, at the significance level α, i.e $\chi_4^{2,\alpha} = 9{,}49$

La région de non-rejet de l'hypothèse nulle s'écrit : $AR = [0; 9{,}49]$

The non-rejection region of the null hypothesis is written:

Example

Pour calculer le rapport critique, on doit tout d'abord calculer la distribution théorique que l'on observerait si le phénomène d'alcoolisation était indépendant du statut socioprofessionnel, autrement dit si la proportion de personnes ayant connu au moins une expérience d'ivresse en 2010 était la même dans chaque catégorie socioprofessionnelle.

To calculate the critical ratio, we must first calculate the theoretical distribution that we would observe if the phenomenon of alcoholism was independent of socio-professional status, in other words if the proportion of people having had at least one experience of alcoholism. drunkenness in 2010 was the same in each socio-professional category.

L'effectif théorique de la cellule de la ligne i et de la colonne j s'obtient simplement en effectuant :

The theoretical size of the cell in row i and column j is obtained simply by performing:

$$n_{ij}^{e} = f_{i\bullet} \times n_{\bullet j} = \frac{n_{i\bullet}}{N} \times n_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

Example

	Employed workers	Students	Unemployed	Retirees	other inactive	Total
Had at least one experience of drunkenness in 2010	2607	962	529	75	131	4304
Had no experience of drunkenness in 2010	12728	2588	2064	2131	2015	21526
Total	15335	3550	2593	2206	2146	25830

Example

On forme le tableau de comparaison des effectifs observés et des effectifs théoriques :

We form the comparison table of observed and theoretical effective:

	n_{ij}	n_{ij}^e	$n_{ij} - n_{ij}^e$	$(n_{ij} - n_{ij}^e)^2$	$\frac{(n_{ij}-n_{ij}^e)^2}{n_{ij}^e}$
$i = 1, j = 1$	2607	$2555,2$	$51,8$	$2679,13$	$1,05$
$i = 2, j = 1$	12728	$12779,8$	$-51,8$	$2679,13$	$0,21$
$i = 1, j = 2$	962	$591,5$	$370,5$	$137248,58$	$232,02$
$i = 2, j = 2$	2588	$2958,5$	$-370,5$	$137248,59$	$46,39$
$i = 1, j = 3$	529	$432,1$	$96,9$	$9396,15$	$21,75$
$i = 2, j = 3$	2064	$2160,9$	$-96,9$	$9396,15$	$4,35$
$i = 1, j = 4$	75	$367,6$	$-292,6$	$85603,79$	$232,88$
$i = 2, j = 4$	2131	$1838,4$	$292,6$	$85603,79$	$46,56$
$i = 1, j = 5$	131	$357,6$	$-226,6$	$51340,12$	$143,58$
$i = 2, j = 5$	2015	$1788,4$	$226,6$	$51340,12$	$28,71$
					$\sum = \mathbf{757,50}$

Example

La valeur observée du rapport critique vaut donc $\chi^2 = 757{,}5 \notin AR$.

On rejette au seuil de significativité de 5% l'hypothèse nulle d'indépendance de l'expérience d'ivresse vis-à-vis du statut socioprofessionnel.

Les expériences d'alcoolisation excessive semblent tout au contraire liées au statut d'occupation des individus.

The observed value of the critical ratio is therefore worth $\chi^2 = 757{,}5 \notin AR$.

At the significance level of 5%, we reject the null hypothesis of the independence of the experience of drunkenness vis-à-vis socio-professional status.

On the contrary, experiences of excessive drinking seem to be linked to the employment status of individuals.

Example

L’examen du tableau initial laisse en effet apparaître que la prévalence des expériences d’ivresse est de 17% sur la population, mais que cette prévalence varie de 3% chez les retraités à 20% chez les chômeurs et 27% chez les étudiants qui sont les deux catégories nettement au-dessus du reste de la population.

C’est certainement la raison pour laquelle les campagnes pour l’éducation à la santé ciblent tout particulièrement cette dernière catégorie.

Examination of the initial table indeed shows that the prevalence of drunkenness experiences is 17% in the population, but that this prevalence varies from 3% among retirees to 20% among the unemployed and 27% among students who are are the two categories clearly above the rest of the population.

C’est certainement la raison pour laquelle les campagnes pour l’éducation à la santé ciblent tout particulièrement cette dernière catégorie.