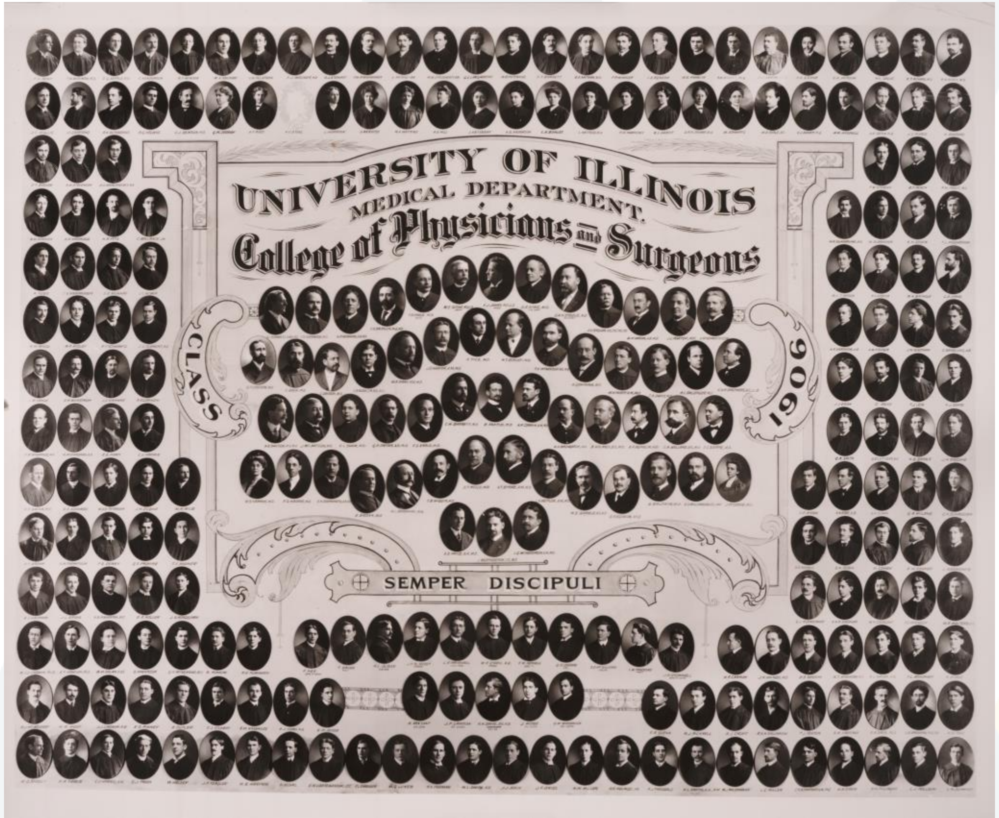
Test d'indépendance du khi carré

Julie Milot
Professeure de mathématique
Département de mathématiques
Collège de Maisonneuve
jmilot@cmaisonneuve.qc.ca

Ressource développée dans le cadre du projet Mathéma-TIC
Financé par le ministère de l'Enseignement supérieur, de la Recherche et de la Science (MESRS)
du Québec dans le cadre du Programme d'arrimage universités-collèges

Mise en situation



Existait-il un lien entre le sexe et le plus haut niveau de scolarité complété en Amérique du Nord en 1995?

Test d’indépendance du khi carré

Test permettant de déterminer s’il existe un lien entre deux variables (dont au moins une est à échelle nominale ou ordinale) dans une population, à partir des données d’un échantillon de cette population.

Étape 1 : Formuler les hypothèses

Hypothèse nulle

H_0 : Il n'existait **pas de lien** entre le sexe et le plus haut niveau de scolarité complété.

H_1 : Il existait **un lien** entre le sexe et le plus haut niveau de scolarité complété.

Hypothèse alternative

Étape 2 : Indiquer le seuil de signification du test

Seuil de signification

$\alpha = 5\ \%$

Probabilité que le test nous révèle qu’il existe un lien entre les deux variables alors que dans les faits, ce lien n’existe pas.

Tableau de contingence

Distribution de 58 adultes nord-américains selon leur sexe et le plus haut niveau de scolarité qu'ils ont complété, 1995.

Sexe	Niveau de scolarité			Total
	Secondaire	Collégial	Universitaire	
Femme	10	7	8	25
Homme	13	11	9	33
Total	23	18	17	58

Source : données fictives

$\frac{25}{58} = 0,43$

Données provenant d'un échantillon de 58 adultes nord-américains, 1995.

Identifiant du répondant	Sexe	Niveau de scolarité
1	Homme	Universitaire
2	Femme	Collégial
3	Femme	Universitaire
4	Homme	Secondaire
5	Femme	Collégial
6	Homme	Collégial
7	Homme	Secondaire
8	Homme	Universitaire
9	Femme	Secondaire
10	Femme	Collégial
11	Homme	Universitaire
12	Femme	Secondaire
13	Femme	Collégial
14	Femme	Universitaire
15	Homme	Collégial
…	…	…
58	Femme	Secondaire

Source : données fictives.

Étape 3 : Déterminer les effectifs théoriques et vérifier les conditions d'application du test

Effectif théorique se trouvant à l'intersection de la i^e ligne et de la j^e colonne

$$T_{ij} = \frac{(\text{total de la ligne } i)}{\text{grand total}} \times (\text{total de la colonne } j)$$

$$T_{12} = \frac{25}{58} \times 18 = 7{,}76$$

Distribution de 58 adultes nord-américains selon leur sexe et le plus haut niveau de scolarité qu'ils ont complété, 1995.

Sexe	Niveau de scolarité			Total
	Secondaire	Collégial	Universitaire	
Femme	10 **9,91**	7 **7,76**	8 **7,33**	25
Homme	13 **13,09**	11 **10,24**	9 **9,67**	33
Total	23	18	17	58

Source : données fictives

1ère ligne

2e colonne

1ère condition d'application : $n \geq 30$

2e condition d'application : $T_{ij} \geq 5 \quad \forall\, i, j$

Khi carré

Somme des carrés des écarts relativisés.

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - T_{ij}\right)^2}{T_{ij}}$$

Distribution de 58 adultes nord-américains selon leur sexe et
le plus haut niveau de scolarité qu'ils ont complété, 1995.

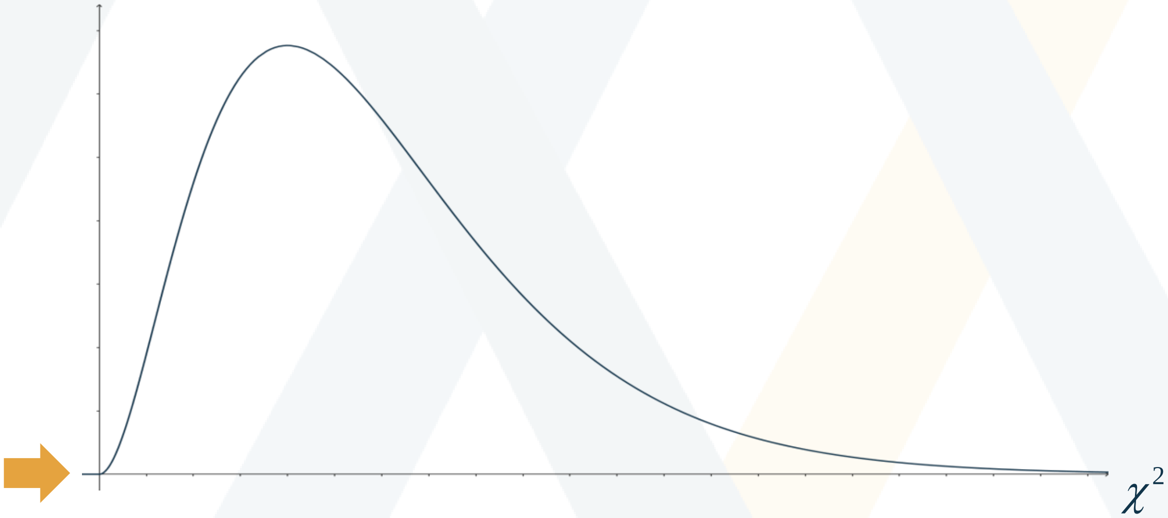
Sexe	Niveau de scolarité						Total
	Secondaire		Collégial		Universitaire		
Femme	10	**9,91**	7	**7,76**	8	**7,33**	25
Homme	13	**13,09**	11	**10,24**	9	**9,67**	33
Total	23		18		17		58

Source : données fictives

Khi carré

Somme des carrés des écarts relativisés.

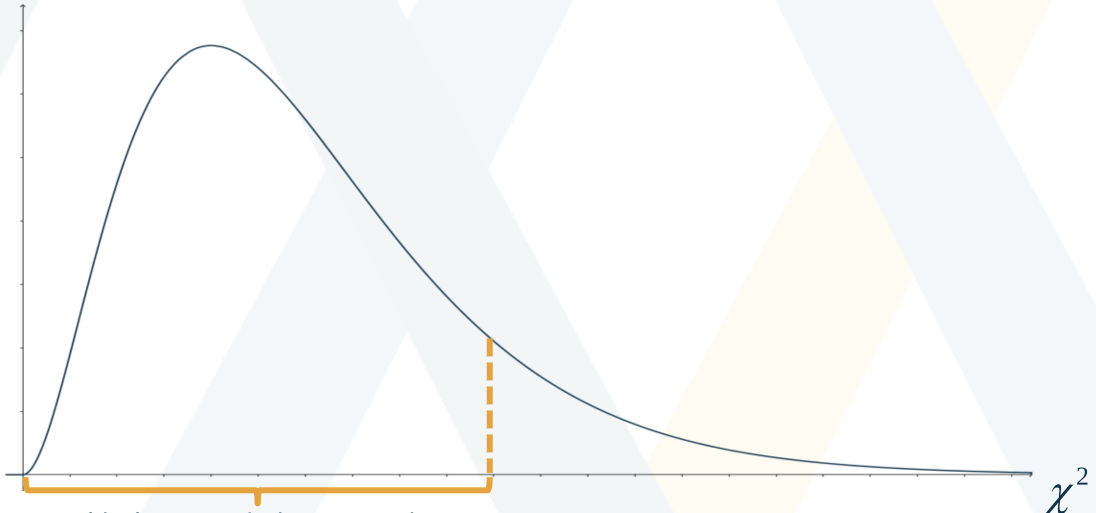
$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - T_{ij}\right)^2}{T_{ij}}$$



Khi carré

Somme des carrés des écarts relativisés.

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - T_{ij}\right)^2}{T_{ij}}$$

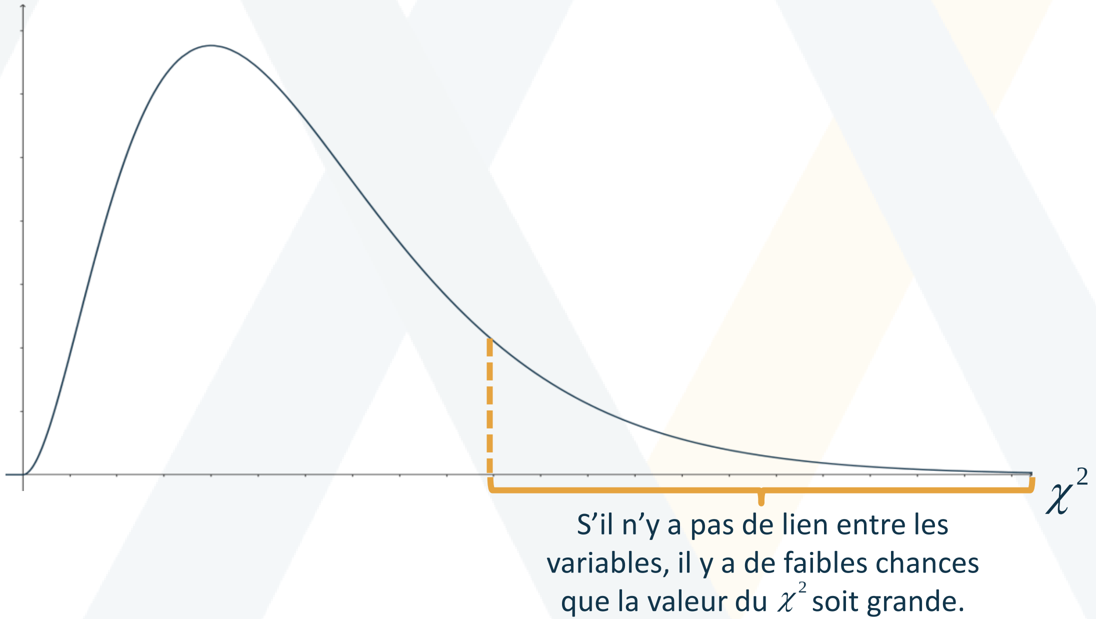


S’il n’y a pas de lien entre les variables, il y a de fortes chances que la valeur du χ^2 soit faible.

Khi carré

Somme des carrés des écarts relativisés.

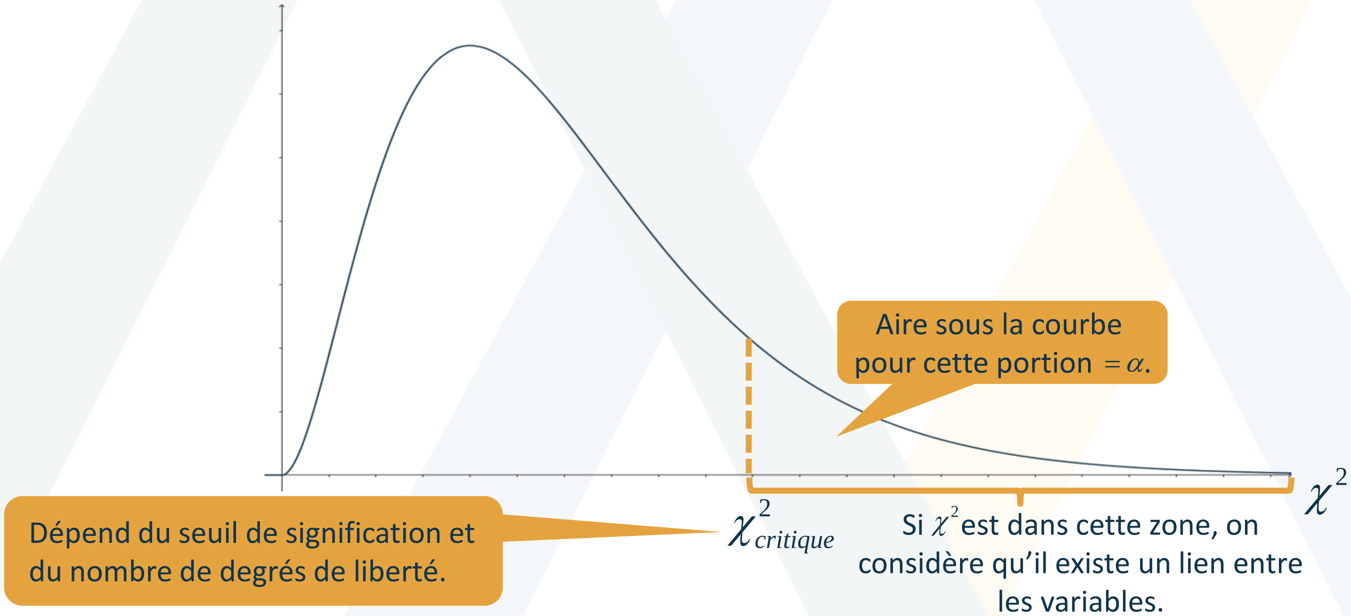
$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - T_{ij}\right)^2}{T_{ij}}$$



Khi carré

Somme des carrés des écarts relativisés.

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - T_{ij}\right)^2}{T_{ij}}$$



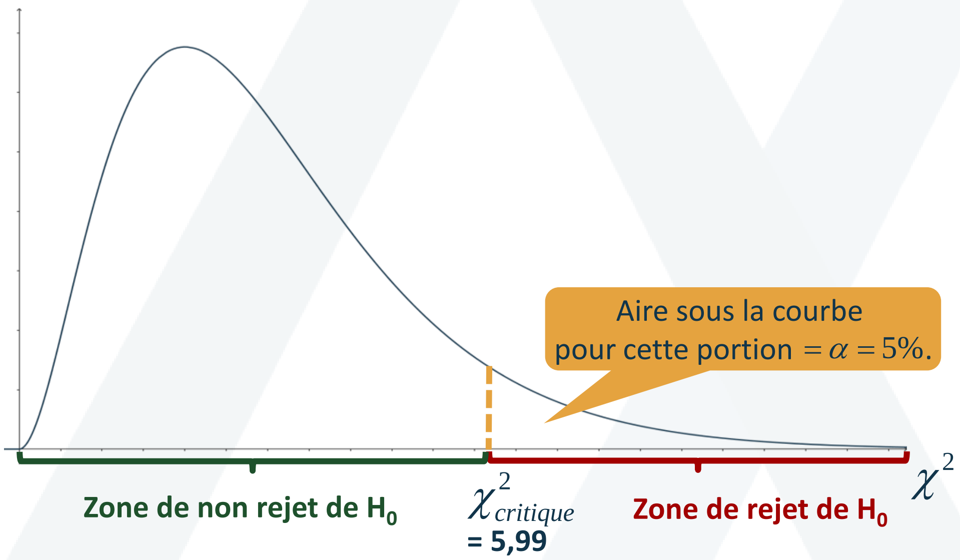
Étape 4 : Déterminer le nombre de degrés de liberté

Nombre de degrés de liberté

ν = (2 – 1) x (3 – 1) = 2

(nombre de lignes -1) x (nombre de colonnes -1)

Étape 5 : Déterminer $\chi^2_{critique}$ et définir la règle de décision



ν \ α	10%	5%	2,5%	1%	0,5%
1	2,71	3,84	5,02	6,54	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,82	9,35	11,35	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,65	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,54	20,09	21,96
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19
11	17,28	19,68	21,92	24,73	26,76
12	18,55	21,03	23,34	26,22	28,30
13	19,81	22,36	24,74	27,69	29,82
14	21,06	23,69	26,12	29,14	31,32
15	22,31	25,00	27,49	30,58	32,80
16	23,54	26,30	28,85	32,00	34,27
17	24,77	27,59	30,19	33,41	35,72
18	25,99	28,87	31,53	34,81	37,16
19	27,20	30,14	32,85	36,19	38,58
20	28,41	31,41	34,17	37,57	40,00
…	…	…	…	…	…

Étape 6 : Calculer χ^2

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Distribution de 58 adultes nord-américains selon leur sexe et le plus haut niveau de scolarité qu'ils ont complété, 1995.

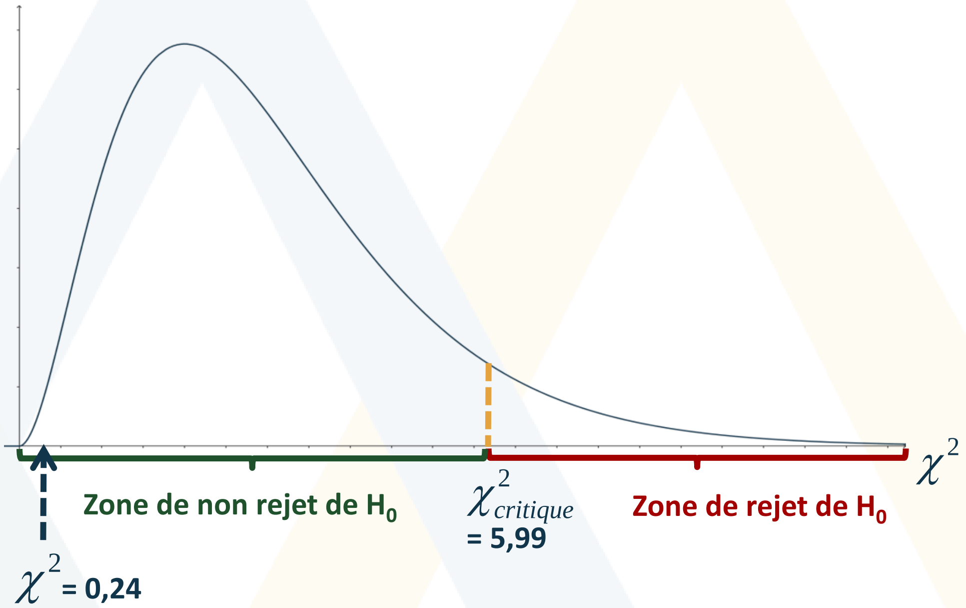
Sexe	Niveau de scolarité						Total
	Secondaire		Collégial		Universitaire		
Femme	10	9,91	7	7,76	8	7,33	25
Homme	13	13,09	11	10,24	9	9,67	33
Total	23		18		17		58

Source : données fictives

$$\chi^2 = \frac{(10-9,91)^2}{9,91} + \frac{(7-7,76)^2}{7,76} + \frac{(8-7,33)^2}{7,33} + \frac{(13-13,09)^2}{13,09} + \frac{(11-10,24)^2}{10,24} + \frac{(9-9,67)^2}{9,67}$$

$$= 0,24$$

Étape 7 : Prendre une décision et l'interpréter



Décision : On ne rejette pas H_0, car $\chi^2 = 0{,}24 \leq 5{,}99 = \chi^2_{critique}$.

Interprétation : Avec un seuil de signification de 5 %, on ne peut affirmer qu'il existait un lien entre le sexe et le plus haut niveau de scolarité complété chez les adultes nord-américains en 1995.

Coefficient de contingence et coefficient de Cramér

Nombres servant à déterminer l'intensité d'un lien statistique existant entre deux variables dont au moins une a une échelle nominale ou ordinale.

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad V = \sqrt{\frac{\chi^2}{n\left(\min\left(\text{nb. lignes}-1;\ \text{nb. colonnes}-1\right)\right)}}$$

La valeur de ces coefficients est toujours comprise entre 0 et 1.

Plus la valeur de ces coefficients est près de 1, plus le lien statistique entre les deux variables est fort.

À l'inverse, plus ces coefficients ont une valeur près de 0, plus le lien est faible.

Résumé

1. H_0 : Il n'y a **pas de lien** entre la variable A et la variable B.
H_1 : Il y a **un lien** entre la variable A et la variable B.

2. $\alpha =$ faible %

3.

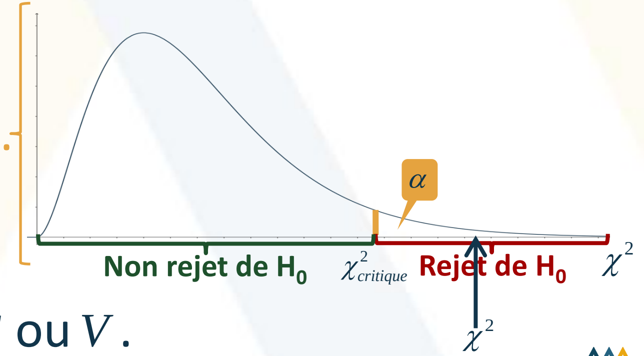
Variable A	Variable B			Total
	Modalité b_1	Modalité b_2	Modalité b_3	
Modalité a_1	O_{11} T_{11}	O_{12} T_{12}	O_{13} T_{13}	Total ligne 1
Modalité a_2	O_{21} T_{21}	O_{22} T_{22}	O_{23} T_{23}	Total ligne 2
Total	Total colonne 1	Total colonne 2	Total colonne 3	Grand total

$T_{ij} = \dfrac{\text{(total ligne } i)}{\text{grand total}} \times \text{(total colonne } j)$

1ère condition d'application : $n \geq 30$ **2e condition d'application :** $T_{ij} \geq 5 \quad \forall i, j$

4. $v =$ (nb lignes -1) x (nb colonnes -1)

5.



6. $$\chi^2 = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

7. Décider
Interpréter la décision
Déterminer la force du lien s'il y a lieu avec C ou V.

Conception du contenu

Julie Milot

Collège de Maisonneuve

jmilot@cmaisonneuve.qc.ca

Révision du contenu

Samuel Bernard et Hélène Lambert

samuel.bernard@collanaud.qc.ca

hlambert@cmaisonneuve.qc.ca

Direction de projet

Samuel Bernard

Bruno Poellhuber

Postproduction

Marie-Ève Lanthier

Musique

Sébastien Belleudy

sebe.bandcamp.com

Crédit images

Flickr, Wikipedia

flickr.com, en.wikipedia.org

Conception graphique

Christine Blais

Production des modèles en LaTeX

Nicolas Beauchemin

nicolas.beauchemin@bdeb.qc.ca

Mathéma-TIC

Mathéma-TIC

Production

Samuel Bernard

CÉGEP RÉGIONAL de LANAUDIÈRE à Terrebonne

Bruno Poellhuber

Université de Montréal

Mathéma-TIC