Analyse de données quantitatives (LPOLS1221)
Séance 9 : Analyse de variance (ANOVA) & tests post-hoc

Rappel théorique

Objectif : L'ANOVA de type simple ou avec 1 facteur à k groupes (*One-way ANOVA*) teste l'existence d'une **association entre une variable quantitative et une variable qualitative avec trois modalités ou plus**. Comme les autres tests d'association, l'ANOVA part de l'hypothèse nulle (H_0) d'absence de relation. Cela se traduit par une absence de différence significative entre les moyennes $H_0: \bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \cdots = \bar{X}_n$.

Conditions d'application :

- Variable dépendante **quantitative** (possible aussi sur des variables ordinales) et **normalement distribuée** si l'échantillon est petit (<50 observations)
- Variable explicative/indépendante **catégorielle avec au moins 3 modalités**
- Condition d'**homoscédasticité** : la variance des groupes à comparer est égale (à vérifier via le test de Levene)
- Condition d'**indépendance** entre les groupes que l'on veut comparer.

Etapes :

1. Analyser et comparer les moyennes de l'échantillon (tableaux descriptifs)
2. Vérifier la condition d'homoscédasticité → **Test de Levene** (H_0 : Toutes les variances sont égales).
    - Signification < seuil : on peut rejeter H_0 et conclure qu'au moins une des variances est différente des autres et donc elles ne sont pas toutes égales.
    - Signification > seuil : on ne peut pas rejeter H_0 et donc on doit procéder selon les prémisses de H_0 (par défaut)

3. **Si l'on ne peut pas conclure à l'existence d'une différence significative entre les variances** (Sig. > seuil), on lit les résultats de l'ANOVA, pour tester l'égalité des moyennes : on utilise le niveau de signification (p-valeur) associé à la statistique F et on l'interprète à partir du seuil d'erreur que l'on est prêt à accepter (0,1% ; 1% ; 5%).

$$F = \frac{\sum N_g(\bar{X}_g - \bar{X}_t)^2/(k-1)}{\sum(\bar{X}_i - \bar{X}_g)^2/(n-k)} = \frac{variance\ inter-groupe}{variance\ intra-groupe}$$

Plus F se rapproche de 0, moins il est probable que l'on puisse rejeter H_0.

- Sig. < seuil : on peut rejeter H_0 et conclure qu'au moins une des moyennes est différente des autres
- Sig. > seuil : on ne peut pas rejeter H_0 d'égalité des moyennes et donc on ne peut pas conclure qu'il y a une association entre les variables.

Si l'on conclut qu'au moins une des moyennes est différente (Sig. < seuil), il faudra voir laquelle/lesquelles. Pour cela, on effectue un test post-hoc avec présomption de

variances égales (par exemple Bonferroni, Schéffé, Tukey), pour comparer les moyennes, deux par deux dans la même logique (Sig. vs. seuil).

Ensuite, nous pouvons calculer l'intensité de l'association entre les deux variables, avec le coefficient R, dont le carré exprime la part de variance expliquée par la variable indépendante (variable de groupes). Il se calcule à la main :

$$R^2 = \frac{\sum N_g(\bar{X}_g - \bar{X}_t)^2}{\sum(X_i - \bar{X}_t)^2} = \frac{somme\ des\ carrés\ inter - groupe}{somme\ des\ carrés\ totale}$$

$$R = \sqrt{R^2} = \sqrt{\frac{\sum N_g(\bar{X}_g - \bar{X}_t)^2}{\sum(X_i - \bar{X}_t)^2}}$$

A titre indicatif, un coefficient R situé autour de 0.1 sera faible, autour de 0.3 il indiquera une effet de taille moyenne et autour de 0.5 il indiquera un effet fort.

4. **Si l'on conclut qu'au moins une des variances est différente** (Sig. < seuil), alors on ne peut pas utiliser une ANOVA (qui suppose l'homoscédasticité). Dans ce cas, on utilise un test robuste à l'héteroscédasticité (Welch ou Brown-Forsythe), qui s'interprète de la même manière que l'Anova.

   Si, lors de ces tests robustes, on conclut qu'au moins une des moyennes est différente (Sig. < seuil), il faudra voir laquelle/lesquelles. Pour cela, on effectue un test post-hoc avec présomption de variances inégales (par exemple Dunnet's T3 ou C), pour comparer les moyennes, deux par deux dans la même logique (Sig. vs. seuil).

Manipulations SPSS

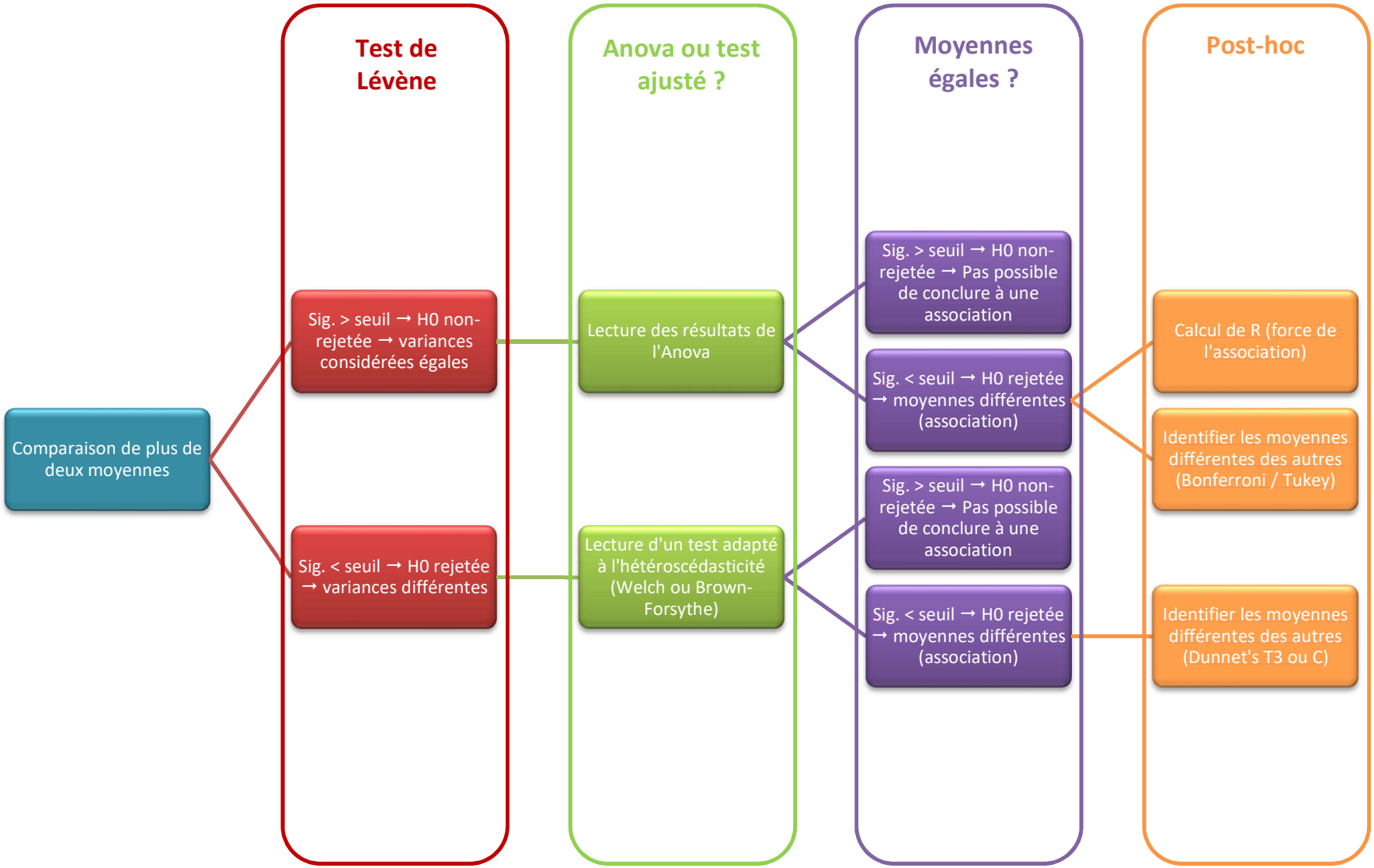
1. « Analyse » (Analyze) → « Comparer les moyennes » (Compare means) → « ANOVA à 1 facteur » (One way ANOVA)

2. Dans l'encadré « Variable(s) dépendantes », insérer la variable quantitative à partir de laquelle on veut calculer et comparer les moyennes observées pour plus de deux groupes spécifiques.

3. Dans l'encadré « critère », insérer la variable catégorielle constitutive des groupes.

4. Dans l'onglet « option », sélectionnez
   - « Caractéristiques » (Descriptives) : fournit quelques statistiques descriptives
   - « Test d'homogénéité des variances » (Homogeneity of variance test)
   - Tests de Welch et Brown-Forsythe : à lire en cas de variances inégales
   - Diagramme des moyennes (bien que l'on préfère le box-plot).

5. Dans l'onglet post-hoc, sélectionnez Tukey ou Bonferonni pour un test post-hoc en cas d'homoscédasticité et le T3 de Dunnett en cas d'hétéroscédasticité.


Test de Lévène
Anova ou test ajusté ?
Moyennes égales ?
Post-hoc
Comparaison de plus de deux moyennes
Sig. > seuil → H0 non-rejetée → variances considérées égales
Sig. < seuil → H0 rejetée → variances différentes
Lecture des résultats de l'Anova
Lecture d'un test adapté à l'hétéroscédasticité (Welch ou Brown-Forsythe)
Sig. > seuil → H0 non-rejetée → Pas possible de conclure à une association
Sig. < seuil → H0 rejetée → moyennes différentes (association)
Sig. > seuil → H0 non-rejetée → Pas possible de conclure à une association
Sig. < seuil → H0 rejetée → moyennes différentes (association)
Calcul de R (force de l'association)
Identifier les moyennes différentes des autres (Bonferroni / Tukey)
Identifier les moyennes différentes des autres (Dunnet's T3 ou C)

Applications

Utilisez la base de données de l'Eurobaromètre 2017 pour réaliser ces exercices.

Nous nous intéressons à la taille des ménages dans les pays fondateurs de l'Union Européenne (**eu6**).

1. La variable portant sur la taille des ménages (**d40abc**) comporte des codes pour les valeurs manquantes qui peuvent biaiser l'analyse. Creez-en une nouvelle, qui n'est pas affectée par de telles valeurs et qui s'appellera **taille_menage**.

2. Où comptent les ménages le plus de membres (**taille_menage)** : dans les grandes villes, dans les petites villes ou bien à la campagne (**d25**) ?

3. Y a-t-il une relation entre la taille du ménage (**taille_menage)** et la classe sociale (**classe_soc** – utilisez la syntaxe de la variable créée au TP7) ?

4. On voudrait savoir si les individus les plus satisfaits de leur vie en général proviennent de ménages nombreux ou peu nombreux. La variable portant sur la satisfaction de vie (**d70**) comporte des codes pour les valeurs manquantes qui peuvent biaiser l'analyse. Creez-en une nouvelle, qui n'est pas affectée par de telles valeurs et qui s'appellera **satisfaction**. Ensuite, testez son association avec la variable **taille_menage**.

5. En termes de taille (**taille_menage**), quels sont les ménages les plus favorables à l'Union Européenne (**opinion_ue** – utilisez la syntaxe de la variable créée au TP3) ?

   Que peut-on conclure par rapport à la taille du ménage en Allemagne, Belgique, France, Italie, Luxembourg et Pays-Bas ?