Etapes d'une étude statistique :

1- définition de la problématique 2- Collecte des données 3- Préparation des données 4-Analyse statistique.

→**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 observés

Echantillonnage : Une partie de la population est observé.

La précision d'une enquête dépond : de la taille de l'échantillon et l'homogénité de la population.

5-Production et diffusion des résultats

Echelle **nominale** : les codes utilisés ne servent qu'à identifier la modalité à laquelle appartient l'individu (+ relation d'ordre entre les individus cas de variable ordinale).

Variables **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.

Graphiques utilisées :

Qualitatives : diagramme sectoriel, diagramme en bâton et courbes cumulées pour les variables ordinales

Quantitatives : var discrète : diagramme en bâton|/ diagramme cumulatif§§ var 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 sou population égales.

Valeurs distinctes : 1-Classer la série 2- déterminer si elle contient un nombre pair ou impair

Valeurs répétées : on utilise les effectifs cumulés croissant. Variable continue : on effectue 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,echelle)

Ecart interquartile: Iq= Q3–Q1 contient les 50% des valeurs les plus centrales 25%<Q2 et 25%>Q2. (ordinale,echelle)

Variance : σ2= 1/N Σni(xi–x/)2 = Σfi(xi–x/)2 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

Intetprétaion : la variabilité des arrivées dans h2 est bcp plus grande que celle des arrivés dans h1.

Les coefficient de variation : ecart_type/moyenne on peut pas comparer les dispersion de 2 séries statistiques qui sont éxprimées dans des unités différentes**.**

Interprétation :cv=0,45 ; cv=0,29.

La dispersion relatives de la distribution des médecins et beaucoup plus grande que celle de la distribution des infermiers. => le groupe des infermiers est bcp plus homogène que le groupe des médecins quand à leur répartition dans les dispensaires.

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-verifier l'existence des outliers. AS=m3/ecart type3 avec m=1/Nsomme(xi-x/)3.

Boite à moustache :

un rectangle d'extrémités Q1et Q3où la médiane est représentée par un trait horizontal à l'intérieur de la boîte,Les 2 moustaches débutent en Q1et Q3 se terminent en la valeur adjacente inférieure et supérieure

respectivement. -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.

Liaison entre 2 variables :

Qualitatives : Chi-deux est nul dans le cas d'indépendance (profils identiques) et d'autant plus important que les profils sont différents entre eux. On établit le tableau de contingence ou bien on juxtapose les courbes en bâton.

Quantitatives ;I- Représentation graphique par nuage de points

II- coefficient de corrélation $r=cov(x,y)/ \sigma(x)\ \sigma(y).cov(x,y)=1/N(Xi-X/)(Yi-Y/)$

Interpretation : $-1<=r<=1$ ➔ relation affine entre x et y il existe une forte liaison linéaire entre X et Y les 2 variables varient dans le meme sens
(coralation positive) ou dans le sens contraire (corelation negative).

Régression linéaire :

Une technique permettant de mettre on relation une variable endogène et n variables exogènes. Cette technique est utilisée lorsque on souhaite prédire une variable.

➔ méthode des moindres carrées : méthode géométrique qui repose sur l'hypothèse suivante : « la relation entre de variables X et Y est linéaire ». L'hypothèse de linéarité est vérifié par :➔ le nuage des points ➔testant la corrélation entre X et Y.

Etapes d'une RL : 1-Identifier la liaison linéaire :➔examinant le diagramme de dispersion (le nuage des points)➔(liaison constante➔pas d'influence, polynomiale

➔Coefficient de corrélation r (tester l'hypothese **r**=0)

RQ : Aussi bien le diagramme de dispersion et le test de corrélation de Pearson suggèrent une relation linéaire entre X et Y

2- Qualité d'ajustement : proche de 1 ➔ un bon modèle. Choisir le modèle avec la plus grande qualité d'ajustement.

R2ajusté exprime combien le modèle restitue l'inertie du nuage initial. A partir de cette ajustement

R2 ajusté vs R2 la relation du R2 ajusté ne tient pas en compte le nombre de variable ni le nombre d'observation.

Tableau d'anova :.teste l'hypothèse h : existe une indépendance entre Y et Xi au niveau alpha, signification<alpha ➔ il existe bien une depondance entre Y et les Xi.(aussi pour h0 R2=0).

Statistiques des résidus : il existe encore des résidus qui se trouvent à l'écart➔ le modèle n'est pas bon, il existe encore des informations contenues dans les résidus bien que on une qualité d'ajustement de n% (ecart type est garnd➔ une grande variabilité entre ce qui est prédit et l'information initiale et les résidus se trouvent à n fois l'écart type).

3- Estimation des paramètres ; h0 ; on teste l'annualité des paramètres ➔(est ce que le paramètre est significative➔influence directement sur Y) ➔Si P_value<alpha on rejette l'annualité du paramètre au niveau alpha.

Coefficient standardisé :

4- validation des paramètres

➔**Normalité et indépendance des Yi** (kolmogrov smirnov||Q-Q plot, histogramme) **h0 : Y suit une distribution normale.**

➔ **Les résidus suivent une distribution normale** (bruit blanc) [**0,1**](leurs PPpplot s'aligne ou pas avec une vrai normale à l'exception de quelques valeurs).

➔ **Indépendances des résidus avec les variables explicatives :** (Zpred,Zresd) + les résidus devrait se comporter de manière aléatoire le long de la bande + la variabilité des résidus n'augmente pas en fonction de l'ampleur des valeurs prévues.

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 salaire d'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.

ACP : Résumer un ensemble vaste de donnée numérique en un ensemble plus petit de valeurs pertinentes. Phase I : mettre en évidence les relations entre les variables ➔etude 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 intraseque qui relie 2 variables en supprimant l'effet linéaire induis par les autres variables. ***Interprétation :*** coefficient proche de 1➔implique d'interrelation transitant 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 le modèle est***

0,9➔merveilleux, 0,8➔méritoire, 0,7➔ Moyen, 0,6➔médiocre, 0,5➔misérable, <0,5➔inacceptable.

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. ***Cas contraire : on peut prédire que cette variable risque de rester seul dans un facteur.***

Est-ce que ça vaut la peine de continuer. !!!!!!!! PHASE IV : Extraction des facteurs.

A- Déterminer le nombre de facteur

Méthode 1 : graphique des valeurs propres.

méthode 2 : variance totale expliquées.

1- <u>Qualité de représentation</u> (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 - <u>corrélation reproduites avec variance totale expliquée :</u> pourcentage des résidus avec tolérance d'erreur de 5%

<u>Interprétation :</u> 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 la variable **EX ;** on remarque que pour le 1^{er} axe est celui pour lequel les cœfficients sont les plus élevés➔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 nbr de var qui ont une corélation importante avec un facteur. **Quartimax=** s'applique lorsque une variable est fortement corrèles avec plusieurs axes à la fois➔ minimser le nbre de facteur requis . **Equimax =** combinaison des 2 autres méthodes. **OBLIMIN** cas extrême non orthogonale.

PAHSE V ; calcul des Scores : synthétiser l'information en nombre de facteur trouvé. Prendre en considération la 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.