# Seance 12: Analyse de la variance

Visseho Adjiwanou, PhD.

24 November 2021

# Rappel

| Type de la variable indépendante | | | Type de la variable dépendante | | |
|---|---|---|---|---|---|
| | | | Catégorielle | | Continue |
| | | | **Nominale** | **Ordinale** | **Intervalle/ratio** |
| | **Catégorielle** | **Nominale** | Tableau croisé<br>Test du chi-carré<br>V de Cramer<br>Diagramme de barre empillée | Tableau croisé | Analyse de la variance<br>Test t<br>État carré<br>Diagramme boxplot |
| | | **Ordinale** | Tableau croisé | Tableau croisé | idem |
| | **Continue** | **Intervalle/ratio** | Transformer la VI en catégorielle | | Corrélation / régression<br>Test t<br>R carré<br>Diagramme de nuage de points |

**Figure 1**

# Analyse de la variance

# Introduction

- La variance jour un rôle important dans la comparaison de moyennes entre deux groupes.
- le test t de Student nous informe sur le degré de confiance que nous plaçons entre deux différences
- Malheureusement, il sert seulement à comparer deux groupes
- Dans bien souvent des cas cependant, nous sommes intéressés à comparer plus de deux groupes:
- Par exemple:
- comparer les niveaux de revenus entre les provinces du Canada
- comparer les réseaux (nombre de numéro de téléphone) de différents groupes d'étudiants

# Introduction

- Une solution est de comparer les groupes deux par deux et utiliser le test t que vous connaissez déjà
- Vous pouvez vous rendre compte que cela devient vite lourd et peut conduire à des erreurs
- Une nouvelle méthode est utilisée dans ce cas: Analyse de la variance, nommée ANOVA

# Analyse de la variance

- L'ANOVA est une statistique inférentielle utilisée lorsque vous travaillez avec des variables d'intervalle / ratio (dépendantes) et des variables catégorielles (nominales ou ordinales comme variable indépendante)
- L'ANOVA nous permet de comparer entre les groupes de la variable catégorielle pour déterminer si la différence entre les moyennes (l'une d'entre elles) est statistiquement significative ou non.
- L'ANOVA unidirectionnelle (one-way ANOVA) fait référence à l'analyse avec une variable indépendante et une variable dépendante, tandis que l'ANOVA bidirectionnelle (two-way ANOVA) fait référence à deux variables indépendantes.
- Cette présentation est limitée à l'ANOVA unidirectionnelle.

# Analyse de la variance

- Exemple1: Peers influence (influence de paire)
- Est-ce que avoir des amis qui fume la cigarette est associée à une grande consommation de cigarettes?
- Variable indépendante : Combien de tes amis fument la cigarette?
- Aucun, certains, Tous
- Variable dépendante: Combien de cigarettes fumes-tu dans une journée typique?
- 1, 2, 3 . . .
- Le but de notre analyse est de déterminer si les différences (le cas échéant) entre ces trois groupes sont significatives et si
- elles peuvent être caractérisées comme une association statistiquement significative.
- L'ANOVA nous permet de le faire.

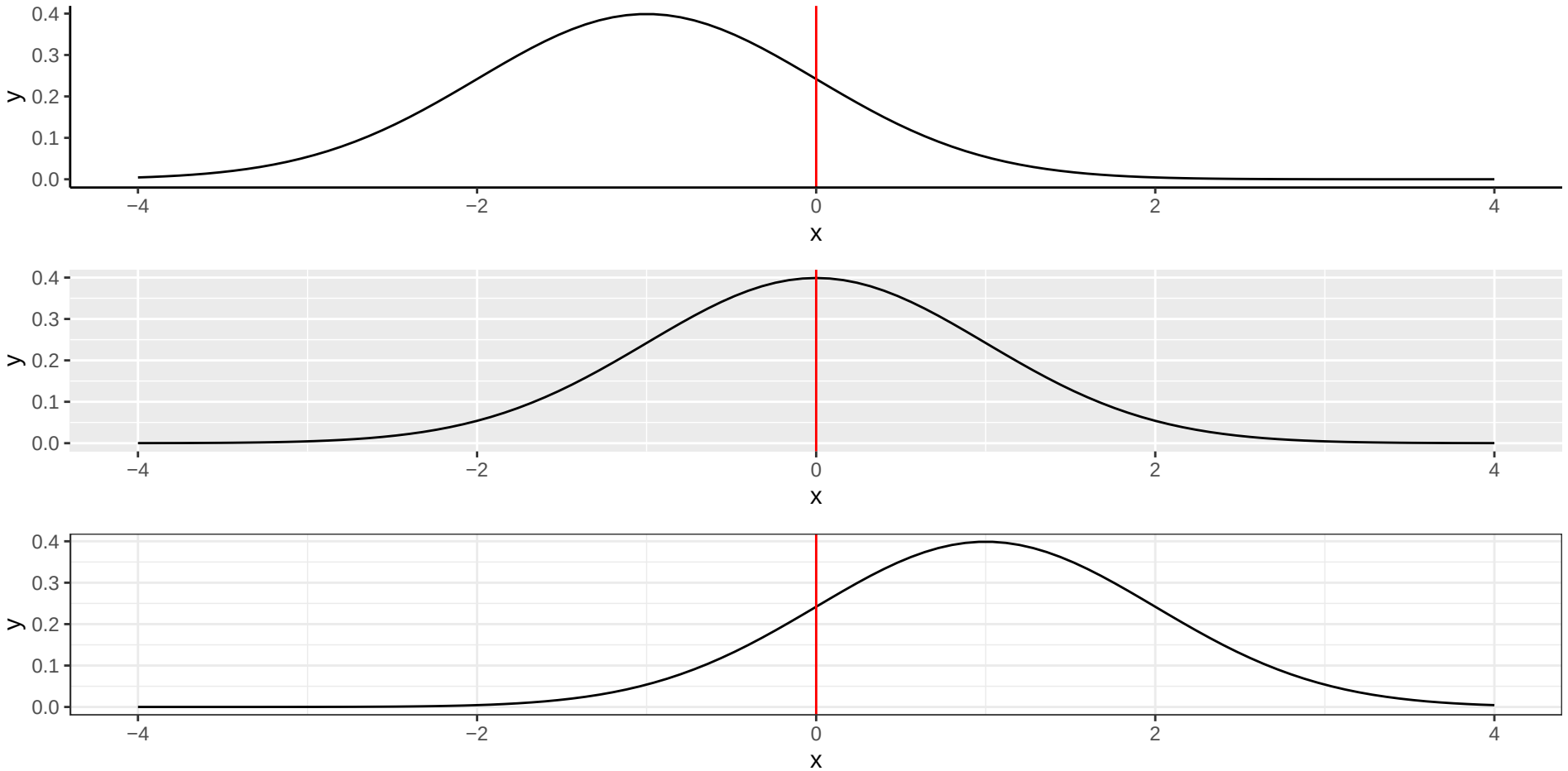
# Analyse de la variance

- Exemple 2: Lien entre le fait de fumer la cigarette et l'acceptation de la politique de la cigarette au campus
- Variable indépendante: As-tu déjà fumé?
- Fume actuellement, Avait fumé, N'a jamais fumé
- Variable dépendante: Sur une échelle de 0 (ne supporte pas) à 10 (support total), que est ton support de la politique?
- 0, 1, 2, 3, ...
- l'erreur-type de la différence de la moyenne était importante pour comparer deux groupes.
- Dans le cas de l'analyse de la variance, on va se référer à la **somme des carrés**

# Un résumé simple de l'ANOVA

- Voici les résultats obtenus dans 3 groupes (de 20 personnes chacun) sur l'interdiction de fumer sur le campus
- Le score va de -4 (en total désaccord) à 4 (support total)

# Un résumé simple de l'ANOVA



# Un résumé simple de l'ANOVA

On voit clairement que :

- Le premier groupe a une grande aversion de la politique alors que le groupe 2 semble neutre et le groupe 3 est en total accord avec la politique
- Vous pouvez aussi voir qu'il y a un certain chevauchement entre les trois groupes (variance)
- La question qu'on se pose est donc la suivante:
- est-ce que le degré de chevauchement est tel que les groupes ne sont pas si différents?
- L'analyse de la variance répond à cette question en décomposant la variance totale dans nos données (l'ensemble des trois groupes) en :
- une variance à l'intérieur de chaque groupe et
- une variance entre les groupes

# Un résumé simple de l'ANOVA

- Vous voyez ainsi que si le rapport de la variation entre les groupes et de la variation à l'intérieur des groupes est assez grand, Nous pouvons affirmer que la différence entre les groupes est significative
- De toute évidence, quelle que soit la variable qui divise les répondants en groupes 1, 2 et 3, elle influence leurs opinions en faveur d'une interdiction de fumer sur le campus.

# Analyse de la variance

- Dans le cas de l'analyse de la variance, on va se référer à la **somme des carrés**
- La somme des carrés est un concept qui nous permet de comprendre la variance et une étape dans le calcul de la variance
- Elle vaut : $\sum(X - \bar{X})^2$
- Divisé par N (ou N-1) nous donne la variance
- La racine carré de la variance nous donne l'écart-type

# Analyse de la variance

Voici les résultats obtenus de l'échantillon de 15 étudiants sur la consommation de la cigarette et la politique "sans cigarette au campus":

| Fumeurs actuels | Anciens fumeurs | Jamais fumé |
|---|---|---|
| 2 | 4 | 6 |
| 3 | 5 | 7 |
| 4 | 6 | 8 |
| 5 | 7 | 9 |
| 6 | 8 | 10 |
| $N_1 = 5$ | $N_2 = 5$ | $N_3 = 5$ |
| $\bar{X}_1 = 4$ | $\bar{X}_2 = 6$ | $\bar{X}_1 = 8$ |

- $N = 15$
- $\bar{X}_{total} = 6$

# Analyse de la variance : Étape 1

1. Somme totale des carrés (total sum of squares, SST)

$$SCT = \sum (X - \bar{X}_{total})^2$$

# Analyse de la variance : Étape 1

| X | X - $\bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 2 | 2- 6 = -4 | $(-4)^2 = 16$ |
| 3 | 3 - 6 = -3 | $(-3)^2 = 9$ |
| 4 | | |
| 5 | | |
| 6 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |

# Analyse de la variance : étape 2

2. La deuxième étape de notre analyse consiste à déterminer le degré de variation au sein de chaque groupe: **Somme des carrés Intra-groupe** (Within-group sum of squares, SSW)

$$SC_{Intra} = \sum (X - \bar{X}_k)^2$$

- k faisant référence à chaque groupe

# Analyse de la variance : étape 2

- Pour le groupe des **fumeurs actuels** (moyenne = 4)

| X | X - $\bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 2 | 2- 4 = -2 | $(-2)^2 = 4$ |
| 3 | 3- 4 = -1 | $(-1)^2 = 1$ |
| 4 | | |
| 5 | | |
| 6 | | |

- Somme des carrés = 10

# Analyse de la variance : étape 2

- Pour le groupe des **anciens fumeurs** (moyenne = 6)

| X | X - $\bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 4 | 4- 6 = -2 | $(-2)^2 = 4$ |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |

- Somme des carrés = 10

# Analyse de la variance : étape 2

- Pour le groupe de **Ceux qui n'ont jamais fumé** (moyenne = 8)

| X | X - $\bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 6 | 6 - 8 = -2 | $(-2)^2 = 4$ |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |

- Somme des carrés = 10

# Analyse de la variance : étape 2

2. La deuxième étape de notre analyse consiste à déterminer le degré de variation au sein de chaque groupe: **Somme des carrés Intra-groupe** (Within-group sum of squares, SSW)

- $SS_{Intra} = 10 + 10 + 10 = 30$

# Analyse de la variance : étape 3

3. **Somme des carrés Inter-groupes** (Between-group sum of squares, SSB)

$$SC_{Inter} = \sum N_k(\bar{X}_k - \bar{X}_{total})^2$$

- La formule nous dit que pour chaque groupe, nous devons multiplier le nombre de cas dans ce groupe par la différence au carré entre la moyenne du groupe et la moyenne totale, puis les additionner.
- Le résultat est la somme des écarts au carré des moyennes du groupe par rapport à la moyenne totale.
- Encore une fois, rappelez-vous que l'indice k fait référence à un groupe spécifique.

# Analyse de la variance : étape 3 (dernière étape)

3. **Somme des carrés Inter-groupes** (Between-group sum of squares, SSB)

| Groupe | $N_k$ | $\bar{X}_k$ | $N_k(\bar{X}_k - \bar{X}_{total})^2$ |
|---|---|---|---|
| 1 | 5 | 4 | $5 * (4-6)^2 = 20$ |
| 2 | 5 | 6 | $5 * (6-6)^2 = 0$ |
| 3 | 5 | 8 | $5 * (8-6)^2 = 20$ |

- $SC_{Inter} = 40$

# Analyse de la variance : étape 3

- Premier constat: $SCT = SC_{Intra} + SC_{Inter}$
- Dans une base de données, il y a une certaine quantité de variance (SCT) qui est partagée entre
- entre groupes (SC_Inter)
- Et à l'intérieur des groupes (SC_Intra)

# Analyse de la variance : étape 4

4. **Calcul des carrés moyens**

- Les sommes des carrés peuvent être basées sur des échantillons de petite ou grande taille entre les groupes, il convient de corriger cela.
- Ceci se fait en calculant les degrés de liberté dans les données
- Degré de liberté totaux = N - 1
- Degré de liberté intergroupes = k - 1
- degré de liberté intragroupe = $N_{total}$ - k

# Analyse de la variance : étape 4

4. **Calcul des carrés moyens**

- On divise chaque somme des carrés par son degré de liberté pour avoir les variances:
- Variance totale $= \frac{\sum(X-\bar{X}_{total})^2}{N-1}$
- Variance intra-groupe $= \frac{\sum(X-\bar{X}_k)^2}{N-k}$
- Variance intergroupe $= \frac{\sum N_k(\bar{X}_k-\bar{X}_{total})^2}{k-1}$
- Ces variances sont souvent appelées **somme moyenne des carrés** ou simplement **carrés moyens**

# Analyse de la variance : étape 4

4. **Calcul des carrés moyens**

- Variance totale = 70/(15 - 1) = 70 / 14 = 5
- Variance intragroupe = 30/(15 - 3) = 30/12 = 2.5
- Variance intergroupe = 40 / (3 -1) = 40 / 2 = 20

# Analyse de la variance : étape 5 (dernière étape)

5. **Calcul de la statistique F**

- $F - ratio = \frac{\text{Variance intergroupe}}{\text{Variance intra-groupe}}$
- Décision : Comparer le F calculé au F lu dans la table de Fisher au degré de significativité souhaité.
- La table de Fisher a deux degré de liberté:
- Le degré de liberté du numérateur (k - 1)
- Le degré de liberté du dénominateur (N - k)

# Analyse de la variance : étape 5 (dernière étape)

5. **Calcul de la statistique F**

- F (calculé) = 20/2.5 = 8
- F (lu, pour alpha = 0.05, dln = 2, dld = 12) = 3.89
- Décision: Nous rejetons l'hypothèse nulle car F calculé est supérieur au F lu

# Intensité de l'association

# Intensité de l'association

- L'intensité de l'association est mesurée par le ratio de corrélation encore noté $E^2$
- $E^2 = \frac{\text{Somme des carrés intergroupe}}{\text{Somme totale des carrés}}$
- $E^2 = \frac{SC_{Inter}}{SCT}$

# Intensité de l'association

- Dans notre exemple, cette intensité vaut: $40/70 = .57$
- C'est une association forte entre le statut de consommation de la cigarette et l'attitude envers l'interdiction de fumer la cigarette sur le campus

# En conslusion

- L'analyse de la variance permet de mesurer l'association entre une variable dépendante (ratio/intervalle) et une variable indépendante catégorielle (nominale ou ordinale)
- Si le nombre de groupe est égale à 2, on se retrouve dans le cas de la comparaison de moyenne

# Tables statistiques usuelles

# Tables usuelles

http://www.hec.unil.ch/mbrulhar/Tables%20statistiques%20usuelles.pdf