I - INTRODUCTION AUX STATISTIQUES

J-P. Croisille

Université de Lorraine

UEL - Année 2012/2013

1- DONNEES

Domaines:

- Données en sciences de la vie
- Données en écologie
- Données en sociologie
- Données en physique expérimentale

Types des données:

- qualitatives/quantitatives
- discrètes/continues
- univariées/multivariées

Données en sciences de la vie

Données de la concentration en créatine phosphokinase (enzyme relative aux fonctionnement musculaire et cérébral) chez 36 volontaires masculins.

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Il s'agit de "données continues univariées".

Questions:

- Les données sont-elles corrélées entre elles ?
- Peut-on repérer des tendances dans ces données ? Quelles sont la valeur moyenne, la médiane de ces données ?
- Peut-on inférer raisonnablement une tendance concernant la concentration en créatine dans l'ensemble de la population ?

Données en écologie

Donnée de l'acidité de 15 lacs des Alpes (valeur du pH).

7.2	7.3	6.1	6.9	6.6
7.3	6.3	5.5	6.3	6.5
5.7	6.9	6.7	7.9	5.8

Données continues univariées.

Questions:

- Quelle est la moyenne de la valeur de l'acidité de tous les lacs des Alpes ?
- Donner une fourchette de l'acidité à "95%".
- Quelle est la valeur médiane de l'acidité de tous les lacs des Alpes ?
- Au vu de l'échantillon de ces 15 lacs, peut-on raisonnablement affirmer que les lacs des Alpes sont trop acides ?
- Prendre la meilleure décision de politique écologique possible (à l'échelle globale) uniquement au vu de ces 15 valeurs.

Données en sociologie

Données multivariées des valeurs moyennes de consommation alimentaire en France de 7 types d'aliments dans 12 types de familles. Abbréviations: $MA =$ ouvriers, $EM =$ employés, $CA =$ cadres. Le chiffre donne le nombre d'enfants dans la famille.

	pain	**lég.**	**fruits**	**viande**	**volaille**	**lait**	**vin**
MA2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
CA2	372	767	562	1948	927	235	433
MA3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
CA3	438	843	689	2345	1148	243	341
MA4	534	660	367	1620	638	414	407
EM4	460	699	484	1856	762	400	416
CA4	385	789	621	2366	1149	304	282
MA5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
CA5	515	1097	887	2630	1167	561	284

Questions:

- La catégorie sociale influence-t-elle le type des aliments consommés ?
- le nombre d'enfants a-t-il davantage d'influence dans une catégorie sociale qu'une autre ?
- Quel type de décision de politique familiale concernant l'alimentation ces données suggèrent-elles ?

Données expérimentales

On mesure la largeur d'une pièce. On réalise 10 mesures, qui donnent

valeur	323	324	325
nombre de fois	1	5	4

Questions:

- L'appareil de mesure (mètre laser) est-il bien réglé ?
- A-t-on pris les mesures à des endroits corrects dans la pièce ?
- Quelle est la "largeur moyenne" de la pièce ? Quelle est l"'incertitude" sur cette valeur moyenne ?

2- La démarche statistique

Deux types de questions:

Questions sur la nature des données, leur validité

- Les données sont basées sur un **échantillon**. L'échantillon est-il correctement constitué ? Y-a-t-il des corrélations entre les individus de l'échantillon ?
- Problème sur l'erreur que l'on peut commettre sur les données: fiabilité et précision d'un appareil de mesures, d'un questionnaire: erreurs systématiques.

Il s'agit d'un travail sur les données elles-même.

Questions sur la population globale sous-jacente aux données

- Evaluation d'une "mesure globale": largeur de la pièce, etc.
- Caractériser la variabilité des données. Sous quelle forme des mesures répétées donnent-elles un résultat identique ?
- Les données permettent-elles de faire un diagnostic fiable pour des décideurs ? Problème de l'existence d'une méthode rationnelle pour extrapoler des renseignements d'ordre global à partir d'un seul échantillon.

Problème de **l'inférence statistique**

Types d'expériences

Principalement deux types de démarches en statistiques:

- *Détermination de paramètres*
  Determiner la valeur numérique d'une quantité physique à partir de mesures expérimentales.
- *Tests d'hypothèses*
  Tester si un modèle particulier ou une prédiction est consistante avec des données.

Terminologie

- *Précision*
  Caractériser comment le résulat d'une expérience approche la véritable valeur.
- *Consistance*
  Caractériser comment des mesures répétées donnent un résultat identique.
- *Erreurs aléatoires:*
  Erreurs dues au fait même de mesurer. On mesure concrètement quelque chose qui n'est qu'une occurence (un exemplaire, un exemple,..) d'une quantité. La largeur d'une pièce par exemple, n'est jamais parfaitement définie. C'est une quantité qui est en fait une moyenne. Effort pour une définition aussi claire que possible de ce que l'on souhaite mesurer.
- *Erreurs systématiques:*
  Erreurs dues à l'appareil d'observation, de mesure. Même si la quantité mesurée existait dans l'absolu, on ne peut pas l'atteindre à cause de l'imperfection de l'instrument de mesure.

3- Histogrammes

Exemple: données sur 32 cerisiers

Arbre	Diam.	Haut.	Volume	Arbre	Diam.	Haut.	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	16.9	66	64.3
11	11.3	79	24.2	27	17.3	81	55.4
12	11.4	76	21.0	28	17.5	82	55.7
13	11.4	76	21.4	29	17.9	80	58.3
14	11.7	69	21.3	30	18.0	80	51.5
15	12.0	75	19.1	31	18.0	80	51.0
16	12.9	74	22.2	32	20.6	87	77.0

Données multivariées, continues.

Diamètre des cerisiers.

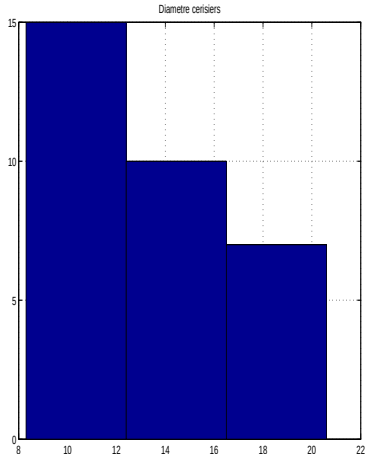


Figure: 3 classes

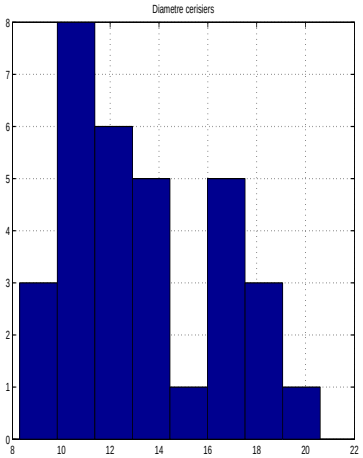


Figure: 8 classes

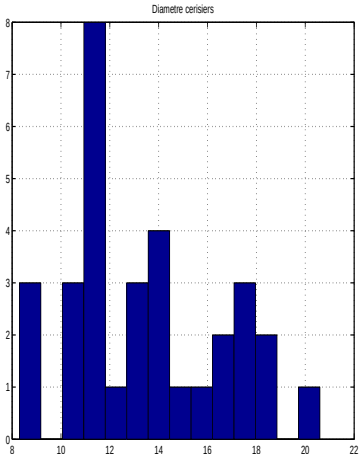


Figure: 14 classes

Hauteur des cerisiers.

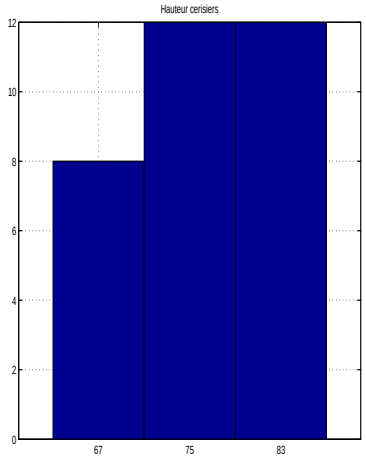


Figure: 3 classes

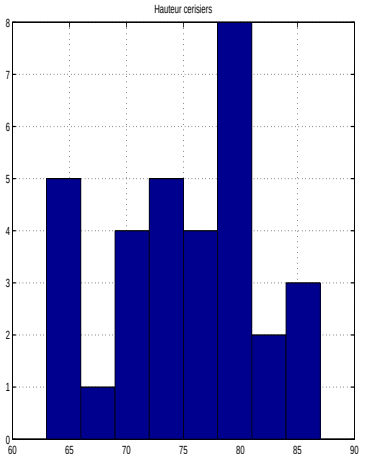


Figure: 8 classes

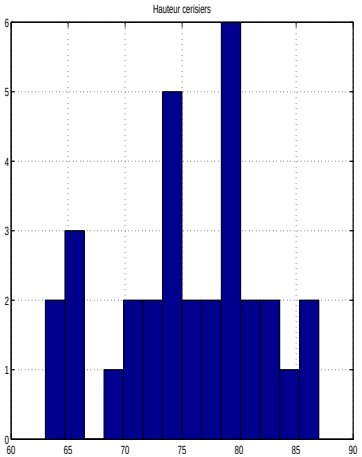


Figure: 14 classes

Volume des cerisiers.

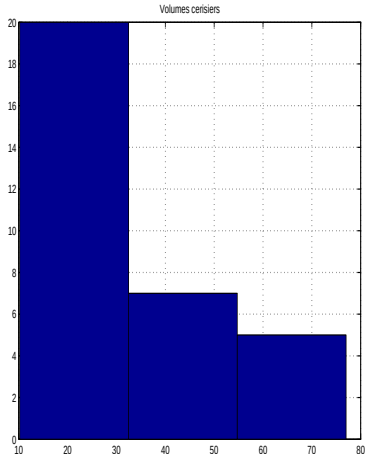


Figure: 3 classes

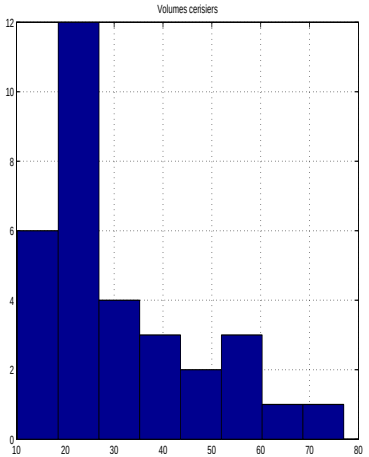


Figure: 8 classes

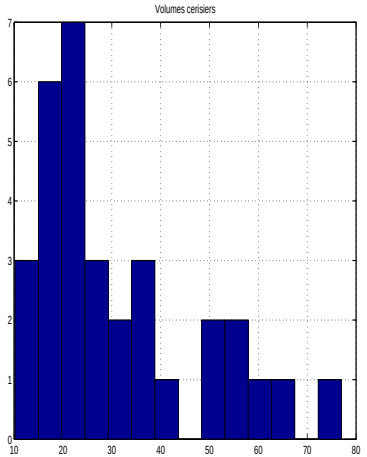


Figure: 14 classes

Regroupement des données.

Avant de représenter un histogramme, il est important de définir précisément les *classes* que l'on représente.

- Le nombre de classes doit être suffisamment petit pour donner un résumé pertinent des données.
- Toute observation doit appartenir à une seule classe.
- Lorsque c'est possible, toutes les classes doivent avoir la même largeur en abscisse sur l'histogramme.

Pour 32 arbres,

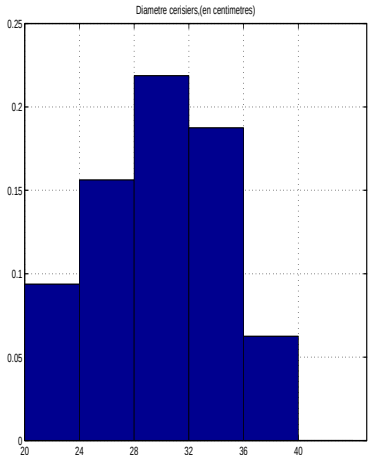
- 3 classes est un nombre trop faible.
- 14 classes est un nombre trop important.
- 8 classes conviennent à peu près pour décrire chacune des caractéristiques de la population des 32 arbres.

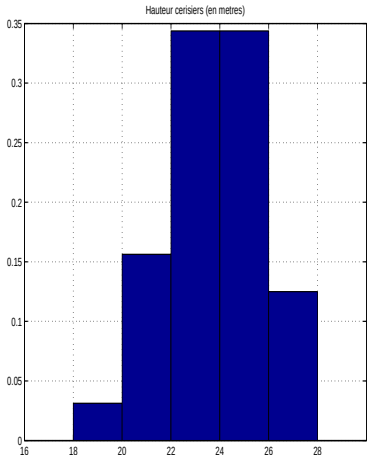
Limites de classes explicites, fréquences

On peut souhaiter des variantes dans la définition de l'histogramme. Par exemple

- Spécifier soi-même les limites des classes.
- Mettre en ordonnée, non pas le nombre d'éléments (la fréquence), mais la fréquence *relative*.

Diamètre, hauteur, volume des cerisiers en fréquence.





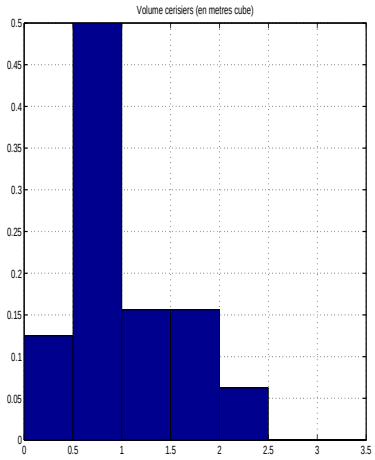


Figure: Diamètre, 20-24-28-32-36-40, cm

Figure: Hauteur, 18-20-22-24-26-28 mètres

Figure: Volume, 0-0.5-1-1.5-2.-2.5 m^3

Terminologie

- **Classe:** Catégorie pour grouper des données.
- **Fréquence:** Le nombre d'observations dans une classe.
- **Distribution des fréquences:** Liste de toutes les classes et de leur fréquences.
- **Fréquence relative:** Rapport de la fréquence d'une classe au nombre total d'observations.
- **Limites inférieure et supérieure d'une classe:** les nombres x_i, x_s tels que la classe soit définie par les conditions $x_i < x < x_s$.
- **Centre d'une classe:** Milieu de la classe $x_m = (x_i + x_s)/2$.
- **Largeur d'une classe:** Valeur $\Delta x = x_s - x_i$.

4- Mesures sur les données

Moyenne empirique

On considère un **échantillon** de n observations, $x_i, i = 1, ..., n$. On définit la **moyenne empirique**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

Exemple: Les salaires de travailleurs saisonniers en été s'établit en deux groupes.

Premier groupe	300	300	300	940	300	300	400	300	400	450	800	450	1050

Deuxième groupe	300	300	940	450	400	400	300	300	1050	300

Moyenne du premier groupe: $m_1 = \frac{6290}{13} = 483.85$.
Moyenne du deuxième groupe: $m_1 = \frac{4740}{10} = 474.00$.
Conclusion: les personnes du premier groupe gagnent davantage en moyenne que celles du second.

Médiane

La **médiane** est définie comme la valeur x_M tel que

$$\mathcal{P}(x < x_M) = 0.5 \quad , \quad \mathcal{P}(x > x_M) = 0.5 \tag{2}$$

La médiane est la valeur x_M qui divise l'échantillon de données en deux parties égales. Si les données sont rangées en une liste de valeurs croissantes, alors

- Si le nombre de données est impair, la médiane est la valeur qui est exactement au milieu de la liste.
- Si le nombre de données est pair, la médiane est la moyenne des deux valeurs centrales.

Médiane (2)

Les données de chaque groupe sont rangées par ordre croissant.

Premier groupe	300	300	300	300	300	300	400	400	450	450	800	940	1050

Deuxième groupe	300	300	300	300	300	400	400	450	940	1050

Médiane du premier groupe: $m_1 = 400$.

Médiane du deuxième groupe: $m_2 = 350$.

Conclusion: les personnes du premier groupe gagnent davantage que ceux du second, car le salaire médian du premier groupe est de 400 et de 350 dans le second.

Le mode

C'est la valeur la plus fréquente pour une distribution discrète: c'est la classe correspondant au pic d'un histogramme. Sa détermination n'est pas aisée et dépend du découpage en classes de l'histogramme.
Pour une répartition parfaitement symétrique, on a

$$\textit{moyenne=médiane=mode}$$

Important: chaque notion a sa logique propre. Il faut choisir la notion la plus pertinente en fonction de la situation que l'on veut décrire.
La moyenne est sensible aux valeurs extrêmes, la médiane l'est moins.

Exemples

1. Un élève a obtenu $14, 11, 16, 20, 9$ comme notes en mathématiques pendant un trimestre. On calcule sa moyenne $14/20$, qui tient compte de toutes les notes avec poids égal. Les bonnes notes et les mauvaises notes interviennent et se compensent dans le calcul de la moyenne.
2. On dispose de la liste des prix des appartements vendus dans une grande ville. La meilleure mesure centrale est la médiane. C'est le prix au dessous duquel la moitié des appartements ont été effectivement vendus. Ce nombre atténue l'effet des appartements très chers et très peu chers. Il renseigne aussi (par exemple) sur le pouvoir d'achat effectif des acheteurs.
3. Au marathon de Paris, il y a 7500 hommes qui terminent et 4000 femmes. Chaque observation est "homme" ou "dame". Le mode "homme" est la mesure centrale adaptée pour l'observation "Concurrent ayant terminé l'epreuve".

Logique des mesures moyennes

- La *moyenne* est relative à la *somme* des valeurs de l'échantillon. Cette somme représente une richesse ou potentiel *global* de la population. Si la moyenne du poids de 100 poulets est de 1.850 kg, cela signifie 185 kg en tout, ce qui correspond à une valeur marchande globale de l'élevage.
- Au contraire, la *médiane* représente une valeur pour le *client, patient, consommateur,...*. Exemple du temps de survie de patients atteints d'un cancer après un certain traitement médical. Il arrive que l'on ait une survie de moins d'un an pour la plupart des patients et une survie de plus de dix ans pour quelques-uns. Dans ce cas la moyenne a peu de sens, et la médiane représente au contraire la survie effective d'un patient typique.

Ecart-type, variance

Considérons un échantillon d'une grandeur physique X. Les valeurs de l'échantillon sont les mesures $x_1, x_2, x_3, ..., x_n$. La moyenne empirique est

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

On définit l'*écart-type empirique* de l'échantillon par

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2} \tag{4}$$

La quantité s^2 s'appelle la *variance empirique*. On a

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 \tag{5}$$

Explication qualitative du $n-1$

Une seule mesure est suffisante pour avoir une information concernant la moyenne, mais il faut au moins deux mesures pour avoir une information sur l'écart à la valeur moyenne. La véritable justification est que l'on choisit le facteur $n-1$ pour que la moyenne de la variance empirique soit la vraie variance.

5- Box-plots

Une représentation graphique très parlante pour examiner une distribution et pour faire des comparaisons entre distributions est la représentation **box-plot** (ou diagramme moustache).
La **valeur médiane d'une distribution** Q_2 partage l'échantillon observée en deux moitiés d'effectif égal: salaire médian, note médiane, performance médiane. On a donc

- La partie supérieure de l'échantillon: ce sont les individus i tels que $x_i > Q_2$.
- La partie infèrieure de l'échantillon: ce sont les individus i tels que $x_i < Q_2$.

Cette notion est généralisée par la notion de quartiles. On divise à nouveau chacune de ces deux moitiés en deux. Ceci partage l'échantillon en 4 parties d'effectif égal. Les 3 valeurs frontières sont notées Q_1,Q_2 =médiane, Q_3 et sont appelées les 3 **quartiles**. On appelle **intervalle interquartile** la quantité IQR $= Q_3 - Q_1$.

Exemple: pression sanguine chez 7 sujets

151	124	132	170	146	124	113

. En ordonnant de façon croissante, on obtient:

113	124	124	132	146	151	170

- Médiane: $Q_2 = 132$ valeur de l'individu du milieu.
- Médiane de la partie infèrieure de l'échantillon: $Q_1 = 124$.
- Médiane de la partie infèrieure de l'échantillon: $Q_3 = 151$.

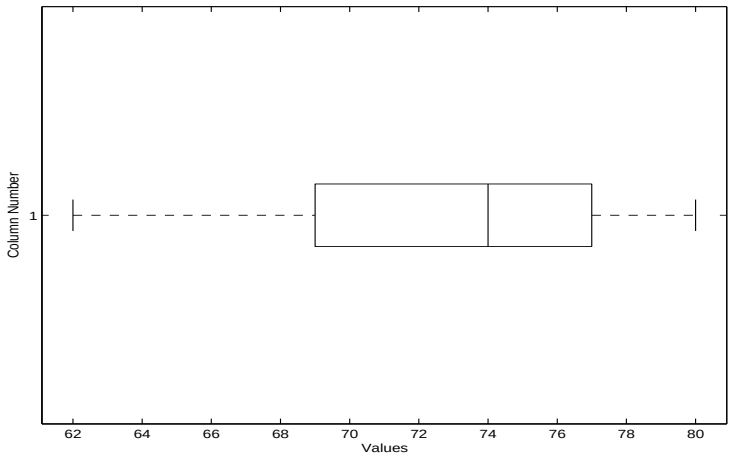
Pression sanguine (en mm Hg) pour sept sujets hommes d'âge moyen

62	64	68	70	70	74	74	76	76	78	78	80

- Médiane: $Q_2 = 74$ valeur de l'individu du milieu.
- Médiane de la partie infèrieure de l'échantillon: $Q_1 = \frac{68+70}{2}$.
- Médiane de la partie infèrieure de l'échantillon: $Q_3 = \frac{76+78}{2}$.

Boxplots

Le minimum, le maximum, la médiane et les 2 quartiles Q_1, Q_3 constituent un résumé de la distribution avec 5 nombres. Ces 5 nombres sont représentés graphiquement par le diagramme “box-plot”.



Boxplots - comparaison

Il est souvent très parlant de représenter en parallèle les boxplots de 3 échantillons de même effectif, mais avec des conditions différentes. Voici la croissance d'un végétal (un radis) dans trois conditions de lumininosité

- *obscurité*

15	20	11	30	33	22	37	10	29	35	8	10	15	25

- *1/2 lumière, 1/2 obscurité*

10	15	22	25	9	15	4	11	20	21	27	20	10	20

- *lumière*

3	5	5	7	7	8	9	10	10	10	10	14	20	21

Boxplots - comparaison (2)

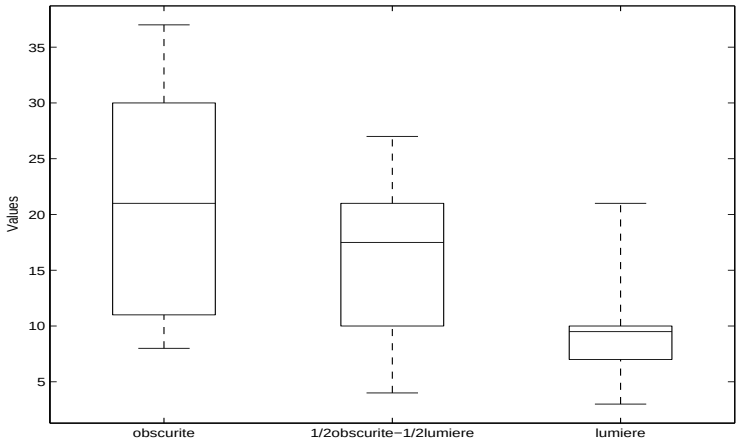


Figure: Boxplot, comparaison de croissance, 14 radis

Valeurs extrêmes, boxplots modifiés

Une *valeur extrême* est une valeur de l'échantillon qui est "aberrante" ou "suspecte". Par convention, on donne habituellement ce nom à une valeur x_i telle que

$$x_i < Q_1 - 1.5\,\mathrm{IQR} \text{ ou } x_i > Q_3 + 1.5\,\mathrm{IQR} \tag{6}$$

On prolonge les "moustaches" seulement jusqu'à un maximum de 1.5 fois l'intervalle interquartile, et on represente par un symbole les individus au delà des limites supérieures et inférieures ainsi définies.

Boxplots - comparaison (2)

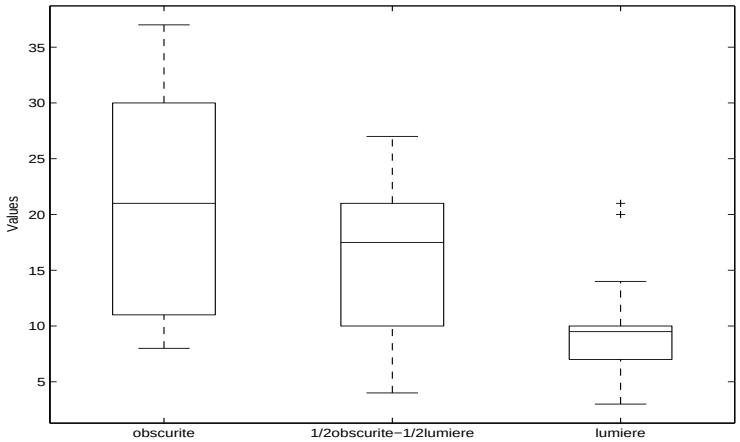


Figure: Boxplot, comparaison de croissance, 14 radis

6- Représentation des données multivariées

Etude de l'origine du chien préhistorique. Corrélations avec différents canidés: le chien moderne, le chacal doré, le loup chinois, le loup indien, le cuon et le dingo. Les données anatomiques sont les suivantes $X_1 =$ largeur mandibule, $X_2 =$ hauteur mandibule, $X_3 =$ longueur de la première molaire, $X_4 =$ largeur de la première molaire, $X_5 =$ distance de la première molaire à la troisième molaire, $X_6 =$ distance de la première molaire à la quatrième molaire.

	X_1	X_2	X_3	X_4	X_5	X_6
ch.mod.	10	21	19	8	32	37
chac.do.	8	17	18	7	30	33
l. chin.	14	27	27	11	42	48
l. ind.	12	24	25	9	40	45
cuon	11	24	21	9	29	38
dingo	10	23	21	8	34	43
ch.pr.	10	22	19	8	32	35

Matrice des données

On range les données dans une matrice notée X.

$$X = \begin{pmatrix} 10 & 21 & 19 & 8 & 32 & 37 \\ 8 & 17 & 18 & 7 & 30 & 33 \\ 14 & 27 & 27 & 11 & 42 & 48 \\ 12 & 24 & 25 & 9 & 40 & 45 \\ 11 & 24 & 21 & 9 & 29 & 38 \\ 10 & 23 & 21 & 8 & 34 & 43 \\ 10 & 22 & 19 & 8 & 32 & 35 \end{pmatrix} \tag{7}$$

- Taille de l'échantillon=n= nombre de lignes.
- Nombre de caractères étudiés=p=nombre de colonnes.

Caractéristiques d'une matrice des données

- Vecteur moyenne empirique:

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{8}$$

où x_i= ième individu.

- Matrice des covarances empiriques $p \times p$ $C = (c_{j,k})$, $1 \leq j, k \leq p$

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)(x_i - m)^T \tag{9}$$

Le coefficient $c_{j,k}$ représente une mesure de la corrélation empirique entre le j-ème et le k-ème caractère.

Caractéristiques d'une matrice des données (2)

Matrice des corrélations: C'est la matrice R également $p \times p$ (p= nombre de caractères) déduite de la matrice des corrélations empiriques par

$$r_{j,k} = \frac{\sum_{i=1}^{n}(x_{ij} - m_j)(x_{ik} - m_k)}{(\sum_{i=1}^{n}(x_{ij} - m_j)^2)^{1/2}(\sum_{i=1}^{n}(x_{ik} - m_k)^2)^{1/2}} \tag{10}$$

Le coefficient $r_{j,k}$ est un coefficient sans dimensions qui représente la corrélation entre les j-ème et k-ème caractère.

Courbes d'Andrew

Chaque courbe représente une observation à p caractères.

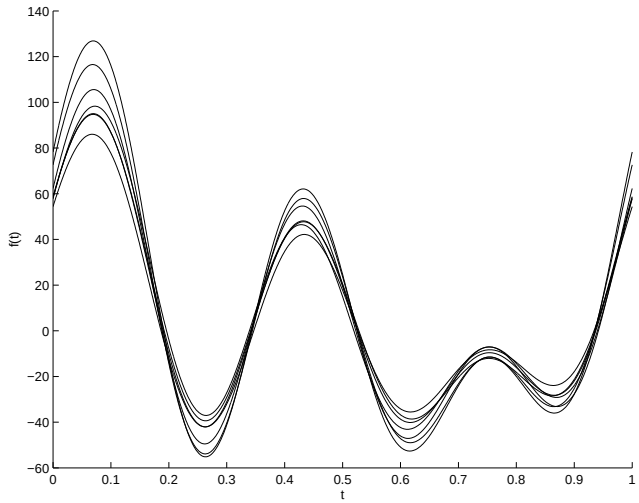


Figure: Courbes d'Andrew, données chien préhistorique

Boxplots multiples

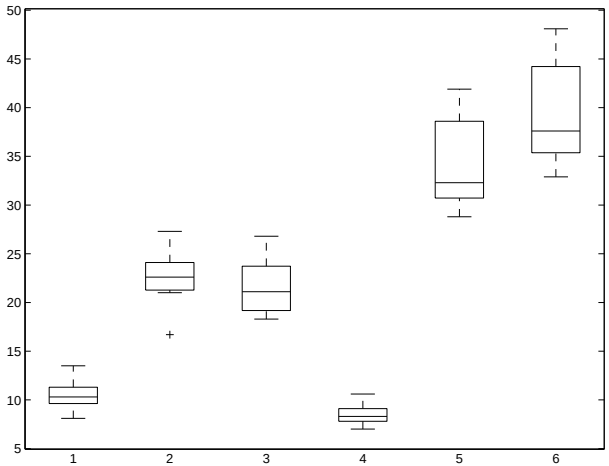


Figure: Boxplots multiples, données chien préhistorique

Représentation en étoile

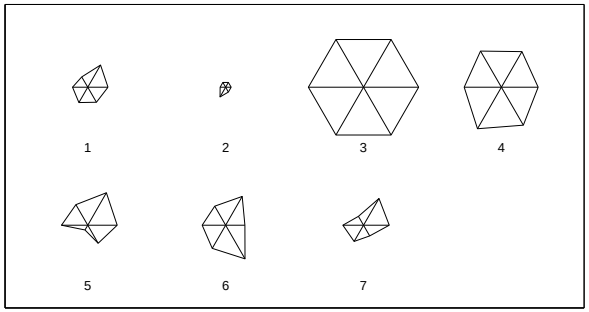


Figure: Représentation en étoile, données chien préhistorique

Visages de Chernoff

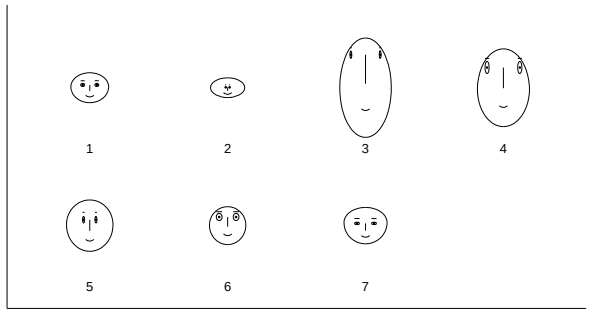


Figure: Visages de Chernoff, données chien préhistorique

Scatterplot

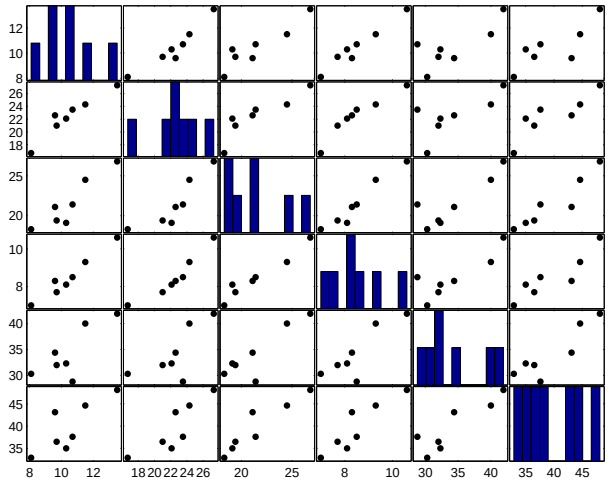


Figure: Scatterplot, données chien préhistorique