SOCI1241-1 Eléments du calcul des probabilités appliquées aux sciences sociales
Exercices complémentaires chapitre 4
Comparaison de 2 proportions (données pairées) – Association entre 2 variables catégorielles

---

Comparaison de 2 proportions (données pairées)

1. Un sondage auprès de 50 élèves d'un lycée (comportant plusieurs milliers d'élèves) a été effectué avant et après la remise des résultats de juin. Deux questions leur ont été posées. La première était posée avant la remise des résultats et leur demandait s'ils pensaient avoir réussi leur année. 40 élèves ont répondu « oui ». La seconde question était posée après la remise des résultats et leur demandait s'ils avaient réussi leur année : 18 élèves ont répondu non. Seuls 28 élèves avaient estimé correctement avoir réussi leur année.
   a. Ces données suggèrent-elles que, dans le lycée, la proportion de personnes pensant réussir leur année est égale à la proportion réelle de réussites?
   b. Donnez un ensemble de valeurs plausibles pour cette différence de pourcentages.
   c. Répondez à la question a) en utilisant un test.

2. Une étude cross-over a pour objectif de comparer l'efficacité de deux médicaments contre les migraines. Codons le succès du médicament par 1 et l'échec par 0. Pour un premier groupe de 50 sujets, le traitement A est administré à la première migraine et le traitement B à la seconde. Les résultats pour cette séquence A-B sont les suivants: 6 ont (1-1), 25 ont (1-0), 10 ont (0-1) et 9 ont (0-0). Pour le deuxième groupe de 50 sujets les traitements A et B sont administrés dans l'ordre inverse. Parmi ces 50 nouveaux sujets, les résultats pour la séquence B-A sont les suivants: 10 ont (1-1), 20 ont (0-1), 12 ont (1-0) et 8 ont (0,0). En oubliant l'ordre d'administration des traitements c'est-à-dire en supposant que le temps entre les deux migraines est suffisamment long pour que le premier traitement n'ait plus d'effet,
   a. Déterminez s'il y a lieu penser que les deux traitements ont la même efficacité à partir d'un intervalle de crédibilité? Explicitez votre conclusion.
   b. Déterminez s'il y a lieu penser que les deux traitements ont la même efficacité en utilisant un test adéquat.

3. On a demandé à 200 personnes, choisies au hasard parmi de nombreux chefs d'entreprise, si elles approuvaient la politique de leur gouvernement: 35% des personnes questionnées la désapprouvent. Quelques mois plus tard, la monnaie du pays est violemment attaquée sur les marchés des changes et subit une dévaluation de 5%. On repose alors la même question aux mêmes chefs d'entreprise: désormais, 60% approuvent la politique du gouvernement. De plus, 50 des 200 personnes questionnées désapprouvent l'action du gouvernement avant et après la dévaluation de la monnaie.
   a. Quelle est la probabilité que la crise monétaire ait augmenté le pourcentage des chefs d'entreprises approuvant la politique du gouvernement ?
   b. Donnez un ensemble de valeurs plausibles pour la différence de pourcentages concernée.
   c. Peut-on dire que l'action du gouvernement durant la crise a significativement modifié la satisfaction des chefs d'entreprises ? Répondez à cette question à l'aide d'un test adéquat.

Association entre 2 variables catégorielles

4. Dans l'étude de Framingham (USA), on a suivi une cohorte d'hommes âgés de 35 à 44 ans, classés en deux groupes à risque en fonction de leur taux de cholestérol de départ (soit >2,5 g/l, soit ≤2,5 g/l). L'étude s'est intéressée à la survenue de l'infarctus du myocarde. 603 hommes sont considérés et les résultats sont les suivants : parmi 135 hommes ayant un taux de cholestérol de départ > 2,5 g/l, 10 ont eu un infarctus du myocarde. De plus, 447 hommes ont un taux de cholestérol de départ ≤2,5 g/l et n'ont pas eu un infarctus du myocarde.
   a. Effectuez une analyse descriptive des données. (suggestion: estimez les pourcentages conditionnels en ligne). Représentez vos résultats graphiquement.
   b. À partir de cette étude, peut-on déterminer si le cholestérol influence l'occurrence de l'infarctus du myocarde?
   c. Estimez et donnez un intervalle de crédibilité pour le rapport de risques. Que peut-on en conclure ?
   d. Estimez et donnez un intervalle de crédibilité pour la différence de risques. Que peut-on en conclure?

5. Lors d'une étude sur les carences en iodure (µg/l) d'enfants mis en crèches, on a voulu déterminer également s'il existait un lien entre les habitudes alimentaires et le niveau de carence en iodure des mères. Pour cela, chaque mère impliquée dans l'étude a rempli un questionnaire portant sur ses habitudes alimentaires. Il y avait notamment la question suivante: « A quelle fréquence mangez-vous du poisson? »(Jamais, au plus1 fois/mois, >1 fois/mois). Voici les résultats obtenus:

	Carence		
Consommation de Poisson	Absence	Modérée	Sévère
Jamais	3	12	15
Au plus 1 fois/mois	12	60	15
Plus d'1 fois/mois	60	96	27

   a. Effectuez une analyse descriptive des données. (suggestion: estimez les pourcentages conditionnels en ligne). Représentez vos résultats graphiquement.
   b. Peut-on, à partir de ces données, suggérer que la consommation de poisson influence le niveau de carence?
   c. Un médecin s'intéresse uniquement au risque d'une carence élevée lorsqu'on ne mange pas de poisson (respectivement: lorsqu'on en mange au plus 1 fois par mois). A l'aide de rapports, comparez ces deux risques avec le risque de carence lors d'une consommation >1 fois/mois. Que peut-on en conclure?

6. Un grand fabricant de cigarettes désire savoir si le fait de fumer est lié à la consommation de chocolat. À cette fin, une enquête a été réalisée auprès de 1000 personnes. Parmi 600 mangeurs de chocolat, 200 personnes ne fument pas. De plus, 100 personnes ne consomment ni chocolat, ni cigarette.
   a. Effectuez une analyse descriptive des données. Représentez vos résultats graphiquement.

b. À partir de ce sondage, peut-on déterminer si le fait de fumer est lié à la consommation de chocolat?
c. Dans la population concernée, comparez les proportions de consommateurs de chocolat chez les fumeurs et chez les non-fumeurs de deux manières différentes.

Test d'ajustement (univarié)

7. La répartition des principaux groupes sanguins en France est reprise dans le tableau ci-dessous :

O Rhésus+	A Rhésus+	B Rhésus+	AB Rhésus+	O Rhésus-	A Rhésus-	B Rhésus-	AB Rhésus-
37%	38.1%	6.2%	2.8%	7%	7.2%	1.2%	0.5%

Le centre de transfusion sanguine de Pau a observé la répartition suivante sur 5000 donneurs :

O Rhésus+	A Rhésus+	B Rhésus+	AB Rhésus+	O Rhésus-	A Rhésus-	B Rhésus-	AB Rhésus-
2291	1631	282	79	325	332	48	12

La répartition palinoise des 8 types groupes-rhésus est-elle différente de la répartition nationale ?

Réponses exercices complémentaires :

Exercice 1.

NB : $n_{21} + n_{12} < 25$ ➔ les conditions d'utilisation ne sont pas remplies en pratique pour appliquer les méthodes vues au cours. Si on applique néanmoins ces méthodes, on obtient :

a) $(\pi_2 - \pi_1 | données) \approx N(-0.16, 0.00589)$ → $P(\pi_2 - \pi_1 < 0|D) = 0.9817$: la plausibilité que la proportion de réussites effectives soit plus faible que la proportion de réussites attendues est >0.95 ➔ suffisant pour conclure que les 2 proportions différent (surestimation de la réussite).

b) $IC\ 95\%\ pour\ \pi_2 - \pi_1$: $[-0.3104;\ -0.0096] \not\ni 0$ → conclusion cf. point a)

c) $Valeur\ calculée\ pour\ le\ test\ de\ McNemar$: $4 > 3.84$ → on rejette l'hypothèse selon laquelle 0 est une valeur plausible pour la différence de proportions (pairées).

Exercice 2.

Indice :

		Traitement B	
		1	0
Traitement A	1	16	45
	0	22	17

$n_{21} + n_{12} \geq 25$ → Conditions d'utilisation OK

a) $(\pi_2 - \pi_1 | données) \approx N(-0.23, 0.006171)$→ $IC\ 95\%\ pour\ \pi_2 - \pi_1$: $[-0.384;\ -0.076] \not\ni 0$: puisque 0 n'est pas une valeur plausible, il y a bien une différence d'efficacité entre les 2 traitements : le traitement B est significativement moins efficace que le traitement A.

b) $Valeur\ calculée\ pour\ le\ test\ de\ McNemar$: $7.89 > 3.84$ → on rejette l'hypothèse selon laquelle 0 est une valeur plausible pour la différence de proportions (pairées).

Exercice 3.

Indice :

		Approbation T2		
		1	0	
Approbation T1	1	100	30	130
	0	20	50	70
		120	80	200

$n_{21} + n_{12} \geq 25$ → Conditions d'utilisation OK

a) $(\pi_2 - \pi_1 | données) \approx N(-0.05, 0.0352^2)$→ $P(\pi_2 - \pi_1 > 0|D) = 0.0778$

b) $IC\ 95\%\ pour\ \pi_2 - \pi_1$: $[-0.12, 0.02] \ni 0$: puisque 0 est une valeur plausible, on ne peut pas conclure à une différence significative (proportion de « satisfaits » avant/après)

c) $Valeur\ calculée\ pour\ le\ test\ de\ McNemar$: $2 < 3.84$ → même conclusion qu'en b)

Exercice 4.

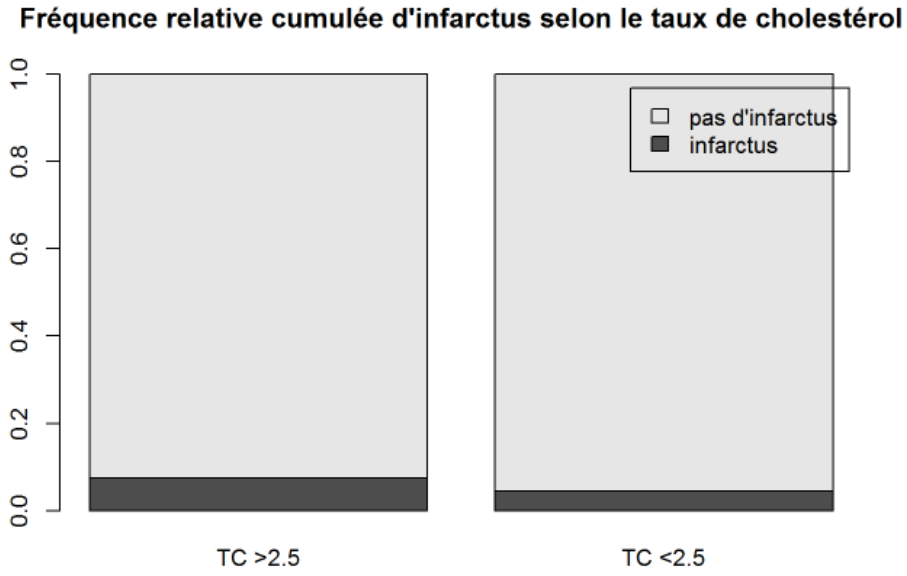
a) Table de contingence :

	Infarctus	Pas d'infarctus
TC >2.5	10	125
TC <2.5	21	447

Table des fréquences relatives de la réponse (en %) conditionnellement à l'explicative :

	Infarctus	Pas d'infarctus
TC >2.5	7.41	92.59
TC <2.5	4.49	95.51

Représentation graphique :



Interprétation :

Dans l'échantillon, on observe que la proportion d'infarctus est un peu plus élevée dans le groupe ayant un taux de cholestérol >2.5g/l mais le risque est faible dans les deux groupes.

b) Test d'indépendance : $G^2 = 1.69 < \chi^2_1(0.95) = 3.84$ ➔ on ne rejette pas l'hypothèse d'indépendance ⇔ avec les données à disposition, on ne peut pas dire que le cholestérol influence l'occurrence de l'infarctus du myocarde. NB : $A_{ij} \geq 5\ \forall i,j$: condition OK

c) $IC\ 95\%\ pour\ \frac{\pi_1}{\pi_2}$: $[0.8, 3.42] \ni 1$ ➔ même conclusion point b)

d) Cf. chapitre 2 : $IC\ 95\%\ pour\ \pi_1 - \pi_2$: $[-0.0188, 0.0772] \ni 0$ ➔ même conclusion

Exercice 5.

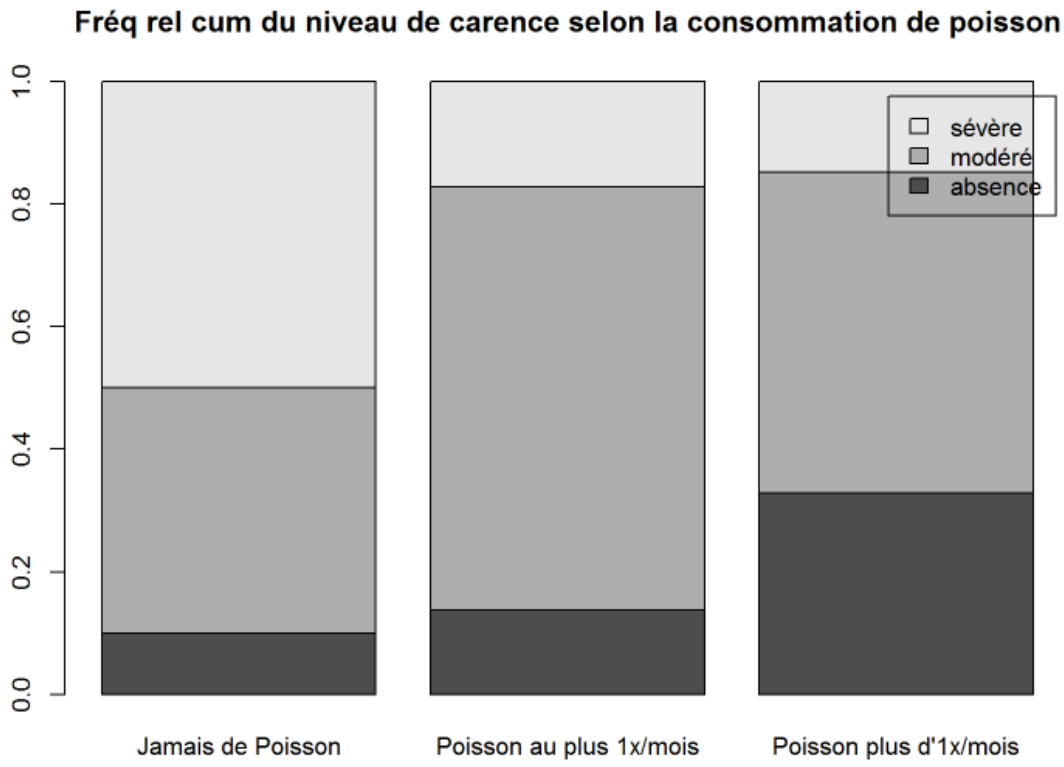
a) Table de contingence :

	Absence de Carence	Carence Modérée	Carence Sévère
Jamais de Poisson	3	12	15
Poisson au plus 1x/mois	12	60	15
Poisson plus d'1x/mois	60	96	27

Table des fréquences relatives de la réponse (en %) conditionnellement à l'explicative :

	Absence de Carence	Carence Modérée	Carence Sévère
Jamais de Poisson	10.00	40.00	50.00
Poisson au plus 1x/mois	13.79	68.97	17.24
Poisson plus d'1x/mois	32.79	52.46	14.75

Représentation graphique :



Interprétation :

Dans l'échantillon, on observe que plus la fréquence de consommation augmente, plus la proportion de carence sévère diminue et inversement, plus la proportion d'absence de carence augmente.

b) Test d'indépendance : $G^2 = 29.6 > \chi^2_4(0.95) = 9.49$ ➔ on rejette l'hypothèse d'indépendance ⇔ il existe un lien entre les carences en iodure et la fréquence de consommation de poisson. NB : $A_{ij} \geq 5\ \forall i, j$: condition OK

c) On commence par « binariser » la variable réponse :

	Carence Sévère	Pas de carence sévère
Jamais de Poisson	15	15
Poisson au plus 1x/mois	15	72
Poisson plus d'1x/mois	27	156

- $IC\ 95\%\ pour\ {}^{\pi_1}/_{\pi_3}$: $[2.02, 5.55] \not\ni 1$ ➔ les mères qui ne mangent jamais de poisson ont entre 2 fois et 5.5 fois plus de chance d'avoir des carences sévères par rapport à celles qui en mangent plus d'une fois par mois.
- $IC\ 95\%\ pour\ {}^{\pi_2}/_{\pi_3}$: $[0.64, 2.02] \ni 1$ ➔ pas de différence significative entre les mères qui mangent du poisson au plus 1x/mois Vs plus d'une fois par mois quant au risque de développer des carences sévères en iodure.

Exercice 6.

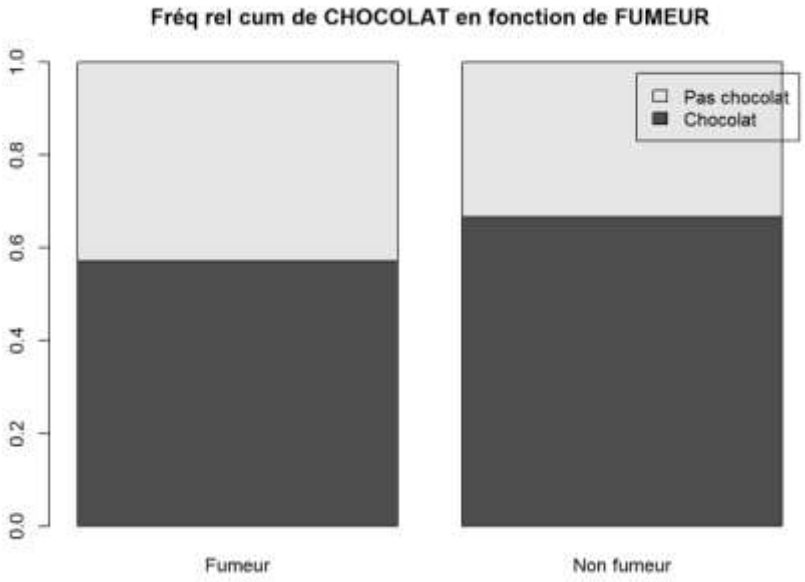
a) Table de contingence :

	Chocolat	Pas chocolat
Fumeur	400	300
Non fumeur	200	100

Table des fréquences relatives de la réponse (en %) conditionnellement à l'explicative :

	Chocolat	Pas chocolat
Fumeur	57.14	42.86
Non fumeur	66.67	33.33

Représentation graphique :



Interprétation :

Dans l'échantillon, on observe que la consommation de chocolat est plus élevée chez les non-fumeurs, ce qui suggère l'hypothèse selon laquelle le fait de fumer diminuerait l'appétence pour le chocolat (à vérifier au moyen d'un test).

b) Test d'indépendance : $G^2 = 8.04 > \chi_1^2(0.95) = 3.84$ ➔ on rejette l'hypothèse d'indépendance ⇔ il existe un lien significatif entre le fait de fumer et la consommation de chocolat. NB : $A_{ij} \geq 5\ \forall i,j$: condition OK

c) On pose :

- $\pi_1 = proportion\ de\ consommateur\ de\ chocolat\ chez\ les\ \boldsymbol{non\ fumeurs}$
- $\pi_2 = proportion\ de\ consommateur\ de\ chocolat\ chez\ les\ \boldsymbol{fumeurs}$

Deux méthodes pour comparer les proportions :

- IC 95% pour RR $= \frac{\pi_1}{\pi_2}$ = [1.053 ; 1.293] $\not\ni 1\ et\ V > 1 \rightarrow \pi_1 > \pi_2$: il y a significativement plus de consommateurs de chocolat chez les non-fumeurs que chez les fumeurs.
- IC 95% pour $\pi_1 - \pi_2$ = [0.03 ; 0.16] $\not\ni 0\ \ et\ V > 0 \rightarrow \pi_1 > \pi_2$ (même conclusion)

Exercice 7.

Fréquences attendues en cas d'ajustement

O Rhésus+	A Rhésus+	B Rhésus+	AB Rhésus+	O Rhésus-	A Rhésus-	B Rhésus-	AB Rhésus-
1850	1905	310	140	350	360	60	25

$A_i \geq 5\ \forall i$: condition OK

$G^2 = 188.32 > \chi_7^2(0.95) = 14.07$ → on rejette l'hypothèse d'ajustement ⇔ La répartition palinoise des 8 types groupes-rhésus est significativement différente de la répartition nationale.