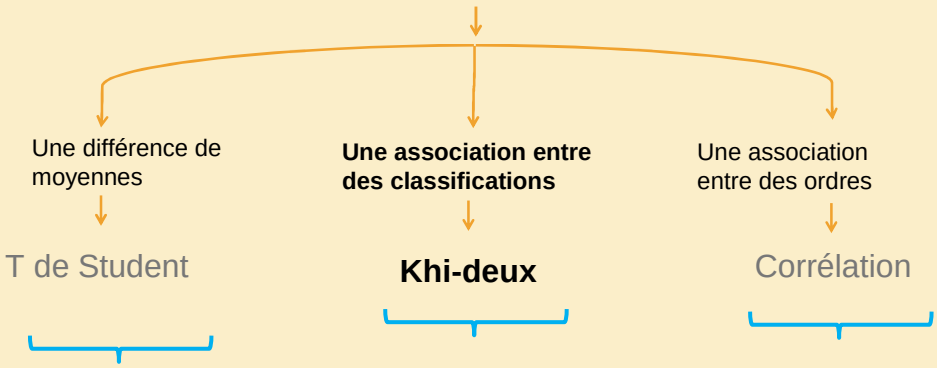
6. TD4 – Khi-deux

Les tests d'inférence statistiques permettent d'estimer le risque d'inférer un résultat d'un échantillon à une population et de décider si on « prend le risque » (si ≤ 0.05 ou 5 %)



Déjà fait : TD 2

Aujourd'hui : TD4

Déjà fait : TD3

6. TD4 – Khi-deux
6.1 Définition

- **Ici, on travaille à partir de fréquences ou d'effectifs issus de variables nominales ou catégorielles.**

Deux types de X^2 (= khi/chi deux/(au) carré).

- X^2 de conformité
  - Une seule variable
  - Permet de savoir si la distribution est conforme à une distribution qu'on connaît (par exemple, celle qu'on aurait pu obtenir par hasard).

- X^2 d'indépendance
  - Mesure le lien entre 2 variables nominales (ex: pays de naissance et couleur des yeux)
  - Permet, à partir de l'observation de l'échantillon, de décider si 2 variables nominales sont indépendantes ou non et de déterminer le risque qu'on prend de généraliser à tort de l'échantillon à la population.

6. TD4 – Khi-deux
6.1 Définition

- Principe : le khi-deux permet de décider si une série d'**effectifs *observés*** diffère significativement d'une série d'**effectifs *théoriques*** (ou ***attendus***)

- La distribution (ensemble des effectifs) théorique est celle qui correspond à l'**hypothèse nulle**.

Hypothèse nulle (H_0) : notion centrale des tests d'inférence statistique.

C'est l'hypothèse selon laquelle il n'y a **pas de lien entre les variables**.

C'est à partir de cette hypothèse qu'est calculée la probabilité *p*.

Hypothèse alternative (H_1): Hypothèse selon laquelle il y a un **lien entre les variables**

6. TD4 – Khi-deux
6.1 Définition
6.2 Khi-deux de conformité

- Enquête menée sur 110 sujets

Parmi ces 4 viennoiseries, laquelle préférez-vous?

	Croissant	Pain au chocolat	Pain aux raisins	Brioche
Effectifs observés	20	60	10	20
Effectifs théoriques/attendus	27,5	27,5	27,5	27,5

Posons H_0: Toutes les viennoiseries sont aussi populaires les unes que les autres.

Si l'hypothèse nulle est vérifiée, on devrait avoir à peu près le même nbre de personnes / catégorie, soit: **110/4** = **27,5 (Effectifs théoriques)**

⇒ Le test de khi-deux va comparer les effectifs observés et les effectifs théoriques.

⇒ Si les 2 types d'effectifs sont proches, l'hypothèse nulle est validée, s'ils sont éloignés, alors l'hypothèse nulle est rejetée.

- **1/ On soustrait les effectifs théoriques aux effectifs observés**

Effectifs observés	Effectifs théoriques	Effectifs observés - Effectifs théoriques
20	27.5	-7.5
60	27.5	32.5
10	27.5	-17.5
20	27.5	-7.5

- **2/ On élève au carré les différences**

  $(-7.5)^2 = 56.25$

  $32.5^2 = 1056.25$

  $(-17.5)^2 = 306.25$

  $(-7.5)^2 = 56.25$

- **3/ On divise ces différences au carré par la fréquence attendue (effectif théorique)**

  56.25/27.5 = 2.05

  1056.25/27.5 = 38.41

  306.25/27.5 = 11.14

  56.25/27.5 = 2.05

- **4/ On en fait la somme**

  2.05+38.41+11.14+2.05 = 53.65

  $X^2 = 53.65$

- **5/ Calcul du degré de liberté (ddl) [nombre de catégories -1]**

  Ici, 4 (viennoiseries) -1 = 3

  ddl = 3

- 6/ On se reporte à une table de Khi-Deux pour avoir la valeur de *p*

p ddl	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245

$X^2 = 53.65$; ddl = 3

Dans la ligne de ddl = 3, on cherche dans la table une valeur de chi-deux immédiatement inférieure à celle qu'on a calculée (53.65) et on en déduit que le p associé au khi-deux calculé est inférieur à celui qui figure dans la table.

Valeur de p ? **p < 0.001**

Que peut-on conclure ? **Globalement, les personnes interrogées n'aiment pas toutes les viennoiseries de la même façon : préférence pour les pains au chocolat**

Remarque: parallèle entre…

- Le khi-deux de conformité compare la distribution d'une **variable nominale** à une valeur fixe (ici la valeur de hasard, déterminée à partir du nombre de catégories de la variable)

- Test t univarié (cf. TD Comparaison de moyennes) compare la distribution d'une **variable quantitative** à une valeur fixe (qui peut aussi être la valeur de hasard)

Mini-TD

On veut caractériser des écoles par leur degré de mixité sociale

école mixte (milieux sociaux équilibrés); **école non mixte** (les familles d'un certain milieu – Fav ou Défav.- sont plus nombreuses).

Pour chaque élève de classes de CM de chaque école, on prend en considération la profession des parents et chaque élève est ensuite affecté à un milieu social

4 milieux sociaux: Défavorisé; Intermédiaire -; Intermédiaire +; favorisé

Nbre d'élèves (effectifs)

École	*Défav*		*Inter-*		*Inter+*		*Fav*		Tot.
du Bois	6	13,9%	6	13,9%	8	18,6%	23	53,40%	43
Mille Chemins	7	17,5%	12	30%	12	30%	9	22,5%	40
Village	7	14%	19	38%	17	34%	7	14%	50
Portail Rouge	17	45,9%	6	16,2%	6	16,2%	8	21,6%	37
Renaudie+ Vercors	21	80,8%	2	7,7%	2	7,7%	1	3,8%	26
Total	58		45		45		48		196

Tableau 6 : Profil social des écoles

Buson, L. (2009). *Variation stylistique entre 5 et 11 ans et réseaux de socialisation scolaire: usages, représentations, acquisition et prise en compte éducative*. Thèse de doctorat, Université Stendhal, Grenoble.

6. TD4 – Khi-deux
6.1 Définition
6.2 Khi-deux de conformité

École	*Défav*		*Inter-*		*Inter+*		*Fav*		Tot.
du Bois	6	13,9%	6	13,9%	8	18,6%	23	53,40%	43
Mille Chemins	7	17,5%	12	30%	12	30%	9	22,5%	40
Village	7	14%	19	38%	17	34%	7	14%	50
Portail Rouge	17	45,9%	6	16,2%	6	16,2%	8	21,6%	37
Renaudie+ Vercors	21	80,8%	2	7,7%	2	7,7%	1	3,8%	26
Total	58		45		45		48		196

Tableau 6 : Profil social des écoles

Mini-TD

- L'auteure a pris une répartition équilibrée comme effectif théorique pour calculer le khi-deux (43/4 pour l'école du Bois).

- Recalculer les khi-deux et trouver les p associés dans la table distribuée (5 groupes d'étudiants, une école par groupe)

- Au vu des khi-deux, quelles écoles peuvent être qualifiées de socialement mixtes ou non mixtes ?

43/4

Calcul du X^2 pour l'école du Bois

Effectifs observés	Effectifs théoriques	Effectifs observés - Effectifs théoriques
6 (defav)	10.75	-4.75
6 (inter-)	10.75	-4.75
8 (inter+)	10.75	-2.75
23 (fav)	10.75	12.25

(Effectifs observés - Effectifs théoriques)²	(Effectifs observés - Effectifs théoriques)² / fréq. attendue	X^2 (somme de la colonne précédente)
22.56	2.10	18.86
22.56	2.10	
7.56	0.70	
150.06	13.96	

p / ddl	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6679	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245

	École du Bois	École Mille Chemins	École Village	École Portail Rouge	Écoles Renaudie + Vercors
Khi-deux	18,86	1,80	9,84	8,95	43,23
	p < 0.001	Non significatif	p < 0.02	p < 0.05	p < 0.001
degré de liberté	3	3	3	3	3

On cherche dans la table une valeur de khi-deux immédiatement inférieure à celle qu'on a calculée et on en déduit que le p associé au khi-deux calculé est inférieur à celui qui figure dans la table.

6. TD4 – Khi-deux
6.1 Définition
6.2 Khi-deux de conformité

École	*Défav*		*Inter-*		*Inter+*		*Fav*		Chi 2	p	conclusion
du Bois	6	13,9%	6	13,9%	8	18,6%	23	53,40%	18,86	<0.001	Non mixte favorisée
Mille Chemins	7	17,5%	12	30%	12	30%	9	22,5%	1,80	ns	mixte
Village	7	14%	19	38%	17	34%	7	14%	9,84	< 0.02	Non mixte intermédiaire
Portail Rouge	17	45,9%	6	16,2%	6	16,2%	8	21,6%	8,95	< 0.05	Non mixte défavorisé
Renaudie+ Vercors	21	80,8%	2	7,7%	2	7,7%	1	3,8%	43,23	<0.001	Non mixte défavorisé
Total	58		45		45		48				

Tableau 6 : Profil social des écoles

Au sens strict, une seule école – Mille Chemins - est vraiment mixte. Le khi-deux est NS et tous les milieux sociaux y sont équitablement représentés.
Dans les 4 autres écoles, un milieu social est plus représenté : favorisé (Bois), intermédiaire (Village) ou défavorisé (Portail Rouge et Renaudie+Vercors).
NB – Au total, l'auteure trouve que les enfants de familles défavorisées scolarisées au contact d'enfants d'autres milieux ont plus de flexibilité stylistique : ils adaptent mieux leur façon de parler à l'interlocuteur.

Buson, L. (2009). *Variation stylistique entre 5 et 11 ans et réseaux de socialisation scolaire: usages, représentations, acquisition et prise en compte éducative*. Thèse de doctorat, Université Stendhal, Grenoble.

EXAMEN

- **Lundi 9 janvier à 10H30 à 12H30 en Amphi 4**
- 1/3 temps supplémentaire: RDV à 09H50 en C213
- Rappel:
  - **Documents papiers autorisés**
    - **Mais PAS d'ordinateur ou téléphone portable**
  - Supports de cours mis en ligne sur *Alfresco Share*

https://espaces-collaboratifs.grenet.fr/share/page/site-index

!!! La semaine prochaine !!!

- **Lundi 12 décembre**
  - **Cours avec T. Mout en F111**
    (8h30-10h30: groupe1; 10h30-12h30: groupe 2)
  - Séance sur l'examen de l'an passé (envoyé par mail pendant les vacances de la Toussaint)

6.2 Khi-deux d'indépendance

- Mesure le lien entre 2 variables nominales (ex: pays de naissance et couleur des yeux)
- Permet, à partir de l'observation de l'échantillon, de décider si 2 variables nominales sont indépendantes ou non.

- On veut savoir s'il y a un lien entre le fait de boire et de fumer chez les étudiants
  - Enquête : On demande à 110 étudiants :
    > fumez-vous régulièrement (plus d'une cigarette / semaine) ?
    > buvez-vous de l'alcool régulièrement (au moins deux verres/ semaine) ?

<u>Quatre catégories de réponses</u>

Combien de personnes boivent et fument ?
Combien de personnes boivent mais ne fument pas ?
Combien de personnes ne boivent pas mais fument ?
Combien de personnes ne boivent pas et ne fument pas ?

1/ On résume les données observées dans une table 2X2 (2 lignes et 2 colonnes)

2/ On calcule les totaux des lignes et des colonnes

	Fume	Ne fume pas	Total
Boit	50	15	65
Ne boit pas	20	25	45
Total	70	40	110

La 1ère étape du test consiste à déterminer combien on peut s'attendre à trouver d'étudiants dans chaque catégorie en supposant qu'il n'y a aucune relation entre les 2 variables (= hypothèse nulle) → Calcul des effectifs théoriques (méthode assez similaire à celle utilisée par le khi-deux de conformité SAUF qu'on ne s'attend pas à avoir 4 cellules identiques. Les effectifs théoriques ne sont donc pas tous égaux).

Logique
Si pas de lien entre FUME et BOIT, la proportion de buveurs et de non buveurs devrait être à peu près la même chez les fumeurs et les non fumeurs : 65/110 et 45/110

3/ Calcul des effectifs théoriques:

On se demande ce qu'on aurait dans chaque case si la proportion restait la même que dans les totaux marginaux : combien y aurait-il de buveurs parmi les 70 fumeurs si il y en avait la même proportion que 65/110 ?

On multiplie le total de chaque ligne par le total de chaque colonne en divisant par le total général (110).

Effectifs observés

	Fume	Ne fume pas	Total
Boit	50	15	65
Ne boit pas	20	25	45
Total	70	40	110

Pour FUME et BOIT : (70*65)/110 = 41.4

Pour NE FUME PAS et BOIT: (40*65)/110 = 23.6

Effectifs théoriques

	Fume	Ne fume pas	Total
Boit	**41.4**	**23.6**	65
Ne boit pas	**28.6**	**16.4**	45
Total	70	40	110

4/ On soustrait les effectifs théoriques aux effectifs observés

Effectifs observés - Effectifs théoriques

	Fume	Ne fume pas
Boit	**50 – 41.4 = 8.6**	**15 – 23.6 = -8.6**
Ne boit pas	**20 – 28.6 = -8.6**	**25 – 16.4 = 8.6**

- **5/ On élève au carré la valeur absolue des différences**
- **6/ On divise ces nombres par l'effectif théorique de chaque cellule**

	Fume	Ne fume pas
Boit	**8.6^2/41.4** **1.8**	**8.6^2/ 23.6** **3.1**
Ne boit pas	**8.6^2/28.6** **2.6**	**8.6^2/16.4** **4.5**

- **7/ On additionne tous ces nombres**
- 1.8 + 3.1 + 2.6 + 4.5 = 12

KHI-DEUX = 12

- **On associe une valeur de p à ce khi-deux de 12: on regarde dans la table Khi-deux au bon ddl**

!!! Dans le cas de khi-deux à plus d'une variable :

ddl = (nbre de lignes -1) x (nbre de colonnes -1)

Ici 2 lignes et 2 colonnes, donc ddl = (2-1) X (2-1) = 1

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245

Avec un ddl de 1, le khi-deux calculé de 12 est supérieur à 10.8276. On en conclut donc que le p associé est inférieur à 0.001.
La probabilité d'obtenir un khi-deux aussi élevé que 12 dans l'échantillon en supposant que les 2 variables FUME et BOIT sont indépendantes dans la population (hypothèse nulle) est inférieure à 1 chance sur 1000.

Conclusion :
Chez les étudiants interrogés, il y a un lien entre le fait de boire et le fait de fumer

Mini-TD

Version simplifiée du fichier sur les erreurs de liaison recueillies chez une fillette entre 2 ans 1 mois et 6 ans 4 mois

Mot2	Erreurs en N	Erreurs en Z	Orientation singulier / pluriel
Ami	1	11	pluriel
Arbre	10	7	pluriel
Enfant	11	14	pluriel
Escalier	6	0	pluriel
Habit	9	10	pluriel
Oiseau	23	32	pluriel
Ane	27	1	singulier
Anorak	6	0	singulier
Arc-en-ciel	11	0	singulier
Avion	12	1	singulier
Eléphant	15	0	singulier
Orage	11	2	singulier
Ours	20	0	singulier

Etape 1 : construisez le tableau à double entrée ci-dessous

	Nombre total d'erreurs /n/	Nombre total d'erreurs /z/
Noms orientés pluriel		
Noms orientés singulier		

Etape 2 : Calculez le khi-deux d'indépendance et concluez

Mot2	Erreurs en N	Erreurs en Z	Orientation singulier / pluriel
Ami	1	11	pluriel
Arbre	10	7	pluriel
Enfant	11	14	pluriel
Escalier	6	0	pluriel
Habit	9	10	pluriel
Oiseau	23	32	pluriel
Ane	27	1	singulier
Anorak	6	0	singulier
Arc-en-ciel	11	0	singulier
Avion	12	1	singulier
Eléphant	15	0	singulier
Orage	11	2	singulier
Ours	20	0	singulier

Étape 1 : construisez le tableau à double entrée ci-dessous

	Nombre total d'erreurs /n/	Nombre total d'erreurs /z/
Noms orientés pluriel	60	74
Noms orientés singulier	102	4

Etape 2 : Calculez le khi-deux d'indépendance et concluez

	Nombre total d'erreurs /n/	Nombre total d'erreurs /z/	
Noms orientés pluriel	60 $ET_1 = (162X134) / 240 = \underline{90,45}$ $D_1 = (EO_1-ET_1)^2 / ET_1$ $D_1 = 10,250995$	74 $ET_2 = (78X134) / 240 = \underline{43,55}$ $D_2 = (EO_2-ET_2)^2 / ET_2$ $D_2 = 21,2905281$	134
Noms orientés singulier	102 $ET_3 = (162x106) / 240 = \underline{71,55}$ $D_3 = (EO_3-ET_3)^2 / ET_3$ $D_3 = 12,958805$	4 $ET_4 = (78x106) / 240 = \underline{34,45}$ $D_4 = (EO_4-ET_4)^2 / ET_4$ $D_4 = 26,9144412$	106
	162	78	240

$X^2 = D_1 + D_2 + D_3 + D_4$
Donc $X^2 \approx 71,41$
DDL = 1
$p < 0.001$

Conclusion
Les erreurs en /n/ et /z/ sont plus ou moins fréquentes selon que les noms sont orientés au pluriel et au singulier.
- Que se passe-t-il exactement pour chaque type de noms ?