MAT4681 Statistique, Été 2022

Examen Final **A**

L'examen dure 3 heures, toute sortie avant le fin est autorisée, et sera définitive.

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une copie des 14 pages d'aide mémoire envoyées avant l'examen, possiblement annotée.

Dans les feuilles qui suivent, il y a 25 questions relatives au cours de statistique. Pour chaque question, cinq réponses sont proposées. Une seule est valide, et vous ne devez en retenir qu'une,

- vous gagnez 1 point par bonne réponse
- vous ne perdez pas de points pour une mauvaise réponse
- vous ne gagnez pas de point pour plusieurs réponses

Aucune justification n'est demandée.

Il est fortement recommandé de lire attentivement la question avant de chercher la réponse.

Pour toute question donnant un ordre de grandeur ("environ...") la bonne réponse sera la plus proche.

La page de réponses est la dernière page du lot que vous lisez présentement : merci de décrocher ladite feuille et de ne rendre que cette dernière, après avoir inscrit votre code permanent en haut à gauche.

Merci de cocher le carré en bleu ou en noir. En cas d'erreur, vous pouvez cocher une autre case en rouge. Seule cette dernière sera alors retenue.

Le surveillant ne répondra à aucune question durant l'épreuve : en cas de soucis sur une question (interprétation possiblement fausse, typo, etc), vous pouvez mettre un court commentaire sur la feuille de réponses.

Règlement no 18 *Tout acte de plagiat, fraude, copiage, tricherie, falsification de document ou création d'un faux document commis par une candidate, un candidat, une étudiante, un étudiant, de même que toute participation à ces actes ou tentative de les commettre, à l'occasion d'un examen, d'un travail ou d'un stage faisant l'objet d'une évaluation ou dans toute autre circonstance, constitue une infraction au sens de ce règlement.* (a. 2.1)
https://r18.uqam.ca/

1 Soit Z une variable aléatoire suivant une loi normale centrée réduite. Que vaut

$$\mathbb{P}\big[Z \in (-2.30\ ;\ 0.14)\big]$$

A) environ 0.445

B) environ 0.545

C) environ 0.645

D) environ 0.745

E) environ 0.845

$\mathbb{P}\big[Z \in (-2.30; 0.14)\big] = \mathbb{P}\big[Z \leqslant 0.14\big] - \mathbb{P}\big[Z \leqslant -2.30\big]$ soit $\Phi(0.14) - \Phi(-2.30)$, soit $0.55567 - 0.01072$.

```
> pnorm(0.14)-pnorm(-2.30)
[1] 0.5449459
```

2 Une étude est menée auprès d'un échantillon aléatoire de 2000 conducteurs québécois, pour savoir s'il y a un lien entre l'âge du conducteur et le risque d'accident. On a obtenu les statistiques suivantes

Répartition des conducteurs de l'échantillon selon l'âge et l'implication dans un accident au cours des 12 derniers mois

	Implication dans un accident		
Âge du conducteur	**Oui**	**Non**	**Total**
De 16 ans à 24 ans	14	194	208
De 25 ans à 44 ans	22	652	674
De 45 ans à 64 ans	18	782	800
65 ans et plus	6	312	318
Total	60	1 940	2 000

Source : Société de l'assurance automobile du Québec. *Bilan routier 2012*, juillet 2013.

Si les deux variables (âge du conducteur et implication dans un accident) étaient indépendantes, combien de personnes de 45 ans à 64 ans devraient être impliquées dans un accident ?

A) environ 6

B) environ 12

C) environ 18

D) environ 24

E) environ 30

Avec $i = 3$ (45-65 ans) et $j = 1$ (impliqué, oui), sous hypothèse d'indépendance, $n^{\perp}_{i,j} = \dfrac{n_{i,\cdot}.n_{\cdot,j}}{n}$ soit $\dfrac{800 \cdot 60}{2000} = 24$.
On peut le valider avec la sortie informatique suivante

```
> > M = matrix(c(14,22,18,6,194,652,782,312),4,2)
> chisq.test(M)$expected
[,1]   [,2]
[1,]  6.24 201.76
[2,] 20.22 653.78
[3,] 24.00 776.00
[4,]  9.54 308.46
```

3 Sur 1000 billes dans un roulement à billes, on observe que le poids d'une bille est de 5.02 et l'écart-type 0.30. On prend 100 billes au hasard, quelle est la probabilité que le poids total (des 100 billes) soit plus grand que 510 ?

A) moins de 0.5%

B) entre 0.5% et 1%

C) entre 1% et 5%

D) entre 5% et 20%

E) plus de 20%

Le poids total est $S = \sum_{i=1}^{100} X_i$, d'après le théorème central limite

$$Z = \sqrt{100} \cdot \frac{\overline{X} - 5.02}{0.3} = \sqrt{100} \cdot \frac{\frac{S}{100} - 5.02}{0.3} \approx \mathcal{N}(0,1) \text{ ou } S = 100 \cdot 5.02 + \sqrt{100} \cdot 0.3Z = 502 + 3Z$$

de telle sorte que

$$\mathbb{P}[S > 510] = \mathbb{P}[3Z > 8] = 1 - \Phi\left(\frac{8}{3}\right) \approx 0.38\%$$

(numériquement, on cherche $1 - \Phi(2.66667)$)

```
> 1-pnorm(8/3)
[1] 0.003830381
```

4 On lance une pièce bien équilibrée 120 fois. Quelle est la probabilité d'avoir une proportion de 'face' qui excède 5/8.

A) moins de 0.5%

B) entre 0.5% et 1%

C) entre 1% et 5%

D) entre 5% et 20%

E) plus de 20%

$S = \sum_{i=1}^{120} X_i \approx \mathcal{B}(120, 1/2)$, donc d'après le théorème central limite

$$Z = \sqrt{120} \cdot \frac{\frac{S}{120} - 0.5}{\sqrt{0.5 \cdot (1 - 0.5)}} \approx \mathcal{N}(0,1) \text{ ou } \frac{S}{120} = \frac{1}{2} + \frac{0.5}{\sqrt{120}} \cdot Z$$

donc

$$\mathbb{P}\Big(\frac{S}{120} > \frac{5}{8}\Big) = \mathbb{P}\Big(\frac{1}{2} + \frac{0.5}{\sqrt{120}} \cdot Z > \frac{5}{8}\Big) = \mathbb{P}\Big(\frac{0.5}{\sqrt{120}} \cdot Z > \frac{1}{8}\Big) = \mathbb{P}\Big(Z > \frac{\sqrt{120}}{8 \cdot 0.5}\Big) = 1 - \Phi\Big(\frac{\sqrt{120}}{4}\Big) = 1 - \Phi(2.738613)$$

```
> 1-pnorm(sqrt(120)/4)
[1] 0.00308495
> 1-pnorm(5/8,.5,.5/sqrt(120))
[1] 0.00308495
> 1-pbinom(5*120/8,120,.5)
[1] 0.002227512
> 1-pbinom(5*120/8,121,.5)
[1] 0.003075404
```

5 On interroge 100 personnes au hasard pour savoir s'ils ont voyagé hors du Québec au cours des 5 dernières années, et 55 ont affirmé que oui. Donner un intervalle de confiance à 99% pour la probabilité qu'un Québécois ait voyagé hors du Québec au cours des 5 dernières années.

A) $\left[0.55 \pm 0.082\right]$

B) $\left[0.55 \pm 0.098\right]$

C) $\left[0.55 \pm 0.116\right]$

D) $\left[0.55 \pm 0.128\right]$

E) $\left[0.55 \pm 0.154\right]$

L'intervalle de confiance de niveau α (ici 1%) est

$$\widehat{p} \pm u_{1-\alpha}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.55 \pm 2.575829 \cdot \sqrt{\frac{0.55 \cdot 0.45}{100}} = 0.55 \pm 2.575829 \cdot \frac{0.4974937}{10}$$

car $\Phi^{-1}(0.995) = 2.575829$.

```
> prop.test(55,100,.4,conf.level=.99)

  1-sample proportions test with continuity correction

data:  55 out of 100, null probability 0.4
X-squared = 8.7604, df = 1, p-value = 0.003078
alternative hypothesis: true p is not equal to 0.4
99 percent confidence interval:
 0.4179543 0.6755886
```

6 On dispose de deux échantillons d'ampoules électriques. Selon les fabricants, les ampoules de type A ont la durée de vie moyenne est 1400 heures, avec un écart-type de 120 heures; et les ampoules de type B, ont une durée de vie moyenne est de 1200 heures avec un écart-type de 80 heures. On considère deux échantillons de 125 ampoules de type A, et 125 de type B. Quelle est la probabilité que la différence entre les deux durées moyennes dépasse 240 heures.

A) moins de 0.5%

B) entre 0.5% et 1%

C) entre 1% et 2%

D) entre 2% et 5%

E) plus de 5%

On a un test de comparaison de moyenne, avec deux groupes A et B. Pour la moyenne

$$\mu_{\overline{x}_A - \overline{x}_B} = \mu_{\overline{x}_A} - \mu_{\overline{x}_B} = \mu_A - \mu_B = 1400 - 1200 - 200$$

et

$$\sigma^2_{\overline{x}_A - \overline{x}_B} = \sigma^2_{\overline{x}_A} + \sigma^2_{\overline{x}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{100^2}{125} + \frac{200^2}{125} = 20^2$$

comme on a beaucoup d'observations,

$$Z = \frac{(\overline{X}_A - \overline{X}_B) - \mu_{\overline{X}_A - \overline{X}_B}}{\sigma_{\overline{x}_A - \overline{x}_B}} = \frac{(\overline{x}_A - \overline{x}_B) - 200}{20} \approx \mathcal{N}(0,1)$$

donc $\overline{X}_A - \overline{X}_B = 200 + 20Z$ On nous demande $\mathbb{P}(\overline{X}_A - \overline{X}_B > 240)$ soit

$$\mathbb{P}\Big(200+20Z > 240\Big) = \mathbb{P}\Big(200+20Z > 240\Big) = \mathbb{P}\Big(200+20Z > 250\Big) = \mathbb{P}\Big(Z > \frac{50}{20}\Big) = 1-\Phi(2.4) \approx 0.008197536$$

7 On a demandé à 400 adultes et 600 adolescents s'ils aimaient une émission de télévision diffusée hier soir : 100 adultes et 300 adolescents ont affirmé l'avoir aimé. Donner un intervalle de confiance à 95% pour la différence de pourcentage entre les deux

A) $25\% \pm 4\% = (21\% \ ; \ 29\%)$

B) $25\% \pm 6\% = (19\% \ ; \ 31\%)$

C) $25\% \pm 8\% = (17\% \ ; \ 33\%)$

D) $25\% \pm 10\% = (15\% \ ; \ 35\%)$

E) $25\% \pm 12\% = (13\% \ ; \ 37\%)$

Comme on l'a vu en cours, considérons deux groupes A et B, dans le cas l'intervalle de confiance de niveau α pour la différence entre les proportions sera

$$\widehat{p}_A - \widehat{p}_B \pm u_{1-\alpha/2}\sqrt{\frac{\widehat{p}_A \cdot (1-\widehat{p}_A)}{n_A} + \frac{\widehat{p}_B \cdot (1-\widehat{p}_B)}{n_B}} = \frac{300}{600} - \frac{100}{400} \pm u_{1-\alpha/2}\sqrt{\frac{1}{600}\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{400}\frac{1}{4}\cdot\frac{3}{4}}$$

soit $0.25 \pm u_{1-\alpha/2} \cdot 0.02975 = 0.25 \pm 0.05832$

8 On veut estimer la proportion de personnes qui utilisent internet au quotidien sous la forme d'un intervalle de confiance à 95%. Pour cela, on a prélevé un échantillon aléatoire de taille $n = 226$ personnes, et on a observé que 196 utilisaient internet au quotidien. Déterminez la borne supérieure de l'intervalle de confiance (symétrique) à 95% de la proportion de personnes qui utilisent internet au quotidien.

A) moins de 90%

B) entre 90% et 91

C) entre 91% et 92%

D) entre 92% et 93%

E) plus de 93%

La fréquence empirique est ici $\widehat{p} = \frac{196}{226} \approx 0.8672$. L'intervalle de confiance (symétrique) à 95%, en utilisant l'approche de Wald, c'est à dire une approximation normale de la loi binomiale est (puique $u_{1-5\%/2} = \Phi^{-1}(97.5\%) = 1.96$)

$$\left[\widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \ ; \ \widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] = \left[0.8672 - 1.96\sqrt{\frac{0.8672 \cdot (1-0.8672)}{226}} \ ; \ 0.8672 - 1.96\sqrt{\frac{0.8672 \cdot (1-0.8672)}{226}}\right.$$

qui était la réponse C.

9 Un pharmacien prétend qu'une molécule est efficace dans 90% des cas pour guérir une maladie dans un délai de 8 heures. Dans un échantillon de 100 personnes qui ont testé la molécule, on note $\hat{p}$ la proportion de personnes guéries. Si on veut tester $H_0 : p = 90\%$ quelle serait la forme de la région de rejet du test, si $H_1 : p < 90\%$, pour un seuil $\alpha = 5\%$?

A) $\hat{p} < 80\%$

B) $\hat{p} < 85\%$

C) $\hat{p} < 90\%$

D) $\hat{p} > 85\%$

E) $\hat{p} > 80\%$

On a ici un test unilatéral, et comme $H_1 : p < 90\%$, on va rejeter H_0 si $\hat{p}$ est trop petit, autrement dit $\hat{p} < k$. Pour trouver le seuil k, on sait que si H_0 était vraie ($p = 90\%$), on devrait avoir $\mathbb{P}[\hat{P} < k] = 5\%$. Or, par approximation normale

$$\hat{P} \approx \mathcal{N}\left(90\%, \frac{90\% \cdot 10\%}{100}\right) \text{ soit } \mathcal{N}\left(0.9, 0.03^2\right)$$

donc, si $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}[\hat{P} < k] = \mathbb{P}\left[\frac{\hat{P} - 0.9}{0.03} < \frac{\hat{k} - 0.9}{0.03}\right] = \mathbb{P}\left[Z < \frac{\hat{k} - 0.9}{0.03}\right] = \Phi\left(\frac{\hat{k} - 0.9}{0.03}\right) = 5\% \text{ soit } \frac{\hat{k} - 0.9}{0.03} = -1.64$$

autrement dit $k = 0.9 - 1.64 \cdot 0.03$, soit 0.8508, qui correspond à la réponse B.

```
> qnorm(.05,.9,sqrt(.9*.1/100))
[1] 0.8506544
```

Les deux questions suivantes reposent sur le même énoncé : On cherche à estimer le moyenne du nombre d'heures de sommeil par patient dans une population de 500 patients traités par un nouveau type de somnifère. Pour cela, on a observé pour 50 patients, choisis aléatoirement dans cette population, le nombre d'heures de sommeil lors d'un enregistrement nocturne. On a obtenu, sur ce échantillon, une moyenne de 8 heures de sommeil, par patient, avec un écart-type (basé sur la variance corrigée) de 1 heure.

10 Quel est l'estimateur de l'écart-type de la moyenne du nombre d'heures de sommeil, par patient ?

A) moins de 0.130 heure

B)entre 0.130 et 0.145 heure

C) entre 0.145 et 0.160 heure

D) entre 0.160 et 0.175 heure

E) plus de 0.175 heure

L'écart-type est ici $\hat{s}_1 = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{50}} = 0.1414$ qui était la réponse B. On peut être un peu plus précis en utilisant les formules dans un contexte d'échantillonnage, comme cela a pu être fait lors des séances d'exercice, et on a alors une petite correction pour tenir compte du faire qu'on a un échantillon de 50 personnes sur un groupe de 500, et l'écart-type serait alors $\hat{s}_2 = \frac{1}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = \frac{1}{\sqrt{50}}\sqrt{1 - \frac{50}{500}} = 0.13416$, ce qui donne encore la réponse B. Bref, quelle que soit l'approche utilisée, on devait avoir la même réponse.

11 Déterminer la borne supérieure de l'intervalle de confiance (symétrique) à 95% pour la moyenne du nombre d'heures de sommeil, par patient, dans toute la population.

A) moins de 8.15 heure

B) entre 8.15 et 8.20 heure

C) entre 8.20 et 8.25 heure

D) entre 8.25 et 8.30 heure

E) plus de 8.30 heure

On va utiliser un intervalle de confiance Gaussien (c'est la seule chose qu'on sache faire), et l'intervalle de confiance à 95% sera alors

$$[\overline{x} - 1.96\hat{s}\ ;\ \overline{x} + 1.96\hat{s}] \text{ ou } [\overline{x} - 2\hat{s}\ ;\ \overline{x} + 2\hat{s}]$$

si on utilise le quantile de la loi de Student à $n-1=49$ degrés de liberté (on a ici un estimateur de la variance). Autrement dit, suivant les formules utilisées on a

$$\begin{cases} 8+1.96\hat{s}_1 = 8.277186 \\ 8+2\hat{s}_1 = 8.282843 \\ 8+1.96\hat{s}_2 = 8.262962 \\ 8+2\hat{s}_1 = 8.268328 \end{cases}$$

qui sont toutes associées à D, qui est la bonne réponse ici.

12 Une étude affirme que la probabilité d'avoir un "accident majeur", pour une année donnée, sur un réacteur nucléaire est de l'ordre de 0.03%. En Europe, il y a 143 réacteurs. Quel serait un ordre de grandeur raisonnable pour qu'il ait au moins un accident majeur sur un réacteur, sur une période de 30 ans (en supposant l'indépendance entre les réacteurs, et les années).

A) environ 65%

B) environ 72%

C) environ 85%

D) environ 95%

E) plus de 99%

On va utiliser ici une approximation par une loi de Poisson : le nombre d'accident suit une loi de Poisson de paramètre $\lambda = np$ avec $p = 0.03\%$ et $n = 143 \times 30$, soit 1.287. La probabilité de n'observer aucun accident sera $\exp(-1.287)$, soit 27.6%, et donc la probabilité d'avoir au moins un accident sera de 72.39%.

13 Un restaurateur se demande s'il existe une relation linéaire entre le temps passé dans l'établissement (x), et le montant de la facture (y). Sur un échantillon de $n = 35$ clients, il a observé

$$\overline{x} = 50 \text{ minutes},\ s_x = 10 \text{ minutes}, \overline{y} = 35\$,\ s_y = 10\$,$$

et comme les données sont appareillées, $r_{xy} = \text{cov}(x, y) = 55$. Il se demande si la relation linéaire observée entre le temps passé dans l'établissement (x), et le montant de la facture (y) est significative (ou pas). On note t la statistique de test. Il fait un test de niveau de confiance $\alpha = 5\%$.

A) $t \geqslant 1.5$ et on ne peut pas conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé

B) $t \leqslant 1.5$ et on ne peut pas conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé

C) $t \leqslant 2.5$ et on peut conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé

D) $t \geqslant 2.5$ et on ne peut pas conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé

E) $t \geqslant 3.5$ et on peut conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé

Mille excuses: dans l'énoncé, je notais "$r_{xy} = \text{cov}(x, y) = 55$", qui n'est pas la corrélation (comme le suggère la notation r) mais la covariance (comme indiqué). J'aurais du l'appeler s_{xy}. La (vraie) corrélation vaut ici

$$R = \frac{s_{xy}}{s_x \cdot s_y} = \frac{55}{10 \cdot 10} = 0.55$$

(on sait que s_x est l'écart-type (1) parce que c'est la notation du cours, (2) on nous dit que l'unité est la même que pour $\overline{x}$ =). On peut alors utiliser la statistique de test

$$T = R\sqrt{\frac{n-2}{1-R^2}} = 0.55 \cdot \sqrt{\frac{35-2}{1-0.55^2}} = 3.783$$

qui soit suivre, sous l'hypothèse $H_0 : r = 0$ une de Student à $n-2$ degrés de liberté, donc si $|t| > 2$, on rejette H_0, autrement dit, on peut conclure qu'il y a une relation linéaire entre le temps passé et le montant dépensé. C'est la réponse E.

14 Afin de tester l'efficacité d'un sérum, on constitue un échantillon de 200 personnes : 100 se voient prescrire le sérum (groupe A) et 100 se voient prescrire un placebo (groupe B)

Frequencies Observed

	Recover	Do Not Recover	TOTAL
Group *A* (using serum)	75	25	100
Group *B* (not using serum)	65	35	100
TOTAL	140	60	200

On cherche à tester H_0, le serum n'a pas d'effet (pas davantage qu'un placebo) par un test du chi-deux. Que vaut la statistique de test ?

A) moins de 2.5

B) entre 2.5 et 3.5

C) entre 3.5 et 4.5

D) entre 4.5 et 5.5

E) plus de 5.5

Si les variables étaient indépendantes, on devrait avoir les comptage suivants,

$$\boldsymbol{N}^{\perp} = \begin{pmatrix} \frac{140 \cdot 100}{200} & \frac{60 \cdot 100}{200} \\ \frac{140 \cdot 100}{200} & \frac{60 \cdot 100}{200} \end{pmatrix} = \begin{pmatrix} 70 & 30 \\ 70 & 30 \end{pmatrix}$$

Et la statistique du chi-deux

$$Q = \frac{(75-70)^2}{70} + \frac{(35-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} = 2.38$$

15 Afin de tester l'efficacité d'un sérum, on constitue un échantillon de 200 personnes : 100 se voient prescrire le sérum (groupe A) et 100 se voient prescrire un placebo (groupe B). On obtient le tableau de contingence suivant, avec $x \in \{0, 1, \cdots, 30\}$

Frequencies Observed

	Recover	Do Not Recover	TOTAL
Group *A* (using serum)	**70 + *x***		100
Group *B* (not using serum)			100
TOTAL	140	60	200

On cherchait à tester H_0 correspondant à l'hypothèse que le serum n'a pas d'effet (pas davantage qu'un placebo) par un test du chi-deux. On sait que H_0 est rejeté au seuil $\alpha = 5\%$. Que vaut x ?

A) 3 ou moins

B) 7 ou moins

C) 10 ou moins

D) 3 ou plus

E) 7 ou plus

Comme pour la question précédente (les totaux par ligne et par colonne sont inchangés),

$$\boldsymbol{N}^{\perp} = \begin{pmatrix} \frac{140 \cdot 100}{200} & \frac{60 \cdot 100}{200} \\ \frac{140 \cdot 100}{200} & \frac{60 \cdot 100}{200} \end{pmatrix} = \begin{pmatrix} 70 & 30 \\ 70 & 30 \end{pmatrix} \text{ alors que } \boldsymbol{N} = \begin{pmatrix} 70+x & 30-x \\ 70-x & 30+x \end{pmatrix}$$

et la statistique du chi-deux vaut

$$Q = \frac{(70+x-70)^2}{70} + \frac{(30-x-30)^2}{30} + \frac{(70-x-70)^2}{70} + \frac{(30+x-30)^2}{30} = \frac{2x^2}{70} + \frac{2x^2}{30} = \frac{2x^2}{21}$$

La limite pour rejeter H_0 est le quantile de la loi du $\chi^2(1)$ de niveau 95%, soit 3.8414, donc

$$\frac{2x^2}{21} > 3.8414 = 1.96^2 \text{ soit } x > \sqrt{\frac{21}{2} \cdot 3.8414} = \sqrt{40.335} = 6.35^2$$

donc x vaut 7 ou plus.

16 On dispose des observations suivantes, supposées suivre une distribution $\mathcal{U}([-\theta, \theta])$, avec $\theta > 0$,

$$\boldsymbol{x} = \{-0.94, -0.51, 0.29, 1.63, -1.19, 1.59\}$$

Que vaut l'estimateur $\widehat{\theta}$ obtenu par la méthode des moments ?

A) moins de 1.7

B) entre 1.7 et 1.95

C) entre 1.95 et 2.25

D) entre 2.25 et 2.4

E) plus de 2.4

Comme $\mathbb{E}[X] = 0$ on utilise la variance, en notant que $\text{Var}[X] = \frac{(2\theta)^2}{12}$ et donc

$$\hat{\theta} = \frac{1}{2}\sqrt{12\text{Var}(\boldsymbol{x})}$$

mais on peut note que $\text{Var}[X] = \mathbb{E}[X^2]$, et on peut utiliser la moyenne des x_i^2,

$$\hat{\theta} = \frac{1}{2}\sqrt{12 \cdot \frac{0.94^2 + 0.51^2 + 0.29^2 + 1.63^2 + 1.19^2 + 1.59^2}{6}} = 1.978497$$

17 En faisant un calcul rapide, un étudiant propose un intervalle de confiance pour la moyenne de la forme $[-2.053749; 2.053749]$, en supposant la variance connue. En se souvenant qu'il n'avait que 30 observations, il envisage une correction, pour tenir compte du fait que la variance est inconnue. Que deviendra alors l'intervalle de confiance (pour le même niveau de confiance) ?

A) $[-2.04\ ;\ 2.04]$

B) $[-2.15\ ;\ 2.15]$

C) $[-2.31\ ;\ 2.31]$

D) $[-2.46\ ;\ 2.46]$

E) $[-2.62\ ;\ 2.62]$

$[-2.053749; 2.053749]$ peut s'écrire $0 \pm u_{1-\alpha/2}\dfrac{\sigma}{\sqrt{30}}$.

18 Une étude menée auprès d'un échantillon de 450 hommes et 500 femmes indique que $\widehat{p}_H = 17\%$ des hommes et $\widehat{p}_F = 13\%$ des femmes dorment moins de 6.5 heures par nuit. Au seuil de signification de 5%, quelle différence minimale entre les deux proportions doit-on avoir pour conclure que le pourcentage de personnes qui dorment moins de 6.5 heures par nuit est plus élevé chez les hommes que chez les femmes ?

A) $\widehat{p}_H - \widehat{p}_F > 1.2\%$

B) $\widehat{p}_H - \widehat{p}_F > 2.6\%$

C) $\widehat{p}_H - \widehat{p}_F > 3.2\%$

D) $\widehat{p}_H - \widehat{p}_F > 3.8\%$

E) $\widehat{p}_H - \widehat{p}_F > 4.6\%$

La différence des proportions suit une loi normale, centrée si H_0 est vraie,

$$\hat{p} = \frac{1}{450+500}\Big(450 \cdot 17\% + 500 \cdot 13\%\Big) = \frac{76.5+65}{950} = 14.9\%$$

et la est variance

$$\sigma^2_{\overline{x}_A - \overline{x}_B} = \sigma^2_{\overline{x}_A} + \sigma^2_{\overline{x}_B} = \left(\frac{1}{450} + \frac{1}{500}\right)\hat{p}(1-\hat{p}) = (2.3\%)^2$$

L'écart maximal toléré est $1.645 \cdot 2.3\% = 3.8\%$.

19 Lors d'un test qualité, on a interrogé 200 personnes pour savoir s'ils préféraient les croustilles A ou les croustilles B. On a eu un peu plus de personne qui ont préféré B, de telle sorte qu'on a le comptage suivant, avec $x \in \{0, 1, 2, \cdots, 100\}$,

A	B
100-x	100+x

Un test statistique (avec un seuil de 5%) n'a pas permis de les départager. Quelle est la valeur maximale de x ?

A) $x \leqslant 8$

B) $x \leqslant 14$

C) $x \leqslant 18$

D) $x \leqslant 22$

E) $x \leqslant 26$

On peut tenter deux approches ici. La première consiste à utiliser un test de proportion classique (avec une approximation normale), ou tout simplement un intervalle de confiance. La seconde consiste à utiliser un test du chi-deux (c'est un peu tordu, mais c'est ce qu'on fera dans la question suivante).

Commençons par le faire en R, juste pour voir

```
> prop.test(92,200)

  1-sample proportions test with
  continuity correction

data:  92 out of 200, null probability 0.5
X-squared = 1.125, df = 1,
p-value = 0.2888
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3899055 0.5316562
 > prop.test(114,200)

  1-sample proportions test with
  continuity correction

data:  114 out of 200, null probability 0.5
X-squared = 3.645, df = 1,
p-value = 0.05624
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4982053 0.6390612
```

Autrement dit, si $x = 14$, on a une p-value qui atteint 5%, c'est donc a priori la réponse B. Maintenant pour le faire, on peut construire un intervalle de confiance à 95%, et chercher le premier moment à 50% n'est plus dedans. Ici, l'intervalle de confiance est

$$\left[\frac{100+x}{200} + 1.96\sqrt{\frac{(100-x)(100+x)}{200^3}}\right]$$

Par exemple si $x = 8$, on a

$$\left[\frac{108}{200} + 1.96\sqrt{\frac{92\times 10)}{200^3}}\right] = \left[0.4709\ ;\ 0.6090\right]$$

qui correspond quasiment au premier intervalle de confiance obtenu avec R. On notera qu'il faudra regarder ici la borne inférieure, et voir quand on va dépasser 50%. Pour $x = 14$

$$\left[\frac{114}{200} + 1.96\sqrt{\frac{114\times 86}{200^3}}\right] = \left[0.50138\ ;\ 0.6386\right]$$

Bref, on retrouve la réponse B.

Sous H_0, on espérerait avoir 100 observations dans chaque groupe, et la statistique du χ^2 s'écrit

$$Q = \frac{(100-x-100)^2}{100} + \frac{(100+x-100)^2}{100} = \frac{2\cdot x^2}{100} = \frac{1}{50}x^2$$

et la limite est le quantile de niveau 95% à 2-1=1 degrés de liberté, soit 3.84, aussi

$$\frac{1}{50}x^2 \leqslant 3.84 \text{ ou } x^2 \leqslant 192.0729 \text{ ou } x \leqslant 13.85904$$

qui correspond à la réponse B.

20 Après avoir lancé 600 fois un dé, on a obtenu le comptage suivant. Les deux premiers comptages manquent (pour 1 et 2), mais on sait qu'il y a eu plus de 2 que de 1. Et comme la somme vaut 200, on sait qu'on peut écrire $100-x$ et $100+x$ respectivement avec $x \in \{0, 1, 2, \cdots, 100\}$.

1	2	3	4	5	6
100-x	100+x	95	101	98	106

Un test du chi-deux, avec un seuil de 5% conclut que le dé est équilibré. Quelle est la valeur maximale de x ?

A) $x \leqslant 8$

B) $x \leqslant 14$

C) $x \leqslant 18$

D) $x \leqslant 22$

E) $x \leqslant 26$

Sous H_0, on espérerait avoir 100 observations de chaque face, et la statistique du χ^2 s'écrit

$$Q = \frac{(100-x-100)^2}{100} + \frac{(100+x-100)^2}{100} + \frac{(95-100)^2}{100} + \frac{(101-100)^2}{100} + \frac{(98-100)^2}{100} + \frac{(106-100)^2}{100} = \frac{2\cdot x^2 + 5^2 + 1 + 2^2 +}{100}$$

et la limite est le quantile de niveau 95% à 6-1=5 degrés de liberté, soit 11.07, aussi

$$0.66 + \frac{1}{50}x^2 \leqslant 11.0705 \text{ ou } x^2 \leqslant 520.525 \text{ ou } x \leqslant 22.815$$

21 On prélève un échantillon de 40 enfants afin d'estimer le nombre de moyen μ de caries dentaires parmi les 1300 élèves d'une école. Les résultats de l'examen dentaire sont présentés dans le tableau suivant

Nombre de caries	0	1	2	3	4	Total
Effectifs (nombre d'enfants)	28	7	2	2	1	40

Déterminez un intervalle de confiance à 95% pour μ.

A) [0.17 ; 0.87]

B) [0.21 ; 0.83]

C) [0.33 ; 0.72]

D) [0.47 ; 0.57]

E) [0.51 ; 0.54]

On va utiliser une approximation normale ici (comme toujours). On va commencer par calculer la moyenne,

$$\frac{1\cdot 7 + 2\cdot 2 + 3\cdot 2 + 4\cdot 1}{40} = \frac{21}{40} = 0.525$$

et la variance, en calculant la moyenne du carré

$$\frac{1^2 \cdot 7 + 2^2 \cdot 2 + 3^2 \cdot 2 + 4^2 \cdot 1}{40} = \frac{49}{40} = 1.225$$

puis en utilisant $s^2 = 1.225 - 0.525^2 = 0.949$. Un intervalle de confiance à 95% sera alors de la forme (faisons simple)

$$\left[0.525 - 2\sqrt{0.94940}\ ;\ 0.525 + 2\sqrt{0.94940}\right] = [0.216\ ; 0.833]$$

Notons qu'on pourrait aussi utiliser (comme cela a pu être fait en séance d'exercices) une petite correction pour tenir compte de l'erreur d'échantillonnage (mais ce n'était pas l'objet du cours)

$$\left[0.525 - 2\sqrt{0.94940}\sqrt{1 - \frac{40}{1300}}\ ;\ 0.525 + 2\sqrt{0.94940}\sqrt{1 - \frac{40}{1300}}\right] = [0.221\ ;\ 0.828]$$

qui est encore très proche de B

Les deux questions suivantes portent sur le même énoncé: On cherche à déterminer le poids d'athlètes à partir de leur taille. Sur un échantillon de $n = 38$ sportifs, on obtient les statistiques suivantes, pour la taille (x, en centimètres) et pour le poids (y, en kilogrammes),

$$\overline{x} = 183 \text{ cm},\ s_x = 8 \text{ cm}, \overline{y} = 70 \text{ kg},\ s_y = 3 \text{ kg},$$

et comme les données sont appareillées, $r_{xy} = \text{cov}(x, y) = 22.8$. On note $y = a + bx$ l'équation de la droite de régression obtenue par moindres carrés.

22 Que vaut b ?

A) moins de 0.35

B) entre 0.35 et 0.40

C) entre 0.40 et 0.45

D) entre 0.45 et 0.50

E) plus de 0.50

La encore, j'ai noté r la covariance, mille excuses. L'estimateur de la pente par moindres carrés est

$$b = \frac{s_{xy}}{s_x^2} = \frac{22.8}{8^2} = 0.35625$$

23 Que vaut l'écart-type de b ?

A) moins de 0.018

B) entre 0.018 et 0.020

C) entre 0.020 et 0.022

D) entre 0.022 et 0.024

E) plus de 0.024

Pour calculer l'écart-type de b, notons que la corrélation vaut ici $R = \frac{s_{xy}}{s_x s_y} = \frac{22.8}{8 \cdot 3} = 0.95$, donc

$$\hat{s}_b = \frac{s_y}{s_x}\sqrt{\frac{1 - R^2}{n - 2}} = \frac{3}{8}\sqrt{\frac{1 - 0.95^2}{38 - 2}} = 0.0195$$

qui correspond à la réponse B.

Les deux questions suivantes portent sur le même énoncé: Un sondage a été mené auprès de 1000 travailleurs québécois. On leur a demandé leur opinion quant à une diminution de dépenses publiques dans les programmes d'aide sociale. On leur a aussi demandé s'ils étaient syndiqués, ou pas, et les données ont été résumées dans le tableau de comptage suivant

	Pour une diminution des dépenses			
	En accord	Indifférent	En désaccord	Total
Syndiqué	200	160	150	510
Non syndiqué	170	140	180	490
Total	370	300	330	1000

24 Que vaut la statistique du chi-deux du test d'indépendance entre l'opinion quant à une baisse des dépenses publiques et le fait d'être syndiqué

A) moins de 3.5

B) entre 3.5 et 4.5

C) entre 4.5 et 5.5

D) entre 5.5 et 6.5

E) plus de 6.5

Pour faire le test du chi-deux, il faut commencer par calculer les nombres qu'on s'attendrait à observer si les données étaient indépendantes, soit $n^\perp$ avec les notations du cours. Rappelons que

$$N = \begin{pmatrix} 200 & 160 & 150 \\ 170 & 140 & 180 \end{pmatrix} \text{ et } N^\perp = \frac{1}{1000}\begin{pmatrix} 510\cdot 370 & 510\cdot 300 & 510\cdot 330 \\ 490\cdot 370 & 490\cdot 300 & 490\cdot 330 \end{pmatrix} = \begin{pmatrix} 188.7 & 153 & 168.3 \\ 181.3 & 147 & 161.7 \end{pmatrix}$$

et la statistique du chi-deux est alors

$$Q = \frac{(200-188.7)^2}{188.7} + \frac{(160-153)^2}{153} + \cdots$$

et on obtient un peu plus de 6 (pour être précis 6.095), qui était la réponse D.

25 Que vaut la p-value du test

A) environ 0.5%

B) environ 1%

C) environ 2%

D) environ 5%

E) environ 8%

La p-value, pour le test du chi-deux, correspond à la probabilité qu'une loi du chi-deux à $(3-1)\times(2-1) = 2$ degrés de liberté dépasse 6. Or la table du chi-deux nous dit qu'il s'agit du quantile de niveau 95% donc la p-value est de l'ordre de 5%.

```
> 1-pchisq(6,2)
[1] 0.04978707
```

On peut d'ailleurs faire le test du chi-deux sous R

```
> chisq.test(matrix(c(200,170,160,140,150,180),2,3))

  Pearson's Chi-squared test

data:  N
X-squared = 6.0955, df = 2, p-value = 0.04747
```

Code permanent :

Sujet : A

question 1	☐ A	☐ B	☐ C	☐ D	☐ E
question 2	☐ A	☐ B	☐ C	☐ D	☐ E
question 3	☐ A	☐ B	☐ C	☐ D	☐ E
question 4	☐ A	☐ B	☐ C	☐ D	☐ E
question 5	☐ A	☐ B	☐ C	☐ D	☐ E
question 6	☐ A	☐ B	☐ C	☐ D	☐ E
question 7	☐ A	☐ B	☐ C	☐ D	☐ E
question 8	☐ A	☐ B	☐ C	☐ D	☐ E
question 9	☐ A	☐ B	☐ C	☐ D	☐ E
question 10	☐ A	☐ B	☐ C	☐ D	☐ E
question 11	☐ A	☐ B	☐ C	☐ D	☐ E
question 12	☐ A	☐ B	☐ C	☐ D	☐ E
question 13	☐ A	☐ B	☐ C	☐ D	☐ E
question 14	☐ A	☐ B	☐ C	☐ D	☐ E
question 15	☐ A	☐ B	☐ C	☐ D	☐ E
question 16	☐ A	☐ B	☐ C	☐ D	☐ E
question 17	☐ A	☐ B	☐ C	☐ D	☐ E
question 18	☐ A	☐ B	☐ C	☐ D	☐ E
question 19	☐ A	☐ B	☐ C	☐ D	☐ E
question 20	☐ A	☐ B	☐ C	☐ D	☐ E
question 21	☐ A	☐ B	☐ C	☐ D	☐ E
question 22	☐ A	☐ B	☐ C	☐ D	☐ E
question 23	☐ A	☐ B	☐ C	☐ D	☐ E
question 24	☐ A	☐ B	☐ C	☐ D	☐ E
question 25	☐ A	☐ B	☐ C	☐ D	☐ E