C.2 Distributions univariées

C.2.1

Un journaliste a relevé le nombre d'interviews qu'il a réalisé la première semaine d'un festival de cinéma. Les résultats obtenus sont les suivants :

jour	nombre d'interviews
L	20
Ma	5
Me	10
J	15
V	10
S	20
D	20

On considère l'interview comme unité statistique et le jour comme variable observée.

1. Déterminer les fréquences relatives des interviews par jour.
2. Est-il possible de représenter les résultats par un histogramme ? Pourquoi ?
3. Faire une représentation graphique appropriée.

On considère à présent le jour comme unité statistique et le nombre d'interviews comme variable observée.

4. Donner la répartition des jours selon le nombre d'interviews.
5. Faire l'histogramme de la distribution en considérant les classes [0 - 12.5],]12.5 - 17.5],]17.5 - 22.5].

C.2.2

Contrairement aux adultes, les enfants ont tendance à se souvenir des histoires sous la forme d'une séquence d'actions plutôt que d'une intrigue globale. Pour décrire un film, ils employent donc constamment l'expression "et puis ...". Une expérimentatrice dotée d'une patience infinie a demandé à 50 enfants de lui raconter un film donné. Entre autres variables, elle a compté le nombre de "et puis ...". Les données sont les suivantes :

18	15	22	19	18	17	18	20	17	12
16	16	17	21	23	18	20	21	20	20
15	18	17	19	20	23	22	10	17	19
19	21	20	18	18	24	11	19	31	16
17	15	19	20	18	18	40	18	19	16

1. Représentez graphiquement la distribution des fréquences de ces données.
2. Quelle est la forme générale de la distribution ?
3. Faites un histogramme pour ces données en utilisant un nombre raisonnable de classes.
4. Calculez le mode, la médiane, la moyenne et la moyenne tronquée à 20% des enfants. Commentez.

C.2.3

Dans le cadre de l'étude décrite dans l'exercice C.2.2, l'expérimentatrice a obtenu le même type de données pour des adultes :

10	12	5	8	13	10	12	8	7	11
11	10	9	9	11	15	12	17	14	10
9	8	15	16	10	14	7	16	9	1
4	11	12	7	9	10	3	11	14	8
12	5	10	9	7	11	14	10	15	9

1. Que pouvez-vous dire après un simple examen de ces nombres ? Les enfants et les adultes de ces deux échantillons semblent-ils se souvenir des histoires de la même manière ?
2. Représentez graphiquement la distribution des fréquences de ces données en utilisant pour les axes la même échelle que celle utilisée pour l'exercice C.2.2. Qu'y a-t-il de différent entre les deux graphiques ?
3. Calculez la distribution des fréquences cumulées des adultes et représentez-la graphiquement.
4. Calculez le mode, la médiane, la moyenne et la moyenne tronquée à 20% des adultes. Commentez.

C.2.4

Les données suivantes représentent un échantillon de taille 12'705 de la distribution des salaires annuels en France en 1980 (source : INSEE, Déclaration Annuelle des Salaires (DAS)). En l'absence d'informations plus précises, nous supposerons que les salaires se répartissent entre 0 et 350'000 francs.

Salaires en milliers de francs	Effectifs
moins de 15	390
[15 - 20[	230
[20 - 25[	454
[25 - 30[	1080
[30 - 35[	1420
[35 - 40[	1565
[40 - 50[	2737
[50 - 60[	1746
[60 - 70[	1009
[70 - 80[	598
[80 - 100[	631
100 et plus	845

Représenter graphiquement ces données et commenter le résultat.

C.2.5

Nous avons évalué à l'aide d'un questionnaire les connaissances en algèbre des étudiants en sciences sociales. Le score théorique minimal, correspondant à des compétences médiocres, vaut 0. Le score théorique maximal, correspondant à de bonnes compétences, vaut 200. Voici les scores obtenus par 95 étudiants :

18	36	39	40	41	44	46	47	48	54
58	60	63	64	65	66	67	69	69	71
71	73	73	74	74	75	77	80	80	83
85	86	86	86	87	87	88	89	90	90
90	90	91	92	92	92	93	94	94	97
98	99	99	100	100	100	101	102	102	102
104	108	109	111	111	112	112	112	114	114
121	121	122	122	123	124	125	126	129	130
131	132	132	132	134	134	141	141	142	143
146	149	151	153	172					

Indications : $\sum_{i=1}^{95} x_i = 9'110$, $\sum_{i=1}^{95} x_i^2 = 963'352$.

1. Déterminer le mode, la médiane et la moyenne de cette distribution.
2. Calculer l'étendue, les quartiles, l'écart interquartile, la variance et l'écart type de cette distribution.
3. Résumer cette distribution par un box-plot.
4. Calculer les 9 déciles $D_1, \ldots, D_9$ de cette distribution.
5. Représenter ces données par un histogramme en utilisant 10 classes de fréquences plus ou moins égales.

C.2.6

Calculez la médiane, les quartiles, la moyenne et l'écart type de la distribution suivante, puis représentez-la par un box-plot.

j	1	2	3	4	5
x_j	10	15	17	21	34
n_j	1	9	12	7	3

C.2.7

Le tableau suivant présente la distribution de 100 déficients mentaux selon le temps utilisé pour accomplir le test des disques de la batterie Standard Bonnardel.

Temps en secondes	Nombre de déficients
$0 \leq X < 60$	12
$60 \leq X < 120$	20
$120 \leq X < 180$	15
$180 \leq X < 240$	24
$240 \leq X < 300$	14
$300 \leq X < 360$	8
$360 \leq X < 420$	5
$420 \leq X \leq 480$	2

Représenter graphiquement ces données et commenter le résultat.

C.2.8

1. Soit les scores x_1, x_2, x_3 et x_4. Leur moyenne vaut 20. Déterminer les valeurs de ces scores. La solution est-elle unique ? Si non, proposer au moins deux solutions possibles.
2. Soit les scores y_1, y_2, y_3 et y_4. Leur variance vaut 25. Déterminer les valeurs de ces scores. La solution est-elle unique ? Si non, proposer au moins deux solutions possibles.
3. Soit les scores z_1, z_2, z_3 et z_4. Leur moyenne vaut 15 et leur variance 36. Déterminer les valeurs de ces scores. La solution est-elle unique ? Si non, proposer au moins deux solutions possibles.

C.2.9

Cet exercice utilise les données présentées dans l'annexe E.1.

Représenter les observations de la variable X par un boxplot. Commenter le résultat obtenu.

C.2.10

Cet exercice utilise les données présentées dans l'annexe E.2.

Représenter les observations de la variable X par un boxplot. Calculer aussi la moyenne et la variance de cette variable et commenter l'ensemble des résultats.

C.2.11

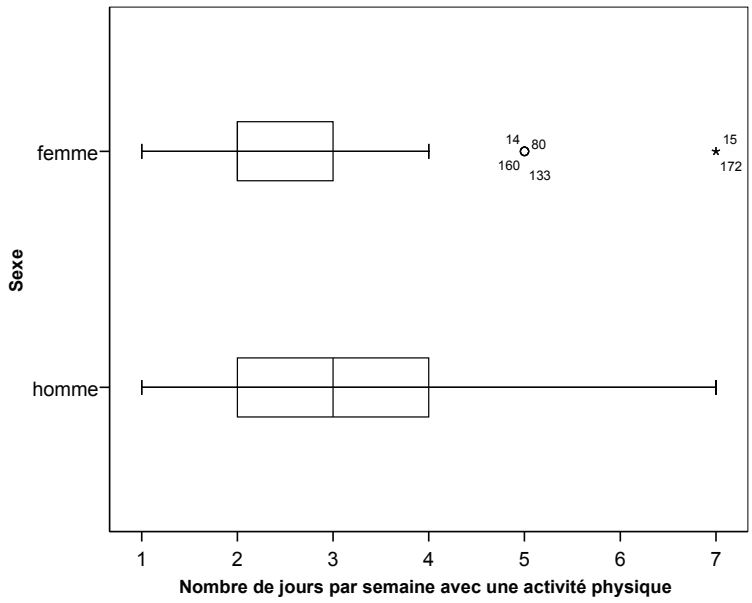
Cet exercice utilise les données présentées dans l'annexe E.2.

En utilisant les classes suivantes : [10.0 - 14.0],]14.0 - 18.5],]18.5 - 24.0],]24.0 - 36.0], construire et commenter un histogramme pour la variable Y.

C.2.12

Nous utilisons un échantillon de 275 jeunes adultes ayant eu 18 ans en 2004. Cet échantillon est représentatif de l'ensemble des jeunes de 18 ans vivant en Suisse (source : Panel Suisse des ménages). Nous disposons d'une variable nommée **Activité** donnant le nombre de jours par semaine durant lesquels le jeune adulte a pratiqué une activité physique (de 0 à 7).

Les deux boxplots suivant représentent la distribution de la variable **Activité** séparément pour les hommes et les femmes de l'échantillon. Comparez et commentez les deux graphiques. *Indications : Pour les femmes, le premier quartile est égal à la médiane. Les données représentées individuellement sur la droite du boxplot des femmes correspondent à des données extrêmes.*



C.2.13

Cet exercice utilise les données présentées dans l'annexe E.4.

Nous ne considérons que les 15 iris de l'espèce numéro 1. Représenter la largeur des pétales de ces 15 fleurs sur un graphique en bâtons, puis représenter la longueur des pétales de ces 15 fleurs sur un deuxième graphique en bâtons et comparer les deux graphiques.

C.2.14

Cet exercice utilise les données présentées dans l'annexe E.5.

La moyenne de la variable "justice" (X) est égale à 83.9%. Or, le taux réel d'acceptation de cette initiative a été de 86.4%. Expliquer cette différence.

C.2.15

Soit les données suivantes :

i	x_i	n_i
1	5	1
2	4	3
3	3	5
4	2	3
5	1	1

Calculer les expressions suivantes :

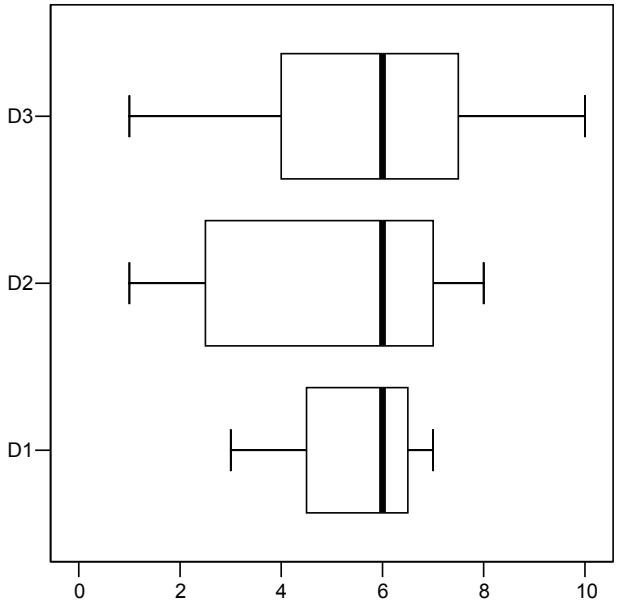
$$n = \sum_{i=1}^{5} n_i \quad , \qquad \sum_{i=1}^{5} \frac{n_i}{n} \quad , \qquad \overline{x} = \frac{1}{n} \sum_{i=1}^{5} n_i x_i \quad , \qquad \left(\sum_{i=1}^{5} n_i \right) \cdot \left(\sum_{i=1}^{5} x_i \right)$$

$$\sum_{i=1}^{5} n_i x_i \quad , \qquad \sum_{i=1}^{5} (n_i x_i)^2 \quad , \qquad \left(\sum_{i=1}^{5} n_i x_i \right)^2 \quad , \qquad \sum_{i=1}^{5} n_i (x_i - \overline{x})$$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{5} n_i (x_i - \overline{x})^2 \quad , \qquad \frac{1}{n} \sum_{i=1}^{5} n_i x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{5} n_i x_i \right)^2$$

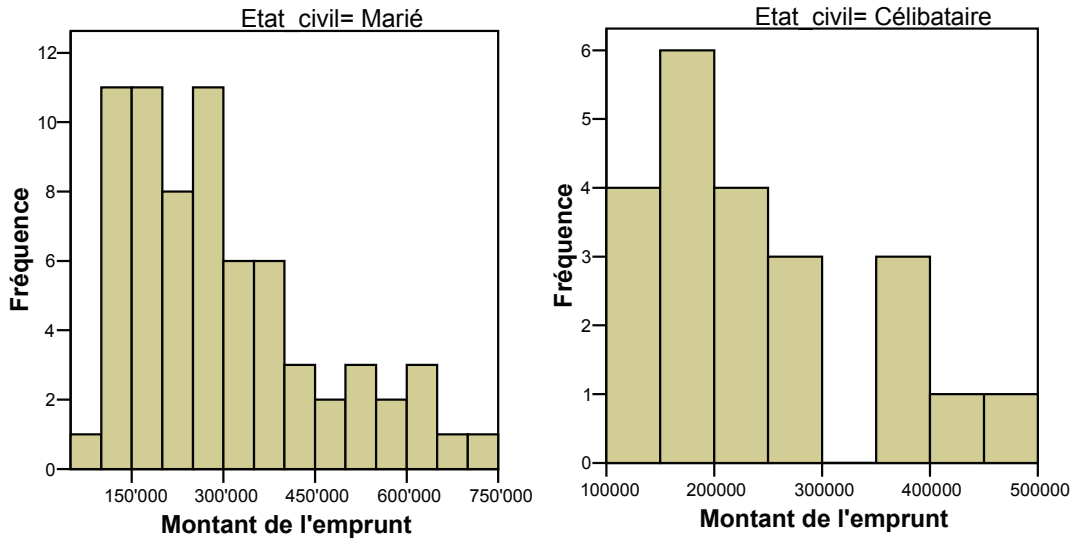
C.2.16

Le graphique suivant présente 3 boxplots correspondant à des notes sur une échelle de 0 à 10 attribuées par le responsable du département marketing d'une entreprise à ses employés en 2001 (D1), 2002 (D2) et 2003 (D3). Ces boxplots ont été réalisés à l'aide de respectivement 3, 15 et 27 observations. Commentez séparément chacun des 3 boxplots, puis comparez-les.



C.2.17

Nous disposons de données relatives à des emprunts effectués auprès d'une banque pour l'achat d'un logement par un échantillon de 100 personnes. Nous avons représenté sur des histogrammes la distribution du montant de l'emprunt en séparant les données en fonction de l'état-civil (marié ou célibataire) des emprunteurs. Commenter et comparer les deux histogrammes.



C.2.18

Le tableau suivant généré par R fournit différentes informations concernant une variable représentant l'année durant laquelle un échantillon de personnes a effectué un emprunt hypothécaire auprès d'une banque en vue de l'acquisition d'un logement. En vous basant sur ce tableau, que pouvez-vous dire quant à la tendance centrale et à la dispersion de cette variable ? Construisez et interprétez aussi le boxplot correspondant.

mean	sd	0%	25%	50%	75%	100%	n
2000	3.74	1994	1995	2000	2002	2005	15

C.2.19

Le tableau suivant donne le nombre de nouveaux cas de tests VIH positifs confirmés en Suisse entre 1993 et 2005 (source : Office Fédéral de la Statistique). Représentez graphiquement ces données afin de mettre en évidence l'évolution de la tendance dans le temps.

Année	Nombre de cas
1993	1600
1995	1012
1997	835
1999	597
2000	581
2001	631
2002	791
2003	756
2004	743
2005	714

C.2.20

Nous utilisons un échantillon de 275 jeunes adultes ayant eu 18 ans en 2004. Cet échantillon est représentatif de l'ensemble des jeunes de 18 ans vivant en Suisse (source : Panel Suisse des ménages). Nous disposons d'une variable nommée **Relation** indiquant le degré de satisfaction dans les relations avec les autres (de 0 : pas du tout satisfait, à 10 : pleinement satisfait).

Le tableau suivant donne la distribution de la variable **Relation** pour les 164 jeunes de l'échantillon ayant répondu à cette question :

Valeur	0	4	5	6	7	8	9	10
Fréquence	1	1	5	10	29	52	37	29

Construisez un graphique en bâtons et un boxplot pour ces données. Commentez.

C.2.21

Les chiffres ci-dessous correspondent aux temps en minutes réalisés par les coureurs d'une équipe universitaire pour parcourir soit un mille, soit un quart de mille :

Temps pour un quart de mille : 0.92 0.98 1.04 0.9 0.99
Temps pour un mille : 4.52 4.35 4.6 4.7 4.5

Utiliser l'écart-type et le coefficient de variation pour résumer la dispersion des données. Sur quel parcours les temps sont-ils les plus réguliers ?

Annexe E

Données

E.1 Médecins

Le tableau suivant donne, pour chaque canton suisse en 1997, le nombre total de médecins (variable X), ainsi que le pourcentage de médecins généralistes par rapport à l'ensemble des médecins du canton (variable Y) :

Canton	X	Y	Canton	X	Y
ZH	2493	34	SH	127	43
BE	1757	32	AR	71	48
LU	470	42	AI	12	58
UR	40	50	SG	637	44
SZ	129	52	GR	308	47
OW	30	60	AG	702	38
NW	33	52	TG	260	50
GL	45	53	TI	549	35
ZG	148	34	VD	1413	33
FR	325	34	VS	431	39
SO	340	44	NE	311	38
BS	671	17	GE	1196	19
BL	448	36	JU	92	46

Source : Annuaire statistique de la Suisse, édition 2000.

E.2 Europe-vitesse

Le 4 mars 2001, les citoyens suisses ont eu à se prononcer sur les deux sujets suivants :

1. Initiative "Oui à l'Europe" demandant l'ouverture immédiate de négociations en vue de l'adhésion de la Suisse à l'Union Européenne.
2. Initiative "30 km/h" demandant la généralisation de la limitation de vitesse à 30 km/h dans l'ensemble des localités suisses.

Le tableau suivant donne, pour chaque canton suisse, les pourcentages d'acceptation de ces deux initiatives (X="Oui à l'Europe", Y="30 km/h") :

Canton	X	Y	Canton	X	Y
ZH	23.7	25.0	SH	17.6	23.7
BE	23.5	22.2	AR	13.5	18.6
LU	15.7	18.1	AI	6.8	10.7
UR	9.4	18.0	SG	14.2	17.6
SZ	10.7	13.1	GR	14.5	25.6
OW	11.2	14.7	AG	17.0	17.1
NW	10.9	13.9	TG	13.7	17.1
GL	13.3	20.7	TI	15.9	16.9
ZG	17.2	16.7	VD	39.4	17.9
FR	27.3	13.1	VS	20.9	14.1
SO	19.8	18.9	NE	48.8	18.2
BS	29.2	35.8	GE	41.1	25.1
BL	22.7	23.1	JU	44.2	15.1

Nous connaissons de plus les résultats intermédiaires suivants :

$$\sum_{i=1}^{26} x_i = 542.2,\ \sum_{i=1}^{26} y_i = 491,\ \sum_{i=1}^{26} x_i^2 = 14'478.6,\ \sum_{i=1}^{26} y_i^2 = 9'968,\ \sum_{i=1}^{26} x_i\, y_i = 10'630.9$$

E.3 Course de l'escalade

La course de l'escalade ayant lieu à Genève au mois de décembre est l'une des compétitions de course à pied les plus populaire de Suisse. Les participants sont répartis en 27 catégories différant notamment par la longueur et la difficulté du parcours. Nous ne nous intéressons ici qu'aux catégories "Femmes I" à "Femmes V" pour lesquelles la distance à courir est de 4780 mètres. Les données sont extraites des résultats officiels de la course du 2 décembre 2000. Le tableau suivant donne pour ces 5 catégories la classe d'âge concernée, le nombre d'inscrites et le nombre de classées (femmes ayant effectivement terminé l'épreuve).

Catégorie	Age	Nombre d'inscrites	Nombre de classées
Femmes I	[19-28]	833	675
Femmes II	]28-38]	1218	968
Femmes III	]38-48]	986	814
Femmes IV	]48-58]	426	360
Femmes V	]58-80]	132	108
Total		3595	2925

De plus, nous connaissons les temps en secondes réalisés par les 3 premières de chaque catégorie :

Femmes I :	1069	1071	1112
Femmes II :	1027	1095	1105
Femmes III :	1077	1093	1118
Femmes IV :	1217	1232	1238
Femmes V :	1349	1359	1428

E.4 Iris

Lors de ses travaux sur la méthode statistique appelée “analyse discriminante”, le grand mathématicien R.A. Fisher a utilisé un célèbre jeu de données décrivant la longueur et la largeur des pétales et des sépales de 150 iris appartenant à 3 espèces différentes. Voici un échantillon extrait de ces données. Les 3 espèces d'iris sont numérotées de 1 à 3 (1 : Iris Setosa, 2 : Iris Versicolor, 3 : Iris Verginica). Le tableau suivant donne les largeurs et longueurs des pétales (en millimètres) de 45 iris (15 de chaque espèce) :

Espèce	Largeur	Longueur	Espèce	Largeur	Longueur	Espèce	Largeur	Longueur
1	2	14	2	13	45	3	24	56
1	2	10	2	16	47	3	23	51
1	2	16	2	14	47	3	20	52
1	1	14	2	12	40	3	19	51
1	2	13	2	10	33	3	17	45
1	2	16	2	10	41	3	19	50
1	2	16	2	15	45	3	18	49
1	4	19	2	10	33	3	21	56
1	2	14	2	14	39	3	19	51
1	2	14	2	12	39	3	18	55
1	4	15	2	15	42	3	15	50
1	2	14	2	13	44	3	23	57
1	2	14	2	15	49	3	20	49
1	1	14	2	11	30	3	18	58
1	3	17	2	13	36	3	21	54

E.5 Justice et trafic

Le tableau suivant donne les pourcentages d'acceptation à deux votations soumises aux citoyens suisses le 12 mars 2000 :

– X **: justice** : arrêté fédéral relatif à la réforme de la justice
– Y **: trafic** : initiative populaire pour la diminution par deux du trafic routier

Canton	X	Y	Canton	X	Y
ZH	89.9	26.7	SH	83.8	21.7
BE	88.4	23.4	AR	84.3	22.1
LU	85.0	20.2	AI	80.4	14.4
UR	73.7	20.6	SG	86.8	20.6
SZ	77.2	14.2	GR	82.2	21.4
OW	69.2	15.2	AG	82.3	18.3
NW	87.1	16.1	TG	83.4	18.3
GL	83.9	21.3	TI	90.1	23.5
ZG	84.7	20.3	VD	87.2	16.6
FR	87.4	14.5	VS	70.7	10.0
SO	82.8	19.5	NE	85.7	17.7
BS	91.7	33.9	GE	92.3	24.6
BL	88.8	22.3	JU	82.3	14.0

Tableau des moyennes et écart-types :

	X	Y
Moyenne	83.9	19.7
Ecart-type	5.7	4.8

Produit croisé :

$$\sum_{i=1}^{26} x_i y_i = 43'362$$

E.6 Scrabble

Le **scrabble** (marque déposée) est un jeu de type mots-croisés dans lequel les joueurs doivent former des mots à partir d'un ensemble de 7 jetons tirés aléatoirement. Dans sa version française, le jeu comporte 102 jetons, les 100 premiers portant la mention d'une lettre et les 2 derniers étant blancs et pouvant remplacer n'importe quelle lettre. Chaque jeton vaut un certain nombre de points. Le tableau suivant donne pour chacun des jetons le nombre de pièces identiques dans le jeu (variable X) ainsi que le nombre de points qui lui est associé (variable Y) :

Jeton	X	Y	Jeton	X	Y	Jeton	X	Y
J	1	8	F	2	4	N	6	1
K	1	10	G	2	2	O	6	1
Q	1	8	H	2	4	R	6	1
W	1	10	P	2	3	S	6	1
X	1	10	V	2	4	T	6	1
Y	1	10	blanc	2	0	U	6	1
Z	1	10	D	3	2	I	8	1
B	2	3	M	3	2	A	9	1
C	2	3	L	5	1	E	15	1