I*Etapes d'une étude statistique:1- définition de problématique **2-** Collecte des données **3-** Préparation des données **4-** Analysestatistique **5 -**Production et diffusion des résultats.

*Analyse « déductive » descriptive: a pour but de synthétiser et de représenter les données observées pour que l'on puisse prendre des décisions facilement par des illustrations graphiques.

*Analyse « inductive » inférence: permet de généraliser et d'étendre dans certaines conditions les conclusions obtenues.Cette phase comporte un certain risque d'erreur.

Enquête: Ensemble des opérations qui ont pour but de collecter de façon organisée des informations relatives à un groupe d'individus ou d'élément observés dans leur milieu. Recensement : toutes les unités de la population sont observées. Echantillonnage : Une partie de la population est observé.La précision d'une enquête dépond : de la taille de l'échantillon et l'homogénéité de la population.

Echelle nominale: Catégoriale. Echelle ordinale: catégoriale +ordre.Quantitatives: Discrète : Ne peut prendre qu'un ensemble limité de valeurs souvent entières.Continue : peut prendre toutes les valeurs d'un intervalle fini ou infini.
Qualitatives: diagramme sectoriel, diagramme en bâton et courbes cumulées pour les variables ordinales.
Quantitatives: Discrète: diagramme en bâton/diagramme cumulatif. Continue: histogrammes/courbes cumulés.

1-Caractéristiques de tendances centrales :

Mode: la valeur du caractère la plus fréquente. Une série peut être bimodale ou même multimodale.
Moyenne: la valeur centrale la plus utilisées, calculable sur les variables quantitatifs.
Médiane: la médiane Q2 est la valeur de la variable statistique discrète x située au milieu du classement ordonné dans le sens croissant des valeurs de x. Permettant de partager la population en 2 sous population égales (50%).
Valeurs distinctes: 1-Classer la série 2- déterminer si elle contient un nombre pair ou impair
Valeurs répétées: on utilise les effectifscumulés croissant.
Variable continue: oneffectue une interpolation linéaire.

2-Mesures de dispersion :

Etendu: c'est la différence entre les valeurs extrêmes du caractère x : e = max(xi) –min (xi) (ordinale, échelle)
Ecart interquartile: Iq= Q3–Q1 contient les 50% des valeurs les plus centrales Q1(25%) et Q3(75%) (Ordinale,échelle)
Variance : $S^2 = 1/N \sum_{i=1}^{k} n_i \quad (x_i - \bar{x})$sert à caractériser l'écart plus ou moins important de l'ensemble des valeurs par rapport à la moyenne. « Degré de variabilité »
Ecart-type: σ=√S^2.Des données présentent une forte variabilité si l' écart-type est grand par rapport à la moyenne.Les coefficients de variation : cv=écart-type/moyenneon ne peut pas comparer la dispersion de 2 séries statistiques qui sont exprimées dans des unités différentes. Cv < 0.15 n'est pas significatif la moyenne est suffisante.
Boite à moustache : un rectangle d'extrémités Q1et Q3où la médiane est représentée par un trait horizontal à l'intérieur de laboîte,Les 2 moustaches débutent en Q1et Q3 se terminent en la valeur adjacente inférieure et supérieurerespectivement. -L'extrémité de la moustache inférieure est la valeur minimum dans les données qui est supérieure à la valeur : Q1 -1,5*(Q3-Q1).-L'extrémité de la moustache supérieure est la valeur maximum dans les données qui est inférieure àla valeur : Q3 +1,5*(Q3-Q1).Les valeurs xi éloignées ou «atypique» vérifiant:xi > Q3 +1,5*(Q3-Q1) ou xi< Q1 -1,5*(Q3-Q1) .sont représentées par une étoile.

3-Mesures de formes : distribution sous forme de cloche 2- distribution symétrique 3- Skewness(AS=0 symétrie) si AS penche à droite=>asymétrie gauche. 4- kurtosis (AP=3),5-verifierl'existence des outliers. AS=m3/ecart type3 avecm=1/Nsomme(xi-x/)3.

Liaison entre 2 variables :

Qualitatives: Chi-deux est nul dans le cas d'indépendance (profils identiques) et d'autant plus important queles profils sont différents entre eux. On établit le tableau de contingence ou bien on juxtapose les courbes enbâton.
Quantitatives :I- Représentation graphique par nuage de points II- coefficient de corrélation r=cov(x,y)/σ(x)σ(y).Cov(x,y)= $1/n \sum_{i=1}^{n} \quad (x_i - \bar{x})(y_i - \bar{y})$Interprétation: r=-1 ou r=1 relation affine entre x et y il existe une forte liaison linéaire entre X et Y (en général -1<r<1)
Qualitative/Quantitative :pour chaque Y qualitatif on crée une boite a moustache et on compare.

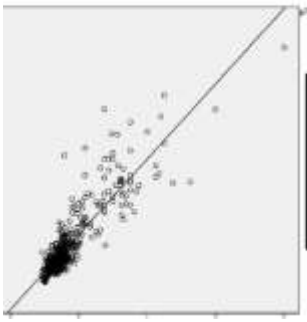
II*Régression linéaire :

Hypothèse 1 : La relation entre X et Y doit être linéaire;
Hypothèse 2 : le nombre d'observations doit être supérieur au nombre de variables;
Hypothèse 3 : les variables exogènes doivent être linéairement indépendantes.
Hypothèse 4 : Normalité et indépendance des Yi
Hypothèse 5 : Homoscedasticité
Hypothèse 6 : Les résidus doivent être Normaux, indépendants, centrés et non corrélées avec les variables explicatives. Les résidus standardisés sont dans l'intervalle [-3,3] => variabilité acceptable
Relation entre la consommation et chacune des variables ?
La matrice de corrélation montre que la consommation est fortement corrélé avec les variables indépendantes (coefficient de pearson ~1)
De plus, **les diagrammes de dispersions** laissent penser que l'hypothèse de linéarité semble acceptable

Est-ce que les variables X1,…X4 sont indépendantes ?
Le tableau des variables introduites montrent que toutes les variables ont été introduite, donc l'introduction de chaque variables n'a pas fait passer la tolérance des autres variables au-dessous du seuil de tolérance (0,3) => les variables exogènes sont faiblement colinéaires
Conclusion ?on peut effectuer une regression linéaire comme méthode descriptive/ajustement linéaire
Méthode utilisée : méthode des moindres carrés ordinaire.
Qualité du modèle : Récapitulatif du modèle : le R^2 ajusté = 0,948 => il restitue 94,8% d'information de la variabilité initiale
Expression de la relation : voir tableau des coefficients
Qu'est ce qui influence le plus la consommation d'essence d'une voiture :
La puissance ; ayant le plus grande coeff de corrélation et la plus grande beta ;
Rq : Bêta : permet de comparer la contribution de chaque variable ; t :doit être >1.96 pour que le coeff d'un var dans l'équation soit significatif
En analysant les résidus, pensez-vous que le modèle ajuste bien les données :
Oui : Les Résidus suivent une distribution normale (graphique de répartition de résidu par une répartition normale (**PP gaussien**))
Non : **L'histogramme (conso,effectif)** n'est pas symétrique donc la distribution n'est pas stable =>nnstablité de Y
Existe de l'information non expliqué ?
Non, **la représentation des prédictions standardisés en fct des résidus standarisé**ne fait apparaitre aucun modèle particulier ce qui confirme l'hypo de val constante de la variance du terme d'erreur (homoscédasticité) et d'indépendance des termes d'erreur.
Interpretation du coefficient de la variable indépendante : Une pente de a implique une augmentation d'une unité en X entrainera une augmentation moyenne de a unité en Y
Les facteurs d'amélioration :
- La prise en compte des autres puissances de la variable niveau d'études : puisque la relation entre celle-ci et le salaired'embauche, comme on a dit au début, est polynomiale.
- Traitement des valeurs aberrantes : toutes les valeurs qui dépasse -3σ et 3σ.on inserrant des variables dummy qui prend 1 pour les valeurs atypiques et 0 ailleurs.
III*ACP: Résumer un ensemble vaste de donnée numérique en un ensemble plus petitde valeurs pertinentes.
Inertie totale = variance totale = p
Part de variance expliquée par la première composante principale $=\lambda_1/p$
Part de variance expliquée par la deuxième composante principale $=\lambda_2/p$
Part de variance expliquée par les deux premières composantes principales $=(\lambda_1+\lambda_2)/p$
Et ainsi de suite pour les autres dimensions...
Phase I : mettre en évidence les relations entre les variables (étude des liens entre les variables.)
A- Matrice de corrélation (pour avoir une idée sur les classes homogènes qu'on peut extraire)
Si les variables son peu corrélés, il est alors inutile de déterminer les facteurs commun puisque les variables partagent peu de caractéristiques en communs.
Si les variables sont fortement corrélés, il parait pertinent de chercher à synthétiser l'information en réduisant le nombre de variables en un petit en un petit nombre de facteur 2 à 2 non corrélés.
PHASE II : évaluation des propriétés du modèle factoriel. *matrice anti-image ; matrice des corrélations partielles.
Les corrélations partielles totales donnent une idée de la force intrinsèque qui relie 2 variables en supprimantl'effet linéaire induis par les autres variables. *Interprétation :* coefficient proche de 1implique d'interrelationtransitant par toutes les variables du modèle
PHASE III : critères de pertinence d'une ACP :
A- Test de KMO (Kaiser Meyer Olkyn). *Interprétation : KMO proche de 1=>forcément en va réduire et lemodèle est* Merveilleux(0,9), méritoire(0,8),Moyen(0,7), médiocre(0,6), misérable(0,5), inacceptable (<0,5). Diag :EchDeMesure,
B- Test de sphéricité de Bartlett. Test l'hypothèse h0 ; matrice de corrélation est égale à la matrice d'identité pour savoir si les variables sont corrélées.
C- MSA calculable variable par variable. Diagonale de l'anti-image *Interprétation*; plus le MSAi est élevé(proche de 1) plus la variable correspondante contribue fortement dans la construction des facteurs. *Cascontraire : on peut prédire que cette variable risque de rester seuldans un facteur.*
Est-ce que ça vaut la peine decontinuer. !!!!!!!!
PHASE IV : Extraction des facteurs.
A- Déterminer le nombre de facteur
Méthode 1 : graphique des valeurs propres. Méthode 2 : variance expliquéestotale.
1- Qualité de représentation (extraction à un espace de n variable)
A)- les variables sont quasiment parfaitement expliqués(bien restitués)
B) – Les variables sont mal restitués<9.

On augmente le nbr de facteur.
RQ : le nbre de facteur requis dépond du % cumulé d'info (grand) et les résidus (petit).
2 - corrélation reproduites avec variance totale expliquée : pourcentage des résidus avec tolérance d'erreur de 5%
Interprétation : avec n facteur on a réussi à restituer presque la totalité de variabilité de l'échantillon.
On peut s'en assurer aussi par la qualité de représentation pour remarquer que tous les variables sont bien restitués.
PHASE IV : Extraction des facteurs=Rendre les facteurs plus interprétables.
A) -Matrice des composantes : qui se lit verticalement il s'agit de la corrélation entre la composante et lavariable EX ; on remarque que pour le 1er axe est celui pour lequel les coefficients sont les plusélevés=>nous sommes dans la présence d'un effet de taille=>Rotation
VARIMAX=s'applique lorsque la plupart des variables sont représentées sur un seul axe =>minimiser le nbrde var qui ont une corrélation importante avec un facteur.Quartimax= s'applique lorsque une variable estfortement corrèles avec plusieurs axes à la fois=>minimser le nbre de facteur requis.
Equimax = combinaisondes 2 autres méthodes.
OBLIMIN cas extrême non orthogonale.
PAHSE V ; calcul des Scores : synthétiser l'information en nombre de facteur trouvé. Prendre en considérationla matrice des coefficients des cordonnées des composantes (qui contient la projection des variables sur les axes)pour voir les signes des variables qui varient dans le même sens ou pas.)

Étape 1 : la relation existante est linéaire



Corrélations

		Salaire actuel	Salaire d'embauche
Salaire actuel	Corrélation de Pearson	1	,880**
	Sig. (bilatérale)		,000
	N	474	474
Salaire d'embauche	Corrélation de Pearson	,880**	1
	Sig. (bilatérale)	,000	
	N	474	474

** La corrélation est significative au niveau 0,01 (bilatéral).

Étape 2 : Qualité du modèle de régression linéaire

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Changement dans les statistiques				
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F
1	,880a	,775	,774	$8,115.358	,775	1622.118	1	472	,000

a. Valeurs prédites : (constantes), Salaire d'embauche
b. Variable dépendante : Salaire actuel

La qualité du modèle de régression est 77,4%

Étape 3 : Estimation des paramètres du modèle de regression linéaire

Coefficients

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	1928,206	888,680		2,170	,031
	Salaire d'embauche	1,909	,047	,880	40,276	,000

a. Variable dépendante : Salaire actuel

Étape 3 : tous les paramètres sont significatifs à 5%. donc l'équation de la droite :
Salaire_actuel = 1928,206 + 1,909 (salaire_embauche)

Modèle non valide

- Nécessité d'une transformation

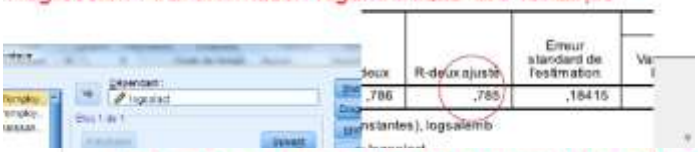
Nuage de points

Étape 4 : Analyse des résidus : résidus n'est pas une distribution gaussienne

Diagnostic des observations

Numéro de l'observation	Erreur Résidu	Salaire actuel	Prév
18	6,074	$103,750	$54,6

Regression : transformation logarithmique des variables



	R-deux ajusté	Erreur standard de l'estimation	Va
	,785	,18415	

(constantes), logsalemb

Regression : transformation logarithmique des variables

Coefficients

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	,705	,232		3,038	,003
	logsalemb	,998	,024	,886	41,593	,000

a. Variable dépendante : logsalact

Diagnostic des observations

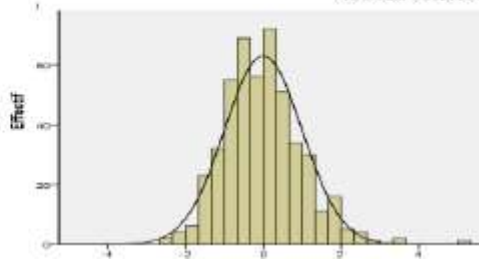
Numéro de l'observation	Erreur Résidu	logsalact	Prévision	Résidu
18	3,481	11,55	10,9087	,64108
218	5,092	11,29	10,3520	,93779
274	3,592	11,34	10,6742	,66143

a. Variable dépendante : logsalact

Statistiques des résidus

	Minimum	Maximum	Moyenne	Ecart-type	N
Prévision	9,7934	11,9739	10,3568	,35218	474
Résidu	-,46515	,93779	,00000	,18396	474
Erreur Prévision	-1,600	4,592	,000	1,000	474
Erreur Résidu	-2,526	5,092	,000	,999	474

a. Variable dépendante : logsalact



Les résidus est une normale. Le modèle est bon
Salaire-actuel = exp (logsalemb*0,998 + 0,705)
On peut l'affiner en y rajoutant la variable dummy_218

Examen analyse de données 2011 (corrigé)

Partie I : statistiques descriptives

a) Population d'analyse :les 200 clients de INTER-WEB

b) Le caractère étudié est la durée de connexion d'un client

c) C'est un caractère quantitatif continu

d) Oui on peut en calculer la moyenne

e) $\bar{X} = \frac{1}{200} \sum_{i=1}^{6} n_i c_i$ avec ni le nombre de clients dans la modalité i et ci le centre de la classe de la modalité i

$\bar{X} = \frac{30+10+90+30+150+100+210+30+270+20+330+10}{200}$

f) Le mode correspond à la valeur de la durée de connexion la plus fréquente et c'est 100

g) La médiane :

$$Q_2 = Binf + \frac{\frac{N}{2} - F}{fme} * E$$

Binf la borne inférieure de la classe médiane

N le nombre total d'observations =200

F la somme des fréquences des classes précédant la classe médiane =10+30+100

=10+30+100

fme la fréquence de la classe médiane

E l'étendu de la classe médiane

Ck+Ck+1/2=(150+210)/2=180 et 180 se trouve dans la classe [180, 240[donc c'est la classe médiane

AN

Partie II : régression linéaire

2.1. variable dépendante : durée de téléchargement

Variable indépendante :taille du fichier

2.2. facile

2.3.

2.4. le coefficient de la variable indépendante est la pente de la droite de régression

2.5. les hypothèses d'une régression linéaire sont :

La normalité et l'indépendance des yi

L'homosédasticité

Les résidus doivent suivre un bruit blanc, c'est-à-dire suivre une loi normale centrée réduite, en d'autre termes appartenir à l'intervalle [-3*écart-type,+3*écart-type]

2.6. Les résidus suivent une loi normale centrée réduite selon l'histogramme dans l'annexe

2.7. non, car la signification de la constante =0,232 >0,05 donc la constante est significative

2.8. AN dans la droite de régression

Partie III : analyse factorielle :

3.1. la matrice de corrélations nécessite de travailler avec les mêmes unités ce qui n'est pas le cas ici, donc on doit procéder à une standardisation et puis établir la matrice de corrélations

3.2. les variables fortement corrélées sont popul et manu

Au pourra au maximum réduire le nombre de variables à 5

3.3. KMO=0,653 médiocre donc la réduction n'est pas très importante, et la signification de Bartlett=0 donc on peut rejeter l'hypothèse d'indépendance des variables de la matrice de corrélations

3.4. le premier axe : 36,603% , le deuxième axe :24,999, cumulé=61%

3.5. la plupart des variables sont représentées sur un seul axe, d'où le recours à une rotation varimax

3.6. manuet popul sont corrélées avec le premier axe, tandis que temp et wind avec le deuxième, et le reste des variables avec le troisième axe.