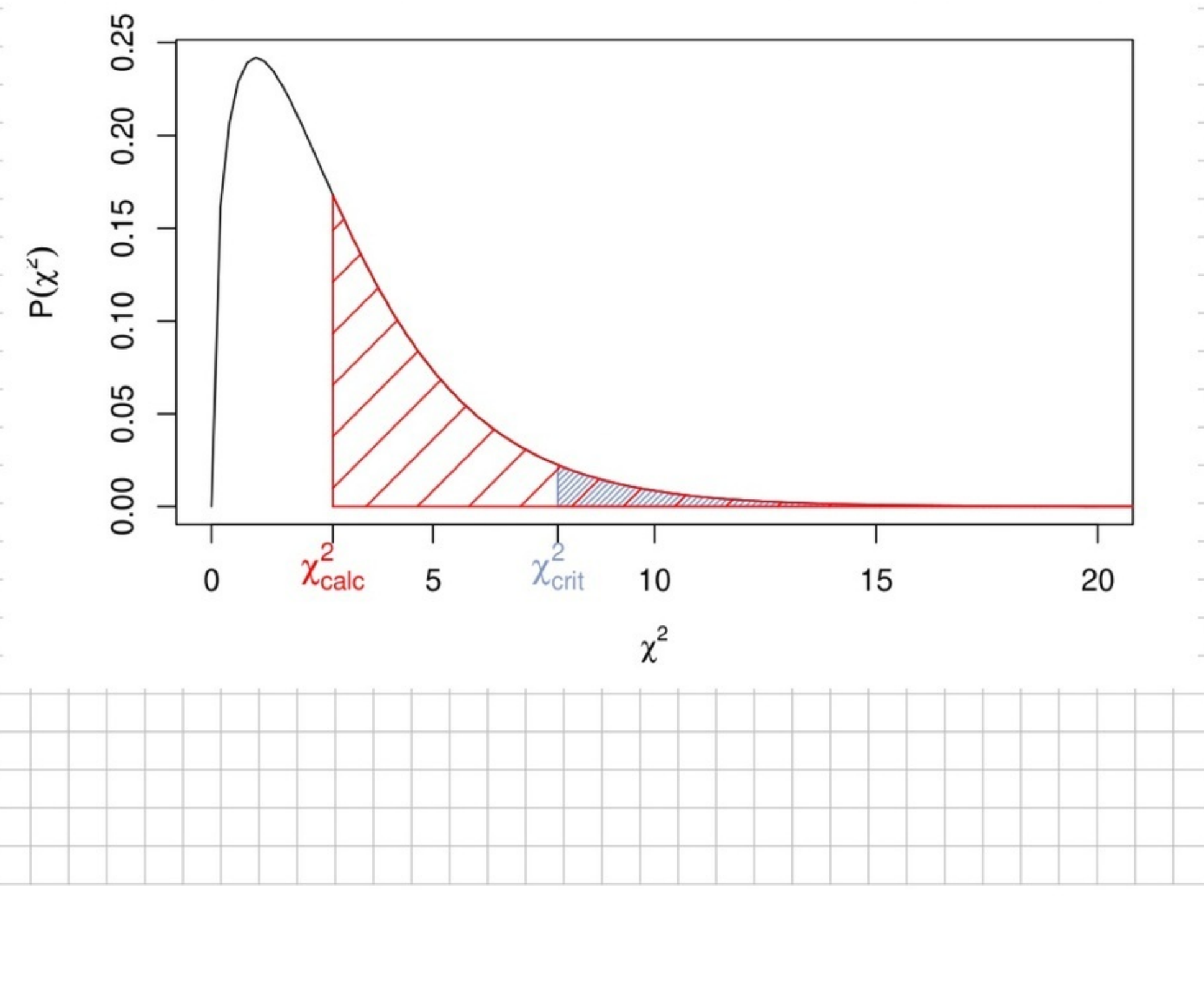
PrivateTeacher

Cours Privés de Science

Chi-Carré : χ^2
Test d'indépendance



Sommaire

Introduction

Le Chi-carré (Chi2) est un nombre qui mesure la dépendance entre des observations faites sur des individus et les groupes auxquels ils appartiennent.

On cherche à savoir si les observations faites au sein d'un groupe sont spécifiques à ce groupe ou si au contraire elles sont partagées par l'ensemble de la population.

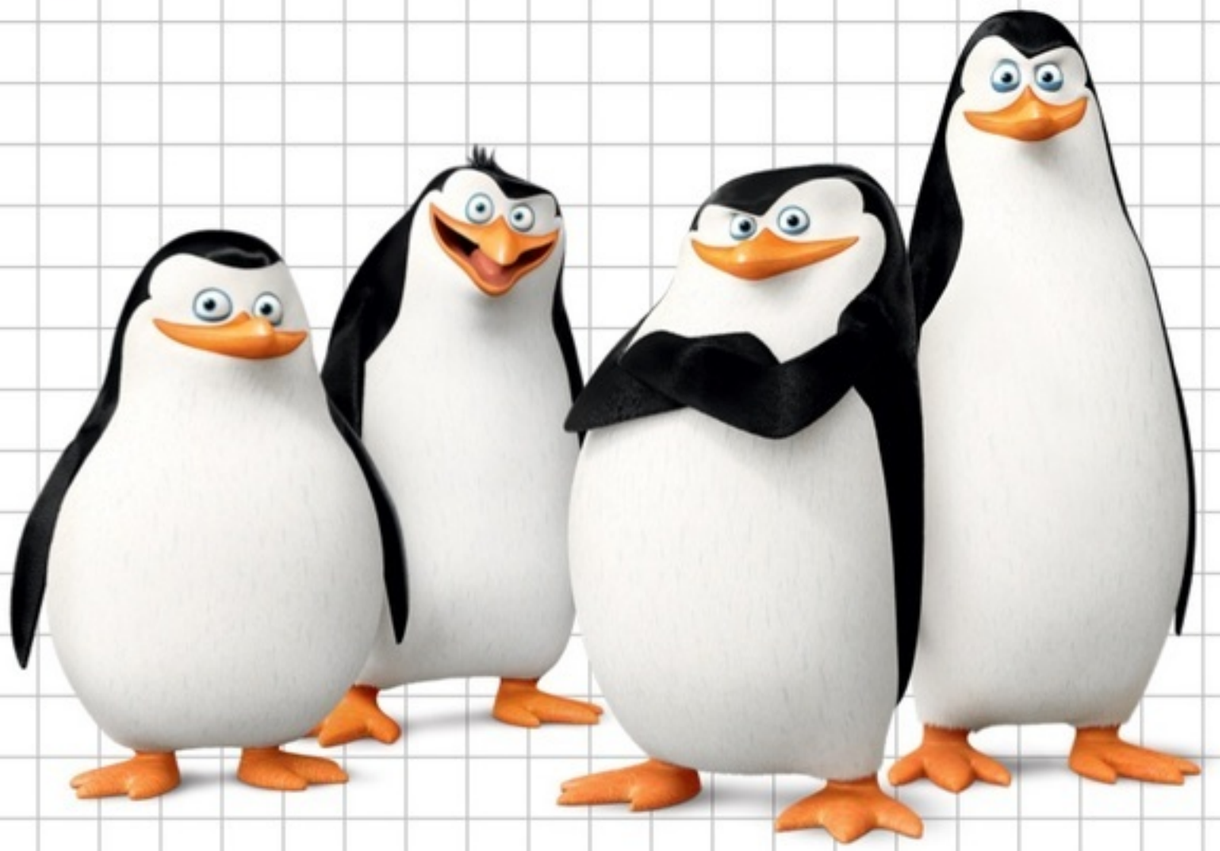
Si les observations sont spécifiques à un groupe en particulier, on peut s'en servir pour identifier ce groupe.

Dans ce cas, les observations possèdent une valeur explicative vis-à-vis du groupe. La variable "groupe" (une variable catégorielle) sera alors dépendante des observations.

Le Chi2 est une statistique qui quantifie ce degrés de dépendance. Pour cette raison, elle permet de conduire un test d'indépendance

Le cas des pingouins

Imaginons que l'on souhaite étudier une population de pingouins dans l'hémisphère Sud. Deux groupes nous intéressent en particulier : les pingouins tendres et bons à manger, et les pingouins coriaces, difficiles à mâcher.



Afin de préserver au mieux la population de pingouins, on cherche à déterminer <u>à quel groupe</u> ils appartiennent <u>avant</u> de les passer sur le grille.

Dans ce but, on choisit 4 caractéristiques qui devraient nous permettre de déterminer <u>à l'avance</u> si un pingouin est tendre ou coriace.

Prendre des mesures

Voici les 4 caractéristiques que l'on décide d'observer sur les pingouins dans le but de déterminer le groupe auquel ils appartiennent :

Observation 1 : taille
Observation 2 : poids
Observation 3 : age
Observation 4 : quantité de poisson mangée par semaine

Les groupes nous l'avons dit sont :

Groupe A : Pingouins coriaces
Groupe B : Pingouins tendres

L'étape suivante consiste à prendre des mesures au sein de la population. On se déguise donc en pingouin pour passer inaperçu et on effectue nos mesures.

Afin d'en faciliter la lecture, on organise ensuite nos résultats sous forme de table qu'on appel tableau croisé ou encore table de contingence.

Table de contingence

Une table de contingence se présente de la manière suivante : sur la première ligne est indiquée le nom de la variable.

Groupe	Taille	Poids	Age	Poisson
Coriace	4	7	1	2
Tendre	5	8	9	3

La première variable est la variable "groupe". Il s'agit d'une variable catégorielle à deux modalités : coriace et tendre.

Les 4 variables suivantes sont des variables numériques dont la valeur est inscrite dans le gpe correspondant.

Chaque valeur représente une moyenne sur plusieurs pingouins. La taille moyenne des pingouins coriaces p. exple, vaut 4

Une mesure d'indépendance

On a imaginé une mesure de l'écart entre les valeurs observée "O" et les valeurs espérée "E"

Cette valeur se nomme Chi-carré : χ^2

$$\chi^2 = \sum_i \chi_i^2$$

i : indice de chaque observation dans chaque groupe (ici, $i = 1, 2, 3, 4, 5, 6, 7, 8$)

χ_i^2 : distance entre la valeur observée et la valeur espérée.

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

O : valeurs observée

E : valeurs espérée

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Calcul du Chi2

0) Le point de départ pour le calcul manuel de la valeur du χ^2 est notre table de contingence.

	Obs 1	Obs 2	Obs 3	Obs 4
Gpe A	4	7	1	2
Gpe B	5	8	9	3

1) On commence par calculer la somme des valeurs, horizontallement et verticallement.

	Obs 1	Obs 2	Obs 3	Obs 4	
Gpe A	4	7	1	2	14
Gpe B	5	8	9	3	25
	9	15	10	5	39

2/ On peut ensuite calculer la valeur espérée de chaque catégorie

	Esp 1	Esp 2	Esp 3	Esp 4	
Gre A	$\frac{9 \cdot 14}{39}$	$\frac{15 \cdot 14}{39}$	$\frac{10 \cdot 14}{39}$	$\frac{5 \cdot 14}{39}$	14
Gre B	$\frac{9 \cdot 25}{39}$	$\frac{15 \cdot 25}{39}$	$\frac{10 \cdot 25}{39}$	$\frac{5 \cdot 25}{39}$	25
	9	15	10	5	39

↓

	Esp 1	Esp 2	Esp 3	Esp 4
Gre A	3.23	5.38	3.58	1.79
Gre B	5.76	9.61	6.41	3.20

3) On calcul enfin les distances χ_i^2 qui séparent les valeurs observées O_i des valeurs espérées E_i

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

	$\chi^2_{1,2}$	$\chi^2_{3,4}$	$\chi^2_{5,6}$	$\chi^2_{7,8}$
Gre A	$\frac{(4-3.23)^2}{3.23}$	$\frac{(7-5.38)^2}{5.38}$	$\frac{(1-3.58)^2}{3.58}$	$\frac{(2-1.79)^2}{1.79}$
Gre B	$\frac{(5-5.76)^2}{5.76}$	$\frac{(8-9.61)^2}{9.61}$	$\frac{(9-6.41)^2}{6.41}$	$\frac{(3-3.20)^2}{3.20}$

	$\chi^2_{1,2}$	$\chi^2_{3,4}$	$\chi^2_{5,6}$	$\chi^2_{7,8}$
Gre A	0.183	0.487	1.859	0.024
Gre B	0.100	0.269	1.046	0.012

On additionne enfin chacune de ces valeurs pour obtenir la valeur finale :

$$\chi^2 = \sum_i \chi_i^2 = 3.984$$

Détail du calcul

$$\chi^2 = \sum_i \chi_i^2$$

$$= \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(4-3.23)^2}{3.23} + \frac{(7-5.38)^2}{5.38} +$$

$$\frac{(1-3.58)^2}{3.58} + \frac{(2-1.79)^2}{1.79} +$$

$$\frac{(5-5.76)^2}{5.76} + \frac{(8-9.61)^2}{9.61} +$$

$$\frac{(9-6.41)^2}{6.41} + \frac{(3-3.25)^2}{3.25}$$

$$= 0.183 + 0.487 + 1.859 + 0.024 +$$
$$0.100 + 0.269 + 1.046 + 0.012$$

$$\chi^2 = 3.984$$

Interprétation

Cette valeur cependant, n'a pas de sens à elle seule.

Si on souhaite l'utiliser, il faut pouvoir la comparer à un point de référence.

C'est de cette manière seulement que l'on pourra dire si sa valeur est grande ou petite.

Il existe au moins deux méthodes :

1) Le V de Cramer

2) Situer la valeur du χ^2 au sein de sa distribution de probabilité.

La première méthode est la plus simple, c'est celle que nous verrons en premier.

La deuxième est plus générale et fait intervenir un test d'hypothèse.

Sa portée est beaucoup plus grande.
Il est donc utile de la bien comprendre.
Nous la verrons donc dans tous les détails.

Le V de Cramer

Le V de Cramer, noté V, est un nombre qui mesure le degrès d'association (le degrés de dépendance) entre une variable catégorielle et des variables numériques.

Sa valeur est compise entre 0 et 1

$$0 < V < 1$$

0 signifie pas de dépendance entre le groupe et les observations.
Les groupes sont similaires.

1 signifie qu'il y a une dépendance entre les groupes et les observations.
Les groupes sont différents.

Le V de Cramer se calcul en divisant le χ^2 par le nombre N d'observations ainsi que par un nombre en rapport avec le nombre de lignes " nlgn " et le nombre de colonnes " ncol " de la table de contingence.

On le calcul à l'aide de la formule :

$$V = \left(\frac{\chi^2}{N \cdot \min(nlgn - 1 ; ncol - 1)} \right)^{1/2}$$

N = nombre d'observations

$nlgn$ = nombre de lignes

$ncol$ = nombre de colonnes

$\min(nlgn - 1 ; ncol - 1)$ =

le plus petit nombre entre $nlgn - 1$ et $ncol - 1$

On le voit, le V de Cramer est un χ^2 normalisé par le nombre de mesures et réduit par une racine carrée.

Dans l'exemple qui nous occupe, les valeurs sont les suivantes :

$N = 8$

$nlgn = 2$

$nlgn - 1 = 1$

$ncol = 4$

$ncol - 1 = 3$

$\min(nlgn-1 ; ncol-1) = \min(1;3) = 1$

enfin, $\chi^2 = 3.984$

on a donc :

$$V = \left(\frac{\chi^2}{N \cdot \min(nlgn-1 ; ncol-1)} \right)^{1/2}$$

$$= \sqrt{\frac{3.984}{8 \cdot 1}} = \underline{\underline{0.7}}$$

Cette valeur s'approche de 1, les deux groupes sont donc plutôt différents.

Distribution de probabilité du Chi2

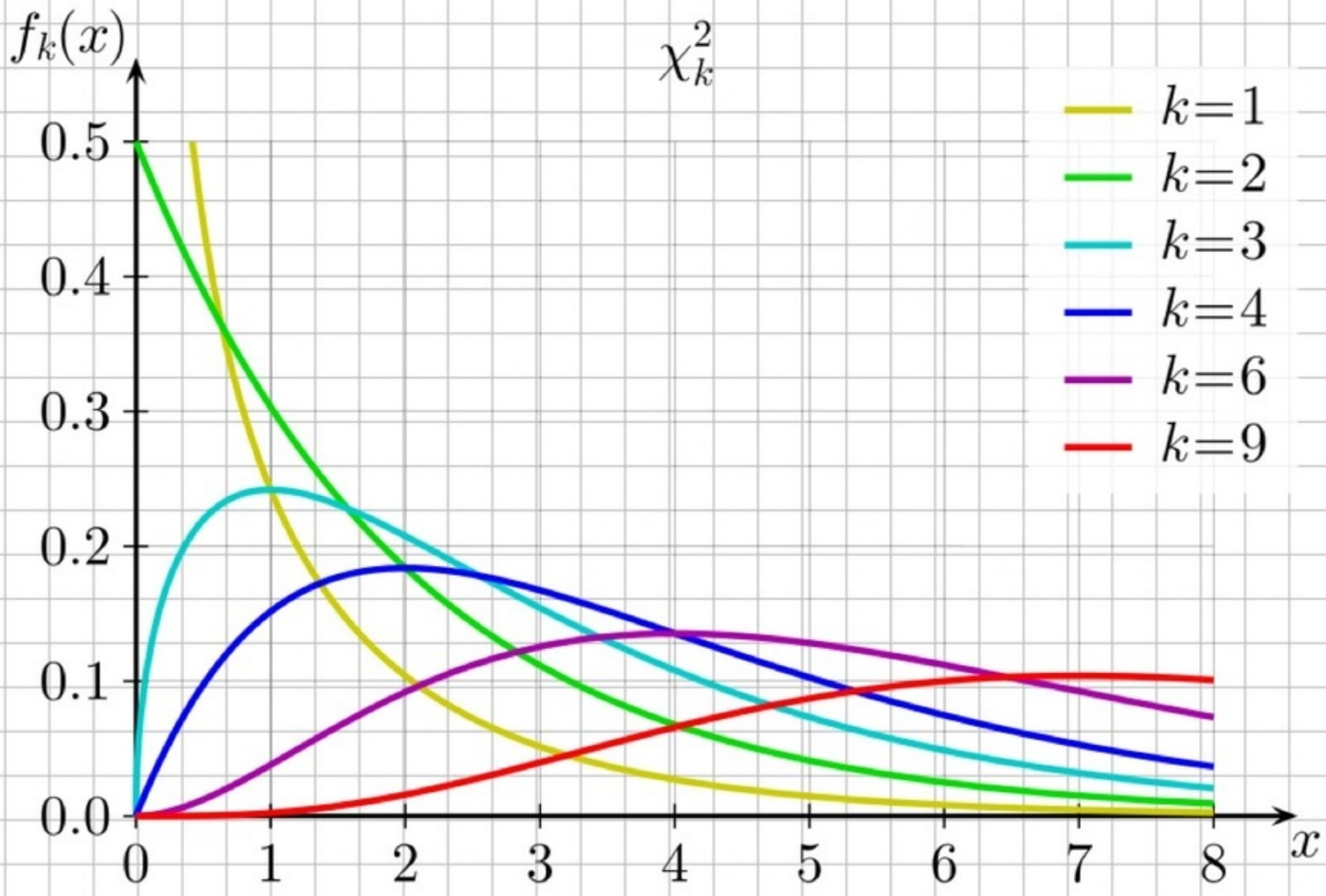
La valeur du χ^2 nous l'avons vu, mesure la somme des écarts entre les valeurs observées et les valeurs espérées.

Il s'agit maintenant de déterminer qu'est ce qu'un écart important et qu'est ce qu'un écart insignifiant

On utilise pour cela la distribution de probabilité des valeurs de χ^2.

Elle nous servira à poser un seuil en terme de probabilité au delà duquel l'écart est jugé suffisamment grand

Il existe plusieurs distributions de χ^2.
Chacune d'entre elle appartient à
la même famille de distribution,
et chacune se distingue des autres par
sa forme donnée par un paramètre k :



Graphique des différentes distributions de probabilité de χ^2 en fonction du paramètre k

Source : Wikipedia

La valeur de k représente le degrés de liberté "df" (degree of freedom) de la distribution.

Degrés de liberté

Voici à présent comment situer la valeur du χ^2 au sein de sa distribution de probabilité.

Le degrès de liberté nous l'avons vu, est un paramètre qui donne sa forme a la distribution

On commence donc par déterminer le degrès de liberté de notre distribution.

Sa valeur est égale au produit entre le nombre de ligne moins 1 et le nombre de colonne moins 1

$$df = (nlgn - 1) \cdot (ncol - 1)$$

Dans notre exemple on a

$$df = (2-1) \cdot (4-1)$$

$$= 1 \cdot 3 = 3$$

On a donc $df = 3$

Schématiquement, la distribution des valeurs de χ^2 pour $df = 3$ se présente de la façon suivante :

0 Ensemble de toutes les valeurs possible de χ^2 pour $df = 3$

$\chi^2 = 0$ signifie que la somme des écarts entre les valeurs observées et les valeurs espérées est nulle.

Autrement dit, le 0 correspond à une situation où la population est homogène càd où la variable "groupe" est indépendante des observations.

Or, c'est probablement encore vrai pour $\chi^2 = 0.5$, $\chi^2 = 1$, $\chi^2 = 1.5$.

Mais jusqu'à quelle valeur de χ^2 peut on considérer que la variable "groupe" est indépendante des observations ?

Jusqu'à quelle valeur de χ^2 peut on considérer que les groupes sont toujours identique ?

Valeur critique et seuil alpha

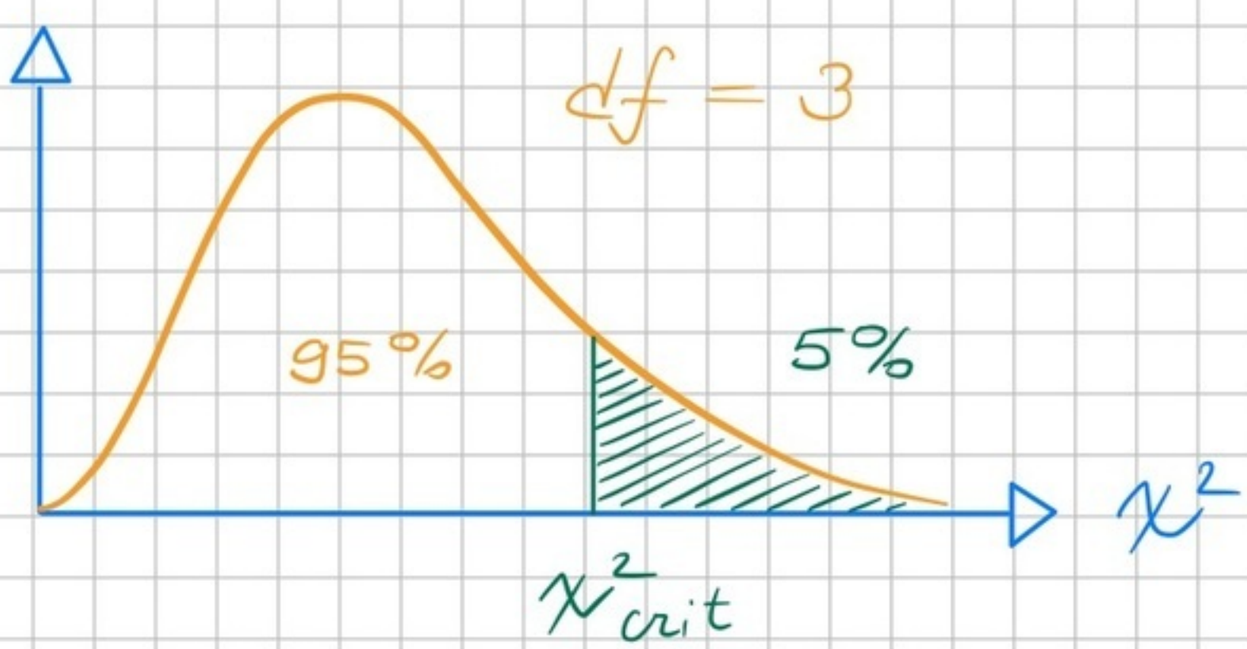
Pour répondre à cette question, on utilise la distribution de probabilité du χ^2 avec $df = 3$

Cela nous permet de définir un seuil au dela duquel les deux groupes ne peuvent plus être considéré identiques.

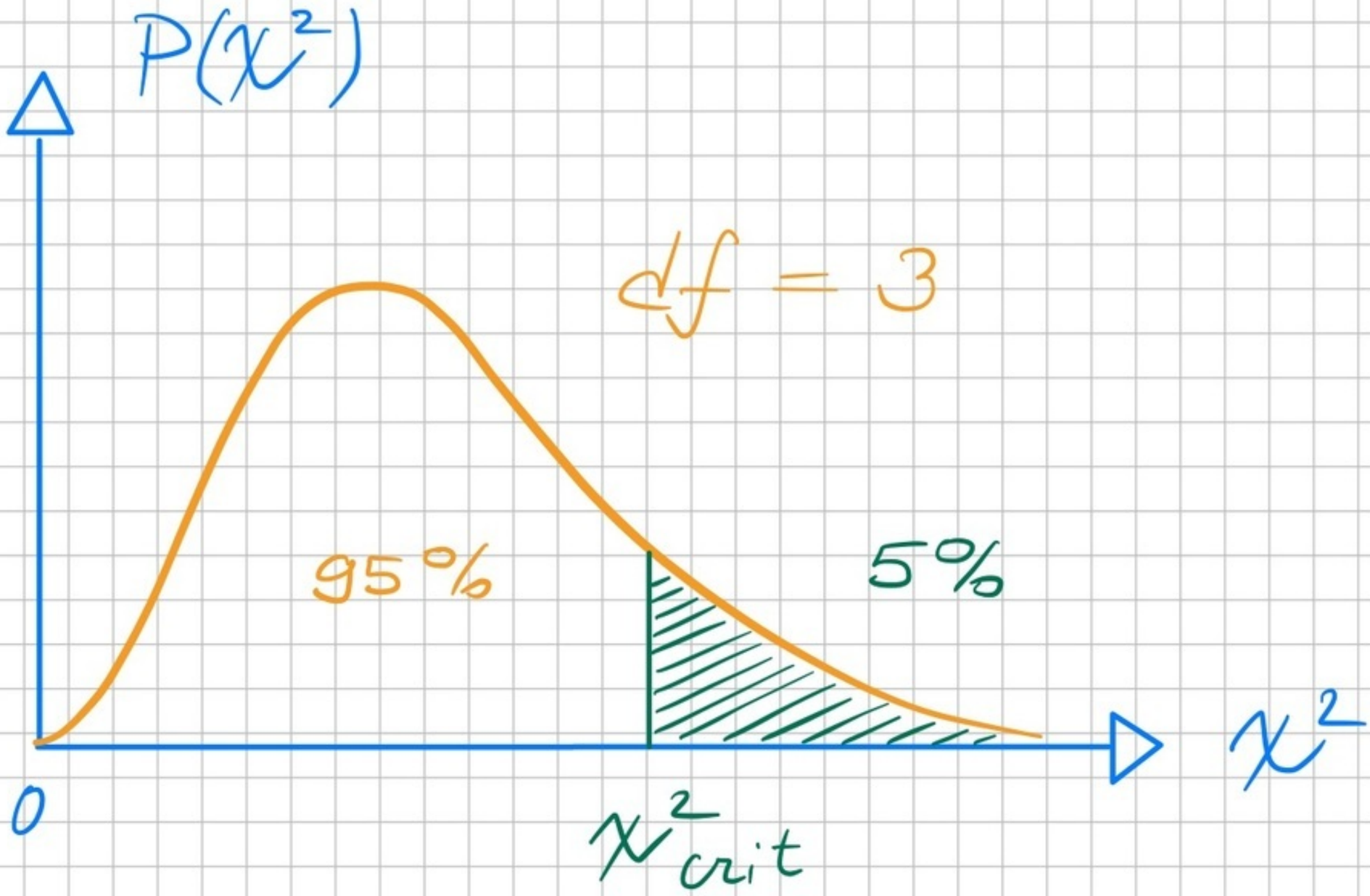
On exprime ce seuil en terme de probabilité et on le nomme <u>seuil α</u>
La plupart du temps on choisit $\alpha = 5\%$

<u>Il signifie la chose suivante :</u>

Moi, statisticien.ne, j'estime que le 5% des valeurs les plus extrêmes de la distribution représentent des valeurs pour lesquelles il est raisonnable de dire que les groupes sont différents



Représentation graphique



Cette région regroupe les valeurs de χ^2 pour lesquelles les gpes sont semblables

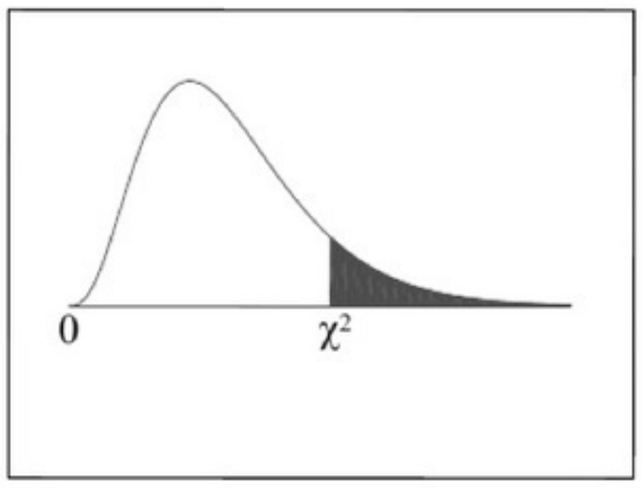
Cette région regroupe les valeurs de χ^2 pour lesquelles les gpes sont différents

La valeur critique χ^2_{crit} est donnée par le seuil α que l'on a choisi. Elle dépend de la forme de la courbe (df)

Pour la trouver, on a recours aux tables

Table de distribution de Chi2

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

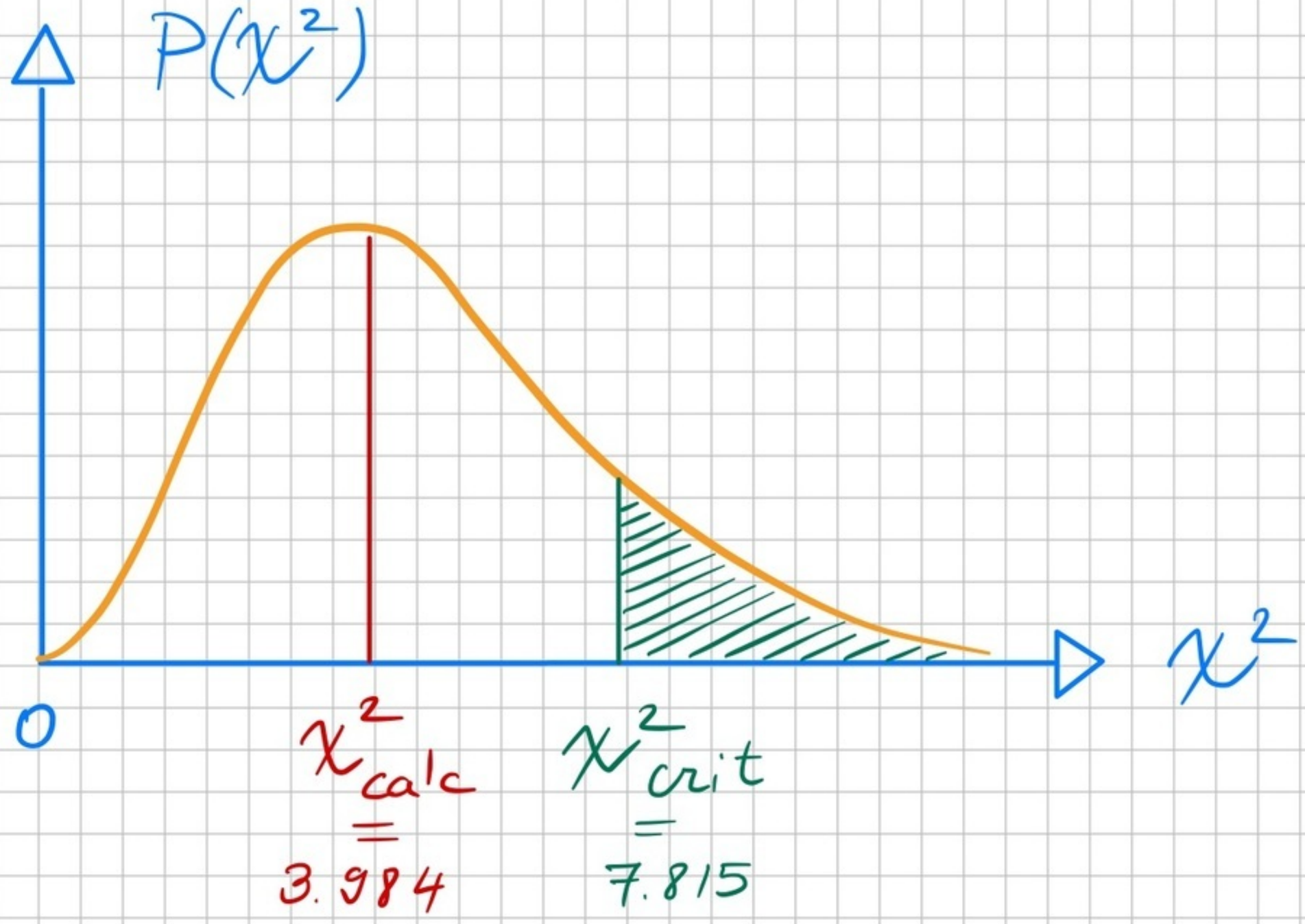
Comparer les valeurs

On le voit, la valeur critique χ^2_{crit} au delà de laquelle les groupes sont considérés différents est la suivante

$$\chi^2_{crit} = 7.815$$

On peut dès lors comparer notre propre valeur χ^2_{calc} à la valeur critique

$$\chi^2_{calc} = 3.984$$



Significativité statistique

Notre valeur χ^2_{calc} on le voit,
se trouve dans la zone des valeurs
de χ^2 pour lesquelles les groupes
sont considérés identiques.

Etant donné les outils mathématiques
auxquels on a eu recours pour
parvenir à ce résultat, on formule
notre réponse de la manière suivante :

> Les groupes doivent être considérés
> différents au seuil $\alpha = 0.05$

Il est important de réaliser que
notre conclusion dépend entièrement
du seuil choisit

Un seuil α plus grand en effet
nous donne un χ^2_{crit} plus petit.

Il est de bonne coutume cependant
de choisir un seuil α qui soit
au minimum de 5%

C'est ainsi que l'on atteint
la significativité statistique.

Remarques

Notons enfin que la conclusion à laquelle nous sommes parvenu à l'aide de la distribution de probabilité n'est pas la même que celle obtenue à l'aide du V de Cramer

Cela tient au fait qu'il n'existe pas de valeur limite de V au delà de laquelle les groupes sont considérés différents.

À la place, nous avons une échelle linéaire entre 0 et 1 qui nous permet d'évaluer le " degrès de similarité "

A nous de choisir quelle valeur signifie une différence significative.

Cette méthode comporte donc un degrès de subjectivité qui n'existe pas avec la méthode du seuil α

Pour cette raison cette dernière doit être considérée comme lui étant supérieur.

Modèle pour le calcul manuel du Chi-2

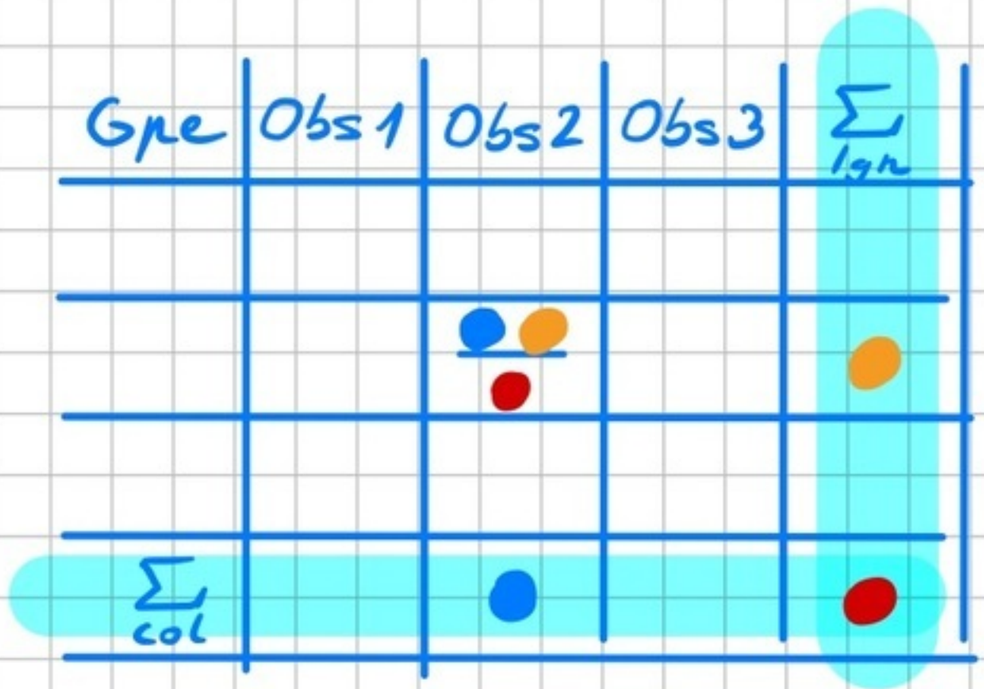
1a) Écrire le nom des différents groupes dans la colonne "Gpe"

1b) Entrer les valeurs des observations faites au sein de chaque groupe.

1c) Calculer la somme des valeurs selon chaque ligne ($\sum_{lgn}$) et selon chaque colonne ($\sum_{col}$)

Gpe	Obs 1	Obs 2	Obs 3	Obs 4	$\sum_{lgn}$
$\sum_{col}$					

2) Pour chaque case du tableau,
calculer les valeurs espérées Esp.
Multiplier la somme de la ligne ● à
la somme de la colonne ● et
diviser par le total du tableau ●



Gpe	Esp1	Esp2	Esp3	Esp4

3) Calculer les écarts χ_i^2 entre chaque valeur observée et sa valeur espérée.

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

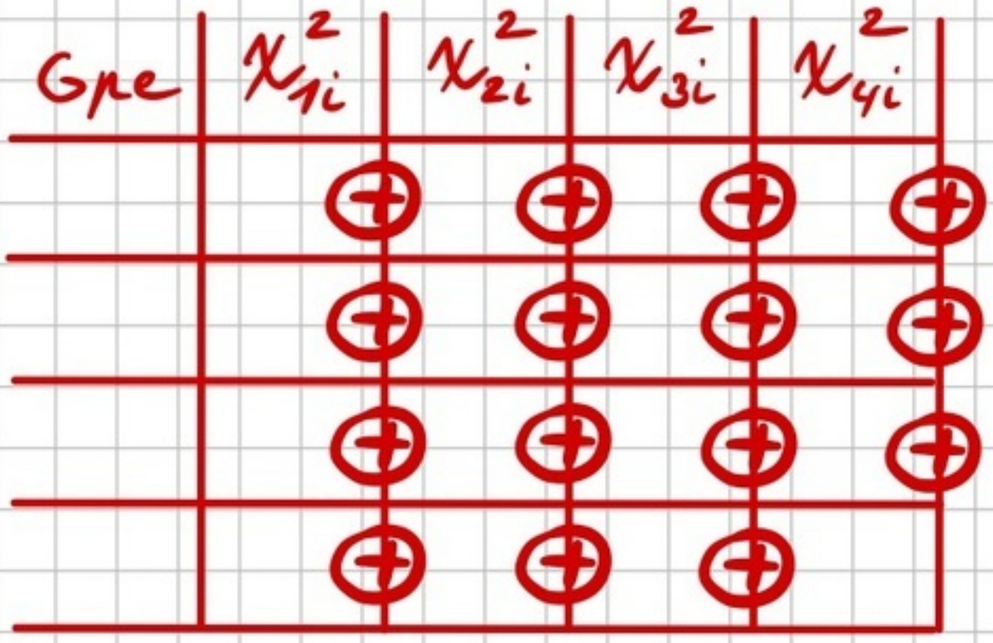
Gpe	Obs1	Obs2	Obs3	Obs4

Gpe	Esp1	Esp2	Esp3	Esp4

Gpe	χ_1^2	χ_2^2	χ_3^2	χ_4^2

4) Additionner enfin toutes les valeurs de χ_i^2 pour obtenir la valeur calculée de χ^2

$$\chi^2 = \sum_i \chi_i$$



χ^2 =