M1 MEEF 2016-2017
Séquence 4

Feuille 4. La statistique.

Exercice 1. On considère n points du plans $((x_i, y_i))_{i\in\{1,\dots,n\}}$. Pour (a, b) deux réels quelconques, on approche ce nuage de point par une droite d'équation $y = ax + b$. On définit la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ par, pour tout $(a, b) \in \mathbb{R}^2$,

$$f(a,b) := \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - (ax_i + b))^2\right].$$

Considérons M la variable aléatoire à valeurs dans $\mathbb{R}^2$ qui charge uniformément les points (x_i, y_i)

$$M \sim \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}.$$

On définit alors X comme étant l'abscisse de M et Y comme étant l'ordonnée de M. On a alors $(X, Y) = M$ et dans cette construction on a pour tout $i \neq j$, $\mathbb{P}(X = x_i, Y = y_j) = 0$. De plus,

$$X \sim \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \qquad \text{et} \qquad Y \sim \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}.$$

1. Que représente $nf(a, b)$?
2. Ecrire $f(a, b)$ en fonction de $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{E}(X^2)$, $\mathbb{E}(Y^2)$ et $\mathbb{E}(XY)$.
3. Justifier que f est différentiable, calculer ses dérivées partielles, exprimer sa différentielle ainsi que son gradient.
4. Trouver tous les points critiques de f.
5. La fonction f admet-elle un maximum global ? Justifier.
6. Montrer que l'on peut écrire f de la façon suivante :

$$f(a,b) = (b + a\mathbb{E}(X) - \mathbb{E}(Y))^2 + \mathrm{Var}(X)\left(a - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\right)^2 + \mathrm{Var}(Y) - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)}.$$

7. En déduire la nature du point critique de la question 4 que l'on notera (a_0, b_0). Cet extremum est-il global ? A quoi correspond graphiquement la droite $y = a_0x + b_0$?
8. Calculer $f(a_0, b_0)$ et exprimer cette quantité en fonction du coefficient de corrélation r et de $\mathrm{Var}(Y)$. En déduire pourquoi on demande que r soit proche de 1 en valeur absolue pour une bonne approximation.

Exercice 2. Le tableau suivant représente la distribution de la taille des coquilles d'individus adultes dans l'espèce *capea nemoralis* (escargots des bois). Deux régions ont été étudiées, A : région de Montluçon, B : région d'Egletons.

Intervalle (en mm)	[16 ;17[	[17 ;18[	[18 ;19[	[19 ;20[	[20 ;21[	[21 ;22[	[22 ;23[
Nombre d'individus de A	0	100	200	950	210	200	45
Nombre d'individus de B	20	40	110	280	80	60	66

Intervalle (en mm)	[23 ;24[	[24 ;25[	[25 ;26[	[26 ;27[	[27 ;28[	[28 ;29[
Nombre d'individus de A	8	0	1	0	1	0
Nombre d'individus de B	1	0	0	0	0	0

1. Pour chacune des régions, déterminer le mode, les fréquences, les fréquences cumulées, la moyenne, la classe médiane, les classes quartiles et l'écart-type.

2. Déterminer la droite d'équation $y = ax + b$ passant par $(19,0,17)$ et $(20,0,73)$ et en déduire une valeur interpolée de la médiane pour la distribution A.

Exercice 3. Dans une ferme industrielle, le service vétérinaire veut modifier le régime alimentaire des vaches, dans le but d'augmenter la production laitière. Pour cela, on a choisi au hasard 15 vaches que l'on a nourries pendant un mois avec l'aliment habituel et l'on a relevé pour chaque vache X la production quotidienne moyenne de lait exprimée en kg. Puis, on a nourri ces mêmes vaches pendant un mois avec le nouvel aliment et on a relevé de même Y la production quotidienne moyenne de chaque vache.

N° de la vache	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X en kg/jour	27,6	23,4	25,2	28,2	28,8	25,8	27	27	29,4	28,2	30	28,2	32,4	29,4	30
Y en kg/jour	28,8	25,6	26,4	28	31,2	27,2	28,8	28	29,6	29,2	28,4	29,6	31,2	32	29,2

On note x_i la production de la vache i avec l'aliment habituel et y_i avec l'aliment nouveau. On note également $N = 15$, X une variable aléatoire de loi $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ et Y de loi $\frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$.

1. Calculer $\mathbb{P}(X = x_1)$ et $\mathbb{P}(Y = y_1)$. Quel est le terme statistique pour désigner $\mathbb{P}(X = 27,6)$?

2. On pose $U = \frac{X-28{,}2}{0{,}6}$, $u_i = \frac{x_i-28{,}2}{0{,}6}$, $V = \frac{Y-28{,}8}{0{,}8}$ et $v_i = \frac{y_i-28{,}8}{0{,}8}$. On sait que

$$\sum_{i=1}^{15} u_i = -4; \qquad \sum_{i=1}^{15} u_i^2 = 190; \qquad \sum_{i=1}^{15} v_i = 1,5; \qquad \sum_{i=1}^{15} v_i^2 = 67,75; \qquad \sum_{i=1}^{15} u_i v_i = 91.$$

   Exprimer ces sommes en termes probabilistes de U et de V.

3. Écrire X et Y en fonction de U et V et en déduire $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X,Y)$.

4. Calculer la valeur moyenne et l'écart type des $(x_i)_{1\leqslant i\leqslant 15}$ et des $(y_i)_{1\leqslant i\leqslant 15}$.

5. Calculer le coefficient de corrélation r entre $(x_i)_{1\leqslant i\leqslant 15}$ et $(y_i)_{1\leqslant i\leqslant 15}$, que peut-on en conclure ?

Exercice 4. On réalise une enquête auprès de 1000 ménages. On s'intéresse à la liaison entre X : « le nombre d'enfants à charge du ménage » et Y : « les dépenses annuelles de fournitures scolaires » (données en dizaine d'euros).

	Y=[0 ;4[	Y=[4 ;10[	Y=[10 ;20[	Y=[20 ;40[
X=1	322	12	2	0
X=2	14	230	116	36
X=3	0	0	20	248

1. Calculer le coefficient de corrélation linéaire r, que peut-on en déduire ?

2. Calculer l'équation de la droite de régression de Y en X : $y = ax + b$.

Exercice 5. Une filature livre des pelotes de laine dont les poids $X_1, X_2, \ldots$ sont supposés i.i.d. de moyenne m et d'écart-type σ.

1. Quel est la moyenne de S_n ? Son écart-type ? Quel théorème permet de justifier que lorsque n est grand, $S_n = X_1 + \cdots + X_n$ renormalisé suit une loi normale ?

2. On suppose que S_n suit une loi normale et on suppose également connaître $\sigma = 4,5g$. On prélève 9 pelotes dans une livraison dont le poids total vaut $507,6g$. Calculer un intervalle de confiance à 95%. Tester la valeur $m = 60g$ pour cet intervalle.
   On précise que, pour la loi normale centrée réduite, le quantile d'ordre $0,975$ *vaut* $q = 1,96$.

3. On considère un nouveau type de pelote. On note x_i le poids de la pelote i. Montrer que
$$\hat{m} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{et} \qquad \hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{m})^2,$$
   sont des estimateurs sans biais de m, respectivement σ^2.

4. On suppose toujours que S_n suit une loi normale. On a mesuré $\hat{m} = 102g$ et $\hat{\sigma}^2 = 7,5g$. En approchant σ par $\hat{\sigma}$, construire un intervalle de confiance autour de $\hat{m}$ à 95%.

Exercice 6. On s'intéresse à la durée de vie de tubes fluorescents fabriqués par une usine. Un organisme indépendant souhaite vérifier que la durée de vie moyenne dépasse $\theta_0 = 1600$ heures. Pour cela il mesure la durée de vie moyenne sur un échantillon de taille $n = 100$ et trouve une valeur de 1575 heures avec un écart-type de 120 heures. On pose $X_1, \ldots, X_n$ la durée de vie de n tubes fluorescents, de moyenne θ et de variance σ^2.

1. Quelles hypothèses sur $(X_i)_{i\geqslant 1}$ sont nécessaires pour assurer, pour $S_n = X_1 + \cdots + X_n$, la convergence en loi suivante :
$$\frac{S_n - n\theta}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathscr{N}(0,1).$$
   On supposera dans la suite l'approximation suivante $N := \frac{S_n - n\theta}{\sigma\sqrt{n}} \sim \mathscr{N}(0,1)$.

2. A l'aide de la fonction caractéristique de la gaussienne de moyenne μ et de variance δ^2 donnée par
$$\varphi_{\mu,\delta^2}(t) = \mathrm{e}^{it\mu}\mathrm{e}^{-\frac{t^2\delta^2}{2}},$$
   retrouver la loi de $\frac{S_n}{n}$.

3. Quelle est l'hypothèse nulle H_0 adaptée pour l'organisme indépendant s'il désire être sûr que la moyenne dépasse θ_0 ?

4. Compléter la construction de la zone de rejet de l'hypothèse nulle, pour $x_\alpha > 0$ un réel qui sera choisi ultérieurement :
$$\mathcal{R} := \left\{ \frac{S_n}{n} \ldots \theta_0 + x_\alpha \right\}.$$

5. On définit l'erreur de première espèce au niveau α par
$$\sup_{\theta\in H_0} \mathbb{P}_\theta\left(\mathcal{R}\right) = \alpha.$$
   ou sous $\mathbb{P}_\theta$, les X_i sont de moyenne θ. En raisonnant qualitativement, pour quelle valeur de θ, la borne supérieure est-elle atteinte ?

6. On approche brutalement σ par l'écart-type empirique valant 120 heures. On donne que le quantile d'ordre $1-\alpha = 0,95$ vaut $q_\alpha = 1,65$. En déduire la valeur minimale de x_α pour $\alpha = 0,05$.

7. On définit également l'erreur de seconde espèce par

$$\sup_{\theta \in H_1} \mathbb{P}_\theta \left(\mathcal{R}^c \right).$$

   Traduire en français la signification de cette erreur. Pour quelle valeur de θ cette borne supérieure est-elle atteinte ? Afin de minimiser cette seconde erreur, vaut-il mieux avoir x_α plutôt grand ou petit ? En déduire une valeur de x_α adaptée.

8. A l'aide de la valeur mesurée pour la durée de vie moyenne, conclure si l'organisme rejette ou non l'hypothèse nulle avec une certitude de 95%.

9. Quelle est la conclusion si l'on exige une précision d'ordre 99% sachant que le quantile d'ordre $0,99$ vaut $q_\alpha = 2,33$?

10. Faire le même travail mais du point de vue de l'usine, la conclusion est-elle la même pour le test à 95% ? à 99% ?