TD PS
Statistiques descriptives
Caractéristiques d'une variable statistique : quantiles, moyennes, courbe de concentration, ...

1. On fait subir à une variable statistique quantitative X pour laquelle on a n observations $x_1, x_2, \cdots, x_n$ une transformation linéaire ,i.e. $Y = aX + b$. Calculer la moyenne de Y et son écart-type en fonction de la moyenne de X et de son écart-type.

2. On dispose d'un échantillon de , taille n sur X quantitative. Déterminer quelle est la valeur de a qui minimise $\frac{\sum_{i=1}^{n}(x_i-a)^2}{n}$.

3. On a n observations $x_1, x_2, \cdots, x_n$ sur une variable statistique quantitative X. Montrer que $\forall a \in \mathbb{R}$, $\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2 + n\ (a - \bar{X})^2$ où $\bar{X}$ est la moyenne empirique de X (formule de Huygens).

4. On réunit deux échantillons $x_1, x_2, \cdots, x_{n_1}$ et $y_1, y_2, \cdots, y_{n_2}$ de tailles respectives n_1 et n_2, de moyennes respectives $\bar{X}$ et $\bar{Y}$, d'écart-types respectifs s_X et s_Y. Calculer G moyenne de l'échantillon résultant et s écart-type de l'échantillon résultant.

5. La gestion de qualité de service du réseau téléphonique est fondée sur des indicateurs qui sont publiés dans l'annuaire statistique des Télécoms. Un de ces indicateurs est l'indice global de qualité de service qui peut prendre la valeur de 0 à 100 et est noté IGQS. Cet indicateur est évalué à partir d'enquête clients pour chacune des 22 régions françaises définies par France-Télécom.
   Voici quelques caractéristiques de la variable IGQS de 1982 à 1986.

Caractéristiques	1982	1983	1984	1985	1986
taille	22	22	22	22	22
moyenne	84.95	86.97	90.23	93.48	95.06
médiane	85.4	88.25	91.65	94.6	96.1
écart-type	7.91	9.22	7.46	5.28	5.16
minimum	52.7	50	61.4	74.6	77.6
maximum	92.4	97.1	98.2	100	100
premier quartile	84.9	85.8	88	92	93.6
troisième quartile	89	91	93.8	96.1	98.1

   Représentez les boites de dispersion (boites à moustache) pour la variable IGQS de 1982 à 1986. Commentez l'évolution.

6. Nous nous intéressons à la consommation d'électricité en fonction de la température. Les données proviennent d'une compagnie d'électricité située au nord de l'état de New-York aux Etats-Unis. Les données dont on dispose sont la consommation moyenne d'électricité par jour pour un mois donné en kilowatt-heure notée *kwh* et la température journalière moyenne en degrés Farenheit pour un mois donné notée *tempf*. La période étudiée va d'août 1989 à février 1994. Les données originales figurent à la fin de l'exercice.

   Le tableau suivant vous donne quelques statistiques calculées pour *kwh* et *tempf*.

Statistiques descriptives	kwh	tempf
n (taille de l'échantillon)	55	55
moyenne	43.275	53.818
écart-type	23.789	15.401
minimum	10.414	24.000
Q_1 (premier quartile)	24.207	39.000
médiane	38.621	
Q_3 (troisième quartile)		69.000
maximum	101.66	79.000
écart-interquartile		
coefficient de variation		

   (a) Complétez les valeurs manquantes dans le tableau.

   (b) La relation entre une température en Farenheit et une température en Celsius s'écrit :
   température en Farenheit = 9/5(température en Celsius) + 32.
   Ainsi lorqu'il fait 0 Celsius, il fait 32 Farenheit et lorsqu'il fait 20 C, il fait 68 F. Nous noterons *tempc* la température exprimée en Celsius (*tempf* est la température exprimée en Farenheit) et donc $tempf = \frac{9}{5}\, tempc + 32$.
   A partir des statistiques précédentes sur *tempf*, déterminer la moyenne de *tempc*, son écart-type, sa valeur minimale et maximale.

   (c) Pour représenter graphiquement l'histogramme de la distribution de la consommation journalière moyenne d'électricité par mois, on va regrouper les données par classe.
   Le tableau suivant vous donne les classes choisies.

classes	[9-14[	[14-24[	[24-34[	[34-44[	[44-54[	[54-64[	[64-84[	[84-104[
effectif	1	12	13	6	6	6		

   Complétez-le et représentez graphiquement l'histogramme en précisant clairement ce que vous mettez en abscisse et en ordonnée.

   (d) Comment qualifieriez-vous cet histogramme (est-il unimodal ou bimodal, est-il symétrique ? etc) Vous pouvez commenter cette distribution en prenant en compte les statistiques descriptives de la question précédente.

Données de l'exercice

Le tableau comporte 8 colonnes (pour l'instant seules les 5 premières colonnes vous intéressent) :

- Dans la colonne 1 se trouve le numéro de l'observation de 1 à 55,
- dans la colonne 2 le numéro du mois de 1 à 12,
- dans la colonne 3 l'année,
- dans la colonne 4 la consommation moyenne journalière d'électricité en kilowatt-heure,
- dans la colonne 5 la température moyenne quotidienne en degré Farenheit,
- dans la colonne 6 le logarithme en base 10 de la consommation d'électricité, c'est-à-dire le log de la valeur de la colonne 4
- en colonne 7 le logarithme en base 10 de la température, c'est-à-dire de la valeur de la colonne 5,
- dans la colonne 8, on a la moyenne mobile centrée d'ordre 12 lorsque celle-ci peut être calculée.

1	8	1989	24.828	73	1.395	1.863	
2	9	1989	24.688	67	1.392	1.826	
3	10	1989	19.310	57	1.286	1.756	
4	11	1989	59.706	43	1.776	1.633	
5	12	1989	99.667	26	1.999	1.415	
6	1	1990	49.333	41	1.693	1.613	
7	2	1990	59.375	38	1.774	1.580	54.3
8	3	1990	55.172	46	1.742	1.663	54.3
9	4	1990	55.517	54	1.744	1.732	54.4
10	5	1990	25.938	60	1.414	1.778	54.8
11	6	1990	20.690	71	1.316	1.851	55.7
12	7	1990	24.333	75	1.386	1.875	56.1
13	8	1990	22.759	74	1.357	1.869	56.0
14	9	1990	24.688	66	1.392	1.820	55.9
15	10	1990	22.759	61	1.357	1.785	56.0
16	11	1990	50.588	49	1.704	1.690	56.3
17	12	1990	79.000	41	1.898	1.613	56.7
18	1	1991	87.188	35	1.940	1.544	56.9
19	2	1991	47.333	41	1.675	1.613	57.0
20	3	1991	38.621	42	1.587	1.623	57.0
21	4	1991	27.931	56	1.446	1.748	56.8
22	5	1991	25.000	69	1.398	1.839	56.5
23	6	1991	17.241	73	1.237	1.863	56.3
24	7	1991	17.000	77	1.230	1.886	56.1
25	8	1991	22.188	74	1.346	1.869	55.9
26	9	1991	18.276	66	1.262	1.820	55.6
27	10	1991	27.241	56	1.435	1.748	55.3
28	11	1991	39.706	47	1.599	1.672	54.7
29	12	1991	81.667	38	1.912	1.580	54.2
30	1	1992	79.688	34	1.901	1.531	53.9
31	2	1992	67.667	37	1.830	1.568	53.6
32	3	1992	65.862	39	1.819	1.591	53.5
33	4	1992	54.839	51	1.739	1.708	53.4
34	5	1992	34.667	60	1.540	1.778	53.3
35	6	1992	23.448	70	1.370	1.845	53.1
36	7	1992	29.586	73	1.471	1.863	53.0
37	8	1992	32.448	71	1.511	1.851	53.0
38	9	1992	30.033	66	1.478	1.820	52.6
39	10	1992	43.586	54	1.639	1.732	52.8
40	11	1992	48.029	45	1.682	1.653	53.1
41	12	1992	52.424	36	1.720	1.556	53.3
42	1	1993	62.333	34	1.795	1.531	53.7
43	2	1993	71.156	32	1.852	1.505	54.1
44	3	1993	44.621	39	1.650	1.591	54.2
45	4	1993	19.357	55	1.287	1.740	54.2
46	5	1993	23.733	64	1.375	1.806	54.3
47	6	1993	10.414	72	1.018	1.857	54.4
48	7	1993	17.875	79	1.252	1.898	53.9
49	8	1993	24.207	75	1.384	1.875	53.5
50	9	1993	24.333	65	1.386	1.813	
51	10	1993	40.935	54	1.612	1.732	
52	11	1993	42.938	48	1.633	1.681	
53	12	1993	65.250	35	1.815	1.544	
54	1	1994	101.300	24	2.006	1.380	
55	2	1994	101.660	32	2.007	1.505	

7. On veut calculer combien il y a dans un pays de médecins par habitant sachant qu'on a les résultats pour les 4 régions également peuplées du pays. On a 1 médecin pour 100 habitants dans la région 1, 1 pour 200 dans la région 2, 1 pour 250 dans la région 3 et 1 pour 500 dans la région 4. Quel est le résultat ?

8. Un salaire augmente de 10% la première année, de 15% la seconde et de 30% la troisième année. De combien a -t-il augmenté en moyenne sur les trois ans ?

9. Un chiffre d'affaires augmente de 5% les deux premières années, de 9% les quatre années suivantes et de 12% la dernière année. Quelle est l'augmentation moyenne annuelle ?

10. Un atelier de reproduction aphotocopié 1000 documents à la vitesse de 800 documents à l'heure et un autre atelier a photocopié 2000 documents de même nature à la vitesse de 700 documents à l'heure. Quelle est la vitesse moyenne pour photocopier les 3000 documents ?

11. Le tableau suivant donne la fréquence en pourcentage et la superficie en pourcentage des exploitations selon la taille de surface agricole utilisée (en ha) en Bretagne en 1990 (Source : tableaux de l'économie bretonne 1993- Pub Insee DR Bretagne). 82607 exploitations se répartissent 1735774 hectares.

Taille en hectares	**fréquence en %**	**superficie en %**
0 à moins de 2	13	1
2 à moins de 5	12	2
5 à moins de 10	10	3
10 à moins de 20	19	13
20 à moins de 30	18	21
30 à moins de 40	13	21
40 à moins de 50	8	17
50 et plus	7	22

   (a) Tracez la courbe de concentration associée à la distribution des surfaces agricoles.
   (b) Parmi ces 3 valeurs, 0.12, 0.46, 0.88, quelle est celle de l'indice de Gini ?
   (c) Quelle est la taille moyenne $\bar{x}$ d'une exploitation ?
   (d) Calculez la médiane et la médiale.
   (e) Il existe d'autres mesures d'inégalité que l'indice de Gini, en particulier l'écart relatif moyen qui est la moyenne arithmétique des valeurs absolues des écarts relatifs à la moyenne. Son expression est la suivante :

$$D = \frac{\sum_{j=1}^{k} f_j \mid x_j - \bar{x} \mid}{\bar{x}} \tag{1}$$

   où
   - f_j est la fréquence relative de la classe j. Ainsi $f_1 = 0.13$
   - k est le nombre de classes, ici $k = 8$;
   - x_j est le centre de la classe.

   Calculer D. Dans quel cas, $D = 0$? Dans quel cas $D = 1$?