Méthodes Statistiques pour l'Ingénieur (MSI)
Durée : 2 heures

Documents autorisés : calculatrice, feuille A4 manuscrite recto/verso. Les exercices sont indépendants.

Exercice 1 (5 points)

Les réponses sont à renseigner sur la feuille jointe au sujet et à remettre dans la copie.

Plusieurs réponses peuvent être vraies pour chaque question. Une question est dite juste si et seulement si toutes les affirmations vraies sont cochées et aucune affirmation fausse n'est cochée.

- *1 question juste : +1/2 de points*
- *1 question fausse : -1/2 de points*
- *1 question sans réponse : 0 point*

Exemples :

- *8 questions justes et 2 sans réponse = 4 points*
- *10 questions justes = 5 points*
- *5 questions justes, 5 questions fausses = 0 points*

Question 1 :

A. La médiane est un quantile.
B. L'écart-type est la racine carrée de la variance.
C. La médiane est sensible aux valeurs extrêmes.
D. La médiane est un indicateur de dispersion.
E. Aucune de ces affirmations n'est vraie.

Question 2 :

Les services de la Poste ont effectué une enquête sur la durée en jours, pour qu'un colis soit remis à son destinataire à partir du jour de dépôt. Ces résultats sont résumés dans le tableau ci-dessous.

Nombre de jours	1	2	3	4	6	7	8	10	11	12
Nombre de colis	106	158	195	96	55	38	62	14	76	50

Le nombre total de colis est 850. La moyenne du nombre de jours pour qu'un colis soit remis à son destinataire à partir du jour de dépôt est de 4,78.

A. Le mode de la série est 195.
B. Le mode de la série est 3.
C. Le mode de la série est 2,5.
D. La médiane est 2.
E. Aucune de ces affirmations n'est vraie.

Question 3 :

A. L'état civil (marié, célibataire, concubin, divorcé, veuf) est une variable qualitative ordinale.
B. L’état civil (marié, célibataire, concubin, divorcé, veuf) est une variable qualitative cardinale.
C. Le niveau d'études de votre père (primaire, secondaire, premier cycle sup, 2ème cycle sup, etc.) est une variable cardinale.
D. Le niveau d'études de votre père (primaire, secondaire, premier cycle sup, 2ème cycle sup, etc.) est une variable ordinale.
E. Aucune de ces réponses n'est vraie.

Question 4 :

A Le code postal du lieu d'habitation est une variable quantitative.
B La durée de vie en jours d'un ordinateur est une variable qualitative ordinale.
C Les variables qualitatives nominales existent.
D Les variables qualitatives binaires existent.
E Aucune de ces affirmations n'est vraie.

Question 5 :

A. La statistique inférentielle est un domaine de la statistique.
B. La statistique référentielle est un domaine de la statistique.
C. La statistique fréquentielle est un domaine de la statistique.
D. Le mot statistique vient du latin *status*, qui signifie Etat.
E. Aucune de ces affirmations n'est vraie.

Question 6 :

A. Dans un échantillonnage stratifié, les strates sont obligatoirement sans individu en commun.
B. Des méthodes d'enquêtes non aléatoires sont souvent utilisées dans les sondages politiques.
C. L'échantillonnage par grappe nécessite d'avoir une liste de l'ensemble de la population.
D. L'échantillonnage systématique nécessite d'avoir une liste de l'ensemble de la population.
E. Aucune de ces réponses n'est vraie.

Question 7 :

Des douaniers font des contrôles sur les véhicules qui leur paraissent suspects.
A. La méthode d'échantillonnage est une méthode aléatoire simple.
B. La méthode d'échantillonnage est une méthode d'échantillonnage stratifié.
C. La méthode d'échantillonnage est une méthode aléatoire systématique.
D. La méthode d'échantillonnage est une méthode des quotas.
E. Aucune de ces affirmations n'est vraie.

Question 8 :

Pour mesurer le temps passé par les clients dans son magasin, un commerçant donne un chronomètre à un client sur dix, dans l'ordre d'entrée au magasin.
A. La méthode d'échantillonnage est une méthode aléatoire simple.
B. La méthode d'échantillonnage est une méthode d'échantillonnage stratifié.
C. La méthode d'échantillonnage est une méthode aléatoire systématique.

D. La méthode d'échantillonnage est une méthode des quotas.
E. Aucune de ces affirmations n'est vraie.

Question 9 :

Pour une enquête sur des élèves de troisième du département du Rhône, on tire au hasard un collège sur 10 du département et on interroge tous les élèves des collèges ainsi sélectionnés.
A. La méthode d'échantillonnage est une méthode aléatoire simple.
B. La méthode d'échantillonnage est une méthode d'échantillonnage stratifié.
C. La méthode d'échantillonnage est une méthode aléatoire par grappe.
D. La méthode d'échantillonnage est une méthode d'échantillonnage à plusieurs degrés.
E. Aucune de ces affirmations n'est vraie.

Question 10 :

A. Un échantillon aléatoire de 1 000 Luxembourgeois est aussi représentatif des Luxembourgeois qu'un échantillon aléatoire de 1 000 Chinois l'est des Chinois.
B. Il existe des méthodes d'échantillonnage non aléatoires utilisées en statistique.
C. Doubler la taille de l'échantillon permet de réduire de moitié les fluctuations.
D. A taille égale, les résultats des échantillons pris dans une population plus homogène seront plus sujets aux fluctuations d'échantillonnage.
E. Aucune de ces affirmations n'est vraie.

EXERCICE 1
NOM :
Prénom :
Groupe :

Les réponses sont à renseigner dans la feuille jointe exclusivement :
Mettre une croix quand l'affirmation est vraie.

N° de question	A	B	C	D	E
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Cadre réservé au correcteur

Nombre de questions justes =
Nombre de questions fausses =
Nombre de questions sans réponse =

d'où la note sur 5 points :

Exercice 2 (3 points)

On considère deux échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_q)$ identiquement et indépendamment distribués *(i.i.d)* de variables aléatoires parentes $X \sim N(m_1, \sigma^2)$ et $Y \sim N(m_2, \sigma^2)$. On suppose que la variance σ^2 est connue.

1. Donner la distribution des statistiques $\bar{X}$ et $\bar{Y}$ (0.5 point).

2. Montrer que $T = \bar{X} - \bar{Y}$ est un estimateur sans biais de $m_1 - m_2$ et calculer sa variance (1 point).

3. En déduire la distribution de T et une fonction pivotale du paramètre $m_1 - m_2$ (c'est-à-dire la fonction obtenue en centrant et réduisant T). (1 point)

4. Déterminer un intervalle de confiance pour le paramètre $m_1 - m_2$ au niveau de confiance $1 - \alpha$= 0.95 (0.5 point).

Exercice 3 (4 points)

L'intensité lumineuse d'une lampe de 100 watts est une variable aléatoire distribuée selon une loi normale avec un écart-type σ de 80 lumens (unité internationale de la puissance lumineuse). Pour un fabricant, on considère que la population des lampes produites doit avoir une intensité lumineuse d'au moins 1400 lumens. Si cette norme n'est pas respectée, la production est mise au rebus.

1. On prélève un échantillon E1 de 65 lampes. L'intensité lumineuse moyenne de cet échantillon est de m=1385. En prenant soin de poser les hypothèses de test, déterminer la valeur moyenne de luminosité qui définit la zone de rejet pour un test de niveau 5%. Peut-on alors conclure à partir de l'échantillon E1, que la production satisfait la norme de qualité (1,5 points) ?

2. Si en réalité l'intensité lumineuse moyenne est de 1380 lumens, quelle est la puissance de ce test (1,5 points) ?

3. On consent de rejeter un lot de cette qualité (c'est-à-dire 1400 lumens) 1 fois sur 100. De plus, on veut rejeter 90 fois sur 100 un lot dont la moyenne est de 1370 lumens. Déterminer la taille de l'échantillon à prélever (1 point).

Exercice 4 (4 points)

Dans une école primaire, tous les élèves ont passé un test d'orthographe. Il s'agissait d'une courte dictée, la même pour tous. On note pour chaque élève le nombre de fautes d'orthographes et la taille des pieds, c'est-à-dire la pointure. Les résultats de cette étude sont résumés dans le tableau suivant :

Nb fautes	29	28	23	19	18	15	14	16	13	12	4	5	2
Pointure	26	27	28	29	30	31	32	33	34	35	36	37	38

1. Tracez le graphique représentant le nuage de points des fautes d'orthographes en fonction de la pointure des élèves, placez le point moyen. Calculez les estimations des coefficients de la droite de régression associée. Tracez la droite sur le nuage de points. Comment interpréter la constante du modèle? Et la pente (2 pts) ?

2. Quelles conditions doivent être réunies pour qu'un coefficient de corrélation et la pente de la droite de régression associée aient la même valeur ? Calculez la valeur du coefficient de corrélation linéaire entre nombre de fautes et pointure (1pt).

3. Que peut-on dire du lien entre corrélation et causalité ? Justifiez à partir de l'exemple (1 pt).

Exercice 5 (4 points)

Soit la variable aléatoire X de densité

$$f(x) = \frac{\lambda^4}{6} x^3 e^{-\lambda x} \, 1_{[0,+\infty[}(x)$$

avec $\lambda > 0$. On dispose d'un échantillon i.i.d $(X_1, \dots, X_n)$ de variable aléatoire parente X.

1. Donner l'estimateur du maximum de vraisemblance du paramètre λ (1 point).

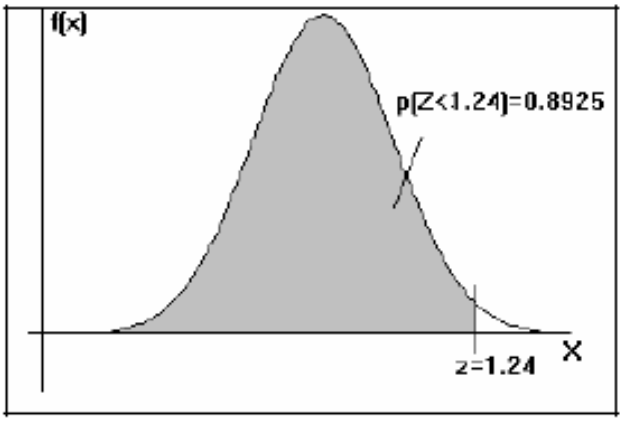
2. Calculer l'information de Fisher et la borne de Cramer-Rao. En déduire une loi approchée pour l'estimateur du maximum de vraisemblance (1 point).

3. Expliquer de façon générale le lien entre les notions précédentes et le concept d'efficacité d'un estimateur. Comment interviennent les statistiques exhaustives dans ce cadre théorique (2 points) ?

ANNEXE : table des lois usuelles

TABLE DE LA LOI NORMALE CENTREE REDUITE

Lecture de la table: Pour z=1.24 (intersection de la ligne 1.2 et de la colonne 0.04), on a la proportion P(Z < 1,24) = 0.8925



P(Z > 1,96) = 0,025
P(Z > 2,58) = 0,005
P(Z > 3,29) = 0,0005

Rappels:

1/ P(Z > z) = 1 - P(Z < z) et 2/ P(Z < -z) = P(Z > z)
Exemple: Sachant P(Z < 1,24) = 0,8925, on en déduit:
1/ (P(Z > 1,24) = 1 - P(Z < 1,24) = 1- 0,8925 = 0,1075
2/ P(Z < -1,24) = P(Z > 1,24) = 0,1075

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

DISTRIBUTION DU KHI2

La table donne les valeurs critiques de χ^2 pour un nombre de degrés de liberté (ddl) et pour un seuil repère donnés (α).

Par exemple:

Pour ddl = 3 et α= 0,05 la table indique $\chi^2 = 7{,}81$

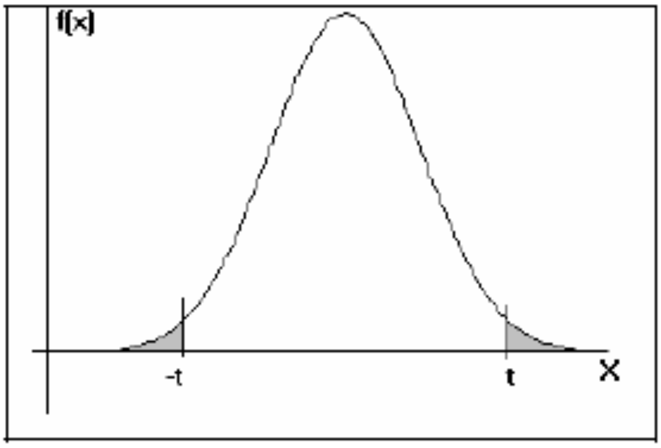
Ceci signifie que: $P(\chi^2_{[3]} > 7{,}81) = 0{,}05$

ddl \ α	0,05	0,01	0,001
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,12
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,72	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,31
21	32,67	38,93	46,80
22	33,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70

DISTRIBUTIONS DU t DE STUDENT

Table des valeurs critiques bilatérales usuelles

Pour une distribution de Student à ddl degrés de liberté et pour une proportion α (.05, .01 ou .001), la table indique t tel que $P(|T| > t) = \alpha$



Exemple: Pour ddl = 5, on a $P(|T| > 2.571) = .05$ (on note $t_{[5].05}$ cette valeur.).

α α/2 ddl	0,05 0,025	0,01 0,005	0,001 0,0005
1	12.706	63.657	636.619
2	4.303	9.925	31.599
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
21	2.080	2.831	3.819
22	2.074	2.819	3.792
23	2.069	2.807	3.768
24	2.064	2.797	3.745
25	2.060	2.787	3.725
26	2.056	2.779	3.707
27	2.052	2.771	3.690
28	2.048	2.763	3.674
29	2.045	2.756	3.659
30	2.042	2.750	3.646
40	2.021	2.704	3.551
60	2.000	2.660	3.460
120	1.980	2.617	3.373
30000	1.960	2.576	3.291

Méthodes Statistiques pour l'Ingénieur
Corrigé de l'examen Final

ENTPE 2A

6 Aout 2013

Exercice 1 : QCM

Pour réussir cet exercice, il fallait maïtriser deux chapitres du cours. Tout d'abord le cours 1 et le TD1 sur les statistiques descriptives pour les questions 1 à 5. Des fichiers pdf intitulés "SEANCE1 intro" et "TD1 stats des" vous avaient été transmis. En ce qui concerne les questions 6 à 10, il s'agit de la deuxième moitié du premier cours donné en amphi sur les méthodes d'échantillonnage. Un fichier pdf nommé "SEANCE1 ECHANTILLONNAGE" a été envoyé à toute la promotion tandis que les deux premiers exercices du TD2 ont exactement le même contenu que ces questions du QCM. Si des difficultés persistent, revoir le TD2. La notation finale a été beaucoup plus généreuse que celle annoncée. Certaines réponses différentes que celles présentes dans le tableau mais bien expliquées ont été récompensées.

Référence complémentaire : Ardilly, P. (2006). Les techniques de sondage. Editions Technip.

	A	B	C	D	E
Q1	X	X			
Q2	X				
Q3				X	
Q4			X	X	
Q5	X			X	
Q6	X	X		X	
Q7					X
Q8			X		
Q9			X		
Q10	X	X			

Table 1 – Correction du QCM

Exercice 2

Exercice très classique de rappel du cours de proba de 1A. Revoir les rappels de proba du premier cours et le chapitre sur les intervalles de confiance.

1)Nous avons classiquement $\bar{X}$ qui suit une loi normale $N(m_1, \frac{\sigma^2}{n})$ et $\bar{Y} \sim N(m_2, \frac{\sigma^2}{q})$.

2) Linéarité de l'espérance.

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = m_1 - m_2. \tag{1}$$

Comme $\bar{X}$ et $\bar{Y}$ sont indépendantes, nous avons

$$Var[\bar{X} - \bar{Y}] = Var[\bar{X}] + Var[\bar{Y}] = \sigma^2(\frac{1}{n} + \frac{1}{q}). \tag{2}$$

3) On centre et on réduit pour obtenir la fonction pivotale

$$\frac{\big(T - (m_1 - m_2)\big)}{\big(\sigma * \sqrt{\frac{1}{n} + \frac{1}{q}}\big)}, \tag{3}$$

qui suit une loi normale centrée-réduite $N(0, 1)$.

4) Intervalle de confiance

$$T - 1.96 * \sigma * \sqrt{\frac{1}{n} + \frac{1}{q}} < m_1 - m_2 < T + 1.96 * \sigma * \sqrt{\frac{1}{n} + \frac{1}{q}}. \tag{4}$$

Exercice 3

Exercice standard sur les test de type bac +3. Bien revoir chapitre sur les tests pour maitriser les différentes facettes. Une recherche basique sur internet sur les test d'hypothèses sur des lois normales vous mènera à un nombre quasi infini de ressources. Je recommande particulièrement ce polycopié dont le contenu est proche du cours prodigué ici :

http ://www-ljk.imag.fr/membres/Olivier.Gaudoin/PolyMSI.pdf

1) Test unilatéral
- (H0) m = 1400
- (H1) m < 1400

Sous $H0$. Zone rejet au seuil 0.05 :

$$-1.64 = \sqrt{65} * \frac{(\bar{X} - 1400)}{80}. \tag{5}$$

D'où $\bar{X} = 1383.72$ et zone de rejet : $[0; 1383.72]$. Comme $\bar{x}>$1383.72 : non rejet de $H0$, production de qualité.

2) Sous $H1$:

$$Z_1 = \sqrt{65} * \frac{(1383,72664 - 1380)}{80} = 0.375564, \tag{6}$$

$P(Z <= Z_1) = 0,6443$.
Puissance de 64.4%.
Erreur de 2ème espèce : 35.6%.

3) On cherche n sachant $\alpha = 0.01$, $1 - \beta = 0.9$. Sous H_1, $m = 1370$.
Sous H_0, on a la zone de rejet

$$-2.33 = \sqrt{n} * \frac{(\bar{X} - 1400)}{80}, \tag{7}$$

donc $\bar{X} = -2.33 * \frac{80}{\sqrt{n}} + 1400$.

Sous $H1$ on a

$$z_1 = \sqrt{n} * \frac{(\bar{X} - 1370)}{80}. \tag{8}$$

$P(Z <= z_1) = 0.9$, ce qui implique $z_1 = 1.28$.
Donc $1.28 = -2.33 + 3/8 * \sqrt{n}$ Soit $N >= 92,67$.
On en déduit que la taille requise pour l ?échantillon est $N = 93$.

Exercice 4

Application directe du chapitre sur la regression linéaire.

Référence complémentaire : Cornillon, P. A., Matzner-Løber, É. (Eds.). (2006). Régression : théorie et applications. Springer.

1) Tracé des points et de la droite de regression :
Le nuage de points ainsi que la droite de regression montrent une liaison linéaire entre le nombre de fautes et la pointure. Cette liaison est confirmée par un coefficient de corrélation de Bravais Pearson d'une valeur de -0.96. Nous avons donc une corrélation négative forte. Les coordonnées du point moyen sont (32, 15.23).
2) Cette question triviale visait juste à faire prendre conscience de la difference entre pente et coefficient de corrélation.
3) Equation de la droite des moindres carré : $y^* = \alpha + \beta x$. Avec

$$\beta = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}, \tag{9}$$

et $\alpha = \bar{y} - \beta\bar{x}$.
D'après les calculs, on a $\alpha = 82.571$ et $\beta = -2.104$

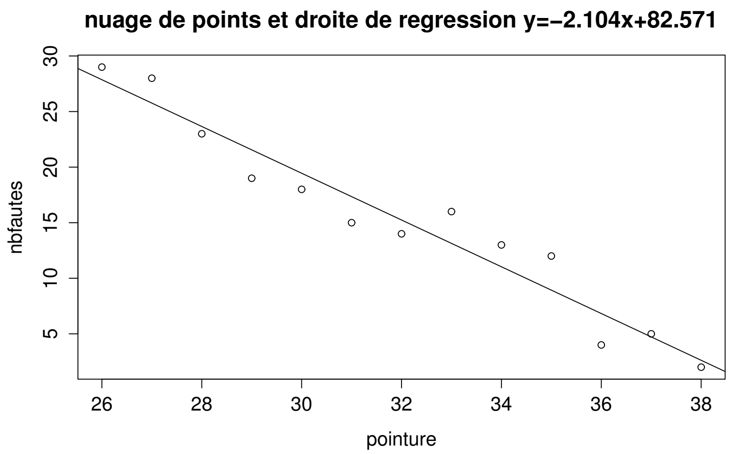


FIGURE 1 – Nuage de points et droite de regression

4) Différence entre corrélation et causalité. La pointure n'explique bien sûr pas le nombre de fautes. Il y a une variable de confusion : l ?âge et donc le niveau (classe) des enfants.

Voilà pour info le code en R :
nbfautes=c(29,28,23,19,18,15,14,16,13,12,4,5,2) ;
pointure=c(26,27,28,29,30,31,32,33,34,35,36,37,38) ;
lm.D9=lm(nbfautes pointure) ;
summary(lm.D9) ;

Call : lm(formula = nbfautes pointure)

Residuals : Min 1Q Median 3Q Max -2.8132 -1.4396 -0.6044 1.9780 3.0824

Coefficients : Estimate Std. Error t value Pr(>|t|) (Intercept) 82.5714 5.2086 15.85 6.36e-09 *** pointure -2.1044 0.1617 -13.02 5.02e-08 *** — Signif. codes : 0 ?*** ? 0.001 ?** ? 0.01 ?* ? 0.05 ?. ? 0.1 ? ? 1

Residual standard error : 2.181 on 11 degrees of freedom Multiple R-squared : 0.939, Adjusted R-squared : 0.9335 F-statistic : 169.4 on 1 and 11 DF, p-value : 5.02e-08

Exercice 5

Exercice là encore standard sur l'estimation du maximum de vraisemblance. Revoir les cours 2 et 3 en amphi. Comme annoncé en amphi, les questions 2 et 3 étaient plus sélectives. Je renvoie au livre de Gilbert Saporta pour plus d'information.

Référence complémentaire : Saporta, G. (2006). Probabilités, analyses des données et statistiques. Editions Technip.
1) La fonction de vraisemblance s'écrit

$$L(x_1, ..., x_n; \lambda) = \Pi_{i=1}^{n} f(x_i) = \Sigma_{i=1}^{n} ln(f(x_i)). \tag{10}$$

On a la log-vraisemblance suivante :

$$lnL(x_1, ..., x_n; \lambda) = 4nln(\lambda) - \lambda n\bar{x} + \Sigma_i ln(\frac{x_i^3}{6}). \tag{11}$$

Il suffit ensuite de développer l'expression de la log-vraisemblance comme suit :

$$\frac{\partial lnL}{\partial \lambda}(x_1, ..., x_n; \lambda) = \frac{4n}{\lambda} - n\bar{x} \tag{12}$$

puis de voir qu'elle s'annule pour

$$\lambda = \frac{4}{\bar{x}} \tag{13}$$

On vérifie bien que la dérivée seconde est négative ($\frac{-4n}{\lambda^2} < 0$) et on en déduit que l'estimateur du maximum de vraisemblance de λ est

$$\hat{\lambda} = \frac{4}{\bar{X}}. \tag{14}$$

2) Le support de densité ne dépend pas du paramètre λ. On obtient l'information de Fisher directement à partir de la dérivée seconde et on a

$$E\big[\frac{\partial^2 lnL}{\partial \lambda^2}(X_1, ..., X_n; \lambda)\big] = -\frac{4n}{\lambda^2}, \tag{15}$$

Soit

$$I_n(\lambda) = \frac{4n}{\lambda^2}. \tag{16}$$

La borne de Cramer-Rao est l'inverse de l'information de Fisher. On a approximativement la loi approchée

$$\hat{\lambda} \sim N(\lambda, \frac{\lambda^2}{4n}). \tag{17}$$

3) Voir cours en amphi ou le livre de Saporta.