Feuille de TD n^0 1. Elaboration des données expérimentales

Exercice 0.1 Le Coucou est un oiseau qui fait couver ses oeufs par oiseaux d'autres espèces de tailles très différentes. Il a été hypothèse qui le Coucou puisse adapter la taille de ses oeufs à la taille du nid dans lequel il pond. Une étude faite sur la taille des oeufs déposés dans les nids de petite taille (Roitelet) ou de plus grande taille (Fauvette) a donné les valeurs (en millimètres) suivantes :

Nids de Roitelet :
19,5 - 22,1 - 21,5 - 20,9 - 22,0 - 21,0 - 22,3 - 21,0 - 20,3 - 20,9 - 22,0 - 22,0 - 20,8 - 21,2 - 21,0
Nids de Fauvette :
22,0 - 23,9 - 20,9 - 23,8 - 25,0 - 24,0 - 23,8 - 21,7 - 22,8 - 23,1 - 23,5 - 23,0 - 23,0 - 23,1
(d'après O.H. Latter, revue *Biometrica*, 1902).

(a) Donner les tableaux des effectifs et des fréquences des deux échantillons.

(b) Donner les tableaux récapitulatifs avec les groupements par classes suivants :
$[19, 20[, [20, 21[, [21, 22[, [22, 23]$ (premier échantillon)
$[20, 21[, [21, 22[, [22, 23[, [23, 24[, [24, 25]$ (second échantillon)

(c) Tracer les histogrammes des effectifs des deux échantillons.

(d) Tracer sur un même graphique les histogrammes des fréquences des deux échantillons.

(e) Donner pour chacun de ces deux échantillons les paramètres de position et de dispersion.

(f) Tester l'hypothèse indiquée au début de l'exercice.

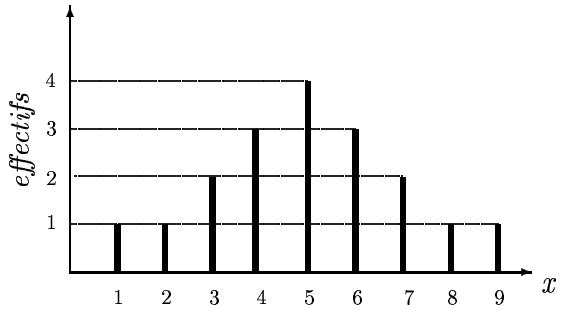
Référence : [B1], page 19.

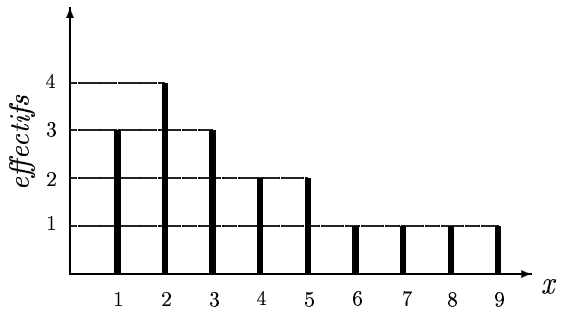
Exercice 0.2 Soit $x_1, \ldots, x_N$ un échantillon de taille N de la variable x. On suppose que $x_1, \ldots, x_h$ sont les valeurs distinctes de x dans l'échantillon. Soit N_k l'effectif de x_k et F_k la fréquence de x_k dans l'échantillon. Vérifier les *formules de Koenig* pour la variance de x :

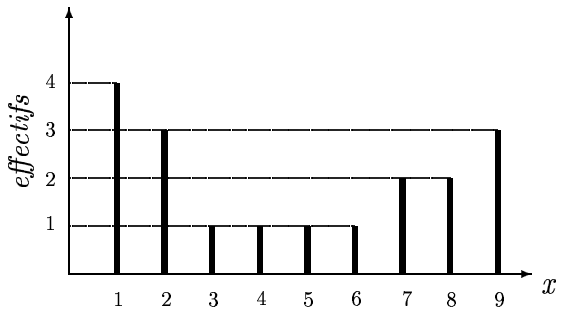
$$v_x = \frac{1}{N}\sum_{k=1}^{N} x_n^2 - \overline{x}^2 \tag{1}$$

$$v_x = \frac{1}{N}\sum_{k=1}^{h} N_k x_k^2 - \overline{x}^2 \tag{2}$$

$$v_x = \sum_{k=1}^{h} F_k x_k^2 - \overline{x}^2. \tag{3}$$







Exercice 0.3 Les graphiques ici à côté sont les diagramme en bâtons des effectifs de trois échantillons différents de la variable x.

(a) Tracer sur les graphiques la moyenne, le mode, la médiane et les quartiles.

(b) Calculer les paramètres de dispersion pour chacun des échantillons.

Exercice 0.4 En Belgique ils existent deux espèces de chênes, le chêne pédonculé (*Quercus robur*) et le chêne sessile (*Quercus petrae*). Plusieurs caràcteres pourraient servir à différencier les deux espèces, comme par exemple la hauteur de l'arbre, le type de feuilles, leur répartition sur la branche, ou le type de fleur ou de fruit. On choit *arbitrairement* le type de feuilles, et plus prècisement la longueur relative du pètiole. Le pètiole est la partie amincie de la feuille reliant le limbe à la tige. La longueur du pètiole varie de 5 mm à 10 mm pour le chêne pédonculé et de 10 mm à 30 mm pour le chêne sessile. La longueur du pètiole dèpend aussi de la longueur et donc de l'âge de la feuille (une feuille jeune est plus petite et a normalement un pètiole plus court). Par contre, le rapport entre la longueur du pètiole et la longueur totale de la feuille est indépendent de l'âge (et de la choix de l'unité de mesure). Des observations ont donnè les résultats suivants.

Données pour le chêne sessile :

0,0100	0,0479	0,0536	0,0548	0,0561	0,0602	0,0603	0,0622	0,0622	0,0626
0,0632	0,0638	0,0641	0,0647	0,0651	0,0667	0,0667	0,0675	0,0678	0,0681
0,0684	0,0690	0,0700	0,0708	0,0708	0,0714	0,0715	0,0719	0,0720	0,0722
0,0725	0,0729	0,0731	0,0736	0,0737	0,0747	0,0749	0,0753	0,0758	0,0768
0,0768	0,0771	0,0776	0,0782	0,0783	0,0789	0,0792	0,0800	0,0800	0,0804
0,0808	0,0811	0,0815	0,0818	0,0820	0,0826	0,0827	0,0830	0,0835	0,0838
0,0844	0,0849	0,0851	0,0855	0,0855	0,0862	0,0866	0,0867	0,0871	0,0873
0,0874	0,0878	0,0879	0,0879	0,0879	0,0881	0,0882	0,0887	0,0891	0,0899
0,0915	0,0916	0,0916	0,0921	0,0922	0,0923	0,0927	0,0933	0,0940	0,0941
0,0941	0,0953	0,0956	0,0958	0,0962	0,0963	0,0976	0,0980	0,0980	0,0986
0,0987	0,0992	0,0997	0,0999	0,1015	0,1023	0,1023	0,1031	0,1034	0,1040
0,1045	0,1047	0,1060	0,1062	0,1062	0,1070	0,1071	0,1077	0,1080	0,1087
0,1115	0,1118	0,1118	0,1125	0,1126	0,1131	0,1137	0,1140	0,1169	0,1173
0,1173	0,1189	0,1193	0,1200	0,1217	0,1293	0,1308	0,1365	0,1379	0,1408

Données pour le chêne pédunculé :

0,0100	0,0178	0,0189	0,0204	0,0213	0,0223	0,0229	0,0232	0,0232	0,0239
0,0247	0,0250	0,0250	0,0250	0,0250	0,0256	0,0259	0,0263	0,0263	0,0263
0,0263	0,0266	0,0266	0,0266	0,0267	0,0270	0,0270	0,0270	0,0270	0,0277
0,0278	0,0278	0,0278	0,0280	0,0282	0,0286	0,0286	0,0289	0,0294	0,0294
0,0297	0,0298	0,0300	0,0303	0,0307	0,0309	0,0312	0,0315	0,0317	0,0317
0,0319	0,0323	0,0323	0,0329	0,0334	0,0334	0,0334	0,0337	0,0341	0,0345
0,0345	0,0348	0,0349	0,0351	0,0351	0,0353	0,0357	0,0357	0,0361	0,0364
0,0366	0,0366	0,0367	0,0370	0,0370	0,0375	0,0375	0,0375	0,0377	0,0378
0,0380	0,0384	0,0385	0,0385	0,0389	0,0389	0,0392	0,0392	0,0392	0,0395
0,0400	0,0400	0,0405	0,0405	0,0406	0,0408	0,0408	0,0411	0,0412	0,0416
0,0417	0,0417	0,0417	0,0417	0,0422	0,0425	0,0425	0,0428	0,0428	0,0430
0,0430	0,0435	0,0435	0,0435	0,0435	0,0438	0,0439	0,0445	0,0445	0,0446
0,0447	0,0448	0,0454	0,0454	0,0459	0,0463	0,0465	0,0470	0,0470	0,0476
0,0476	0,0482	0,0484	0,0487	0,0488	0,0491	0,0491	0,0491	0,0496	0,0497
0,0500	0,0500	0,0500	0,0500	0,0508	0,0510	0,0515	0,0522	0,0526	0,0526
0,0534	0,0535	0,0538	0,0545	0,0545	0,0549	0,0556	0,0556	0,0560	0,0564
0,0566	0,0568	0,0571	0,0571	0,0582	0,0589	0,0589	0,0594	0,0600	0,0606
0,0615	0,0617	0,0625	0,0656	0,0667	0,0672	0,0678	0,0714	0,0740	0,0750

Pour chacun des échantillons :

(a) Grouper les données par classes d'amplitude constant $0,01$ et calculer les effectifs et les fréquences des classes.

(b) Tracer les histogrammes des fréquences.

Pour une classe $[a, b[$ (ou $[a, b]$) on appelle $(a+b)/2$ le centre de la classe. Soient $I_1, \ldots, I_N$ les classes d'un échantillon de la variable x. On indique par X_k le centre de la classe I_k et par N_k son effectif. La moyenne de l'échantillon peut être calculée approximativement par la formule

$$\overline{x} \approx \frac{1}{N} \sum_{k=1}^{N} N_k X_k \,.$$

De la même façon, la variance et l'écart-type de l'échantillon peuvent être calculés approximativement par les formules

$$v_x \approx \frac{1}{N}\sum_{k=1}^{N} N_k(X_k - \overline{x})^2 \qquad \text{et} \quad \sigma \approx \sqrt{v_x}\,.$$

(c) Calculer approximativement la moyenne et l'écart-type des deux espèces de chênes.

(d) Comparer les résultats obtenus.

Référence : [D+], pages 145–164.

Exercice 0.5 Soient $P_1(x_1, y_1), \ldots, P_N(x_N, y_N)$ des points "presque" alignés. Notre objectif est de trouver la "meilleure' droite $y = ax + b$ représentent l'alignement. Il s'agit donc de trouver une formule pour calculer a et b. Le critère des *moindres carrés ordinaires* (MCO) conduit à une droite unique, appellée *droite des moindres carrés*, pour laquelle la somme des carrés des écarts verticaux

$$\sum_{k=1}^{N} \big(y_k - (ax_k + b)\big)^2$$

est minimale.

Pour $\alpha, \beta \in \mathbb{R}$ on pose

$$E(\alpha, \beta) = \sum_{k=1}^{N} \big(y_k - \alpha x_k - \beta)\big)^2.$$

(a) On suppose d'abord α fixe. Montrer que $E(\alpha, \beta)$ est un trinôme du second degré en β de la forme

$$A\beta^2 + 2B\beta + C$$

où A, B, C ne dépendent pas de β et $A > 0$, et que donc il est minimum pour la valeur de β égale à

$$b = -\frac{B}{A} = \overline{y} - a\overline{x}. \qquad (*)$$

(b) On suppose maintenant que $\beta = b$ comme dans (*). Montrer que $E(\alpha, a)$ est un trinôme du seconde degré en α de la forme

$$\widetilde{A}\alpha^2 + 2\widetilde{B}\alpha + \widetilde{C}$$

où $\widetilde{A}, \widetilde{B}, \widetilde{C}$ ne dépendent pas de α et $\widetilde{A} > 0$, et que donc il est minimum pour la valeur de α égale à

$$a = -\frac{\widetilde{B}}{\widetilde{A}} = \frac{1}{N}\frac{\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})}{v_x}. \qquad (**)$$

(c) Vérifier que a et b donnés par (*) et (**) rendent $E(\alpha, \beta)$ minimum : pour toutes α,β

$$E(\alpha, \beta) \geq E(a, b)\,.$$

(d) Le nombre

$$c_{x,y} := \frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})$$

est appellée la *covariance des variables x et y*. Montrer la formule

$$c_{x,y} := \frac{1}{N}\sum_{k=1}^{N} x_k y_k - \overline{x}\,\overline{y}\,.$$

On conclue :

La droite des moindres carrés est la droite d'équation $y = ax + b$ avec

$$a = \frac{c_{x,y}}{v_x} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}\,.$$

Exercice 0.6 Dans les années 50, une fuite de déchets radioactifs issue d'une zone de stockage près de Hanford (Etat de Washington) s'est répandue dans la rivière Columbia. On à calculè pour 9 comtés situés en aval, dans l'Etat de l'Oregon, un indice d'exposition (fonction de la distance à Hanford, de la distance qui sépare le lieu de résidence d'un citoyen "type" de la rivière, etc.). De même, on a calculé la mortalité par cancer (nombre de décès annuels pour 100 000 habitants entre 1959 et 1964). Les données sont les suivantes (d'après Fadeley, *Journal of environmental health*, 1965) :

Comté	*Indice d'exposition* x	*Mortalité par cancer* y
Clatsop	8,3	210
Columbia	6,4	180
Gilliam	3,4	130
Hood River	3,8	170
Morrow	2,6	130
Portland	11,6	210
Sherman	1,2	120
Umatilla	2,5	150
Wasco	1,6	140

1. Calculer la droite des moindres carrés.
2. Estimer la mortalité par cancer si $x = 5,0$.
3. Représenter graphiquement les 9 points comtés et les réponses en (a) et (b).

Référence : T.H. Wonnacott et R.J. Wonnacott, *Statistique*, page 416.