Correction Session 2 (au fi. mi. DU)

COURS

Soit $(X_1, \dots, X_n)$ un échantillon simple au hasard. On note
$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ la "variance empirique" ou "d'échantillon".
celle-ci est un estimateur de la variance de la variable aléatoire parente $X_i \sim \mathcal{D}(\mu, \sigma^2)$.
S_n^2 est un estimateur convergent mais biaisé de σ^2.

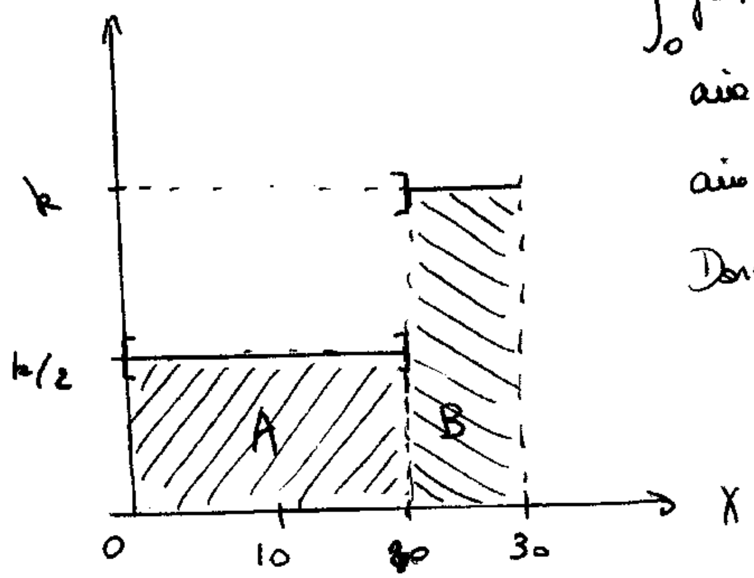
Exercice 1

① $f(x) = \begin{cases} \frac{k}{2} & \text{si } x \le 20 \\ k & \text{si } x > 20 \end{cases}$ la fonction est définie pour $x \in [0, 30]$.

f est une densité si $\int_{\text{support } X} f(x)\,dx = 1$

l'aire située sous le graphe de la densité doit correspondre à 1.

Traçons le graphe de la densité :



$\int_0^{30} f(x)\,dx = \text{aire } A + \text{aire } B$

aire $A = 20 \times \frac{k}{2} = 10k$

aire $B = (30 - 20) \times k = 10k$

Donc $\int_0^{30} f(x)\,dx = 10k + 10k = 20k = 1$

D'où $\boxed{k = 1/20}$

Ex.1 ②

la fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)\,dt$

- Pour $x \leq 20$ $F(x) = \int_0^x \frac{k}{2}\,dt = \left[\frac{kt}{2}\right]_0^x = \frac{kx}{2} - \frac{k\times 0}{2} = \frac{kx}{2}$

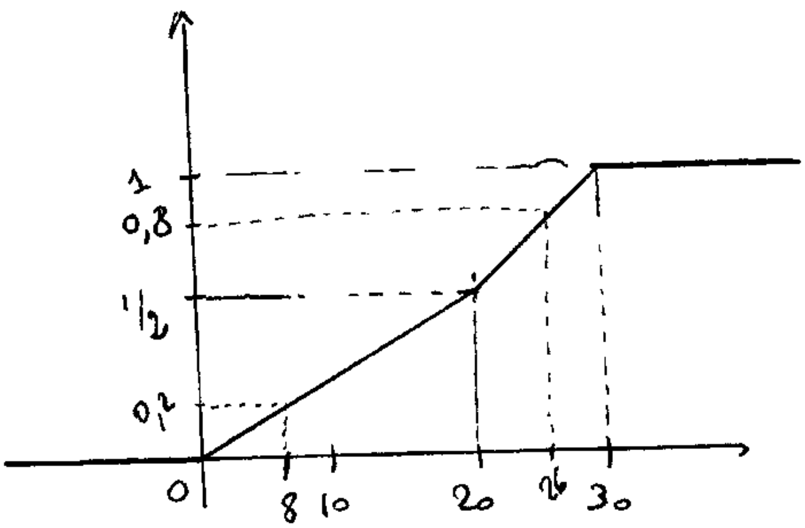
- Pour $x > 20$ $F(x) = \int_0^{20} \frac{k}{2}\,dx + \int_{20}^{x} k\,dt = \frac{k\times 20}{2} + \left[kt\right]_{20}^{x}$

$$F(x) = 10k + kx - 20k = kx - 10k$$

or $k = 1/20$ donc

$$\boxed{F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x}{40} & \text{si } x \leq 20 \\ \frac{x}{20} - \frac{1}{2} & \text{si } x > 20 \\ 1 & \text{si } x \geq 30 \end{cases}}$$

③ graphe de la fonction de répartition F :
il est composé de segment de droites
quand $x = 20$ $F(20) = 20/40 = 1/2$
$F(20) = \frac{20}{20} - \frac{1}{2} = 1/2$. $\Big\}$ on a continuité en 20.



Ex.1 (4)

La loi n'est pas symétrique, la médiane vaut 20 car $F(20) = 1/2$

(5) Le quantile $Q_{20\%}$ est défini tel que $F(Q_{20\%}) = 0,2$

$Q_{20\%} = F^{-1}(0,2)$ il est situé sur le premier segment de la fonction puisque $20\% < 50\%$.

Donc $\frac{Q_{20\%}}{40} = 0,2$ $\boxed{Q_{20\%} = 8}$ $P(X \leq 8) = 20\%$.

De la même façon $Q_{80\%} = F^{-1}(0,8)$ deuxième segment.

Donc $\frac{Q_{80\%}}{20} - \frac{1}{2} = 0,8 \Leftrightarrow \frac{Q_{80\%}}{20} = 0,8 + \frac{1}{2} = 1,3$

$Q_{80\%} = 1,3 \times 20 = 26$ $\boxed{Q_{80\%} = 26}$ $P(X \leq 26) = 80\%$

(6) $Q_{0,05} = F^{-1}(0,05)$

$\Leftrightarrow \frac{Q_{0,05}}{40} = 0,05 \Leftrightarrow Q_{0,05} = 0,05 \times 40 = 2$

$P(X < Q_{0,05}) = 0,05$

Il y a 5% de chances pour que X soit inférieur à 2.

Exercice 2

(1) La population mère est constituée des 4100 électeurs d'une ville.

L'échantillon simple au hasard de taille n.

La variable aléatoire parente X vaut 1 lorsque l'électeur se déclare favorable au candidat B.

$X \sim \mathcal{B}(p)$ où p est la probabilité qu'un individu soit favorable au candidat B.

Ex. 2

① suite

Pour une Bernoulli on sait que $E(X) = p$ et $V(X) = p(1-p)$

Par le théorème central limite :

$\sqrt{n}\,\dfrac{\overline{X}_n - \mu}{\sigma} \underset{\infty}{\sim} N(0,1)$ où $E(X_i) = \mu$ et $V(X_i) = \sigma^2$

X_i étant une suite i.i.d de v.a.

ainsi $\overline{X}_n \underset{\infty}{\sim} N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$

Comme $V(X) = p(1-p) \Rightarrow \sigma = \sqrt{p(1-p)}$ et $E(X) = \mu = p$.

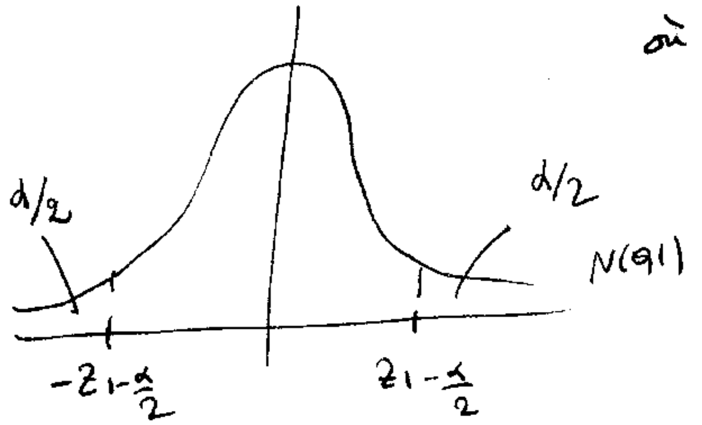
et donc dans le cas discret $\overline{X}_n$ est aussi la fréquence empirique F_n on a $\boxed{F_n \underset{\infty}{\sim} N\left(p, \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right)}$

② On cherche un intervalle de confiance de niveau $1-\alpha$ pour la proba inconnue p tel que $P(a \leq p \leq b) = 1-\alpha$

on centre et réduit F_n : $\sqrt{n}\,\dfrac{F_n - p}{\sqrt{p(1-p)}} \underset{\infty}{\sim} N(0,1)$.

ainsi $P\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\dfrac{F_n - p}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$

où $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$



Ex. 2 (2) suite

$\times \frac{\sqrt{p(1-p)}}{\sqrt{n}}$: $P\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le F_n - p \le z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$

$-F_n$: $P\left(-F_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le -F_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$

$\times(-1)$: $P\left(F_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le F_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$

On remplace $p(1-p)$ par son estimateur $F_n(1-F_n)$

$$\boxed{P\left(F_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{F_n(1-F_n)}{n}} \le p \le F_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{F_n(1-F_n)}{n}}\right) = 1-\alpha}$$

(3) Application de la formule

Pour un niveau $1-\alpha = 90\% = 0,9 \Rightarrow \alpha = 10\% = 0,1$

Donc $1-\frac{\alpha}{2} = 0,95$ Donc $z_{1-\frac{\alpha}{2}} = z_{0,95} = \Phi^{-1}(0,95) = 1,645$

Taille de l'échantillon $n = 40 + 142 + 352 + 28 + 46 + 344 = 952$

fréquence empirique $F_n = \frac{28+46+344}{952} = \frac{418}{952} \simeq 43,91\%$

$z_{1-\frac{\alpha}{2}}\sqrt{\frac{F_n(1-F_n)}{n}} = 1,645 \times \sqrt{\frac{0,4391 \times 0,5609}{952}} \simeq 0,02646$

$P(0,4126 \le p \le 0,4656) = 99$

(4) Il y a 90% de chances pour que le candidat B récolte entre 41,26% et 46,56% des voix.

Avec un seuil d'erreur de 10% le candidat B ne remportera pas les élections

Ex 2 suite

(5) $\hat{P}(X=1/Y=y_1) = \dfrac{28}{28+40} = \dfrac{28}{68} = 0,41176$.

on estime que 41,18 % des cadres sont favorables au candidat B.

$\hat{P}(X=1/Y=y_2) = \dfrac{46}{46+142} = \dfrac{46}{188} = 0,24467$

On estime que 24,47 % des professions intermédiaires sont favorables au candidat B.

$\hat{P}(X=1/Y=y_1) > \hat{P}(X=1/Y=y_2)$ les cadres apparaissent davantage favorables au candidat B que les professions intermédiaires

(6) On souhaite savoir s'il y a indépendance entre la catégorie socio professionnelle et les préférences électorales. si Y est la v.a désignant la CSP et X celle désignant la préférence pour B. on cherche à tester : $\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendants} \\ H_1 : X \text{ et } Y \text{ ne sont pas indépendants} \end{cases}$

on utilise pour cela le test du Khi-deux.

Ex 2

⑤ suite

Calculons les effectifs théoriques sous hypothèse d'indépendance (H_0): n_{ij}^t

Y_j \ X_i	$X = x_0$	$X = x_1$	
$y = y_1$	38,14	29,86	68
$y = y_2$	105,45	82,55	188
$y = y_3$	390,40	305,60	696
	534	418	952

$$n_{ij}^t = \frac{n_i \times n_j}{n}$$

La stat de test $D = \sum_{ij} \frac{(n_{ij} - n_{ij}^t)^2}{n_{ij}^t} \underset{H_0}{\sim} \chi^2((3-1)\times(2-1))$

$\chi^2(2)$

Dans Matrice des distances d_{ij} $\frac{(n_{ij} - n_{ij}^t)^2}{n_{ij}^t}$

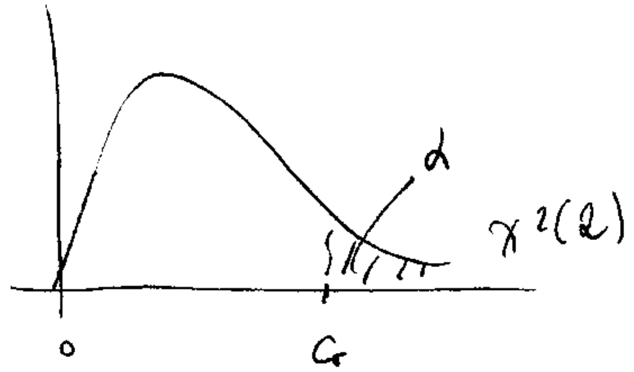
Y_j \ X_i	x_0	x_1
y_1	0,09	0,11
y_2	12,66	16,18
y_3	3,78	4,83

$D = 37,65$.

seuil $\alpha = 10\% = 0,1$

valeur critique $C_r = 4,6$

zone de rejet $\mathcal{C}_r = [C_r, +\infty[$



$D > 4,6$

donc on rejette H_0

Il n'y a pas indépendance avec 10% de chances de se tromper