Barème

Université de Cergy-Pontoise - L2 Economie – Finance – CMI – DU - S2 - Année 2017-2018

EXAMEN D'ANALYSE STATISTIQUE

DEUXIEME SESSION

Tous documents interdits - Calculatrice non programmable autorisée

Questions de cours 15

1) Voici plusieurs variables. Indiquez, pour chacune d'entre elles, la notation utilisée en cours et s'il s'agit d'une variable aléatoire ou pas. /1

- Ecart-type empirique
- Fréquence empirique

2) Expliquez la logique d'un test statistique de conformité à une valeur de type /4

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta > 0 \end{cases}$$

Plus précisément à quoi correspondent les notions d'erreurs de type I, II, la puissance du test et expliquez comment la région critique est calculée.

Exercice 1 : Construction d'un estimateur par la méthode du maximum de vraisemblance 16

On dispose d'un échantillon simple au hasard $(X_1, X_2, \ldots, X_n)$ issu du Processus Générateur de Données suivant : $X_i \rightarrow N(\theta; 1)$ pour $i=1,\ldots,n$.

1) Indiquez la fonction de vraisemblance correspondant à ce cas. /1

2) Après avoir résolu les conditions nécessaires d'optimisation, donnez l'expression de l'estimateur de l'espérance θ par la méthode du maximum de vraisemblance. /2

3) Résolvez les conditions suffisantes. /1

4) Est-ce un estimateur sans biais ? /1

5) L'estimateur est-il efficace ? (démontrez). /1

Indications :
La fonction de densité de la loi normale d'espérance μ et d'écart-type σ s'écrit :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

La quantité d'information de Fisher se calcule comme $I_n(\theta) = -E\left[\frac{d^2 \ln L}{d\theta^2}\right]$, la borne de Rao-Cramer correspondant à l'inverse de cette quantité.

Exercice 2 : Choix de marques /12

Afin d'optimiser son positionnement stratégique à New York, une grande marque de vêtements (Ralf) réalise une étude comparative. Elle réalise un tirage aléatoire simple d'une population de référence constituée de plusieurs milliers de clients New Yorkais potentiels et les interroge sur leurs préférences en termes de marques. Les effectifs enquêtés se répartissent dans le tableau de contingence ci-dessous.

Préférence / Catégorie Socio – Prof (CSP)	Marque Ralf	Marque Lacorte	Autres Marques
Etudiants	352	300	243
Autres Catégories	280	310	123

Une campagne publicitaire de la marque Ralf cible spécifiquement les étudiants. Ainsi, la marque Ralf souhaiterait savoir si elle parvient à attirer davantage la préférence des jeunes que des personnes issues d'autres catégories socio-professionnelles. Peut-on considérer que la marque Ralf a un positionnement privilégié auprès des étudiants ? En d'autres termes est-ce que la proportion de personnes préférant Ralf au sein de la population étudiante est différente de la proportion de personnes préférant Ralf au sein des populations ayant une autre CSP ?

Vous répondrez à cette question en construisant un intervalle de confiance représentant la différence entre la proportion d'étudiants préférant Ralf et la proportion d'autres catégories préférant Ralf.

Pour rédiger votre réponse respectez les étapes et notations suivantes :

Etape 0 : Déterminez les 2 populations mères de référence ici. Indiquez la taille des échantillons respectifs *pour ce problème spécifique*. Indiquez la proportion d'étudiants préférant la marque Ralf (cette fréquence empirique sera notée F_{nA}) et la proportion de personnes d'autres catégories préférant la marque Ralf (cette fréquence empirique sera notée F_{nB}) /2

Etape 1 : Modélisez. Définissez les variables aléatoires X_i^A et X_i^B indiquez leur loi de probabilité. /1

Etape 2: En supposant les échantillons indépendants, donnez un estimateur adéquat de la différence entre la proportion d'étudiants préférant Ralf et la proportion d'autres catégories préférant Ralf et indiquez sa loi de probabilité (preuve <u>*complète*</u>) /3

Etape 3 : Après avoir centré et réduit, construisez l'intervalle de confiance de niveau 99%. /2

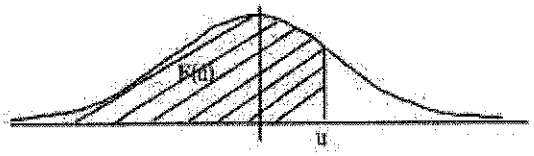
Etape 4 : Faire une phrase d'interprétation. /1

2) Avec un niveau de confiance de 90%, le résultat serait-il le même ? /1

3) Lorsque le niveau de confiance de l'intervalle baisse, que constatez-vous et pourquoi ? /1

4) Discutez de la validité de l'hypothèse d'indépendance dans ce cas-là. /1

Variable NORMALE CENTREE REDUITE



$U \approx N(0,1)$ $\quad P(U \leq u) = F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2} dx$

$F(-u) = 1 - F(u)$ $\qquad P(|U| \leq u) = 2\,F(u) - 1$

TABLE de F(u) en fonction de u :

u	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	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Correction Session 2 (eco Fi)

Questions de cours

① - Ecart type empirique : S_n, aléatoire

- Fréquence empirique : F_n, aléatoire

② VOIR COURS

Exercice 1

① $\mathcal{L}(x_1, \dots, x_n ; \mu) = \prod_{i=1}^{n} f(x_i, \mu)$ si l'échantillon est i.i.d.

où $f(x_i ; \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$

② $\hat{\mu}$ est solution de $\max_{\mu} \mathcal{L}(x_1, \dots, x_n ; \mu) \Leftrightarrow \max_{\mu} \ln \mathcal{L}(x_1, \dots, x_n ; \mu)$

où $\ln \mathcal{L}(.)$ est la log-vraisemblance noté $\ell(x_1, \dots, x_n ; \mu)$

$$\ell(x_1, \dots, x_n ; \mu) = \sum_{i=1}^{n} \ln f(x_i ; \mu) = \sum_{i=1}^{n} \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\ell(x_1, \dots, x_n ; \mu) = -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

(CN) : $\frac{d\ell(x_1, \dots, x_n ; \hat{\mu})}{d\mu} = 0 \Leftrightarrow -\frac{1}{2}\sum_{i=1}^{n} 2(x_i - \mu) \times (-1) = 0$

$\Leftrightarrow \sum(x_i - \hat{\mu}) = 0 \Leftrightarrow \sum x_i - n\hat{\mu} = 0 \Leftrightarrow \boxed{\hat{\mu} = \frac{\sum x_i}{n}}$

③ (CS) : $\frac{d^2\ell(x_1, \dots, x_n ; \mu)}{d\mu^2} < 0$ $\quad \frac{d^2\ell(.)}{d\mu^2} = -n < 0$ OK

Exercice 1 (suite)

(4) $\hat{\mu}$ est sans biais si $E(\hat{\mu}) = \mu$

$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$ car $E()$ est un opérateur linéaire.

$E(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mu$ car $E(X_i) = \mu$

$E(\hat{\mu}) = \frac{1}{n} \times n \times \mu = \mu$ $\hat{\mu}$ est sans biais.

(5) $\hat{\mu}$ est efficace si sa variance atteint la borne de Rao Cramer

- $V(\hat{\mu}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$ car les X_i sont indépendants par hypothèse.

  or $V(X_i) = 1$

  D'où $V(\hat{\mu}) = \frac{1}{n^2}\sum 1 = \frac{n \times 1}{n^2} = \boxed{\frac{1}{n}}$

- Borne de Rao-Cramer $= \frac{1}{I_n(\mu)}$

$I_n(\mu) = -E\left(\frac{d^2 \ln \mathcal{L}}{d\mu^2}\right) = -E\left(\frac{d^2 \ell}{d\mu^2}\right) = -E(-n) = n$

Borne $= \frac{1}{n} = V(\hat{\mu})$ Donc $\hat{\mu}$ est efficace sa variance est la plus petite possible puisqu'elle atteint la borne de Rao Cramer.

Exercice 2

(1) **Etape 0 :**

On a 2 populations mères

- Pop. A : ensemble des étudiants à New York
- Pop. B : ensemble des personnes d'autres catégories à New York

- On dispose d'échantillons iid tirés de ces 2 populations mères

de tailles $n_A = 352 + 300 + 243 = \underline{895}$

$n_B = 280 + 310 + 123 = \underline{713}$

- on dispose de statistiques d'échantillon : proportion d'étudiants préférant la marque Raff : $f_{n_A} = \frac{352}{895} \simeq 0,3933 \simeq \underline{39,33\ \%}$

et proportion d'autres personnes préférant la marque Raff

$f_{n_B} = \frac{280}{713} \simeq 0,3927 \simeq \underline{39,27\ \%}$

Etape 1 :

Soient les variables aléatoires X_i^A et X_i^B représentant respectivement le fait de préférer la marque Raff pour un étudiant (respectivement une personne d'une autre catégorie).

$\begin{cases} X_i^A = 1 \text{ si l'individu } i \text{ préfère Raff} \\ 0 \text{ sinon} \end{cases}$

idem pour X_i^B.

Ces variables aléatoires binaires suivent une loi de Bernoulli

Exercice 2 (suite)

$X_i^A \sim \mathcal{B}(p_A)$ où p_A et p_B représentent la proba qu'un
$X_i^B \sim \mathcal{B}(p_B)$ client de la pop A (et B pour p_B)
préfère la marque Ralf.
Ces probas sont inconnues.

on sait que $E(X_i^A) = p_A$ $\quad V(X_i^A) = p_A(1-p_A)$
$E(X_i^B) = p_B$ $\quad V(X_i^B) = p_B(1-p_B)$.

(Caractéristiques de la loi de proba = théorique).

étape 2:

les fréquences empiriques sont des estimateurs de la proportion dans la population mère (ou de la proba)

ainsi F_n^A est aussi la moyenne empirique $\overline{X}_n^A = \frac{1}{n_A}\sum_{i=1}^{n} X_i^A = F_n^A$

estimateur de p_A noté $\hat{p}_A = \overline{X}_{n_A} = F_n^A$

en effet $E(F_n^A) = E(\overline{X}_{n^A}) = E\left(\frac{1}{n_A}\sum_{i=1}^{n_A} X_i^A\right) = \frac{1}{n_A}\sum_{i=1}^{n_A} \underbrace{E(X_i^A)}_{p_A}$.

$= \frac{1}{n_A} \times n_A \times p_A = p_A$

F_n^A est un estimateur sans biais de p_A.

Par le théorème central limite nous connaissons la loi de proba asymptotique (quand la taille de l'échantillon tend vers $+\infty$) de la moyenne empirique $\overline{X}_{n_A}$ et donc de la fréquence empirique F_n^A.

Exercice 2 (suite)

Etape 2 (suite) :

TCL : $\overline{X}^A_{n_A} \underset{\infty}{\sim} \mathcal{N}\left(\mu_A, \frac{\sigma_A}{\sqrt{n_A}}\right)$ ici $\overline{X}_{n_A} = F_n^A$

$\mu_A = p_A$

Donc dans le cas de la fréquence empirique on a : $\sigma_A = \sqrt{V(X;A)} = \sqrt{p_A(1-p_A)}$

$$\boxed{F_n^A \underset{\infty}{\sim} \mathcal{N}\left(p_A ; \sqrt{\frac{p_A(1-p_A)}{n_A}}\right)}$$

de la même façon pour $\boxed{F_n^B \underset{\infty}{\sim} \mathcal{N}\left(p_B ; \sqrt{\frac{p_B(1-p_B)}{n_B}}\right)}$

- La différence de proportion dans la population mère s'écrit $p_A - p_B$ et s'estime par $F_n^A - F_n^B$

en effet $E(F_n^A - F_n^B) = p_A - p_B$

car $E(F_n^A - F_n^B) = \underbrace{E(F_n^A)}_{p_A} - \underbrace{E(F_n^B)}_{p_B}$.

$F_n^A - F_n^B$ est un estimateur sans biais de $p_A - p_B$.

- La loi de proba asymptotique de $F_n^A - F_n^B$ est la loi normale car c'est une combi. lin. de lois normales.

j'ai donc $\boxed{F_n^A - F_n^B \underset{\infty}{\sim} \mathcal{N}(p_A - p_B, \ ?)}$.

$V(F_n^A - F_n^B) = V(F_n^A) + V(F_n^B)$ si les échantillons A et B sont indépendants ce que l'on suppose.

Exercice 2 (suite)

Etape 2 suite

On $V(f_n^A) = \frac{p_A(1-p_A)}{n_A}$ $\quad V(F_n^B) = \frac{p_B(1-p_B)}{n_B}$.

donc $V(F_n^A - F_n^B) = \frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}$

D'où $\boxed{F_n^A - F_n^B \underset{\infty}{\sim} N\left(p_A - p_B ; \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}\right)}$

Etape 3

On cherche un intervalle $[A,B]$ tel que $P(A < p_A - p_B < B) = 0{,}99$

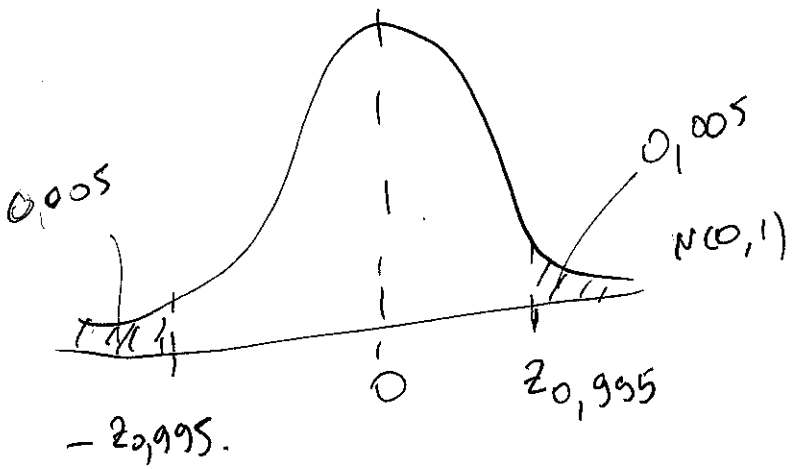
On centre et on réduit la loi de l'intervalle de $p_A - p_B$:

$$Z = \frac{F_n^A - F_n^B - (p_A - p_B)}{\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}} \underset{\infty}{\sim} N(0,1)$$

Donc $P(-z_{0,995} < Z < z_{0,995}) = 0{,}99$

Dans la table on lit que

$z_{0,995} = \Phi^{-1}(0{,}995) = \underline{2{,}575}$



$P(-2{,}575 < Z < 2{,}575) = 0{,}99$.

Exercice 2 (suite)

Etape 3 (suite)

les écarts types des fréquences empiriques sont estimés en remplaçant $p_A(1-p_A)$ par $F_{m_A}(1-F_{m_A})$.

D'où $Z = \dfrac{F_m^A - F_m^B - p_A - p_B}{\sqrt{\dfrac{F_{m_A}(1-F_m^A)}{n_A} + \dfrac{F_m^B(1-F_m^B)}{n_B}}}$.

on récrit l'intervalle : $P(-2,575 < Z < 2,575) = 0,95$

$\Leftrightarrow P\left(-2,575 < \dfrac{F_m^A - F_m^B - p_A - p_B}{\sqrt{\dfrac{f_{m_A}(1-f_m^A)}{n_A} + \dfrac{f_m^B(1-f_m^B)}{n_B}}} < 2,575\right) = 0,95$

$\Leftrightarrow P\left[F_m^A - F_m^B - 2,575 \times \sqrt{\dfrac{F_{m_A}(1-f_m^A)}{n_A} + \dfrac{F_m^B(1-f_m^B)}{n_B}} < p_A - p_B\right.$

$\left. < \underbrace{F_m^A - f_m^B}_{\simeq 0,00059} + 2,575 \times \underbrace{\sqrt{\dfrac{f_m^A(1-f_m^A)}{n_A} + \dfrac{f_m^B(1-f_m^B)}{n_B}}}\right] = 0,95.$

$= \sqrt{0,00026 + 0,00033} = 0,02451$

$\Leftrightarrow \boxed{IC(p_A - p_B) = [-0,0625 ; 0,0637]}$

Exercice 2 (suite)

Etape 4 Il y a 99% de chance pour que la différence entre la proportion d'étudiants préférant Ralf et la proportion d'autres catégories préférant Ralf soit comprise entre −6,25% et +6,37%. Cet intervalle comprenant le 0 on ne peut pas dire que Ralf parvienne spécifiquement à attirer la préférence des étudiants.

---

(2) Avec un niveau de confiance de 0,90 on utilise

$z_{0,95} = 1,645$

$IC_{(p_A - p_B)\ 90\%} = [0,00059 - 1,645 \times 0,02451 ; 0,00059 + 1,645 \times 0,02451]$

$= [-0,0397 ; 0,0409]$

(3) Lorsque le niveau de confiance baisse l'intervalle s'éclaircit car on est moins restrictif.

(4) Nous avons supposé l'indépendance entre les échantillons A et B : il n'y aurait donc pas de lien entre la proba qu'un étudiant préfère Ralf et la proba que quelqu'un d'une autre catégorie préfère Ralf. Cette hypothèse faisable, ne semble pas plus forte que de supposer l'indépendance au sein de chaque échantillon.