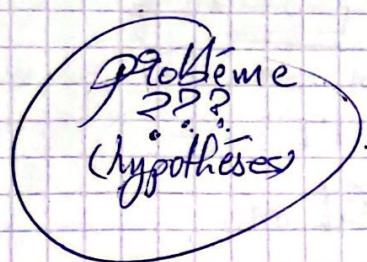


Statistiques et analyse de données

Introduction sur le module.

- La statistique c'est une science qui permet d'étudier une phénomène (aperçu du phénomène) / problème (^{on pose des Q})
- Les statistiques, ensemble des résultats qui nous permet d'étudier un phénomène. exemple: population d'une usine ...
- Analyse des données; La compréhension des données. (faire une étude) à fin de prendre des décisions



processus:

1. Collecte des données

exemples: Sismique Houzez

- impact - zone fortement agravee...

méthode: Questionnaires

- Observation

- outils - capteurs

exemple 2: Société

- Etude de marché / Satisfaction

faire des Interviews pour savoir

Tendance des Etudes

2. Prétraitement

- signifie la structuration / transforme à une forme numérique pour les traiter numériquement + sous forme Tableaux

Variable	Individu

Données
Brutes

Structurées

Tableau

Individu

Variable

Questionnaires / interview

1. Q
R₁
2. Q
R₂
...
m. Q
R_m

Traduction →

accrue dépend à notre
problème + importance d'expert

indiv	Var 1	Var m
1	R ₁ réponse	—
...	—	—
n	—	R _m

- chaque Q est traduit à une variable significative.
 - chaque Individu représente une feuille Questionnaire
- ⇒ Data Utilisable.

! Population → on prend un Echantillon

Tres Grande

Taille = n

— doit être représentatif

à fin de minimiser l'Erreur

But final

Généraliser les Etats
d'un échantillon sur
toute la population

Echantillon → population

estimation

3 - Résumage (paramètre, Graphique ...)

4 - Analyse

× les méthodes descriptive : comprendre

× Inférentielle prédition, c'est dans le futur
Là météo [l'Anticipate]

1 - Statistique descriptive

- a) Le recensement (الاستبيان)
- a) Les Interviews
- a) Les Questionnaires
- b) Q → Table
- c) Analyse.

Il est préférable d'utiliser un échantillon taille $n > 30$

loi normale

$n < 30$ loi Student

- Estimation
- Hypothèse : formulation du problème ou de ...

↳ se fait sous la forme suivante :

Q → Test d'hypothèse

$$\begin{array}{l} \text{Hypothèse} \\ \text{nul } H_0: \quad \text{Test d'hypothèse} \\ \text{(rejet)} \end{array}$$

$H_0: \quad \text{H} = \theta_0 \quad (\text{H} - \theta_0 = 0 \text{ c'est à dire dans le}$

$H_1: \quad \text{H} = \theta_1 \quad (\text{H} - \theta_0 \neq 0 \text{ c'est à dire hors de})$

Hypothèse

Alternative

$\text{Je vérifie ce que j'ai trouvé c'est}$

$\text{celui que j'ai calculé ou pas ?}$

Et on dit : sous l'hypothèse H_0 soit vraie, α : seuil de rejet / l'erreur

$$(1\%, 5\%) = 10\% \\ 0,01 \quad 0,05 \quad 0,1$$

$T^* = \text{statistique de Test}$

Les loi qu'on a:

Loi normale
 Loi Student
 Loi χ^2
 loi Fisher

Si $T_{obs} > T_{critique}$ \rightarrow Calcul Empérique

Types de données

Qualitative

continu
le temps

discrete

number of

Quantitative

~~example~~ nominaux
couleur des yeux
name of

Binary

ordinal
taille des vêtements
rating
ranking

2. Statistique descriptive:

categoriel: effectifs...

a/- Donnée Quantitative

1- Je prends dans le comportement d'une variable \rightarrow Chart pour trouver la loi sur laquelle on va se baser

b/- Représentation graphique.

- type des Graphes: - histogramme : continuiter d'intervalle.
- ... scatter diagram.

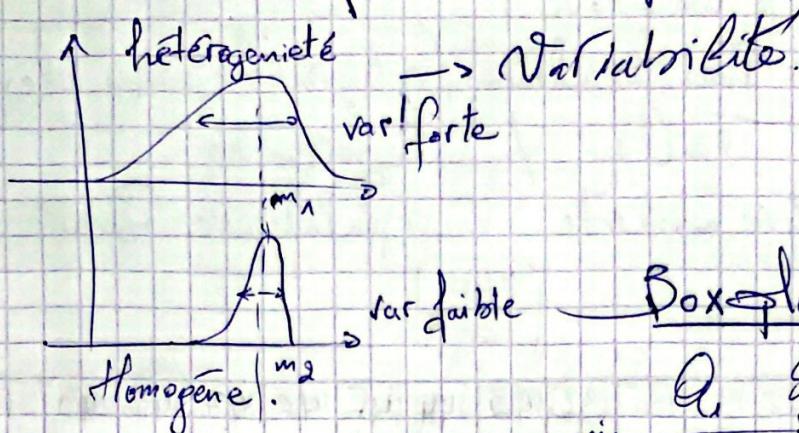
Sur chart = qualitatif

c/- Représentation Numérique.

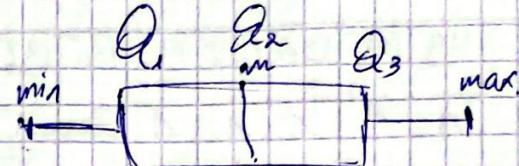
• Relation entre Central Tendency et variability \rightarrow
lors de notre représentation graphique on cherche le
comportement

• En distribution, le pic est la moyenne.

→ les individus sont proches ou loign à la moyenne.



Boxplot:



$n_1, n_i \in \text{SRQ}$
⇒ outliers

rang
de variabilité (SRQ)

Maintenant on passe d'un tableau l'individu/variable.

Vers 3

Catégories	classes	effectifs	fréquence	$f_{c,d}$	$f_{c,d}$
A	n_A	n_A	$f_A = n_A/N$	f_A	$\frac{1}{N} (100\%)$
Z	n_Z	n_Z	$f_Z = n_Z/N$	f_Z	$1 - f_A$

$A = n_A \rightarrow$ opposition de A

$N = n_A + \dots + n_Z =$ effectif total = n.

Pourquoi $f_{c,c}$ et $f_{c,d}$? Utile pour une Q qui se pose pour un ensemble de classes.
[catégorie]

R Studio

Régression linéaire Simple :

- Collecte de donnée
- Prétraitement.

- Analyse 1 : analyse univariée (S. descriptive)

ou se base sur

La représentation graphique.

et

Tendance Central
Median Mode

taille
échantillons $\rightarrow n$

- Analyse 2 :

Le problème \Rightarrow mesuré la satisfaction

Théorie
Centrale
Limite

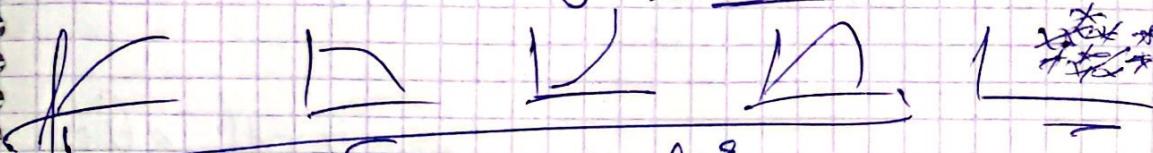
\hookrightarrow analyse Bivariée.

- Analyse 3 : analyse Multivariée. $X = (X_1, \dots, X_n)$

Pour faire une analyse de doit avoir le type de variable.

1er cas : les 2 vars qt (x, Y)

\Rightarrow Régression linéaire;



trouver une relati

mage de put.

Scatterplot.

\rightarrow une liaison linéaire entre les variables,

la pente $\rightarrow a \approx b$ Coeff.

1 - graphe, scatter plot

2 - Estimation de a et b : $y = a_n + b_n + \epsilon$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow ?$$

V. indépendante

Precipitation \rightarrow

La récolte \rightarrow

Variable dépendante

G_1 et G_2 , linearité \oplus et \ominus .

2 - Mesure la corrélation;

\rightarrow mesurer la liaison entre 2 vars

$$\text{Cov}(x, y) = E((x_i - \bar{x})(y_i - \bar{y})) = \frac{1}{n} \sum_i (x_i y_i - \bar{x} \bar{y})$$

$$\rightarrow \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{n} \sqrt{y}}$$

$$|\text{Corr}| \leq 1$$

$r \approx 1$
liaison forte \oplus

$r \approx 0$
pas de liaison
 $r \approx -1$
forte \ominus

Test d'hypothèse \Rightarrow Mesurer la significativité du test

$$\left\{ \begin{array}{l} H_0: \rho = 0 \rightarrow (\text{pas significatif}) \\ H_1: \rho \neq 0 \rightarrow (\text{significatif}) \end{array} \right.$$

$$\left\{ \begin{array}{l} \mathcal{D} = 0 \\ \mathcal{D} \neq 0 \end{array} \right.$$

$\alpha \ll$ erreur de 1^{er} espèce (Rejeter H_0 alors

H_0 est vrai) $\alpha = 1\%, 5\%, 10\%$

Rejeter H_0 si $|T_{obs}| > |T_{crit}|$

Fatdean des 3 de statistiques

iques

Rejeter l'hypothèse :

$$\Leftrightarrow T_{\text{obs}} > T_{\text{crit}}$$

$$\Leftrightarrow P\text{-value} \leq \alpha$$

Si $P_1 = 0,02561 < 0,05 \Rightarrow$ corrélation significative
 $P_2 = 0,0684 \Rightarrow$ corrélation pas significative.

$$\begin{cases} \alpha = 0,05 \\ \text{par défaut} \end{cases}$$

$$[y - a] = \frac{\text{Cov}(x, y)}{\text{var}(x)}, \quad \boxed{b = \bar{y} - a\bar{x}}.$$

$$y = \hat{a}x + \hat{b} \quad ? \text{ test d'hyp sur } a \text{ et } b$$

a non significative $\Rightarrow y$ non significative.

$$\Rightarrow a = 0 \quad (y \text{ n'est pas significative})$$

Comment valider le modèle ?

4) - Validité du modèle:

de voir

① R^2 : coeff de détermination.

$R^2 \approx 1 \Rightarrow$ Bon/pas mal.

$R^2 \approx 0 \Rightarrow$ modèle non significatif.

Valeur \rightarrow test d'hypothèse.

② Analyse des résidus ; vérifie la normalité, homogénéité

$$\begin{array}{c} e_i \\ | \\ x_i \end{array} \quad \begin{array}{c} e_i \\ | \\ y_i \end{array}$$

Mass corporelle	60,2	62	62,9	36,1	54,6	48,5	42	47,4	50,6	42
Métabolisme	1679	1792	1666	995	1425	1396	1418	1362	1502	1256
	47,7	40,3	57,9	46,9	33,1	51,9	42,4	54,1	51,1	41,2
	1614	1189	1767	1439	913	1460	1124	1052	1377	1204

	estimate	error	t-value	p-value
--	----------	-------	---------	---------

intercept

n_1

n_2

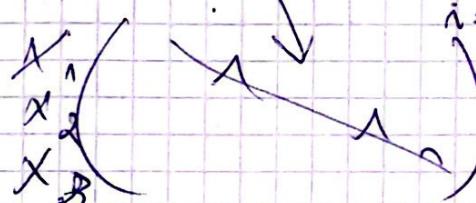
n_3

Valeur

$+b_0 \quad a_i = 0$

$\quad a_i \neq 0$

$$y = a_1 n_1 + a_2 n_2 + a_3 n_3 + b_0 \\ = \sum_{i=1}^3 a_i n_i + b_0$$

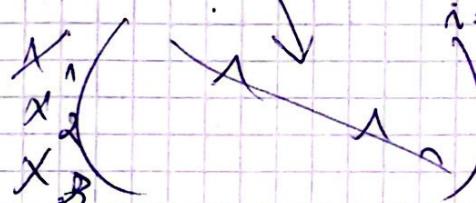


La matrice de Corrélation

Tableau Summary

lm(y ~ n1, n2, ...)

$$\text{corr}(x_1, x_2) = r_{12}$$



→ Odt à faire dans RLM → doit être non corrélés.

y appart

n --- v. indep

x_1 doit pas être lié à x_2 ...

Corrélation de n_i et n_j faible

doit être
 $\text{corr}(x_i, x_j) \approx 0$
faible

$J_{ij} /$
 $\text{corr}(x_i, x_j) \neq 0$

on peut pas y - n

La régression.

RLM comme RS

les regresseurs doivent pas être corrélés entre eux.

A la solution ? Eliminer l'un des vars (si on a une idée déjà) ou proposer une autre analyse / technique.

→ Analyse en composant principale (ACP).

objectif : 1. regroupement des Elts corrélés.
vars d'individu \Rightarrow La dimension diminue

2.

$$X = (x_1, \dots, x_p) \Rightarrow C_{Pi} = \sum_j a_{ij} x_j$$

je dois faire un.

Une composante principale.

on doit garder [la trace] de l'individu

max des infos

Il existe au moins deux vars corrélées.

1. Matrice de Corrélat factorisable.

$$X \cdot \mathbb{1}_{au\;mois\;i+j} / \text{corr}(x_i, x_j) \leq 0 | X'X$$

2) - Calcule les valeurs + vecteurs propres

en identifie les Composantes en pratiquant

Calcule les valeurs propres.

$C_{Pi} \rightarrow x_1$

$C_{P2} \rightarrow x_2$

$C_{PP} \rightarrow x_p$

ordre décroissant

d'importance

$R^P \rightarrow R^D$

on doit découper (en gardant le max d'info)

75-80% \rightsquigarrow

$$X'XU = \lambda U$$

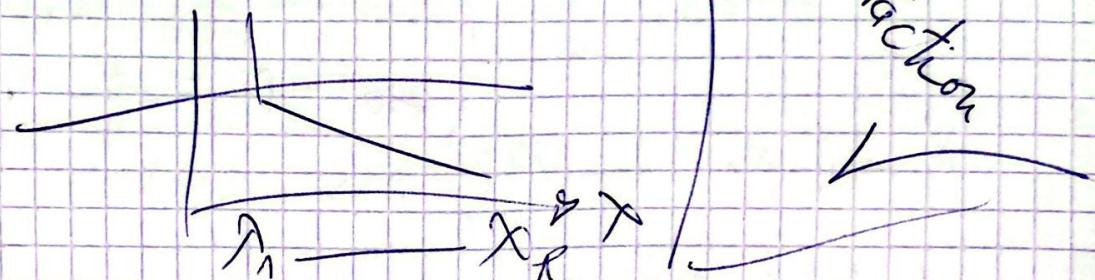
$|x_1 + \dots + \lambda d \approx 80\%|$ ou bien on dit, Indice de Keyzer
on utilise la valeur de λ ,
on prend valeur sup à 1.

3) Extraction des composantes principales

les conditions

- X communis ; 80% (tac de X) \Rightarrow faire le même échelle.
- Indice Keyser $\lambda > 1$ (normé) \rightarrow $\lambda = \frac{x_i - \bar{x}}{s_x}$
- Screen coude

$\Phi \rightarrow CP^1$

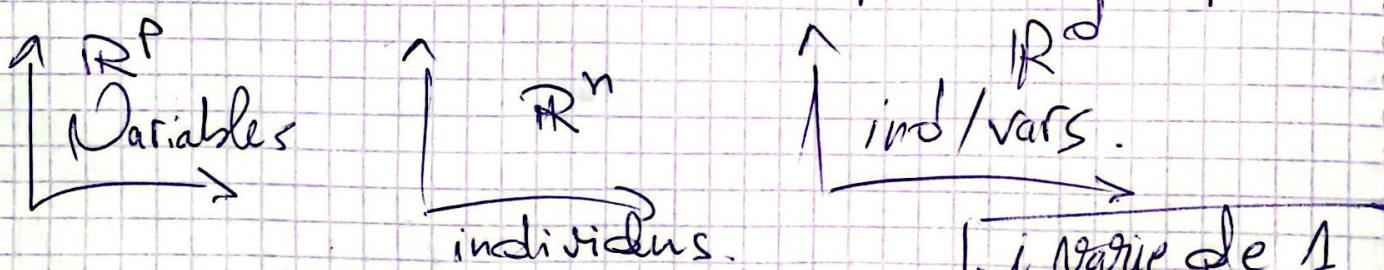


\rightarrow Après l'E des composantes

Ach faywge3? les CPS sont les nouvelles vars
 droite non pas une valeur

$$\begin{array}{c|c|c|c} x_1 & x_2 & \dots & x_p \end{array} \xrightarrow{\text{axe}} \begin{array}{c|c} CP_1 & CP_2 \end{array}$$

La variable de quelle composante se compose.



\rightarrow passer d'une dim large (espace factorisable) $\xrightarrow{\text{a P}}$
 à P. Combin des x_i composantes initiales.

ACP

Réduire l'espace
de dimension

$P \leq k$ petit (α, β)

ils ne pensent
pas
ensemble = point isolé

Identifier les regroupements
isolation



regroupement

soit soit
sur les individus soit sur les vars

Tableau

ind	x_1	x_2	...	x_n
1				
...				
n				

on doit minimiser la variante

R^u (vars)

R^p (individus)

projection

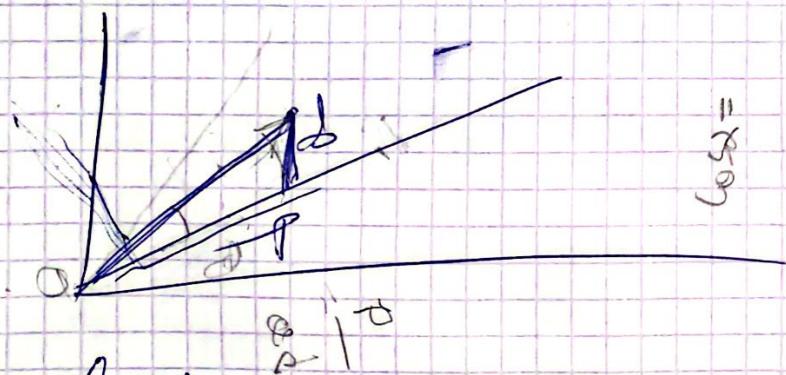
. Reg linéaire, on parle sur une droite

. Maintenant, on parle sur l'espace, (ensemble à la classification)

- min la variance interne
- max — externe

distance

la projection



Plan de représentation qu'on cherche: (C_{Pi})

entre l'axe ACPi, et le
vecteur d'individu.

{ Composante
principale }

on parle ici sur la Corrélation

et la Qualité de Représentation (bien présenté ou pas)

exemple Classification Variable qui porte valeur -



II - La Représentation des individus + vars
dans l'esp. plan (CPI)

SI - Interprétation

variable qui contribue dans la construction d'une axe
(de quoi?)

Autre info la distance /
l'angle (cosine)
entre le vecteur
de variable de
l'ordonnée à l'axe
et avec l'axe de
projection

La projection d'une pt
sur un axe, c'est quoi?

→ Ses coordonnées selon
son vecteur. Si va être grande

Cos doit approcher à 1.
mli fait koun l'angle de
projection s'agit tend vers 0

Qualité ; sa représentation est bien \uparrow/\downarrow la cos².

Pour les vars, c'est la même chose, proche des bords du
cercle de la corrélation, les valeurs éloignées sont mal
matrice de Corrélation

fait pas partie.



feature selection, ACP fait partie de la classification (s'il en
mon dataset)

Qualité sur un axe, on a la représentation, [plan]
on essaie de faire la somme, Cpi par rapport à qj

- La proximité dans l'espace, homogénéité, l'ensemble qui se proches coorelent négativement + positivement.

II. Corrélation 1.

$$\text{Rep} \quad \text{Corr}(x_1, x_2) = 0 \quad \text{Corr}(x_1, x_2) \\ \downarrow \qquad \qquad \qquad \rightarrow x_1 \qquad \qquad \qquad \rightarrow x_2 \\ \text{L} \rightarrow \text{ne sont pas corréler.}$$

Fisher on a trouvé la Corrélation et tous, maintenant on s'orient à la réalité, c'est quoi la réalité ?
(Garder les fins) Les données

Très important

lorsque

les cont. cont. qui sont les

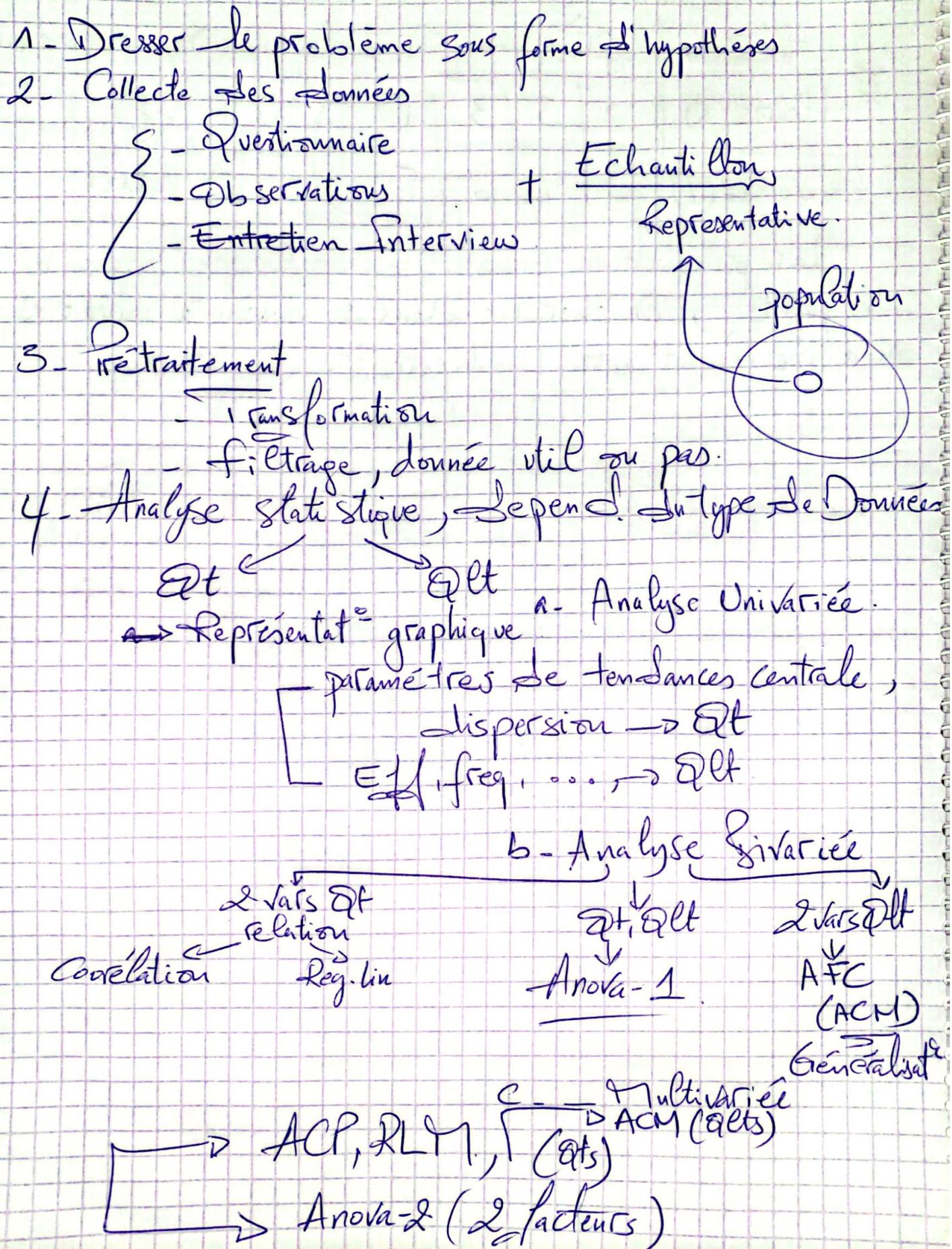
vars qui sont fortement contribuer à une construction
(CPj) (x_1, x_3, α) . \rightarrow CPj non significative
taille \rightarrow age genre $a_1 x_1 + a_2 x_2 + a_3 x_3 + \epsilon$

Interpretation:

- * Signification des composantes
- * Contribution
- * Présenter bien ou pas.

Recap:

Processus



* Les mesures de Tendance Centrale nous donne une idée sur, * Représentation Graphique, distribution, Qu'est ce que nous donne comme une information par rapport au loi normale.

format Asymétrique, ça veux dire quoi?

- La Symétrie \rightarrow Homogénéité ou peut la lire en 2 sens



• écart \uparrow , certain variabilité.

• se comporte à la même manière par rapport à l'homogénéité symétrique

La distribution, et comment les vars se comporte par rapport à la moyenne.

les outlier \rightarrow par rapport à l'I de confiance \bar{x}, σ, \dots

* 1 - valeur de référence
écart type grand \rightarrow la signification petit \rightarrow Distribution

moyenne
médiane.

• Eff fréquence / répond à la Q, Eff n'est plus plus parlant que la fréquence
f_{CC}, f_{CD}, info sur une Catégorie précise +
 \rightarrow regroupement de Catégorie.

d'A.B; Obj: Etudier un Ensemble de vars,
Qu'est ce que représente c'est pas par rapport
à notre sujet
corrélation, liaison entre les vars.
Là Sigmoid, Reg linéaire

C'est quoi? Trouver une
droite qui va ajuster les
données, minimiser l'erreur

+ La prediction, deux vars sont
liées, deux var, on parle de l'effet

Anova I: Obj deux var, on parle de l'effet
on parle sur une var

mesurable l'réaction
uperplan

Lorsqu'on parle sur la minimisation, E ↓
j'essaye de réduire l'erreur

L'équilibre! Anova-simple



Sigma: K-Groupe

$\begin{bmatrix} n \\ K \end{bmatrix}$ au $n \times 3$
effectif K' on aura l'effet de
taille

AFC: Obj: mesurer la correspondance entre
deux vars qt.
modalité i tamasha maa dyal j'en pas?

3 - ACP

[RLY1] entre les variables x du passé
Corrélat entre y et les regressions x
on a une corrélation.

1 ACP

Le Régrageur

Les vars qui caractérisent les individus.

ANOVAbles au lieu de 1 facteur, 2 facteurs

+ d'Interaction]
+ Effet de ①]
+ ————— ②]
+ value → loi de Student

- test l'hypothèse
- La var

F-statistic → Fisher

analyse des résidus - normalité -

3 dim
C1, C2
C1, C3
C2, C3

TP ACP:

AXE 1: Consommation alimentaire selon le cadre socio-professionnel.
— 2: Il caractérise plus la consommation d'innovatives.

Si on a λ_{dim}^2 , on peut pas continuer l'analyse
avoir un λ_{dim}

Axe 1 plus important