Traitement et analyse des données

Enseignant : Dr MAHOUI Karim

Démarche

- Rappel sur les échelles de mesure : types de variables
  - Faire des aller-retour
  - Importance des définitions et du processus d'opérationnalisation
- Types d'analyse:
  - 1ère classification : analyses univariée, bivariée et multivariée
  - 2e classification: statistique descriptive/exploratoire, statistique inférentielle
- Pour les analyse univariée et bivariée : voir cours de Dimitri Coll, Chargé de cours HEC de Montréal
- Pour le choix des tests (approfondissement du cours ci-dessus), Voir les deux sources suivantes (voir videos) : site de claude Goulet (planet psy) :lien suivant : http://pagesped.cahuntsic.ca/sc_sociales/psy/psy.htm ainsi que le site biostatgv sur le panorama des tests d'hypothèses http://marne.u707.jussieu.fr/biostatgv/?module=tests
- Analyse multivariée: voir panorama des methodes (différentes figures).Voir fichier de wikipédia pour la classification des différentes methodes
- Pour la mise en oeuvre de la regression simple et multiple : utiliser soit spss (voir le site de l'université de Sherbrook): http://spss.espaceweb.usherbrooke.ca/ très pratique; soit Eviews, voir des applications dans le domaine FCI sur le site en anglais!) de Dave Smant : https://sites.google.com/site/davesmant/courses/working-with-eviews/intro-eviews-programming
- Pour les analyses multivariées, je préconise XLSTAT (facilité d'uilisation)

Rappel de quelques notions

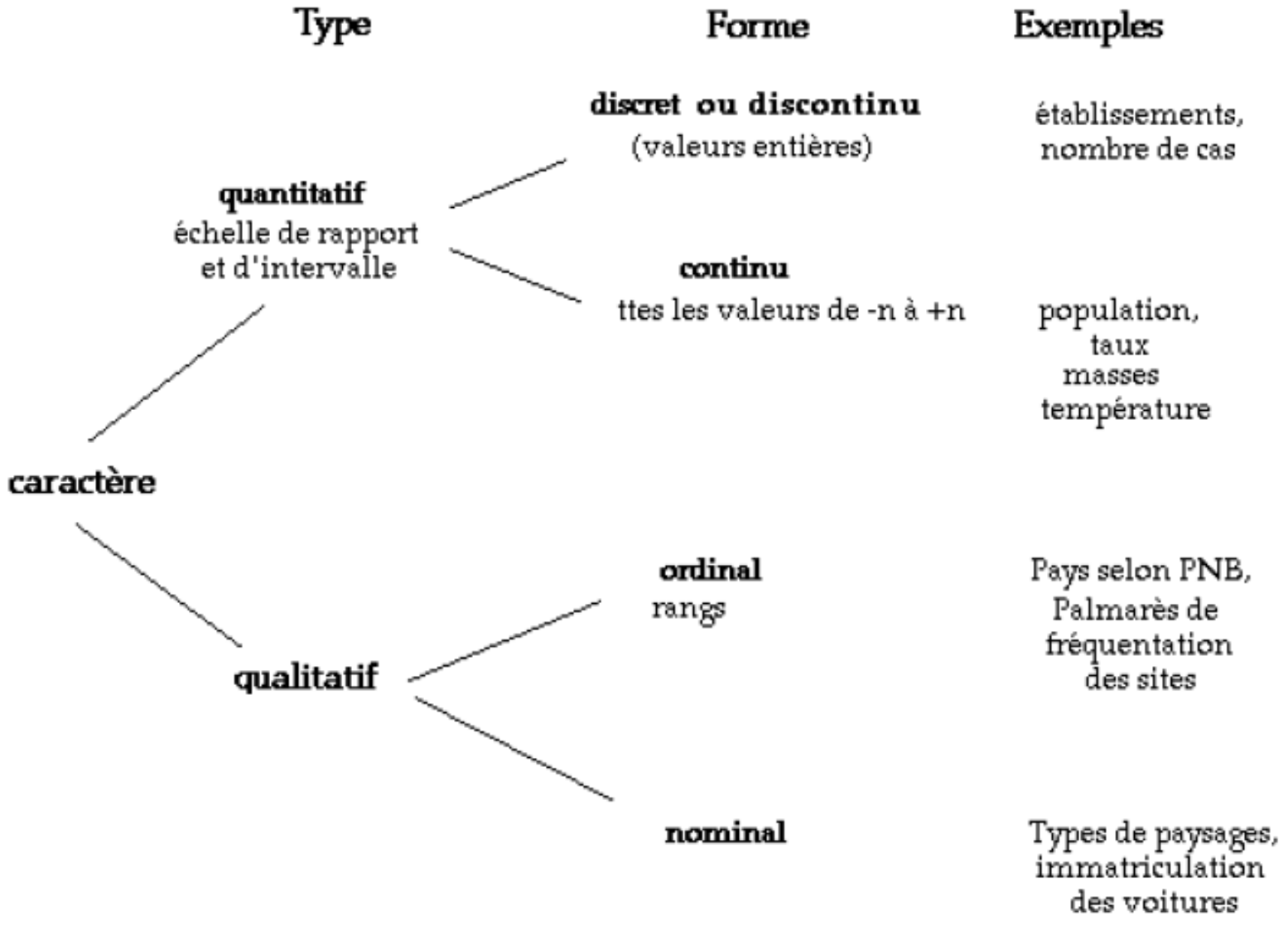
- **Variable**: une caractéristique des objets de l'étude. Ex. le revenu d 'une personne; le taux de chômage d 'une commune, wilaya, région, pays…
- Sur la base d'une hypothèse, on peut ensuite distinguer entre:
  - Variable indépendante (X, ou causale, ou explicative ou exogène)
  - Variable dépendante (Y, ou expliquée ou endogène)
  - Variable intermédiaire

Rappels de quelques notions

- **Indicateurs**: caractéristiques mesurables qui permettent de situer les objets étudiés sur des dimensions
- Indice: combinaison de plusieurs indicateurs qui permet de mesurer une dimension ou un concept
- Opérationnalisation:

Concept → dimension → indicateurs

Typologie des échelles de mesures



Tests statistiques – Définition et principes

- Les tests statistiques font partie de ce que l'on appelle la statistique inférentielle.
- Au contraire de la statistique descriptive, on va utiliser des lois de probabilités afin de prendre une décision dans une situation faisant intervenir une part de hasard. Effectivement, dans les tests statistiques, on ne va pas travailler sur une population mais sur un échantillon.
- Les tests statistiques sont ainsi souvent utilisées pour isoler une partie de la population d'une influence. On forme ainsi une population témoin.

Par exemple :

- Dans le domaine médical, on isole 2 échantillons : le premier soumis a un médicament et le second non soumis. On observe ainsi l'effet du médicament.

http://bienvenudansladata.com/tag/tests-statistiques

Typologie des test

1ère typologie: test paramétrique et test non paramétrique

•Tests paramétriques : Test des paramètres de la série en faisant l'hypothèse d'une distribution (souvent normal)
•Tests non paramétriques : Test de la série sans hypothèse de distribution.

	Avantages	Inconvénients
Test paramétrique	Plus puissants	Conditions d'applications contraignant
Test non paramétrique	Champs d'application plus vastes : • Echantillons de faibles tailles • Données qualitatives	Souvent moins documentés

Source : http://bienvenudansladata.com/tag/tests-statistiques

- **Finalité du test**
  - Conformité : La valeur observée dans mon échantillon est bien celle attendues (correspondant à un standard)
  - Adéquation : La série suit la même distribution qu'une loi choisie à priori (souvent la loi normale)
  - Homogénéité : Les échantillons proviennent de la même population, ie la variable d'intérêt a le même comportement sur l'ensemble des échantillons
  - Indépendance : il existe une liaison entre les variables.
- **Type de variables**
  - Qualitatives
  - Quantitatives
- **Nombre et le type d'échantillons**
  - Un seul échantillon
  - Deux ou plus échantillons
    - Appariés
    - Indépendants

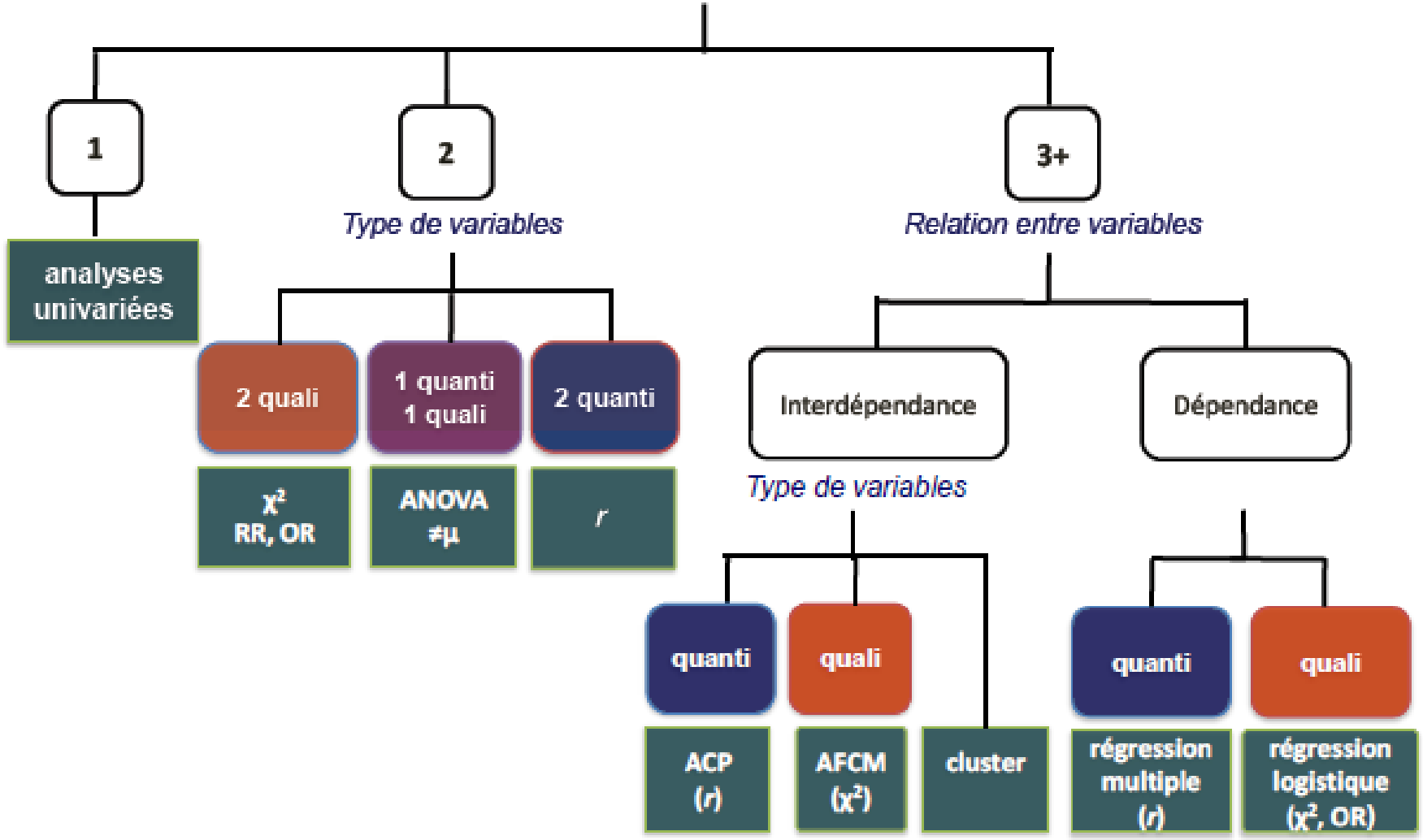
Test paramétrique

	Qualitative	Quantitative
Conformité		Test de moyenne (Test Normal ou de Student) Test de Variance (Test de Fisher) Test de proportion (Test de Student ou McNemar) ANOVA
Adéquation		
Homogénéïté		Test de moyenne (Test Normal ou de Student) Test de Variance (Test de Fisher) ANOVA
Indépendance		Coefficient de corrélation de Pearson

Test non paramétrique

	Qualitative	Quantitative
Conformité		Test de Mann-Whitney
Adéquation		Test de Kolmogorov-Smirnov Test du khi-2 Test de Shapiro
Homogénéïté	Test du khi-2	Test des rangs de Wilcoxon Test de Mann-Whitney Test de McNemar
Indépendance	Test du khi-2	Rho de Spearman Tau-a de Kendall

2nd typologie: nbre de variables



Autres arbres de décision

- Xlstat guide : https://help.xlstat.com/s/article/guide-de-choix-de-test-statistique?language=fr
- Site Biostat TGV : https://biostatgv.sentiweb.fr/?module=tests
- Site Bioinfo : https://bioinfo-fr.net/tests-statistiques-suivez-lguide
- Site claude goulet: https://pagesped.cahuntsic.ca/sc_sociales/psy/methosite/consignes/decision.htm

Analyse des données

1. Analyse statistique univariée des données

Source : Dimitri Coll, Chargé de cours HEC de Montréal

L 'ANALYSE STATISTIQUE UNIVARIÉE

- Décrire et synthétiser les résultats de la recherche en analysant les variables une à la fois.

- Dans le cas de variables non métriques, on utilise des distributions de fréquences.

- Dans le cas de variables métriques, on utilise les statistiques descriptives (mesures de tendance centrale et de dispersion)

L ’ANALYSE STATISTIQUE UNIVARIÉE

Variable non métrique

Échelles	Tendance centrale	Dispersion	Graphique
Nominale	Mode	-	Pie chart Histogramme
Ordinale	Mode Médiane	-	Pie chart Histogramme
Métrique	Mode Médiane Moyenne	Écart type Étendue	Histogramme

2. Analyse statistique bivariée des données

LES TYPES DE RELATION

Les analyses bivariées :

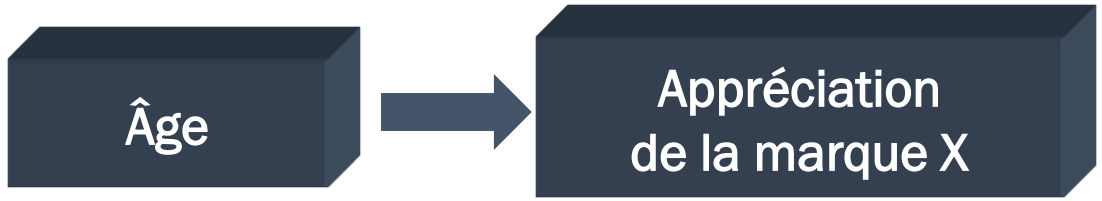
- Vérifient les relations entre deux variables
- 2 types de relation bivariée :
  - les relations de **dépendance** (plus fréquentes)
    - Deux variables :
      - variable indépendante
      - variable dépendante
  - les relations **d'interdépendance**
    - Dans une relation d'interdépendance les 2 variables s'influencent mutuellement

Une variable peut jouer le rôle de variable dépendante dans un contexte donnée et agir à titre de variable indépendante dans un autre.

LES TYPES DE RELATION

Les analyses bivariées :

- Exemple de relation de dépendance



Exemple : L'âge permet-il d'expliquer l'attitude envers la marque X ?

- Exemple de relation d'interdépendance

Exemple : Quelle est la différence d'appréciation entre la marque X et la marque Y ?

DÉMARCHE DE L'ANALYSE BIVARIÉE

Déterminer l’échelle de mesure des questions

↓

Déterminer le test approprié

↓

Faire le test

↓

Interpréter le résultat

LE CHOIX D ’UNE TECHNIQUE D’ANALYSE APPROPRIÉE

ÉCHELLE DE MESURE DE LA PREMIÈRE VARIABLE	ÉCHELLE DE MESURE DE LA DEUXIÈME VARIABLE		
	Nominale ou Ordinale		D’intervalles ou de ratio
Nominale ou Ordinale	Tableau croisé		Comparaison de moyennes
D’intervalles ou de ratio	Comparaison de moyennes		Corrélation ou régression

	Deux variables non métriques	Une variable non métrique et une variable métrique	Deux variables métriques
Type d'analyse	Tableau croisé	Comparaison de moyennes	Corrélation ou régression
Tests statistiques	χ^2 γ (si les variables sont ordinales)	*t* (2 moyennes) *F* (2 moyennes ou plus)	*t* (corrélation, régression) *F* (régression)
Force de la relation	*V* de Cramer γ gamma (si les variables sont ordinales)	η eta	*r* (corrélation) *R* (régression)

	Deux variables non métriques	Une variable non métrique et une variable métrique	Deux variables métriques
Interprétation	Fréquences et pourcentages dans le tableau	Moyennes de groupe	Ordre de grandeur et signe du coefficient
À surveiller	Fréquences observées et théoriques % de cell avec moins de 5 rep	Taille des groupes, valeurs extrêmes	Dispersion des variables (linéarité) et valeurs extrêmes

DÉMARCHE DE L'ANALYSE BIVARIÉE

Déterminer l'échelle de mesure des questions

Déterminer le test approprié

Faire le test

Contrepartie : H1

- Évaluer la significativité de la relation entre les variables
- Déterminer la force de la relation
- Déterminer le sens de la relation

- Déterminer H_0
- Tester H_0
- Accepter/Rejeter H_0

Interpréter le résultat

L ’analyse des tableaux croisés : Le test d’indépendance Chi2

L ’ANALYSE DES TABLEAUX CROISÉS

Procédure de test

- On pose l’hypothèse nulle :

- H0 : Il n’y a pas de relation entre les deux variables.
- H1 : Il y a une relation entre les deux variables.

- On rejette l’hypothèse nulle (on conclut que la relation existe dans la population) si :

L ’ANALYSE DES TABLEAUX CROISÉS

La statistique khi²

Les fréquences théoriques sont les fréquences que l’on obtiendrait si les variables dépendante et indépendante ne sont pas associées (indépendantes).

Logique du test d 'indépendance du khi²

- La statistique x2
  - un indice de la distance entre les fréquence théoriques et les fréquences observées.

- Plus la valeur de x2 est grande, plus on croit que les deux variables sont associées.
  - Rejet de l'hypothèse H0

- la relation existe dans la population lorsque la valeur de x2 est trop improbable,
  - plus précisément lorsque la probabilité d'observer une telle valeur est inférieure à 0,05 (règle de la valeur p).
  - sous l'hypothèse que les deux variables sont indépendantes

L ’ANALYSE DES TABLEAUX CROISÉS

- La valeur p (ou seuil de significativité)…
- … correspond au % de chance que H0 ($\chi 2 = 0$) soit vrai.
- NOTEZ: La relation entre les variable est significative lorsque $p \leq 0{,}05$

L ’ANALYSE DES TABLEAUX CROISÉS

La force de la relation

V de Cramer :

$$0 \leq V \leq 1$$

Interprétation qualitative de la statistique *V*

		V	$\geq$	0,70	relation très forte
0,50	$\leq$	V	$\leq$	0,69	relation forte
0,30	$\leq$	V	$\leq$	0,49	relation modérée
0,10	$\leq$	V	$\leq$	0,29	relation faible
0,01	$\leq$	V	$\leq$	0,09	relation très faible
		V	$=$	0,00	relation nulle

Cas des variables ordinales

L ’ANALYSE DES TABLEAUX CROISÉS : CAS DES VARIABLES ORDINALES

- L’analyse du x2 appropriée pour des variables nominales ou ordinales
- Lorsque les deux variables sont mesurées à l’aide d’une échelle ordinale, on peut procéder à une analyse complémentaire à l’aide de la statistique gamma (γ).
- La statistique γ mesure le sens et la force de la relation entre deux variables ordinales dans une relation linéaire :
- $-1 \leq \gamma \leq 1$
-
- Important : Nécessité d’un échantillon assez grand
- L'interprétation se fait à partir du schéma d'interprétation du V de Cramer

L ’ANALYSE DES TABLEAUX CROISÉS

Important…

- Il est important de s’assurer que les fréquences à l’intérieur du tableau sont suffisamment grandes.

Règles :

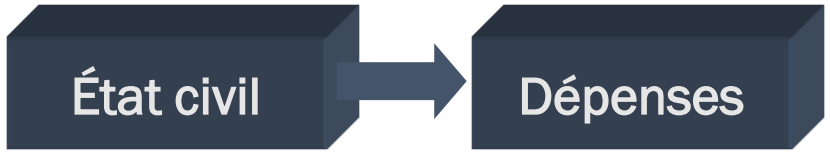
- - Oij ≥ 1
- - Tij ≥ 1
- - max. entre 25 % et 30 % de Tij ¸ < 5

Regroupement de catégories (recodification)

Comparaison de deux moyennes indépendantes

COMPARAISON DE DEUX MOYENNES INDÉPENDANTES

Exemple



Variables non métriques

2 personnes différentes

Homme	Femme
350 $	245 $
325 $	195 $
...	...
290 $	220 $

Variables métriques

$\overline{X}_1 = 318\ \$$ $\overline{X}_2 = 222\ \$$

COMPARAISON DE DEUX MOYENNES INDÉPENDANTES : LA PROCÉDURE DE TEST

On pose l'hypothèse nulle :

H_0 : Il n'y a pas de relation entre les deux variables.

H_1 : Il a une relation entre les deux variables

On rejette l'hypothèse nulle (on conclut que la relation existe dans la population) si :

Test bilatéral	Test unilatéral
$t > t_{0,025}$	$t > t_{0,05}$ (à droite)
ou $< -t_{0,025}$	ou $t < t_{0,05}$ (à gauche)

dans les deux cas, $\nu = n_1 + n_2 - 2$ (degrés de liberté).

COMPARAISON DE DEUX MOYENNES INDÉPENDANTES : LA FORCE DE LA RELATION

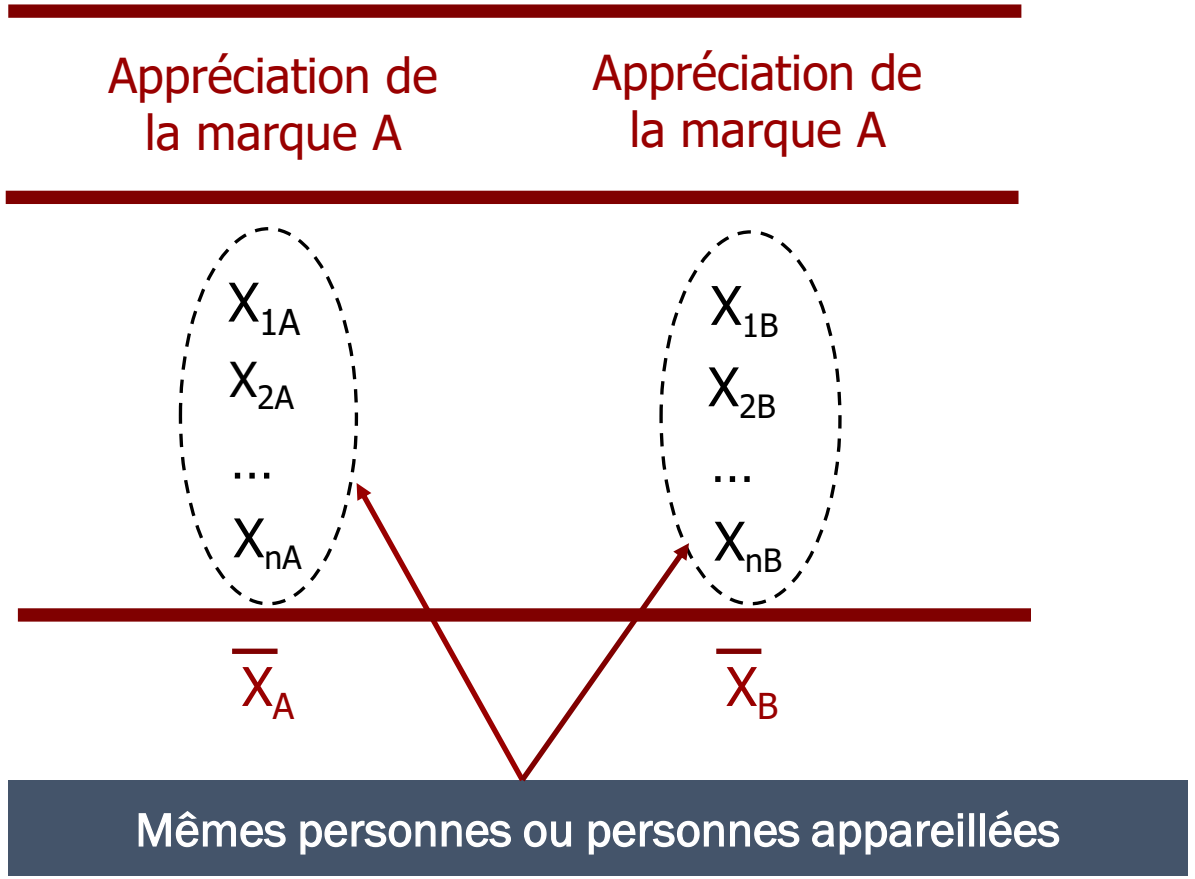
On peut mesurer la force de la relation entre les deux variables par le biais de l'indice suivant, qu'on appelle la statistique eta « η »:

$$\eta = \sqrt{\frac{t^2}{t^2 + n_1 + n_2 - 2}}$$

$$0 \leq \eta \leq 1$$

Aucune relation — Relation parfaite

COMPARAISON DE DEUX MOYENNES APPAREILLÉES



COMPARAISON DE PLUSIEURS MOYENNES INDÉPENDANTES (ANOVA)

Exemple

Marié	Célibataire	Divorcé
350 $	245 $	375 $
325 $	195 $	350 $
…	…	…
290 $	220 $	310 $
$\overline{X}_m$ = 318 $	$\overline{X}_c$ = 222 $	$\overline{X}_d$ = 379 $

COMPARAISON DE PLUSIEURS MOYENNES INDÉPENDANTES : LA PROCÉDURE DE TEST

On pose l'hypothèse nulle :

H0 : Il n'y a pas de relation entre les deux variables.

H1 : Il a une relation entre les deux variables

On rejette l'hypothèse nulle (on conclut que la relation existe dans la population) si :

COMPARAISON DE DEUX PLUSIEURS MOYENNES INDÉPENDANTES : LA LOGIQUE DU TEST

On conclut que la relation existe dans la population

- lorsque la valeur de F est improbable,

plus précisément…

- lorsque la probabilité d'observer une telle valeur est inférieure à 0,05 (règle de la valeur p).

COMPARAISON DE PLUSIEURS MOYENNES INDÉPENDANTES : LA FORCE DE LA RELATION

On peut mesurer la force de la relation entre les deux variables par le biais de l'indice suivant, qu'on appelle la statistique eta :

$$\eta = \sqrt{\frac{SCG}{SCT}}$$

SCG = somme des carrés entre les groupes

SCT = somme des carrés totale

$0 < \eta < 1$

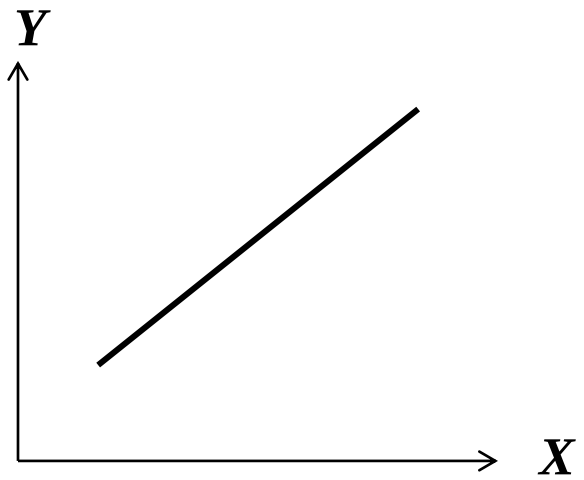
Statistique non fournie par spss

L’analyse de corrélation

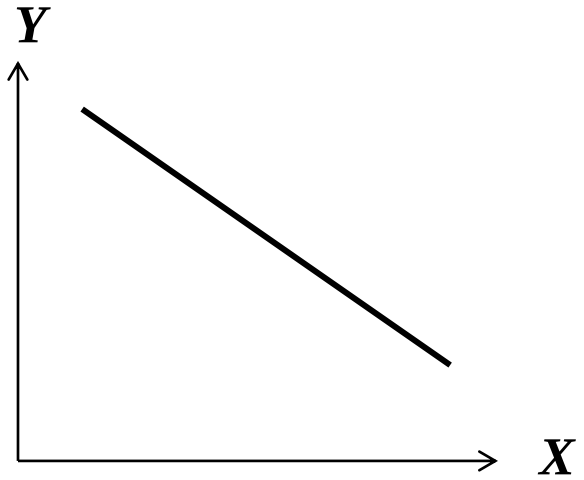
L ’ANALYSE DE CORRÉLATION

- Variables mesurées avec des échelles métriques.
- Établir si l’augmentation des valeurs d’une des deux variables entraîne systématiquement l’augmentation ou la diminution des valeurs de l’autre variable.
- Relation linéaire

L ’ANALYSE DE CORRÉLATION

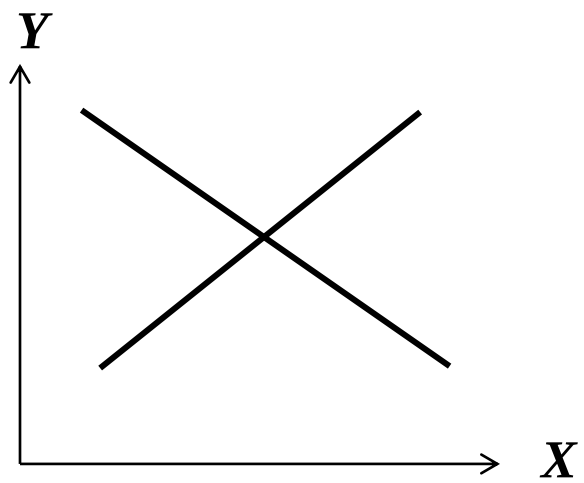


Covariation positive : $r > 0$

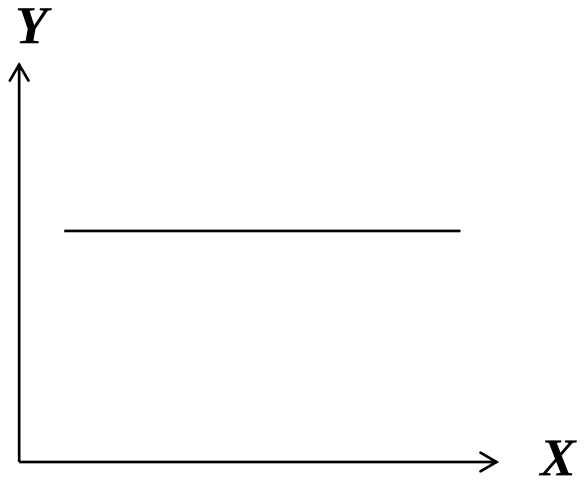


Covariation négative : $r < 0$

L ’ANALYSE DE CORRÉLATION



Covariations significatives



Absence de covariation

COEFFICIENT DE CORRÉLATION DE PEARSON

- Varie :

$$-1 \leq r \leq +1$$

- L'interprétation peut se faire à partir du schéma d'interprétation du V de Cramer

L 'ANALYSE DE CORRÉLATION : LA PROCÉDURE DE TEST

On pose l'hypothèse nulle :

H_0 : Il n'y a pas de relation entre les deux variables

. *H_1 : Il a une relation entre les deux variables*

On rejette l'hypothèse nulle (on conclut que la relation existe dans la population) si :

Test bilatéral	Test unilatéral
$t > t_{0,025}$ ou $< -t_{0,025}$	$t > t_{0,05}$ (à droite) ou $t < t_{0,05}$ (à gauche)

dans les deux cas, $\nu = n-2$ (degrés de liberté).

L ’ANALYSE DE CORRÉLATION : LA STATISTIQUE t

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

r = Coefficient de corrélation

n = Taille de l’échantillon

L ’ANALYSE DE CORRÉLATION : LA LOGIQUE DU TEST EN t

- On conclut que la relation existe dans la population lorsque la valeur de t (sous l’hypothèse que les deux variables sont indépendantes) est trop improbable, plus précisément lorsque la probabilité d’observer une telle valeur est inférieure à 0,05 (règle de la valeur p).

CONCLUSION SUR L ’ANALYSE STATISTIQUE BIVARIÉE

- L’importance de l’interprétation.
- La signification statistique versus la signification pratique.
- Les relations non significatives.
- La force de la relation.

Conclusion sur les analyses univariée et bivariée

- Première étape de l'analyse à proprement parler, la description des données permet de représenter les valeurs observées sur les différents individus de l'échantillon. L'**analyse univariée**, qui examine une seule variable à la fois, repose sur la **description** (fréquences, tendance centrale, dispersion, distribution), la visualisation graphique des variables et, éventuellement, sur l'**inférence**, c'est-à-dire la comparaison à des valeurs de référence connues pour déterminer si un échantillon diffère significativement d'une population plus large. L'**analyse bivariée** permet d'aller plus loin par l'étude des relations entre deux variables, grâce aux **tris croisés** et aux principaux tests d'analyse bivariée : **tests d'association** (khi-deux) et **tests de comparaison** (test *t*, test *U* de Mann-Whitney, etc.). Pour aller encore plus loin dans l'analyse, il faudra mettre en place des **analyses multivariées** abordées dans les chapitres suivants.