INSA Rouen – STPI 2
Cahier d'exercices (TD et TP)

Stéphane Canu, Remi Flamary et Karina Zapién Arreola

Janvier 2012

Résumé

Ce recueil d'exercices est principalement destiné aux élèves ingénieurs qui suivent un premier cours sur les principes du traitement de l'information, mais s'adresse aussi à tous ceux qui souhaitent parfaire leurs connaissances dans l'un ou l'autre des sujets traités à savoir : la description des données, l'étude de la relation entre deux variables et les test statistiques de student et du chi 2. Il a été établi, à l'aide des tous les enseignants qui y ont participé, d'après un certain nombre d'exercices de l'UTC (UV SY02) [1]

Table des matières

1. http://www4.utc.fr/~sy02

1 Questions de cours

Exercice 1 — Demandes à l'entourage du roi

1. qu'est-ce qu'un OPM ?
2. quelle est la différence entre une moyenne et une espérance ?
3. quelles sont les différences entre une variable qualitative et une variable quantitative
4. quelles sont les différences entre une probabilité et une fonction densité de probabilité

Exercice 2 — Question de vocabulaire

Nous avons vu en cours le tableau suivant :

Théorique	*Empirique*
$\mathbb{P}(x)$ mesure de probabilité	$\widehat{\mathbb{P}}(x)$mesure empirique
$F(x)$ fonction de répartition	$\widehat{F}(x)$
$\mathbb{E}(X)$ espérence	$\bar{X}$
σ^2	$\widehat{\sigma^2}$ variance empirique
M médianne	$\widehat{M}$ médianne empirique

1. précisez en vous aidant du tableau les principales différences entre variables qualitatives et variables quantitatives.
2. Soit X une variable aléatoire discrète à valeur dans l'ensemble $\{0, 1, 3, 4\}$ et $(X_1, X_2, ..., X_i, ..., X_n)$ un échantillon (une collection de n variables aléatoires i.i.d.)
   a) donnez un exemple d'une telle variable aléatoire et un exemple d'échantillon pour $n = 10$,
   b) définissez $\mathbb{P}(x)$, $F(x)$, $\mathbb{E}(X)$, σ^2 et M
   c) calculez $\widehat{\mathbb{P}}(x)$, $\bar{X}$, $\widehat{\sigma^2}$ et $\widehat{M}$ à partir de votre échantillon,
   d) Comment appelle t'on les quantités empiriques $\widehat{F}(x)$ et $\bar{X}$ associées à la fonction de répartition et à l'espérence.
   e) Construisez $\widehat{F}(x)$ pour votre échantillon.
3. Soit Y une variable aléatoire continue sur $\mathbb{R}$ et $(Y_1, Y_2, ..., Y_i, ..., Y_n)$ un échantillon (une collection de n variables aléatoires i.i.d.)
   1. donnez un exemple d'une telle variable aléatoire et un exemple d'échantillon pour $n = 10$,
   2. définissez $\mathbb{P}(x)$, $F(x)$, $\mathbb{E}(X)$ et M
   3. calculez $\widehat{\mathbb{P}}(x)$ et $\widehat{F}(x)$ à partir de votre échantillon.

Exercice 3 — C'est la rose

Parmi ces résultats vus en cours lequel vous semble le plus important et pourquoi ?

1. $\mathrm{a} = (X^\top X)^{-1} X^\top \mathrm{y}$
2. $\widehat{\sigma}_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \mathrm{x})^2$
3. quand deux variable aléatoires x et y sont indépendantes on a $\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$
4. $\min_{\mathrm{u,v}} \|X - \mathrm{u}\mathrm{v}^\top\|_F^2 \quad \Leftrightarrow \quad X^\top X \mathrm{v} = \lambda \mathrm{v}$
5. $c_i = \dfrac{H_{ii}}{p(1 - H_{ii})^2} \dfrac{\widehat{e}_i^2}{s^2}$

Attention : il ne faut donner qu'une et une seule réponse

2 Statistiques descriptives

2.1 Variables Statistiques

Exercice 4 Fréquence, cumulées, fonction de répartition et densité empiriques

— Les deux tableaux suivants représentent : A gauche l'étude du taux de cholestérol sur un échantillon de 100 personnes et à droite TW est une mesure de précipitation sur la partie ouest de Tasmanie et SC une mesure de précipitation sur le sud toujours de la Tasmanie. La variable SEEDED précise si ces précipitations ont été obtenues après avoir ensemencés les nuages (S) ou non (U).

Taux de Cholestérol (gr/l)	effectifs
$[1.0; 1.4[$	6
$[1.4; 1.6[$	13
$[1.6; 1.8[$	16
$[1.8; 2.0[$	22
$[2.0; 2.2[$	18
$[2.2; 2.4[$	10
$[2.4; 2.6[$	6
$[2.6; 2.8[$	4
$[2.8; 3.0[$	3
$[3.0; 3.4[$	2

SEEDED	SEASON	TW	SC
S	AUTUMN	3,50	1,40
U	AUTUMN	0,78	0,79
S	WINTER	0,75	0,36
U	WINTER	2,01	1,27
S	WINTER	4,61	2,16
U	WINTER	1,90	0,55
U	WINTER	1,37	0,85
S	WINTER	0,90	0,65
U	SPRING	2,1	1,08
S	SPRING	3,00	3,10
S	SPRING	1,46	0,64
U	SPRING	2,79	1,30

— On a compté le nombre de personnes faisant la queue aux caisses d'un supermarché ce qui nous a donné l'échantillon :

caisse	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
nb de personnes	1	3	4	0	7	2	0	4	2	4	3	2	3	5	99	4	0	3	7

1. Pour chacune de ces variables (cholestérol, SEASON, TW et nb de personnes) préciser le type de variable considéré et donner lorsque c'est possible un tableau récapitulant :
   a) les fréquences
   b) les fréquences cumulées
   c) la fonction de répartition empirique

Exercice 5 de quoi s'agit-il ?

Chaque fois que vous prenez l'avion, un enregistrement est ajouté à un fichier central. Cet enregistrement contient une centaine de champs dont les champs suivants :

nom de la variable	*exemple*
identifiant	242 335 976 54 24
nom	Peutu
prénom	Stefen
date de naissance	13 mars 1938
nationalité	Australienne
aéroport d'arrivée	Canberra
prix du billet (en euros)	1124
mode de payement du billet	carte de crédit - liquide - chèque
nombre de bagages enregistrés	2

pour chacune des variables du fichier « voyageur », préciser le domaine et de quel type de variable il s'agit :
— d'une variable aléatoire ou non ,
— d'une variable qualitative ou quantitative,

— si elle est discrète ou continue,

Exercice 6 — Mon choix

Afin de mieux connaître la provenance des étudiants, on leur a demandé de remplir un questionnaire dans lequel ils ont indiqué leur bac d'origine (L, S ou ES). On a obtenu les résultats suivants :

S	S	L	S	S	S	L	ES	ES	ES	L	S	S
S	ES	S	L	S	S	S	L	ES	ES	S	ES	S
ES	S	L	S	S	L	S	S	L	ES	ES	L	L
S	S	S	ES	S	L	S	S	S	L	S	S	S
S	ES	S	S	S	S	S	S	S	S	L	ES	ES
L	L	S	S	S	ES	S	S	S	S	S	L	S

1. Précisez la nature de la variable observée et les objectifs de l'étude.
2. Calculez la distribution de probabilité de la variable observée, en effectif et en fréquence.

Exercice 7 — Du beurre de pâté de graisse

Voici un ensemble de relevés du niveau de cholestérol des enseignants de l'INSA :

86	90	98	87	88	92	98	92	91	94	90	93	101
97	97	95	93	97	76	96	90	98	94	89	89	95

1. Précisez la nature de la variable observée et les objectifs de l'étude,
2. Calculez les effectifs, fréquence et fréquence cumulée, tracer la fonction de répartition empirique.
3. Discrétiser la variable aléatoire en 4 modalités.
4. Tracer les effectifs, fréquences et la fonction de répartition empirique après discrétisation.
5. Supposons que les observations dont nous disposons sont issues d'une loi normale :
   (a) Quelle est la probabilité qu'une future observation se trouve à une distance de plus de 4 écart type de la moyenne.
   (b) Donnez un intervalle dans lequel se trouvera (avec grande probabilité) une future observation.
6. Répondre aux deux questions précédentes dans le cas où l'on en connait pas la loi dont sont issues les observations.

Exercice 8 — Millionnaire

- Voici le détail des gains possibles lorsque l'on tourne la roue de la fortune à la télévision.

Gains	10^6	600 k€	500 k€	400 k€	300 k€	200 k€	100 k€
Positions	8	14	16	16	16	16	14

Si l'on est invité à faire tourner cette roue, quel est le gain moyen que l'on peut attendre ?

- La française des jeux imprime des liasses de 500.000 tickets qu'elle vend 10€ et parmi lesquels on compte le nombre suivant de gagnants :

Gains	TV	50 k€	10 k€	1 k€	500 €	100 €	50 €	20 €	10 €
Tickets	1	8	12	58	240	1500	5500	30 000	87500

Gratter TV donne le droit d'aller à la télévision pour faire tourner la roue de la fortune...
Quel est la probabilité d'acheter un ticket gagnant ? Quel est le gain minimum, maximum et moyen pour la française des jeu sur une liasse de 500.000 tickets ?

Exercice 9 — des queues de cerises

Les cerises Momenrency du sud ouest de la France ont un poids suivant une distribution normale centrée sur $\mu = 5,02$ grammes et de variance $\sigma^2 = 0,30$ grammes.

1. Trouvez la probabilité pour qu'un échantillon de 100 cerises prises au hasard ait un poids total :
   1. compris entre 496 et 500 grammes ?
   2. strictement supérieur à 510 grammes ?
2. Si on tire au hasard deux lots de 1000 cerises chacun, quelle est la probabilité que leur poids diffère de plus de 2 grammes ?
3. Donnez un intervalle en gramme, centré autour de la moyenne, dans lequel 75 % des cerises sont.
4. Que devient cet intervalle si l'on abandonne l'hypothèse de normalité ?

2.2 Description des variables

Exercice 10 — Centrage et Jivaro

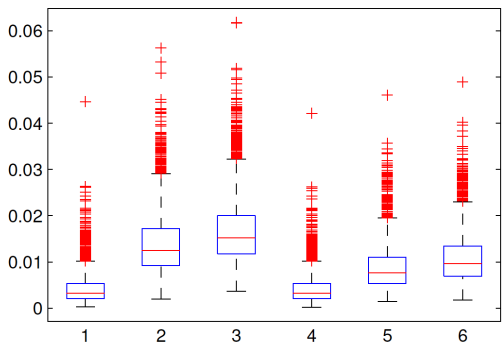
On a compté le nombre de personnes faisant la queue aux caisses d'un supermarché ce qui nous a donné l'échantillon :

caisse	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
nb de personnes	1	3	4	0	7	2	0	4	2	4	3	2	3	5	99	4	0	3	7

Sur la ligne du haut on trouve le numéro de la caisse et sur celle d'en dessous le nombre de personnes faisant la queue.

1. calculez la moyenne et la variance empirique de cet échantillon,
2. dessinez la boite à moustaches de cet échantillon,
3. transformez cet échantillon en un nouvel échantillon équivalent mais centré et réduit, et proposez une interprétation à cette nouvelle variable (un nom par exemple),
4. discrétiser la variable aléatoire en 3 modalités et dessiner l'histogramme associé.

Exercice 11 — Bigote



La figure ci-dessus présente pour 6 méthodes différentes (abscisse) les performances en terme de probabilité d'erreur sur des jeux de test (ordonnée). Plus la probabilité d'erreur est faible, meilleure est la méthode.

1. quels commentaires vous inspire cette figure ? quelle(s) est (sont) les meilleure(s) méthode(s) ? quelles sont celles qui sont équivalentes et pourquoi ?

Exercice 12 — Beurre de paté de graisse (bis)

L'étude du taux de cholestérol sur un échantillon de 100 personnes a donné les résultats suivants :

Taux de Cholestérol (gr/l)	effectifs
$[1.0; 1.4[$	6
$[1.4; 1.6[$	13
$[1.6; 1.8[$	16
$[1.8; 2.0[$	22
$[2.0; 2.2[$	18
$[2.2; 2.4[$	10
$[2.4; 2.6[$	6
$[2.6; 2.8[$	4
$[2.8; 3.0[$	3
$[3.0; 3.4[$	2

1. Tracer l'histogramme. Justifier les valeurs données.
2. Tracer la courbe de la fonction de répartition empirique.
3. Déterminer le mode et la moyenne de cette distribution.
4. Dessiner la boite à moustache de ces observations.

Exercice 13 — La moyenne des vitesses et la vitesse moyenne

Un alpiniste est en train de monter une montagne, sa vitesse change au fur et à mesure qu'il gravit la montagne. On a mesuré sa vitesse sur des portions de 1 km dont la pente était constant. Il a maintenu une vitesse constante pour chaque km, qui était mesuré et enregistré dans le tableau suivant :

Portion	km 0 à 1	km 1 à 2	km 2 à 3	km 3 à 4	km 4 à 5
Vitesse (en km/h)	10	5	3,33	2,5	2

1. calculez la moyenne et la médiane des vitesses,
2. calculez le temps de parcours. En déduire la vitesse moyenne,
3. montrez que la moyenne harmonique

$$c = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{v_i}}$$

est le résultat de la minimisation suivante

$$\min_c \sum_{i=1}^{n}\left(\frac{1}{v_i} - \frac{1}{c}\right)^2$$

4. calculez la moyenne harmonique des vitesses. Comparez la à la moyenne des vitesses et à la vitesse moyenne et commentez.

Exercice 14 — Dégraissage

Une entreprise emploie 200 ouvriers dont le salaire annuel est de 15 000 euros et deux cadres dont le salaire annuel est de 45 000 euros. A la suite d'une restructuration de l'entreprise et de nombreux licenciements, le responsable du personnel peut annoncer aux syndicats que les salaires ont progressé de plus de 15 % en moyenne et dans le même temps pour rassurer les actionnaires de l'entreprise en faisant état d'une augmentation limitée à 5% pour les cadres et 10% pour les ouvriers.

1. Comment cela est-il possible ?

Exercice 15 La moyenne des taux et le taux moyen 6 points

L'inflation à été très importante en Argentine. Nous allons supposer qu'il est prévue une évolution de l'inflation donnée par an dans le tableau suivant :

Année	2013	2014	2015	2016	2017
inflation (en $\%$)	$i_1 = 250$	$i_2 = 100$	$i_3 = 200$	$i_4 = 50$	$i_5 = 400$

A une une inflation de i $\%$ on associe le taux d'inflation $t = 1 + i/100$ de sorte que le prix en fin d'année puisse être calculé par une simple multiplication.

1. calculez les taux d'inflation de 2013 à 2017, leur moyenne et leur médiane,
2. calculez le prix fin 2017 d'un article coutant 100 pesos début 2013. En déduire le taux d'inflation moyen sur la période,
3. montrez que la moyenne géométrique
$$c = \sqrt[n]{\prod_{i=1}^{n} t_i}$$
est le résultat de la minimisation suivante
$$\min_c \sum_{i=1}^{n} (\log t_i - \log c)^2$$
4. calculez la moyenne géométrique des taux d'inflation. Comparez la à la moyenne des taux et au taux moyen et commentez.

Exercice 16 Boites à moustaches

X	0	0,5	0,75	1,5	2	2	2,5	2,75	3	3	3,5	3,5	4	4,5	4,5
Y	1,5	1,5	0	-2,5	0,5	3,5	3	3,5	1	4,5	6,5	5	12	1,5	11,5

1. Comparez les boites à moustaches des deux variables
2. En faisant un hypothèse de normalité, donnez un intervalle dans lequel on va trouver 90 % des valeurs de X
3. Donnez un intervalle dans lequel on va trouver 80 % des valeurs de Y

2.3 Couples de variables

Exercice 17 Corrélations

X	0	0,5	0,75	1,5	2	2	2,5	2,75	3	3	3,5	3,5	4	4,5	4,5
Y	1,5	1,5	0	-2,5	0,5	3,5	3	3,5	1	4,5	6,5	5	12	1,5	11,5

1. Représentez graphiquement le nuage de points, avec ses deux boites à moustaches
2. Calculez la moyenne du nuage,
3. Représentez les enveloppes convexes successives,
4. Trouvez graphiquement les médianes de Jarvis et de Tukey,
5. Calculez le coefficient de corrélation et la covariance du nuage de points.
6. qu'en pensez vous en terme de relation entre les variables et de points aberrants

Exercice 18 Chant de nuage !

`http://lib.stat.cmu.edu/datasets/cloud`

SEEDED	SEASON	TW	SC
S	AUTUMN	3,50	1,40
U	AUTUMN	0,78	0,79
S	WINTER	0,75	0,36
U	WINTER	2,01	1,27
S	WINTER	4,61	2,16
U	WINTER	1,90	0,55
U	WINTER	1,37	0,85
S	WINTER	0,90	0,65
U	SPRING	2,1	1,08
S	SPRING	3,00	3,10
S	SPRING	1,46	0,64
U	SPRING	2,79	1,30

$$\sum_{i=1}^{12} TW_i = 25,17$$

$$\sum_{i=1}^{12} TW_i^2 = 68,34$$

$$\sum_{i=1}^{12} SC_i = 14,15$$

$$\sum_{i=1}^{12} SC_i^2 = 23,32$$

$$\sum_{i=1}^{12} TW_i \times SC_i = 37,22$$

TW est une mesure de précipitation sur la partie ouest de Tasmanie et SC une mesure de précipitation sur le sud toujours de la Tasmanie. La variable SEEDED précise si ces précipitations ont été obtenues après avoir ensemencés les nuages (S) ou non (U).

1. précisez le type des variables,
2. calculez pour chaque variable sa moyenne et son écart type,
3. dessinez une mesure de probabilité associée à la variable « SEASON »,
4. dessinez la boite à moustache de la variable « SC », en précisant les valeurs numériques,
5. construisez un histogramme de la variable « SC »,
6. dessinez la médiane du couple de variables « TW » et « SC », et le *bagplot* en deux dimensions,
7. calculez la corrélation et la covariance entre les variables TW et SC,
8. pensez vous qu'il y ai des données aberrantes et pourquoi.

Exercice 19 Descriptions bidimensionelles

Lors d'une certaine expérience, des sujets sont soumis à deux stimuli. On observe leur temps de réaction mesuré en secondes. Les résultats obtenus sont récapitulés dans le tableau suivant :

Sujets	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Stimulus 1	35	29	30	31	33	32	33	34	32	33	32	30	28	34	33	33
Stimulus 2	35	36	34	34	36	37	33	35	35	35	35	33	36	37	34	35

On désigne par X la variable : temps de réaction au stimulus 1 et par Y la variable : temps de réaction au stimulus 2.

1. Cas des données individualisées
   (a) Tracer le nuage d'individus (des points) et placer le « centre de gravité ».
   (b) Calculer la covariance de X et Y.
   (c) Calculer le coefficient de corrélation linéaire entre X et Y.
2. Propriétés du coefficient de corrélation linéaire.
   Que devient le coefficient de corrélation linéaire précédent si :
   (a) On ajoute 3 à chaque valeur de X ?
   (b) On multiplie chaque valeur de Y par 2 ?
   (c) On ajoute le couple (32,35) (centre de gravité) aux observations ?
   (d) On ajoute plusieurs fois le couple (32,35) (centre de gravité) aux observations ?
3. Cas d'une organisation en tableau de contingence.
   On considère le tableau de contingence donnant la distribution conjointe du couple (X,Y).
   (a) Donner la distribution conjointe du couple de variables (X,Y) sous la forme d'un tableau de contingence des effectifs et des fréquences.
   (b) Déduire du tableau précédent les distributions marginales et rappeler ce quelles représentent.
   (c) A partir de ces deux distributions marginales est-il possible de reconstruire le tableau de contingence ? Donner un (contre)exemple.
   (d) Combien de sujets ont mis un temps de réaction de 33 s au 1er stimulus et un temps de réaction de 35 s au 2eme stimulus ?
   (e) Combien de sujets ont mis un temps de réaction inférieur à 30 s au 1er stimulus et un temps de réaction supérieur à 35 s au 2eme stimulus ?
   (f) Quel est le nombre (le pourcentage) des sujets qui ont réagi moins vite au 1er stimulus qu'au 2eme stimulus ?
   (g) On considère la distribution des temps de réaction au 1er stimulus de tous les individus ayant mis un temps de réaction de 35 s au 2eme stimulus.
      i. Comment appelle-t-on une telle distribution ? Combien a-t-on de distributions conditionnelles différentes ?
      ii. Donner une représentation graphique de cette distribution, calculer sa moyenne et sa variance. Comment désigne-t-on ces indices ?
4. Autres tableau de contingence et tableau à deux étapes
   On décide de regrouper les observations de X en 3 classes [28 30] ;]30 33] et]33 35] que l'on appelle groupe 1, groupe 2 et groupe 3. (ou encore rapide, moyen et lent).
   (a) Dresser un tableau à deux étapes donnant les distributions de Y par groupe.
   (b) Dresser un tableau de contingence en tenant compte de ce nouveau groupement.
   (c) Calculer les moyennes conditionnelles par groupe et rappeler comment on en déduit la moyenne globale de Y.
   (d) Calculer les variances conditionnelles par groupe ainsi que les écart-types.
5. Opérations sur les variables
   A chaque individu on associe le temps moyen calculé à partir de ses deux temps de réaction au 1er et au 2eme stimulus. On définit ainsi une nouvelle variable Z.

(a) Exprimer Z en fonction de X et Y. Déduire la moyenne de Z à partir des moyennes de X et Y.

(b) Donner la distribution de Z et calculer sa variance. Etait-il possible de déduire cette variance à partir des variances de X et Y ?

Exercice 20 — Sudoku !

Nous avons le tableau de contingence incomplet suivant :

X \ Y	$y_1 = 1$	$y_2 = 4$	$y_3 = 8$	Marginal de X $\mathbb{P}(X)$
x_1	0,1			
x_2				
Marginal de Y $\mathbb{P}(Y)$	0,3			1

De plus, nous disposons de l'information suivante :

$$\mathbb{P}(Y = y_1 | X = x_2) = \frac{1}{2}$$

$$\mathbb{P}(Y = y_3 | X = x_1) = \frac{1}{2}$$

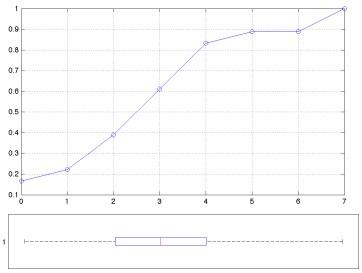
$$\mathbb{E}[Y | X = x_2] = 3$$

1. compléter le tableau de contingence,
2. les variables X et Y sont-elles indépendantes ?

Astuce : Pour le calculer plus rapidement, suivez l'ordre suivant :

X \ Y	$y_1 = 1$	$y_2 = 4$	$y_3 = 8$	Marginal de X $\mathbb{P}(X)$
x_1	.1	⑤	④	③
x_2	①	⑥	⑦	②
Marginal de Y $\mathbb{P}(Y)$	.3	⑧	⑨	1

de plus, les variables ⑥et ⑦s'obtiennent en résolvant un système de deux équations à deux inconnues.



2.4 Description multi variable

Exercice 21 — Lasse épée

Voici la copie d'une session Matlab. On y trouve des lignes de code qui commencent par le chargement d'un tableau de données X comportant $p = 3$ variables et $n = 15$ observations. Rappel : la fonction matlab `[U,d] = eig(M)` calcule les vecteurs propres (U) et les valeurs propres associées (sur la diagonale de d) de la matrice M.

1. dessinez la fonction de répartition empirique de la première variable,
2. donnez la boite à moustache correspondant à la première variable,
3. représentez graphiquement la médiane de Tukey et le sac médian, des deux premières variables
4. y a-t'il un point aberrant ?
5. quelle est la moyenne et la variance de Xn ?
6. quelle est la première composante principale et quelle est sa relation avec les trois variables originales ?
7. représentez le nuage de points dans les deux premiers axes de l'analyse en composantes principales (ACP).
8. comment caractériser la qualité de la représentation du nuage de points dans les deux premiers axes de l'ACP ?

```
load X
X = 8.5635   -0.4249   -1.1302
   10.6189   -0.6752    0.5311
   10.3912   -0.5847    0.2538
   10.4251   -0.6200    0.7893
   11.1107   -0.5394    0.6410
    9.1614   -0.5497    0.1627
   10.5240   -0.4315    0.2750
   10.8754   -0.4242    0.4208
    8.4377   -0.3782   -1.3738
    9.0995   -0.3888   -1.2308
   10.6006   -0.5199    0.9124
   11.1059   -0.5205    0.8106
    8.7338   -0.3190   -1.1281
    9.8828   -0.5151    0.2720
    8.3042   -0.3916   -1.5087

mean(X) = 9.8557   -0.4855   -0.0868

cov(X) = 1.0553   -0.0647    0.8529
        -0.0647    0.0101   -0.0692
         0.8529   -0.0692    0.8099

Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X));
cov(Xn) =  1.0000   -0.6251    0.9226
          -0.6251    1.0000   -0.7635
           0.9226   -0.7635    1.0000

[U,d]=eig(X'*X)
U =    0.0493   -0.0010   -0.9988
       0.9982   -0.0334    0.0493
       0.0334    0.9994    0.0006
diag(d) =   0.1      11.5     1475.4
```

```
[Un,dn]=eig(Xn'*Xn)
Un = -0.6025    0.5467    0.5814
      0.2211    0.8143   -0.5366
      0.7669    0.1948    0.6115
diag(dn) =   0.7704     5.5677     35.6618

V = X*U
V = -0.0395   -1.1243   -8.5747
    -0.1326    0.5423  -10.6390
    -0.0627    0.2623  -10.4073
    -0.0784    0.7987  -10.4425
     0.0309    0.6470  -11.1234
    -0.0914    0.1713   -9.1772
     0.0974    0.2782  -10.5323
     0.1270    0.4234  -10.8828
    -0.0072   -1.3692   -8.4469
     0.0195   -1.2267   -9.1084
     0.0343    0.9182  -10.6128
     0.0552    0.8159  -11.1176
     0.0746   -1.1259   -8.7396
    -0.0177    0.2787   -9.8961
    -0.0318   -1.5034   -8.3143

Vn = Xn*Un
Vn =
    0.0019   -0.4233   -1.7634
   -0.3375   -0.9941    1.8629
   -0.2415   -0.4430    1.0630
    0.1174   -0.5950    1.6345
   -0.2342    0.3898    1.4921
    0.4790   -0.8342    0.1184
    0.0349    0.8706    0.3364
   -0.0308    1.1484    0.5954
   -0.0293   -0.1650   -2.2492
   ...
```

Exercice 22 — Retour de l'épée fatiguée — 10 points

À la suite de l'énoncé, vous trouverez la copie d'une session Matlab. On y trouve des lignes de code qui commencent par le chargement d'un tableau de données X comportant $p = 4$ variables et $n = 25$ observations.

En annexe de cet énoncé, vous trouverez deux figures représentant le nuage points des deux premières variables du tableau X. Ces figures vous serviront à répondre aux deux premières questions de cet exercice.

Rappel : la fonction matlab `[U,d] = eig(M)` calcule les vecteurs propres (U) et les valeurs propres associées (sur la diagonale de d) de la matrice M.

1. représentez les enveloppes convexes successives et indiquez où se situe la médiane de Tukey du nuage de points,
2. représentez le sac médian du nuage de points,
3. y a-t'il des points aberrants ? quel impact pour la suite de l'analyse statistique ?
4. quelle(s) est/sont, selon vous, l'(es) étape(s) indispensable(s) à appliquer avant d'effectuer une ACP ?
5. quelle sont les moyennes et variances de X et de A ?
6. quelle relation permet de représenter les données par rapport à un axe de l'ACP ?
7. caractérisez la qualité de la représentation du nuage de points par rapport à la première composante de l'ACP ?
8. représentez les variables dans le plan des deux premières composantes principales. Quelle interprétation donnez vous à cette représentation ?

```
load X

X =

   -0.5410   -0.6331    0.3745    0.7852
    0.6451   -0.1519    0.9507    0.1997
    0.7272   -1.2195    0.7320    0.5142
    0.3686   -2.1837    0.5987    0.5924
   -1.5703    0.9193    0.1560    0.0465
   -2.1499    0.3417    0.1560    0.6075
   -0.0876   -2.6452    0.0581    0.1705
   -2.0000    1.5000    0.8662    0.0651
   -2.6466   -1.4144    0.6011    0.9489
    1.1217   -2.5342    0.7081    0.9656
    1.8794    1.0015    0.0206    0.8084
   -0.1118    2.3506    0.9699    0.3046
   -3.2037   -2.0683    0.8324    0.0977
   -2.8480    1.5765    0.2123    0.6842
    1.6140   -3.1144    0.1818    0.4402
   -2.1306   -3.5424    0.1834    0.1220
    3.0000   -1.0000    0.3042    0.4952
    2.0483   -3.5186    0.5248    0.0344
   -1.3114   -4.6953    0.4319    0.9093
   -3.6899   -3.7823    0.2912    0.2588
   -3.0942   -4.3754    0.6119    0.6625
   -1.5000    4.0000    0.1395    0.3117
   -2.9431    3.4100    0.2921    0.5201
   -3.9392   -4.3146    0.3664    0.5467
    6.3000   -8.6000    0.4561    0.1849

mean(X) = -0.6425   -1.3878    0.4408    0.4510
[n p] = size(X);
```

```
cov(X) = 5.9875   -2.1368    0.0211   -0.0569
        -2.1368    8.4077   -0.0333    0.0090
         0.0211   -0.0333    0.0813   -0.0071
        -0.0569    0.0090   -0.0071    0.0890

A = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X));
cov(A) = 1.0000   -0.3012    0.0302   -0.0779
         -0.3012    1.0000   -0.0403    0.0104
          0.0302   -0.0403    1.0000   -0.0832
         -0.0779    0.0104   -0.0832    1.0000

[U,d]=eig(X'*X)
U =-0.0076    0.0937    0.9617   -0.2573
    0.0027    0.1026    0.2478    0.9634
    0.7205    0.6870   -0.0761   -0.0556
   -0.6934    0.7133   -0.0887   -0.0512

diag(d) =  2.1989  8.4937  147.8644  259.4253

[Vp,valp]=eig(A'*A)
Vp =-0.7035   -0.1053    0.1665   -0.6828
    -0.6864   -0.1321   -0.2734    0.6608
    -0.0668    0.7115   -0.6691   -0.2041
    -0.1718    0.6821    0.6707    0.2354

diag(valp) = 16.5600 22.1583 25.4257 31.8561
```

Exercice 23 — Tout (ou presque) en un — 16 points

La tableau `X` (donné au verso) présente les données d'un fichier `cereal.mat`, présent sur votre bureau, contenant les mesures des ingrédients de 77 produits de céréales pour petit déjeuner. Sur chaque type de céréale, 8 variables ont été mesurées. Ces variables sont (dans l'ordre) :

Calories	la teneur en calories (le nombre pour un repas)
Carbo	la teneur en carbohydrates (en gramme pour un repas)
Cups	le nombre de tasse recommandé par repas
Fat	la teneur en graisse (en gramme pour un repas)
Fiber	la teneur en fibres alimentaires (en gramme pour un repas)
Potass	la teneur en potassium (en mg pour un repas)
Protein	la teneur en proteine (en gramme pour un repas)
Sugars	la teneur en sucre (en gramme pour un repas)

A chaque type de céréale a été associé son nom contenu dans la variable `Name`. Ainsi la première céréale est `Name(1) -> '100% Bran'`.

1. Nous allons commencer par nous intéresser à la deuxième variable (`Carbo`) uniquement.
   a) construisez un tableau avec les fréquences, les fréquences cumulées de la deuxième variable (`Carbo`).
   b) tracez sa fonction de repartition empirique
   c) déterminez graphiquement la médiane et les quartiles
   d) en déduire la DIQ
   e) dessinez la boite à moustache associée sur une feuille
   f) s'il y a une variable aberrante enlevez la et recalculez la mediane
2. Nous allons maintenant nous intéresser au couple de variables 1 et 8 (`Calories` et `Sugars`).
   a) représenter le nuage des points de ces deux variables à laide de votre ordinateur
   b) trouvez graphiquement la médiane de Tuckey du nuage de points donné figure 1.
   c) calculez la covariance et la corrélation de ce couple de variables
   d) cette valeur de corrélation vous semble t'elle significativement importante ?
   e) pourquoi la variable `Cv` n'a t'elle pas exactement la même valeur que `C(2,1)` ?

   ```
   n = length(x1);
   c1 = mean(x1); c8 = mean(x8);
   Cv = x1'*x8/n - c1*c8
   C = cov(x1,x8);
   ```
3. Nous allons maintenant traiter du tableau `X` dans sa globalité.
   a) centrez et réduisez les variable. Expliquez pourquoi cette étape est nécessaire avant d'effectuer une analyse en composantes principale (ACP)
   b) calculez la matrice de variance covariance des données. Commentez.
   c) calculez les composantes (ou facteurs ou directions) principales.
   d) projetez les individus sur les deux composantes les plus informatives (si vous voulez mettre les noms des céréales associées c'est stylé)
   e) la figure 2 donne la représentation des variables sur ces deux premières composantes. À la vue de cette figure, quelle interprétation donnez vous aux deux axes principaux. Quel est le rôle de la variable `Cups` ?

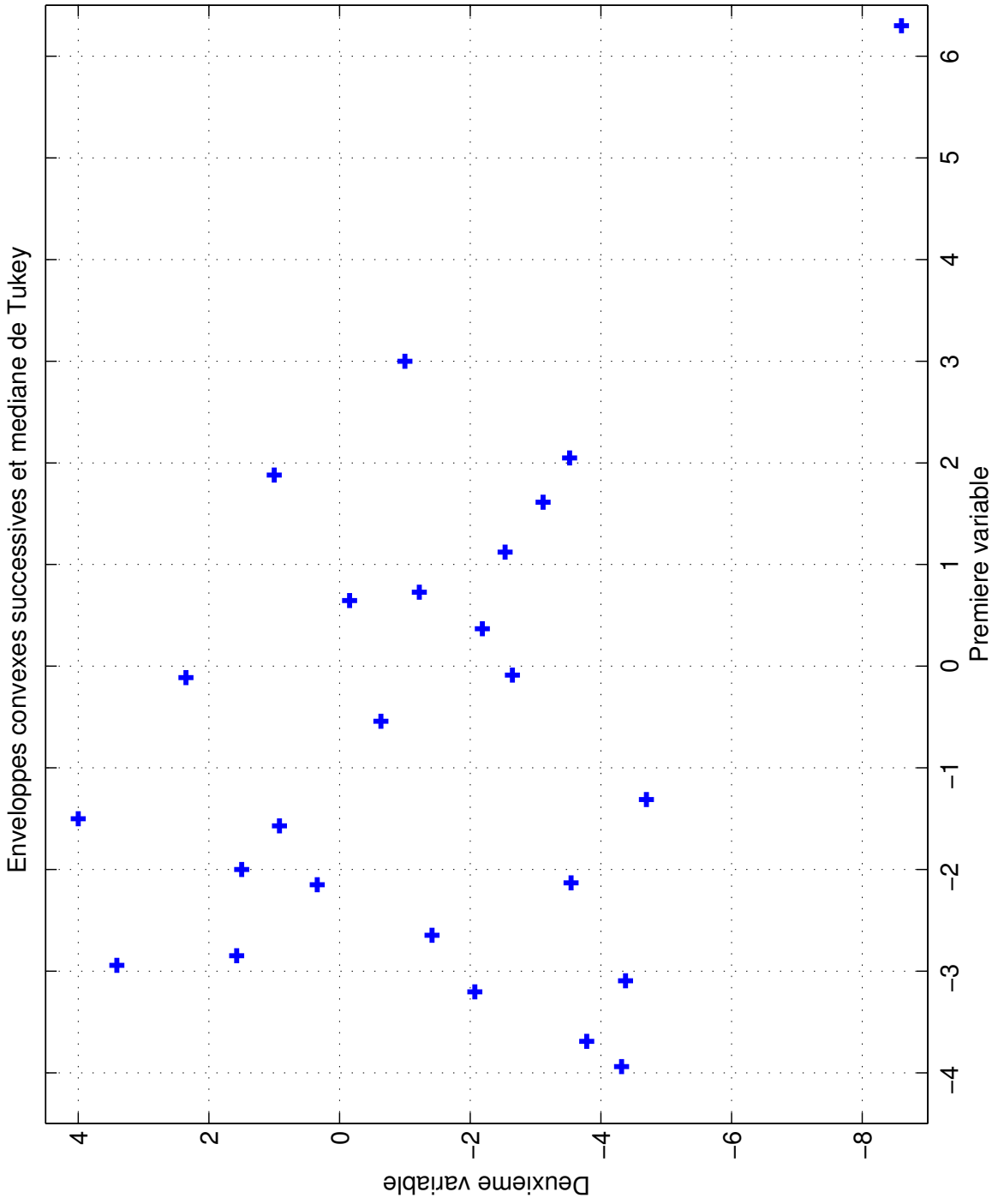


FIGURE 1 – Représentez graphiquement la médiane de Tukey et les enveloppes convexes successives.

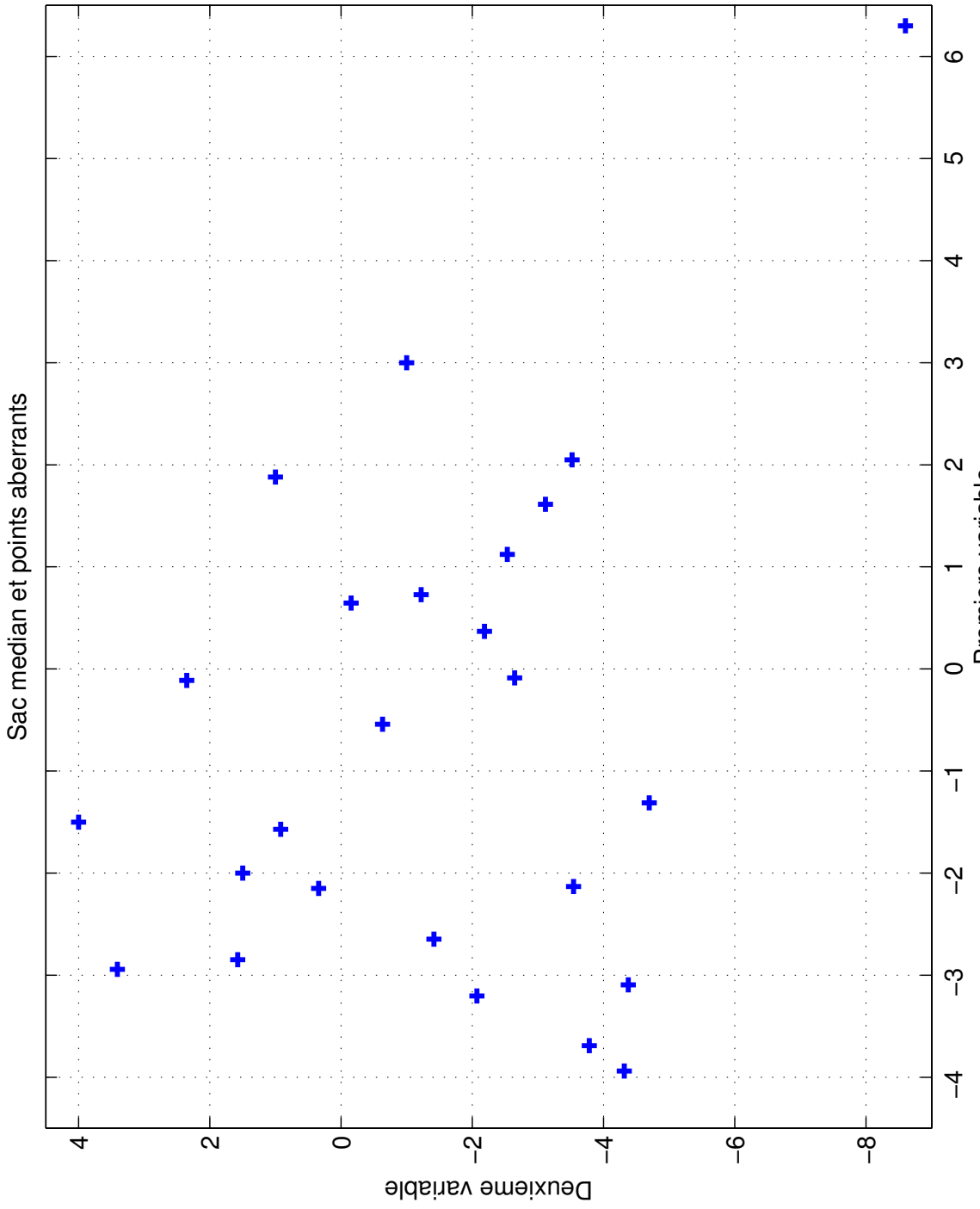


FIGURE 2 – Tracez le sac médian et repérez les éventuelles points aberrants sur cette figure.

```
X = [
   70.0000    5.0000    0.3300    1.0000   10.0000  280.0000    4.0000    6.0000
  120.0000    8.0000   -1.0000    5.0000    2.0000  135.0000    3.0000    8.0000
   70.0000    7.0000    0.3300    1.0000    9.0000  320.0000    4.0000    5.0000
   50.0000    8.0000    0.5000         0   14.0000  330.0000    4.0000         0
  110.0000   14.0000    0.7500    2.0000    1.0000   -1.0000    2.0000    8.0000
  110.0000   10.5000    0.7500    2.0000    1.5000   70.0000    2.0000   10.0000
  110.0000   11.0000    1.0000         0    1.0000   30.0000    2.0000   14.0000
  130.0000   18.0000    0.7500    2.0000    2.0000  100.0000    3.0000    8.0000
   90.0000   15.0000    0.6700    1.0000    4.0000  125.0000    2.0000    6.0000
   90.0000   13.0000    0.6700         0    5.0000  190.0000    3.0000    5.0000
  120.0000   12.0000    0.7500    2.0000         0   35.0000    1.0000   12.0000
  110.0000   17.0000    1.2500    2.0000    2.0000  105.0000    6.0000    1.0000
  115.0000   13.0000    0.7500    3.0000         0   45.0000    1.0000    9.0000
  110.0000   13.0000    0.5000    2.0000    2.0000  105.0000    3.0000    7.0000
  110.0000   12.0000    1.0000    1.0000         0   55.0000    1.0000   13.0000
  110.0000   22.0000    1.0000         0         0   25.0000    2.0000    3.0000
  100.0000   21.0000    1.0000         0    1.0000   35.0000    2.0000    2.0000
  110.0000   13.0000    1.0000         0    1.0000   20.0000    1.0000   12.0000
  110.0000   12.0000    1.0000    1.0000         0   65.0000    1.0000   13.0000
  110.0000   10.0000    0.5000    3.0000    4.0000  160.0000    3.0000    7.0000
  100.0000   21.0000    1.0000         0    1.0000   -1.0000    3.0000         0
  110.0000   21.0000    1.0000         0    1.0000   30.0000    2.0000    3.0000
  100.0000   11.0000    0.7500    1.0000    2.0000  120.0000    2.0000   10.0000
  100.0000   18.0000    0.7500         0    1.0000   80.0000    2.0000    5.0000
  110.0000   11.0000    1.0000    1.0000    1.0000   30.0000    2.0000   13.0000
  110.0000   14.0000    0.7500         0    1.0000   25.0000    1.0000   11.0000
  100.0000   14.0000    0.8000         0    3.0000  100.0000    3.0000    7.0000
  120.0000   12.0000    0.6700    2.0000    5.0000  200.0000    3.0000   10.0000
  120.0000   14.0000    0.6700         0    5.0000  190.0000    3.0000   12.0000
  110.0000   13.0000    0.7500    1.0000         0   25.0000    1.0000   12.0000
  100.0000   11.0000    0.8800         0         0   40.0000    2.0000   15.0000
  110.0000   15.0000    0.7500    1.0000         0   45.0000    1.0000    9.0000
  100.0000   15.0000    0.8800    1.0000    3.0000   85.0000    3.0000    5.0000
  110.0000   17.0000    0.2500         0    3.0000   90.0000    3.0000    3.0000
  120.0000   13.0000    0.3300    3.0000    3.0000  100.0000    3.0000    4.0000
  120.0000   12.0000    1.0000    2.0000    1.0000   45.0000    1.0000   11.0000
  110.0000   11.5000    0.7500    1.0000    1.5000   90.0000    3.0000   10.0000
  110.0000   14.0000    1.3300         0         0   35.0000    1.0000   11.0000
  110.0000   17.0000   -1.0000    1.0000    1.0000   60.0000    2.0000    6.0000
  140.0000   20.0000    0.7500    1.0000    2.0000   95.0000    3.0000    9.0000
  110.0000   21.0000    1.5000    1.0000         0   40.0000    2.0000    3.0000
  100.0000   12.0000    0.6700    2.0000    2.0000   95.0000    4.0000    6.0000
  110.0000   12.0000    1.0000    1.0000         0   55.0000    2.0000   12.0000
  100.0000   16.0000   -1.0000    1.0000         0   95.0000    4.0000    3.0000
  150.0000   16.0000   -1.0000    3.0000    3.0000  170.0000    4.0000   11.0000
  150.0000   16.0000   -1.0000    3.0000    3.0000  170.0000    4.0000   11.0000
  160.0000   17.0000    0.6700    2.0000    3.0000  160.0000    3.0000   13.0000
  100.0000   15.0000    1.0000    1.0000    2.0000   90.0000    2.0000    6.0000
  120.0000   15.0000    0.6700    1.0000         0   40.0000    2.0000    9.0000
  140.0000   21.0000    0.6700    2.0000    3.0000  130.0000    3.0000    7.0000
   90.0000   18.0000   -1.0000         0    3.0000   90.0000    3.0000    2.0000
  130.0000   13.5000    0.5000    2.0000    1.5000  120.0000    3.0000   10.0000
  120.0000   11.0000    0.6700    1.0000    6.0000  260.0000    3.0000   14.0000
  100.0000   20.0000    1.0000         0    1.0000   45.0000    3.0000    3.0000
   50.0000   13.0000    1.0000         0         0   15.0000    1.0000         0
   50.0000   10.0000   -1.0000         0    1.0000   50.0000    2.0000         0
  100.0000   14.0000    0.5000    1.0000    2.0000  110.0000    4.0000    6.0000
  100.0000   -1.0000    0.6700    2.0000    2.7000  110.0000    5.0000   -1.0000
  120.0000   14.0000    0.7500    1.0000    5.0000  240.0000    3.0000   12.0000
  100.0000   10.5000    0.5000    2.0000    2.5000  140.0000    3.0000    8.0000
   90.0000   15.0000    0.5000         0    2.0000  110.0000    2.0000    6.0000
  110.0000   23.0000    1.1300         0         0   30.0000    1.0000    2.0000
  110.0000   22.0000    1.0000         0         0   35.0000    2.0000    3.0000
   80.0000   16.0000   -1.0000         0    3.0000   95.0000    2.0000         0
   90.0000   19.0000    0.6700         0    4.0000  140.0000    3.0000         0
   90.0000   20.0000    0.6700         0    3.0000  120.0000    3.0000         0
  110.0000    9.0000    0.7500    1.0000    1.0000   40.0000    2.0000   15.0000
  110.0000   16.0000    1.0000         0    1.0000   55.0000    6.0000    3.0000
   90.0000   15.0000   -1.0000         0    3.0000   90.0000    2.0000    5.0000
  110.0000   21.0000    1.0000    1.0000         0   35.0000    2.0000    3.0000
  140.0000   15.0000    1.0000    1.0000    4.0000  230.0000    3.0000   14.0000
  100.0000   16.0000    1.0000    1.0000    3.0000  110.0000    3.0000    3.0000
  110.0000   21.0000    0.7500    1.0000         0   60.0000    2.0000    3.0000
  110.0000   13.0000    1.0000    1.0000         0   25.0000    1.0000   12.0000
  100.0000   17.0000    0.6700    1.0000    3.0000  115.0000    3.0000    3.0000
  100.0000   17.0000    1.0000    1.0000    3.0000  110.0000    3.0000    3.0000
  110.0000   16.0000    0.7500    1.0000    1.0000   60.0000    2.0000    8.0000 ]
```

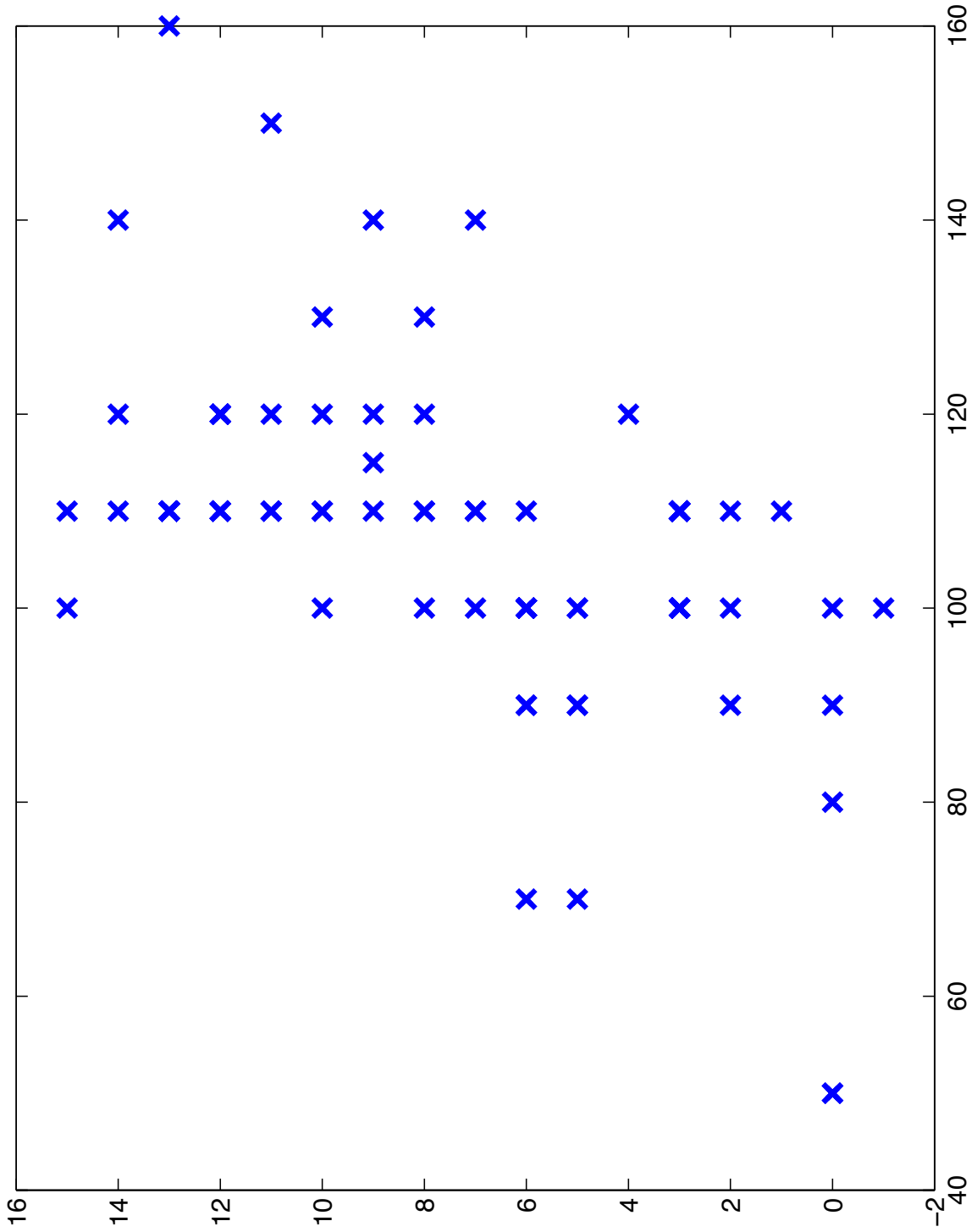


FIGURE 3 – Trouvez la médiane de Tuckey de ce nuage de points.

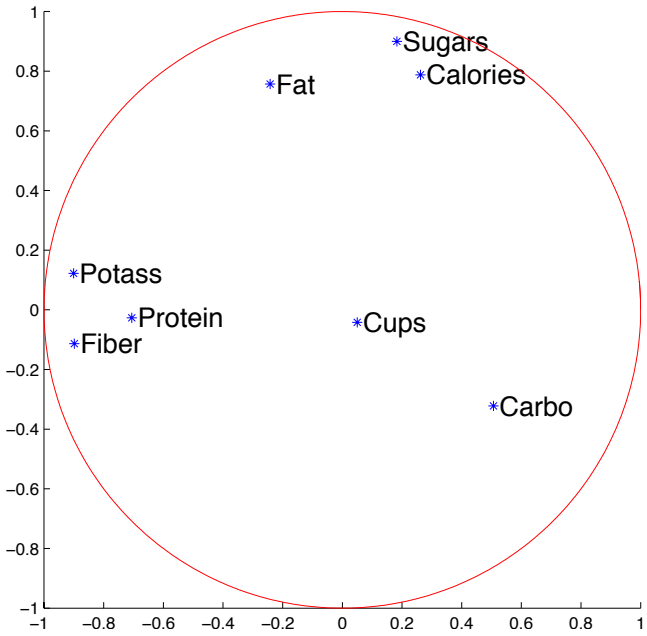


FIGURE 4 – Représentation des variables sur ces deux premières composantes de l'ACP du tableau X des céréales.

3 Régression linéaire

Exercice 24 — Calculs de base

Le but de cet exercice est de tester la réalité d'une relation linéaire entre le chiffre d'affaire et le nombre de salarié d'une entreprise. Les données mesurées sont les suivantes :

Année	*Nombre de salariés*	*Chiffre d'affaire*
1957	294	634
1959	314	728
1961	383	819
1963	402	938
1965	475	1136
1967	786	1317

1. représenter le nuage de points, poser le modèle et estimer les paramètres,
2. quelle est la qualité de ce modèle ? Y a-t'il vraiment une dépendance linéaire entre nombre de salariés et le chiffre d'affaire ?
3. estimer le chiffre d'affaire d'une entreprise l'an prochain si elle employe 800 salariés.
4. étudier les résidus et la contribution de chacun des points de mesure ? Si il se trouvait des points abérrants, les éliminer et recalculer les paramètres du modèle.
5. quel est le rôle du temps.

Exercice 25 — Calculs incroyables

Montrez que :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{c}(n\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i)$$

avec un c que l'on précisera.

Exercice 26 — Collage des modèles

On considere le modèle continu suivant :

$$y = \begin{cases} ax & si\ x < 1 \\ bx + c & si\ x \geq 1 \end{cases}$$

1. estimer a, b et c par la méthode des moindres carrés pour un échantillon $\{(x_i, y_i)\}$, $i = 1, ...n$. Privilégier l'optimalité par rapport aux pentes.
2. donner la solution matricielle de ce modèle, en définissant $\mathbf{y}, \mathbf{X}$ et $\boldsymbol{\alpha}$.

Exercice 27 — Régression Polynomiale

La régression linéaire étant un outil maîtrisé, nous proposons de traiter la régression polynomiale. Ainsi, y peut être exprimé sous la forme :

$$y = \sum_{k=0}^{p} c_k x^k \quad \text{avec : } \mathbf{c} = [c_p, \ldots, c_k, \ldots, c_0] \in \mathbb{R}^{p+1} \tag{1}$$

On considère que $\mathbf{c}$ est un vecteur colonne. Pour faire la régression, on a n couples (x_i, y_i) que l'on utilisera sous la forme de deux vecteurs colonne X et Y .

1. Donner le vecteur ligne $\mathbf{x}$ en fonction du réel x tel que $y = \mathbf{xc}$, en déduire la relation matricielle correspondant à l'équation 1 entre Y et et Z avec :

$$Z = \begin{bmatrix} x_1^p & \ldots & x_1 & 1 \\ \vdots & \ddots & & \vdots \\ x_N^p & \ldots & x_N & 1 \end{bmatrix} = [X^k\ X^{k-1}\ \ldots X\ \ 1] \tag{2}$$

2. En utilisant la formule de régression en matricielle du cours donner c en fonction de Y et Z.
3. Les données sont :

X	Y
1	2
2	5
3	10
4	17
5	26

FIGURE 5 – Données X et Y

a) Effectuer le régression linéaire entre X et Y et déterminer le coefficient de corrélation.
b) Effectuer la régression polynomiale (degré 2) entre X et Y et déterminer le coefficient de détermination. les coefficients c_k pourront être arrondis pour simplifier les calculs.
c) Combien de points faut il au minimum pour effectuer une régression polynomiale de degré p ?

Aide pour les calculs

$$\left(\begin{bmatrix} 979 & 225 & 55 \\ 225 & 55 & 15 \\ 55 & 15 & 5 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 0.0714 & -0.4286 & 0.5 \\ -0.4286 & 2.6714 & -3.3 \\ 0.5 & -3.3 & 4.6 \end{bmatrix}$$

Exercice 28 — données abérrantes (outliers)

L'échantillon suivant représente le poids et la taille de 13 étudiants tirés au hasard.

	Poids (Kg)	*Taille*
1	70	155
2	63	150
3	72	180
4	60	135
5	66	156
6	70	168
7	74	178
8	65	160
9	62	152
10	64	198
11	67	145
12	65	139
13	68	152

1. poser le modèle de régression linéaire de la taille en fonction du poids et estimer les paramètres,
2. étudier les résidus et la contribution de chacun des points de mesure ? S'il se trouvait des points abérrants, les éliminer et recalculer les paramètres du modèle.
3. estimer le poids moyen d'un étudiant mesurant 1 mètre 68 ? Quelle est la précision de cette prédiction ?

Exercice 29 — restons simple

Cherchant à expliquer la température d'un four (y) en fonction de l'hygrométrie (x) un boulanger a réalisé dix huit expériences différentes. Il a exécuté un programme de régression linéaire simple ($y = ax + b + \varepsilon$) sur ses données et il a trouvé une estimation de la pente de la droite de $\widehat{a} = 1.0355$, un coefficient $\widehat{b} = 0.2353$ et un coefficient de détermination $R^2 = 0.8534$. Le logiciel a aussi donné les résultats suivants :

Point	x	y	e	r	ei	c
1	-0.20	-0.04	-0.06	-0.28	-0.08	0.01
2	0.72	1.30	0.32	1.35	0.34	0.06
3	0.08	0.08	-0.24	-1.03	-0.27	0.05
4	1.39	1.37	-0.31	-1.45	-0.40	0.30
5	0.22	0.27	-0.18	-0.78	-0.20	0.02
6	0.87	1.36	0.22	0.95	0.24	0.04
7	1.38	1.44	-0.22	-1.03	-0.28	0.15
8	0.94	1.28	0.07	0.32	0.08	0.01
9	0.36	0.62	0.01	0.03	0.01	0.00
10	1.22	1.29	-0.21	-0.96	-0.26	0.09
11	0.74	1.41	0.41	1.73	0.44	0.11
12	0.13	0.12	-0.24	-1.04	-0.27	0.05
13	0.21	0.26	-0.19	-0.82	-0.21	0.03
14	0.56	1.16	0.35	1.46	0.37	0.06
15	-0.35	-0.04	0.09	0.42	0.11	0.02
16	-0.30	-0.16	-0.09	-0.40	-0.11	0.02
17	0.48	0.93	0.20	0.84	0.21	0.02
18	0.48	0.83	0.10	0.40	0.10	0.00

où e désigne l'erreur d'estimation (les résidus), r les résidus standardisés, ei les résidus studentisés et c les distance de Cook (encore appelées contributions).

1. A propos du modèle :
   a) quelle différence faites vous entre a et $\widehat{a}$?

b) Qu'est-ce que ε dans le modèle ?

2. Que pensez vous des résultats de cette régression linéaire ?

Exercice 30 — Passons aux choses sérieuses

Le voisin du boulanger, lui même boulanger à la retraite, explique que le modèle linéaire n'est pas le bon. Le bon modèle est le suivant :

$$y = \begin{cases} ax + bx^2 + \varepsilon & \text{si } x \leq 0,6 \\ c + \varepsilon & \text{si } x > 0,6 \end{cases}$$

1. En supposant que le modèle est continu, posez le problème sous la forme matricielle

$$\mathrm{y} = X\alpha + \epsilon$$

où X est une matrice et $\mathrm{y}, \alpha, \epsilon$ des vecteurs que l'on précisera.

2. A partir des données du tableau de l'exercice précédent (les x et les y) donnez une estimation des paramètres a, b et c.

indication : $$\begin{pmatrix} 3.7927 & 1.8979 \\ 1.8979 & 1.1579 \end{pmatrix}^{-1} = \begin{pmatrix} 1.4670 & -2.4046 \\ -2.4046 & 4.8052 \end{pmatrix}$$

Exercice 31 — Calculs non reinaux mais matriciels — 6 points

Soit W une matrice diagonale de taille n et de terme général w_i.

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & w_2 & \dots & 0 & \dots & \dots & 0 \\ \vdots & \dots & \ddots & \ddots & \ddots & \dots & \vdots \\ 0 & \dots & 0 & w_i & 0 & \dots & 0 \\ \vdots & \dots & \ddots & \ddots & \ddots & \dots & \vdots \\ 0 & \dots & \dots & 0 & \dots & w_{n-1} & 0 \\ 0 & \dots & \dots & 0 & \dots & 0 & w_n \end{pmatrix}$$

1. Donner la forme du terme général f_i du vecteur $\mathrm{f} = W\boldsymbol{e}$ en fonction des w_i et des e_i, où $\boldsymbol{e}$ est un vecteur de $\mathbb{R}^n$ de terme général e_i.
2. Donner également la forme générale de $\boldsymbol{e}^\top W \boldsymbol{e}$ en fonction des w_i et des e_i.
3. Pour des vecteurs $\mathrm{x} = (x_1, \dots, x_n)^\top$, $\mathrm{y} = (y_1, \dots, y_n)^\top$ et $\mathrm{w} = (w_1, \dots, w_n)^\top$ donnés, calculer la solution du problème suivant :

$$\min_{a,b} \; J(a,b) \qquad \text{avec } J(a,b) = \tfrac{1}{2} \sum_{i=1}^{n} w_i \big(y_i - (a + bx_i)\big)^2$$

4. Calcul matriciel :
   a) Écrire matriciellement $J(\alpha)$ avec $\alpha = (a,b)^\top$ et en fonction des vecteurs $\mathrm{x}, \mathrm{y}, \mathbb{1}$ et de la matrice W (où $\mathbb{1} = (1, \dots, 1)^\top$ est un vecteur de 1 de taille n).
   b) En déduire $\nabla_\alpha J$, le gradient de J par rapport à α.
   c) Donner l'expression de α optimal (solution du problème de minimisation) en fonction de x, y, $\mathbb{1}$ et W.
   d) Donner le code Matlab permettant de résoudre ce problème.

Exercice 32 — La régression mixte

Lors d'une série d'expériences on a mesuré le rendement de différentes parcelles de blé soumise à différentes quantités d'engrais pour deux variétés de blé, un blé OGM et un blé non OGM. Les résultats des expériences sont rapporté dans le tableau ci-contre.

y Rendement	x Engrais	v Variété
112	3.2	OGM
143	4.2	OGM
134	4.1	OGM
124	3.7	OGM
112	3.4	OGM
142	4.2	N
128	3.8	N
116	3.9	N
104	2.9	N
113	3.2	N

1. Poser le modèle linéaire du rendement comme une fonction des la quantité d'engrais et de la variété utilisée. Donner la matrice X du modèle associé.
2. Calculer les coefficients d'influence linéaire de chacune des composantes.

Exercice 33 — La régression pondérée

Afin d'étalonner un système permettant de calculer la turbidité (notée y) à partir du pH (notée x) une campagne de mesures est réalisée dans trois centres : Rouen, Rennes et Lyon. Mais le niveau d'équipement et de compétence de chacun des centres exige de traiter différemment les données venant de centres différents. Une manière de formaliser le problème consiste à associer à chaque couple de mesures (x_i, y_i) une pondération p_i traduisant la confiance relative que l'on souhaite accorder à cette mesure (qui est la confiance que l'on accorde au centre qui l'a produite). Le problème des moindres carrés se pose alors de la manière suivante :

$$\min_{a,b} \sum_{i=1}^{n} p_i \left(a x_i + b - y_i\right)^2$$

1. Trouvez les formules permettant de calculer les valeurs de a et b réalisant la minimisaiton des moindres carrés pénalisés.
2. le tableau suivant résume les mesures effectuées dans chacun des centres. Ainsi par exemple à Rouen on a réalisé 10 mesures pour lesquelles la somme des pH obtenu est de 5. Les mesures venant de Rouen ont une confiance de 3, celles de Rennes une confiance de 2 et celles de Lyon une confiance de 1.

	p_i	n_j	$\sum_{i=1}^{n_j} x_i$	$\sum_{i=1}^{n_j} y_i$	$\sum_{i=1}^{n_j} x_i^2$	$\sum_{i=1}^{n_j} x_i y_i$
centre de Lyon	1	$n_1 = 30$	16	2	16	6
centre de Rennes	2	$n_2 = 20$	12	2,5	6	5
centre de Rouen	3	$n_3 = 10$	5	1	4	3

   calculez les valeurs de a et b réalisant la minimisation du critère des moindres carrés à partir des données ci-dessus
3. quelle turbidité peut-on attendre d'une rivière lorsque son pH est de 7 ?

Exercice 34 — Croissance et Régression

D'après Monsieur Marcotte, professeur au laboratoire de Géophysique et Géostatistique de l'École Polytechnique de Montréal, il y a dans Excel une fonction « *Growth* » (ou croissance en français) qui permet d'estimer les paramètres c et d du modèle :

$$z = c\, d^x$$

1. indiquez comment « linéariser » ce modèle et ainsi obtenir des estimés pour c et d avec un programme de régression linéaire. Indiquez clairement le vecteur y et la matrice X qui seront soumis au programme de régression, les coefficients obtenus par la régression et le lien avec les coefficients recherchés.
2. les prédictions obtenues avec ce modèle minimiseront-elles la somme des carrés des erreurs $\sum_{i=1}^{n}(y_i - z_i)^2$. Justifiez votre réponse.

Exercice 35 — Fuite dans la pampers

Un climatologue nous a fourni les données suivantes de température et de concentration en ozone mesurées au centre de la ville de Caen, à la même heure, les lundi de printemps 2007.

x : Température	y : Ozone
23.1	81.4
18.9	52.6
16.4	48.4
14.6	38.0
20.3	39.4
25.4	34.2
21.5	61.9
22.1	77.3
17.3	34.0
14.8	49.8
25.5	97.0
23.1	67.8
24.2	92.7

1. dessinez la fonction de répartition empirique de la variable Ozone,
2. donnez la boite à moustache correspondant à la variable Ozone,
3. représentez graphiquement la médiane de Tukey et le sac médian, y a-t'il un point aberrant ?
4. calculez la covariance et le coefficient de corrélation du nuage de points,
5. pour expliquer la concentration d'ozone en fonction de la température, les climatologues proposent le modèle suivant :

$$y = \begin{cases} b & \text{si } x < c \\ ax + d & \text{si } x \geq c \end{cases}$$

   a) réécrire le modèle en éliminant d en utilisant la contrainte de continuité du modèle
   b) donnez l'estimateur de a et b au sens des moindres carrés, sachant que $c = 21$
   c) évaluez la contribution de chacun des points à la régression
   d) (subsidiaire ++) si maintenant on suppose que l'on ignore la valeur de c, proposez une méthode permettant de l'estimer conjointement avec a et b.

Exercice 36 — La régression multiple

1. chargez le tableau `X` présente les données d'un fichier `cereal.mat`, présent sur votre bureau, contenant les mesures des ingrédients de 77 produits de céréales pour petit déjeuner. Sur chaque type de céréale, 8 variables ont été mesurées. Ces variables sont (dans l'ordre) :

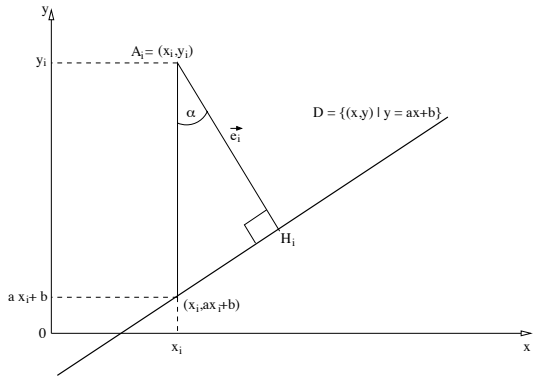
Calories	la teneur en calories (le nombre pour un repas)
Carbo	la teneur en carbohydrates (en gramme pour un repas)
Cups	le nombre de tasse recommandé par repas
Fat	la teneur en graisse (en gramme pour un repas)
Fiber	la teneur en fibres alimentaires (en gramme pour un repas)
Potass	la teneur en potassium (en mg pour un repas)
Protein	la teneur en proteine (en gramme pour un repas)
Sugars	la teneur en sucre (en gramme pour un repas)

   A chaque type de céréale a été associé son nom contenu dans la variable `Name`. Ainsi la première céréale est `Name(1) -> '100% Bran'`.
2. effectuer la régression de la variable FAT (la quatrième) en fonction de toutes les autres. S'il y en a, éliminez les observation aberrante et calculez une estimation de σ^2
3. quelles est la variable la plus explicative de FAT ?
4. en utilisant la méthode du C_p, sélectionnez les variables les plus explicative de la variable FAT.

Exercice 37 La régression orthogonale

Etant donné le nuage de points en deux dimensions $A_i = (x_i, y_i)$ nous nous proposons de déterminer la droite (D) telle que la somme des carrées des distances $(A_i H_i)$ soit minimale, H_i étant la projection orthogonale de A_i sur (D). Soit à rendre minimale la quantité :

$$d^2 = \sum_{i=1}^{n} \|\vec{e_i}\|^2$$



1. (a) Montrez que la norme du vecteur $\vec{e_i}$ s'écrit $\|\vec{e_i}\| = k(y_i - ax_i - b)$ où k est une constante que l'on précisera
   (b) Montrez que la somme des carrés des erreurs peut se décomposer comme la somme de deux termes $d^2 = \sum_{i=1}^{n} \|\vec{e_i}\|^2 = B^2 + C^2$ dont l'un ne depend des points (x_i, y_i) qu'à travers $\bar{x}$ et $\bar{y}$ les moyennes des x_i et des y_i ($\bar{x} = \sum_{i=1}^{n} x_i, \bar{y} = \sum_{i=1}^{n} y_i$)
   (c) En déduire que la droite (D) passe par le point moyen $(\bar{x}, \bar{y})$
2. Analyse géométrique. Soit z_i le point de coordonnées $(x_i - \bar{x}, y_i - \bar{y})$.
   (a) Reformulez le problème après avoir translaté l'origine ou point $(\bar{x}, \bar{y})$ (on note que dans ce cas la solution ne dépend plus de b. Pour a fixé, b ne dépend que du couple $(\bar{x}, \bar{y})$).
   (b) On considère maintenant que (Δ) est la droite vectorielle définie par $\Delta = \{(x, y) | \alpha x + \beta y = 0\}$ avec $\alpha^2 + \beta^2 = 1$. Montrez que tous les points de (Δ) sont othogonaux au vecteur normé $w = (\alpha, \beta)$ et colinéaire au vecteur $u = (\beta, -\alpha)$. Montrez que dans ces nouvelles coordonées on a :

$$\vec{\varepsilon_i} = (\alpha x_i + \beta y_i)\vec{w} = (\alpha x_i + \beta y_i) \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

   (c) En déduire que

$$B^2 = \sum_{i=1}^{n} \|\vec{\varepsilon_i}\|^2 = \alpha^2 V(x) + \beta^2 V(y) + 2\alpha\beta C(xy)$$

   où $V(x), V(y)$ et $C(x, y)$ sont trois termes que l'on précisera.
   (d) montrez que le minimum de B^2 est donnée par :

$$\begin{pmatrix} \alpha = \cos t \\ \beta = \sin t \end{pmatrix} \quad \text{avec} \quad \tan 2t = \frac{2C(xy)}{V(x) - V(y)}$$

   (e) Donnez une interprétation géométrique du résultat.
3. Application numérique

X	0	0,5	0,75	1,5	2	2	2,5	2,75	3	3	3,5	3,5	4	4	4,5
Y	1,5	0,5	0	0,5	0,5	2	3	3,5	3	4,5	5	6	8	8,5	9,5

4. Analyse matricielle du problème :
   (a) Le vecteur $\vec{\varepsilon_i}$ est aussi le résultat de la projection de $\vec{z_i}$ sur la droite vectorielle engendrée par le vecteur $\vec{w}$. On a donc :

$$\vec{\varepsilon_i} = k\vec{w}$$

   Montrez que $k = \frac{\vec{z_i}^\top \vec{w}}{\vec{w}^\top \vec{w}}$. On utilisera le fait que la projection est aussi le vecteur qui réalise le minimum de la distance entre $\vec{z_i}$ et la droite vectorielle.

(b) montrez que

$$\|\vec{\varepsilon_i}\|^2 = \frac{\vec{w}^\top \vec{z_i} \vec{z_i}^\top \vec{w}}{\vec{w}^\top \vec{w}}$$

(c) En déduire que

$$B^2 = \sum_{i=1}^{n} \|\vec{\varepsilon_i}\|^2 = \frac{\vec{w}^\top M \vec{w}}{\vec{w}^\top \vec{w}}$$

où M est une matrice que l'on précisera

(d) Montrez que le minimum de B^2 est donnée par la solution du problème de minimisation sous contraintes suivant :

$$\begin{cases} \min\limits_{w} \ \vec{w}^\top M \vec{w} \\ \text{sous la contrainte} \quad \|\vec{w}\|^2 = 1 \end{cases}$$

Ecrire le Lagrangien associé à ce problème de minimisation sous contraintes

(e) Montrez que les solutions du problème de minimisation sous contrainte précédent sont les vecteurs propres d'une matrice que l'on précisera. Donnez une interprétation géométrique du résultat.

Indications (rappels)
— $\cos\alpha = \frac{1}{1+a^2}$
— $\cos^2 t = \frac{1+cos2t}{2}$

Exercice 38 — Chose promise...

voici le début d'une session Matlab :

```
[n,m] = size(v);
           n = 100
           m = 1
[n,m] = size(u);
          n = 100
           m = 1
[n,m] = size(x);
           n = 100
            m = 1
```

1. pour chacune des instructions suivante indiquez :
   a) la signification de l'instruction (en une phrase),
   b) la nature du résultat et sa taille.

```
M = [ones(n,1) u v];
[n,p] = size(M);
y = (M'*M)\(M'*x);
a = M*y
e = x-a
H = M*inv(M'*M)*M'
H*H
sum(e)
```

4 Test statistiques

4.1 Test du Chi2

Exercice 39 — Faut toujours mettre sa ceinture

Une étude a été menée en 1990-91 sur les facteurs pouvant influer sur le port de la ceinture de sécurité par les conducteurs et les passagers de voitures de tourisme et de véhicules utilitaires

On s'intéresse tout d'abord à l'effet du type d'occupation du véhicule (conducteur seul, conducteur + passagers avant, conducteur + passagers arrière, conducteur + passagers avant et arrière) sur le port de la ceinture par le conducteur. On dispose de 8374 observations concernant cette partie de l'étude. Les données sont les suivantes :

	Port ceinture	non port de ceinture
Seul	2825	3468
Cond. + pass. avant	729	815
Cond. + pass. arrière	80	113
Cond. + pass. av. et arr.	168	176

1. Tester l'existence d'un lien entre les deux variables « Type d'occupation » et « Port de la ceinture » avec un risque de première espèce de 0,05

Exercice 40 — Sur le χ^2

1. Lors d'un test χ^2, on a trouvé $D(O,T) = 0$. Quelles conclusions peut-on en tirer :
   — ce résultat est impossible, car D ne peut prendre que des valeurs positives.
   — on garde l'hypothèse Nulle, mais cela n'est pas concluant ;
   — on en déduit que l'hypothèse Nulle est sans doute la correcte.
   — on en déduit que l'hypothèse alternative est sans doute la correcte.
   — on ne peut rien décider car la décision de garder l'hypothèse nulle ou l'hypothèse alternative dépend de la valeur de D mais aussi de la valeur de n le nombre total d'observations.
2. Même question pour $D(O,T) = 0,01$ et $D(O,T) = 10$.

Exercice 41 — Qui sera carré et bleu ?

D'après la loi de Mendel sur les gènes dominants et récessifs, si l'on tire un pois (le légume) au hasard on a trois chances sur quatre d'observer un pois rond contre une chance sur quatre pour qu'il ne le soit pas (on dit alors qu'il est anguleux). De même, on a trois chances sur quatre d'observer un pois jaune contre une chance sur quatre pour qu'il soit vert.

1. montrez, en admettant l'indépendance des deux caractères, que la probabilité de tirer un pois rond et jaune est de 9/16, celle de tirer un poids rond et vert égale à celle de tirer un pois anguleux et jaune est de 3/16.
2. dans mon jardin j'ai observé 556 pois dont 315 ronds et jaune, 108 ronds et vert, 101 anguleux et jaune et 32 anguleux et vert. Dois-je en conclure que Mendel à raison ?
3. un généticien souhaite savoir s'il y a oui ou non un lien entre deux gènes. Pour ce faire il décide d'observer les deux gènes sur 1000 individus tiré au hasard. Le tableau suivant résume les résultats obtenus.

Genotype	BB	Bb	bb
AA	57	140	101
Aa	39	223	225
aa	5	54	156

   Peut-on décider que les gènes sont indépendants ?
4. ce même généticien décide de tester au hasard 1.000.000 de relations de ce type avec des gènes. En imaginant qu'il n'y ai aucune relation entre les gènes testés, en appliquant la stratégie vue en cours, combien de fois se sera t'il trompé (en moyenne) ?

Exercice 42 — Des grands et des petits

On aimerait savoir si il existe un lien entre la taille des gens et leur lieu de naissance (Nord ou Sud de la France). Pour cela, on a mesuré 15 personnes nées dans le nord et 15 personnes nées dans le sud, voici leur taille :

Sud	181.6 176.3 170.8 173.5 163.0 187.0 170.6 188.0 172.6 178.7 155.5 163.0 182.5 163.6 175.8
Nord	176.4 178.6 166.5 167.3 189.8 179.6 172.0 172.3 188.6 179.4 185.1 184.0 187.6 184.0 166.6

On divisera la population en trois types :
- Grand : 180 cm $\leq$ taille
- Moyen : 170 cm$\leq$ taille <180 cm
- Petit : taille < 170 cm

1. Faire le tableau de contingence
2. Y-a-t-il un lien entre la taille et le lieu de naissance ?

Exercice 43 — Attention, il faut les bonnes données !

Pendant toute une années monsieur Carlos a noté les prévisions météo et a observé le temps qu'il faisait le lendemain. Le tableau suivant rapporte ses chiffres :

Observé \ Prédit	Pluie	Soleil
Pluie	80	40
Soleil	15	230

1. A partir du tableau ci-dessus, peut on affirmer que les observations du lendemain sont indépendantes des prévisions ?
2. Monsieur Carlos prétend que la météo se trompe une fois par semaine. Peut on accepter l'hypothèse de Monsieur Carlos ? On supposera que si la probabilité de se tromper est de p, la fréquence observée des erreurs est distribuée statistiquement comme une loi normale d'espérance p et de variance $\frac{p(1-p)}{n}$.

Exercice 44 — Fish distribution

Pour des grandes valeur de λ (disons $\lambda \geq 1000$), la loi de poisson peut être approchée par une loi normale d'espérance λ et de variance λ (et oui le même et unique λ).

1. donnez un exemple d'une variable aléatoire suivant une loi de Poisson.
2. si W est une variable aléatoire suivant une loi de poisson de paramètre λ que l'on sait supérieure à 1000. Proposez une approximation de la loi des variables aléatoires X, Y et Z suivantes

$$X = \frac{W - \lambda}{\lambda} \qquad Y = \frac{W - \lambda}{\sqrt{\lambda}} \qquad Z = \frac{\overline{W} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

   avec $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$ la moyenne empirique de n réalisations de cette variable aléatoire.
3. avec ces hypothèses, quelle serait la loi des variables aléatoires suivantes :

$$A = \frac{1}{\lambda}\sum_{i=1}^{n}(W_i - \lambda)^2 \qquad B = \frac{1}{\lambda}\sum_{i=1}^{n}(W_i - \overline{W})^2$$

Exercice 45 — Roselyne et les lions

Les résultats observés de l'évolution d'une bilardose congrogène (une certaine maladie) à la suite de l'emploi de l'un ou l'autre des traitements daloricine (D) et mitrovial (M) figurent dans le tableau ci-dessous en pourcentage :

	guérison	amélioration	état stationnaire
daloricine (D)	36 %	16 %	10 %
mitrovial (M)	20 %	8 %	10 %

1. L'étude ayant portée sur 150 cas, peut-on en conclure que les traitements D et M sont significativement différents quant à leur efficacité ?
2. quelle serait la réponse si les données n'avaient portée que sur 1500 cas (dix fois plus) ?
3. selon vous, est-il exact de dire que la probabilité de guérir quelque soit le traitement est de 0,6 ?

Exercice 46 — le temps ne fait rien à l'affaire

A la suite d'un traitement, pour 70 patients jeunes on a observé 40 cas d'amélioration alors que pour 100 patients âgés on a en a observé 50. Peut on considérer qu'il y a un lien significatif entre l'âge du patient et l'effet du traitement ?

Exercice 47 — le Kid vampire

Les centre de transfusion sanguine diffusent le tableau de contingence en pourcentage qui donne la répartition en France des principaux groupes sanguin :

Groupe	0	A	B	AB
R+	37,01	38,09	6,20	2,80
R-	7,02	7,18	1,20	0,50

1. selon vous y a t'il dépendance entre le groupe sanguin et le facteur rhésus dans la population ? (Il y a eu 480 000 donneurs)
2. selon vous , est-il exact de dire que la distribution de probabilité des groupes sanguins est la suivante : $\mathbb{P}(G = 0) = \mathbb{P}(G = A) = 0,448$, $\mathbb{P}(G = B) = 0,072$ et $\mathbb{P}(G = AB) = 0,032$
3. quelle serait la réponse si les données n'avaient portée que sur 48 000 donneurs (dix fois moins) ?

Exercice 48 — C'est comme la divination

On propose un questionnaire comprenant 10 questions avec chacune deux réponses possibles, l'une vraie et l'autre fausse. Pour tester si une personne interrogée répond correctement on adopte la règle suivante : si 7 réponses ou plus sont correctes on admet que la personne n'a pas répondu au hasard.

1. Donner les deux hypothèses du tests
2. Quelle est la p-valeur du test d'une personne qui a donné 8 réponses correctes ?
3. Quelle est la probabilité de décider qu'une personne à répondu correctement alors qu'elle a répondu au hasard ?
4. Que devient cette probabilité lorsque chacune des questions posées comporte trois réponses dont une seule est vraie ?

4.2 Test de student

Exercice 49 QCM

1. La distribution des moyennes d'échantillons indépendants suit une loi normale
   Quelles sont la(es) affirmation(s) exacte(s) ?
   a) Quelque soit la distribution de la variable si l'effectif des échantillons est de grande taille
   b) Uniquement si la distribution de la variable suit une loi normale
   c) Si la variable suit une loi log normale et que l'effectif des échantillons est de grande taille
   d) Si l'effectif des échantillons est de petite taille et que la distribution de la variable suit une loi normale
   e) On doit réaliser un test de conformité de la distribution des moyennes à la loi normale pour savoir si la distribution des moyennes d'échantillons indépendants suit une loi normale

2. Dans un test statistique classique, on formule des hypothèses :
   Quelles sont la(es) affirmation(s) exacte(s) ?
   a) L'hypothèse nulle est formulée dans le but de la rejeter
   b) Les hypothèses alternatives sont formulées dans le but de les accepter
   c) L'hypothèse nulle est formulée dans le but de l'accepter
   d) Les hypothèses alternatives sont formulées dans le but de les rejeter
   e) La prise en compte a priori d'une partie seulement des hypothèses alternatives permet d'envisager des tests unilatérau

3. Dans un test statistique, Le risque de première espèce (risque alpha) :
   Quelles sont la(es) affirmation(s) exacte(s) ?
   a) Est le risque de rejeter l'hypothèse nulle alors qu'elle est vrai
   b) Correspond, lors d'un essai thérapeutique, dans un test sur le critère d'efficacité d'un médicament, au risque de conclure que le médicament est efficace alors qu'il ne l'est pas
   c) Correspond, lors d'un essai thérapeutique, dans un test sur le critère de tolérance clinique ou biologique d'un médicament au risque de conclure que le médicament a une bonne tolérance alors que celle-ci est mauvaise
   d) Permet de définir la puissance d'un test : puissance = 1 - alpha
   e) Est toujours connu

4. Dans un test statistique, Le risque de deuxième espèce (risque beta) :
   Quelles sont la(es) affirmation(s) exacte(s) ?
   a) Est le risque de rejeter l'hypothèse nulle alors qu'elle est vrai
   b) Correspond, lors d'un essai thérapeutique, dans un test sur le critère d'efficacité d'un médicament, au risque de conclure que le médicament est efficace alors qu'il ne l'est pas
   c) Correspond, lors d'un essai thérapeutique, dans un test sur le critère de tolérance clinique ou biologique d'un médicament au risque de conclure que le médicament a une bonne tolérance alors que celle-ci est mauvaise
   d) Est le risque d'accepter l'hypothèse nulle alors qu'elle est fausse
   e) Est toujours maîtrisé

Exercice 50 Statistiques propres

Il est connu (par certains) que les frais annuels des familles mexicaines en lessive pour les machines à laver est une variable aléatoire qui suit une loi normale de variance connue et égale à 100 pesos. Le directeur de marketing de « jabon pueblo » (une célèbre marque de détergents) souhaite augmenter la consommation de lessive en faisant une campagne publicitaire annuelle. Cette campagne sera réalisée si les dépenses moyennes par famille sont égales ou supérieures à 60 pesos. Sur un échantillon de 625 familles, on a observé une dépense moyenne de 60,9 pesos par année.

1. Formulez les deux hypothèses du test que vous souhaitez mettre en œuvre.
2. Donnez la p-valeur du test et le seuil de décision du test.
3. Déterminez si la campagne publicitaire doit être réalisée ou non, avec un risque de première espèce de 0,05.

4. Quels changements doit-on faire si la variance est inconnue ?

Exercice 51 — Ogé aime mon santo

Afin de tester la toxicité d'une variété de maïs OGM, un groupe de 10 rats a été nourri avec ce maïs. Un groupe témoin de 12 rats a été nourri avec du maïs non-OGM. Après 90 jours, on mesure le poids du foie de chaque rat. Les résultats ont été les suivants : sur le groupe OGM on a trouvé un poids moyen de 16,9 grammes avec un écart type de 1,21 gramme sur le groupe nonOGM on a trouvé un poids moyen de 16,2 grammes avec un écart type de 1,1025 gramme

Peut on en conclure que la consommation d'OGM modifie le poids du foie ?

Exercice 52 — L'étudiante va au bois

Dans deux types de forêts distincts, on a mesuré en mêtre les hauteurs dominantes respectivement de 13 et 14 peuplements du même age choisis au hasard et indépendamment.

x : type 1	y : type 2
23,4	22,5
24,4	22,9
24,6	23,7
24,9	24,0
25,0	24,4
26,2	24,5
26,3	25,3
26,8	26,0
26,8	26,2
26,9	26,4
27,0	26,7
27,6	26,9
27,7	27,4
28,5	

$$\sum_{i=1}^{n_1} x_i = 366,1$$
$$\sum_{i=1}^{n_2} y_i = 326,9$$
$$\sum_{i=1}^{n_1} x_i^2 = 9602$$
$$\sum_{i=1}^{n_2} y_i^2 = 8251$$

En admettant un risque de première espèce de 0,05,

1. les deux forêts sont-elles identiques si l'on suppose que la variance des hauteurs mesurées est connue ($\sigma^2 = 1,7$) ?

2. les deux forêts sont-elles toujours identiques si l'on suppose que la variance des hauteurs mesurées est inconnue ?

Exercice 53 — Les lectures de l'étudiant

Dans une étude sur la motivation et l'apprentissage, on demande à des élèves de quatrième de lire un document scientifique. Les sujets sont répartis en deux groupes distincts. Dans l'un des deux groupe (témoin), les titres de section du document sont inintéressants, tandis que dans l'autre groupe (groupe test), ils sont accrocheurs. Seuls les titres changent d'un groupe à l'autre, et ils sont choisis de la même longueur. On relève ensuite le temps de lecture par un score T . On trouve les valeurs suivantes :

témoin	test	témoin	test
4	5	8	6
5	5	9	5
6	4	8	4
7	5	12	4
6	5	11	3
7	6	10	1
6	5	9	2
5	6	8	12
2	7	8	4
6	6	9	3

En admettant un risque de première espèce de 0,05 et la normalité des distributions, peut on dire qu'il y a une différence entre les deux groupes ?

Exercice 54 La paire fait l'affaire

Nous avons mesuré les performances en terme de temps de réaction de sujets avant et après un entraînement extrêmement sévère.

Personne	1	2	3	4	5	6	7	8	9	10
Avant	14	9	23	15	19	38	43	29	15	17
Après	10	11	20	10	15	30	32	30	16	17

1. A la vue de ce tableau, peut on en déduire que l'entraînement est efficace ?
2. Refaites ce calcul en considérant maintenant que les mesures « avant et après » s'effectuent sur des personnes différentes, tirées indépendamment au hasard.
3. Donnez les p-valeurs de l'échantillon dans les deux cas

Exercice 55 Un bon bol d'air pour l'étudiant

La capacité respiratoire d'une personne est une variable aléatoire X, distribuée suivant une loi normale de paramètres inconnus. On tire deux échantillons i.i.d. de taille 10 de cette loi. Le premier groupe de mesures a été effectué avant d'avoir soumis 10 individus à un traitement dont on cherche à démontrer l'efficacité. Le second groupe de mesures a été pris après avoir effectué le traitement. On cherche à savoir si ce traitement est efficace ou non.

1. proposez une stratégie de décision.
2. après avoir observé les résultats suivants, et pour un risque de première espèce de 5%, que décidez vous ?

individu	1	2	3	4	5	6	7	8	9	10
groupe 1	102	84	91	72	93	91	135	115	101	94
groupe 2	102	115	105	107	101	114	111	120	102	101

3. un expert du domaine vous certifie que la variance de la mesure est de $\sigma^2 = 111$. En quoi cela change t'il votre raisonnement ?

Exercice 56 avant/après

Afin d'évaluer un programme de réhabilitation, on soumet les douze premiers bénéficiaires à une même série de tests avant de rentrer dans le programme puis six mois plus tard à leur sortie du programme. Ces tests visent à évaluer la capacité à réaliser des tâches comme emballer des paquets, laver des assiettes... Les scores sont des notes qui vont de 0 (ne sait rien faire) jusqu'à 40 (travail excellent). Le tableau suivant récapitule les résultats obtenus :

bénéficiaire	score avant	score après
1	11	13
2	6	10
3	5	7
4	35	37
5	22	21
6	6	8
7	34	35
8	25	24
9	14	18
10	39	45
11	25	23
12	16	19

1. A votre avis, ce programme améliore t'il les capacités des bénéficiaires ?
2. Ce programme améliore t'il mieux les capacités des bénéficiaires ayant un score initial inférieur à 20 ?

Exercice 57 — Les deux font la paire

Est-ce-que la concentration de cholestérol, triglycérides et d'autres substances se modifie si des échantillons de sang sont conservés pendant un certain temps ? Évidemment, la réponse à cette question est une information importante pour l'organisation du travail de laboratoire. Dans une étude publiée, les échantillons de sang de 10 sujets d'une certaine population ont été analysés immédiatement après la prise de sang et 8 mois après.

Avant :	74	80	75	136	104	102	90	100	95	84
Après :	66	85	71	132	104	105	89	102	101	84

On se demande si les deux mesures de chaque échantillon sont suffisamment éloignées pour qu'on puisse décider qu'il y a un effet de la conservation. Préciser l'ensemble des choix que vous êtes amenés à faire.

Exercice 58 — bras droit – bras gauche

Afin de tester un traitement on soumet le bras gauche de quinze patients à un traitement expérimental. A l'issue de ce traitement on mesure pour chaque patient la même quantité sur son bras droit et sur son bras gauche. Le tableau suivant récapitule les résultats obtenus :

sujet	bras droit	bras gauche
1	8,97	6,58
2	5,90	4,53
3	4,43	3,86
4	1,67	3,00
5	6,84	4,32
6	2,98	2,91
7	12,98	14,88
8	11,22	9,76
9	2,60	11,68
10	6,98	1,19
11	9,01	12,07
12	14,99	13,92
13	5,72	4,54
14	15,39	11,18
15	14,37	13,10

1. A votre avis, ce traitement a t'il un effet ?
2. Si l'on admet l'hypothèse que le traitement n'a pas d'effet, quelle est la probabilité d'observer que six fois sur quinze la mesure sur le bras gauche est inférieure à celle du bras droit. A partir de ce raisonnement, vous parait t'il raisonnable de dire que le traitement à un effet (justifiez).
3. Quelle est la p valeur si l'on admet que le traitement diminue la mesure de 0,5 (on parle d'espérance).

Exercice 59 — Test de base

Le but de cet exercice est de tester la réalité d'une relation linéaire entre les variables x et y puis entre x et z :

x	y
1.00	1.14
2.00	1.24
3.00	1.09
4.00	1.25
5.00	1.71
6.00	1.57

1. Estimez les coefficients de la régression $y = ax + b$
2. Estimez la variance du bruit
3. Peut-on faire l'hypothèse que a=0 :
   a) Si vous connaissez la variance de bruit $(0, 03)$
   b) Si vous ne connaissez pas la variance du bruit

Exercice 60 — Modèle Optimiste

A partir d'un échantillon $\{(x_i, y_i)\}_{i=1,\dots,n}$. On considère le modèle suivant :

$$y_i = a|x_i| + b + \xi_i \qquad i = 1, n$$

où ξ_i est un bruit représenté par une variable aléatoire d'espérance nulle, et a et b deux paramètres inconnus.

1. Construire l'estimateur de a et b au sens des moindres carrés.
2. Calculez l'estimation de a et b à partir des données suivantes :
   on commencera par représenter le nuage de points.

x	y
-37	82
-83	163
29	52
-182	370
-157	304
202	411
-7	8
262	530
-24	55
17	25

3. Est-il raisonnable de considérer que $a = 2, 4$ (avec une erreur de premier ordre de 5 %).

5 Travaux pratiques

Exercice 61 — Mise en bouche

Ce TP vise à vous familiariser avec des données réelle et à l'évolution des caractéristiques d'un signal au cour du temps.

1. Prise en main des outils
   Ce TP vise à vous familiariser avec les outils et les données que vous devrez utiliser lors de la rédaction de votre rapport M8. Il vous sera en effet demandé à la fin du semestre de rendre un rapport où vous utiliserez les notions vues en cours.
   a) Moodle
      Les notes de cours ainsi que des sujet d'examen sont disponible sur Moodle (`https://moodle.insa-rouen.fr/course/view.php?id=169`).
      Les notes de cours seront mises à jour au fur et à mesure de l'année, faites attention à avoir la bonne version.
   b) Matlab
      Nous utiliserons le logiciel Matlab pour l'étude des données. Un tutoriel sur ce logiciel est disponible sous moodle (`https://moodle.insa-rouen.fr/course/view.php?id=154`).
2. Etude de données
   a) Chargement des données
      Nous allons travailler sur des données de mesure de pression sanguine disponibles à l'adresse : `http://www.stat.ucla.edu/projects/datasets/`.
      — Télécharger le fichier "cardiac.dat", les informations concernant ce fichier sont disponible (fichier html "cardiac-explanation")
      — Charger ce fichier sous matlab (en utilisant l'interface graphique ou la fonction *csvread*.
      — Extraire les données de pression sanguine (deuxième colonne) et les visualiser en texte.
   b) Fonction de répartition empirique
      Nous allons estimer la fonction de répartition réelle en utilisant les données. Pour cela nous utiliserons la fonction *sort* ainsi que la fonction *plot*.
      — calculer la probabilité empirique pour qu'un sujet ait moins de 120 (utiliser les fonctions *find* et *length*),
      — calculer la probabilité pour qu'un sujet ait exactement 120,
      — dessiner la fonction de répartition empirique associée aux données
      — dessiner l'histogramme brut des valeurs
      — la variable considérée est elle discrète ou continue ?
   c) Histogramme
      Une bonne visualisation de la répartition des données est l'histogramme. Coder une boucle qui permet d'obtenir le nombre d'échantillon compris dans des intervalles réguliers (une dizaine par exemple). Pour cela on utilisera la fonction *find*, *length* et *bar*.
      A partir de ce nouveau codage :
      — dessiner la nouvelle fonction de répartition empirique
      — dessiner l'histogramme associé
      — la variable considérée est elle discrète ou continue ?

Exercice 62 — Analyse en Composantes principales

Dans ce TP, il vous est demandé d'effectuer une analyse en composantes principales. Cette dernière vous permettra d'étudier des données de grandes dimension

1. Données
   Vous disposez pour ce TP de 3 fichiers de données :
   — **Mes_donnees_ACP.mat** : ce fichier contient des données pour illustrer le fonctionnement de l'ACP. Commencer le TP en utilisant ces données.
   — **iris.mat** : ce fichier contient pour chaque individu des mesures physiques sur des iris. Comme il y a plusieurs sortes d'iris différent cherche à visualiser ces différences.
   — **digit.mat** : ce fichier contient des données de grandes dimensions ($\approx$12000 individus et $\approx$700 dimensions) correspondant à des images. Le chiffres 0 et 1 sont écrits dans ces images.

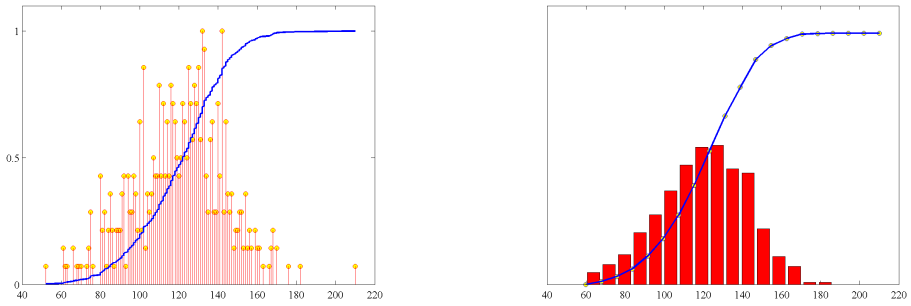


FIGURE 6 – exemples de visualisation à réaliser dans ce TP

Commencer le TP en utilisant le premier fichier. Lorsque vous aurez analysé les données du premier fichier, vous pourrez analyser les autres.

2. Etude des données
   Vous devez créer un fichier de script qui contient plusieurs parties. Lorsque vous aurez fini l'étude avec les données du premier fichier, vous pourrez analyser les données suivantes.

   a) Chargement des données

   1. Charger les données en utilisant la commande **load** suivie du nom du fichier à charger. Les données sont dans la variable *X*.
   2. Visualiser les nuages de points ainsi que les histogramme de ces variables en utilisant la fonction **my_multiplot**.

   b) Analyse en Composante Principale

   1. Faire une normalisation des données et la stocker dans la variable *Xn*. Pour cela, vous devez soustraire sa moyenne à chaque variable et diviser le reste par l'cart type de la variable.
   2. Calculer les matrices *U,V et D* en utilisant la fonction matlab **svd**. Ces matrices contiennent les composantes des individus projeté sur les vecteurs propres pour *U*, les valeurs propres dans la diagonale de *D*, et finalement les vecteurs propres dans les lignes de *V*.

   c) Visualisations

   1. Visualiser (**plot**) les valeurs singulières décroissantes. il y a-t-il beaucoup de grandes valeurs ?
   2. Visualiser la projection des individus sur les axes principaux (matrice *U*). Tracer pour les axes (1,2), (1,3), (2,3) et (4,5).
   3. Visualiser les vecteurs propres (matrice *V*) sur les axes de l'ACP (1,2), (1,3) et (2,3).

   d) Analyse et interprétation

   1. combien d'axe proposez vous de concerner et pourquoi ?
   2. interprtez la projection des individus sur les composantes principales
   3. interprtez les contributions des variables aux composantes principales

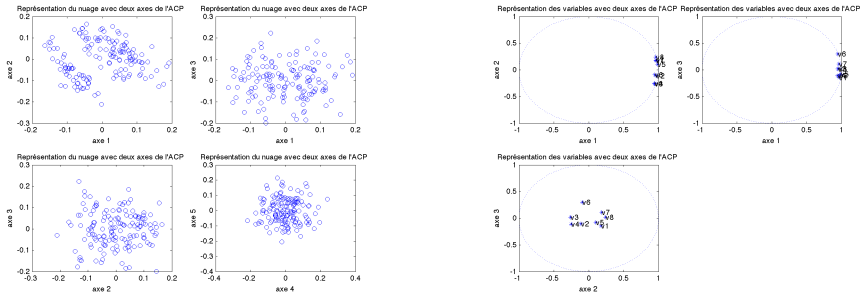


FIGURE 7 – exemples de visualisation à réaliser dans ce TP

Exercice 63 — Régression linéaire

1. Récupérez le jeu de données *Auto MPG data set*[2]
2. Après lecture du fichier de description des données précisez les variables explicative et les variables à expliquer
3. Proposez une méthode pour traiter les valeurs manquantes. Vous pourrez utiliser l'instruction suivante :

```
ind = find(isnan(X(:,3)));
```

4. Effectuez la régression linéaire
   a) estimez les coefficients de la régression linéaire
   b) estimez le coefficient de détermination R^2
   c) calculez les résidus, les leviers et les contributions de chacune des observations
   d) y'a t'il des observations hors épure ?
   e) quelles sont les variables vraiment explicatives ?

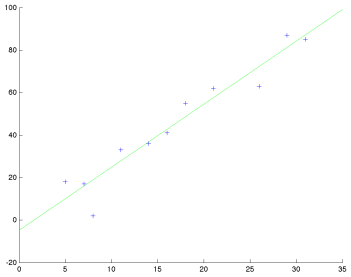


FIGURE 8 – exemple de visualisation à réaliser dans ce TP

Exercice 64 — Test statistiques

1. que fait l'instruction `randn` en matlab ?
2. distribution des moyennes
   a) générer un échantillon de $n = 9$ réalisations d'une variable aléatoire normale centrée réduite.
   b) transformer cet échantillon en un échantillon tiré selon une loi normale d'espérance $m = 10$ et d'écart type $s = 3$
   c) Quelle est la moyenne et la variance de cet échantillon
3. la variance
   a) générez maintenant $\ell = 10^6$ échantillons de taille $n = 9$ d'espérance $m = 10$ et d'écart type $s = 3$. Quelle est la moyenne des variances calculées pour chaque échantillon ?

2. `archive.ics.uci.edu/ml/datasets/Auto+MPG`

4. comparer les moyennes
   a) générer deux échantillon de $n = 32$ réalisations d'une variable aléatoire tiré selon une loi normale d'espérance $m = 10$ et d'écart type $s = 3$
   b) calculez la distance entre les deux moyennes et les deux variances
   c) combien de fois devez vous générer deux échantillons pour que la distance entre les moyennes dépasse 1 (puis 2)
   d) même question si le second échantillon est tiré avec une espérance $m_1 = 9,75$ puis $m_2 = 8,5$,
5. le chi2 avec : $p(1) = 0.2; p(2) = 0.8; q(1) = 0.4; q(2) = 0.6;$
   a) générez un tableau Tp de probabilités de deux lignes et deux colonnes telle que T(i,j) = p(i) * q(j)
   b) générez un tableau T de deux lignes et deux colonnes de $n = 200$ couples d'observations tirées selon les probabilités du tableau Tp
   c) Calculez la distance du chi2 du tableau T

Remarques

Remarques :
— Le compte rendu comprend deux pages : un recto qui décrit ce que vous avez fait et un verso qui décrit se que vous en avez pensé.
— Répétez les expériences plusieurs fois de manière à pouvoir obtenir des résultats moyens quand vous estimez le temps d'exécution.
— Le compte rendu devra être envoyé avec le code.
— Le code doit être indenté.
— Il doit être dans un format non modifiable (pdf par exemple).

Tables de la loi de Student : Cette table nous donne les valeurs de t telle que $\mathbb{P}(T > t)$ lorsque T suit une loi de student à ν degrés de liberté

ν	0,10	0,05	0,025	0,01	0,005	0,001
1	3,078	6,314	12,706	31,821	63,657	318,313
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,782
8	1,397	1,860	2,306	2,896	3,355	4,499
9	1,383	1,833	2,262	2,821	3,250	4,296
10	1,372	1,812	2,228	2,764	3,169	4,143
11	1,363	1,796	2,201	2,718	3,106	4,024
12	1,356	1,782	2,179	2,681	3,055	3,929
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
31	1,309	1,696	2,040	2,453	2,744	3,375
32	1,309	1,694	2,037	2,449	2,738	3,365
33	1,308	1,692	2,035	2,445	2,733	3,356
34	1,307	1,691	2,032	2,441	2,728	3,348
35	1,306	1,690	2,030	2,438	2,724	3,340
36	1,306	1,688	2,028	2,434	2,719	3,333
37	1,305	1,687	2,026	2,431	2,715	3,326
38	1,304	1,686	2,024	2,429	2,712	3,319
39	1,304	1,685	2,023	2,426	2,708	3,313
40	1,303	1,684	2,021	2,423	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
100	1,290	1,660	1,984	2,364	2,626	3,174
∞	1,282	1,645	1,960	2,326	2,576	3,090

[]