Logiciels pour la statistique : ANOVA avec SAS®

Contexte.

- comparaison de moyennes d'une variable aléatoire quantitative dans différents groupes (variable qualitative).
- 2 groupes $\Longrightarrow$ test de Student.
- $>$2 groupes $\Longrightarrow$ ANOVA.

Contexte de l'ANOVA.

- Lors de la comparaison de plus de 2 groupes, il n'est plus possible d'utiliser le test de Student.
- ANOVA : ANalysis Of Variance : on analyse quelle part de la variance totale peut être expliquée par le fait que les observations proviennent de différents groupes.

- X variable quantitative continue.
- Y variable qualitative à $k > 2$ niveaux (marqueur de groupe).
- On note $X_1, \cdots, X_k$ la variable X pour chacun des groupes.
- $\mu_1, \cdots, \mu_k$: moyenne de X selon le groupe.
- $\sigma_1, \cdots, \sigma_k$: écart-type de X selon le groupe.

ANOVA.

Modèle : $\forall 1 \leq j \leq k,\ X_j = \mu_j + \varepsilon$ où ε : variable suivant une loi normale centrée.

Sur les observations : $\forall 1 \leq j \leq k\ \forall 1 \leq i \leq n_k,\ x_{i,j} = \overline{x}_j + e_{i,j}$

Hypothèses :

- H_0 : $\mu_1 = \mu_2 = \cdots = \mu_k$.
- H_1 : au moins une moyenne est différente des autres.

Conditions d'application de l'ANOVA.

1. Chaque échantillon est composé d'observations indépendantes.
2. Les k échantillons sont indépendants.
3. $\sigma_1 = \cdots = \sigma_k$.
4. $\forall 1 \leq j \leq k,\ X_j \sim \mathcal{N}(\mu_j, \sigma_j)$.

Cela fait pas mal de choses à vérifier, notamment il faudrait faire k tests de normalité.

$\Longrightarrow$ Résidus

Conditions d'application de l'ANOVA.

1. L'échantillon des résidus est composé d'observations indépendantes.
2. Homoscédasticité : la variance des résidus est la même pour tous les groupes.
3. $\varepsilon \sim \mathcal{N}(0, \sigma)$ où σ^2 est la variance commune des résidus.

Vérification des conditions d'application.

1. Protocole de récolte des données.
2. Test de Shapiro-Wilk sur les résidus.
3. Test de Bartlett d'égalité des k variances.

Remarque

- *On teste la normalité en premier car s'agit d'une condition d'application du test de Bartlett.*
- *Le test de Bartlett peut se faire directement sur les X car ils ont la même variance que les résidus.*
- *L'ANOVA est relativement robuste à la non-normalité des résidus donc un Q-Q plot peut être intéressant.*
- *L'ANOVA n'est pas robuste à l'hétéroscédasticité.*

Que faire si les conditions d'application ne sont pas vérifiées.

1. ANOVA pour données répétées, modèles mixtes.
2. Q-Q plot : test relativement robuste à la non-normalité.
   - test non-paramétrique : Kruskal-Wallis.
   - transformation des données.
3. Kruskal-Wallis ? Nécessite que les formes des distributions soient similaires.

Test de Bartlett.

Hypothèses :

- $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.
- H_1 : au moins une variance est différente des autres.

Statistique de décision :

- notée B,
- Sous H_0, B suit asymptotiquement une loi du χ^2 à $k-1$ degrés de liberté.
- En pratique : tous les effectifs ≥ 5.

Conditions d'application :

1. observations indépendantes.
2. normalité dans chaque groupe (résidus).

Test de Bartlett sous SAS®.

Il peut être lorsqu'on fait une ANOVA grâce à des options de la PROC utilisée.

Remarque

Comme on utilise ce test pour vérifier les conditions d'application de l'ANOVA, il faut que la p-value soit > 0.05.

Test de l'ANOVA.

Hypothèses :

- H_0 : $\mu_1 = \mu_2 = \cdots = \mu_k$.
- H_1 : au moins une moyenne est différente des autres.

Statistique de décision :

- notée F,
- Sous H_0, F suit une loi de Fisher à $k-1$, $N-k$ degrés de liberté où :
  - k est le nombre de groupes (ie de modalité de Y).
  - N est l'effectif total.

ANOVA sous SAS®.

Première méthode

Utilisation de PROC ANOVA.

- ANOVA à un facteur.
- ANOVA à plusieurs facteurs si les effectifs sont équilibrés.

PROC ANOVA.

```
PROC ANOVA DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
run ;
```

Avec le test de Bartlett :

```
PROC ANOVA DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        means Parcelle / HOVTEST=bartlett ;
run ;
```

ANOVA sous SAS®.

Seconde méthode

Utilisation de PROC GLM.

- ANOVA à un facteur.
- ANOVA à plusieurs facteurs.

PROC GLM.

```
PROC GLM DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
run ;
```

Avec le test de Bartlett :

```
PROC GLM DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        means Parcelle / HOVTEST=bartlett ;
run ;
```

Vérification de la normalité.

Avec la PROC ANOVA : on ne peut pas extraire les résidus :

```
PROC UNIVARIATE DATA=mais1 normal ;
    var Hauteur ;
    class Parcelle ;
run ;
```

Vérification de la normalité.

Avec la PROC GLM : on peut pas extraire les résidus :

```
PROC GLM DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        output out=mais2 r=residus ;
run ;

PROC UNIVARIATE DATA=mais2 normal ;
        var residus ;
run ;
```

Après l'ANOVA ?

2 possibilités selon le résultat de l'ANOVA :

- On accepte H_0 : il n'y a pas de différence significative entre les moyennes des différents groupes.
  - Conclusion : la variable Y n'a pas d'effet sur la variable X.
  - Si on fait une erreur : deuxième espèce : β. Puissance a posteriori ?
- On rejette H_0 : les moyennes ne sont pas toutes égales.
  - Pour quels groupes y-a-t'il une différence ?
  - Il faut refaire des tests dit tests post-hoc ou tests de comparaisons multiples.
  - Comparaison de tous les groupes ou uniquement avec un groupe de référence ?

Comparaisons multiples sans groupe de référence.

- On envisage toutes les comparaisons possibles.
- k groupes $\Longrightarrow (k-1)+(k-2)+\cdots+2+1$ comparaisons.
- Pour chaque comparaison, on effectue un test de Student.
- Nécessité de faire une correction sur les p-valeurs des tests : prendre en compte le fait que les mêmes échantillons sont utilisés plusieurs fois.
- Corrections possibles :
  - Bonferroni : divise la p-value par le nombre de tests effectués. Trop conservatrice.
  - Tukey.

Comparaisons multiples avec groupe de référence.

- On envisage les comparaisons entre le groupe de référence et les autres groupes.
- k groupes $\Longrightarrow (k-1)$ comparaisons.
- Pour chaque comparaison, on effectue un test de Student.
- Nécessité de faire une correction sur les p-valeurs des tests : prendre en compte le fait que les mêmes échantillons sont utilisés plusieurs fois.
- Corrections possibles :
  - Bonferroni.
  - Dunnett.

Comparaisons multiples sous SAS®.

```
PROC ANOVA DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        means Parcelle / TUKEY ;
        run ;
```

Comparaisons multiples sous SAS®.

```
PROC ANOVA DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        means Parcelle / DUNNETT ;
run ;

PROC ANOVA DATA=mais1 ;
        class Parcelle ;
        model Hauteur=Parcelle ;
        means Parcelle / DUNNETT('Nord') ;
run ;
```

Comparaisons multiples sous SAS®.

```
PROC GLM DATA=mais1 ;
    class Parcelle ;
    model Hauteur=Parcelle ;
    means Parcelle / TUKEY ;
    run ;
```

Comparaisons multiples sous SAS®.

```
PROC GLM DATA=mais1 ;
      class Parcelle ;
      model Hauteur=Parcelle ;
      means Parcelle / DUNNETT ;
run ;

PROC GLM DATA=mais1 ;
      class Parcelle ;
      model Hauteur=Parcelle ;
      means Parcelle / DUNNETT('Nord') ;
run ;
```

ANOVA à 2 facteurs.

```
PROC GLM DATA=mais1 ;
      class Parcelle Verse_Traitement ;
      model Hauteur=Parcelle Verse_Traitement ;
      lsmeans Parcelle Verse_Traitement / ADJUST=TUKEY ;
      output out= mais2 r=residus ;
run ;

PROC UNIVARIATE DATA=mais2 normal ;
      var residus ;
run ;
```

ANOVA à 2 facteurs avec interaction.

```
PROC GLM DATA=mais1 ;
      class Parcelle Verse_Traitement ;
      model Hauteur=Parcelle Verse_Traitement
Parcelle*Verse_Traitement ;
      lsmeans Parcelle Verse_Traitement
Parcelle*Verse_Traitement/ ADJUST=TUKEY ;
      output out=mais2 r=residus ;
run ;

PROC UNIVARIATE DATA=mais2 normal ;
      var residus ;
run ;
```