

Mon lutin d'Analyse de données

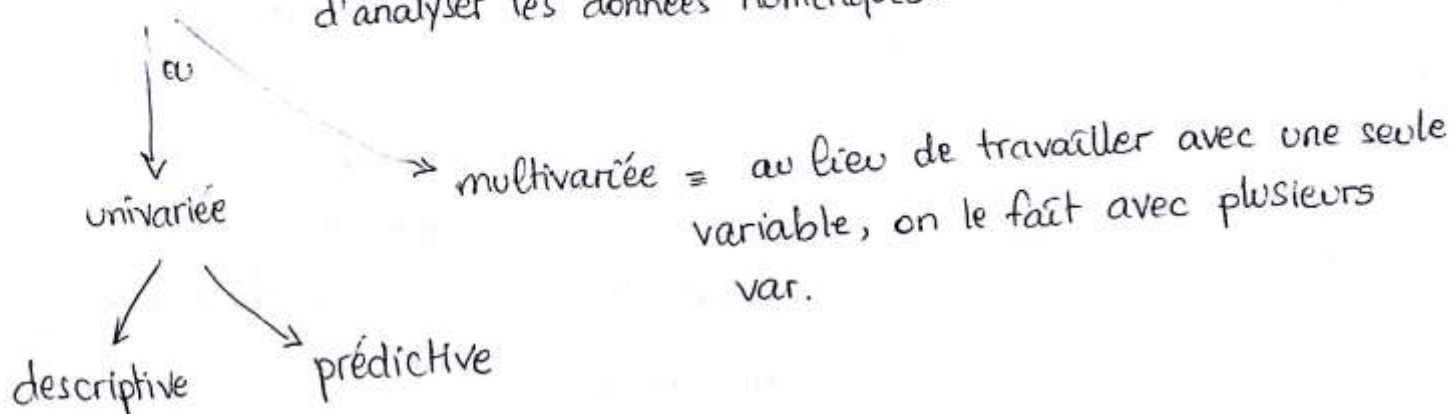
Table des matières

| | |
|-------------------------|----|
| Cours | 4 |
| Exam 2014..... | 25 |
| Corrigé Exam 2014 | 31 |

Analyse de données

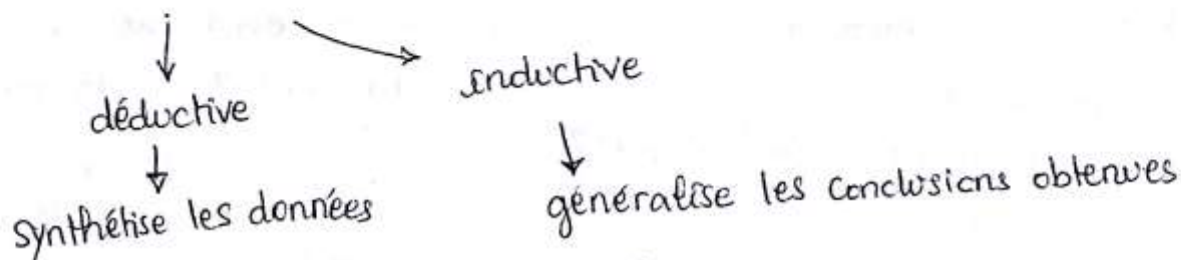
①

Statistique = ens. de méthodes qui permettent de rassembler et d'analyser les données numériques.



• Toute étude statistique passe par ces 5 étapes :

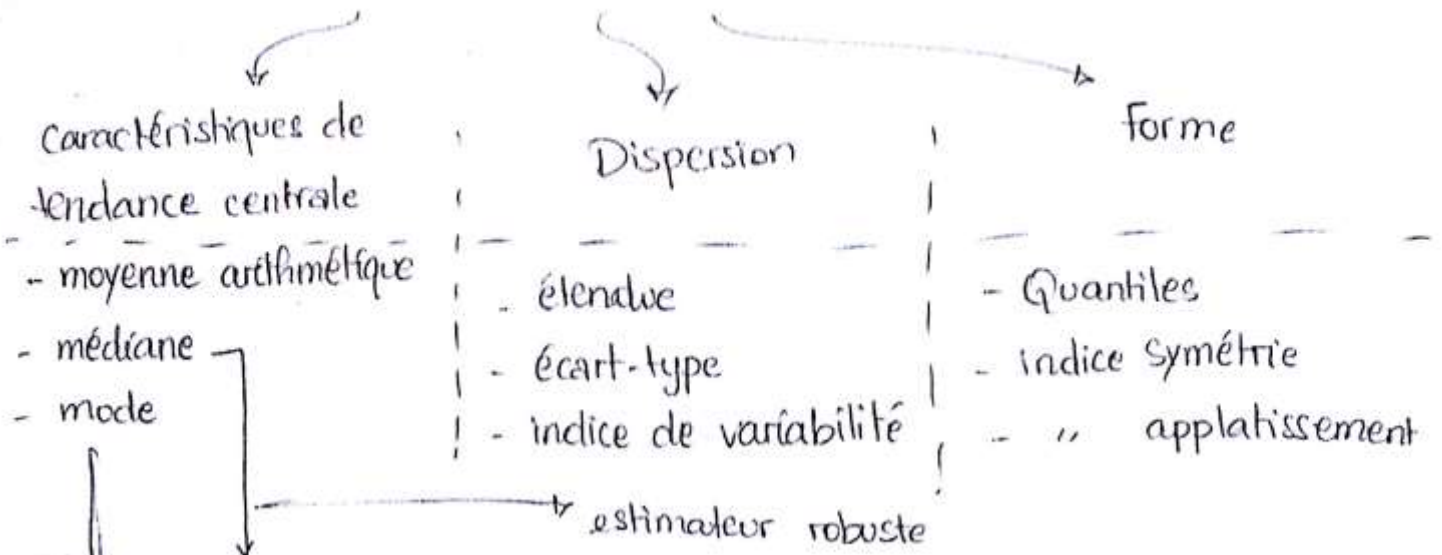
- 1 - Définition de la problématique de l'étude.
- 2 - Collecte de données
- 3 - Préparation des données (données de mauvaise qualité → nettoyer)
- 4 - Analyse statistique



5 - Production et diffusion des résultats

- Données
 - ordinales → peut être classée du plus petit au plus gd.
 - nominales → ne peut pas être ordonnée du + petit au + gd.
→ pas de relation d'ordre
 - continues → intervalles entre les valeurs sont égaux.

c Indicateurs Statistiques



partage les données en deux sous-ensembles de même fréquence.

la val la + fréquente.

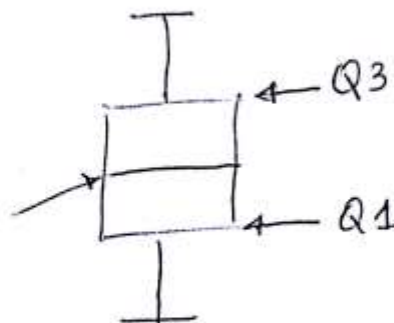
écart-type = en moyenne, de combien chaque score diffère de la moyenne?

étendu = différence entre les valeurs extrêmes.

Notion de boîte à moustaches (Box plot)

= résumé graphique d'une distrib^o

médiane
(n'est pas nécessairement au centre)
de la boîte à moustaches



Notion de quartiles



Q1 = val. en dessous de laquelle se trouvent 25% des observations.

Q3 = val. en dessous de laquelle se trouvent 75% des observations.

le coeff de variat^o

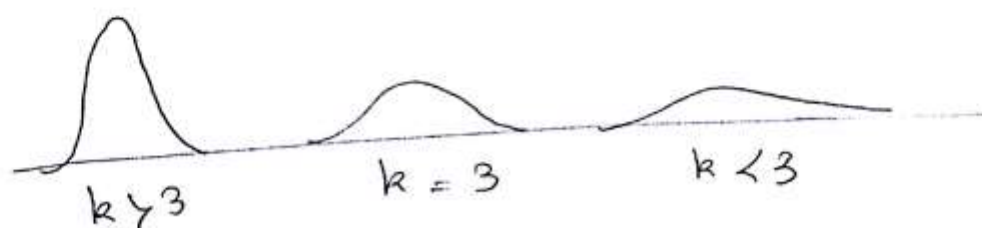
$$CV = \frac{\sigma}{\bar{x}}$$

Distribution normale \rightarrow majorité des sujets sont regroupés de façon symétrique autour de la moyenne. (2)

skewness \rightarrow coeff d'asymétrie (S)

- $S < 0$
distribut°
penche à gauche
- $S = 0$
distribut°
symétrique
- $S > 0$
distribut°
penche à droite

kurtosis \rightarrow coeff d'applatissage (K)



Liaison entre deux var quantitatives

\rightarrow Notion de corrélation (manière dont 2 var quantitatives continues varient simultanément)

\rightarrow Représentat° graphique (nuage de points)

\rightarrow coeff de corrélat° = $\{-1, 1\}$ \Rightarrow relation presque affine entre x et y .
proche de

~ On ne doit pas uniquement se fier à la moyenne. On doit également s'intéresser à l'écart-type par ex, qui exprime bien la différence de variabilité.

21/04/15

Chapitre II

Régression linéaire simple et multiple

- Indice de liaison entre les variables \rightarrow coeff de corrélation, statistique du χ^2
- régression linéaire \rightarrow influence des variables sur une autre et la modéliser (on prédit une ~~var~~) var à partir d'autres var)

Indeed, on a une var endogène (c.à.d la variable à expliquer) et plusieurs var. exogènes (c.à.d les var. explicatives).

Par exemple, on essaie d'expliquer le prix de ventes en considérant plusieurs facteurs (superficie, emplacement, étage, ...)

$$\text{prix vente} = f(+ \text{ facteurs})$$

Généralement, on a :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

var.
endogène
qu'on cherche
à expliquer

coefficients
à estimer

variables
exogènes qui nous aideront
à expliquer

partie aléatoire
qu'on ne peut
pas contrôler
(erreur)

- On s'intéresse à la régression linéaire simple.

C'est-à-dire qu'on a une variable quantitative à expliquer par une seule valeur quantitative.

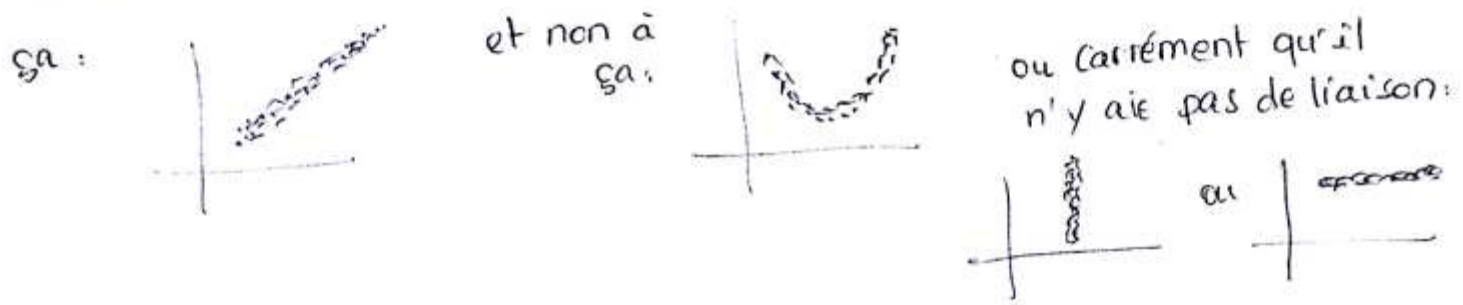
L'équation devient donc,

~~$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$~~

$$y = \beta_0 + \beta_1 x + \varepsilon$$

on cherche à évaluer ces deux constantes

Mais, avant d'essayer de trouver les val de β_0 et β_1 , il faut d'abord vérifier le type de liaison qu'on a et s'assurer qu'on a bel et bien une liaison linéaire (Il faut que l'allure ressemble à



• on aboutit à la conclusion que l'équation de la droite de régression linéaire simple s'écrit comme suit :

$$y_i = \underset{\substack{\uparrow \\ \text{pente}}}{a} x_i + \underset{\substack{\uparrow \\ \text{cte}}}{b} + \underset{\substack{\nwarrow \\ \text{erreur (elle résume toute l'info qu'on n'a pas prise en considération par la régression)}}}{\varepsilon_i}$$

on a défini la droite par la méthode des moindres carrés (la somme des carrés $y_i - \hat{y}_i$ doit être minimale)

on pose plusi + 1^{re} hyp dont :

$$E(\varepsilon_i) = 0, \quad \text{var}(\varepsilon_i) = \sigma^2, \quad \text{cov}(x_i, \varepsilon_i) = 0,$$

| | | |
|---|--|--|
| \Downarrow ε_i sont iid (indépendants et identiquement distribués) | \Downarrow hyp. d'homoscédasticité | \Downarrow erreur indépendante de la var exogène |
|---|--|--|

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \varepsilon_i \equiv N(0, \sigma)$$

coming back to la droite de régression linéaire simple.

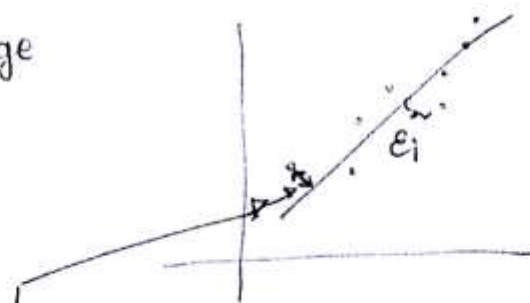
- elle doit approcher au mieux le nuage de points.

- le critère des moindres carrés



minimiser la somme des carrés des écarts entre les vraies valeurs de y et les valeurs prédites.

on minimise $S = \sum \epsilon_i^2$



le fait qu'on ait une régression avec cte \Rightarrow la droite de régression ne passe pas nécessairement par le centre.

on a trouvé que pour déterminer les val de a et b , on utilise les relations suivantes.

$$\begin{cases} \hat{a} = \frac{\text{Cov}(x, y)}{\sigma^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

et: $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + \hat{b})$

• SCR: $SCR = \sum \hat{\epsilon}_i^2$

$SCR = 0 \Rightarrow$ cas parfait (on n'a pas d'erreur)

Mais comme la réalité est loin d'être parfaite, on voudrait savoir à partir de quel moment on pourrait dire qu'une régression est bonne / mauvaise ?

\Rightarrow on décompose la variance de Y .

Q4

④

$$SCT = SCR + SCE$$

indique la variabilité totale de Y (l'information disponible dans les données)

variabilité non expliquée par le modèle (l'écart entre les valeurs observées de Y et celles prédites par le modèle)

indique la variabilité expliquée par le modèle. (variation de Y expliquée par X)

En s'appuyant sur les définitions ci-dessus, il est clair que,

- Puisque SCR indique l'écart entre les valeurs observées et celles prédites, alors si $SCR = 0$, alors l'écart = 0 donc les valeurs observées sont exactement celles prédites.

$SCR = 0 \Rightarrow$ Modèle parfait \Rightarrow Droite de régression passe par tous les points du nuage.

- Puisque SCE indique la variation de Y expliquée par X, alors si $SCE = 0$, cela veut dire que X n'explique rien par rapport à Y.

- Autre notion qu'il faut absolument connaître est celle du coefficient de détermination.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

indique la proportion de la variance de Y expliquée par le modèle

Question standard d'examen:

Interprétation : $R^2 = 81\%$?

\Rightarrow C'est le pourcentage d'information qu'on peut récupérer à partir du modèle de régression.

C'est-à-dire : en ayant x et en utilisant la droite de régression, on peut récupérer 81% de l'info qui réside dans y.

On en déduit donc que, Plus R^2 est proche de 1, mieux c'est car on pourra récupérer $n + 15$ données connaissant X .

Plus R^2 est proche de 0, moins d'info on pourra récupérer.

30/04/15

* Propriété des estimateurs

Il faut vérifier 2 props importantes :

1. est-ce que l'estimateur est sans biais ?

(estimateur sans biais \Rightarrow moy obtenue = vraie valeur)

2. est-ce qu'il est convergent ?

(estimateur convergent \Rightarrow + l'échantillon est gd, + il est précis)

1. estimateur sans biais :

$$E(\text{param}) = \text{param.} \Rightarrow$$

hypothèses à vérifier :

- variable exogène n'est pas stochastique (= n'est pas aléatoire)
- $E(\varepsilon_i) = 0$

2. estimateur convergent :

$$\text{variance} \xrightarrow{+\infty} 0 \Rightarrow$$

hyp. à vérifier

- $\text{var}(\varepsilon_i) = \sigma^2$ (hyp. d'homoscédasticité)
- $\text{cor}(\varepsilon_i, \varepsilon_j) = 0$

Un estimateur précis \Rightarrow variances petites \Rightarrow

- var erreur faible
- dispersion forte
- grand échantillon

Remarque :

Les estimateurs des moindres carrés de la régression sont sans biais et convergents.

- On a calculé la qualité du modèle à partir de (10) l'échantillon (En calculant R^2 ;) , mais il faut s'assurer que c'est le cas pour l'ensemble de la population. ⑤

Pour savoir si la régression est globalement significative, on introduit une nouvelle variable très importante: La statistique F

$$F = \frac{CME}{CMR} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$

$$F = \frac{R^2/1}{\frac{(1-R^2)}{n-2}}$$

La statistique F est utilisée au lieu de R^2 pour indiquer si l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.

Important: $F > F_{1-\alpha}(1, n-2) \Rightarrow$ on est dans la région critique $\Rightarrow H_0$ rejetée

on peut également utiliser la valeur de la p-value (α')

$\alpha' < \alpha \Rightarrow$ Mêmes conclusions as above
(Modèle globalement significatif au risque α)

- Test de significativité de a (Pour vérifier l'influence réelle de X sur Y):

hypothèses: $\begin{cases} H_0: a = 0 \\ H_1: a \neq 0 \end{cases}$ (pente nulle \Rightarrow aucune influence) (*)

si $|t_{\hat{a}}| > t_{1-\frac{\alpha}{2}} \Rightarrow H_0$ rejetée

avec $t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}$

$$\hat{\sigma}_{\hat{a}} = \sqrt{\frac{\hat{\sigma}_e^2}{\sum (x_i - \bar{x})^2}}$$

$$\hat{\sigma}_e^2 = \frac{SCR}{n-2}$$

sur la diapo 12,
on a:
t théorique = 2,306
et
t(\hat{a}) = 5,609
et
t(\hat{b}) = 1,105
t(\hat{a}) > t théorique
donc a \neq 0
t(\hat{b}) < t théorique
donc b = 0

Question d'exam (exemple diapo 22)

Quelle est l'équation de la régression avant le test de significativité ?

$$y = \frac{ax}{0,71} + \frac{b}{4,39}$$

Après le test :

$$y = \frac{ax}{0,71} \quad (\text{On a ignoré le } b \text{ car on a trouvé } b=0)$$

On doit utiliser l'équation après le test. (voir ~~verso~~ feuille précédente (*))

• Supposons qu'on a un échantillon de données et qu'on veut faire la régression. Il faut :

1. vérifier le nuage de points et le coeff de corrélation.
(Pour voir si on a une relation linéaire ou pas. If not, yesni si on n'a pas de relation linéaire, ce n'est pas la peine de continuer because it doesn't make sense whatsoever !)
2. Calculer a et b , la qualité du modèle, l'inférentiel (si a est vraiment $\neq 0$, pareil pour b) pour ainsi voir si ce qu'on est en train de calculer est vrai pour toute la population et non pour l'échantillon uniquement.
+ vérification du modèle après l'avoir généré.

Pour concrétiser tout ce qu'on a vu jusque là, prenons l'exemple qu'on nous a donné sous SPSS. Ouvrez les diapos (à partir de la diapo n° 86).

On a dit qu'il faut tout d'abord commencer par vérifier qu'on a bel et bien une relation linéaire entre nos variables

C'est pour cette raison qu'on cherche à avoir le
diag. de dispersion.

(slide 27)

Nous remarquons que le nuage de points V correspond à celui
de deux variables li binat'hoon relation linéaire.

Donc on peut effectivement définir une droite de régression (slide 28)

(On) Pour vérifier si on a une relat° linéaire entre le prix de vente
et le prix d'achat, on peut également calculer le coeff de
corrélation. C'est justement ce qu'on a fait sur le slide 90.

okay, so now we are dead sure that we can apply ~~had~~ la
régression linéaire simple. La question qui se pose d'aba

hiya : quelle est l'équat° de cette droite ?

Bon, on sait déjà qu'elle sera sous la forme,

$$y = \beta_0 + \beta_1 x + \epsilon$$

On procède à l'estimat° de β_0 et β_1 (par grâce à la figure des
coefficients sur le slide 93. (It's obvious that, vu le fait qu'on
cherche à estimer le prix de vente en fonction du prix d'achat,
 y représente le prix de vente et x le prix d'achat).

Selon le tableau du slide 93, on déduit que: $\beta_1 = 1,775$
et $\beta_0 = -43,615$.

Donc $\text{prix_vente} = -43,615 + 1,775 \text{ prix_achat}$.

À ce stade, nous sommes en mesure d'interpréter la droite
de régression (Question d'examen):

Une augmentation d'une unité du prix d'achat
va conduire en moyenne à une augmentation
de 1,775 unités du prix de vente.

Ensuite, on passe à l'inférence (slide 96).

On a $p\text{-value} = 0,000$ donc $p\text{-value} < 5\%$ which means we're cool.

Ensuite, pour mesurer la qualité du modèle, on avait dit qu'on utilisait le R^2 . Therefore, on a choisi la figure sur le slide 97 pour en extraire la qualité du modèle. Actually, on utilise le R^2 ajusté au lieu du R^2 car il ne dépend pas du nombre de variables. On avait dit que plus R^2 s'approche de 1, mieux c'est. Ici, on a R^2 ajusté = 0,916 so machi ghir mezyane :D ! Qualité du modèle men tiraz raf3!

Enfin, on utilise la table d'Anova. Mais sara7a je ne sais pas comment l'utiliser. Tout ce que je sais, c'est que le F nous permet de savoir si ce qu'on a trouvé est valable pour toute la population. (La table d'Anova nous permet également de confirmer que β_0 et β_1^{ne} sont pas = 0).

Rappel

- vérifier si on a une relat° linéaire.
 - Graphiquement: à partir du nuage de points.
 - Pas " (nsit comment on l'appelle ρ): à partir du coefficient de corrélation (il faut qu'il soit proche de 1).
- estimer les paramètres a et b .
- vérifier si on a un bon modèle (R^2)
- Inférence :
 - statistique t_b ($\neq 0$ par toute la population)
 - a et $b \neq 0$.

qualité du modèle $\Rightarrow R^2$ ajusté

qualité du modèle global $\Rightarrow F$

Comment valider le modèle ?

- relation linéaire entre Y et X .
- normalité de Y .
- l'analyse des résidus
 - normalité de ε .
 - hétéroscédasticité
 - auto-corrélation.

↓

Si on n'arrive pas à trouver une relation linéaire entre Y et X , il faut utiliser un autre type de régression \Rightarrow Faire des transformations sur X et Y afin de trouver une relation linéaire entre les nouvelles valeurs.

ventes = $-12 \times \text{prix} + 1000$
 Lecture en termes d'évolution:
 si le prix augmente de 1€, les ventes vont diminuer de 12 unités.

◦ Supposons qu'on a : $Y = bX^a$ (Modèle log-linéaire)

Obviously, ceci n'est pas une relation linéaire. On doit donc introduire des transformations :

$$e(\log Y = a \log X + B) \rightarrow \begin{aligned} & y = e^B \cdot x^a \\ & \log Y = \alpha, \\ & \log X = \beta \\ & \text{nous donne,} \\ & \alpha = a\beta + B \end{aligned}$$

◦ Modèle exponentiel :

$$Y = e^{ax+b}$$

→ on introduit le log,

$$\log Y = ax + b$$

◦ Modèle logarithmique : $Y = a \ln(X) + b$

vous devez trouver Y , x that's fine ita
ma kanch x unit !

◦ vérification des hypothèses :

- y suit une loi normale

→ Graphiquement, à partir de la distribution

→ Sinon, à partir des indicateurs (Skewness $AS \approx 0$
Kurtosis $AP \approx 3$)

- hyp. sur le terme aléatoire ε .

→ $E(\varepsilon_i) = 0$
→ $V(\varepsilon_i) = \sigma^2$ } homoscedasticité

→ $\text{cov}(x_i, \varepsilon_i) = 0$ → erreur indép.

→ $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ → non auto-corrél^o des erreurs

→ $\varepsilon_i \sim N(0, \sigma)$ → normalité des erreurs.

2x/05/15

⑧

ACP fait partie de la méthode d'apprentissage non supervisé.

↓ but
Avoir des groupes de données
↓ comment?

↓
obj: - structurer les données.
- synthétiser les données
Groupement de données

Projeter les données dans un espace à dimension plus réduite.
we're humans, ghir 3 dimensions et on n'arrive pas à les visualiser clairement. Let alone more!

Mais lorsqu'on projette, il y a de grandes chances que le graphique obtenu ne reflète pas la réalité.

L'analyse factorielle vise donc à représenter les données sur des espaces de dim plus petites et qui soient tout de même pertinents.

ACP est l'analyse factorielle qu'on fait lorsque toutes les données sont quantitatives.

Le 1^{er} axe principale ^{par rapport à l'axe Δ_1} est l'axe qui minimise l'énergie ~~des~~ nuage de points (comme c'est montré sur le slide 22), c'est également un axe qui passe par le centre du nuage de pts.

Un autre objectif est celui de maximiser l'inertie du nuage projeté sur l'axe Δ_1 → pour maximiser l'allongement sur la droite.

• On récupère le max d'infos à travers le 1^{er} axe (Question d'exam)
Par ex, sur la diapo 30, on a trouvé que la 1^{ère} composante principale explique 73,5 % de la variance totale.

Si on n'a pas de corrélation \Rightarrow l'info récupérée sur $\Delta 2$ sera totalement indépendante par rapport à celle de $\Delta 1$.

Pour connaître le pourcentage d'info récupéré sur le 2^{ème} axe, il suffit de calculer $\frac{\lambda_2}{P}$ (but yeah, she'll normally give us all the values we need)

Donc si la valeur propre est très petite, on ne va pas récupérer beaucoup d'informations. C'est l'un des critères qu'on utilise pour déterminer la dimension de notre espace. Indeed, ~~on~~ on trouve qu'il est incensé d'ajouter toute une dimension pour n'en récupérer que 2-3%.

• On suppose que notre espace est de dimension 3. Pour représenter les données, on projette sur l'axe qui correspond au facteur 1 et celui qui correspond au facteur 2 (on ignore le facteur 3). Puis on projette sur l'axe du fact 1 et ^{l'axe du fact} 3, puis sur l'axe du fact 2 et l'axe du fact 3.

Dans chaque repère à 2 dimensions, je fais la projection pour trouver les groupes. Mais est-ce qu'on aura vraiment les mêmes groupes? Probably not!

—
est-ce qu'on peut utiliser l'ACP? \rightarrow Oui, (allées) les données sont toutes de type continu.

- supposons que mes données sont décrites de manière indépendante
 \hookrightarrow pas de redondance \rightarrow on ne peut pas faire une ACP.

- Si dans la matrice de corrélation on a de grandes valeurs, on doit faire de l'ACP car les variables sont fortement corrélées.
- Si la moyenne est presque la même et les écarts sont très différents, \Rightarrow on doit centrer.

Étapes d'une analyse factorielle

1. Examen de la matrice de corrélation.
2. Extraction des facteurs.
3. Transformation par rotation des facteurs.
4. Calcul des scores.

~~[Il faut vérifier la mat de corrélation]~~

L'ACP est un cas particulier de l'analyse factorielle.

1. Au début, il faut vérifier si on a besoin de centrer et réduire. When is that really needed?

Si on n'a pas la même grandeur \rightarrow il faut centrer

Si on n'a pas la même unité \rightarrow il faut réduire.

2. Ensuite, on vérifie la matrice de corrélation (Est-ce qu'il y a une forte corrélation ou pas entre les variables) sur le slide et par exemple, nous avons la matrice de corrélat° qui correspond à notre exemple de voitures.

Pour chaque variable, on cherche ~~celle~~ la variable avec laquelle elle a le plus grand coeff de corrélation (C'est ce qui est entouré en rouge sur les slides).

- On vérifie également la matrice des corrélations, also known as matrice anti-image. Celle-ci supprime les effets linéaires induits par les autres variables.

coeffs proches de 0 \Rightarrow \exists relations transitant par toutes les variables du modèle.

- Ensuite, on effectue le test de KMO pour voir si l'ACP est pertinente ou pas.

\Rightarrow test KMO

0 \leftarrow \downarrow \leftarrow 1 parfait, on va tout réduire
Ce n'est même pas la peine de faire ACP, ach had elfdi7a!

si KMO est proche de 0,7, on dit que c'est bon!

- Après, on vérifie les MSA_i . Ce sont des variables de corrélation qui se trouvent sur la diagonale de la matrice anti-image. (voir diapo 100 for instance)

Les valeurs qui ont une grande valeur MSA_i vont participer à la construction des facteurs.

- Then, on effectue le test de sphéricité de Bartlett. C'est un test statistique qui sert pour prouver que la matrice de corrélation est vraiment différente de la matrice d'identité. (You know, puisque la mat. de corrélat° est calculée à partir d'un échantillon, il se peut qu'il y ait des fluctuations qui font que l'échantillon n'est pas fiable).

Une fois on a fini tout ce calvaire et on s'est assuré (10) que (100) all is fine, il faudra déterminer combien d'axes on va garder.

Ce qu'il faut voir c'est pour chaque axe, combien on va récupérer.

Par exemple sur le graphe slide 103, on remarque sur le graphique des valeurs propres qu'à partir de la 4^{ème} valeur propre, le gain est minime.

on peut également se référer au tableau slide 105, où l'on remarque que le 1^{er} axe explique 77% de la variance totale.

Au bout de 3 axes, la variance (cumulée) totale est de 96,7% so we don't really need to add a new dimension.

Pour ce qui est du tableau de corrélations reproduites, on doit lire le pourcentage de résidus. (résidu minimal)

Il faudra trouver un équilibre entre les 3 (pourcentage d'infos, valeurs propres, résidus).

01/06/15

Rappel

Étapes de l'Analyse Factorielle

objectif: éliminer la redondance entre plusieurs variables

- KMO \rightarrow calculé à partir de la matrice de corrélation et de la matrice anti-image.
- Sphéricité \rightarrow si mat. vraiment \neq Id
- MSA_i \rightarrow diagonale de la mat. anti-Img
- pourcentage des résidus \rightarrow du à partir de la mat anti corrélation.

Qualité de représentation

| var | initial | extraction |
|-----|---------|------------|
| | | |

\rightarrow c'est le pourcentage d'info que je peux récupérer si je fais une projection sur l'axe.

si on a des variables fortement corrélées avec le 1^{er} axe (comme c'est le cas sur le slide 107), on ne peut interpréter les autres axes \rightarrow on fait donc des rotations.

sur chaque ligne, on cherche la valeur maximale.

L'axe sera donc représenté par cette ~~(valeur)~~ variable dont la valeur est maximale sur l'axe en question.

facteur ss val de corrélation max \Rightarrow on doit faire une rotation.

Si on avait sur la matrice des coefficients du slide 115 la valeur $-0,814$ pour la vitesse, cela voudrait dire que la valeur de ma vitesse va diminuer.

Il faut interpréter les variables qu'on a à la fin de la matrice des coeffs.

Si un axe est représenté par une seule var \Rightarrow non, on ne doit pas l'éliminer. It simply means that la variable est en question est indépendante.

All in all, il faut :

- spécifier si l'ACP est pertinent ou pas (4 paramètres)
- trouver les facteurs et ceux qu'on doit garder (3 things)
- si on aura une rotation ou pas.
- coordonnées et projection dans le nouvel espace.

Et puis à l'examen, pour une figure telle celle sur le slide 116, il faudra spécifier quels facteurs on a sur les axes.

Examen

Nom et Prenom :

Année Universitaire : 2013 - 2014

Date : 21/05/2014

Filière : Ingénieur

Durée : 90 min

Semestre : S4

Période : P2

Module : M4.6 - Management de la Donnée

Elément de Module : M4.5.1 - Analyse de Données

Professeur : H. Benbrahim

Consignes aux élèves ingénieurs:

- Une feuille A4 recto verso de synthèse est autorisée.
- Toute tentative de fraude sera sanctionnée par la note **zéro**.
- La clarté et la simplicité des réponses est obligatoire.
- **IL FAUT TOUJOURS SPECIFIER LE NUMERO DE LA FIGURE QUE VOUS AVEZ UTILISE POUR DONNER OU JUSTIFIER LE RESULTAT.**
- **IL Y A DES FIGURES QUI MANQUENT (2 OU 3 OU 4 FIGURES). C'EST FAIT EXPRES. SI VOUS NE TROUVEZ PAS UNE FIGURE, MENTIONNEZ QUE VOUS AVEZ BESOIN DE TEL OU TEL FIGURE.**

Un magazine français a publié un comparatif des 12 principaux Smartphones disponibles sur le marché en novembre 2010. Chaque Smartphone est décrit et évalué par les points suivants :

- **Prise en main** : facilité de prise en main (note sur 20)
- **Communication** : qualité des communications téléphoniques (note sur 20)
- **Organisation** : fonctionnalités d'organisation : agenda, carnet d'adresses, etc. (note sur 20)
- **Divertissement** : offre en divertissement : jeux, etc. (note sur 20)
- **Navigation** : offre en logiciel de navigation : Maps, GPS, etc. (note sur 20)
- **Prix** : prix public hors abonnement (en euros)
- **Autonomie** : autonomie en communication (en heures)

Source : Le Point no 2045, 11 novembre 2010 : Le guide du numérique 2011.

| Marque | Modèle | PriseEnMain | Communication | Organisation | Divertissement | Navigation | Prix | Autonomie |
|---------------|----------------|-------------|---------------|--------------|----------------|------------|------|-----------|
| Apple | iPhone4 | 17 | 14 | 17 | 17 | 17 | 630 | 10 |
| Samsung | GalaxyS | 18 | 16 | 16 | 13 | 15 | 599 | 6 |
| Sony-Ericsson | XperiaX10 | 18 | 14 | 16 | 15 | 14 | 600 | 5 |
| Samsung | Omnia7 | 18 | 16 | 13 | 15 | 14 | 600 | 6 |
| LG | Optimus7 | 15 | 16 | 15 | 15 | 14 | 499 | 5 |
| Nokia | N8 | 14 | 14 | 16 | 15 | 15 | 549 | 7 |
| Motorola | MilestoneXT720 | 13 | 15 | 16 | 15 | 13 | 399 | 8 |
| RII | Torch9800 | 13 | 18 | 14 | 13 | 13 | 569 | 12 |
| HTC | Desire | 14 | 14 | 17 | 12 | 14 | 430 | 6 |
| ACER | Stream | 15 | 14 | 15 | 12 | 14 | 500 | 5 |
| Samsung | Wave | 15 | 15 | 14 | 13 | 12 | 499 | 5 |
| HP | PalmPré2 | 13 | 15 | 13 | 14 | 10 | 449 | 4 |

Partie I: Statistique Descriptive:

I-1 Quelles est, en moyenne, la caractéristique la mieux notée ? La moins bien ?

I-2 Par rapport à quelle caractéristique les smartphones étudiés différent-ils le plus ? Calculer les indicateurs nécessaires.

I-3 Etudier la boîte à moustache de l'attribut prix des smartphones.

Partie II: Régression Linéaire Simple:

On aimerait construire une droite de régression pour expliquer le prix des smartphones en fonction de toutes les autres variables.

2.1 Y a-t-il une liaison linéaire entre le prix et les autres variables? Justifiez :

La régression pas à pas a convergé et a aboutit au modèle 2 donné en Annexe.

2.2 Quelle est la qualité du modèle ? et quelle est son interprétation ?

2.3 Expliquez à un enfant de 12 ans l'apport de ce résultat exprimé en terme du prix du smartphone.

2.4 La statistique F (D dans la table d' ANOVA) pour la régression est égale à 15,275. Expliquez précisément quelle est l'hypothèse testée et quelle est la conclusion ?

2.5 Donnez l'équation de la droite avant justification de la significativité :

2.6 En considérant un risque d'erreur de 5%, cette régression est-elle significative?
Donner la nouvelle équation de la régression, justifier.

2.7 Analyser les résidus et vérifier si les hypothèses de validation du modèle de régression sont vérifiées en justifiant par les différentes figures données en annexe.

2.8 Quelle est le prix possible pour un smartphone qui a les caractéristiques suivantes :

| | | | |
|--------------------|--------------------|-------------------|---------------------|
| Prise en main = 15 | communication = 16 | Organisation = 15 | Divertissement = 15 |
| Navigation = 14 | Prix = 499 | Autonomie = 5 | |

Partie III : Analyse en Composantes Principales

3.1 Les opérations de centrage et de réduction sont-elles nécessaire ? Justifier

3.2 L'analyse factorielle est-elle pertinente ? justifier votre réponse en utilisant tous les critères possibles.

3.3 En effectuant une analyse bivariée sur les variables, proposez un scénario possible de regroupement de variables. On vérifiera par la suite si ce regroupement est maintenu ou pas après l'extraction des facteurs.

3.4 L'extraction de facteurs a été faite à l'aide de l'Analyse en composantes principales

a) Quels sont les pourcentages d'inerties expliquées par l'ACP obtenue par l'extraction de :

- 2 facteurs

- 3 facteurs

- 4 facteurs

- 5 facteurs

b) quels nombres de facteurs retenez-vous? justifiez avec toutes les justifications possible.

3.5 Pour des raisons quelconques, 3 facteurs ont été retenus.

a) Est-il nécessaire de faire une rotation? Justifiez :

b) donnez une interprétation de ses facteurs (sans ou avec rotation)

3.6 Quelle sont les groupes qu'on peut générer après cette ACP ?

3.7 Comment peut-on décrire les smartphones :

Iphone4 :

Torch9800 :

Omnia7 :

Desire :

Statistiques descriptives

| | N | Minimum | Maximum | Moyenne | Ecart type |
|---------------------|----|---------|---------|---------|------------|
| PriseEnMain | 12 | 13 | 18 | 15,33 | 2,015 |
| Communication | 12 | 14 | 18 | 15,17 | 1,267 |
| Organisation | 12 | 13 | 17 | 15,17 | 1,403 |
| Diversissement | 12 | 12 | 17 | 14,08 | 1,505 |
| Navigation | 12 | 10 | 17 | 13,75 | 1,712 |
| Prix | 12 | 399 | 630 | 526,08 | 75,150 |
| Autonomie | 12 | 4 | 12 | 6,42 | 2,314 |
| N valide (listwise) | 12 | | | | |

Figure 1

Figure 2

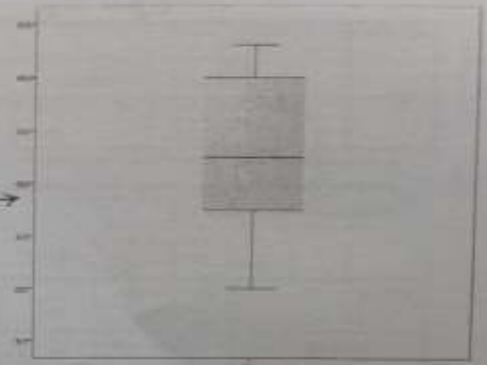


Figure 3

Corrélations

| | Prix | PriseEnMain | Communication | Organisation | Diversissement | Navigation | Autonomie |
|-------------------------------|-------|-------------|---------------|--------------|----------------|------------|-----------|
| Corrélation de Pearson | | | | | | | |
| Prix | 1,000 | ,756 | ,060 | ,051 | ,337 | ,311 | ,416 |
| PriseEnMain | ,756 | 1,000 | -,227 | ,133 | ,230 | ,553 | -,130 |
| Communication | ,060 | -,227 | 1,000 | -,033 | -,151 | -,398 | -,346 |
| Organisation | ,051 | ,133 | -,033 | 1,000 | ,195 | ,700 | ,173 |
| Diversissement | ,337 | ,230 | -,151 | ,195 | 1,000 | ,397 | ,198 |
| Navigation | ,311 | ,553 | -,398 | ,700 | ,397 | 1,000 | ,418 |
| Autonomie | ,416 | -,130 | -,346 | ,173 | ,198 | ,418 | 1,000 |
| Sig. (unilatérale) | | | | | | | |
| Prix | | ,002 | ,427 | ,432 | ,131 | ,317 | ,089 |
| PriseEnMain | ,002 | | ,229 | ,333 | ,236 | ,031 | ,344 |
| Communication | ,437 | ,228 | | ,014 | ,320 | ,100 | ,135 |
| Organisation | ,438 | ,333 | ,014 | | ,334 | ,006 | ,290 |
| Diversissement | ,101 | ,238 | ,320 | ,304 | | ,101 | ,260 |
| Navigation | ,017 | ,031 | ,100 | ,005 | ,121 | | ,088 |
| Autonomie | ,089 | ,344 | ,135 | ,293 | ,229 | ,088 | |

Regression Prévision standardisée

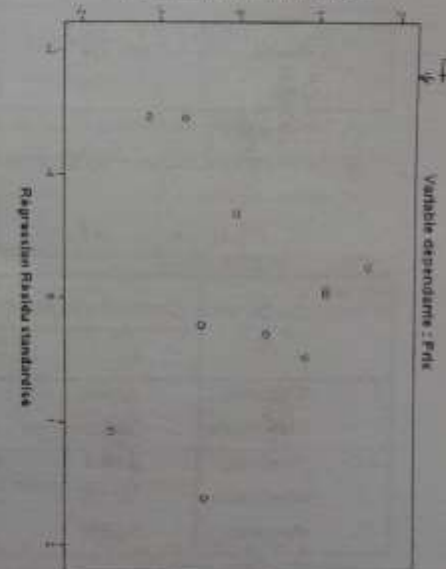


Figure 4

Figure 5

Modèle régressif avec statistiques

| Modèle | R | R ajusté | R ajusté corrigé | Erreur standard de l'estimation | Changement dans les statistiques | | | | Durée de l'ajustement |
|--------|-------------------|----------|------------------|---------------------------------|----------------------------------|-----------------------|-------|-------|-----------------------|
| | | | | | Variation de R | Variation de R ajusté | durée | durée | |
| 1 | ,756 ^a | ,572 | ,520 | 54,574 | ,572 | 13,358 | 1 | 10 | 2014 |
| 2 | ,817 ^a | ,641 | ,616 | 44,137 | ,348 | 15,275 | 1 | 6 | 2014 |

a. Valeurs prédites : (constantes), PriseEnMain

b. Valeurs prédites : (constantes), PriseEnMain, Autonomie

c. Variable dépendante : Prix

Figure 6

ANOVA^a

| Modèle | | Somme des carrés | ddl | Moyenne des carrés | F | Sig. |
|--------|------------|------------------|-----|--------------------|--------|-------------------|
| 1 | Régression | 35574,480 | 1 | 35574,480 | 13,256 | ,004 ^a |
| | Réidu | 2058,437 | 10 | 205,844 | | |
| | Total | 62122,917 | 11 | | | |
| 2 | Régression | 52261,297 | 2 | 26130,649 | 23,848 | ,000 ^b |
| | Réidu | 985,615 | 9 | 109,523 | | |
| | Total | 62122,917 | 11 | | | |

a. Valeurs prédites : (constantes), PriseEnMain

b. Valeurs prédites : (constantes), PriseEnMain, Autonomie

c. Variable dépendante : Prix

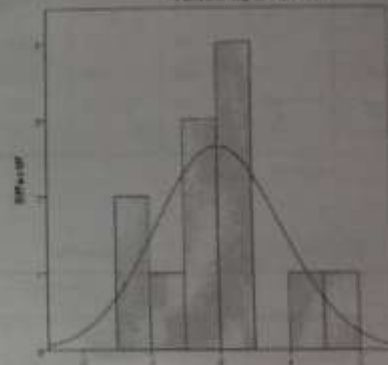
Figure 8

Coefficients^a

| Modèle | | Coefficients non standardisés | | Coefficient standardisés | t | Sig. |
|--------|-------------|-------------------------------|-----------------|--------------------------|-------|------|
| | | B | Erreur standard | | | |
| 1 | (Constante) | 82,590 | 119,257 | | ,785 | ,450 |
| | PriseEnMain | 28,201 | 7,717 | ,756 | 3,655 | ,004 |
| 2 | (Constante) | 84,317 | 65,298 | | ,838 | ,541 |
| | PriseEnMain | 30,730 | 4,595 | ,824 | 6,154 | ,000 |
| | Autonomie | 18,590 | 4,740 | ,521 | 3,938 | ,004 |

a. Variable dépendante : Prix

Variable dépendante : Prix



Source : 1/2014
Durée : 1000

Figure 7

Regression Residu standardisés

Diagramme graphique P-P de régression de résidu standardisés

Variable dépendante : Prix

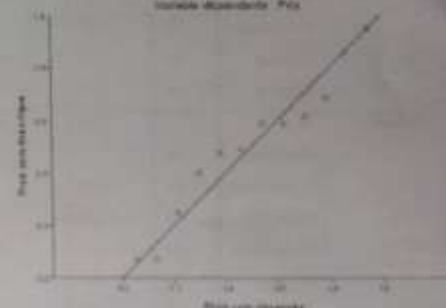


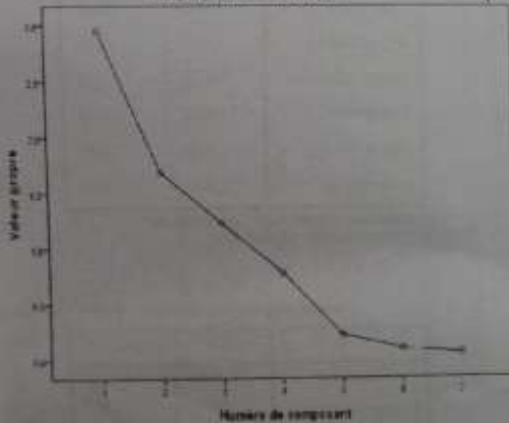
Figure 5

Figure 10

| Composante | Variance totale expliquée | | | | | | | | |
|------------|---------------------------|------------------|-----------|--|------------------|-----------|--|------------------|-----------|
| | Valeurs propres initiales | | | Extraction Somme des carrés des facteurs retenus | | | Somme des carrés des facteurs retenus pour la rotation | | |
| | Total | % de la variance | % cumulée | Total | % de la variance | % cumulée | Total | % de la variance | % cumulée |
| 1 | 2,947 | 42,098 | 42,098 | 2,947 | 42,098 | 42,098 | 2,314 | 32,052 | 32,052 |
| 2 | 1,680 | 24,124 | 66,222 | 1,680 | 24,124 | 66,222 | 2,050 | 28,201 | 60,253 |
| 3 | 1,225 | 17,500 | 83,722 | 1,225 | 17,500 | 83,722 | 1,497 | 21,378 | 81,631 |
| 4 | ,779 | 11,124 | 94,846 | | | | | | |
| 5 | ,218 | 3,085 | 97,931 | | | | | | |
| 6 | ,085 | 1,204 | 99,135 | | | | | | |
| 7 | ,060 | ,708 | 100,000 | | | | | | |

Méthode d'extraction : Analyse en composantes principales.

Graphique de valeurs propres

Matrice des composantes^a

| | Composante | | |
|----------------|------------|-------|-------|
| | 1 | 2 | 3 |
| PriseEnMain | ,889 | ,088 | -,662 |
| Communication | -,437 | ,829 | ,072 |
| Organisation | ,828 | -,566 | ,447 |
| Diversissement | ,550 | ,163 | ,030 |
| Navigation | ,937 | -,052 | ,186 |
| Prix | ,755 | ,536 | -,284 |
| Autonomie | ,350 | ,597 | ,683 |

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites.

Matrice des composantes après rotation^a

| | Composante | | |
|----------------|------------|-------|-------|
| | 1 | 2 | 3 |
| PriseEnMain | ,918 | ,109 | -,279 |
| Communication | -,680 | -,840 | ,413 |
| Organisation | ,846 | ,834 | ,191 |
| Diversissement | ,468 | ,207 | ,262 |
| Navigation | ,809 | ,643 | ,363 |
| Prix | ,924 | -,050 | ,285 |
| Autonomie | ,126 | -,004 | ,964 |

Méthode d'extraction : Analyse en composantes principales.

Méthode de rotation : Varimax avec normalisation de Kaiser.

a. La rotation a convergé en 5 itérations.

Figure 14 → Corrélations reproduites pour 3 facteurs

b. Les résidus sont calculés entre la covariance observée et la covariance reproduite. Il y a 6 (28,0%) résidus non redondants avec des valeurs absolues supérieures à 0,05.

Figure 15 → Corrélations reproduites pour 4 facteurs

b. Les résidus sont calculés entre la covariance observée et la covariance reproduite. Il y a 2 (9,0%) résidus non redondants avec des valeurs absolues supérieures à 0,05.

Figure 16 → Matrice des composantes^a

| | Composante | | | |
|----------------|------------|-------|-------|-------|
| | 1 | 2 | 3 | 4 |
| PriseEnMain | ,628 | ,068 | -,662 | -,194 |
| Communication | -,422 | ,828 | ,072 | -,003 |
| Organisation | ,828 | -,566 | ,447 | -,141 |
| Diversissement | ,550 | ,163 | ,030 | ,016 |
| Navigation | ,937 | -,052 | ,186 | -,145 |
| Prix | ,755 | ,536 | -,284 | -,108 |
| Autonomie | ,350 | ,597 | ,683 | -,129 |

Méthode d'extraction : Analyse en composantes principales.

a. 4 composantes extraites.

Figure 17 → Indice KMO et test de Bartlett

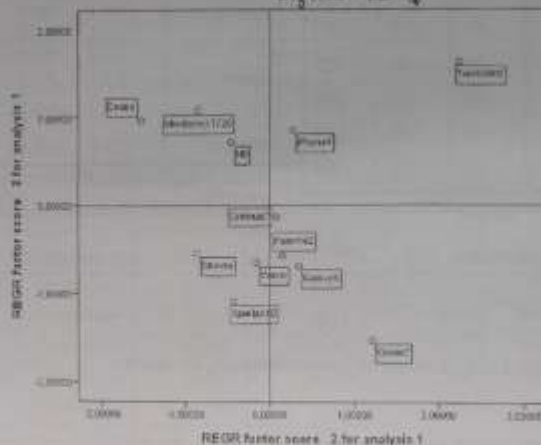
| | | |
|---|---------------------------|--------|
| Méthode de sélection de l'échantillonnage de Kaiser-Meyer-Olkin | | ,878 |
| Test de sphéricité de Bartlett | Khi-deux approché | 41,782 |
| | ddl | 21 |
| | Signification de Bartlett | ,004 |

Figure 18 → Qualité de représentation pour 3 facteurs

| | Initial | Extraction |
|----------------|---------|------------|
| PriseEnMain | 1,000 | ,935 |
| Communication | 1,000 | ,883 |
| Organisation | 1,000 | ,811 |
| Diversissement | 1,000 | ,330 |
| Navigation | 1,000 | ,915 |
| Prix | 1,000 | ,939 |
| Autonomie | 1,000 | ,945 |

Méthode d'extraction : Analyse en composantes principales.

Figure 19



Qualité de représentation pour 4 facteurs

| | Initial | Extraction |
|----------------|---------|------------|
| PriseEnMain | 1,000 | ,972 |
| Communication | 1,000 | ,890 |
| Organisation | 1,000 | ,931 |
| Diversissement | 1,000 | ,998 |
| Navigation | 1,000 | ,837 |
| Prix | 1,000 | ,950 |
| Autonomie | 1,000 | ,962 |

Méthode d'extraction : Analyse en composantes principales.

Exam 2014

- Même si la réponse est fausse, si vous avez indiqué que vous avez utilisé la figure n° kda dans votre raisonnement et si c'est effectivement la figure qu'on devra utiliser, vous aurez quand même un petit 0,25. Better than nothing!
- Il y a des réponses dont les figures ne sont pas représentées dans l'annexe, so if you don't find them, it doesn't mean that you're wrong :D! (Par ex, quand $\text{fact} < \text{it} \times \alpha$) peut-on procéder à l'ACP? On a besoin de kda mais sa fig. isn't available).

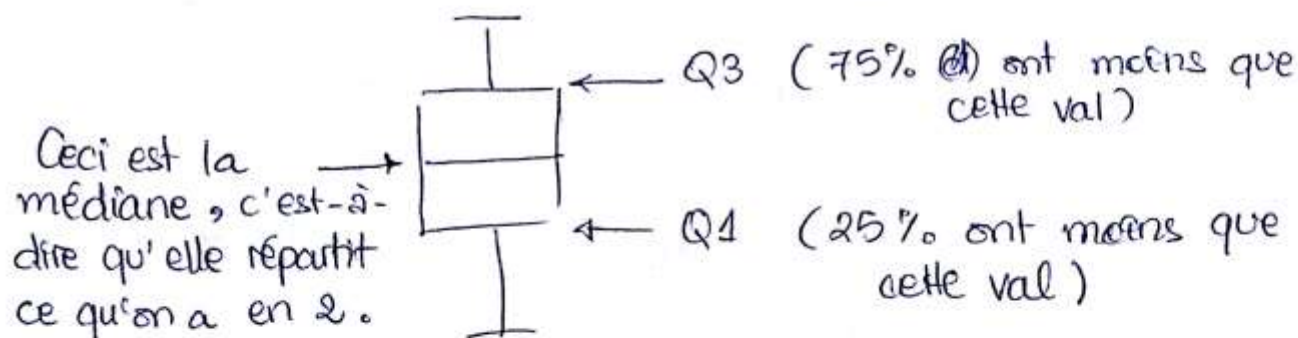
Exam des smartphones:

- I-1. Il suffit de calculer la moyenne des notes de chaque ligne et (d'en) de définir la variable avec la plus grande moyenne et celle avec la plus petite moyenne.
- I-2. Puisqu'on a des unités différentes (heures, euros, notes) \rightarrow on calcule les coeff's de variation de chacun and then we talk.
- I-3. Les infos qu'on peut tirer d'une boîte à moustaches $\rightarrow Q_1, Q_2, Q_3 \rightarrow$ moy. et quartiles.
- \hookrightarrow prix ds quel intervalle.
 - \hookrightarrow sym ou pas? (Sorry, can't remember what I meant by sym)
 - \hookrightarrow La moustache d'en bas est plus grande que celle d'en haut.

Aah! Je viens de me rappeler de ce que veut dire sym but I'm too lazy to cross it.

Sym veut dire symétrique.

Au fait, je ne me rappelle plus si je vous l'ai déjà dit mais une boîte à moustaches means this:



(what is above)

(50%) Exactly 50% ont plus que cette val et l'autre moitié est inférieure à cette val.

(Going back to) Plus la boîte à moustaches est étendue, plus l'étendue des valeurs (max - min) est large. So yeah, I remember dans un TP, on avait à comparer entre 2 boîtes à moustaches (I lost that TP's papers, I'm such a mess sometimes --) et l'une des remarques qu'on avait faites était le fait qu'une boîte à moustaches était beaucoup plus large que l'autre, cela voulait simplement dire que les val sont plus "étendues" (Can't remember AGAIN le terme qu'on utilise pour ça)

MaBlinach, on ferme cette grande parenthèse à propos de la boîte à moustaches.

Dans notre exemple, on remarque que Q_2 n'est pas au centre de la boîte à moustaches \Rightarrow This n'est pas symétrique!

Partie II

2.1. Corrélation multiple \Rightarrow on regarde la corrélation ~~de chaque~~ ^{du} prix avec chaque variable.

on a besoin de la matrice de corrélation et du graphe.

Pour la variable de communication, et comme vous pouvez le remarquer sur la matrice de corrélation (Figure 3), le coeff de corrélation = 0,06.

D'autre part, le test de signifi. sur le même tableau pour la var. comm. est: $\alpha = 0,427$.

Or $0,427 > 0,05 \rightarrow$ donc l'hypothèse nulle est vérifiée.
 \Rightarrow corrélation = 0

Donc la communication ne va pas influencer le prix.

2.2. Pour mesurer la qualité du modèle, on se réfère à la variable R^2 ajusté (sur la Figure 5).

R^2 -ajusté du modèle 2 (on a pris celle du modèle parce que c'est ce qui a été dit dans l'énoncé) est égal à 0,806.

Interprétation? \rightarrow ça veut tout simplement dire que 80,6 est le pourcentage d'informations qu'on peut récupérer.

2.3. Inutile d'essayer d'expliquer au petit enfant en introduisant les notions de pourcentage. That's too complicated for the kind. Une bonne réponse serait comme suit :

On peut déduire le prix du téléphone à partir des

options / caractéristiques qu'on a.

(43)

2.4. hyp testée \rightarrow ? (kayna fle cours)
zerbat zliya thadi : p
Que fait F ? \rightarrow test de significativité globale.
(est-ce que J vrmt $\neq 0$)

2.5. équation de la droite \rightarrow Mnine ghanjibouha?

\Rightarrow Table des coefficients (figure 8)
nous donne (Regardez le modèle 2, la 1^{ère} colonne)

$$\text{prix } y = 54,317 + 30,739 \times \text{prise en main} + 16,999 \times \text{Autonomie.}$$

2.6.
Après le test de sig. (dernière colonne du même tableau)
on remarque que $\text{sig}(cte) = 0,541 > 0,05$, donc on
ne va pas la garder!

La nouvelle (dron) équation de la régression est désormais,

$$\text{prix } y = \textcircled{X} 30,739 \times \text{prise en main} + 16,999 \times \text{Autonomie.}$$

2.7. il faut vérifier les hypothèses ~~des~~ suivantes,

- normalité des résidus
 - " " y
 - autoscédasticité
 - non autocorrélation
 - distrib^o de l'erreur
 - " " y
- } normale

et ce à partir
des graphes
(Il faut spécifier
les graphes..
But I don't know'em
x(C)

2.8. On a l'équation est.

$$\text{prix } x = \text{prise en main} \times 30,739 + 16,999 \times \text{Autonomie}$$

donc on ne prendra en compte que les caractéristiques, prise en main et autonomie.

$$\text{Prix} = 15 \times 30,439 + 5 \times 16,999$$

Partie III

3.1. Ces opérations sont nécessaires car les unités sont différentes.

3.2. On doit utiliser,

- mat. corrélat°
- KMO
- sphéricité de Bartlett.

3.3. La matrice de corrélation nous dira s'il existe une forte corrélation entre les variables.


3.4. Nombre de facteurs à retenir \Rightarrow var. totale expliquée.

on prendra en compte:

- % d'info.
- diag. des val. propres
- corrélat° reproduite.
- qualité de représentation.

Petite remarque :

Normalement, on prendra 4 facteurs. Dans la question suivante, on nous dit qu'on va travailler avec 3 facteurs. Haaanya, machi nchoufouha elle a dit 3 on remet en question notre raisonnement. Don't follow like a sheep.

okay? okay 

3.5. a. Est-ce que la rotation est nécessaire veut dire (14)
est-ce qu'on arrive à interpréter tous les facteurs?
Oui, donc la rotation n'est pas nécessaire.

b. Interprétation → on va utiliser la matrice des
composantes (Figure 12)

Pour chaque variable, on va déterminer la valeur max.

| | |
|---------------|------------------------|
| prise en main | → 1 ^{er} axe |
| comm | → 2 ^{ème} axe |
| orga | → 1 ^{er} |
| Diver | → 1 ^{er} |
| Navig | → 1 ^{er} |
| prix | → 1 ^{er} |
| Autonomie | → 3 ^{ème} |

Le 1^{er} axe est donc représenté par la prise en main,
l'orga, diver, ...

on remarque que la comm. est toute seule, donc elle est
indépendante des autres.

3.7 → outliers → Torch et Omnia.
et sur l'axe 1, (les ~~autres~~)



pour
les téléphones
qui se trouvent ici,

on ne doit pas dire qu'ils ont une grande valeur par
rapport à l'axe 1 mais il faut plutôt dire de quel
facteur il s'agit. Il faut aussi voir si la val +/-
pour voir si la var va augmenter/diminuer.