Séance 3: Bref Survol de l'analyse descriptive

Partie 2: Association entre variables

Visseho Adjiwanou, PhD.

Département de Sociologie - UQAM

Au programme aujourd'hui

1. Association entre deux variables catégorielles
2. Association entre une variable catégorielle et une variable continue
3. Association entre deux variables continues
4. A faire pour la semaine prochaine

Statistiques bivariées

Introduction

- Le cours précédent nous a permis de décrire les caractéristiques d'un échantillon en analysant ses paramètres de positions et de dispersions.
- Mais, en tant que sociologues, nous allons loin que cela. Ce qui nous intéresse, ce sont les relations causales dans lesquelles nous faisons l'hypothèse qu'une variable indépendante affecte une variable dépendante.

Introduction

- Par exemple:
  - Est-ce que le vaccin A protège contre la maladie X?
  - Les classes de petite taille augmentent-elles les résultats des tests standardisés des élèves?
  - Les soins de santé universels amélioreraient-ils la santé et les finances des pauvres?
  - L'éducation réduit-elle le nombre d'enfants?
  - La rémunération des gens sur Wikipedia augmentera-elle leur productivité?
  - Est-ce que l'augmentation du salaire minimum réduit l'activité économique?
  - Est-ce le statut marital influe sur le bonheur?
- La réponse à ces questions va au-delà des logiciels statistiques. Elle commence et finit avec vous.
- les logiciels/statistiques ne nous donnent qu'une indication.

Conditions de la causalité

- L'une des conditions de la causalité est l'existence d'association entre la variable indépendante et la variable dépendante.
- L'analyse de cette association dépend du type des variables indépendantes et dépendantes
  1. Deux variables factorielles ou catégorielles (nominale ou ordinale) : Recours aux tableaux bivariés
  2. Variable dépendante continue (ratio ou d'intervalle) et variable indépendante factorielle : Différences de moyennes: Analyse de la variance (ANOVA)
  3. Deux variables continues : Technique de régression
- Les graphiques permettent aussi de mettre en exergue ces relations.

Répondre à 6 questions

1. Y a-t-il une **relation** entre deux variables ?
2. Quelle est l'**intensité** de cette relation ?
3. Quelles sont la **direction** et la **forme** de cette relation?
4. Pouvons-nous **généraliser** la relation à la **population** de laquelle est tiré l'**échantillon**?
5. La relation est-elle vraiment **causale**?
6. Quelles sont les **variables intermédiaires** qui relient les variables **indépendante** et **dépendante**?

Aujourd'hui, nous allons nous intéresser aux trois premières questions.

Mesure de l'association entre deux variables catégorielles

Tableaux bivariés

Tableaux bivariés

- Évaluer la relation entre les variables dépendantes et les variables indépendantes (ou entre deux variables)
- Donne une première indication de l'effet d'une variable indépendante sur la variable dépendante
- Les lignes et les colonnes n'ont pas la même signification
- Les variables dépendantes sont mises dans la colonne
- Les variables indépendantes sont mises en ligne
- L'interprétation fait référence à l'utilisation de fréquences marginales

Remarque: Dans votre livre de cours, les colonnes et les lignes sont renversées. Ce n'est pas grave du moment où on comprend la logique

Tableau croisé, effectif marginal, pourcentage marginal, pourcentage ligne, pourcentage colonne

Variable A (VI)	Variable B (VD)		
	Oui	Non	Total
Oui	A	B	A+B
Non	C	D	C+D
	A+C	B+D	A+B+C+D

Des fréquences au pourcentage marginal

Fréquences marginales

Variable A (VI)	Variable B (VD)			
	Oui	Non	%	Taille
Oui	A/(A+B)	B/(A+B)	100	N1 = A+B
Non	C/(C+D)	D/(C+D)	100	N2 = C+D

Fréquences marginales ou pourcentages marginaux

- Les fréquences marginales ne sont pas appropriées pour les comparaisons
- Par exemple, quel groupe ne fume pas le plus dans cette comparaison?
  - 6 personnes qui ont une faible éducation ne fume pas
  - 11 personne qui ont une éducation élevée ne fume pas
- Qu'est-ce qu'il faut pour pouvoir bien faire la comparaison?
- D'où l'importance des pourcentages marginaux

Exemple

- Existe-il une relation entre le lieu de résidence et la connaissance sur le VIH/Sida?

Place de résidence	Connaissance approfondie du SIDA		Taille
	Oui	Non	N
Urbain	2348	1954	4302
Rural	7094	11624	18718
N	9442	13578	23020

Exemple

- Mauvais tableau de pourcentage

Place de résidence	Connaissance approfondie du SIDA	
	Oui	Non
Urbain	24.9	14.4
Rural	75.1	85.6
N	100	100

Exemple

- **Bon tableau**

Place de résidence (VI)	Connaissance approfondie du SIDA (VD)			
	Yes	No	%	N
Urban	54.6	45.4	100	4302
Rural	37.9	72.1	100	18718

- **Mauvaise interprétation**: «Parmi les femmes vivant en zone rurale, seules 37,9% ont une connaissance approfondie du sida et 72,1% non; par conséquent, vivre dans des zones rurales vous rend moins susceptible d'avoir une connaissance complète du SIDA.

Exemple

- **Bon tableau**

Place de résidence (VI)	Connaissance approfondie du SIDA (VD)			
	Yes	No	%	N
Urban	54.6	45.4	100	4302
Rural	37.9	72.1	100	18718

- **Correcte interprétation**: La conclusion selon laquelle le lieu de résidence a un effet sur la connaissance du SIDA doit reposer sur une comparaison entre zones urbaines et rurales. Plus précisément, nous comparons les 54,6% avec les 37,9%. On note que les femmes en zones urbaines sont plus susceptibles que les femmes en zones rurales d'avoir une connaissance approfondie du sida.
- La comparaison des sous-groupes est donc essentielle pour la

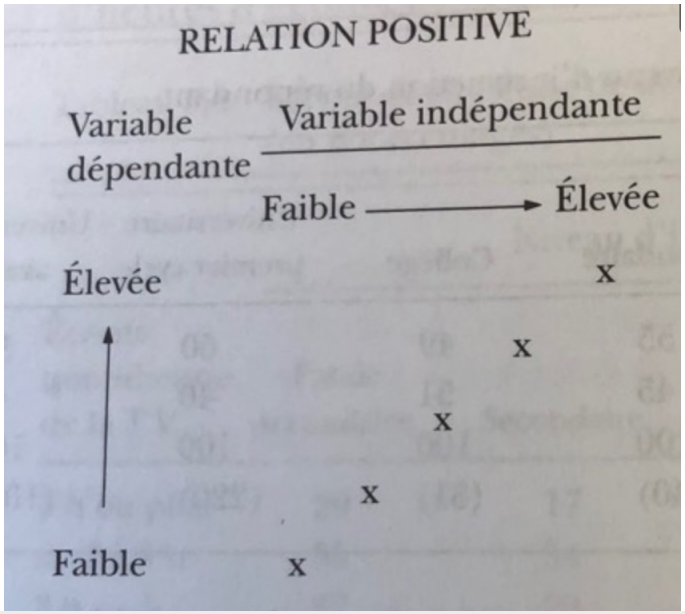
Direction de la relation

Direction de la relation

Il existe trois type de relation entre les variables

1. **Relation positive** : est une relation dans laquelle les scores les plus élevés d'une variable sont associées aux scores les plus élevés de l'autre variable

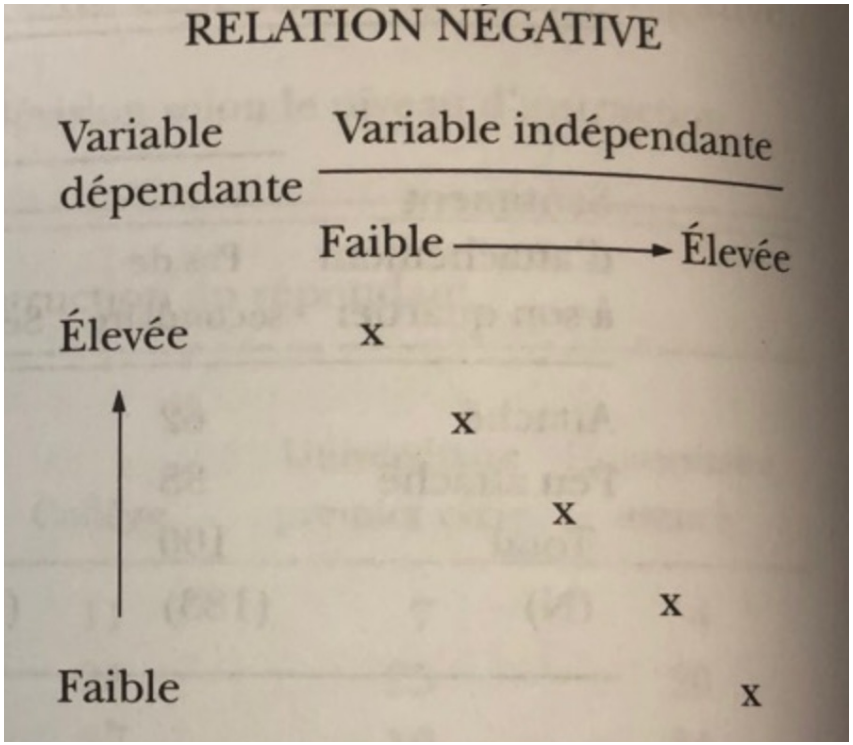
- Exemple?



Direction de la relation

2. Dans une **relation négative** les scores les plus élevés d'une variable sont liés aux scores les plus faibles de l'autre.

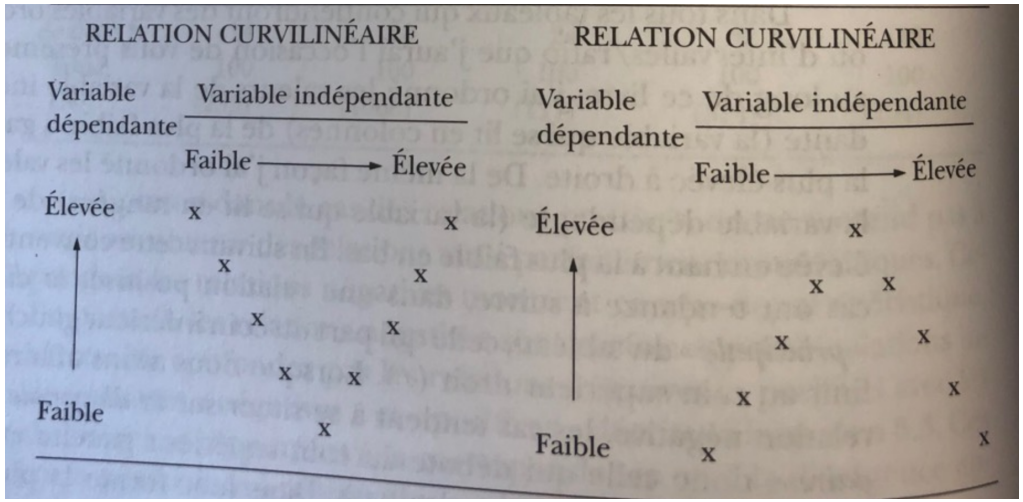
- Exemple?



Direction de la relation

3. Une **relation curvilinéaire** peut prendre différentes formes, mais les plus simples sont les relations dans lesquelles les cas avec des valeurs fortes et faibles pour la variable indépendante ont des valeurs similaires pour la variable dépendante.

- A souvent la forme d'un **V** ou d'un **V renversé**
- Exemple?



Problèmes avec les tableaux bivariés

- Si vous avez des effectifs faibles dans les tableaux, essayer de regrouper les modalités
- Effectif faible (moins de 30 par exemple)
- Les deux variables ne sont pas forcement catégorielles. Que faire dans ce cas?
  - Les regrouper en catégories
  - Problème : perte d'information

Intensité de la relation

Intensité de la relation

- Une relation peut-être nulle

	Y a-t-il une vie après la mort?			
Sexe	Oui	Non	Total	N
Homme	82,0	18,0	100,0	759
Femme	83,0	17,0	100,0	978

Intensité de la relation

- Une relation peut-être modérée

	Possédez-vous une arme à feu			
Sexe	Oui	Non	Total	N
Homme	48,0	52,0	100,0	848
Femme	35,0	65,0	100,0	1066

Intensité de la relation

- ou forte

	Peur de se promener dans son quartier?			
Sexe	Oui	Non	Total	N
Homme	26,0	74,0	100,0	846
Femme	55,0	45,0	100,0	1057

Intensité de la relation

- Nous sommes arrivés à ces différentes conclusions en regardant l'écart entre les pourcentages.
- Mais, que faire si les variables dépendante et indépendantes ont plusieurs modalités?
- Nous le verrons bientôt, mais d'abord, essayons de mieux comprendre l'association entre deux variables.

Chi-carré

Tests statistiques

- Jusque là, nous avons pu voir l'association au sein de nos données
- Mais ces données proviennent d'un échantillon unique
- Comment s'assurer que ce résultat ne change pas si nous changeons d'échantillon?
- C'est-à-dire qu'il ne soit dû qu'au hasard?
- Autrement, comment s'assurer que la relation est aussi vraie au sein de la population?

Logique des tests statistiques

- Déterminer la probabilité de découvrir une relation dans notre échantillon quand il y en a au sein de la population.
- Si cette probabilité est petite (1/20 ou 5%, d'òu l'idée du seuil de 5%), et si nous découvrons une relation au sein de l'échantillon, nous pourrons conclure qu'il existe probablement une relation dans la population.
- Alors il s'agit de tester la supposition qu'il n'existe pas de relation dans la population. On appelle cela l'**hypothèse nulle**, notée H_0.
- Ainsi, rejeter l'hypothèse nulle, revient à dire qu'il existe une relation dans la population.
- On parle de relation **statistiquement significative**

Logique des tests statistiques

- A l'inverse, si les chances de trouver une relation dans l'échantillon alors qu'il n'y en a pas dans la population sont élevées (supérieures à 1 sur 20), nous NE pouvons croire en toute confiance à l'existence d'une relation dans la population.
- La relation trouvée au sein de l'échantillon est probablement dûe au hasard seul.
- Dans ce cas, nous disons que nous NE rejetons PAS l'hypothèse nulle
- Il n'y a pas de relation au sein de la population.

REmarque: On dit **ne pas rejeter l'hypothèse nulle** et non **accepter l'hypothèse nulle**.

Logique des tests statistiques

- **Niveau de significativité** = probabilité de retrouver grâce au hasard une relation au sein d'un échantillon, en dépit de l'absence de relation dans la population. Il est noté α.
- On n'utilisera souvent les niveaux de significativité de 5%, 1% et 0.1%.
- Le fait que la signification statistique repose sur une probabilité implique que nous ne pouvons jamais être absolument certains de faire le bon choix.
- Pour le faire, il faut les données de la population directement.
- Ainsi, on peut commettre deux types d'erreurs.

Logique des tests statistiques

- Erreur de type I ou erreur **alpha** : rejet de l'hypothèse nulle alors qu'elle est vraie. La probabilité d'une erreur de type 1 est α
- Erreur de type II ou erreur **bêta** : Non rejet de l'hypothèse nulle alors qu'elle est fausse. La probabilité de l'erreur de type II n'est pas par contre (1-α)
- Si les chances de l'erreur de type I augmentent, les chances de l'erreur de type II diminuent et vice versa

Décision concernant H_0	Si H_0 est vraie	Si H_0 est fausse
Rejeter H_0	Erreur de type I	Pas d'erreur
Conserver H_0	Pas d'erreur	Erreur de type II

Logique des tests statistiques

- Comment trouver le niveaux de significativité?
- A partir de la distribution d'échantillonnage (distribution d'une statistique quelconque - par exemple la moyenne- de tous les échantillons possibles d'une taille donnée)
- La distribution d'échantillonnage et la méthode pour déterminer la signification statistique dépendent de la nature des données que nous analysons.
- Pour les données disposées en tableau, le test du chi-carré est utilisé. Il repose sur la distribution d'échantillonnage du chi-carré.

Le test du chi-carré

-le chi-carré, χ^2, est un nombre du compare les fréquences observées dans un tableau bivarié aux fréquences auxquelles on devrait s'attendre s'il n'y avait pas du tout de relation entre les deux variables dans la population (les fréquences anticipées).

- Sa formule est:

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a}$$

où :

$$f_a = (\frac{\text{Total de la colonne}}{N}) * (\text{Total de la rangée})$$

- f_a est la fréquence anticipée d'une cellule;
- f_o est la fréquence observée et
- N est le nombre total de cas

Le test de chi-carré

- Il existe une table de la distribution d'échantillonnage du χ^2, c'est-à-dire une table qui donne les probabilités d'obtenir un χ^2 au moins aussi grand qu'une certaine valeur si, dans la population de laquelle fut tiré l'échantillon, il n'y a pas de relation entre les deux variables.
- Cette probabilité dépend de ce qu'on appelle les **degrés de liberté (dl ou ddl)**.
- Dans le cas du chi-carré, le degré de liberté vaut (r - 1)(c - 1), r et c étant le nombre de rangées et de colonnes dans le tableau.
- Ce tableau donne la valeur minimale du chi-carré nécessaire pour obtenir un résultat statistiquement significatifs aux seuils (au niveau de significativité) de 0.05, 0.02, 0.01 et 0.001.

Exemple

Niveau d'instruction	**Attitude face à la violence**		
	Favorable	Pas favorable	Total
Moins que le secondaire	4	6	10
Secondaire	13	11	24
Postsecondaire	12	4	16

Exemple

Tableau des fréquences anticipées

- S'il n'y avait pas de relation entre les deux variables, on devrait s'attendre à ce que le tableau soit comment?

Association entre une variable quantitative (dépendante) et une variable qualitative (indépendante)

Différence de moyennes

On recourt à la différence de moyenne dans le cas où:

- La variable indépendante est dichotomique (homme/femme, jeune/vieux, marié/célibataire)
- La variable dépendantes est quantitative (ratio/intervalle)

Par exemple: - Analyser l'association entre le genre/sexe et le revenu

Différence de moyennes

- Calculer les moyennes de la variable dépendante pour les deux groupes.
- S'apprécie mieux avec un diagramme appelé diagrammes en boîtes. (deuxième partie du cours)

Analyse de la variance

Association entre une variable quantitative (indépendante) et une variable qualitative (dépendante)

Association entre deux variables quantitatives

Régression