Correction du TD

Correction exercice 2

Afin de cerner les caractéristiques des personnes fréquentant les urgences d'un hôpital, on décide de réaliser une enquête. Le questionnaire suivant est administré aux 100 premières personnes se présentant au service des urgences. Le type de chacune des variables relevées à l'aide du questionnaire est présenté en dernière colonne.

Enoncé	Réponse(s)	Type de la variable
Quel âge avez-vous ? (en années)		Quantitative continue, ou discrète (car l'information est demandée en années)
Quel est votre sexe ?	1-Homme 2-Femme	Qualitative binaire
Habitez-vous dans la région ?	0- Non 1-Oui	Qualitative binaire
Comment êtes-vous arrivé ici ?	1- Seul 2-Accompagné par un ami ou un membre de la famille 3-En ambulance 4-Avec les pompiers 5-Autre	Qualitative nominale
Combien de temps avez-vous attendu avant d'être pris en charge ? (en minutes)		Quantitative continue, ou discrète (car l'information est demandée en minutes)
Combien de personnels médicaux avez-vous vu (infirmière, médecin,...) ?		Quantitative discrète
Etes-vous satisfait de l'accueil ?	1- Très satisfait 2-Satisfait 3-Moyennement satisfait 4-Pas du tout satisfait	Ordinale

Correction exercice 3

1. Rappel des éléments du texte permettant de compléter les tableaux :
   - 100 nouveaux cas de M sont dénombrés.
   - 30 cas concernent des femmes. Parmi elles, 2 décèdent dans l'année et 10 guérissent.
   - 70 cas concernent des hommes. Parmi eux, 18 décèdent dans l'année et 20 guérissent.

Les autres chiffres se déduisent par soustraction ou addition.

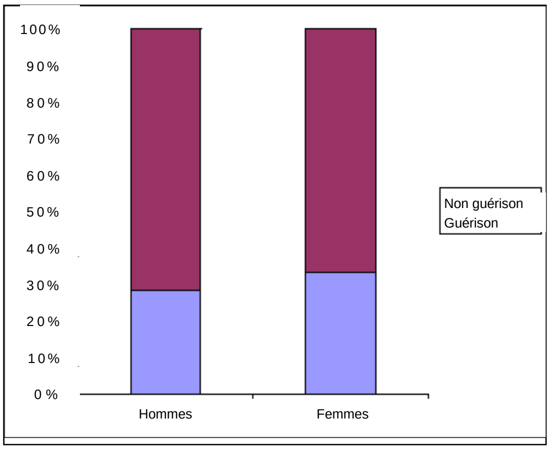
Tableau 1 : Répartition de la guérison en fonction du sexe des 100 personnes.

	Guérison	Non guérison	Total
Hommes	20	50	70
Femmes	10	20	30
Total	30	70	100

Tableau 2 : Répartition des décès en fonction du sexe des 100 personnes.

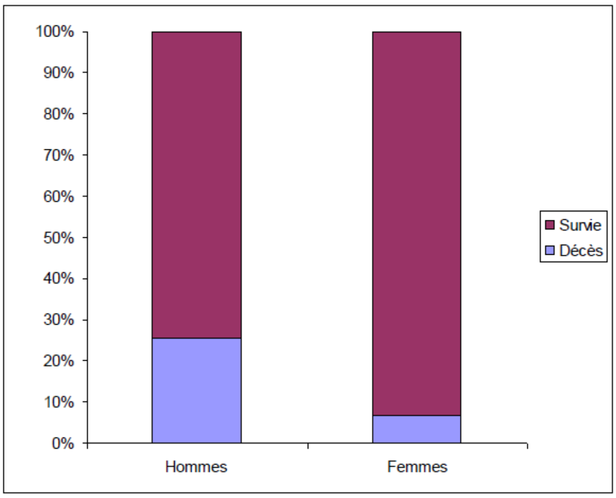
	Décès	Survie	Total
Hommes	18	52	70
Femmes	2	28	30
Total	20	80	100

2. Plusieurs graphiques sont éventuellement possibles.



Graphique 1 : Répartition selon le statut guéri ou non guéri des 70 hommes et 30 femmes

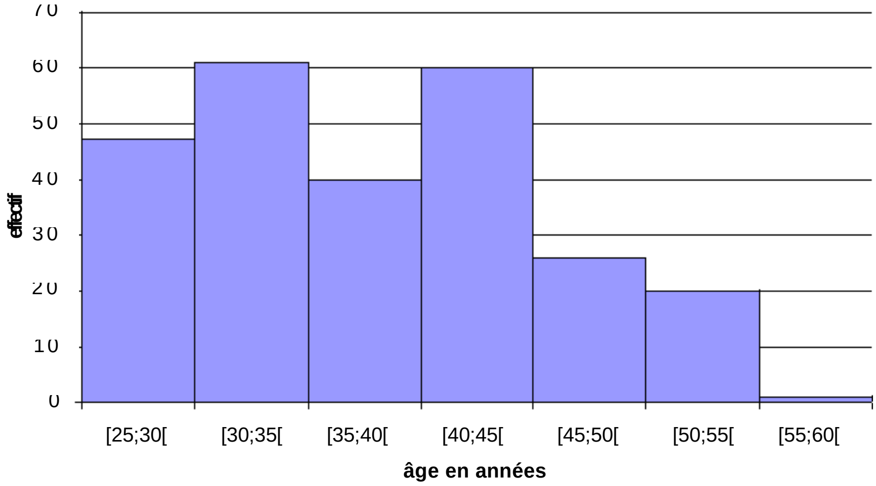
Commentaire : D'après le graphique 1, le taux de guérison ne semble pas vraiment différent selon le sexe.



Graphique 2 : Répartition selon le statut décédé ou non décédé des 70 hommes et 30 femmes

Commentaire : Le taux de décès selon le sexe semble différent avec un taux de décès plus élevé chez les hommes que chez les femmes.

Correction exercice 4



Graphique 4 : Répartition des 255 pompiers secouristes de l'étude en fonction de leur âge

2. Le mode de cette série statistique est égal à 41 ans, car il y a 23 pompiers ayant cet âge. On peut remarquer que le graphique obtenu après regroupement donne une information différente car on peut visualiser deux classes modales : [30 ; 35[et [40 ; 45[.

La médiane de la série statistique est égale à 36 ans.

Détail :

Tableau 3 : répartition des 255 pompiers secouristes selon leur âge

Age	Effectif	%	% cumulé	Age	Effectif	%	% cumulé	Age	Effectif	%	% cumulé
25	12	4,7	4,7	37	5	2,0	52,6	48	5	2,0	89,0
26	13	5,1	9,8	38	14	5,5	58,1	49	7	2,7	91,7
27	14	5,5	15,3	40	9	3,5	61,6	50	10	3,9	95,6
28	8	3,1	18,4	41	23	9,0	70,6	51	3	1,2	96,8
30	9	3,5	21,9	42	12	4,7	75,3	52	4	1,6	98,4
32	18	7,1	29,0	43	9	3,5	78,8	53	2	0,8	99,2
33	15	5,9	34,9	44	7	2,7	81,5	54	1	0,4	99,6
34	19	7,5	42,4	45	8	3,1	84,6	55	1	0,4	100,0
36	21	8,2	50,6	47	6	2,4	87,0				

La moyenne de la série statistique est égale à 37,27 ans et la variance vaut 59,26 ans².

Détail :

$$\text{Moyenne} = \frac{25\times12+26\times13+27\times14+28\times8+30\times9+\ldots+52\times4+53\times2+54\times1+55\times1}{255} = \frac{9503}{255} = 37{,}27$$

Variance =

$$\frac{\sum_{i=1}^{k} n_i x_i^2 - \frac{\left(\sum_{i=1}^{k} n_i x_i\right)^2}{N}}{N} = \frac{(25^2\times12+26^2\times13+\ldots+54^2\times1+55^2\times1)-\frac{9503^2}{255}}{255} = \frac{369257-\frac{9503^2}{255}}{255} = 59{,}26$$

(Remarque : on peut aussi considérer qu'en moyenne les personnes déclarant avoir 25 ans, par exemple, on en réalité 25,5 ans, dans ce cas la moyenne des âges est de 37,77 ans et la variance est inchangée)

Exercice 5

Notons X le nombre de lymphocytes CD4. Selon l'exercice X suit une loi normale $N(400,100^2)$.
Posons U = (X - 400)/100. U suit donc une loi normale centrée réduite N(0,1).

La probabilité pour une femme issue de la population P d'être très immunodéprimée est donc égale à la probabilité que X soit inférieure à 200/ml :

$P_1 = P(X < 200) = P(U < -2)$ donc d'après la table de la loi normale centrée réduite

$-2{,}17 < -2 < -1{,}96$ donc $P(U < -2{,}17) < P(U < -2) < P(U < -1{,}96)$

$0{,}04/2 < P(U < -2) < 0{,}05/2$

$0{,}02 < P(U < -2) < 0{,}025$

En utilisant la fonction : LOI.NORMAL (x, espérance, écart type, cumulative) sous Excel de Microsoft ou sous Calc de Open Office on obtient la valeur exacte attendue, donc ici LOI.NORMAL(200, 400,100,VRAI) et on obtient dans ce cas 0,0227.

La probabilité pour une femme issue de la population P d'avoir entre 200 et 350 CD4/ml est égale à :

$P_2 = P(200 < X < 350) = P(-2 < U < -0{,}5) = P(U < -0{,}5) - P(U < -2) = 0{,}31 - 0{,}02 = 0{,}29$

Exercice 6

On note X la pression artérielle systolique (PAS), X suit une loi normale $N(130,20^2)$. On pose U = (X-130)/20, U suit alors une loi normale centrée réduite N(0,1).

1. La probabilité d'observer une PAS strictement supérieure à 150 mmHg est égale à $Pr(X > 150) = Pr(U > 1) \approx 0{,}32/2 = 0{,}16$

2. La probabilité d'observer une PAS strictement supérieure à 121 mmHg est égale à $Pr(X > 121) = Pr(U > -0{,}45) = 1-Pr(U < -0{,}45) \approx 1-0{,}65/2 = 0{,}675$

----

3. La probabilité d'observer une PAS comprise entre 121 et 150 mmHg est égale à
$Pr(121 < X < 150) = Pr(X > 121) - Pr(X > 150) \approx 0{,}675 - 0{,}16 = 0{,}515$

4. La valeur seuil des 5 % des mesures les plus basses se détermine de la manière suivante :
Il faut déterminer la valeur de y pour laquelle $P(X < y) = 0{,}05$

=> en utilisant la transformation en loi normale centrée réduite, on obtient

$$P(U < (y-130)/20) = 0{,}05 \text{ donc } (y-130)/20 = -1{,}64 \text{ d'où } y = 97{,}2$$

Il y a donc 5% des valeurs de PAS dans cette population qui sont inférieures à 97,2mmHg.

5. La valeur seuil des 5 % des mesures les plus hautes se détermine de la manière suivante :

Il faut déterminer la valeur de y pour laquelle $P(X > y) = 0{,}05$
=> en utilisant la transformation en loi normale centrée réduite, on obtient

$$P(U > (y-130)/20) = 0{,}05 \text{ donc } (y-130)/20 = 1{,}64 \text{ d'où } y = 162{,}8$$

Il y a donc 5% des valeurs de PAS dans cette population qui sont supérieures à 162,8 mmHg.

Correction inférence

Exercice 1.

1. $m_2 = 160/50 = 3{,}2$ et $s_2^2 = (524- (160^2)/50)/49 = (12{,}5/49) = 0{,}255$ donc $s_2 = 0{,}505$

2. Comme $n>30$, l'intervalle de confiance au risque de 5% est donné par :
$IC95\%(\mu_2) = [m_2 \pm 1{,}96\ s_2/\sqrt{n}] = [3{,}2 \pm 1{,}96*0{,}505/\sqrt{50}] = [3{,}06 ; 3{,}34]$

3. Il s'agit d'un test de comparaison de deux moyennes observées avec deux effectifs supérieurs à 30 (cas grands échantillons).
On appelle X_1 le poids à la naissance d'un nouveau-né dans la population dont est issu l'échantillon E_1. X_1 a pour moyenne théorique μ_1 et pour écart-type σ_1.
On appelle X_2 le poids à la naissance d'un nouveau-né dans la population dont est issu l'échantillon E_2. X_2 a pour moyenne théorique μ_2 et pour écart-type σ_2.
A partir de deux échantillons de taille $n_1 = n_2 = 50$, on veut tester si $\mu_1 = \mu_2$.
On appelle M_1 (resp. M_2) la moyenne de la variable X_1 (resp. X_2) dans un échantillon de taille n_1 (resp. n_2). On estimera σ_1 par l'écart-type estimé s_1 et σ_2 par l'écart-type estimé s_2.

Vérification des conditions d'application du test : $n_1 \geq 30$ et $n_2 \geq 30$

Hypothèses :

H_0 : il n'y a pas de différence significative entre les deux moyennes observées ou $\mu_1 = \mu_2$

H_1 : Les deux moyennes observées sont significativement différentes ou $\mu_1 \neq \mu_2$

Choix du risque de première espèce : $\alpha = 0{,}05$

Statistique de test :

Sous H_0, on a : $U = \dfrac{M_1 - M_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \approx N(0,1)$

Région critique : $]-\infty, -1{,}96[\cup]1{,}96, +\infty[$

Calcul de u, réalisation de la statistique de test U :

$$u = \frac{2{,}9 - 3{,}2}{\sqrt{\dfrac{0{,}5^2}{50} + \dfrac{0{,}5^2}{50}}} = -3$$

Conclusion : u appartient à la région critique donc on rejette H_0. La moyenne du poids des 50 nouveaux-nés de mères non fumeuses est significativement plus élevée que la moyenne du poids des 50 nouveaux-nés de mères fumeuses (3,2 kg versus 2,9 kg avec $p < 0{,}01$).

4. La question précédente a permis de mettre en évidence une différence significative entre les deux échantillons. Une des différences existant entre les deux échantillons est la différence de statut tabagique, mais peut être que d'autres différences existent. Avant de pouvoir parler d'influence il faudrait s'assurer qu'aucune autre différence existe.

Exercice 2

Dans P, X = apports énergétiques quotidiens suit une loi $N(3500,1000^2)$.

Dans P, π_0 = proportion d'hommes souffrant de problèmes artériels = 20%.

1. Echantillon tiré au hasard de taille n = 100

On veut déterminer l'intervalle dans lequel la moyenne observée m de l'échantillon à 95% de chances de se trouver et de même pour la proportion observée f, il s'agit donc de construire des intervalles de pari.

- Intervalle de fluctuation de m : n > 30, IP95% (M) = $[\mu_0 \pm u_{0,05} * \sigma / \sqrt{n}]$ = [3304 ; 3696]

  Il y a 95% de chances que m $\in$ [3304 ; 3696].
- Intervalle de fluctuation de f : n π_0 = 20 et n(1-π_0) = 80 > 5 et

  IP95% (f) = $[\pi_0 \pm u_{0,05} * \sqrt{\pi_0(1-\pi_0)/n}]$ = [0,12 ; 0,28]

  Il y a 95% de chances que f $\in$ [12% ; 28%].

2. Test d'ajustement d'une moyenne et d'une fréquence observées.

Dans E : n = 256, m = 3650 kcal et f = 16%

Comparons l'apport énergétique quotidien moyen dans l'échantillon à celui de la population.

On appelle X la variable aléatoire apport énergétique quotidien.

On note μ sa moyenne théorique et σ son écart-type dans la population d'hommes déclarant pratiquer une activité sportive. On note μ_0 = 3500 kcal la moyenne des apports énergétiques connue dans la population. On appelle M la variable aléatoire moyenne de X sur un échantillon de taille n = 256.

Choix des hypothèses :

H_0 : Il n'y a pas de différence significative entre l'apport moyen observé et l'apport moyen théorique : $\mu = 3500$

H_1 : Il y a une différence significative entre l'apport moyen observé et l'apport moyen théorique : $\mu \neq 3500$.

Choix du risque α : α = 0,05

Statistique de test :

Sous H_0, la statistique de test $U = \frac{M - 3500}{\sigma / \sqrt{n}}$ suit une loi N(0,1).

Région critique : $]-\infty ; -1{,}96] \cup [1{,}96 ; +\infty[$

Calcul de u, réalisation de la statistique de test U :

u = 2,4

Conclusion :

U $\in$ RC, on rejette H_0 avec un risque inférieur au seuil de 5% de se tromper. Donc, l'apport énergétique moyen dans l'échantillon est significativement supérieur à celui de la population (p < 0,01).

Comparons la fréquence des problèmes artériels de l'échantillon à celle de la population.

On note :

- π_0 =0,2 la proportion des problèmes artériels dans la population
- π la proportion des problèmes artériels dans la population d'hommes déclarant pratiquer une activité sportive.
- F la variable aléatoire « fréquence des problèmes artériels dans un échantillon de taille n = 256 »

Vérification des conditions d'application : $n\pi_0$= 256*0,2 = 51,2 et $n(1 - \pi_0)$ = 204,8 > 5

Choix des hypothèses :

H_0 : Il n'y a pas de différence significative entre la fréquence observée et la proportion théorique, $\pi = 0,2$

H_1 : Il y a une différence significative entre la fréquence observée et la proportion théorique, $\pi \neq 0,2$

Choix du risque α : $\alpha = 0,05$

Statistique de test :

Sous H_0, la statistique de test $U = \dfrac{F - 0,2}{\sqrt{0,2(1-0,2)/n}}$ suit une loi N(0,1).

Région critique : $]-\infty\ ; -1,96] \cup [1,96\ ; +\infty[$

Calcul de u, réalisation de la statistique de test U :

u = - 1,6

Conclusion :

$u \notin RC$, on ne rejette pas H_0.Donc, au seuil de 5%, la fréquence des problèmes artériels dans l'échantillon n'est pas significativement différente de celle de la population.