

Licence Economie Gestion 2nde année. 2023-2024.

TD de STATISTIQUES.

TD N° 1-2 Variables aléatoires continues

Exercice 1 Calculer :

a) $\int_1^2 (5 + 6x^2) dx$

b) $\int_1^3 \sqrt{x+1} dx$ (avant de chercher une primitive, on aura intérêt à écrire l'expression avec une puissance adéquate)

c) $\int_1^4 x \sqrt{x} dx$ (même remarque)

d) $\int_1^4 \frac{1}{(x+1)^3} dx$ (même remarque) e) $\int_1^{+\infty} \frac{1}{(x+1)^3} dx$

f) $\int_1^4 \frac{1}{x+1} dx$

g) $\int_1^4 \exp(2x) dx$ h) $\int_{-\infty}^4 \exp(2x) dx$

i) la dérivée de la fonction s définie ainsi $x \in \mathbb{R} \mapsto s(x) = x e^{3x} - \frac{e^{3x}}{3}$, puis

$\int_1^4 3x \exp(3x) dx$

Exercice 2 Y est une variable aléatoire continue, de densité de probabilité f :

$$\begin{cases} f(x) = a(x+1)^{-4} & \text{pour } x \geq 0 \\ f(x) = 0 & \text{pour } x < 0 \end{cases}$$

a) Calculer a et représenter graphiquement f .

b) Déterminer la fonction de répartition F et en construire une représentation graphique.

c) Calculer $P(1 \leq Y \leq 2)$, $P(1 < Y \leq 2)$, $P(Y < 1)$, $P(Y > 2)$, $P(-3 < Y < 2)$

d) Prouver l'égalité (pour tout $x \neq -1$) $\frac{x}{(x+1)^4} = \frac{1}{(x+1)^3} - \frac{1}{(x+1)^4}$ et Calculer

l'espérance de Y , $E(Y) = \int_{-\infty}^{\infty} t f(t) dt$.

e) Déterminer la médiane de Y : m tel que $P(Y \leq m) = P(Y > m) = 1/2$.

Exercice 3 Soit X une variable aléatoire continue de fonction de répartition F_X :

$$\begin{cases} F_X(t) = 0 & \text{pour } t < 0 \\ F_X(t) = t^3/3 & \text{pour } 0 \leq t \leq 1 \\ F_X(t) = t - t^2/6 - 1/2 & \text{pour } 1 < t \leq 3 \\ F_X(t) = 1 & \text{pour } t > 3 \end{cases}$$

a) Vérifier que F_X possède les propriétés caractéristiques d'une fonction de répartition.

b) Déterminer une densité de probabilité pour X .

c) Déterminer le mode et la médiane d' X .

d) Calculer l'espérance d' X .

e) Calculer variance d' X .

f) Calculer $P(X < 2)$, $P(0 < X < 2)$, $P(2 < X < 4)$.

g) On pose $Y = \ln(X)$. Justifier que ce soit possible. Déterminer la densité et la fonction de répartition d' Y .

Exercice 4 Montrer que g : $\begin{cases} g(x) = x + c & \text{pour } -2 \leq x \leq 2 \\ g(x) = 0 & \text{pour } x < -2 \text{ et pour } x > 2 \end{cases}$

n'est pas une densité de probabilité, quelle que soit la valeur du réel c .

Exercice 5 Soit X une v.a. de densité de probabilité f :

$$\begin{cases} f(x) = \alpha & \text{pour } x \in [-3, 3] \\ f(x) = 0 & \text{pour } x \in]-\infty, -3[\cup]3, +\infty[\end{cases}$$

a) Déterminer α

On s'intéresse à la variable aléatoire $Y = X^2$.

b) Montrer que la variable aléatoire X^2 admet une densité de probabilité, que l'on déterminera.

c) Calculer l'espérance de X^2 en utilisant la loi de probabilité d' X^2

d) Calculer l'espérance de X^2 en utilisant la loi de probabilité d' X .

Exercice 6 X est une variable aléatoire de fonction de répartition F définie ainsi :

$F(t) = \ln(t)$ sur $[1, e]$, $F(t) = 0$ pour $t < 1$, $F(t) = 1$ pour $t > e$.

a) Déterminer une densité de probabilité pour X .

b) Montrer que l'on peut définir la variable aléatoire $Y = \ln(X)$ et déterminer sa densité de probabilité.

c) Montrer que, de façon générale, si X est une variable aléatoire de fonction de répartition strictement croissante et continue, alors $Y = F(X)$ suit une loi uniforme sur $[0; 1]$ (c'est-à-dire de densité constante sur $[0; 1]$, nulle ailleurs).

Exercice 7 $X_1, X_2, \dots, X_n$ sont n variables aléatoires, indépendantes, suivant toutes la même loi de probabilité, d'espérance 1 et de variance 3. Calculer :

a) $E(X_1^2)$, b) $E(X_1 X_2)$, c) $\text{Var}(X_1 X_2)$, d) $\text{Var}(X_1 + X_2 + \dots + X_n)$,

e) $\text{Cov}(2X_1 + X_2, X_1)$ f) $\text{Cov}(X_1 X_2, X_1)$

On pourra utiliser le résultat suivant : si 2 variables aléatoires U et V sont indépendantes, alors toute fonction de U et toute fonction de V sont aussi indépendantes.

Exercice 8 Pour le tirage d'une loterie, un point P peut se déplacer sur un demi-cercle de diamètre $[AB]$, de rayon 1 m, centré en O .

On note α la mesure, entre 0 et π , de l'angle $(\vec{OA}, \vec{OP})$ H la projection orthogonale de P sur la droite (OA) , et $X = \overline{OH}$ l'abscisse de H sur la droite (OA) . On suppose que cette abscisse est une variable aléatoire, admettant une densité de probabilité f constante sur $[-1; 1]$

$$\begin{cases} f(x) = \lambda & \text{pour } x \in [-1, 1] \\ f(x) = 0 & \text{pour } x \in]-\infty, -1[\cup]1, +\infty[\end{cases}$$

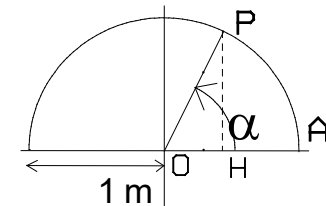
a) Déterminer le réel λ .

b) Déterminer la fonction de répartition d' X

c) Déterminer la fonction de répartition G de la v.a. $Y = \alpha$.

d) Déterminer la densité de probabilité g d' Y

e) Déterminer les modes d' Y .



Licence économie gestion 2nde année. 2023 2024. TD de Statistiques. Mme Volatier et Mr Beau.

Séances de TD 3 à 10.

Lois usuelles

Ex 9-

A

a) X désigne ici une variable aléatoire suivant une loi Gaussienne standard, $\mathcal{N}(0,1)$

Donner: $P(X > 1,7)$ $P(0 < X \leq 0,1)$ $P(-0,3 < X < 0,1)$ $P(X < -0,6)$

b) X désigne ici une variable aléatoire suivant une loi Gaussienne d'espérance 3 et de variance 4, $\mathcal{N}(3,4)$

Donner $P(X < 4)$ $P(1 < X < 4)$ $P(X > 2)$ $P(X \leq 0)$ $P(X^2 < 4)$

c) X désigne ici une variable aléatoire suivant une loi chi-deux de paramètre 10, χ_{10}^2

Donner $P(X > 5)$ $P(X < 16)$

d) X désigne une variable aléatoire suivant une loi de Student de paramètre 15, t_{15} .

Donner $P(X > 2,1)$ $P(X < 3)$ $P(1,3 < X < 3)$

B - chercher des nombres t tel que :

a) X désigne ici une variable aléatoire suivant une loi Gaussienne standard, $\mathcal{N}(0,1)$, Y une variable aléatoire suivant une loi Gaussienne d'espérance 1 et variance 4, $\mathcal{N}(1,4)$,

$P(X < t) = 0,9$ $P(X < t) = 0,01$ $P(X > t) = 0,7$ $P(-t < X < t) = 0,95$

$P(0 < Y < t) = 0,5$

c) X désigne ici une variable aléatoire suivant une loi chi-deux de paramètre 5, χ_5^2

$P(X > t) = 0,3$ $P(X < t) = 0,02$

$P(t < X < t') = 0,90$ (chercher 2 valeurs t et t')

d) X désigne une variable aléatoire suivant une loi de Student de paramètre 20, t_{20} .

$P(X > t) = 0,05$ $P(X < t) = 0,05$ $P(-t < X < t) = 0,9$

$P(t < X < t') = 0,99$ (chercher t et t')

C – X et Y sont deux variables aléatoires Gaussiennes indépendantes satisfaisant à :

$E(X) = 2$, $\text{Var}(X) = 3$, $E(Y) = 1$, $\text{Var}(Y) = 2$.

a) Montrer que X, Y forment un vecteur Gaussien.

b) Montrer que $X+Y$ suit une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ et trouver la valeur des deux paramètres μ et σ^2

c) Calculer $\text{cov}(X+Y, 2X+3Y)$

d) Trouver un nombre a tel que $X+Y$, et $aX+Y$ soient indépendantes.

Convergences

Ex 10- On projette de créer des espaces clos sur le campus pour favoriser la qualité de vie des écureuils. On aimerait connaître l'avis des étudiants. Pour cela, on réalise **un sondage** sur un nombre restreint, n, d'étudiants. On interroge n étudiants, choisis au hasard. Ils n'ont que 2 réponses possibles, favorable ou défavorable. On numérote les réponses de 1 à n. Pour chaque étudiant interrogé, de n°k, on définit une variable aléatoire R_k en posant :

$R_k = 1$ si la réponse est favorable, $R_k = 0$ sinon.

La proportion p d'étudiants favorables parmi tous les étudiants du campus reste inconnue, mais pour réaliser les calculs de cet exercice, on supposera qu'elle est de **p = 55%**.

a) Pour chaque n° k, la variable aléatoire R_k suit une loi usuelle, vue en première année, laquelle? Que vaut

son paramètre ?

Rappeler quelles sont $E(R_k)$ et $Var(R_k)$.

b) On suppose que **les variables aléatoires R_k sont indépendantes** dans leur ensemble. Qu'est-ce que cela signifie en termes de probabilités ? Qu'est-ce que cela signifie concrètement ? Comment doit-on choisir les étudiants interrogés pour que cette hypothèse soit satisfaite ?

c) On définit la variable aléatoire $S_n = R_1 + R_2 + \dots + R_n$.

Quel est le lien entre S_n et le nombre de réponses favorables parmi les étudiants interrogés ?

Montrer que **S_n suit une loi binomiale**, dont on précisera les paramètres. Montrer que la variable aléatoire $Z_n = S_n / n$ est la proportion d'étudiants favorables aux espaces clos, parmi les étudiants interrogés.

d) Pour $n = 4$, évaluer $E(Z_n)$, $Var(Z_n)$, et $P(Z_n < 50\%)$

e) Mêmes questions pour $n = 6$.

f) Evaluer $E(Z_{200})$, $Var(Z_{200})$.

g) Que penser de l'évaluation de $P(Z_{200} < 0,50)$?

Que penser de l'évaluation d'un entier n tel que $P(Z_n < 0,50) < 0,10$?

A quelles préoccupations concrètes correspondent ces questions ?

h) A l'aide de l'inégalité de Bienaymé Tchebichev, montrer que

$P(Z_{200} < 0,50)$ est majorée par un nombre que l'on précisera.

i) Que se passe-t-il pour cette probabilité lorsque le nombre n d'étudiants interrogés croît ? Comment peut-on l'interpréter concrètement ?

j) On désigne par $F_{S_{200}}$ la fonction de répartition de la variable aléatoire S_{200} . Exprimer la probabilité $P(Z_{200} < 0,50)$ en fonction de $F_{S_{200}}$.

k) Pour tout entier n positif on définit une nouvelle variable aléatoire

$$U_n = \frac{\sqrt{n} \cdot (Z_n - 0,55)}{\sqrt{0,55 \cdot (1 - 0,55)}}. \text{ Exprimer } P(Z_{200} < 0,50) \text{ en fonction de la fonction de répartition de } U_{200}.$$

l) On suppose ici que pour $n \geq 200$, une bonne approximation de la fonction de répartition de U_n est fournie par la fonction de répartition de la loi normale standard $\mathcal{N}(0,1)$. Pourquoi ? A l'aide de la table de cette loi, déduire une approximation de $P(Z_{200} < 0,50)$ m) Comparer les résultats de k) et l).

Ex 11- Un service client reçoit des appels téléphoniques. On suppose que leurs durées sont indépendantes et suivent une même loi de probabilité. L'observation suggère que l'espérance de ces appels est 6 minutes et leur variance 4 minutes².

a) A l'aide de l'inégalité de Bienaymé Tchebitchev, majorer la probabilité que la durée d'un appel dépasse 11 minutes.

a) A l'aide de l'inégalité de Bienaymé Tchebitchev, majorer la probabilité que la durée moyenne de 100 appels dépasse 6,25 minutes – soit 6 minutes et 15 secondes –.

b) Justifier l'utilisation d'une approximation par une loi normale et évaluer la probabilité que la durée moyenne de 100 appels dépasse 6,25 minutes.

Ex 12 - On lance un dé n fois et on considère les variables aléatoires $S, X_1, X_2, \dots, X_n$ suivantes :

S : nombre de résultats pairs obtenus en n lancers.

Pour chaque entier k entre 1 et n , $X_k = 1$ si le résultat du k ème lancer est pair, $X_k = 0$ sinon.

a) Quelles sont les lois de probabilité des variables aléatoires $X_1, X_2, \dots, X_n$?

b) Quelle est la loi de probabilité de S ? Quelles sont $E(S/n)$ et $Var(S/n)$?

c) L'inégalité de Bienaymé et Tchebitchev indique un nombre A solution de

$$\left(P\left(\frac{1}{2} - 0,1 < \frac{S}{200} < \frac{1}{2} + 0,1\right) \right) \geq A. \text{ Trouver ce minorant } A.$$

d) A l'aide de l'inégalité de Bienaymé et Tchebitchev, évaluer un nombre n de lancers à partir duquel

$$\frac{1}{2} - 0,05 < \frac{S}{n} < \frac{1}{2} + 0,05, \text{ avec une probabilité d'au moins } 0,95$$

e) évaluer un nombre n satisfaisant cette même condition, mais cette fois en utilisant une approximation par une loi $\mathcal{N}(0,1)$.

Ex 13— X est une variable aléatoire de loi $\mathcal{N}(0,1)$. Pour tout entier positif n on définit une variable aléatoire $A_n = (-1)^n X$.

a) Montrer que pour tout entier n , A_n suit la loi $\mathcal{N}(0,1)$.

b) Montrer que la suite de variables aléatoires $(A_n)_{n \in \mathbb{N}}$ **converge en loi** vers X

c) Evaluer $P(\{A_n - X > 0,1\})$ pour n pair, puis pour n impair

d) Montrer que la suite de variables aléatoires $(A_n)_{n \in \mathbb{N}}$ **ne converge pas en probabilité** vers la variable aléatoire X .

Ex 14— Prouver que si $(X_n)_{n \in \mathbb{N}}$ est une suite de variables aléatoires convergeant en loi vers une variable aléatoire X de loi de probabilité continue, alors

$(\exp(X_n))_{n \in \mathbb{N}}$ converge en loi vers la variable aléatoire $\exp(X)$.

Ex 15— Une compagnie aérienne assure quotidiennement un vol avec une capacité de 400 passagers. En pratique, 15% des clients qui ont réservé ne prennent finalement pas l'avion.

a) 300 réservations ont été faites pour un jour. Quelle est la probabilité qu'au moins 280 de ces clients prennent effectivement l'avion?

b) Quel serait le nombre maximum de places à réserver pour que la probabilité $P(\text{nombre de passagers} > 400)$ soit inférieure à 0,001 ?

Ex 16— Un grossiste en poissonnerie se fournit auprès d'un éleveur. Pour les poissons de cette espèce, il existe un produit de synthèse qu'on administre aux alevins pour les faire grandir plus vite. Ce produit est efficace, mais puisqu'il n'est pas naturel les consommateurs estiment que les poissons qui y ont été exposés sont de moindre qualité. Lorsqu'on utilise ce produit, les alevins mâles sont un peu fragilisés, et on trouve en moyenne 60% de femelles chez les poissons adultes, au lieu de 50% sans le produit.

L'éleveur a indiqué au grossiste qu'il n'utilise pas ce produit. Bien sûr **le grossiste le croit a priori**, mais une quantité significativement plus élevée de femelles dans les poissons reçus pourrait l'amener à douter. Le grossiste reçoit un lot de N poissons, parmi lesquels il compte f femelles. Il suppose que ces poissons ont été choisis parfaitement au hasard dans la production et que le fait que certains soient des femelles n'influe pas sur la probabilité que d'autres le soient.

Pour un lot de N poissons reçus, on cherche pour quelles valeurs de f on pourrait envisager que les poissons aient été exposés au produit incriminé.

a) Ici $N = 100$. Il y a $f = 57$ femelles. En supposant que la proportion de femelles dans la production totale soit de 50%, on cherche à évaluer la probabilité qu'il y ait au moins 57 femelles dans l'échantillon.

Modéliser le problème, puis évaluer $P(f \geq 57)$ à l'aide d'une approximation par la loi Gaussienne, soigneusement justifiée.

Peut-on penser que l'excès de poissons femelles constaté est assez peu probable pour envisager que l'éleveur n'ait pas respecté ses engagements?

b) Même question pour $N=200$, $f = 115$, puis $N=200$, $f = 125$.

c) Le grossiste adopte l'attitude suivante. Pour $N = 200$, il fixe un nombre limite $f_{\max}$ qui ne pourrait être atteint ou dépassé, sans produit, qu'avec une probabilité 0,01.

Si ce nombre $f_{\max}$ est dépassé, il investit dans le conseil d'une société qui analysera les poissons pour savoir s'ils ont été en contact avec le produit.

Déterminer $f_{\max}$.

d) On peut juger que dans ce calcul, l'hypothèse que la proportion soit de 50% est favorisée par rapport à l'hypothèse que la proportion soit de 60%. Pourquoi peut-on émettre ce jugement?

Pourquoi dans cette situation favorise-t-on la première hypothèse?

Ex 17— On modélise la durée de vie d'une abeille ouvrière par une variable aléatoire T de densité de probabilité

$$f: \quad f(t) = 0 \text{ pour } t < 0 \quad f(t) = \lambda t e^{-\alpha t} \text{ pour } t \geq 0.$$

- a) Sachant que l'espérance de T est $E(T) = 30$ jours, déterminer λ et α .
 b) On considère 1000 ouvrières, et la variable aléatoire de durée de vie moyenne:
 $Y = (T_1 + T_2 + \dots + T_{1000})/1000$
 Déterminer l'espérance d'Y, $E(Y)$, et variance d'Y, $Var(Y)$.

- c) Déterminer la valeur de t pour laquelle : $P(|Y - 30| \geq t) \leq 0,01$

- c1) en utilisant l'inégalité de Bienaymé - Tchebichev ,
 c2) puis en utilisant une approximation par une loi normale.

Ex 18- Voici quelques extraits de tables de la loi Binomiale. Pour une variable aléatoire Y suivant la loi B (n ; 0,6) les tableaux suivants donnent $F_Y(k) = P(Y \leq k)$

n = 30	k	10	12	15	17	18	20	25	30
	$P(Y \leq k)$	0,0029	0,0212	0,1754	0,4215	0,5689	0,8237	0,9985	1,0000

n = 900	k	400	500	520	540	570	600
	$P(Y \leq k)$	0,0000	0,0037	0,0926	0,5127	0,9814	1,0000

- a) En écrivant que Y est la somme de n variables aléatoires indépendantes de loi de Bernoulli de paramètre 0,6, montrer qu'on peut évaluer $P(Y \leq k)$ à l'aide d'une approximation par une loi Gaussienne standard.

Indiquer quelles valeurs cette approximation indique pour les cas suivants :

n = 30, k = 12, 15, 25 n = 900, k = 500, 540, 570

- b) L'approximation vous paraît-elle convenable ?

- c) Faire les mêmes calculs en écrivant, à la place de $P(Y \leq k)$: $P(Y \leq k+1/2)$. Pour la loi Binomiale, cela change-t-il quelque chose ? L'approximation par une loi normale paraît-elle meilleure ? Cette manipulation est appelée « correction de continuité ».

- d) L'inégalité de Berry-Esseen indique que, pour une suite de variables aléatoires $X_1, X_2, \dots, X_n, \dots$, indépendantes, de même loi, d'espérance μ et de variance σ^2 , en notant Φ la fonction de répartition de la loi N (0,1) et G_n la fonction de répartition de la variable aléatoire $\frac{\sqrt{n}(X_n - \mu)}{\sigma}$:

$$\sup_{x \in \mathbb{R}} |G_n(x) - \Phi(x)| \leq \frac{33}{4} \frac{E(|X - \mu|^3)}{\sigma^3} \frac{1}{\sqrt{n}}$$

Qu'indique cette inégalité pour une loi Binomiale de paramètres n=30 et p=0,6?

Et pour une loi Binomiale de paramètres n = 900 et p = 0,6 ?

Estimation

Ex 19- On cherche à estimer la proportion p de ménages équipés de congélateur dans l'Essonne. Dans le Nord de l'Essonne on a tiré n_1 ménages au hasard, dans le Sud de l'Essonne, on a tiré n_2 ménages au hasard, $n_2 < n_1$. On suppose ici que la proportion p_1 de ménages équipés dans le Nord de l'Essonne est égale à la proportion p_2 de ménages équipés dans le Sud de l'Essonne, et à la proportion p de ménages équipés dans toute l'Essonne.

On propose ces 3 estimateurs de p : F_1 = proportion de ménages équipés dans le premier échantillon, F_2 = proportion de ménages équipés dans le second échantillon, $F_3 = \frac{n_1 F_1 + n_2 F_2}{n_1 + n_2}$.

- a) Calculer les espérances et les variances de F_1, F_2, F_3 .
 b) Ces estimateurs ont-ils un biais ?
 c) Quel critère statistique utilise-t-on pour choisir un estimateur parmi d'autres ?
 d) Selon ce critère, quel est le meilleur estimateur de p parmi F_1, F_2, F_3 ?

Ex 20 On dispose d'un échantillon $X_1, X_2, \dots, X_{20}$ d'une loi géométrique de paramètre p inconnu.

8	11	3	5	8	8	6	15	17	3
4	15	8	9	2	8	17	5	5	37

voilà les observations :

Rappel : si une variable aléatoire X suit la loi géométrique de paramètre p , alors $E(X) = 1/p$, $\text{Var}(X) = (1-p)/p^2$, et pour tout entier $k \geq 1$ $P(X=k) = p(1-p)^{k-1}$.

- Rappeler ce qu'est un échantillon
- proposer un estimateur du paramètre p en utilisant la méthode des moments.
- Proposer une estimation de p .

Ex 21

On étudie le nombre de pannes d'un type d'appareils dans un intervalle de temps T . On suppose que le nombre de pannes de l'appareil n° i , noté N_i , suit une loi de Poisson de paramètre cT . On étudie k appareils et on note ($n_1, n_2, \dots, n_k$) le nombre de pannes observées durant le temps T pour chaque appareil.

1°) On veut estimer cT .

- Que représente cT ?
- On considère la statistique $\bar{N}$. Calculer son espérance et sa variance.
- Montrer que $\bar{N}$ est un estimateur sans biais et consistant de cT .
- Notons $\bar{n}$ la réalisation de $\bar{N}$. Préciser les différences entre $\bar{n}$, $\bar{N}$, et cT .

2°) On veut maintenant estimer c .

- Que représente concrètement le paramètre c dans la situation étudiée?
- Trouver un estimateur sans biais et consistant de c .

3°) Voici les nombres de pannes constatées pour $T=1000$ heures de fonctionnement sur 30 appareils. Donner une estimation de c correspondant à la réponse précédente.

5	4	6	2	0	3	4	2	4	6	2	7	3	2	4
4	5	3	2	3	2	0	4	2	3	3	1	1	5	4

Ex 22- - Une usine met en place une nouvelle chaîne de production de photocopieuses. Elles sont testées en fin de chaîne et doivent subir un second réglage si elles produisent plus de 5 bourrages de papier.

Voici les nombres de bourrages constatés lors du test de 30 photocopieuses choisies au hasard.

2	3	6	0	1	3	1	10	7	8	1	3	7	1	0
1	3	0	4	1	5	2	0	3	0	10	4	1	0	2

- Donner un estimateur de la proportion de photocopieuses défectueuses,
- Calculer une estimation de cette proportion.
- Donner un **intervalle de confiance** à 95% de cette proportion.
- En tout 800 photocopieuses sont sorties de cette chaîne. Est-il réaliste de prévoir qu'au plus 220 d'entre elles doivent être réglées à nouveau?

Ex 23-

Un sondage cité par un quotidien, la dépêche du midi, le 21 02 2022 – avant la guerre en Ukraine –, <https://tinyurl.com/sondage2102>,

« crédite Emmanuel Macron de 24,5 % des intentions de vote, en baisse de 0,5 %, alors que Marine Le Pen dépasse Eric Zemmour, 16,5% (...) contre 16 % – »

Ce sondage portait sur 1502 personnes. On fera les calculs suivants en imaginant un sondage qui fournirait les mêmes proportions en suivant la méthode aléatoire, c'est-à-dire **en interrogeant des personnes choisies au hasard et supposées indépendantes**.

En réalité, ce sondage utilise la méthode des quotas, qui est moins coûteuse à mettre en œuvre. Une méthode de redressement est de plus appliquée pour remédier aux manques de sincérité éventuels des réponses.

Depuis 2016, une loi impose en France de publier une marge d'erreur avec chaque sondage. Faute de formule applicable à la méthode des quotas et au redressement, cette marge est estimée par la demi largeur de l'intervalle de confiance de niveau 0,95 d'un sondage aléatoire pour un échantillon de même taille. Avant cette loi, seule la taille de l'échantillon était publiée.

- a) Après avoir modélisé le problème, proposer un intervalle de confiance pour chacun des 3 résultats.
 b) Que peut-on penser de leurs comparaisons ?
 c) Que penser des formulations « 24,5 % », « 16,5% »?

Ex 24- Un loueur de bicyclettes s'intéresse au nombre H d'heures de location par jour. Il suppose que cette variable aléatoire suit une loi Gaussienne $N(\mu, \sigma^2)$. Il a relevé ces valeurs 20 jours consécutifs.

58	135	144	165	137	138	221	89	197	86
152	230	157	221	161	199	125	80	131	192

- a) **En se fondant sur ces données**, indiquer un estimateur et une estimation des paramètres μ et σ^2 par la méthode des moments (on peut montrer que l'estimateur indiqué par la méthode des moments pour σ^2 a un léger biais, pour l'éviter on multiplie cet estimateur par $\frac{2}{n}$).
 Une journée n'est rentable qu'à partir de 50 heures de location.
 b) Evaluer la probabilité qu'il y ait moins de 50 heures de location dans la journée.
 c) Déterminer un intervalle de confiance, au niveau de confiance 0,9 de μ et σ^2 .
 d) Comment modifier la réponse de b) pour tenir compte de c) ?
 e) Quels commentaires pouvez vous faire sur les données fournies, du point de vue de la méthode statistique ?

Tests d'hypothèses sur une ou deux proportions

Ex 25 : Cet exercice est une introduction aux tests, l'énoncé est plus détaillé qu'à l'examen. L'exercice 28 vous présentera un énoncé d'examen.

Le responsable d'approvisionnement d'un petit groupe de supermarchés a commandé 275 boîtes de jeux de construction, à sa centrale d'achats. Alors que sa commande va être livrée, son interlocuteur de la centrale d'achats lui annonce qu'à la suite d'une erreur, les boîtes qu'il va recevoir sont issues d'un stock (noté ici S_1) dans lequel 20% des boîtes, réparties au hasard, ont un défaut d'emballage : les inscriptions sont en langue Paumutu septentrionale. Il faudra les renvoyer. On lui annonce aussi qu'il a tout de même relativement de la chance : un autre stock (noté ici S_2) comprenait 25% de ces boîtes.

stock de boîtes	S_1	S_2	chez le responsable
nombre de boîtes	non précisé	non précisé	275
proportion de boîtes défectueuses	20%	25%	inconnue pour les questions a) à d) 62/275 question e)

Ce responsable cherche à savoir si l'on peut raisonnablement accepter que les boîtes viennent du stock S_1 ou s'il peut penser qu'elles viennent de S_2 , contrairement à ce qu'a annoncé son collègue de la centrale d'achats. Il formule donc 2 hypothèses :

H_0 : les boîtes proviennent du stock S_1

H_1 : les boîtes proviennent du stock S_2

Si la proportion D de boîtes défectueuses parmi les 275 reçues est inférieure ou égale à un seuil d_{max} qu'il va fixer, le responsable acceptera H_0 , sinon il rejettera H_0 et acceptera H_1 .

On cherche ici à déterminer d_{max} , puis à exécuter le test.

a) Modéliser le problème, en introduisant les variables aléatoires adéquates et en exprimant la proportion D en fonction d'elles.

b) En déduire la valeur de la constante d_{max} telle que

$$0,05 = P_{H_0 \text{ vraie}}(D > d_{max}) \quad ,$$

c'est-à-dire qu'il y ait une probabilité 0,05, évaluée en supposant l'hypothèse H_0 vraie, que la proportion de boîtes défectueuses observée soit supérieur à $d_{\max}$.

c) Quel est alors le risque de rejeter l'hypothèse H_0 dans le cas où elle est vraie ?

d) Dédurre de la valeur de $d_{\max}$ trouvée en b) la probabilité que la proportion observée soit inférieure à $d_{\max}$, cette fois dans le cas où H_1 est vraie.

e) On observe finalement 62 boîtes défectueuses parmi les 275 livrées. Que doit-on décider ?

f) Dans ce test, quels sont :

- le risque d'erreur de 1ère espèce - la statistique utilisée - la règle de décision - le risque d'erreur de 2nde espèce - la puissance

g) parmi les 2 hypothèses H_1 et H_0 , une d'elles est favorisée a priori. Laquelle ? Comment cela se traduit-il dans la construction du test ? Pourquoi favoriser celle-là et pas l'autre ?

Ex 26 On croit avoir remarqué que dans certaines conditions, chez une espèce de mouche, il y a légèrement moins de mâles que de femelles. On note p : proportion de mâles. On veut tester l'hypothèse :

$H_0: p = 0,5$ contre l'hypothèse $H_1: p = 0,48$. On appelle α le risque d'erreur de 1ère espèce, β celui de seconde espèce. On observe 1 000 mouches prises au hasard.

a) Construire un test pour un risque de 1ère espèce de 0,05 (proposer une statistique, indiquer la loi qu'elle suit, proposer une règle de décision).

b) Sur une population de 1000 mouches, on compte 487 mâles. Quelle est la conclusion du test ?

c) Evaluer le risque d'erreur de seconde espèce.

d) En quoi peut-on juger que l'hypothèse H_0 est favorisée par rapport à H_1 ?

e) Pourquoi a-t-on favorisé cette hypothèse ?

g) On note ici n le nombre de mouches observées. Pour un risque d'erreur de 1ère espèce $\alpha = 0,05$, exprimer en fonction de n la règle de décision, puis le risque d'erreur de 2nde espèce. Montrer que ce risque diminue lorsque n augmente. Pour quelle valeur de n aurait-on un risque de 2nde espèce de 0,10 ?

f) Quelles seraient les différences avec le cas suivant : $H_0: p = 0,5$ contre $H_1: p < 0,48$?

Ex 27 Inspiré d'un examen

Le Centre de Danse du Marais Poitevin organise un festival de danse international. Parmi les spectacles certains sont fondés sur des danses traditionnelles, en particulier : le Funana, la Capoeira, le Otea.

Une enquête a été faite auprès de 300 amateurs de Capoeira, 178 d'entre eux ont déclaré connaître le Funana. Dans une enquête similaire auprès de 200 autres personnes amateurs d'Otea, 130 connaissaient le Funana.

Le Funana serait-il plus connu des amateurs de Otea que des amateurs de Capoeira?

Modéliser le problème, en indiquant les théorèmes et les approximations utilisées, leurs conditions d'utilisation, et comment celles-ci se traduisent, pratiquement, sur les données.

Construire un test, en utilisant un risque d'erreur de 1ère espèce de 0,05.

Conclure.

Ex 28 inspiré du même examen :

Les deux années dernières, environ 12% des spectateurs avaient parcouru plus de 300 km pour venir au festival. Les organisateurs se sont fixé pour l'avenir un objectif de 17% de spectateurs venant de plus de 300km.

Sur les 440 premières réservations, 65 émanaient de personnes qui parcourront plus de 300 km pour venir.

Le directeur du festival voudrait savoir s'il peut raisonnablement penser que son objectif va être atteint.

Il vous demande de construire un test statistique, en vous fondant sur les 400 premières réservations, avec un risque de 1ère espèce de 0,10.

1. A vous de modéliser le problème, de construire un test statistique, en choisissant l'hypothèse nulle, l'hypothèse adverse, en justifiant ce choix, et enfin de conclure.
2. Evaluer le risque d'erreur de 2nde espèce
3. Quelles seraient les différences si l'objectif était de 15% au lieu de 17% ?
4. Quelles seraient les différences si la proportion à tester était « plus de 17% » à la place de « 17% » ?
5. Vous annoncez le résultat du test au directeur du festival. Pour tenir compte de 3. quel commentaire pourriez vous ajouter ?
6. Du point de vue de la méthode statistique, la construction de ce test sur cet exemple présente un défaut. Lequel ?

Séries chronologiques

Ex 29 Voici les ventes, par année, de la fourgonnette JMN Turbo-Cebicev sur l'Essonne.

Année	2018	2019	2020	2021	2022	2023
Nbre de modèles vendus	998	950		650	490	502

Le nombre de modèles vendus en 2020 n'est pas connu.

- a) Déterminer la tendance linéaire
- b) En déduire une prévision pour le nombre de modèles vendus en 2025.

Ex 30 Voici le chiffre d'affaires, par trimestre, d'une librairie.

Année	2021				2022				2023			
Trimestre	JFM	AMJ	JOS	OND	JFM	AMJ	JOS	OND	JFM	AMJ	JOS	OND
Ventes (x 1000 €)		6	11	6	5	8	14	7	6	9,5	12	7

Remarquez qu'il manque 2 trimestres : le 1er de 2021 et le 3ème de 2023.

- a) Déterminer la tendance linéaire.
- b) Comment juger de l'opportunité de tenir compte de variations saisonnières ?
- c) Calculer les coefficients saisonniers.
- d) Etablir la série corrigée de variations saisonnières.
- e) Etablir la série ajustée.
- f) Etablir une extrapolation des ventes au troisième trimestre 2020.

Tests d'adéquation

Ex 31

On a observé, pendant 100 jours, le nombre d'infractions routières graves à un carrefour.

Nombre d'accidents	0	1	2	3	4 et plus
Nombre de jours	35	38	20	5	2

- a) Tester l'ajustement à une loi de Poisson de paramètre 2
- Rappeler l'hypothèse nulle, l'hypothèse adverse, la règle de décision.

Rappels : une variable aléatoire X suit une loi de Poisson si X ne prend que des valeurs entières positives ou nulles, avec pour tout k entier positif ou nul : $P(X=k) = \frac{\lambda^k}{k!} \exp(-\lambda)$. Dans ce cas $E(X) = \lambda$, $Var X = \lambda$.

- b) Tester l'ajustement à une loi de Poisson dont on estimera le paramètre.

Ex 32 Une entreprise de plusieurs milliers de salariés propose à ses employés de regrouper les RTT en choisissant une

semaine de congés parmi 5 semaines possibles, notées ici A, B, C, D.

Voici ce qu'on a obtenu pour les 100 premières réponses, supposées indépendantes :

Semaine	A	B	C	D	E	Ensemble
Nombre de réponses pour cette semaine	16	14	23	26	21	100

Peut-on considérer que les choix des différentes semaines sont équiprobables ? Répondre à l'aide d'un test du χ^2 d'adéquation.

Ex 33 On relève la production en une journée de 30 chèvres, exprimée en litres.

4,69	3,88	5,85	3,85	4,30	4,82	4,20	4,86	4,94	4,49	2,53	4,90	5,72	5,60	4,02
3,48	4,55	5,86	3,86	3,89	1,89	3,69	4,60	5,07	3,72	4,78	4,24	5,33	4,30	5,09

Indications : $\sum_{i=1}^{30} l_i = 133$; $\sum_{i=1}^{30} l_i^2 = 613$

On répartit ces données en 4 classes, remplir le tableau suivant

Production	Moins de 3 litres	[3 , 4[	[4, 4,5[	[4,5; 5[	5 litres et plus
Effectifs					

- Donner des estimateurs sans biais de μ et σ^2 , et en déduire des estimations.
- En supposant que les valeurs de μ et σ^2 soient égales à ces estimations, quelle est, pour chaque classe du tableau, la probabilité que la production soit dans cette classe ?
- Tester l'adéquation des données à une loi Gaussienne, à l'aide d'un test de Chi2.
- On préfère souvent utiliser un autre test pour une loi Gaussienne et plus généralement pour une loi continue (test de Kolmogorov, test de .Shapiro Wilk dans le cas d'une loi Gaussienne), quel défaut pourriez vous imputer à la méthode utilisée dans cet exercice ?