Statistique descriptive

Corrigé de l'examen blanc de décembre 2011

Question 1 (13 points)

Y = Variable d'intérêt = taux de réussite en bachelier dans les universités francophones

1ère partie de la question :

N = 50.000 ; n = 200 ; nombre de succès (réussite) = 110

a. $f = \frac{n}{N} = \frac{200}{50000} = 0.004$

b. Dans le cadre de l'estimation d'une proportion : ME = $2 \times \sqrt{(1-f)\frac{p(1-p)}{n}}$

- Si on n'a pas d'information *a priori* sur p, on prend le cas le plus défavorable (p = 0.5 → ME max). Ici, on sait qu'avec 200 étudiants sondés, p = 0.55 (110/200). Toutefois, cette proportion pourrait changer et se rapprocher encore de 0.5 avec plus ou moins d'individus sondés → dans le doute, on prend p = 0.5 comme référence.
- Si $(1-f) \to 1$ alors, ME $\approx 2 \times \sqrt{(\frac{p(1-p)}{n}}$; Ici, $(1-f) = 0.996 \approx 1$
- Dans ces deux conditions, il faut prendre n $\geq \frac{1}{x^2}$ (x étant la marge d'erreur maximale que le sondeur est prêt à accepter → dans ce cas, x = 3% = 0.03)
- n $\geq \frac{1}{0.03^2} = 1111.11$ → puisque chaque individu apporte plus de précision aux estimations, ON ARRONDIT À L'UNITÉ SUPÉRIEURE (1111,11 est un minimum pour avoir une ME$\leq$ à 3%, or 0.11 individu = « no sens » → on en prend 1112) Si on remplace la formule initiale avec ces données on a :

  ME = $2 \times \sqrt{(1-0.004)\frac{0.5(1-0.5)}{1112}} = 0.02993 \approx 3\%$

2ème partie de la question :

H = 4	N_h	N_h/N	n_h	p_h
Bruxelles	N_B = 15000	N_B/N = 0.3	n_B = 65	p_B = 0.4
Louvain-la-Neuve	N_{LN} = 12000	N_{LN}/N = 0.24	n_{LN} = 40	p_{LN} = 0.2
Namur	N_N = 9250	N_N/N = 0.185	n_N = 37	p_N = 0.3
Liège	N_L = 13750	N_L/N = 0.275	n_L = 58	p_L = 0.28
	N = 50000		n = 200	

a. Dans un sondage stratifié, la variable auxiliaire est la variable qui est utilisée pour délimiter les strates. Idéalement, cette variable doit être fortement liée à la variable d'intérêt (Y) puisque le sondage stratifié a de l'intérêt quand Y est fortement hétérogène au niveau de la population dans son ensemble. Aussi, en définissant des strates sur base des différentes catégories de la variable auxiliaire, l'espoir du sondeur est de réaliser un sondage aléatoire simple au sein de groupes plus homogènes et donc, en définitive, d'améliorer la précision de ses estimations.

b. $fh = \frac{nh}{Nh}$

$fB = \frac{65}{15000} = 0.00433$

$fLN = \frac{40}{12000} = 0.00333$

$fN = \frac{37}{9250} = 0.004$

$fL = \frac{58}{13750} = 0.00422$

Ils mesurent la probabilité pour un individu appartenant à la strate h d'être sélectionné dans l'échantillon n_h.

c. Uniquement au sein de la strate BXL → sondage aléatoire simple

- $p_B = 0.4$ (donné dans l'énoncé)
- IC $(\pi) = p \pm 2 \times \sqrt{V(p)}$ avec $V(p) = (1-f)\frac{p(1-p)}{n}$

  → $V(p) = (1 - 0.00433) \times \frac{0.4 \times (1-0.4)}{65} = 0.003676$

  → IC $(\pi) = 0.4 \pm 2 \times \sqrt{0.003676}$ = [0.28 ; 0.52]

d. Toutes strates confondues → sondage stratifié

- IC $(\pi) = pst \pm 2 \times \sqrt{\hat{V}(pst)}$

  avec $pst = \sum_{h=1}^{H} \frac{Nh}{N} ph$ et $\hat{V}(pst) = \sum_{h=1}^{H} (\frac{Nh}{N})^2 (1-f)\frac{ph(1-ph)}{nh}$

  → p_{st} = 0.4 x 0.3 + 0.2 x 0.24 + 0.3 x 0.185 + 0.28 x 0.275 = 0.3005

  → $\hat{V}(p_{st})$ = 0.3² x (1 - 0.00433) x (0.4 x (1 - 0.4))/65 + 0.24² x (1 - 0.00333) x (0.2 x (1 - 0.2))/40 + 0.185² x (1 - 0.004) x (0.3 x (1 - 0.3))/37 + 0.275² x (1 - 0.00422) x (0.28 x (1 - 0.28))/58 = 0.001016

  → IC $(\pi) = 0.3005 \pm 2 \times \sqrt{0.001016}$ = [0.237 ; 0.364], ce qui signifie que, en gardant 5% de chances de se tromper, le taux de réussite des étudiants de bachelier dans les universités francophones est situé entre 23.7% et 36.4%. (TOUJOURS SE DEMANDER CE QUE SIGNIFIE UN RÉSULTAT CHIFFRÉ EN FRANÇAIS ET SAVOIR L'EXPRIMER - suppose la compréhension du concept et un retour à l'énoncé/question de départ).

e. En allocation proportionnelle : $n_h = n \times \frac{Nh}{N}$

n_B = 200 x 0.3 = 60

n_{LN} = 200 x 0.24 = 48

n_N = 200 x 0.185 = 37

n_L = 200 x 0.275 = 55

f. c_L = 20€ ; c_N = 35€ ; c_{LN} = 40€ ; c_B = 55€

C = 30000€

- 1ère étape : on calcule le nombre global d'individus qu'on va pouvoir sonder avec notre budget :

$$n = \frac{C}{\sum_{h=1}^{H} \frac{Nh}{N} ch} \rightarrow n = \frac{30000}{0.3 \times 0.55 + 0.24 \times 40 + 0.185 \times 35 + 0.275 \times 20} = 787.92$$

Étant donné qu'on ne peut pas dépasser C, ON ARRONDIT le nombre d'individus sondés À L'UNITÉ INFÉRIEURE → n = 787

- 2ème étape : on descend d'un niveau pour répartir ces n individus dans les strates de manière proportionnelle (en référence à la population) :

$n_h = n \times \frac{Nh}{N}$ → n_B = 787 x 0.3 = 236.1

n_{LN} = 787 x 0.24 = 188.88

n_N = 787 x 0.185 = 145.6

n_L = 787 x 0.275 = 216.4

Pour terminer, on arrondit de manière à avoir 787 individus au total :

n_B = 236 ; n_{LN} = 189; n_N = 146; n_N = 216

Question 2 (9 points)

Y = Variable d'intérêt = temps passé sur Facebook par jour

- Calcul des quartiles

- Distribution empirique chez les garçons :

Y	120	150	165	170	180	200	240	260
Fréq absolues	4	4	2	1	5	1	2	1
Fréq cumulées	4	8	10	11	16	17	19	20

(ATTENTION : ORDONNER LA SÉRIE AU PRÉALABLE + CALCULER LES FRÉQUENCES CUMULÉES)

- Médiane = Q2 : n = 20 (pair) → $(y_{\frac{n}{2}} + y_{\frac{n}{2}+1})/2$

On fait la moyenne de la 10ème et de la 11ème observation : $(y_{10} + y_{11})/2$ = (165 +170)/2 = 167.5

- Q1 = médiane des observations qui se situent en-dessous de Q_2 : n = 10 (pair)
On fait la moyenne de la 5ème et de la 6ème observation : $(y_5 + y_6)/2$ = (150 + 150)/2 = 150

- Q3 = médiane des observations qui se situent au-dessus de Q2 : n = 10 (pair)
On fait la moyenne de la 15ème et de la 16ème observation : $(y_{15} + y_{16})/2$ = (180 + 180)/2 = 180

- Distribution empirique chez les filles :

Y	30	45	60	75	90	120	150	200	240	300
Fréq absolues	3	1	3	1	3	4	1	1	2	1
Fréq cumulées	3	4	7	8	11	15	16	17	19	20

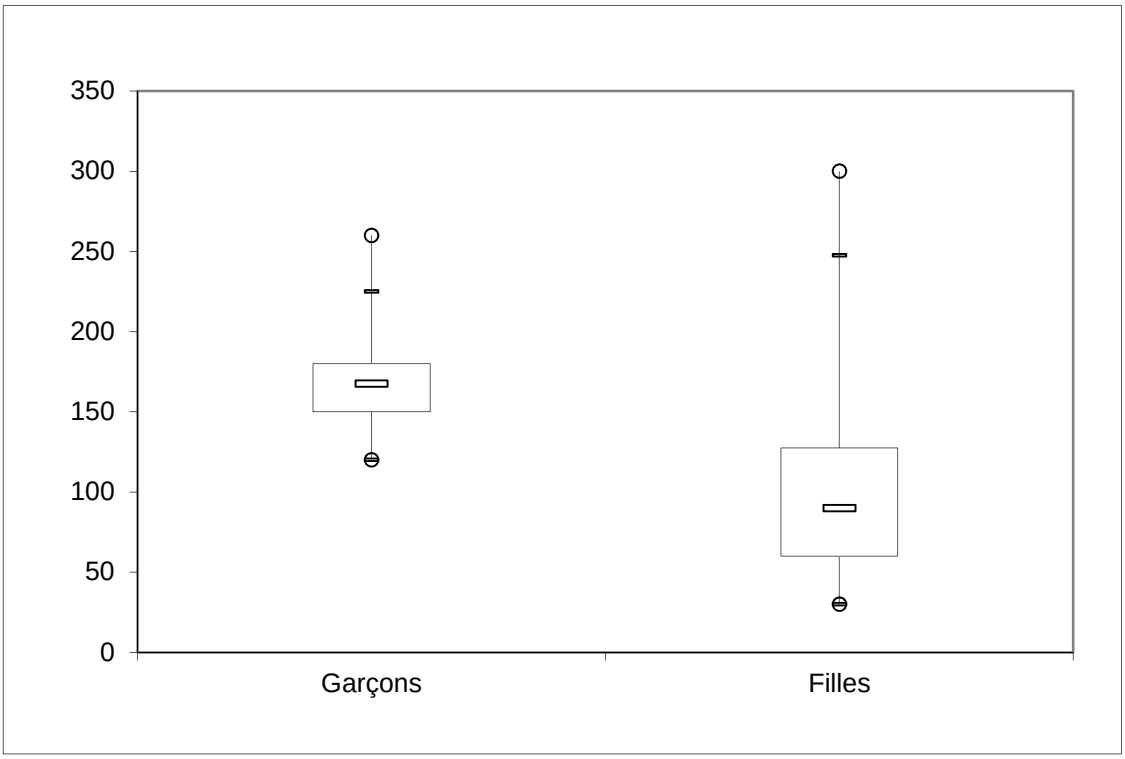
  - Médiane = Q_2 : n = 20 (pair) → $(y_{\frac{n}{2}} + y_{\frac{n}{2}+1})/2$
    On fait la moyenne de la 10ème et de la 11ème observation : $(y_{10} + y_{11})/2$ = (90 + 90)/2 = 90
  - Q_1 = médiane des observations qui se situent en-dessous de Q_2 : n = 10 (pair)
    On fait la moyenne de la 5ème et de la 6ème observation : $(y_5 + y_6)/2$ = (60 + 60)/2 = 60
  - Q_3 = médiane des observations qui se situent au-dessus de Q_2 : n = 10 (pair)
    On fait la moyenne de la 15ème et de la 16ème observation : $(y_{15} + y_{16})/2$ = (120 + 150)/2 = 135

- Construction d'un graphique permettant de visualiser le lien entre les 2 variables :
  - ➔ Boîtes à moustaches du temps passé quotidiennement sur Facebook par sexe (sur une même échelle – axe vertical - pour pouvoir comparer les 2 distributions).
  - ➔ La boîte à moustaches utilise 5 valeurs qui résument des données :
    - la patte inférieure = **la valeur minimum dans les données qui est supérieure à Q1 -1,5 x (Q3-Q1)**
    - la valeur du 1er quartile Q1 (25% des effectifs - trait inférieur de la boîte),
    - la valeur du 2ème quartile Q2 (50% des effectifs - trait horizontal à l'intérieur de la boîte),
    - la valeur du 3ème quartile Q3 (75% des effectifs - trait supérieur de la boîte),
    - la patte supérieure = **la valeur maximum dans les données qui est inférieure à Q3 +1.5 x (Q3-Q1).**

  Ici, le calcul des trois quartiles a déjà été réalisé. Il reste à calculer les « pattes » :

  - Chez les garçons :
    - Patte inférieure : valeur minimum des données (min) = 120 → comparer cette valeur avec Q1-1,5 x EIQ : 150 – 1.5 x (180-150) = 105. Min > Q1-1,5 x EIQ → patte inférieure = 120.
    - Patte supérieure : valeur maximum des données (max) = 260 → comparer cette valeur avec Q3 + 1,5 x EIQ : 180 + 1.5 x 30 = 225. Max > Q3 + 1,5 x EIQ → patte supérieure = 225.

- Chez les filles :
    - o Patte inférieure : valeur minimum des données (min) = 30 → comparer cette valeur avec Q1-1,5 x EIQ : 60 – 1.5 x (135 – 60) = -52.5. Min > Q1-1,5 x EIQ → patte inférieure = 30.
    - o Patte supérieure : valeur maximum des données (max) = 300 → comparer cette valeur avec Q3 + 1,5 x EIQ : 135 + 1.5 x 75 = 247.5. Max > Q3 + 1,5 x EIQ → patte supérieure = 247.5.



c. <u>Chez les garçons :</u>

- $\bar{y} = \frac{1}{n} \sum_{k=1}^{K} n_k y_k = \frac{1}{20} (4 \times 120 + 4 \times 150 + 2 \times 165 + 1 \times 170 + 5 \times 180 + 1 \times 200 + 2 \times 240 + 1 \times 260) = 171$
- $\hat{\sigma}^2 = \frac{1}{n} (\sum_{k=1}^{K} n_k y_k^2) - \bar{y}^2 = \frac{1}{20} (4 \times 120^2 + 4 \times 150^2 + 2 \times 165^2 + 1 \times 170^2 + 5 \times 180^2 + 1 \times 200^2 + 2 \times 240^2 + 1 \times 260^2) - 171^2 = 1546.5$
- $s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{20}{19} \times 1546.5 = 1627.89$

<u>Chez les filles :</u>

- $\bar{y} = \frac{1}{n} \sum_{k=1}^{K} n_k y_k = \frac{1}{20} (3 \times 30 + 1 \times 45 + 3 \times 60 + 1 \times 75 + 3 \times 90 + 4 \times 120 + 1 \times 150 + 1 \times 200 + 2 \times 240 + 1 \times 300) = 113.5$

- $\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{k=1}^{K} n_k\, y_k^2\right) - \bar{y}^2 = \frac{1}{20}\left(3 \times 30^2 + 1 \times 45^2 + 3 \times 60^2 + 1 \times 75^2 + 3 \times 90^2 + 4 \times 120^2 + 1 \times 150^2 + 1 \times 200^2 + 2 \times 240^2 + 1 \times 300^2\right) - 113.5^2 = 5655.25$
- $s^2 = \frac{n}{n-1}\,\hat{\sigma}^2 = \frac{20}{19} \times 5655.25 = 5952.89$

d. Type des variables :

- Sexe : variable qualitative nominale car « fille » et « garçon » indiquent la catégorie à laquelle l'individu appartient sans que l'on puisse quantifier la distance entre les deux catégories ni leur donner un ordre.
- Temps passé sur Facebook : variable quantitative continue car elle prend une valeur numérique et peut en principe prendre une infinité de valeurs possibles entre deux bornes et est généralement exprimée à l'aide d'une unité de mesure (ici : minutes).

Statut des variables :

- Sexe : variable explicative
- Temps passé sur Facebook : variable réponse

Justification : le sexe est susceptible d'influencer, d'introduire une différence dans le temps passé quotidiennement sur FB et non pas l'inverse (le sexe est un « donné » qui ne peut pas être influencé). Dans ce cas, le sens de la relation est très clair.

Question 3 (4 points)

Association entre 2 variables catégorielles (chapitre 1 slides 44 à 64)

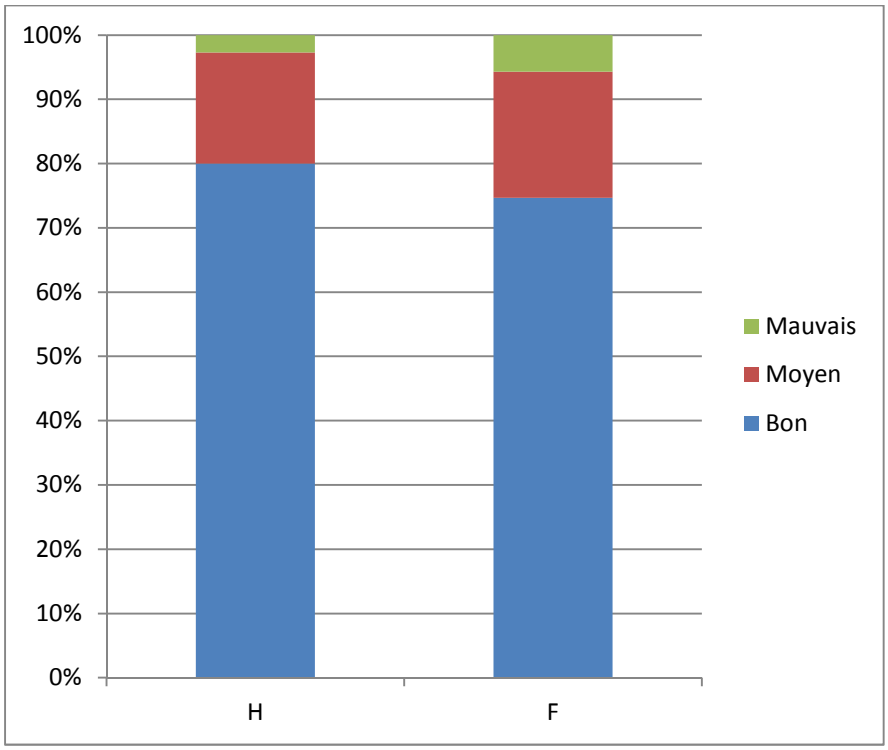
- 1ère étape : distinguer la variable explicative et la variable réponse → Genre = variable explicative ; état de santé autoévalué = variable réponse (pour les mêmes raisons que celles évoquées *supra*).
- 2ème étape : construire la table de contingence en plaçant la variable explicative en lignes et la variable réponse en colonnes.

	État de santé autoévalué			
Genre	Bon	Moyen	Mauvais	TOTAL
Hommes	699	151	24	874
Femmes	675	177	52	904
TOTAL	1374	328	76	1778

- 3ème étape : calcul des pourcentages (fréquences relatives de la variable réponse pour chaque modalité de la variable explicative)

	État de santé autoévalué			
Genre	Bon	Moyen	Mauvais	TOTAL
H	(699/874) x100 = **80%**	(151/874) x100 =**17.3%**	(24/874) x100= **2.7%**	100%
F	(675/904) x100=**74.7%**	(177/904) x100=**19.6%**	(52/904) x100=**5.7%**	100%

- 4ème étape : représentation graphique
  - En abscisse : les différentes catégories de la variable explicative avec rappel des effectifs pour chacune.
  - En ordonnée : les fréquences relatives cumulées (de 0 à 100%) pour les différentes catégories de la variable réponse.



Question 4 (4 points)

X = Vitesse des véhicules (en km/h) ; $X \sim N\ (\mu = 112;\ \sigma^2 = 4.9^2)$; $Z = \frac{X-\mu}{\sigma} \sim N\ (0,1)$

a. $P\ (X > 120) = P\left(Z > \frac{120-112}{4.9}\right)$

$= P\ (Z > 1.63)$

$= 1 - P\ (Z \leq 1.63)$ → Dans la table $P\ (Z \leq 1.63) = 0.9484$

$= 1 - 0.9484 = 0.0516$

➔ 5,16% des véhicules dont la vitesse est mesurée par un radar automatique ne respectent pas la limitation des 120km/h.

b. Il faut ajouter/soustraire une certaine valeur (v) à la moyenne μ pour trouver les bornes (-t ; t) de l'intervalle au sein duquel on trouve 70% des données :

$$P(\mu - v \leq X \leq \mu + v) = 0.7 \quad \Leftrightarrow P\left(\frac{(\mu - v) - \mu}{\sigma} \leq Z \leq \frac{(\mu + v) - \mu}{\sigma}\right) = 0.7$$

$$\Leftrightarrow P\left(\frac{-v}{\sigma} \leq Z \leq \frac{v}{\sigma}\right) = 0.7$$

$$\Leftrightarrow P\left(Z \leq \frac{v}{4.9}\right) - P\left(Z \leq \frac{-v}{4.9}\right) = 0.7$$

$$\Leftrightarrow P\left(Z \leq \frac{v}{4.9}\right) - \left(1 - P\left(Z \leq \frac{v}{4.9}\right)\right) = 0.7$$

$$\Leftrightarrow 2P\left(Z \leq \frac{v}{4.9}\right) - 1 = 0.7$$

$\Leftrightarrow P\left(Z \leq \frac{v}{4.9}\right) = \frac{0.7+1}{2} = 0.85$ → Dans la table $P\left(Z \leq \frac{v}{4.9}\right) = 0.85 \Leftrightarrow \frac{v}{4.9} \approx 1.04$

➔ v = 1.04 x 4.9 ≈ 5.1

➔ 70% des vitesses (centrées sur la moyenne) mesurées par le radar sont comprises entre : [μ - 5.1 ; μ+5.1] = [106.9 ; 117.1] km/h.