Université de Limoges
Faculté des Sciences et Techniques
Département de Mathématiques

Mai 2023

MASTER DE MATHÉMATIQUES ET APPLICATIONS, PARCOURS ACSYON

OUTILS STATISTIQUES

Projet n°2 : Who Plays Video Games?

Rami MECHI & Cleque Marlain MBOULOU

I- Investigations

L'objectif de ce laboratoire est d'étudier les réponses des participants dans le cadre de l'étude dans le but de fournir des informations utiles aux concepteurs des nouveaux laboratoires informatiques concernant les étudiants.

Video Game Survey:

Video games

Variable	Description
time	Time spent playing video games in week prior to survey in hours.
like	Like to play video games 1=Never played, 2=Very much, 3=Somewhat, 4=Not really, 5=Not at all
where	Where play video games 1=Arcade, 2=Home on a system, 3=Home on a computer 4=Home on computer and system, 5=Arcade and Home(system or computer) 6=Arcade and home (both system and computer)
freq	How often play video games 1=Daily, 2=Weekly, 3=Monthly, 4=Semesterly
busy	Play if busy 0=no, 1=yes
educ	Is playing video games educational? 0=no, 1=yes
sex	Sex 0=female, 1=male
age	Age in years
home	Computer at home 0=no, 1=yes
math	Hate math 0=no, 1=yes
work	Hours worked for pay in the week prior to survey
own	Own PC 0=no, 1=yes
cdrom	Owned PC has a CDROM 0=no, 1=yes
email	HAve email account 0=no, 1=yes
grade	Grade expect in course 4=A, 3=B, 2=C, 1=D, 0=F

1. **Estimation de la proportion d'étudiants ayant joué à un jeu vidéo la semaine précédant l'enquête (voir code aussi) :**

   Nous commençons par fournir une estimation de la fraction d'étudiants ayant joué à un jeu vidéo la semaine précédant l'enquête.
   Nous fournissons à la fois une estimation ponctuelle et une estimation par intervalle pour cette proportion.

Estimation ponctuelle :

Pour cela, nous comptons le nombre d'étudiants ayant joué à un jeu vidéo. Autrement dit, les étudiants ayant une valeur **variables time** strictement supérieure à 0. On la note $nb\_etudiants\_joue = 34$. On a un échantillon de taille 91 donc la proportion p est égale à $\frac{34}{91}$ soit environ 37.36%.

Estimation par intervalle :

Pour fournir une estimation par intervalle, nous pouvons utiliser une approche d'estimation par intervalle de confiance. Étant donné que nous avons une grande taille d'échantillon (91 étudiants), nous pouvons utiliser l'approximation de la distribution normale pour construire un intervalle de confiance.

Supposons que nous souhaitons construire un intervalle de confiance à 95% pour la proportion d'étudiants ayant accès à internet à domicile.

Calcul de l'intervalle de confiance :

L'intervalle de confiance peut être calculé en utilisant la formule : $IC = p \pm Z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$, où Z est le score critique correspondant au niveau de confiance choisi. Pour un niveau de confiance de 95%, le score critique Z est approximativement égal à 1.96.

$$IC = 0.3736 \pm 1.96 \cdot \sqrt{\frac{0.3736 \cdot (1-0.3736)}{91}}$$

L'intervalle de confiance à 95% pour la proportion d'étudiants ayant accès à internet à domicile est donc approximativement $[0.2742318, 0.473021]$. Cela signifie que nous pouvons être confiants à 95% que la proportion réelle se situe dans cet intervalle.

2. (a) Nous vérifions comment la quantité de temps passée à jouer à des jeux vidéo la semaine précédant l'enquête se compare à la fréquence de jeu déclarée (quotidienne, hebdomadaire, etc.).
   la frequece est défine telle que :
   1=Quotidien, 2=Hebdomadaire, 3=Mensuel, 4=Semestriel, 99 = « NA » enquêtes manquantes

freq	time
1	40.0
2	71.1
3	1.0
4	1.0
99	0.0

   (b) Comment l'examen de la semaine précédant l'enquête pourrait-il affecter nos estimations précédentes et cette comparaison ?
   Changement temporaire de comportement : En raison de l'examen, les participants peuvent avoir réduit leur temps de jeu pendant cette période spécifique. Cela peut entraîner une sous-estimation du temps moyen de jeu pour la semaine précédant l'enquête.

   Influence de l'examen sur la fréquence de jeu déclarée : L'examen peut avoir modifié la routine et les habitudes des participants, ce qui peut également influencer la fréquence de jeu déclarée. Certains participants peuvent avoir joué moins fréquemment en raison de l'examen, tandis que d'autres peuvent avoir cherché un moyen de se détendre et ont augmenté leur fréquence de jeu.

3. **Estimation par intervalle de la durée moyenne passée à jouer à des jeux vidéo :**
   Nous effectuons une estimation par intervalle de la durée moyenne passée à jouer à des jeux vidéo la semaine précédant l'enquête. Nous gardons à l'esprit la forme générale de la distribution de l'échantillon. Une étude de simulation peut aider à déterminer la pertinence d'une estimation par intervalle. (**Voir code**)
   Intervalle de confiance à 95% : $[0.47975, 2.5422]$

4. **Questions sur les "attitudes" :**
   Nous examinons les questions relatives aux "attitudes". En général, pensez-vous que les étudiants aiment jouer à des jeux vidéo ?
   La variable '$like$' nous donne les propositions suivantes :

1 =**Jamais joué**, 2 =**Beaucoup**, 3 =**Un peu**, 4 =**Pas vraiment**, 5 =**Pas du tout 99 = « NA » enquêtes manquantes**
0n a la distribution suivante :

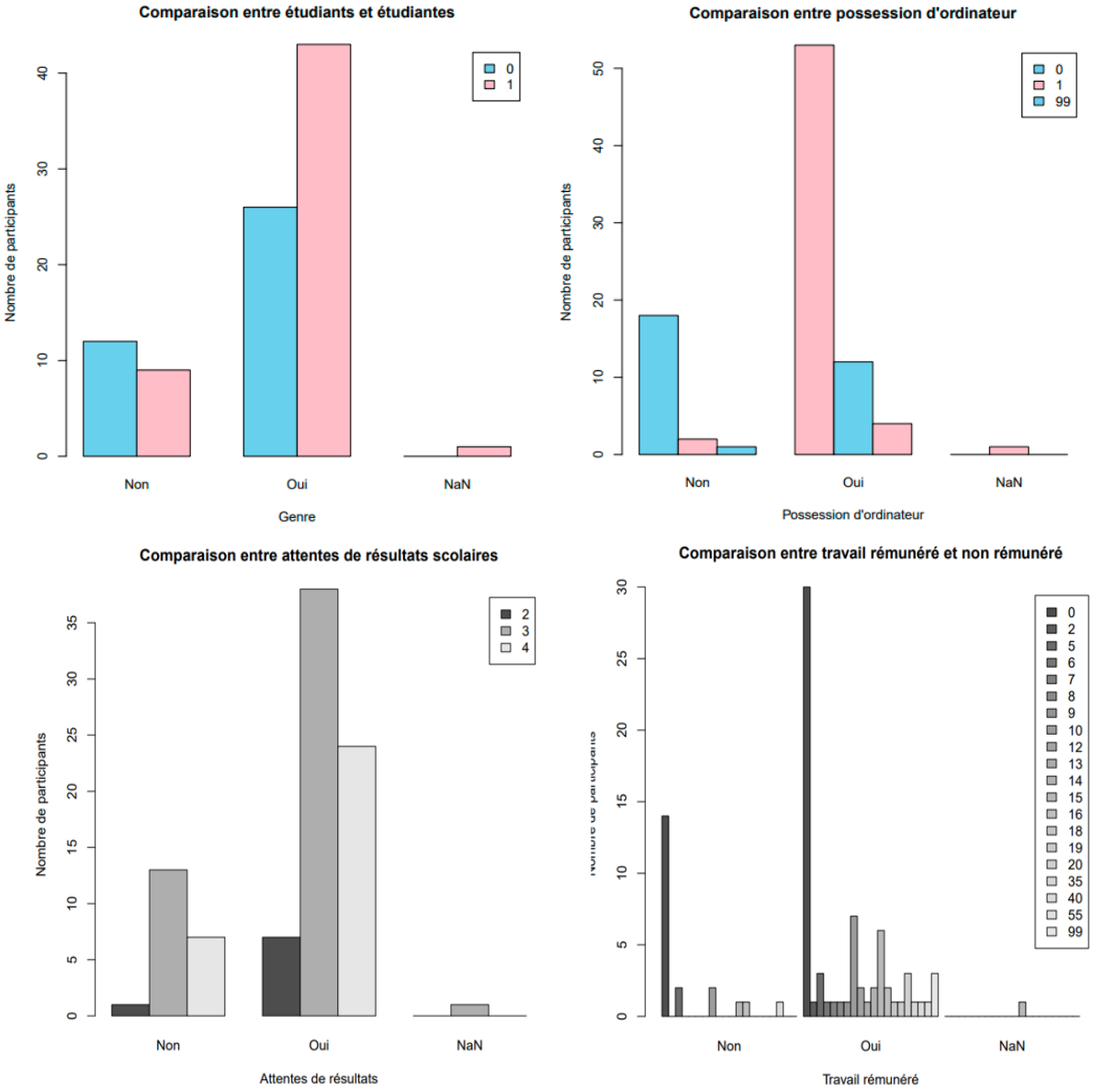
Attitudes	Fréquences
1	23
2	46
3	13
4	7
5	1
99	1

Si vous deviez dresser une liste restreinte des raisons les plus importantes pour lesquelles les étudiants aiment (ou n'aiment pas) les jeux vidéo, que mettriez-vous sur cette liste ?

Raisons pour lesquelles les étudiants aiment les jeux vidéo :
— Divertissement et plaisir : Les jeux vidéo offrent une expérience de divertissement interactive et immersive.
— Défi et compétition : Les jeux vidéo peuvent stimuler l'esprit compétitif et offrir des défis intellectuels ou stratégiques.
— Exploration et découverte : Les jeux vidéo permettent aux joueurs d'explorer des mondes virtuels et de découvrir de nouveaux environnements, histoires et personnages.
— Interaction sociale : Certains jeux vidéo offrent des fonctionnalités multijoueurs en ligne, ce qui permet aux étudiants de jouer avec leurs amis ou de rencontrer de nouvelles personnes.
— Évasion et relaxation : Les jeux vidéo peuvent servir de moyen d'évasion du stress quotidien et offrir une forme de relaxation ou de détente.

Raisons pour lesquelles les étudiants n'aiment pas les jeux vidéo :
— Manque d'intérêt : Certains étudiants peuvent tout simplement ne pas trouver les jeux vidéo intéressants ou attrayants.
— Temps passé excessif : Les jeux vidéo peuvent être chronophages, ce qui peut entraîner une diminution du temps consacré à d'autres activités ou responsabilités.
— Effets négatifs sur la santé : Certains étudiants peuvent percevoir les jeux vidéo comme ayant des effets néfastes sur leur santé physique ou mentale.
— Coût élevé : L'achat de jeux vidéo et de matériel associé peut représenter un investissement financier important.
— Stigmatisation sociale : Certains étudiants peuvent éviter les jeux vidéo en raison de la stigmatisation sociale associée à cette activité.

5. **Recherche de différences entre ceux qui aiment jouer à des jeux vidéo et ceux qui n'aiment pas :**
Pour cela, nous utilisons les questions de la dernière partie de l'enquête et comparons les étudiants de sexe masculin et féminin, ceux qui travaillent pour un salaire et ceux qui ne travaillent pas, ceux qui possèdent un ordinateur et ceux qui n'en possèdent pas, ou encore ceux qui s'attendent à avoir des A dans le cours et ceux qui ne le pensent pas. Les représentations graphiques et les tableaux croisés sont particulièrement utiles pour faire ce type de comparaisons.



6. **Étude amusante sur les notes attendues par les étudiants :**
   Nous approfondissons l'étude sur la note que les étudiants s'attendent à obtenir dans le cours. Nous examinons dans quelle mesure cela correspond à la distribution cible utilisée pour l'attribution des notes, avec 20% d'A, 30% de B, 40% de C et 10% de D ou moins.

Attentes	table_attentes	distribution_cible
A	0.00	0.2
B	0.09	0.3
C	0.57	0.4
D	0.34	0.1
Non spécifié	0.00	0.2

Si les non-répondants étaient des étudiants en échec qui ne venaient plus aux séances de discussion, cela pourrait potentiellement modifier le tableau en affectant la proportion d'étudiants dans chaque catégorie d'attente. Cependant, sans plus d'informations sur les non-répondants et leur relation avec les attentes des étudiants, il est difficile de déterminer l'ampleur exacte de cette influence.

Résumé des résultats

Nous avons mené différentes investigations dans le cadre de cette étude sur les jeux vidéo. Voici un résumé de nos principales conclusions :

— La proportion d'étudiants ayant joué à un jeu vidéo la semaine précédant l'enquête est estimée à $\frac{34}{91} \approx 0.3736$ avec un intervalle de confiance de $[0.2742318, 0.473021]$.
— Le temps passé à jouer à des jeux vidéo la semaine précédant l'enquête est généralement très élevé .
— L'estimation par intervalle de la durée moyenne passée à jouer à des jeux vidéo la semaine précédant l'enquête est $[0.47975, 2.5422]$.
— Les étudiants montrent en général une attitude positive envers les jeux divertissement et plaisir, défi et compétition, évasion et relaxation,... .
— Nous avons observé des différences significatives entre les étudiants qui aiment jouer à des jeux vidéo et ceux qui n'aiment pas en termes de genre (les hommes jouent plus au jeux vidéo) ; de possession d'ordinateur (ceux ayant un ordinateur jouent plus) ; d'attente des resultats scolaires (ceux qui s'attendent à avoir une note moyenne "B" jouent plus) ; de rémunération (ceux qui sont les mieux rémunérés jouent plus) .

Nous espérons que ces résultats fourniront des informations utiles aux concepteurs des nouveaux laboratoires informatiques.

II- Les exercices

1. Considerons la population suivante de 6 unités :

$$x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 4, x_5 = 4, x_6 = 5$$

(a) Trouvons la distribution exacte de $\overline{x}$ pour un échantilon aléatoire simple de taille 2 de cette population.
On a : $\{x_i, x_j\}$ avec $i < j$ pour $i = 1, 2, \ldots, 5$ et $j = 2, 3, \ldots, 6$, sont les populations possibles de tailles 2. Nous en avons $C_2^6 = 15$
Pour chaque échantilons $\{x_i, x_j\}$, on calcul la moyenne $\overline{x} = \dfrac{x_i + x_j}{2}$

$\{x_1, x_2\}$	$\{x_1, x_3\}$	$\{x_1, x_4\}$	$\{x_1, x_5\}$	$\{x_1, x_6\}$	$\{x_2, x_3\}$	$\{x_2, x_4\}$	$\{x_2, x_5\}$	$\{x_2, x_6\}$
1.5	1.5	2.5	2.5	3	2	3	3	3.5

$\{x_3, x_4\}$	$\{x_3, x_5\}$	$\{x_3, x_6\}$	$\{x_4, x_5\}$	$\{x_4, x_6\}$	$\{x_5, x_6\}$
3	3	3.5	4	4.5	4.5

Par conséquent, la distribution exacte de $\overline{x}$ pour un échantillon aléatoire simple de taille 2 de cette population est :

$\overline{x}$	1.5	2	2.5	3	3.5	4	4.5
$\mathbb{P}(\overline{x})$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{5}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$

(b) Utilisons cette distribution exacte pour calculer l'espérance et l'écart-type de la moyenne de l'échantillon.
Espérance : $\mathbb{E}(\overline{x}) = \sum \overline{x_i}\mathbb{P}(\overline{x} = \overline{x_i})$
$\mathbb{E}(\overline{x}) = 1.5 \times \dfrac{2}{15} + 2 \times \dfrac{1}{15} + 2.5 \times \dfrac{2}{15} + 3 \times \dfrac{5}{15} + 3.5 \times \dfrac{2}{15} + 4 \times \dfrac{1}{15} + 4.5 \times \dfrac{2}{15} = 3$

Ecart-type : $SD(\overline{x}) = \sqrt{\mathbb{V}(\overline{x})}$ et $\mathbb{V}(\overline{x}) = \mathbb{E}(\overline{x}^2) - \mathbb{E}(\overline{x})^2$
$\mathbb{E}(\overline{x}^2) = 1.5^2 \times \dfrac{2}{15} + 2^2 \times \dfrac{1}{15} + 2.5^2 \times \dfrac{2}{15} + 3^2 \times \dfrac{5}{15} + 3.5^2 \times \dfrac{2}{15} + 4^2 \times \dfrac{1}{15} + 4.5^2 \times \dfrac{2}{15} = 9.8$
Donc :

$$SD(\overline{x}) = \sqrt{9.8 - 9} = \frac{2\sqrt{5}}{5}$$

(c) Espérance et écart-type théorique de $\overline{x}$

$\mathbb{E}_{th}(\overline{x}) = \mu = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1+2+2+4+4+5}{6} = 3$

$SD(\overline{x}) = \frac{1}{\sqrt{n}}\sigma\frac{\sqrt{N-n}}{\sqrt{N-1}}$ avec $\sigma^2 = \frac{1}{N}\sum_{i=1}^{6}(x_i-\mu)^2$ et $n=2$

$\sigma^2 = 2$

$SD_{th}(\overline{x}) = \frac{1}{\sqrt{2}}2\frac{\sqrt{6-2}}{\sqrt{6-1}} = \frac{2\sqrt{5}}{5}$

En comparant nos réponses à celles trouvées en utilisant la distribution exacte de $\overline{x}$, nous constatons qu'elles sont identiques, ce qui confirme la validité des formules.

2. Considérons la population suivante de 5 unités :

$$x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 4, x_5 = 4$$

(a) Trouvons la distribution exacte de $\overline{M}$ pour un échantilon aléatoire simple de taille 3 de cette population.
On a : $\{x_i, x_j, x_k\}$ avec $i < j$ pour $i = 1, 2, \ldots, 4$, $j = 2, 3, \ldots, 5$ et $k = 3, 4$, sont les populations possibles de tailles 3. Nous en avons $C_3^5 = 10$
Pour chaque échantilons $\{x_i, x_j, x_k\}$, on calcul la médiane $\overline{M}(x_i, x_j, x_k)$

$\{x_1,x_2,x_3\}$	$\{x_1,x_2,x_4\}$	$\{x_1,x_2,x_5\}$	$\{x_1,x_3,x_4\}$	$\{x_1,x_3,x_5\}$
2	2	2	2	2

$\{x_1,x_4,x_5\}$	$\{x_2,x_3,x_4\}$	$\{x_2,x_3,x_5\}$	$\{x_2,x_4,x_5\}$	$\{x_3,x_4,x_5\}$
4	2	2	4	4

Par conséquent, la distribution exacte de $\overline{M}$ pour un échantillon aléatoire simple de taille 3 de cette population est :

$\overline{M}$	2	4
$\mathbb{P}(\overline{M})$	$\frac{7}{10}$	$\frac{3}{10}$

(b) Utilisons cette distribution exacte pour calculer l'espérance et l'écart-type de la médiane de l'échantillon.
Espérance
$\mathbb{E}(\overline{M}) = 2\times\frac{7}{10} + 4\times\frac{3}{10} = \frac{13}{5}$
Ecart-type
$\mathbb{E}(\overline{M}^2) = 4\times\frac{7}{10} + 16\times\frac{3}{10} = \frac{38}{5}$
$SD(\overline{M}) = \frac{38}{5} - \frac{13^2}{5^2} = \frac{21}{25}$

3. Pour un échantillon aléatoire simple de taille 5 dans une population de 100 sujets, notons $I(1), I(2), \cdots, I(5)$ les indices des premier, deuxième, troisième, quatrième et cinquième sujets échantillonnés. Calculons :

(a) $\mathbb{P}(I(1) = 100), \cdots, \mathbb{P}(I(5) = 100)$
D'après la page 9 du sujet, $\mathbb{P}(I(i) = 100) = \frac{1}{N-i+1}$ pour $i = 1, \cdots, 5$ et $N = 100$ dans ce cas.

$\mathbb{P}(I(1)=100)$	$\mathbb{P}(I(2)=100)$	$\mathbb{P}(I(3)=100)$	$\mathbb{P}(I(4)=100)$	$\mathbb{P}(I(5)=100)$
$\frac{1}{100}$	$\frac{1}{99}$	$\frac{1}{98}$	$\frac{1}{97}$	$\frac{1}{96}$

(b) $\mathbb{P}(\texttt{Le sujet 100 soit dans l'échantillon}) = \frac{1}{100}+\frac{1}{99}+\frac{1}{98}+\frac{1}{97}+\frac{1}{96} \simeq 0.0510$

(c) $\mathbb{E}(I(1))$. Puis que $I(1)$ peut prendre 100 valeurs avec une probabilité identique de $\frac{1}{100}$, alors :
$\mathbb{E}(I(1)) = \frac{1}{100}\sum_{i=1}^{100} i = 50.5$

(d) $\mathbb{P}(I(1) = 100 \text{ et } I(2) = 2)$. On déduit toujours de la page 9 que :
$\mathbb{P}(I(1) = 100 \text{ et } I(2) = 2) = \frac{1}{100} \times \frac{1}{99} = \frac{1}{9900}$

(e) $\mathbb{P}(I(1) = 10, I(2) = 20, I(3) = 30, I(4) = 40, \text{ et } I(5) = 50) = \Pi_{i=1}^{5} \frac{1}{100 - i + 1} = 1.106 \times 10^{-10}$

(f) $\mathbb{P}(\text{Le } 10th, 20th, 30th, 40th, \text{ et } 50th \text{ soient dans l'échantillons})$
$= (\frac{1}{100} + \frac{1}{99} + \frac{1}{98} + \frac{1}{97} + \frac{1}{96}) \times (\frac{1}{99} + \frac{1}{98} + \frac{1}{97} + \frac{1}{96}) \times (\frac{1}{98} + \frac{1}{97} + \frac{1}{96}) \times (\frac{1}{97} + \frac{1}{96}) \times \frac{1}{96}$
$\simeq 1.398 \times 10^{-8}$

(g) $\mathbb{P}(\text{Le } , 10th \text{ et } 20th, \text{ soient dans l'échantillon} = (\frac{1}{100} + \frac{1}{99} + \frac{1}{98} + \frac{1}{97} + \frac{1}{96}) \times (\frac{1}{99} + \frac{1}{98} + \frac{1}{97} + \frac{1}{96}) = 2.093 \times 10^{-3}$

4. Supposons qu'un échantillon aléatoire simple de 2 unités soit prélevé dans la population décrite dans l'exercice 1. Trouvons :

(a) $\mathbb{P}(x_{I(2)} = 5)$.
On rappel que $\mathbb{P}(x_{I(2)} = 5)$ représente la probabilité que la deuxième unité échantillonnée dans un échantillon aléatoire simple de taille 2 à partir de la population donnée dans l'exercice 1 ait une valeur de 5 (x_6). On cherche alors à determiner le nombre de couples (x_i, x_6) avec $i \neq 6$.
On a :
$\mathbb{P}(x_{I(2)} = 5) = \sum_{x_i \neq x_6} \mathbb{P}(x_{I(2)} = x_6 | x_{I(1)} \neq x_6)$
Sachant qu'on a 15 échantillons possible, donc 30 couples possible. Les couples vérifiant ce critère sont 5 : $\{(x_1, x_6), (x_2, x_6), (x_3, x_6), (x_4, x_6), (x_5, x_6)\}$. Donc :
$\mathbb{P}(x_{I(2)} = 5) = \frac{5}{30} = \frac{1}{6}$

(b) $\mathbb{E}(x_{I(1)})$
On a :

$x_{I(1)}$	1	2	4	5
$\mathbb{P}(x_{I(1)})$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

D'où :
$\mathbb{E}(x_{I(1)}) = \frac{1}{6} + 2 \times \frac{2}{6} + 4 \times \frac{2}{6} + 5 \times \frac{1}{6} = 3$

(c) $\mathbb{P}(x_{I(2)} = 2, \text{ et } x_{I(1)} = 2)$
$\mathbb{P}(x_{I(2)} = 2, \text{ et } x_{I(1)} = 2) = \mathbb{P}(x_{I(2)} = x_2, \text{ et } x_{I(1)} = x_3) + \mathbb{P}(x_{I(2)} = x_3, \text{ et } x_{I(1)} = x_2) = \frac{2}{30} = \frac{1}{15}$

5. On considère les quantités suivantes : $x_1, x_{I(1)}, \mu, I(1), N, \overline{x}$ et n. Déterminons celles qui sont des variables ou pas.
x_1 : est un valeur fixe représentant la première unité de la population.
$x_{I(1)}$: est une variable aléatoire représentant la première unité de l'échantillon.
μ :est une constante représentant la moyenne de la population.
$I(1)$: constante représentant l'indice de la première unité de l'échantillon.
N : est une constante représentant la taille de la population.
x : est une variable aléatoire représentant une unité quelconque de la population.
n : est une constante représentant la taille de l'échantillon.
En résumé, $x_{I(1)}$, $I(1)$ et x sont des variables aléatoires, tandis que μ, x_1, n et N sont des constantes.

6. Trouvons $\mathbb{E}(x_{I(1)}x_{I(2)})$.
Le produit $x_{I(1)}x_{I(2)}$ à deux valeurs possibles, 0 ou 1, car $x_{I(1)}, x_{I(2)} \in \{0, 1\}$. Puisque les variables $I(1)$ et $I(2)$ sont indépendantes. On a :

$\mathbb{P}(x_{I(1)}x_{I(2)} = 1) = \mathbb{P}(x_{I(1)} = 1, x_{I(2)} = 1) = \mathbb{P}(x_{I(1)} = 1)\mathbb{P}(x_{I(1)} = 1) = \pi \times \pi = \pi^2$.

On a :

$\mathbb{E}(x_{I(1)}x_{I(2)}) = 1 \times \mathbb{P}(x_{I(1)}x_{I(2)} = 1) + 0 \times \mathbb{P}(x_{I(1)}x_{I(2)} = 0) = \pi^2$

Déterminons la covariance de $x_{I(1)}$ et $x_{I(2)}$.
$cov(x_{I(1)}, x_{I(2)}) = \mathbb{E}((x_{I(1)} - \mathbb{E}(x_{I(1)}))(x_{I(2)} - \mathbb{E}(x_{I(2)})))$

Or $\mathbb{E}(x_{I(1)}) = 1 \times \mathbb{P}(x_{I(1))} = 1) + 0 \times \mathbb{P}(x_{I(1)}) = 0 = \pi$.

De même, $\mathbb{E}(x_{I(2)}) = \pi$, donc,

$cov(x_{I(1)}, x_{I(2)}) = \mathbb{E}((x_{I(1)} - \pi)(x_{I(2)} - \pi)) = \mathbb{E}(x_{I(1)}x_{I(2)}) - \pi(\mathbb{E}(x_{I(1)}) + \mathbb{E}(x_{I(1)})) + \pi^2$
$= \pi^2 - 2 \times \pi^2 + \pi^2 = 0$

Ainsi $cov(x_{I(1)}, x_{I(2)}) = 0$ dans ce cas.

7. Construisons un intervalle de confiance à 95% pour la proportion d'étudiants de la classe de statistiques qui possèdent un PC. L'intervalle de confiance est donné par :

the interval

$$I_\alpha = \left[f_n - z_\alpha \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}, f_n + z_\alpha \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}} \right],$$

where $f_n = \left(\sum_{i=1}^{n} x_i\right)/n$ is the observed frequency when we have observed $X_i = x_i \in \{0, 1\}$.

$f_n = \frac{67}{91} \approx 0.736$, $z_\alpha = 1.96$, et $n = 91$. Donc :

$$I = [0.736 - 1.96\frac{\sqrt{0.736(1-0.736)}}{\sqrt{91}}, 0.736 + 1.96\frac{\sqrt{0.736(1-0.736)}}{\sqrt{91}}]$$
$$I \approx [0.646, 0.827]$$

8. Trouvons un intervalle de confiance à 95% pour l'âge moyen des étudiants de la classe. La variance étant connu, l'intervalle de confiance est donné par :

$$\left[\overline{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Avec $\overline{X_n} = 19.5$, $z_\alpha = 1.96$, $\sigma = 1.85$ et $n = 91$. On a :

$$I = [19.5 - 1.96\frac{1.85}{\sqrt{91}}, 19.5 + 1.96\frac{1.85}{\sqrt{91}}]$$
$$I \approx [19.12, 19.88]$$

9. Trouvons la valeur de π qui donne la plus grande variance.
On sait que notre variable aléatoire (**'avoir ou pas un PC '**) suit une loi de Bernouilli, il s'en suit que répeter cette exprience autant de fois qu'il y ait des étudiant (notons N le nombre d'étudiants dans la classe) donne une variable aléatoire qui suit une loi Binomiale, or ici π représente la fréquence d'avoir un PC on a alors que notre variable aléatoire suit une loi Binomiale de paramètre N et π : $\mathcal{B}(N, \pi)$, il s'en suit que la variance est donnée par :

$$V(X) = N\pi(1-\pi)$$

Pour trouver alors la valeur de π qui maximise la variance, étudions la fonction suivante : $f(\pi) = \pi(1-\pi)$. On a $f^{'}(\pi) = 0 \Rightarrow \pi = \frac{1}{2}$ et on étudiant les variations de f on se rend compte qu'avant $\frac{1}{2}$ la fonction croit et après elle décroit ce qui revient à dire que f admet un maximum en $\pi = \frac{1}{2}$, on en déduit alors que $\pi = \frac{1}{2}$ donne la plus grande variance.

10. Trouvons la taille minimale de l'échantillon nécessaire pour obtenir un intervalle de confiance à 95 % pour le pourcentage de la population qui est au plus large de 4 points de pourcentage.
    (a) On a $N = 32000$, utilisons le pourcentage d'étudiants qui possèdent un PC dans l'enquête de la classe de statistiques (qu'on note $\overline{x}$) pour trouver la taille minimale de l'échantillon (qu'on note n) que nous cherchons. Pour cela on va utiliser la formule pour estimer l'erreur standard dûe à l'estimation de la proportion des étudiants ayant un PC, on a alors :

$$SE(\overline{x}) = \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}}$$

Nous voulons avoir un intervalle de confiance de 95% donc pour cela posons $SE(\overline{x}) \leq \alpha$, avec $\alpha = 5\%$ et cherchons n. On a alors :

$$n \geq \frac{N\overline{x}(1-\overline{x}) + N\alpha^2}{\overline{x}(1-\overline{x}) + N\alpha^2}$$

donc $n \geq \dfrac{32000 \times \frac{67}{91}(1-\frac{67}{91}) + 32000 \times (\frac{5}{100})^2}{\frac{67}{91}(1-\frac{67}{91}) + 32000 \times (\frac{5}{100})^2}$
on a alors $n \geq 78.48$
Il suffit alors de prendre $n = 79$ comme taille minimale de l'échantillon.

(b) Au lieu de prendre la valeur de la fréquence donnée par l'enquete de la classe de statistique, nous allons prendre la valeur qui maximise la variance et donc $\overline{x} = \dfrac{1}{2}$ calculée dans la question précedente. On a alors :

$$n \geq \frac{32000 \times \frac{1}{2}(1-\frac{1}{2}) + 32000 \times (\frac{5}{100})^2}{\frac{1}{2}(1-\frac{1}{2}) + 32000 \times (\frac{5}{100})^2}$$

on a alors $n \geq 100.68$
Il suffit alors de prendre $n = 101$ comme taille minimale de l'échantillon

11. Continuons avec l'exercice 10 et supposeons que l'enquête consiste à estimer deuxcaractéristiques des étudiants universitaires. On pense qu'une caractéristique a une prévalence d'environ 50% de la population et l'autre seulement 10%. Pourchacune des conditions suivantes, trouvons la taille d'échantillon requise

(a) Les deux estimations doivent être exactes à environ 1% (c'est-à-dire l'erreur type dechaque estimateur doit être d'au plus 1%). Pour ce faire, calculons les deux erreurs standards pour les deux estimations en prenant cette fois $\alpha = 1\%$ et trouvons la valeur de n à chaque fois. On a alors :
$n_1 \geq \dfrac{32000 \times \frac{1}{2}(1-\frac{1}{2}) + 32000 \times (\frac{1}{100})^2}{\frac{1}{2}(1-\frac{1}{2}) + 32000 \times (\frac{10}{100})^2}$
donc $n_1 \geq 2319.77$
et $n_2 \geq \dfrac{32000 \times \frac{1}{10}(1-\frac{1}{10}) + 32000 \times (\frac{10}{100})^2}{\frac{1}{10}(1-\frac{1}{10}) + 32000 \times (\frac{1}{100})^2}$
donc $n_2 \geq 876.35$
Il suffit alors de prendre $n = 2320$ comme taille requise pour l'échantillon.

(b) Dans cette question, chaque statistique doit être précise à environ $\dfrac{1}{10}$ de son paramètre de population. Pour cela, calculons les erreurs standards en prenant $\alpha = 5\%$ pour la première caractéristque et $\alpha = 1\%$ pour la deuxième, on obtient alors :
$n_1 \geq 100.68$ et $n_2 \geq 876.35$
Il suffit alors de prendre $n = 877$ comme taille requise pour l'échantillon.

12. Dans le cadre d'une enquête auprès des étudiants d'une grande université, les étudiants de troisième cycle et les étudiants de premier cycle doivent être interrogés séparément. Un échantillon aléatoire simple de 100 des 4000 étudiants diplômés. Étant donné qu'il y a 8 fois plus d'étudiants en licence, cherchons la taille de l'échantillon pour l'enquête auprès des étudiants en licence.
Supposons que les écarts types des deux groupes d'étudiants sont identiques .
Pour ce faire, prenons les deux erreurs stahdards pour les deux échantillons donnéé par la formule suivante :

$$SE(N) = N\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

avec $N = N_1 = 4000$ pour la première erreur standard pour le premier échatillon $n = n_1 = 100$) et $N = N_2 = 8 \times N_1$ pour le deuxième échantillon ($n = n_2$: ce qu'on cherche) . Pour trouver n_1 il suffit d'égaliser les deux erreurs standards **(car d'après l'énoncé les deux estimations doivent avoir la même précisions)** , ainsi on a :

$$SE(N_1) = SE(N_2) \Rightarrow N_1\frac{\sigma}{\sqrt{n_1}}\sqrt{\frac{N_1-n_1}{N_1-1}} = N_2\frac{\sigma}{\sqrt{n_2}}\sqrt{\frac{N_2-n_2}{N_2-1}}$$

$$n_2 = \frac{N_2^3}{\frac{N_1^2}{n_1}\frac{N_1-n_1}{N_1-1}(N_2-1)+N_2^2}$$

D'où :

$$n_2 = \frac{(8\times 4000)^3}{\frac{4000-100}{4000-1}(8\times 4000-1)+(8\times 4000)^2}$$

donc $n_2 \approx 5446$, on en déduit alors que pour avoir la même précision en ayant le même ecart-type il ne suffit pas de prendre 8 fois la taille du premier échantillon.

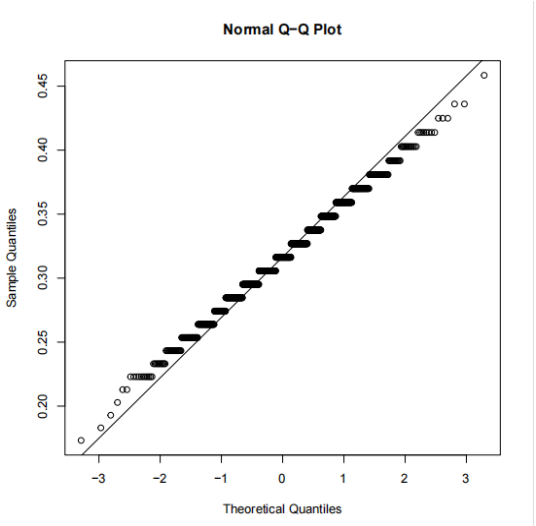
13. (a) Dans cette question notre but est de montrer que π est un estimateur non biaisé, pour cela calculons son espérance :
$E(\overline{\pi}) = E(\frac{\overline{v}}{\frac{131}{314}} = \frac{1}{\frac{131}{314}} \times E(\overline{v})$
Or notre variable aléatoire qui dit si oui ou non qu'on a affaire à la fois à une femme et qui a joué dans la semaine aux jeux vidéos, suit une loi de Bernouilli de paramètre $\overline{v}$? Soit F l'évenement : 'être une femme' et J l'évènement : 'jouer aux jeux vidéos' on a alors que :
$\overline{v} = P(F\cup J) = P_F(J)\times P(F) = \pi \times \frac{131}{314}$ Or on a $E(\overline{v} = \overline{v} = \pi \times \frac{131}{314}$
Il s'ensuit que $E(\overline{\pi}) = \frac{1}{\frac{131}{314}} \times \pi \times \frac{131}{314} = \pi$
On en déduit alors que l'estimateur $\overline{\pi}$ est un estimateur non biaisé de π.

(b) L'erreur standard de $\overline{\pi}$ est donnée par :

$$SE(\overline{\pi}) = \frac{\sqrt{\overline{\pi}(1-\overline{\pi})}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}}$$

On a alors $SE(\overline{\pi}) = \frac{\sqrt{\frac{314\overline{v}}{131}(1-\frac{314\overline{v}}{131})}}{\sqrt{91-1}}\sqrt{\frac{314-91}{314}} = 6.796\times 10^{-4}(\sqrt{(314\overline{v}(131-314\overline{v})}$

14. Voir le code R dans l'autre fichier joint.



15. Pour estimer la probabilité qu'il y ait exactement 38 étudiantes parmi les 91 étudiants de l'échantillon en utilisant une approximation par la courbe normale, nous devons appliquer une correction de continuité. Cela est nécessaire car la distribution représentée par l'histogramme de probabilité, est discrète, tandis que la courbe normale est continue.
Voir le code R (**exo15()**) dans l'autre fichier joint pour la correction de continuité et le calcul de la probabilité.
Le resultat que nous avons trouvé n'est pas similaire à celui dans le texte, on s'y attendait un peu à cause des corrections de la continuité effectuées pour le calcul de la probabilité.
16. Utilisons la méthode du bootstrap pour trouver un intervalle de confiance à 95% pour la proportion d'étudiants possédant des PC. Ensuite, nous comparerons cet intervalle de

confiance bootstrappé à l'intervalle obtenu en supposant que la proportion d'échantillon suit approximativement une distribution normale. Voir le code en utilisant la fonction **exo16()**

Intervalle de confiance bootstrappé : $[0.60975, 0.78]$
Intervalle de confiance basé sur la distribution normale : $[0.6101815, 0.7898185]$

17. Analysons chaque affirmation :
    (a) Il y a 95% de chances que le pourcentage d'étudiants de la classe STAT 21 qui s'attendent à obtenir un A dans le cours soit compris entre 26% et 42%
    Vraie. Étant donné que nous avons un intervalle de confiance à 95%, il y a 95% de chances que le vrai pourcentage d'étudiants qui s'attendent à obtenir un A dans le cours soit contenu dans l'intervalle (26%, 42%).
    (b) . Il y a 95% de chances que le pourcentage d'échantillon soit dans l'intervalle (26%, 42%)
    Fausse. Le pourcentage de l'échantillon est fixe à 34% (31 sur 91). Par conséquent, il n'y a aucune chance que le pourcentage de l'échantillon soit dans l'intervalle (26%, 42%).
    (c) Environ 95% des étudiants de STAT 21 seront inclus dans un intervalle de confiance à 95%
    Vraie. Avec un intervalle de confiance à 95%, nous sommes confiants à 95% que l'intervalle contient le vrai pourcentage de la population. Par conséquent, environ 95% des étudiants de STAT 21 seront inclus dans cet intervalle.
    (d) Sur cent intervalles de confiance à 95% pour un paramètre de population, nous nous attendons à ce que 95 d'entre eux contiennent le paramètre de population.
    Vraie. Lorsque nous construisons cent intervalles de confiance à 95%, nous nous attendons à ce que 95 d'entre eux contiennent le vrai paramètre de population, compte tenu du niveau de confiance. Cela est dû à la propriété statistique des intervalles de confiance.
    En résumé, les affirmations vraies sont : a, c et d. L'affirmation b est fausse.
18. Considerons la population de l'exercice 1 :

$$x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 4, x_5 = 4, x_6 = 5$$

Trouvons les estimations pour un échantillons de taille 2. Voir code **exo18()**

(a) $\mathbb{E}(\hat{x}) = 3$ et $V(\hat{x}) = 1.2$
(b) $\mathbb{E}(\tilde{x}) = 0$ et $V(\tilde{x}) = NAN$
(c) $\mathbb{E}(x^*) = 6$ et $V(x^*) = 4.8$

19. Pour un échantillon aléatoire simple de taille n d'une population de taille N, considérons l'estimation suivante de la moyenne de la population :

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_{I(i)},$$

où les w_i sont des poids fixes. Montrons que tous les estimateurs de l'Exercice 18 sont des cas particuliers de $\overline{x}_w$. Prouvons que pour que l'estimateur soit non biaisé, les poids doivent se sommer à 1.
En examinant les estimateurs de l'Exercice 18 :
a. $\hat{x} = x_{I(1)}$ est un cas particulier de $\overline{x}_w$ où $w_1 = 1$ et tous les autres $w_i = 0$.
b. $\tilde{x} = 2x_{I(1)} - x_{I(2)}$ est un cas particulier de $\overline{x}_w$ où $w_1 = 2$, $w_2 = -1$ et tous les autres $w_i = 0$.
c. $x^* = 2\overline{x}$ est un cas particulier de $\overline{x}_w$ où tous les w_i sont égaux à $\frac{1}{n}$.
Maintenant, pour prouver que l'estimateur $\overline{x}_w$ est non biaisé, nous devons montrer que son espérance est égale à la moyenne de la population.
L'espérance de $\overline{x}_w$ peut être calculée comme suit :

$$E(\overline{x}_w) = E\left(\sum_{i=1}^{n} w_i x_{I(i)}\right).$$

Puisque les w_i sont des poids fixes, nous pouvons les sortir de l'espérance :

$$E(\overline{x}_w) = \sum_{i=1}^{n} w_i E(x_{I(i)}).$$

Comme chaque $x_{I(i)}$ est un échantillon aléatoire simple de la population, leur espérance est égale à la moyenne de la population, $E(x_{I(i)}) = \mu$.
En utilisant cette information, nous avons :

$$E(\overline{x}_w) = \sum_{i=1}^{n} w_i \mu = \mu \sum_{i=1}^{n} w_i.$$

Pour que $\overline{x}_w$ soit non biaisé, nous devons avoir $E(\overline{x}_w) = \mu$. Cela implique que $\sum_{i=1}^{n} w_i$ doit être égal à 1.
Ainsi, pour que l'estimateur $\overline{x}_w$ soit non biaisé, les poids w_i doivent se sommer à 1.
Cela démontre que tous les estimateurs de l'Exercice 18 sont des cas particuliers de $\overline{x}_w$ et que les poids doivent se sommer à 1 pour que l'estimateur soit non biaisé.

20.

21. La covariance entre $x_{I(1)}$ et $x_{I(2)}$ est donnée par :

$$Cov(x_{I(1)}, x_{I(2)}) = Var(\bar{x})\frac{n}{n-1} - \frac{1}{n}\sigma^2$$

En utilisant le fait que $Var(\bar{x}) = 0$ ($n = N$), nous avons :

$$Cov(x_{I(1)}, x_{I(2)}) = -\frac{1}{N-1}\sigma^2$$

Maintenant en remplaçant $Cov(x_{I(1)}, x_{I(2)})$ par son expression, dans $Var(\bar{x})$ on a :

$$Var(\bar{x}) = \frac{1}{n}\sigma^2 - \frac{n-1}{n}\frac{\sigma^2}{N-1} = \frac{1}{n}\sigma^2(1 - \frac{n-1}{N-1}) = \frac{1}{n}\sigma^2\frac{N-n}{N-1}$$

Ainsi :

$$Var(\bar{x}) = \frac{1}{n}\sigma^2\frac{N-n}{N-1}$$

22. Pour montrer que lorsque les x_i dans une population sont des 0 et des 1, et que la proportion de 1's dans la population est π, alors la variance σ^2 est égale à $\pi(1-\pi)$, on peut utiliser la formule de la variance :

$$Var(x) = E(x^2) - [E(x)]^2 = \sigma^2$$

Dans notre cas, puisque les x_i sont soit 0 soit 1, nous avons :

$$E(x) = 0 \cdot \mathbb{P}(x = 0) + 1 \cdot \mathbb{P}(x = 1) = \pi$$

De plus, nous avons :

$$E(x^2) = 0^2 \cdot \mathbb{P}(x = 0) + 1^2 \cdot \mathbb{P}(x = 1) = \pi$$

Ainsi, en utilisant la formule de la variance, nous obtenons :

$$Var(x) = \pi - \pi^2 = \pi(1-\pi) = \sigma^2$$

23. on peut utiliser la formule de la variance :

$$Var(x) = E(x^2) - [E(x)]^2 = \sigma^2$$

Dans notre cas, puisque les x_i sont soit a soit b, nous avons :

$$E(x) = a \cdot \mathbb{P}(x = a) + b \cdot \mathbb{P}(x = b) = ap + b(1 - p)$$

De plus, nous avons :

$$E(x^2) = a^2 \cdot \mathbb{P}(x = a) + b^2 \cdot \mathbb{P}(x = b) = a^2p + b^2(1 - p)$$

Ainsi, en utilisant la formule de la variance, nous obtenons :

$$Var(x) = a^2p + b^2(1 - p) - (ap + b(1 - p))^2 = \sigma^2$$

Donc après réécriture, $\sigma^2 = (b - a)^2p(1 - p)$

24. Il manque la question à l'énoncé!