Chapitre 4

La validation des hypothèses d'ANOVA à un facteur

Dans le modèle standard d'ANOVA, on a fait quelques hypothèses. Pour que les résultats de l'analyse effectuée soient fiables, il est nécéssaire que ces hypothèses soient vérifiées. En pratique, il faut valider ces hypothèses à l'aide d'outils statistiques. Dans ce chapitre, on présente quelques procédures pratiques pour valider les hypothèses sous-jacentes d'une ANOVA. Une procédure répons à une question bien particulière.

4.1 Planification d'une expérience complètement randomisée

Dans une analyse de variance à un facteur l'expérience est complètement randomisée. C'est-à-dire que les unités expérimentales sont réparties entre les modalités du facteur à l'étude au hasard. Souvent une bonne planification fait en sorte que les hypothèses de base sont respectées.
Pour discuter de planification d'expérience il faut définir les termes suivants:

1 L'**unité expérimentale** est l'entité qui reçoit un traitement;

2 Un **traitement** est une combinaison de modalités des facteurs à l'étude; dans une analyse de variance à un facteur un traitement est simplement une modalité du facteur;

3 La **randomisation** fait en sorte que le traitement soit assigné à une unité expérimentale au hasard.

Dans l'expérience qui compare deux diètes pour des rats, supposons que les 20 rats de l'expérience arrive tous ensemble dans une grande cage. Supposons que l'on dispose de 20 cages individuelles, 10 qui fournissent la diète 1 et 10 la 2.

Planification 1: On pourrait prendre prendre les 10 premiers rats de la grosse cage et les mettre dans des cages individuelles pour la diète 1. Les 10 restants seraient alors associés à la deuxième diète. L'effet diète est ici confondu avec l'ordre de sortie de la cage. C'est peut-être les rats les plus actifs qui sont sortis en premier. Ainsi les 2 échantillons ne sont pas identiques au début de l'expérience.

Planification 2: On utilise la randomisation. C'est-à-dire que l'on assigne au hasard les traitements aux unités expérimentales. Pour ce faire on permute au hasard 10 "1" et 10 "2". Les instructions R pour faire cela sont:

```
sample(c(rep(1,10),rep(2,10)),20,replace=FALSE)
 [1] 2 1 1 1 1 2 2 1 1 2 2 1 1 1 2 2 1 2 2 2
```

Le résultat donne l'assignation des rats: le premier tiré reçoit la deuxième diète. Ceux tirés en position 2 à 5 reçoivent la 1; les positions 6 et 7 reçoivent la 2 etc...

La randomisation cherche à faire en sorte que les I échantillons soient, nonobstant les différences de traitement, aussi semblables que possible. Si une expérience est mal planifiée, l'interprétation d'un résultat significatif peut être problématique. Il est peut-être causé par une planification déficiente. Dans l'expérience sur les rats, ceux choisis en premier étaient peut-être plus en santé. C'est peut-être la raison pour laquelle les deux échantillons ont des moyennes différentes.

Si on soupçonne qu'un facteur auxiliaire a un impact sur le résultat d'une expérience on peut incorporer ce facteur dans la planfication pour s'assurer que les échantillons soient "balancés" pour ce facteur. Ce facteur auxiliaire est appelé *bloc*. Le schéma expérimental est appelé un *schéma randomisé avec blocs*.

4.2 Est ce que les I échantillons aléatoires sont indépendants les uns des autres?

Dans la plus part des situations, la réponse à cette question dépend la façon avec laquelle on a récolté les données. L'indépendance des échantillons, appelée aussi *l'indépendance inter-échantillonale*, est donc une conséquence directe du scénario déchantillonage. Une situation standard dans laquelle cette hypothèse est violée est le cas des données. C'est à dire lorsque chaque observation dans un échantillon est reliée à une observation dans chacun des autres échantillon.

EXEMPLE **4.1** *Un chercheur en sciences médicales veut comparer deux médicaments pour réduire le taux de glycémie chez les personnes agées. Il prend des couples de personnes agées et administre à chacun de deux membres du couple un des deux médicaments. Les données ainsi récoltées ne sont clairement pas indépendantes puisque les données d'un couple sont reliées entre elles. En effet, le couple partage le quotidien. Il se peut qu'un couple fasse très attention à son alimentation alors qu'un autre couple mange un peu n'importe quoi.*

4.3 Les observations sont-elles identiquement distribuées à l'intérieur de chaque échantillon?

Ici aussi, c'est le plan d'expérience qui permet de répondre à cette question. En pratique, une situation standard pour laquelle cette hypothèse n'est pas vérifiée est lorsque les données sont obtenues séquentiellement dans le temps: d'abord Y_{i1}, puis Y_{i2}, ensuite Y_{i3}, etc. Lorsque la loi des Y_{ij} évolue dans le temps, nos données ne sont pas identiquement distribuées. Pour détecter cette situation, on peut effectuer un graphe de $\{Y_{i1}, Y_{i2}, \cdots, Y_{in_i}\}$ en fonction de $j = 1, 2, \cdots, n_i$ pour $i = 1, 2, \cdots, I$. Si ce graphe montre une tendance quelconque, on peut penser que cette hypothèse n'est pas vérifiée.

4.4 Est-ce que les observations sont indépendantes les unes des autres à l'intérieur de chaque échantillon?

Encore une fois, c'est le scénario expérimental qui rend cette hypothèse raisonnable. Le cas où les données sont recoltées séquentiellement soulève un doute concernant la véracité de cette hypothèse. En effet, il se peut que les données soient autocorrélées; c'est-à-dire que Y_{ij} soit corrélée avec $Y_{i(j+1)}$. On peut détecter cette situation en traçant le nuage de points $(Y_{ij}, Y_{i,j+1}), j = 1, 2, \cdots, n_i - 1$, ou en calculant les coefficents d'autocorrélation. Pour pouvoir répondre positivement à la question, le nuage de point ne doit montrer aucune tendance et les autocorrélations ne doivent pas être significativement différentes de 0.

4.5 Est-ce que les observations proviennent d'une loi normale?

L'hypothèse de normalité est cruciale pour l'ANOVA. En pratique, la validation de cette hypothèse est une étape importante lors de l'analyse.
Généralement, l'hypothèse de la normalitè est vérifiée sur l'ensemble des données et non pas sur chaque échantillon séparément. D'où la nécéssité de ramener toutes les observations à la même échelle pour avoir une population homogène sur laquelle on va effectuer les différents tests de normalité.
Pour $i = 1, 2, \cdots, I$, on a $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$. Définissons les résidus e_{ij} par $Y_{ij} - \mu_i$. On a alors $e_{ij} \sim \mathrm{N}(0, \sigma^2)$. Ces résidus sont estimés par $\hat{e}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i.}$.
Clairement, les $\hat{e}_{ij}$ sont normalement distribuées puisque c'est la diffrence de deux variables aléatoires normalement distribuées. Calculons l'espérance et la variance de ces estimateurs des résidus.
On a : $E[\hat{e}_{ij}] = E[Y_{ij} - \bar{Y}_{i.}] = \mu_i - \mu_i = 0$
D'autre part, on a:

$$\begin{aligned} V[\hat{e}_{ij}] &= V[Y_{ij} - \bar{Y}_{i.}] \\ &= V[Y_{ij}] + V[\bar{Y}_{i.}] - 2\mathrm{Cov}[Y_{ij}, \bar{Y}_{i.}] \end{aligned}$$

$$
\begin{aligned}
&= V[Y_{ij}] + V[\bar{Y}_{i.}] - 2\mathrm{Cov}[Y_{ij}, \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik}] \\
&= V[Y_{ij}] + V[\bar{Y}_{i.}] - 2\mathrm{Cov}[Y_{ij}, \frac{1}{n_i} Y_{ij}] \\
&= V[Y_{ij}] + V[\bar{Y}_{i.}] - 2\frac{1}{n_i} V[Y_{ij}] \\
&= \sigma^2 + \frac{\sigma^2}{n_i} - 2\frac{\sigma^2}{n_i} \\
&= \sigma^2(1 - \frac{1}{n_i})
\end{aligned}
$$

On en déduit que $\hat{e}_{ij} \sim \mathrm{N}(0, \sigma^2(1 - \frac{1}{n_i}))$, qu'on peut encore écrire:

$$\frac{\hat{e}_{ij}}{\sqrt{\sigma^2(1 - \frac{1}{n_i})}} \sim \mathrm{N}(0, 1)$$

Cette dernière relation n'est malheuseuement pas utilse pour nous puisqu'en pratique, la variance théorique σ^2 est inconnue. Elle est estimée par $\hat{\sigma}^2 = MSE$. Si MSE et $\hat{e}_{ij}$ étaient indépendants, on aurait eu:

$$\hat{\varepsilon}_{ij} = \frac{\hat{e}_{ij}}{\sqrt{MSE(1 - \frac{1}{n_i})}} \sim t_{N-I}$$

Malheuseument, ceci n'est pas le cas. En effet, MSE et $\hat{e}_{ij}$ ne sont pas indépendantes. Mais lorsque les n_i sont suffisament grands, on peut approximer la loi de $\hat{\varepsilon}_{ij}$ par une normale standard.

Les différents tests de normalité seront alors performés sur l'ensemble des résidus studentisés $\{\hat{\varepsilon}_{ij},\ i = 1, 2, \cdots, I,\ j = 1, 2, \cdots, n_i\}$.

Différents tests pour vérifier la normalité d'un ensemble de données $\{X_1, X_2, \cdots, X_n\}$ existent dans la littérature. Certains sont basés sur la densité $\phi(t) = e^{-\frac{t^2}{2}}/\sqrt{2\pi}$ comme le diagramme en boite, *boxplot* en anglais, le diagramme en tige et feuilles, *sterm and leaf plot* en anglais, l'histogramme et les tests d'ajustement khi-deux. D'autres sont basés sur la fonction de répartition $\Phi(t) = \int_{-\infty}^{t} \phi(x)dx$. Dans ce chapitre, on présente un apperçu de ces dernières.

On définit la fonction de répartition d'une variable aléatoire continue X par $F(x) = P(X \leq x)$. En pratique, cette fonction est estimée, à partir d'un échantillon aléatoire $\{X_1, X_2, \cdots, X_n\}$

par la fonction de répartition empirique définie par

$$\begin{aligned}\hat{F}_n(x) &= \frac{\text{nombre d'observations plus petites ou égales à } x}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}\end{aligned}$$

Lorsque n tend vers l'infini, la fonction de répartition empirique $\hat{F}(\cdot)$ tend vers la vraie fonction de répartition en tout point x tel que $0 < F(x) < 1$. Soit x un tel réel. Notons que pour $i = 1, \cdots, n$, la variable aléatoire $1_{\{X_i \leq x\}}$ suit une loi de Bernoulli avec un paramètre égal à $F(x)$. Applicons le théorème central limite sur les $Y_i = 1_{\{X_i \leq x\}}$. On obtient le résultat asymptotique suivant:

$$\sqrt{n}\{\hat{F}_n(x) - F(x)\} \sim \mathrm{N}(0, F(x)(1 - F(x)))$$

Soit $\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$ l'échantillon de statistiques d'ordre obtenu en ordonnant l'échantillon initial $\{X_1, X_2, \cdots, X_n\}$. Par définition, cet échantillon vérifie $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Il est facile de voir qu'on a alors $\hat{F}_n(X_{(i)}) = i/n$ pour $i = 1, 2, \cdots, n$.

4.5.1 Les coefficients d'asymétrie et d'aplatissement

Le coefficient d'asymétrie (skewness) de l'échantillon $X_1, \ldots, X_n$ est donné par

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{3/2}}.$$

Certains logiciels calculent plutôt un estimateur corrigé pour le biais

$$G_1 = \frac{\sqrt{n\,(n-1)}}{n-2}\, g_1,$$

Le coefficient d'aplatissement (kurtosis) est donné par

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2} - 3.$$

Certains logiciels calculent un estimateur corrigé pour le biais, voir

http://en.wikipedia.org/wiki/Kurtosis.

La valeur théorique de ces deux statistiques sont 0 lorsque les données sont normales.

4.5.2 La droite d'Henry

Ce test visuel est basé sur le nuage de n points $(\Phi^{-1}(i/(n+1)), X_{(i)})$. En effet, si léchantillon $\{X_1, X_2, \cdots, X_n\}$ provient d'une loi normale, on aurait $F(\cdot) = \Phi(\cdot)$ et $\Phi(X_{(i)}) \simeq \hat{F}(X_{(i)}) = i/n$ qu'on peut écrire encore $\Phi^{-1}[i/n] \simeq X_{(i)}$. Le nuage de points formera alors à peu près une droite. On utilise $i/(n+1)$ à la place de i/n pour éviter $\Phi^{-1}(0)$, qui n'existe pas. Le logiciel SAS utilise le nuage de points

$$\{(\Phi^{-1}[\frac{i-\frac{3}{8}}{n+\frac{1}{4}}], X_{(i)}), i = 1, 2, \cdots, n\}.$$

4.5.3 Le test de Shapiro et Wilk

Ce test est une approche plus approfondie du test précédent. Si le nuage de points $\{X_{(i)}, \Phi^{-1}[i/(n+1)]\}$ forme une droite, alors le coefficient de corrélation défini par

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

où $u_i = X_{(i)}$ et $v_i = \Phi^{-1}[i/(n+1)]$, ne sera pas loin de 1.mCeci équivaut à dire que r^2 ne sera pas loin de 1. On rejette alors la normalité si r^2 est loin de 1. Il existe des tables pour la distribution de r^2 sous H_0. Ces tables nous servent à calculer la valeur critique à un seuill donné, par exemple à 5% et à calculer la $p - value$ associée à un jeu de données.

4.5.4 Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est basée sur une distance entre la fonction de répartition empirique $\hat{F}(\cdot)$ et la fonction de répartition qu'on veut tester, ici $\Phi(\cdot)$. Si léchantillon

$\{X_1, X_2, \cdots, X_n\}$ provient d'une loi normale, on devrait avoir $\hat{F}(t) \simeq \Phi(t)$ pour tout réel t. En particulier, la statistique D définie par

$$D = \sup_{t\in\mathcal{R}}|\hat{F}_n(t) - \Phi(t)|$$

doit être petite. Le test de Kolmogorov-Smirnov consiste donc à rejetter la normalité si la statistique D est trop grande. Il existe des tables pour la loi D sous H_0. Ces tables nous servent à calculer la valeur critique à un seuill donné, par exemple à 5% et à calculer la $p - value$ associée à un jeu de données.

4.6 Est-ce que les variances théoriques des échantillons sont égales ou pas?

La vérification de l'hypothèse d'homogénéité des variances est une étape importante lors de la réalisation d'une ANOVA. Il existe dans la littérature plusieurs procédures pour effectuer le test $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$ vs H_1 : les variances ne sont pas toutes égales. Cependant, plusieurs de ces tests requièrent la normalité ou légalité des tailles des échantillons ($n_1 = n_2 = \cdots = n_I$). Dans ce chapitre, on présente les tests les plus utilisés dans la pratique. D'une part parcequ'ils sont programmés par SAS et d'autres part parque qu'ils sont les moins restrictifs.

4.6.1 Le test de Levene

Le test de Levene date du début des années soixantes. Il consiste à effectuer une analyse de la variance sur des données transformées. En effet, pour $i = 1, 2, \cdots, I$ et $j = 1, 2, \cdots, n_i$, définissons Z_{ij} par $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$. Le test de Levene consiste à effectuer une ANOVA sur les variables transformées Z_{ij}. Ainsi, on rejette l'hypothèse d'homogénéité des variances au seil α si $F_{obs} > F_{\alpha,I-1,N-I}$ où F_{obs} est défini par

$$F_{obs} = \frac{\sum_{i=1}^{I} n_i(\bar{Z}_{i.} - Z_{..})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})^2/(N-I)}.$$

Ce test est effectué par SAS à l'aide de l'instruction:

```
means traitement / hovtest = LEVENE ;
```

C'est ce test qui est effectué par défaut.

4.6.2 Le test de Brown et Forsythe

Ce test est une variante du test précdant. Ici, on définit les Z_{ij} par $|Y_{ij} - \tilde{Y}_{i.}|$ où $\tilde{Y}_{i.}$ est la médiane de léchantillon $\{Y_{i1}, Y_{i2}, \cdots, Y_{in_i}\}$. Le reste de la procèdure demeure inchangé. Ce test est effectué par SAS à l'aide de l'instruction:

```
means traitement / hovtest = BROWN ;
```

4.6.3 Le test de Bartlett

Le test de Bartlett, considéré comme un test de rapport de vraisemplance est basé sur la statistique L défini par

$$L = \frac{(S_1^2)^{\frac{n_1-1}{N-I}}(S_2^2)^{\frac{n_2-1}{N-I}} \cdots (S_I^2)^{\frac{n_I-1}{N-I}}}{\frac{n_1-1}{N-I}S_1^2 + \frac{n_2-1}{N-I}S_2^2 + \cdots + \frac{n_I-1}{N-I}S_I^2}$$

REMARQUE **4.1** *Le dénominateur et le numérateur de la statistique L définie par l'équation ci-haut sont les moyennes arithmétiques et géométriques respectives de $\{S_1^2, S_2^2, \cdots, S_I^2\}$ pondérées par $w_1 = (n_1 - 1)/(N - I), w_2 = (n_2 - 1)/(N - I), \cdots, w_I = (n_I - 1)/(N - I)$. Ces poids vérifient $w_1 + w_2 + \cdots + w_I = 1$.*

On rejette l'hypothèse d'homogénéité des variances si L est trop grand. Il existe des tables pour la distribution exacte de L. Néanmoins, en pratique on utilise l'approximation suivante. Posons:

$$B = \frac{-(N - I)\log(L)}{c}$$

avec

$$c = 1 + \frac{(\sum_{i=1}^{I} \frac{1}{n_i-1}) - \frac{1}{N-I}}{3(I-1)}.$$

Sous H_0, lorsque les tailles des échantillons $n_1, n_2, \cdot, n_I$ tendent vers l'infini, on obtient asymptotiquement

$$B \sim \chi^2_{I-1}.$$

On rejette donc H_0 si $B > \chi^2_{I-1,\alpha}$. Ce test est effectué par SAS à l'aide de l'instruction:

```
means traitement / hovtest = BARTLETT ;
```

4.7 Résumé

Il y a donc trois hypothèses à vérifier:

- L'indépendance intra et inter échantillons.
- La normalité des erreurs expérimentales.
- L'égalité des variances.

Pour la première hypothèse on cherche à préciser la planification de l'expérience. S'agit-il d'une expérience complètement randomisée? Quelles sont les unités expérimentales? Peut-être que l'ordre dans lequel les données ont été récoltées est associé à leur valeur.
L' hypothèse de normalité n'est pas cruciale pour la validité du test F d'homogénéité des moyennes. Si les tailles d'échantillons n_i sont grandes, la distribution de la statistique F de la table ANOVA suit approximativement une distribution $F_{I-1,\sum n_i}$ même si les données ne sont pas normales. La non normalité des données compromet la puissance du test cependant. Si cette hypothèse est violée, le test de Kruskal-Wallis, basé sur les rangs, est souvent plus puissant que le test F de la table ANOVA. C'est le cas lorsque les données contiennent des valeurs extrêmes (outliers). On peut également s'assurer que quelques valeurs extrêmes n'ont pas une influence indue sur les résultats en refaisant les analyses après avoir exclu ces données.
L'égalité des variances n'est pas vraiment cruciale pour la validité du test F d'homogénéité. Si l'expérience est à peu près balancée et si les tailles d'échantillons sont grandes on peut montrer que le test F est valide même si les variance sont inégales. Le test de Welch (disponible

dans SAS avec `means facteur/welch;`) tient compte de variances inégales. Certains auteurs suggèrent de multiplié le seuil observé du test F par 2 s'assurer de préserver le seuil lorsque les variance sont inégales.

Dans une analyse de variance, la variabilité des données ne doit pas être associée à leurs valeurs moyennes. Si les données sont des dénombrements avec une loi de Poisson alors la variance est à peu près égale à la moyenne. Il y a un lien moyenne-variance et les hypothèses sous-jacentes à l'ANOVA sont violées. Dans ce cas, deux solutions sont possibles. On peut faire une transformation pour chercher à stabiliser la variance et traiter les données avec une ANOVA (pour la Poisson la transformation racine carée fait le travail). On peut également utiliser un modèle linéaire généralisé construit spécifiquement pour la distribution de Poisson. Le mode de variation de certaines variables, que l'on décrit en termes relatifs plutôt qu'en termes absolus, peut aussi suggérer une transformation. Un modèle ANOVA postule des variations absolues additives. Des variations relatives sont en fait multiplicatives. Une transformation logarithmique peut alors s'imposer pour que les hypothèses du modèles ANOVA soient vérifiées. Une transformation complique l'analyse car c'est souvent sur l'échelle originale que les résultats doivent être interprétés.