L1 Mathématiques, Informatique Appliquées aux

Sciences Humaines et Sociales

UE 102A - Statistique descriptive

EXERCICES

2017-2018

Pascal Sarda

Feuille TD 1

I. Initiation au logiciel R.
Quelques fonctions statistiques :

Fonction	Description
`summary()`	Donner diverses caractéristiques numériques
`sum()`	Calculer la taille de l'échantillon
`cumsum()`	Calculer les effectifs cumulés
`unique()`	Donner les modalités
`table()`	Calculer les effectifs
`mean()`	Calculer la moyenne
`weighted.mean()`	Calculer une moyenne pondérée
`max(), min(), range()`	Calculer les valeurs max, min et les deux
`var()`	Calculer la variance (échantillon)
`sd()`	Calculer l'écart-type (échantillon)
`quantile()`	Calculer les quantiles

Quelques fonctions graphiques :

Fonction	Description
`plot()`	Fonction générale
`boxplot()`	Obtenir une boîte à moustaches
`barplot()`	Tracer un diagramme en colonnes
`pie()`	Tracer un camembert
`hist()`	Tracer un histogramme

Exercice 1.1. Récupérer et rentrer des données.

1. Récupérer sur l'espace numérique de travail le fichier "films.ods" en faire une copie "films.csv" (en utilisant le séparateur de champ ";").
   Créer un data.frame *films* contenant les données contenues dans le fichier "films.csv".
2. Créer la liste d'entiers de 1 à 20 puis la liste 5, 10, 15, 20, 25.
3. Créer le vecteur numérique de composantes 1.2, 36, 5.33, -26.5.
4. Créer le vecteur de chaînes de caractères Marron, Bleu, Vert, Autre.
5. Créer le vecteur logique T, F T, T, F.
6. Créer le vecteur d'entiers de 1 à 12. Le transformer en une matrice 3 lignes et 4 colonnes puis en un tableau de données (data frame) de mêmes dimensions. Renommer les colonnes A, B, C et D.

Exercice 1.2. Manipuler des vecteurs.

1. Afficher la troisième composante du vecteur créé à la question 1 de l'exercice précédent puis les composantes de rang 4 à 7 et enfin les cinquième et huitième composantes.
2. Afficher les composantes supérieures à 10 du vecteur créé à la question 2 de l'exercice précédent.

3. Créer les vecteurs deux vecteurs $x = (2.3, 3.5, 6, 14, 12)$ et $y = (3.2, 5, 0.7, 1, 3.5)$. Concaténez ces vecteurs en un vecteur z.
4. Afficher les composantes du vecteur x pour lesquelles les composantes de y sont inférieures à 3.4.
5. Créer un vecteur contenant les entiers de 1 à 12. Remplacer :
   — la cinquième valeur par 2 ;
   — toutes les valeurs supérieures ou égales à 10 par 3 ;
   — la quatrième et la neuvième valeurs par 1.

Exercice 1.3. On a compté le nombre d'arbres plantés sur les parcelles d'un lotissement. On a obtenu les données suivantes :

1, 2, 4, 1, 6, 3, 2, 1, 2, 0, 1, 2, 2, 1, 3, 0, 3, 2, 1, 2, 2, 3, 2, 3

1. Rentrer ces données sous la forme d'un vecteur nommé *arbres* et afficher ce vecteur.
2. Donner la taille de l'échantillon.
3. Trier les valeurs de ce vecteur par ordre croissant.
4. Afficher les effectifs de la variable nombre d'arbres (utiliser la fonction *unique*).
5. Calculer les effectifs puis les effectifs croissants.
6. Calculer les fréquences puis les fréquences cumulées.

II. Statistique descriptive unidimensionnelle.

Exercice 1.4. Variables quantitatives discrètes - Représentations graphiques - Caractéristiques numériques.

On reprend les données de l'exercice précédent.

1. Tracer le diagramme en bâtons des effectifs puis celui des fréquences.
2. Tracer la diagramme cumulatif des effectifs (utiliser d'abord la fonction *plot* puis la fonction *ecdf*).
3. Tracer la boîte à moustaches.
4. Afficher les caractéristiques numériques de la variable. Commenter.
5. En utilisant les fonctions adéquates, calculer successivement :
   — la moyenne, l'intervalle de valeurs (étendue), la médiane ,
   — les quartiles, la variance, l'écart-type (pour les deux dernières valeurs, calculer les valeurs pour la population).

Exercice 1.5. Au poste de péage, on compte le nombre de voitures se présentant sur une période de 5mn. Sur 100 observations de 5mn, on obtient les résultats suivants :

Nombre de voitures	1	2	3	4	5	6	7	8	9	10	11	12
Effectifs	2	8	14	20	19	15	9	6	2	3	1	1

Faire une étude complète de ces données sur le logiciel R.

Exercice 1.6. Étude des données *iris*. Variable quantitative continue.

1. Décrire les données iris (taille de l'échantillon, variables étudiées, types des variables).
2. Calculer les caractéristiques numériques de la variable *Sepal.Length* à l'aide de la fonction *summary*.
3. Utiliser la fonction *hist* pour faire un regroupement en classes de la série : quel est le nombre de classes obtenues ? Quelle est leur amplitude.
4. Donner les effectifs, effectifs cumulés, fréquences et fréquences cumulées pour ce regroupement.
5. Tracer l'histogramme des fréquences, puis le polygone cumulatif.
6. À partir de ce regroupement, calculer les caractéristiques numériques de la variable : moyenne, variance et écart-type.
7. Faire un autre regroupement en 5 classes d'amplitudes respectives 0.5, 1, 1, 0.5 et 1 et débutant à la valeur 4. Tracer l'histogramme correspondant.

Exercice 1.7. Reprendre l'exercice précédent en analysant la variable *Petal.Width* pour l'espèce *versivolor* (pour la dernière question choisir le découpage).

Exercice 1.8. Le tableau ci-dessous donne la répartition des individus d'une population suivant leur âge et leur sexe.

Âge	[0,2[	[2,6[	[6,10[	[10,20[	[20,30[	[30,40[	[40,50[
Hommes	11	40	49	80	100	70	60
Femmes	10	39	48	75	95	70	65

1. Créer deux vecteurs, un pour les hommes et l'autre pour les femmes, avec les données ci-dessus : on utilisera, pour résumer les valeurs d'une classe, son centre.
2. Faire une première description des données : étendue, moyenne, variance, écart-type.
3. Utiliser la fonction *hist* sans choisir les bornes de classes. Commenter.
4. Reprendre les classes proposées dans le tableau, tracer l'histogramme et le polygone cumulatif.

Exercice 1.9. Deux types de distributions très différentes.

1. Le tableau suivant indique la répartition de l'ancienneté du personnel cadre dans une entreprise (source : Bressoud, E., Kahané, J.-C.) :

Ancienneté	Effectifs
$[6,5; 8[$	3
$[8; 9,5[$	8
$[9,5; 11[$	12
$[11; 12,5[$	19
$[12,5; 14[$	9
$[14; 15,5[$	5
$[15,5; 17[$	5

Saisir les données puis tracer l'histogramme en utilisant les bornes de classes du tableau.
Représenter la boîte à moustaches.
Calculer la moyenne puis la médiane.

2. Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon le SAU (surface agricole utilisée) exprimée en hectares (source INSSE et Besse, P. Wikistat) :

SAU (en ha)	Fréquences (%)
$[0; 5[$	24
$[5; 10[$	11
$[10; 20[$	12
$[20; 35[$	19
$[35; 50[$	9
$[50; 200[$	5

Saisir les données puis tracer l'histogramme en utilisant les bornes de classes du tableau. Représenter la boîte à moustaches.
Calculer la moyenne puis la médiane.

3. Comparer les deux distributions : forme de l'histogramme, tracer sur le même graphique les boîtes à moustaches des variables centrées et réduites (c'est-à-dire des variables moins leur moyenne divisées par l'écart-type).

Exercice 1.10. Données iris. Étude de la variable *Species*.

1. Décrire la variable : nombre et noms des modalités.
2. Donner le tableau des effectifs puis celui des fréquences.
3. Tracer un diagramme en colonnes puis un diagramme en secteurs.

Exercice 1.11. Le tableau ci-dessous donne la répartition de la couleur des yeux dans une population.

Couleur des yeux	Effectifs
Marron	52
Bleus	31
Verts	9
Autre	8

1. Quel est le type de la variable étudiée ?
2. Quelle est la taille de l'échantillon ?
3. Calculer les fréquences.
4. Représenter la distribution de la variable par un diagramme en colonnes puis un diagramme en secteurs.

Exercice 1.12. (d'après Grais, B. Exercices corrigés de statistique descriptive, DUNOD). Le tableau suivant donne, par type de produits, la consommation finale d'énergie primaire observée en 1977 au Royaume-Uni et au Japon (unité : million de tonnes-équivalent pétrole).

Type d'énergie	Royaume-Uni	Japon
Combustibles solides	21,8	32,4
Produits pétroliers	72,6	172,9
Gaz	32,3	8,4
Électricité primaire	20	40,5

1. Quel est le type de la variable étudiée ?
2. Que représentent ici les ”effectifs” ?
3. Tracer un diagramme en colonnes représentant la distribution des types de produits pour chaque pays, puis un diagramme en secteurs.
4. Comparer la consommation du Royaume-Uni à celle du Japon (volume et structure).

Feuille TD 2

Quelques fonctions utiles :

Fonction	Description
table()	Calculer des effectifs
barplot()	Diagramme en barres
round()	Calculer des valeurs arrondies
colnames()	Affecter des noms de colonnes
rownames()	Affecter des noms de lignes
margin.table()	Calculer les effectifs marginaux
t()	Calculer la transposée d'une matrice
%*%	Effectuer un produit matriciel
chisq.test()	Test du khi-deux

Exercice 2.1. 400 clients d'un supermarché ont accepté de donner leur âge et de dire s'ils avaient ou non acheté un produit présenté en tête de gondole. Les âges sont regroupés en trois catégories : jeunes, adultes et anciens.

Âge / Achat	Jeunes	Adultes	Anciens
Oui	36	99	96
Non	24	121	24

1. Saisir les effectifs observés dans une matrice avec les noms de lignes et de colonnes identiques à ceux du tableau ci-dessus.
2. Faire une représentation graphique.
3. Calculer la taille de l'échantillon n.
4. Calculer les effectifs marginaux $ni.$ (lignes) et $n.j$ (colonnes).
5. Établir la table des fréquences (en pourcentage arrondis à 2 décimales).
6. Établir la table des profils lignes, en faire une représentation graphique.
7. Établir la table des effectifs théoriques : utiliser les vecteurs des effectifs marginaux $ni.$ et $n.j$ puis un produit matriciel %*% ainsi que la transposée t.
8. Calculer les écarts entre les effectifs observés et les effectifs théoriques puis chaque contribution du khi-deux.
9. Calculer la valeur du khi-deux.
10. Retrouver le khi-deux directement en utilisant la fonction *chisq.test*.
11. En se basant sur les résultats précédents, faire une analyse de la liaison entre les variables X et Y.

Exercice 2.2. Reprendre les données *films* de la feuille précédente. On définit la variable X, "fréquentation cinéma" , à 3 modalités : faible (0 ou 1 film), moyenne (2 à 4 films), forte (5 ou 6 films). La variable Y est le sexe.

1. Dresser la table de contingence donnant les effectifs conjoints des variables X et Y ainsi que les effectifs marginaux.
2. Déterminer les profils-colonnes et faire une représentation graphique.
3. Calculer le khi-deux et faire une analyse de la liaison entre X et Y.

Exercice 2.3. Trois échantillons de plantes sont traités avec trois engrais différents. Les résultats obtenus sont les suivants.

Y / X	Engrais A	Engrais B	Engrais C
Ont fleuri	34	73	63
N'ont pas fleuri	16	12	12

1. Calculer les effectifs marginaux et la taille de l'échantillon.
2. Calculer les fréquences de la variable "Floraison" conditionnellement au type d'engrais (profils colonnes). Commenter.
3. Calculer les effectifs théoriques dans le cas d'une non liaison entre les variables X et Y.
4. Donner la valeur du χ^2 puis du C de Cramer. Commenter.

Exercice 2.4. Dans une population on a relevé pour chaque individu la catégorie socio-professionnelle CSP, (variable X) et le type de produits consommés Produits (variable Y). Les effectifs obtenus sont donnés dans la table ci-dessous :

Produits / CSP	Prod1	Prod2	Prod3	Prod4
Agriculteurs	84	97	60	55
Cadres	40	75	45	45
Employés	96	85	84	66

1. Calculer les effectifs marginaux
2. Calculer les profils lignes et commenter.
3. Calculer les effectifs théoriques dans le cas d'une absence de liaison entre les deux variables statistiques.
4. Calculer le χ^2.

Feuille TD 3

Quelques fonctions utiles :

Fonction	Description
cov()	Calculer la covariance (échantillon)
cor()	Calculer le coefficient de corrélation linéaire
lm()	Ajuster un modèle linéaire
abline()	Ajouter un droite sur un graphique
coef()	Calculer les coefficients produits par le modèle
fitted()	Calculer les valeurs ajustées produites par un modèle
residuals()	Calculer les résidus
layout(), par(mfrow=...)	Partitionner la fenêtre graphique
log()	Calculer un logarithme

Exercice 3.1. Le tableau suivant donne la quantité d'engrais utilisé (variable X, en kg) et le rendement de maïs (variable Y en quintal) sur des parcelles de terrain similaires.

N° de parcelle	Q$^{\text{té}}$ d'engrais (X)	Rendement (Y)
1	20	16
2	24	18
3	28	23
4	22	24
5	32	28
6	28	29
7	32	26
8	36	31
9	41	32
10	41	34

1. Créer un data frame *mais* avec les données ci-dessus et représenter le nuage de point de la variable Y en fonction de la variable X.
2. Afficher un sommaire des caractéristiques numériques des variables X et Y.
3. Calculer la covariance entre X et Y en utilisant son expression puis directement avec la fonction *cov*. Que remarque-t-on ?
4. Calculer le coefficient de corrélation linéaire à l'aide de son expression puis avec la fonction *cor*.
5. Créer l'objet *mais.lm* contenant les résultats obtenus par un ajustement linéaire de la variable Y sur la variable X. Quelle est la classe de cet objet ? Montrer qu'il s'agit d'une sous-classe de la classe *list*.
6. Combien de composantes y a-t-il dans *mais.lm* ? Quels sont leurs noms ?
7. Quels sont les coefficients de la droite de régression de Y sur X ? Tracer la droite de régression sur le graphique de la question 1.
8. Quelle est la valeur ajustée pour le 3$^{\text{ème}}$ individu ?
9. Afficher les résidus ainsi que leurs caractéristiques numériques.

10. Partitionnez la fenêtre graphique en 4 : 2 lignes et 2 colonnes. Faire un plot de l'objet *mais.lm*. Commenter le premier graphique.

Exercice 3.9. Pour mesurer le degré d'intoxication de brochets par des pesticides, on a mesuré le taux de DDT dans la chair de brochets de différents âges, capturés dans une même rivière :

Âge	TxDDT
2	0,20
2	0,25
2	0,18
3	0,19
3	0,29
3	0,28
4	0,31
4	0,33
4	0,36
5	0,71
5	0,38
5	0,47
6	1,10
6	0,83

1. Créer un date frame *brochet* avec les données ci-dessus.
2. En reprenant les étapes de l'exercice précédent, réaliser la régression linéaire de la variable *TxDDT* sur la variable $\hat{A}ge$.
3. Ajouter une colonne au data frame *brochet* contenant le log de la variable *TxDDT*.
4. Réaliser la régression linéaire de la variable *log(TxDDT)* sur la variable $\hat{A}ge$. Comparer avec les résultats obtenus à la question 2.

Feuille TD 4

Quelques fonctions utiles :

tapply()	Calculer une fonction sur plusieurs groupes
boxplot()	Représenter une boîte à moustaches
aov()	Évaluer un modèle d'analyse de la variance

Exercice 4.1. (d'après Bertrand, F., Maumy-Bertrand M., Initiation à la statistique avec R, DUNOD). Des arbres ont été plantés en trois endroits différents. La hauteur des arbres a été mesurée après plusieurs années. Le tableau ci-dessous donne ces hauteurs.

Forêt 1	Forêt 2	Forêt 3
23,4	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24
25	22,5	24
26,2	23,5	24,5

1. Enregistrer les données relatives au tableau ci-dessus dans un data frame (fichier "Données Feuille 4", première feuille de calcul).
2. Calculer les caractéristiques numériques pour l'ensemble des arbres (moyenne, variance, médiane, quartiles).
3. Calculer les caractéristiques numériques pour chaque forêt.
4. Faire une représentation graphique du type boîtes parallèles. Commenter.
5. Calculer la variance inter-classes puis le rapport de corrélation. Conclure.
6. Calculer le rapport de corrélation directement à l'aide la fonction *aov*.

Exercice 4.2. Le tableau suivant présente des mesures de la hauteur (en mm) de la plante Saede brassica, réalisées dans plusieurs milieux différents, sur des prélèvements échantillonnés aléatoirement. Un chercheur désire comparer ces données afin de connaître l'effet du milieu sur la taille de S. brassica.

Milieu 1	Milieu 2	Milieu 3	Milieu 4	Milieu 5
12	141	56	87	241
15	146	67	105	264
12	135	43	79	225
18	147	78	123	257
24	154	45	114	248
32		69		248
31				236
15				

1. Enregistrer les données relatives au tableau ci-dessus dans un data frame (fichier "Données Feuille 4", seconde feuille de calcul). Le data frame comprendra deux colonnes, l'une avec les valeurs de la variable X, hauteur, et l'autre avec celles de la variable Y, milieu.
2. Reprendre la démarche de l'exercice précédent pour faire une analyse de la liaison entre la hauteur et le milieu de la plante.

Feuille 4
Séries chronologiques

Exercice 4.1. (d'après Badia, J., Bastida, R. Haït, J.-R.. Statistique sans mathématique, Technoqup). (NB. on pourra utiliser un tableur pour résoudre cet exercice, *cf.* TP) Le tableau ci-dessous indique les chiffres d'affaires d'une société :

Année	Trimestre	CA
1985	1	5595
	2	4490
	3	6430
	4	7685
1986	1	5735
	2	4630
	3	6565
	4	7830
1987	1	5870
	2	4765
	3	6695
	4	7950
1988	1	5990
	2	5190
	3	7210
	4	8810
1989	1	7200
	2	6395
	3	8175
	4	9270
1990	1	7165
	2	5910
	3	7690
	4	8795
1991	1	7210
	2	6410
	3	8390
	4	9990
1992	1	7995
	2	7200
	3	9205
	4	10800

1. Représenter la série chronologique du chiffre d'affaires.
2. Déterminer la tendance par la méthode des moyennes mobiles d'ordre 4 (puis d'ordre 2).
3. Calculer les coefficients saisonniers en utilisant un modèle additif.
4. Déterminer la série ajustée.
5. Calculer la série corrigée des valeurs saisonnières.

Exercice 4.2. (d'après Bressoud, E, Kahané, J.-C. Statistique descriptive, Pearson). Le tableau ci-dessous indique les entrées trimestrielles, en millions, dans les salles de cinéma en France :

Trimestre	2004	2005	2006
1	50,46	45,34	51,63
2	51,46	41,86	51,06
3	41,07	35,14	35
4	52,34	52,99	50,76

1. Représenter la série chronologique. Quel modèle suggère cette représentation ?
2. Déterminer la tendance linéaire par la méthode des moindres carrés.
3. Calculer les coefficients saisonniers.
4. Déterminer la série ajustée.
5. Représenter sur un même graphique la série brute, la tendance obtenue à la question 2 et la série ajustée.
6. Calculer la série corrigée des valeurs saisonnières.
7. Proposez des prévisions de fréquentations trimestrielles pour l'année 2007.

Exercice 4.3. (d'après Badia, J., Bastida, R. Haït, J.-R.. Statistique sans mathématique, Technoqup). Le tableau ci-dessous indique les chiffres d'affaires de deux sociétés :

Année	Trimestre	Société 1 CA	Société 2 CA
1990	1	15	7
	2	13	9
	3	18	13
	4	22	15
1991	1	18	8
	2	15	11
	3	20	17
	4	23	20
1992	1	18	10
	2	16	14
	3	21	22
	4	25	26
1993	1	20	12
	2	18	16
	3	23	24
	4	27	28

1. Représenter la série chronologique pour chaque société. Commenter.
2. Étude du chiffre d'affaires pour la société 1. On utilisera un modèle additif.
   (a) Déterminer la tendance linéaire par la méthode des moindres carrés.
   (b) Calculer les coefficients saisonniers.
   (c) Déterminer la série ajustée.
   (d) Calculer la série corrigée des valeurs saisonnières.
   (e) Proposez des prévisions du chiffre d'affaires pour l'année 1994.
3. Étude du chiffre d'affaires pour la société 2. On utilisera un modèle multiplicatif. Reprendre les questions (a)-(e) dans ce cadre.