Master 2 SISE

CORRIGE DE L'EXAMEN : INFERENCE STATISTIQUE & ANOVA - ANCOVA

Année Universitaire 2022-2023 - février 2023 Durée 1h 30

Exercice 1 : (Barème de notation : a) 1.5 pt b) 1.5 pt c) 1.5 pt d) 1.5 pt e) 1.5 pt = 7.5 pts)

Statut vaccinal	Non-vacciné	Vacciné
Nombre de personnes testées positives	225	50
Nombre de personnes testées	300	200

a) La proportion de personnes non-vaccinées testées positives est supérieure à 70% :

Conditions d'application du test : grande taille d'échantillon $n_{NV} = 300$.

La statistique de test : $\frac{\widehat{P}_{NV} - p_{NV}}{\sqrt{\frac{p_{NV} q_{NV}}{n_{NV}}}} \rightarrow N(0\,,\,1)$

Echantillon Rhône : $n_{NV} = 300$; $\widehat{p}_{NV} = \frac{225}{300} = 75.00\%$

Risque d'erreur : $\alpha = 5\% \Rightarrow u_\alpha = u_{5\%} = 1.645$ cf. table N(0 , 1)

Hypothèses statistiques

Comparaison d'une proportion à une valeur donnée

Test Unilatéral - Risque à droite :

$$\begin{cases} H_0 : p_{NV} \leq p_0 = 70\% \\ H_1 : p_{NV} > p_0 = 70\% \text{ la proportion est supérieure à } 70\% \end{cases}$$

Statistique de test sous l'hypothèse nulle $H_0 : p_{NV} = p_0 = 70\%$:

$$u_0 = \frac{\widehat{p}_{NV} - p_0}{\sqrt{\frac{p_0 q_0}{n_{NV}}}} = \frac{0.75 - 0.70}{\sqrt{\frac{0.70 \times 0.30}{300}}} = 1.89$$

Conclusion : la valeur de la statistique de test $u_0 = 1.89 > u_{5\%} = 1.645$ est dans la zone de rejet de H_0. On peut donc conclure avec un risque d'erreur $\alpha = 5\%$ que la proportion de personnes non-vaccinées testées positives est supérieure à 70%.

b) Intervalle de confiance :

Conditions d'application du test : grand échantillon aléatoire de taille $n_{NV} = 300$.

Estimation de la proportion de personnes non-vaccinées testées positives : $\widehat{p}_{NV} = \frac{225}{300} = 75\%$.

La statistique de test : $\frac{\widehat{P}_{NV} - P}{\sqrt{\frac{p_{NV} q_{NV}}{n_{NV}}}} \rightarrow N(0,1)$

Marge d'erreur : $E = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{P}_{NV}\widehat{q}_{NV}}{n_{NV}}} = 1.96\sqrt{\frac{0.75\times 0.25}{300}} = 4.90\%$.

Fractile de la loi normale : $\alpha = 5\%$; $u_{\frac{\alpha}{2}=2.5\%} = \pm 1.96$, cf. table N(0 , 1).

Intervalle de confiance de niveau $1-\alpha = 95\%$ de $P_{NV} \in [70.10\% \; ; \; 79.90\%]$

c) Niveau de confiance : $p_{NV} \in [69.86\% \; ; \; 80.14\%]$:

Grand échantillon $n_{NV} = 300$.

Marge d'erreur : $E = u_{\frac{\alpha}{2}} \sqrt{\frac{p_{NV}q_{NV}}{n_{NV}}} = \frac{(0.8014-0.6986)}{2} = 0.0514$

$$\Rightarrow \; u_{\frac{\alpha}{2}} = \frac{E}{\sqrt{\frac{n_{NV}}{p_{NV}q_{NV}}}} = \frac{0.0514}{\sqrt{\frac{300}{0.75\times 0.25}}} = 2.055$$

cf. table N(0 , 1) : $P(U \leq 2.055) = \Phi(2.055) = 98\% = 1 - \frac{\alpha}{2} \Rightarrow \; \frac{\alpha}{2} = 2\%$

$$\Rightarrow \; \alpha = 4\% \; \Rightarrow \; 1-\alpha = 96\%.$$

On attribue à cet intervalle un niveau de confiance de $1-\alpha = 96\%$.

d) Taille minimale de l'échantillon de personnes non-vaccinées à tester :

Niveau de confiance : $1-\alpha = 95\%$

Risque d'erreur : $\alpha = 5\% \; \Rightarrow \; u_{\frac{\alpha}{2}} = u_{2.5\%} = \pm 1.96$ cf. table N(0 , 1).

Estimation ponctuelle de la proportion de personnes non-vaccinées testées positives : $\widehat{p} = 75\%$ avec $\widehat{q} = 25\%$,

Marge d'erreur : $E^* = u_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\widehat{q}}{n^*}} \leq 0.0245$

$$\Rightarrow \; n^* \geq (\frac{u_{\frac{\alpha}{2}}}{E})^2 \widehat{p}\widehat{q} = (\frac{1.96}{0.0245})^2 \times 0.75 \times 0.25 = 1200$$

Il faudrait tester $n^* \geq 1200$ personnes non-vaccinées.

Propriété : réduire de moitié la précision : $E^* = \frac{E}{2} = 0.0245 \Rightarrow n^* = 2^2 n = 4 \times 300 = 1200$

e) Comparaison de deux proportions :

Hypothèses statistiques : Comparaison de proportions - Test unilatéral risque à droite :

$$\begin{cases} \text{H}_0 : p_{NV} \leq p_V \; \Leftrightarrow \; p_{NV} - p_V \leq 0 \\ \text{H}_1 : p_{NV} > p_V \; \Leftrightarrow \; p_{NV} - p_V > 0 \end{cases}$$

Conditions d'application du test : échantillons indépendants provenant de deux populations normales - Grands échantillons $n_{NV} = 300$ et $n_V = 200$.

La statistique de test : $$\frac{(\widehat{P}_{NV} - \widehat{P}_V) - (p_{NV} - p_A)}{\sqrt{\frac{\widehat{p}_{NV}\widehat{q}_{NV}}{n_{NV}} + \frac{\widehat{p}_V\widehat{q}_V}{n_V}}} \; \rightarrow \; N(0 \; , \; 1)$$

Echantillon des non-vaccinés : $n_{NV} = 300$; $\widehat{p}_{NV} = \frac{225}{300} = 75\%$

Echantillon des vaccinés : $n_V = 200$; $\widehat{p}_V = \frac{50}{200} = 25\%$

Ecart observé : $\widehat{p}_{NV} - \widehat{p}_V = 75\% - 25\% = 50\%$

Risque d'erreur : $\alpha = 5\% \; \Rightarrow \; u_{\alpha=5\%} = u_{5\%} = 1.645$ cf. table N(0 , 1)

Statistique de test sous $H_0 : p_{NV} = p_V$

$$u_0 = \frac{(\widehat{p}_{NV} - \widehat{p}_V) - 0}{\sqrt{\frac{\widehat{p}_{NV}\widehat{q}_{NV}}{n_{NV}} + \frac{\widehat{p}_V\widehat{q}_V}{n_V}}} = \frac{0.50}{\sqrt{\frac{0.75\times0.25}{300} + \frac{0.25\times0.75}{200}}} = 12.65$$

Règle de décision et conclusion :

La valeur de la statistique de test u_0, $u_0 = 12.65 > 1.645$, est dans la zone de rejet de H_0. On peut donc considérer avec un risque d'erreur $\alpha = 5\%$ que la proportion de personnes non-vaccinées testées positives est significativement supérieure à celle des personnes vaccinés positives.

********** ◦◦◦ **********

Exercice 2 : (Barème de notation : a) 1.5 pt b) 2.5 pts c) 1.5 pt d) 2.5 pts = 8 pts)

Type	Delta	Omicron
Nombre de patients positifs	50	25
Durée moyenne de séjour observé (jours)	13.6	9.9
Variance observée de la durée de séjour	2.64	1.36

a) Test paramétrique - Comparaison d'une moyenne à une valeur donnée :

Hypothèses statistiques : Test unilatéral risque à gauche .

$$\begin{cases} H_0 : m_O \geq 10 \text{ jours} \\ H_1 : m_O < 10 \text{ jours} \end{cases}$$

Seuil de signification et conditions d'application du test : $\alpha = 5\%$, petit échantillon $n_O = 25$, la variance σ^2 de la durée de séjour en réanimation des personnes positives au variant Omicron est inconnue.

Statistique de test : $\frac{\overline{X}_{n_O} - m_O}{\frac{s^*_O}{\sqrt{n_O}}} \rightarrow T_{n_O - 1 = 24 d.d.l.}$

Statistique de test sous $H_0 : m = 10$ $\quad t_0 = \frac{\overline{X}_{n_O} - m_O}{\frac{s^*_O}{\sqrt{n_O}}} = \frac{\overline{X}_{n_O} - m_O}{\frac{s_O}{\sqrt{n_O - 1}}} = \frac{(9.9 - 10)}{\sqrt{\frac{1.36}{24}}} = -0.420$

Conclusion : fractile de la loi Student $T_{24 d.d.l.}$ (cf.table): $t_{5\%} = -1.711, t_0$ appartient à la zone de non-rejet de H_0 ($t_0 = -0.420 > t_{5\%} = -1.711$), on peut donc conclure avec risque d'erreur $\alpha = 5\%$ qu'il n'y a pas de différence significative. La durée moyenne de séjour en réanimation des personnes positives au variant Omicron n'est pas significativement inférieur à 10 jours.

b) Risque de 2ème espèce - Puissance du test :

β : la probabilité de Non-Rejet de l'hypothèse nulle $H_0 : m_0 = 9.9$ alors qu'en réalité le séjour en réanimation des patients contaminés par le variant Omicron est $H_1 : m_1 = 9.30$.

Il faut déterminer la valeur moyenne limite x^* de la région de Non-Rejet de H_0 :

$$P(\overline{X}_n \leq x*) = \alpha = 5\% \Rightarrow P(T_{24} \leq \frac{x^* - m_0}{\frac{s^*_n}{\sqrt{n}}}) = 0.05$$

$$\Rightarrow P(T_{24} \leq \frac{x^* - 9.9}{0.2380}) = 0.05 \Rightarrow \frac{(x^* - 9.9)}{0.2380} = -1.711 \text{ cf. Table de Student à n = 24 d.d.l..}$$

$$\Rightarrow x^* = 9.7308 \text{ jours}$$

Le risque de 2ème espèce :

$\beta = P[Non\,Rejet \mathrm{H}_0 / H_1 : m_1 = 9.30]$ = P$[\overline{X}_n > x^* = 9.49 / H_1 : m_1 = 9.30]$

$= P[T_{24} > \frac{(9.7308 - 9.30)}{0.2380}] = P(T_{24} > 1.8101) = 1 - P(T_{24} \leq 1.8101)$

$= 1 - F(1.8101) = 1 - 0,9586 = 4.14\%$ cf. table Student à n = 24 d.d.l..

Ainsi, dans pratiquement 4.14% des cas, on acceptera l'hypothèse selon laquelle le séjour moyen en réanimation des patients contaminés par le variant Omicron est égal à 9.9 jours alors qu'en réalité il est de 9.3 jours. La Puissance du test : $1 - \beta = P[Rejet\ \mathrm{H}_0 / H_1 : m_1 = 9.3]$ = 1 - 0.0414 = 95.86%.

c) Comparaison d'une variance à une valeur donnée :

Test unilatéral risque à droite

Données : $n_D = 50$, $s_D^2 = 2.64$, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

Hypothèses statistiques

$$\begin{cases} H_0 : \sigma_D^2 \leq 2^2 \\ H_1 : \sigma_D^2 > 2^2 \end{cases}$$

La statistique de test : $\frac{(n_D - 1) S_D^{*2}}{\sigma_D^2} \rightarrow \chi^2_{(n_D - 1 = 49\ d.d.l.)}$

Valeurs tabulées du khi-deux à 49 degrés de liberté (cf. table du khi-deux) :

$k_{\alpha=5\%} = \chi^2_{0.95\,;\,49} = 66.339$

Statistique de test sous l'hypothèse nulle H_0 : $\sigma_D^2 = 4$

$k_0 = \frac{(n_D - 1) s_D^{*2}}{\sigma_D^2} = \frac{n_D s_D^2}{\sigma_D^2} = \frac{50 \times 2.64}{4} = 33$

Conclusion : vu que $k_0 = 33 < k_{\alpha=5\%} = 66.339$, la valeur k_0 appartient à la zone de non-rejet de H_0. On peut donc conclure avec un risque d'erreur $\alpha = 5\%$, que la variance de la durée de séjour en réanimation des personnes positives au variant Delta n'est pas significativement supérieure à 4 jour, donc l'écart-type de la durée de séjour en réanimation des personnes positives au variant Delta n'est pas significativement supérieure à 2 jour.

d) Test paramétrique - Comparaison de 2 moyennes :

Hypothèses statistiques : unilatéral risque à droite

$$\begin{cases} H_0 : m_D \leq m_O \\ H_1 : m_D > m_O \end{cases}$$ la durée moyenne de séjour pour le variant Delta est plus longue que celle du variant Omicron : $m_O < m_D$

Seuil de signification et conditions d'application du test : $\alpha = 5\%$, les variances σ_D^2 et σ_O^2 sont inconnues, la taille de l'échantillon variant Delta est grande ($n_D = 50$ et la taille de l'échantillon variant Omicron est petite $n_O = 25$).

Il faut d'abord comparer les variances des durées de séjour :

Étude de la variabilité des durées de séjour en réanimation - Comparaison de variances :

Conditions d'application du test : petits échantillons indépendants provenant de deux populations normales de variances inconnues σ_D^2 et σ_O^2.

La statistique de test : $\frac{\sigma_O^2 s_D^{*2}}{\sigma_D^2 s_O^{*2}} \rightarrow F_{(n_D-1=49\,;\,n_O-1=24)}$

Solution 1 - Test d'hypothèses

Hypothèses statistiques - Test bilatéral symétrique

$$\begin{cases} \mathrm{H}_0 : \frac{\sigma_O^2}{\sigma_D^2} = 1 \Leftrightarrow \sigma_D^2 = \sigma_O^2 \text{ : égalité des variances.} \\ \mathrm{H}_1 : \frac{\sigma_O^2}{\sigma_D^2} \neq 1 \Leftrightarrow \sigma_D^2 \neq \sigma_O^2 \text{ : les variances sont différentes.} \end{cases}$$

Seuil de signification : $\alpha = 5\%$

Statistique de test sous l'hypothèse nulle H_0 d'égalité des variances $\frac{\sigma_N^2}{\sigma_A^2} = 1$:

$$f_0 = \frac{s_D^{*2}}{s_O^{*2}} = \frac{50\times24\times2.64}{49\times25\times1.36} = \frac{48\times2.64}{49\times1.36} = 1.902$$

Les valeurs critiques (cf. table de Fisher) :

$f_2 = f_{(2.5\%\,,\,49\,,\,24)} = 2.11 \quad ; \quad f_1 = f_{(97.5\%\,,\,49\,,\,24)} = \frac{1}{f_{(2.5\%\,,\,24\,,\,49)}} = \frac{1}{1.94} = 0.5155.$

Conclusion : $f_1 = 0.5155 \leq f_0 = 1.902 \leq f_2 = 2.11$ on ne rejette pas l'hypothèse H_0. On peut donc conclure, avec un seuil de signification $\alpha = 5\%$, qu'il n'y a pas de différence significative entre les variances des durées de séjour en réanimation ; on peut les supposer comme égales $\sigma_D^2 \approx \sigma_O^2$.

Solution 2 - Intervalle de confiance

Les valeurs critiques (cf. table de Fisher) :

$f_2 = f_{(2.5\%\,,\,49\,,\,24)} = 2.11 \quad ; \quad f_1 = f_{(97.5\%\,,\,49\,,\,24)} = \frac{1}{f_{(2.5\%\,,\,24\,,\,49)}} = \frac{1}{1.94} = 0.5155.$

$$f_1 \leq \frac{\sigma_O^2 s_D^{*2}}{\sigma_D^2 s_O^{*2}} \leq f_2 \Rightarrow f_1 \frac{s_O^{*2}}{s_D^{*2}} \leq \frac{\sigma_O^2}{\sigma_D^2} \leq f_2 \frac{s_O^{*2}}{s_D^{*2}} \Rightarrow 0.5155 \frac{1.36}{2.64}\frac{49}{48} \leq \frac{\sigma_O^2}{\sigma_D^2} \leq 2.11 \frac{1.36}{2.64}\frac{49}{48}$$

$$\Rightarrow 0.2711 \leq \frac{\sigma_O^2}{\sigma_D^2} \leq 1.1096$$

$$I.C._{1-\alpha=95\%}\ \frac{\sigma_O^2}{\sigma_D^2} \in [0.2711\ ;\ 1.1096]$$

Conclusion : $1 \in [0.2711\ ;\ 1.1096]$, non-rejet de l'hypothèse nulle H_0. On peut donc conclure, avec un seuil de signification $\alpha = 5\%$, qu'il n'y a pas de différence significative entre les variances des durées de séjour ; on peut les supposer comme égales $\sigma_D^2 \approx \sigma_O^2$.

Les variances des taux des deux procédés étant égales $\sigma_D^2 \approx \sigma_O^2$, on choisit donc la statistique de test suivante :

Statistique de test : $\frac{(\overline{X}_D - \overline{X}_O) - (m_D - m_O)}{s^* \sqrt{\frac{1}{n_D} + \frac{2}{n_O}}} \rightarrow T_{(n_D+n_O-2=73\ d.d.l.)}$

Estimation ponctuelle de la variance commune de $\sigma^2 = \sigma_D^2 = \sigma_O^2$:

$$s^{*2} = \frac{n_D \times s_D^2 + n_O \times s_O^2}{n_D + n_O - 2} = \frac{(50\times2.64 + 25\times1.36)}{73} = 2.274 = 1.51^2$$

Calcul de la statistique de test sous H_0 : $m_D = m_O$

$$t_0 = \frac{(13.6-9.9)-0}{1.51\sqrt{\frac{1}{50}+\frac{1}{25}}} = \frac{3.70}{1.51\sqrt{\frac{3}{50}}} = \frac{0.7}{0.3699} = 10.002$$

Fractile de la loi de Student : $t_{5\%} = 1.666$, cf. table de Student.

Conclusion du test : $t_0 = 10.002 > t_{5\%} = 1.666$ appartient à la zone de rejet de H_0. On peut donc conclure avec un risque d'erreur $\alpha = 5\%$, qu'il y a une différence significative ; les patients atteints du variant Omicron ont des durées de séjour à l'hôpital moins longues que celles des patients atteints du variant Delta, l'affirmation du chef de service est donc vraie.

********** ○○○ **********

Exercice 3: **(Barème de notation : a) 1.5 pt b) 1.5 pt c) 1.5 pt = 4.5 pts)**

a) Effet de la marque du constructeur sur le prix de la voiture :

Le modèle utilisé est un modèle d'Analyse de la variance à un facteur contrôlé "Marque du constructeur" à deux niveaux "Française Etrangère" Le modèle est de type 1 : à effets fixes, les traitements sont fixés ou contrôlés par l'expérimentateur ou le chercheur. Les conditions semblent être vérifiées : tests de normalité du prix selon les niveaux du facteur "Marque du constructeur" et le test d'homoscédasticité : La normalité du prix par niveau du facteur contrôlé : tests de Shapiro-Wilk (SAS 7 et SAS 2). Les p-value $Pr < W$ sont toutes supérieures au risque $\alpha = 5\%$ (Normalité du prix des voitures de marque étrangère (p-value = 31.90%), des voitures de marque française (p-value = 27.30%), d'où le Non-rejet de l'hypothèse nulle de normalité dans les deux cas.

- L'homoscédasticité (égalité des variances) : test de Levêne (SAS 1) dont la p-value $(Pr > F = 1.24) = 27.58\%$ est supérieure au risque $\alpha = 5\%$, d'où le Non-rejet de lhypothèse nulle d'égalité des variances des prix des voitures de marque française et étrangère. Le modèle d'ANOVA n'est pas significatif dans son ensemble. Le test de Fisher (SAS 8) dont la p-value $(Pr > F = 0.97) = 33.46\%$ est supérieure au risque $\alpha = 5\%$, d'où le non-rejet de l'hypothèse nulle d'égalité des prix selon la marque du constructeur. Il n'y a aucune disparité des prix selon la marque (SAS 9 Tukey Groupement).

b) Effet de la marque du constructeur et de la puissance fiscale sur le prix :

Il s'agit là d'un modèle d'Analyse de la variance à 2 facteurs contrôlés "Marque du constructeur" et "Puissance fiscale" avec interaction des niveaux "française Etrangère" et "4cv 5cv 6cv".

En plus des conditions selon le facteur "Marque du constructeur" (question a), il faut également vérifier la normalité du prix selon les niveaux du facteur "Puissance fiscale" ainsi que le test d'égalité des variances (homoscédasticité).

- La normalité du prix par niveau du facteur contrôlé : tests de Shapiro-Wilk (SAS 3, SAS 10 et SAS 4). Les p-value $Pr < W$ sont toutes supérieures au risque $\alpha = 5\%$ (Normalité du prix des voitures selon la puissance fiscale 4cv (p-value = 92.780%), 5cv (p-value = 59.30%) et 6cv (p-value = 70.26%), d'où le Non-rejet de l'hypothèse nulle de normalité dans les trois cas.

- L'homoscédasticité (égalité des variances) : test de Levêne (SAS 12) dont la p-value $(Pr > F = 15.79) < 0.0001$ est inférieure au risque $\alpha = 5\%$, d'où le rejet de l'hypothèse nulle d'égalité des variances. Il y a hétéroscédasticité, dans ce cas, cette condition n'est pas vérifiée.

Le modèle explicatif dANOVA à 2 facteurs contrôlés avec interaction est significatif dans son ensemble. Le test de Fisher (SAS 11) dont la p-value $(Pr > F = 27.33) < 0.0001$ est inférieure au risque $\alpha = 5\%$, d'où le rejet de l'hypothèse nulle d'égalité des prix selon la marque du constructeur et la puissance fiscale

Selon le test de Fisher (SAS 11, 2ème partie), il y a un effet de la puissance fiscale $[(Pr > F = 61.18) < 0.0001]$, un effet d'interaction $[(Pr > F = 4.96) < 0.0172]$ et à un degré moindre $\alpha = 5.8\%$, de la marque du constructeur $[(Pr > F = 4.01) < 0.0582$.

c) On ajoute une nouvelle co-variable explicative "Cylindrée de la voiture" - au modèle précédent.

Il s'agit là d'un modèle d'Analyse de la Covariance à 2 facteurs contrôlés avec interaction des niveaux des facteurs "Marque du constructeur" et "Puissance fiscale" et une variable explicative continue "Cylindrée".

Les conditions à vérifier : en plus des conditions selon le facteur "Marque du constructeur" (question a) et selon le facteur "Puissance fiscale" (question b), il faut également vérifier la normalité de la co-variable "Cylindrée" : le test de Shapiro-Wilk (SAS 6), la p-value $Pr < W = 41.68\%$ est supérieure au risque $\alpha = 5\%$ d'où le Non-rejet de l'hypothèse nulle de normalité.

Ce modèle ANCOVA avec interaction est significatif dans son ensemble. Le test de Fisher (SAS 5) dont la p-value $(Pr > F = 27.33) < 0.0001$ est inférieure au risque $\alpha = 5\%$, d'où le rejet de l'hypothèse nulle d'égalité des prix selon la marque du constructeur et la puissance fiscale. De plus, seule la puissance fiscale a un effet significatif sur le prix.