Université de Lyon 2
Licence 3 MIASHS IDS

Statistique Inférentielle
Année 2020-21

TD 2 : Tests

Exercice 1 Un ingénieur risque crédit, employé dans une société spécialisée dans le crédit à la consommation, veut vérifier l'hypothèse selon laquelle la valeur moyenne des mensualités des clients de son portefeuille est de 200 euros. Un échantillon aléatoire de 144 clients, prélevé aléatoirement dans la base de données, donne une moyenne empirique $\bar{x} = 193.74$ et une estimation non biaisée de l'écart-type $s = 48.24$.

(i) Quelles sont les hypothèses statistiques associées à la problématique de l'ingénieur et quel type de test faut-il mettre en oeuvre pour l'aider à prendre une décision statistiquement correcte ?

Réponse : On a un modèle pour ces données qui peut être présenté de la manière suivante. Les données sont des variables Gaussiennes i.i.d. $X_1, \ldots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$. L'ingénieur suppose que $\mu = 200$. On pose donc

— $H_0 : \mu = \mu_0 = 200$,
— $H_1 : \mu \neq \mu_0$

(ii) Peut-il conclure, au risque 5%, que la valeur moyenne postulée des remboursements est correcte ?

Réponse : On effectue le test statistique correspondant à déterminer si H_0 est réfutée ou non. Pour cela, on réalise un test de Student. On fait l'hypothèse que H_0 est vraie et on regarde si cette hypothèse est contredite par les données. Sous l'hypothèse H_0, on a

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}.$$

On regarde donc dans la table de la loi de Student à $n-1$ degrés de liberté, les quantiles $t_{n-1,\frac{\alpha}{2}}$ et $t_{n-1,1-\frac{\alpha}{2}}$ associés aux niveaux $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ respectivement. On trouve les valeurs numériques $t_{n-1,1-\frac{\alpha}{2}} \approx 1.98$ en prenant pour valeur des degrés de liberté 120 au lieu de 143. L'intervalle de non rejet de H_0 est donc

$$I_\alpha = [-t_{n-1,1-\frac{\alpha}{2}}; t_{n-1,1-\frac{\alpha}{2}}] = [-1.98; 1.98] \tag{1}$$

On calcule alors la valeur de

$$\begin{aligned} \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} &= \frac{193.74 - 200}{\sqrt{\frac{48.24^2}{144}}} \\ &= -1.56 \end{aligned}$$

et on constate que cette valeur tombe dans l'intervalle : on ne rejette donc pas H_0, on ne peut pas contredire l'ingénieur.

(iii) Faîtes le schéma des régions de rejet et de non rejet de l'hypothèse nulle H_0 en y notant les valeurs critiques calculées à la question précédente.

Réponse : voir Figure 1 ci-dessous.

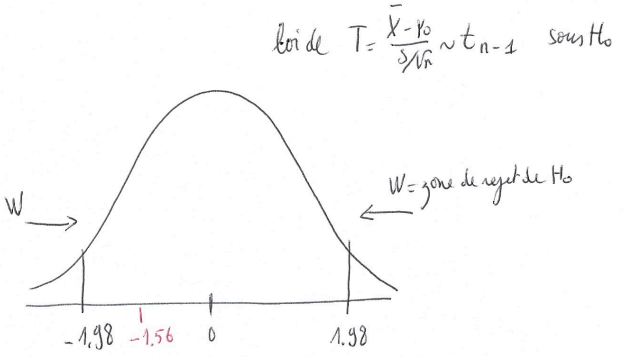

FIGURE 1 – Quantiles de Student et zones de rejet

(iv) Représenter sur ce schéma la p-value associée à ce test. Que vaut-elle ?

Réponse : voir Figure 2 ci-dessous.

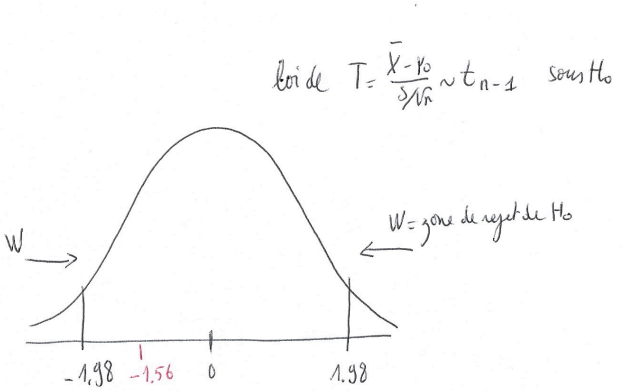

FIGURE 2 – Quantiles de Student et zones de rejet : la p-value correspond a l'aire en rouge hachurée

La-pvalue est donc égale à la probabilité qu'une variable aléatoire de Student $T \sim t_{n-1}$ soit inférieure à -1.56 ou supérieure à 1.56.

En regardant les tables, pour la loi t_{120} (car on n'a pas la t_{143}), on trouve $t_{120,0.95} = 1.658$ et $t_{120,0.90} = 1.289$. Une interpolation grossière nous donne $t_{120,0.94} \simeq 1.56$, d'où $p(T < -1.56) = p(T > 1.56) \simeq 0.06$ et donc une p-value de 0.12.

(v) En utilisant la p-value, quelle aurait été la réponse à la question 2 pour un risque de première espèce $\alpha = 10\%$.

Réponse : Comme le risque $\alpha = 10\%$ est plus petit que la p-value, même conclusion : on ne rejette pas H_0, on ne peut contredire l'ingénieur.

Exercice 2 Dans le cours de Statistique Inférentielle en L3 MIASHS, il y a cette année 28 femmes sur 64 étudiants. En considérant cette promotion comme représentative des étudiants en informatique et statistique, peut-on affirmer que ce type de formation intéresse autant les hommes que les femmes ?

Réponse : Le modèle pour ces données est le suivant. Les observations sont modélisées par des variables i.i.d. $X_1, \dots, X_n$ de type Bernoulli $\mathcal{B}(1, p)$. Cette variable est la réponse à la question du genre de la i^{ieme} personne interrogée. La proportion est

$$F = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2}$$

et on peut faire l'approximation que F suit une loi Normale. Son espérance est

$$\begin{aligned} \mathbb{E}[F] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\, \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] \end{aligned}$$

et comme

$$X_i = \begin{cases} 1 & \text{avec proba. } p \\ \\ 0 & \text{avec proba. } 1-p. \end{cases}$$

alors

$$\mathbb{E}[X_i] = 0(1-p) + 1 \cdot p = p.$$

Ainsi, on obtient que

$$\mathbb{E}[F] = \frac{1}{n} \sum_{i=1}^{n} p = p.$$

En ce qui concerne la variance de F,

$$\begin{aligned}\mathrm{Var}(F) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)\end{aligned}$$

or

$$\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$$

mais comme

$$X_i^2 = \begin{cases} 1 & \text{avec proba. } p \\ \\ 0 & \text{avec proba. } 1-p. \end{cases}$$

ce qui donne $\mathbb{E}[X_i^2] = 0\cdot(1-p) + 1\cdot p = p$ et ainsi

$$\mathrm{Var}(X_i) = p - p^2 = p(1-p).$$

Donc

$$\mathrm{Var}(F) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{1}{n^2}\; n\; p(1-p) = \frac{1}{n}p(1-p).$$

Ainsi, l'approximation de la loi de F par une Gaussienne provenant du théorème central limite est de la forme

$$F \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right). \tag{3}$$

La question correspond à faire le test de savoir si $p = .5$ ou pas. On pose alors

- $H_0 : p = p_0 = .5$
- $H_1 : p \neq p_0$

On se place alors sous l'hypothèse H_0 et on va vérifier si les données contredisent ou pas cette hypothèse. Sous H_0 on a

$$\frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx \mathcal{N}(0,1). \tag{4}$$

Les quantiles $u_{\frac{\alpha}{2}}$ et $u_{1-\frac{\alpha}{2}}$ pour $\alpha = 5\%$ sont donnés par $u_{1-\frac{\alpha}{2}} = 1.96$ et $u_{\frac{\alpha}{2}} = -1.96$. L'intervalle de non rejet de H_0 est donc $I_\alpha = [-1.96; 1.96]$. De plus,

$$\frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{28/64 - .5}{\sqrt{\frac{.5(1-.5)}{64}}} = -1. \tag{5}$$

Cette valeur étant dans I_α, on ne rejette pas H_0. On ne peut pas affirmer que ce type de formation n'intéresse pas autant les hommes que les femmes.

Exercice 3 Dans le cadre d'une étude sur la capacité d'apprentissage des adolescents, on recrute un échantillon de 30 adolescents pour une série de tests. Afin d'avoir un échantillon relativement homogène d'un point de vue du QI (Quotient Intellectuel), le protocole statistique impose que l'écart-type du QI n'excède pas les 20 points.

							QI							
131	108	85	96	86	126	128	107	119	87	103	110	125	77	90
109	109	129	95	117	107	102	83	114	72	99	103	97	109	97

(i) Donnez une estimation sans biais de l'écart-type s de la population dont provient l'échantillon de 30 adolescents.

Réponse : On trouve $s^2 = 245.93$, ce qui donne $s = 15.68$.

(ii) L'hypothèse selon laquelle l'écart-type ne doit pas excéder 20 est-elle acceptable pour un risque de première espèce $\alpha = 0.05$? Quelles hypothèses devez-vous faire sur les données pour pouvoir répondre à cette question ?

Réponse : Il faut tester si $\sigma \leq 20$, c'est à dire $\sigma^2 \leq 400$. On suppose que les valeurs sont i.i.d. Gaussiennes $\mathcal{N}(\mu, \sigma^2)$. Dans ce cas, on a

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

L'hypothèse H_0 est "$\sigma^2 = 400$" et H_1 est "$\sigma^2 < 400$". On se place sous H_0 et on fixe $\alpha = 0.05$. On trouve la valeur $\chi^2_{29,0.05} = 17.7$ dans les tables, et l'intervalle de rejet de H_0 est

$$I_\alpha =]+\infty, 17.7]$$

comme

$$\frac{(n-1)s^2}{400} = 29 \cdot 245.93/400 = 17.83.$$

n'est pas dans l'intervalle, on ne rejette pas H_0, on ne peut pas affirmer que l'écart-type est conforme au seuil toléré.

On aurait aussi pu tester $H_0 : \sigma^2 = 400$ contre $H_1 : \sigma^2 > 400$. On aurait alors comparé la valeur de la statistique de test, 17.83 au $\chi^2_{29,0.95} = 42.6$. Là non plus on n'aurait pas pu

rejeter H_0, on n'aurait donc pas pu conclure que l'échantillon n'est pas conforme à ce qui est attendu. Quelle est la différence entre choisir $H_1 : \sigma^2 > 400$ et $H_1 : \sigma^2 < 400$: dans le premier cas, en cas de rejet de H_0, on maîtrise le risque de se tromper en concluant que l'échantillon d'étudiants n'est pas conforme, et donc de faire recommencer le choix de cet échantillon aux commanditaire de l'étude. Dans le cas $H_1 : \sigma^2 < 400$, on maîtrise le risque de se tromper en homologuant un échantillon d'étudiants qui serait non conforme, et donc en laissant se réaliser une étude dont les résultats pourraient être biaisé par cet échantillon non conforme. Certainement que ce second choix est plus judicieux, car le risque plus important que pour le premier choix : mieux vaut refaire faire une étude que d'en homologuer une dont les résultats pourraient être faux.

Exercice 4 La loi SRU impose aux communes de disposer de 25% de logement sociaux. Un département assure que, même s'il existe des disparités locales, elle respecte en moyenne le quota de logement sociaux. Les quotas relevés dans 10 villes du département sont les suivants :

Pourcentage de logements sociaux	25.6	24.5	24.3	25.0	29.5	24.1	24.8	24.7	25.2	24.9

A partir de cet échantillon de données, diriez-vous que les communes de ce département respecte la loi (risque $\alpha = 2.5\%$) ?

Réponse 1 : On peut odéliser ces données par une suite $X_1, \dots, X_n$ i.i.d. de variables Gaussiennes $\mathcal{N}(\mu, \sigma^2)$. En effet, les fréquences observées sont obtenues à partir d'un grand nombres de logements pour chaque ville et le Théorème Central Limite nous permet de justifier cette approximation Gaussienne. L'hypotèse à tester est $H_0 : \mu = 25$ contre $H_1 : \mu < 25$. En faisant ce choix, on maitrise le risque de se tromper en affirmant que le département ne respecte pas la loi. C'est une affirmation qui peut être lourde de conséquence, il est important de maitriser le risque associé. Le choix inverse $H_1 : \mu > 25$ nous aurait permis de maîtriser de risque de se tromper en affirmant que le département respecte la loi... Selon moi, il le premier choix est préférable. On ne connaît pas la variance mais on peut l'estimer (sans biais). On trouve $\bar{x} = 25.26$ et $s^2 = 2.41$. On sait de plus que

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

On a alors l'hypothèse H_0 consistant à assumer que $\mu = 25$ et l'hypothèse H_1 consistant à assumer que $\mu \neq 25$. Sous H_0, on a donc

$$\mathbb{P}\left(\frac{\bar{X} - 25}{\sqrt{\frac{S^2}{n}}} \leq t_{n-1,1-\alpha} \right) = 1 - \alpha.$$

On trouve les valeurs des quantiles $t_{9,1-.0125} = 2.69$ (interpolation). L'intervalle de rejet de H_0 est donc

$$I_\alpha = [2.69; +\infty[$$

La valeur numérique est

$$\frac{\bar{x}-25}{\sqrt{\frac{s^2}{10}}} = \frac{25.26-25}{\sqrt{\frac{2.41}{10}}} = 0.53.$$

ce qui ne nous permet pas de rejeter H_0. Nous ne pouvons affirmer que le département ne respecte pas la loi.

Réponse 2 : sans remarquer que les données étant des fréquences et donc pouvaient être modélisées par des Gaussiennes, on aurait réaliser un test non paramétrique car l'échantillon est de petite taille. Parmi les tests non paramétriques, Wilcoxon est le plus puissant, mais le test du signe est très rapide à réaliser... Corrigeons les deux. Tout d'abord, on enlève la donnée qui est pile sur 25, car elle n'apporte rien quand à la symétrie de la distribution par rapport à 25 (ce que l'on teste via un test non paramétrique). Le test du signe est très simple. Il suffit de compter le nombre de villes au dessus de 25 : 3 sur 9. La p-value correspondante, pour un test bilatéral, est 0.5078. Pour un test unilatéral, il faut diviser la p-value par 2 : 0.2539. Néanmoins, cette p-value reste forte, on ne rejette pas H_0. Pour le test de Wilcoxon, il faut retrancher 25 à toutes les données et calculer les rangs signés : classement par ordre croissante de valeurs absolues :

x_i	25.6	24.5	24.3	25.0	29.5	24.1	24.8	24.7	25.2	24.9
$x_i - 25$	0.6	-0.5	-0.7	0	4.5	-0.9	-0.2	-0.3	0.2	-0.1
r_i^+	**6**	5	7	0	**9**	8	2.5	4	**2.5**	1

La statistique de Wilcoxon est la somme de r_i^+ des x_i positif (en gras ci-dessus), et donc vaut ici : $W = 17.5$. Pour $n = 9$, au risque 5%, les tables de Wilcoxon nous indique de rejetter $H0$ si $W \leq 5$ ou si $W \geq 40$. Or on veut faire un test unilatéral avec $H_1 : \mu < 25$. Si H_1 est vraie, alors on aura beaucoup de $x_i - 25$ négatif et par conséquent un W qui devrait être petit : la zone de rejet pour cette hypothèse alternative est donc, au risque 2.5%, $W \leq 5$. Ici ce n'est pas le cas, nous ne rejetons pas $H0$

Exercice 5 Une société de vente à distance demande à l'un de ses ingénieurs marketing de modéliser le nombre d'appels téléphoniques par heure reçus sur le standard dédié aux commandes, dans le but d'optimiser la taille de celui-ci. Les nombres d'appels, relevés sur une période de 53 heures, ont été les suivants :

Nombre d'appels x_i	0	1	2	3	4	5	6	7	8	9 et plus
Occurrences N_i	1	4	7	11	10	9	5	3	2	1

(i) Estimer l'espérance et la variance du nombre d'appels. Quelle type de loi semble le mieux décrire ce nombre d'appel ?

Réponse : On a $\bar{x} = 4$ et $s^2 = 3.92$. On a donc que l'estimation de l'espérance est proche de l'estimation de la variance. Cela suggère que nous avons affaire à une loi de Poisson. Pour confirmer, on dessine un diagramme en bâtons :
Ce diagramme ressemble, par sa forme, à celui de la densité d'une loi de Poisson. Par ailleurs, comme les données représentent un nombre d'occurrences dans une plage de temps fixe, cela correspond typiquement à la loi de Poisson.

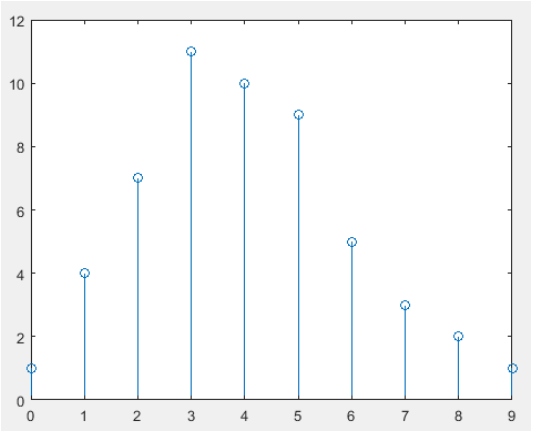

FIGURE 3 – Effectifs par nombre d'appels.

(ii) Tester l'ajustement à cette loi au risque 5%.

Réponse : On rappelle que pour une variables aléatoire discrète X de loi f, prenant ses valeurs dans $\{v_i\}_{i=1}^K$, la variable

$$Z = \sum_{i=1}^{K} \frac{(np_i - N_i)^2}{np_i} \tag{6}$$

où p_i est la probabilité de prendre la valeur v_i, et n est la taille de l'échantillon, suit une loi du χ^2_{K-1}. On peut donc faire le test suivant. On définit :

— H_0 : la loi des données est une loi de Poisson $\mathcal{P}(4)$,

— H_1 : la loi des données n'est pas une loi de Poisson.

Comme on doit estimer la valeur du paramètre de la loi de Poisson en question (que l'on prend comme $\lambda = \bar{x} = 4$, on enlève un degré de liberté supplémentaire.

Il nous faut donc calculer les p_i. On peut soit utiliser la définition de la loi de Poisson, soit les tables (attention elles donnes les valeurs cumulées), soit R et la fonction `dpois(i,4)`. On fera bien attention pour la case "9 et +" de bien calculer la proba $p(x \geq 9)$ et non pas juste 9, de sorte que la somme des p_i soit bien égale à 1. Nous allons représenter cela dans le tableau suivant

x_i	0	1	2	3	4	5	6	7	8	9 et plus
N_i	1	4	7	11	10	9	5	3	2	1
p_i	0.018	0.073	0.147	0.195	0.195	0.156	0.104	0.060	0.030	0.022
np_i	0.954	3.869	7.791	10.335	10.335	8.268	5.512	3.180	1.590	1.116

Attention : il faut que chaque effectif théorique soit supérieur à 5. Ce n'est pas le cas ici, on doit donc regrouper des catégories contiguës.

x_i	2 et moins	3	4	5	6	7 et plus
N_i	12	11	10	9	5	6
p_i	0.238	0.195	0.195	0.156	0.104	0.112
np_i	12.614	10.335	10.335	8.268	5.512	5.936

On peut finalement calculer la statistique D^2, et on trouve $d^2 = 0.197$. La loi sous H_0 est une χ^2_{6-1-1} (il faut bien actualiser le nouveau nombre de classes après regroupement). Or le

quantile vaut $\chi^2_{4,0.95} = 9.49$, d^2 est bien en dessous de ce quantile, on ne rejette pas l'hypothèse de $\mathcal{P}(4)$.

(iii) Sachant qu'une hôtesse d'accueil téléphonique peut traiter jusqu'à 7 appels par heure, combien d'hôtesses doit-on employer pour pouvoir répondre à 95% des appels téléphoniques ?

Nous avons vu à la question précédente que le nombre d'appels par heure pouvait être modélisé par une loi de Poisson $\mathcal{P}(4)$. Pour déterminer le nombre minimal d'hôtesse d'accueil téléphonique nécessaire au traitement de 95% des appels, on doit d'abord regarder quel est le nombre d'appels qui peut intervenir dans 95% des cas, *i.e.* on recherche le quantile d'ordre 0.95 d'une $\mathcal{P}(4)$, qui est égal à 8.
Or une hôtesse peut prendre au maximum 7 appels en charge par heure, donc il faudra au minimum deux hôtesses pour prendre en charge ces 8 appels moyens par heure.

Exercice 6 Sur 2000 personnes interrogées à Lyon, 1040 affirment utiliser régulièrement les transports en commun. Sur 1500 personnes interrogées à Paris, 915 affirment la même chose. Est-ce que les résultats permettent de soutenir que les Lyonnais et les Parisiens utilisent autant les transports en commun, au risque de première espèce de 5%.

On souhaite savoir si la proportion d'individus empruntant les transports en commun est la même à Lyon et à Paris. Ce problème revient à tester l'indépendance de deux variables qualitatives *Ville* et *Utilisation des transports en commun* prenant chacune deux modalités. Nous cherchons donc à tester l'hypothèse :

- H_0 : les deux variables sont indépendantes. Ce qui revient au même de dire que la proportions d'individus empruntant les transports en commun est la même à Lyon et à Paris.
- contre H_1 : les deux variables sont corrélées. Il y a bien une dépendance entre les deux variables.

On commence par traduire les données sous la forme d'une table de contingence :

	Lyon	Paris	Total
Oui	1040	915	1955
Non	960	585	1545
Total	2000	1500	3500

Pour tester cette indépendance, il nous faut donc comparer les effectifs théoriques sous l'hypothèse H_0 aux effectifs observés dans notre échantillon. Sous cette même hypothèse, la statistique de test d^2 définie par :

$$d^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}},$$

où $n_{.j} = \sum_{i=1}^{k} n_{ij}$ et $n_{i.} = \sum_{j=1}^{r} n_{ij}$

suit une loi du $\mathcal{X}^2$ à $(k-1)(r-1)$ degrés de liberté lorsque les effectifs sont de tailles suffisantes soit plus grand que 5. On sera amené à rejeter l'hypothèse H_0 si $d^2 > \mathcal{X}^2_{(k-1)(r-1),1-\alpha}$.

Dans notre cas, la statistique de test est $d^2 = 28.16$ et $\mathcal{X}^2_{1,1-\alpha} = 3.84$. On va donc rejeter l'hypothèse H_0.

Exercice 7 Un data scientist d'une société d'assurance est chargé d'étudier l'impact d'une campagne de publicité réalisée dans 7 régions dans lesquelles la société est déjà implantée. Pour ceci, il a extrait de la base de données, pour un certain nombre d'agents généraux de chaque région, le nombre de nouveaux clients récoltés :

Région	1	2	3	4	5	6	7
Nb d'agents généraux	9	7	7	6	7	6	6
Nb moyen de nouveaux clients	26.88	22.34	19.54	18.95	27.17	25.87	25.72
Variance du nb de nouveaux clients	13.54	12.59	12.87	13.42	13.17	12.56	12.64

L'ingénieur statisticien décide alors de réaliser une analyse de variance afin de tester si le facteur région a une influence sur le nombre de nouveaux clients récoltés. On appelle X_k^i le nombre de nouveaux clients du $i-me$ agent général de la région k. Soit n_k le nombre d'agents généraux de la région k, et K le nombre de régions ($K = 7$). Nous supposons que les variables aléatoires X_k^i sont normales, de moyenne μ_k et de variance σ.

Le tableau de l'analyse de la variance s'écrit comme suit en fonction des X_k^i

Région	1	2	3	4	5	6	7
Nb de nouveaux clients agent 1	X_1^1	X_2^1	X_3^1	X_4^1	X_5^1	X_6^1	X_7^1
Nb de nouveaux clients agent 2	X_1^2	X_2^2	X_3^2	X_4^2	X_5^2	X_6^2	X_7^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Le problème consiste donc à tester

$$H_0 : \mu_1 = ... = \mu_K = \mu \quad \text{contre} \quad H_1 : \exists 1 \leq i, j \leq K, \text{ t.q. } \mu_i \neq \mu_j.$$

Soient :

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad \bar{X} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad n = \sum_{k=1}^{K} n_k.$$

(i) Interpréter $\bar{X}_k$ et $\bar{X}$.

$\bar{X}_k$ désigne la moyenne associée au groupe k, *i.e.* le nombre moyen de nouveaux clients enregistrés par agent dans la région k.
$\bar{X}$ désigne la moyenne globale, donc sur l'ensemble de l'échantillon, *i.e.* il s'agit du nombre moyen de nouveaux clients enregistrés par agent dans l'ensemble des régions (donc indépendamment de la région).

(ii) En remarquant que $X_k^i - \bar{X} = X_K^i - \bar{X}_k + \bar{X}_k - \bar{X}$, démontrer la formule de l'analyse de variance

$$\underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X})^2}_{=V_T^2} = \underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)^2}_{=V_R^2} + \underbrace{\frac{1}{n}\sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})^2}_{=V_A^2}$$

qui représente la décomposition de la variance totale V_T^2 en la variance V_A^2 due au facteur A (Variance inter-groupe) plus la variance résiduelle V_R^2 (Variance intra-groupe).

Remarquons que l'indication conduit au développement suivant :

$$(X_k^i - \bar{X})^2 = (X_K^i - \bar{X}_k)^2 + (\bar{X}_k - \bar{X})^2 + 2(\bar{X}_k - \bar{X})(X_K^i - \bar{X}_k).$$

Nous pouvons alors réécrire la variance totale V_T^2 comme suit :

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X})^2 &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)^2 + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{X} - \bar{X}_k)^2 \\ &\quad + \frac{2}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)(\bar{X}_k - \bar{X}).\end{aligned}$$

Etudions maintenant chaque terme séparément.
Le premier terme du membre de droite représente la variance intra-groupe V_R^2.
Le deuxième terme peut être réécrit

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{X} - \bar{X}_k)^2 = \frac{1}{n}\sum_{k=1}^{K}(\bar{X} - \bar{X}_k)^2\sum_{i=1}^{n_k}1 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{X} - \bar{X}_k)^2 = V_A^2.$$

Enfin, il nous reste à montrer que le dernier terme est nul, pour cela il suffit de remarque que :

$$\frac{2}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)(\bar{X}_k - \bar{X}) = \frac{2}{n}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})\underbrace{\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)}_{\text{=0 comme somme de variables centrées}}.$$

En effet, cela se développe comme suit :

$$\frac{2}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)(\bar{X}_k - \bar{X}) = \frac{2}{n}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})\left(\sum_{i=1}^{n_k} X_k^i - n_k\bar{X}_k\right).$$

$$\frac{2}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)(\bar{X}_k - \bar{X}) = \frac{2}{n}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})\left(\sum_{i=1}^{n_k} X_k^i - n_k\frac{1}{n_k}\sum_{i=1}^{n_k} X_k^i\right).$$

$$\frac{2}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)(\bar{X}_k - \bar{X}) = \frac{2}{n}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})\underbrace{\left(\sum_{i=1}^{n_k} X_k^i - \sum_{i=1}^{n_k} X_k^i\right)}_{=0}.$$

Ce qui montre le résultat demandé.

(iii) Calculer V_T^2, V_A^2 et V_R^2

Nous disposons, pour chaque région k des valeurs moyennes $\bar{X}_k$, qui sont fournies dans le tableau. Nous avons également un échantillon de de 48 agents répartis sur 7 régions différentes.
On commence par déterminer la valeur moyenne globale de notre échantillon $\bar{X} = \frac{1}{n}\sum_{k=1}^{K} n_k \bar{X}_k = 23.93$.
On en déduit alors la valeur de la variance inter-groupe $V_A^2 = 10.31$ d'après la relation définie en (ii).
Pour déterminer la variance intra-groupe V_R^2, il suffit de remarquer que cette variance peut également être définie comme la moyenne des variances au sein des différents groupes, ce que l'on peut réécrire :

$$V_R^2 = \underbrace{\frac{1}{n}\sum_{k=1}^{K} n_k \times \left(\underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k} \left(X_k^i - \bar{X}_k\right)^2}_{\text{variance au sein du groupe } k} \right)}_{\text{moyenne des variances}}.$$

Nous disposons des informations relatives à la moyenne des variances au sein des différentes régions, nous trouvons donc $V_R^2 = 13$.
Enfin, d'après la relation $V_T^2 = V_R^2 + V_A^2$, nous avons $V_T^2 = 23.31$.

(iv) Finaliser l'analyse de variance pour juger si la campagne de publicité a eu le même impact dans toutes les régions.

Les variables aléatoires $\frac{nV_A^2}{\sigma^2}$ et $\frac{nV_T^2}{\sigma^2}$ sont distribuées selon une loi du $\mathcal{X}^2$ à respectivement $K-1=6$ et $n-K=41$ degrés de liberté (donc la variable aléatoire $\frac{nV_T^2}{\sigma^2}$ suit une loi du $\mathcal{X}^2$ à $n-1$ degrés de liberté).

On considère l'hypothèse H_0 traduisant l'égalité des moyennes, *i.e.* le nombre de nouveaux clients par agent est le même dans chaque région : $\mu_1 = ... = \mu_K$ **VS** l'hypothèse H_1 : il existe au moins une région dont la valeur moyenne est différente d'une valeur moyenne d'une autre région.

Finalement, sous cette hypothèse H_0, la statistique de test F définie par

$$F = \frac{\frac{\frac{nV_A^2}{\sigma^2}}{K-1}}{\frac{\frac{nV_R^2}{\sigma^2}}{n-K}} = \frac{\frac{V_A^2}{K-1}}{\frac{V_R^2}{n-K}} \sim F_{K-1,n-K},$$

suit une loi de Fisher à $K-1$ et $n-K$ degrés de liberté.

Dans ce cas, on rejette l'hypothèse H_0 si la statistique F est supérieure au quantile de la $F_{K-1,n-K}$ d'ordre $1-\alpha$. La valeur de ce quantile est égale à 2.33 et la valeur de la statistique de test $F = 5.42$, ce qui nous conduit à rejeter l'hypothèse H_0 : **il existe donc une région** i **pour laquelle le nombre moyen de nouveaux clients est sensiblement différent du nombre moyen de nouveaux clients dans les autres régions**.
Notez bien que ce test ne permet pas d'affirmer que toutes les moyennes sont différentes, mais simplement qu'au moins une des valeurs moyennes est différente des autres.