Table des matières

Première partie

Analyse descriptive des séries chronologiques

1 Notations

Définition: Série chronologique

Suite finie de données quantitatives indexée par le temps.
Si on considère une série chronologique de longueur n :

- t_1,...,t_n désigne les n instants successifs d'observation
- y_i sera la valeur mesure à l'instant t_i (en considérant les dates d'observations équidistantes).

2 Modèles de décomposition déterministes

Deux modèles sont étudiés :

1. Le modèle additif
2. Le modèle multiplicatif

combinant chacun :

1. Une tendance f_i
2. Une composante saisonnière s_i
3. Une composante résiduelle e_i

Définition: Modèle additif

Le modèle additif prédit une étiquette sous la forme suivante :

$$y_i = f_i + s_i + e_i,\ i = 1..n$$

avec :

$$\sum_{j=1}^{p} s_j = 0 \text{ et } \sum_{i=1}^{n} e_i = 0$$

Où p désigne une période.
On utilise ce modèle quand, en reliant minima et maxima, on obtient deux droites parallèles.

Définition: Modèle additif

Le modèle multiplicatif prédit une étiquette sous la forme suivante :

$$y_i = f_i(1 + s_i)(1 + e_i),\ i = 1..n$$

avec :

$$\sum_{j=1}^{p} s_j = 0 \text{ et } \sum_{i=1}^{n} e_i = 0$$

Où p désigne une période.
On utilise ce modèle quand, en reliant minima et maxima, on obtient une sorte de cône.

3 Ajustement de la tendance

3.1 Ajustement linéaire

Formule: Méthode des moindres carrés

Elle vient de la recherche des paramètres $a, b \in \mathbb{R}$ minimisant la fonctionnelle suivante :

$$\sum_{i=1}^{n} (y_i - (at_i + b))^2$$

ce qui nous donne :

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(t_i - \bar{t})}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \\ \hat{b} &= \bar{y} - \hat{a}\bar{t} \end{aligned}$$

Formule: Méthode des deux points

Cette méthode consiste à choisir arbitrairement deux points par lesquels on fait passer une droite.
La réalisation de cette méthode se fait en général en prenant deux sous-suites, et en prenant les points médians de chaque sous-série.
Cette méthode s'avère efficace en présence de points aberrants, chose que la méthode des moindres carrés ne prend pas en compte.

Propriété: Appréciation des régression linéaire

Un moyen de qualifier la qualité de la regression linéaire est d'utiliser le coefficiet de corrélation linéaire, noté r, et défini par :

$$r = \frac{\text{cov}(y,t)}{\sigma_y \sigma_t}$$

En effet, en réécrivant l'expression, on peut montrer que :

$$r^2 = \frac{\text{Variance expliquée}}{\text{Variance totale}}$$

3.2 Ajustement polynomial

Formule: Polynôme des moindres carrés

On minimise la même fonction que précédemment, mais en cherchant cette fois non plus a et b d'une régression linéaire mais $a_i,\ i = 0, .., d$ d'un polynôme de degré d. En notant :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } T = \begin{pmatrix} 1 & t_1 & \cdots & t_1^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^d \end{pmatrix}$$

On obtient :

$$\theta^{MC} = \begin{pmatrix} a_0 \\ \vdots \\ a_d \end{pmatrix} = ({}^tTT)^{-1} \times {}^tTY$$

3.3 Ajustement non linéaire

On a deux cas :

1. Soit on se ramène à un ajustement linéaire via un changement de variable
2. Soit on cherche à déterminer les coefficients restants via une méthode à d points (d étant le nombre de paramètres à estimer), ou en minimisant le carré des erreurs (ce qui ne donne pas toujours une formule explicite).

4 Lissage par moyenne mobile

Définition: Moyenne mobile simple

On note $MM(k)$ la série des moyennes mobiles d'ordre k de la série $(y_j)_{j=1...n}$, et on a :
— lorsque k est pair et vaut $2m$:

$$MM(k)_j = \frac{y_{j-m+1} + ... + y_j + y_{j+1} + ... + y_{j+m}}{2m}$$

— lorsque k est impair et vaut $2m+1$:

$$MM(k)_j = \frac{y_{j-m} + ... + y_j + y_{j+1} + ... + y_{j+m}}{2m+1}$$

pour $j = m+1, ..., n-m$.

Définition: Moyenne mobile centrée

La série notée $MMC(k)$ uniquement pour k pair et définie par :

$$MMC(k)_j = \frac{MM(k)_{j-1} + MM(k)_j}{2}$$

Propriété:

— La série $MM(p)$ ou $MMC(p)$ ne possède plus de composante saisonnière de période p.
— Une moyenne mobile atténue l'aplitude des fluctuations irrégulières d'une chronique.

5 Décomposition d'une série chronologique

Formule: Étapes de la décomposition

1. La désaisonnalisation
   (a) Lissage par moyennes mobiles : on construit la série des moyennes mobiles d'ordre p, la saisonnalité (centrées si p pair).
   (b) Constrution de la série des différences / quotients : observation - série des moyennes mobiles ou obs / MM
   (c) Calcul des coefficients saisonniers non centrés : moyennes des différences pour chaque saison
   (d) Centrage des coefficients saisonniers : moyennes des p coefficients non centrés, puis on centre les coefficients saisonniers.
   (e) Constrution de la série corrigée des variations saisonnières : observation - composante saisonnière (selon, bien sûr, la saison) ou division.
2. La série lissée des prévisions
   (a) Ajustement d'une tendance : regression linéaire (ou autre) sur la CVS
   (b) Construction de la série lissée des prévisions : résultat de la régression + coefficient saisonnier. $= \hat{y}_i$ (ou $= \hat{f}_i(1+\hat{s}_i)$)

Deuxième partie

Modèle linéaire gaussien simple

1 Différents rappels

Rappel : Différentes lois de probabilité

— Loi du χ^2 : On prend $Z_1, ..., Z_n$ n variables aléatoires indépendantes et de même loi $\mathcal{N}(0,1)$. Alors $S_n = \sum_{k=1}^{n} Z_k^2$ suit une loi du chi-deux à n degrés de liberté, ce qu'on note $S_n \hookrightarrow \chi_n^2$.

$\mathbb{E}(S_n) = n$ et $\mathbb{V}(S_n) = 2n$.

Le théorème de Cochran nous dit que si X, Y et Z sont trois variables alétoires positives telles que $Z = X + Y$ et que $Z \hookrightarrow \chi_n^2$ et $X \hookrightarrow \chi_p^2$ alors $Y \hookrightarrow \chi_{n-p}^2$ et on a indépendance entre X et Y.

Dans la suite, les deux variables sont indépendantes :
— Loi de Student : Si $U \hookrightarrow \mathcal{N}(0,1)$ et $V \hookrightarrow \chi_n^2$ alors $\frac{U}{\sqrt{\frac{V}{n}}} \hookrightarrow T_n$
— Loi de Fisher : Si $U \hookrightarrow \chi_p^2$ et $V \hookrightarrow \chi_q^2$ alors $\frac{U/p}{V/q} \hookrightarrow F(p,q)$
— Le carré d'une Student T_n est une loi de Fisher $F(1,n)$.

2 Définition du modèle

Définition: MLG

Dans le cadre du modèle linéaire gaussien simple, les données $x_1, ..., x_n$ ne sont pas des réalisations de variables aléatoires et on suppose que les données $y_1, ..., y_n$ sont les réalisations de n variables aléatoires $Y_1, ..., Y_n$ qui sont liées aux données $x_1, ..., x_n$ de la manière suivante :

$$\forall i \in \{1, ..., n\}, Y_i = \alpha x_i + \beta + \varepsilon_i$$

où $\alpha, \beta \in \mathbb{R}$ et où $\varepsilon_1, ..., \varepsilon_n$ sont n variables aléatoires indépendantes et même loi $\mathcal{N}(0, \sigma^2)$.

Définition: des estimateurs

On définit trois estimateurs :
— un estimateur de α qu'on notera A et qui vaut :

$$A = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

— un estimateur de β qu'on notera B :

$$B = \bar{Y} - A\bar{x}$$

(Ces deux estimateurs sont obtenus en minimisant la quantité $f(A, B) = \sum_{i=1}^{n}(Y_i - Ax_i - B)^2$)
— Un estimateur du paramètre σ^2 des ε_i qu'on note S^2 :

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - Ax_i - B)^2$$

Petite convention d'écriture :

$$d_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Proposition: Loi des différents estimateurs

Sous les hypothèses du modèle linéaire gaussien, A et B sont des estimateurs sans biais et convergents en probabilité des paramètres α et β, et on a :

$$A \hookrightarrow \mathcal{N}\left(\alpha, \frac{\sigma^2}{nd_x^2}\right)$$

$$B \hookrightarrow \mathcal{N}\left(\beta, \frac{\sigma^2(d_x^2 + \bar{x}^2)}{nd_x^2}\right)$$

Sous les mêmes hypothèses, S^2 est un estimateur sans biais de σ^2 et on a :

$$\frac{(n-2)S^2}{\sigma^2} \hookrightarrow \chi^2_{n-2}$$

Enfin, on a S^2 indépendant de A, B et $\bar{Y}$.

3 Intervalle de confiance

Proposition: Variables aléatoires pour les intervalles

Pour construire les intervalles de confiance, étant donné qu'on ne connaît pas σ^2, on a utiliser son estimateur S et "studentiser" les variables. En effet, vu que S^2 est indépendant de A et B, et vu les lois que chacun d'entre eux suit, cela est tout à fait réalisable ! Ainsi, sous les hypothèses du modèle linéaire Gaussien :

$$\frac{(A-\alpha)\sqrt{nd_x^2}}{S} \hookrightarrow T_{n-2} \text{ et } \frac{(B-\beta)\sqrt{nd_x^2}}{S\sqrt{d_x^2+\bar{x}^2}} \hookrightarrow T_{n-2}$$

(Pour retrouver ces estimateurs, il suffit de centrer et réduire A et B, puis de réutiliser la méthode de construction d'une Student)
Pour S^2, on utilise directement le fait qu'elle suive une loi du chi-deu à n-2 degrés de liberté.

Formule: Intervalles de confiance de α, β et σ^2

$$IC_{1-\delta}(\alpha) = \left[a \pm \frac{st_{n-2,\delta/2}}{\sqrt{nd_x^2}}\right]$$

$$IC_{1-\delta}(\beta) = \left[b \pm st_{n-2,\delta/2}\frac{\sqrt{\bar{x}^2+d_x^2}}{\sqrt{nd_x^2}}\right]$$

$$IC_{1-\delta}(\sigma^2) = \left[\frac{(n-2)s^2}{k_{1-\delta/2}}; \frac{(n-2)s^2}{k_{\delta/2}}\right]$$

4 Tests dans le modèle linéaire gaussien

On prend toujours comme hypothèse le modèle linéaire gaussien.

4.1 Test significatif du lien linéaire

Définition: Statistique de ce test

On veut tester l'hypothèse

$$H_0 : \ll \alpha = 0 \gg$$

contre l'alternative

$$H_1 : \ll \alpha \neq 0 \gg$$

On a pour cela la statistique suivante :

$$\frac{(A-\alpha)\sqrt{nd_x^2}}{S} \sim T_{n-2}$$

On construit donc la statistique de test suivante :

$$Z = \frac{A\sqrt{nd_x^2}}{S} \underset{H_0}{\sim} T_{n-2}$$

qui sous H_1 ne suit plus la même loi.

Proposition: Zone de rejet et stratégie

On fixe un risque δ et on calcule $t_{n-2,\delta/2}$ tel que :

$$\mathbb{P}(|Z| < t_{n-2,\delta/2}) = 1 - \delta$$

La zone de rejet de H_0 au risque δ est alors de la forme $\{|Z| > t_{n-2,\delta/2}\}$.
On calcule une réalisation de Z :

$$z = \frac{a\sqrt{nd_x^2}}{s} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

et on décide ainsi :
— si $|z| \leq t_{n-2,\delta/2}$, on accepte H_0 au risque δ.
— si $|z| > t_{n-2,\delta/2}$, on rejette H_0 au risque δ.

4.2 Test d'un modèle linéaire spécifique

Définition: Statistique de ce test

On veut tester l'hypothèse

$$H_0 : \ll \alpha = \alpha_0 \text{ et } \beta = \beta_0 \gg$$

contre l'alternative

$$H_1 : \ll \alpha \neq \alpha_0 \text{ ou } \beta \neq \beta_0 \gg$$

On a pour cela la statistique suivante :

$$\frac{\sum_{i=1}^n ((A-\alpha)x_i + (B-\beta))^2/2}{\sum_{i=1}^n (Y_i - Ax_i - B)^2/(n-2)} \sim F(2, n-2)$$

On construit donc la statistique de test suivante :

$$Z = \frac{\sum_{i=1}^n ((A-\alpha_0)x_i + (B-\beta_0))^2/2}{\sum_{i=1}^n (Y_i - Ax_i - B)^2/(n-2)} \underset{H_0}{\sim} F(2, n-2)$$

qui sous H_1 ne suit plus la même loi.

Proposition: Zone de rejet et stratégie

On fixe un risque δ et on calcule $f_{2,n-2,\delta}$ tel que :

$$\mathbb{P}(Z < f_{2,n-2,\delta}) = 1 - \delta$$

La zone de rejet de H_0 au risque δ est alors de la forme $\{Z > f_{2,n-2,\delta}\}$.
On calcule une réalisation de Z :

$$\begin{aligned} z &= \frac{n-2}{2} \frac{\sum_{i=1}^n ((a-\alpha_0)x_i + (b-\beta_0))^2}{\sum_{i=1}^n (y_i - ax_i - b)^2} \\ &= \frac{n-2}{2} \frac{n(b-\beta_0)^2 + 2n\bar{x}(a-\alpha_0)(b-\beta_0) + (a-\alpha_0)^2 \sum_{i=1}^n x_i^2}{(\sum_{i=1}^n y_i^2 - n\bar{y}^2) - a^2 (\sum_{i=1}^n x_i^2 - n\bar{x}^2)} \end{aligned}$$

et on décide ainsi :
— si $|z| \leq f_{2,n-2,\delta}$, on accepte H_0 au risque δ.
— si $|z| > f_{2,n-2,\delta}$, on rejette H_0 au risque δ.

5 Prévision d'une valeur

On cherche ici à estimer une valeur inconnue y_0 à partir d'une donnée x_0. On va associer à y_0 une variable aléatoire Y_0 définie par :

$$Y_0 = Ax_0 + B + \varepsilon_0$$

avec $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$. On va également dire que y_0 est la réalisation d'une variable aléatoire $\hat{Y}_0$ définie par :

$$\hat{Y}_0 = Ax_0 + B$$

Ainsi, puisque $\mathbb{E}(Y_0) = \alpha x_0 + \beta = \mathbb{E}(Y_0)$, alors $\hat{Y}_0$ est un estimateur de $\mathbb{E}(Y_0)$. Par conséquent, $\hat{y}_0$ est à la fois une estimation de l'espérance de Y_0 et une prévision de y_0.

5.1 Intervalle de confiance pour l'espérance de Y_0

Théorème:

Dans le cadre du MLG :

$$\frac{\hat{Y}_0 - \mathbb{E}(Y_0)}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nd_x^2}}} \sim T_{n-2}$$

Proposition:

A partir de ce résultat, on peut bâtir l'intervalle de confiance pour le paramètre inconnu $\mathbb{E}(Y_0) = \alpha x_0 + \beta$. Au niveau de confiance $(1 - \delta\%)$ cet intervalle a pour expression :

$$IC_{1-\delta}(\mathbb{E}(Y_0)) = \left[y_0 \pm t_{n-2,\delta/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nd_x^2}} \right]$$

5.2 Intervalle de prévision pour une observation Y_0

Théorème:

Dans le cadre du MLG :

$$\frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nd_x^2}}} \sim T_{n-2}$$

Proposition:

A partir de ce résultat, on peut bâtir l'intervalle de confiance pour le paramètre inconnu $\mathbb{E}(Y_0) = \alpha x_0 + \beta$. Au niveau de confiance $(1 - \delta\%)$ cet intervalle a pour expression :

$$IC_{1-\delta}(Y_0) = \left[y_0 \pm t_{n-2,\delta/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nd_x^2}} \right]$$

6 Test par comparaison de modèles

6.1 Test du caractère significatif de la liaison linéaire

On va tester deux modèles :

$$\begin{array}{rcccl} M_1 & : & Y_i & = & \beta + \varepsilon_i \qquad (\varepsilon_i) \text{ iid de loi } \mathcal{N}(0,\sigma^2) \\ M_2 & : & Y_i & = & \alpha x_i + \beta + \varepsilon_i \quad (\varepsilon_i) \text{ iid de loi } \mathcal{N}(0,\sigma^2) \end{array}$$

On cherche donc à tester l'hypothèse nulle

$$H_0 : \ll \text{modèle } M_1 \gg$$

contre l'alternative

$$H_1 : \ll \text{modèle } M_2 \gg$$

Théorème:

Sous les hypothèse du MLG :

$$Z = \frac{\sum_{i=1}^{n}(\bar{Y} - Ax_i - B)^2/1}{\sum_{i=1}^{(} Y_i - Ax_i - B)^2/(n-2)} \underset{H_0}{\sim} F(1, n-2)$$

Démonstration :

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\bar{Y} - Ax_i - B)^2 + \sum_{i=1}^{n}(Y_i - Ax_i - B)^2$$

Or :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - Ax_i - B)^2 \sim \chi^2_{n-2}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y}) \underset{H_0}{\sim} \chi^2_{n-1}$$

D'après le théorème de Cochran, on a :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\bar{Y} - Ax_i - B)^2 \underset{H_0}{\sim} \chi^2_1$$

Avec indépendance entre les deux variables aléatoires.

Proposition:

On prend une réalisation de la variable aléatoire Z qu'on note z. La zone de rejet au risque δ est de la forme $\{Z > f_{1,n-2,\delta}\}$.

Ce test vient de l'analyse de la variance. On voit si, en passant du premier modèle au second, l'apport à la variance est significatif ou non. (Il suffit de regarder l'expression du numérateur et du dénominateur pour s'en convaincre !)

Troisième partie

ANOVA 1

But : tester l'égalité de p moyennes ($p \geq 2$).

1 Données et modèle

1.1 Données

On cherche à étudier l'effet d'un facteur A, qu'on supposera à p niveaux, sur une variable quantitative Y. On suppose que le facteur A influe uniquement sur les moyennes sur les moyennes des distributions de chacun des p groupes et non sur leur variance.

Niveau du facteur A	A_1	A_2	...	A_p
	y_{11}	y_{21}	...	y_{p1}
	⋮	⋮	...	⋮
	⋮	y_{2n_2}	...	⋮
	⋮		...	y_{pn_p}
	y_{1n_1}			
Effectifs	n_1	n_2	...	n_p
Moyennes empiriques	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	...	$\bar{y}_{p\bullet}$

1.2 Modèle

On fait les hypothèses suivantes :
— Pour tout $i \in \{1, ..., p\}$ et pour tout $j \in \{1, ..., n_i\}$, la donnée y_{ij} est la réalisation d'une variable aléatoire Y_{ij} de loi $\mathcal{N}(\mu_i, \sigma^2)$
— Les variables (Y_{ij}) sont globalement indépendantes.
ce qu'on résume par :

$$Y_{ij} = \mu_i + \varepsilon_{ij}, (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Définition: Dimension

Dans le contexte de l'ANOVA, on appelle dimension l'espace dans lequel vit l'espérance des variable aléatoires (Y_{ij}). Cette dimension est égale à la différence entre :
— le nombre de paramètres d'espérance envisagés dans la modélisation
— et le nombre de contraintes d'identifiabilité nécessaires

Remarque : Ici, le modèle est de dimension p, car on a p paramètres (les μ_i) à estimer et aucune contrainte. On notera ce modèle (M_p).

2 Test de l'effet du facteur

2.1 Introduction des hyposthèses

On veut tester l'absence d'effet du facteur sur les moyennes. On va donc tester l'hypothèse nulle :

$$H_0 :\ll \mu_1 = ... = \mu_p \gg$$

contre l'alternative :

$$H_1 :\ll \exists (i, j) \text{ tel que } \mu_i \neq \mu_j \gg$$

Sous l'hypothèse H_0, on a :

$$Y_{ij} = \mu + \varepsilon_{ij} \text{ avec } (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Ce modèle est de dimension 1, on le notera donc (M_1).
Tester l'absence d'effet du facteur A sur Y, c'est tester :

$$H_0 :\ll \text{Modèle } (M_1) : Y_{ij} = \mu + \varepsilon_{ij} \gg$$

contre l'alternative :

$$H_1 :\ll \text{Modèle } (M_p) : \mu_i + \varepsilon_{ij} \gg$$

2.2 Estimation des paramètres

Proposition: *Dans le modèle complet* (M_p)

Dans ce modèle, on doit estimer les (μ_i) et σ^2 :
— On estime μ_i (pour tout $i = 1, ..., p$) par $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\bullet}$
— On prédit pour tout (i, j), Y_{ij} par $\hat{Y}_{ij} = \hat{\mu}_i$
— Les résidus (estimations des ε_{ij}) sont définis par les $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i\bullet}$
— La somme des carrés résiduelle vaut $SCR(M_p) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$
— Enfin, on estime σ^2 par $S^2 = \frac{SCR(M_p)}{n-p}$

Proposition: *Dans le modèle* (M_1)

Dans ce modèle, on doit estimer μ et σ^2 :
— On estime μ par $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{\bullet\bullet}$
— On prédit pour tout (i, j), Y_{ij} par $\hat{Y} = \hat{\mu}$
— Les résidus (estimations des ε_{ij}) sont définis par les $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{\bullet\bullet}$
— La somme des carrés résiduelle vaut $SCR(M_1) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$
— Enfin, on estime σ^2 par $S^2 = \frac{SCR(M_1)}{n-1}$

Proposition: *Statistique de test*

Dans le cadre du modèle complet d'ANOVA 1, on a :

$$SCR(M_1) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \underbrace{(\hat{Y}_{ij}(M_p) - \hat{Y}_{ij}(M_1))}_{\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}}{}^2 + SCR(M_p)$$

et $\frac{1}{\sigma^2} SCR(M_p) \sim \chi^2_{n-p}$. De plus, sous H_0, $\frac{1}{\sigma^2} SCR(M_1) \underset{H_0}{\sim} \chi^2_{n-1}$. Donc :

$$Z = \frac{(SCR(M_1) - SCR(M_p))/(p-1)}{SCR(M_p)/(n-p)}$$

suit une loi de Fisher $F(p-1, n-p)$ (sous l'hypothèse H_0).
La zone de rejet de H_0 au risque δ est de la forme $\{Z > f_{p-1,n-p,\delta}\}$.

On peut voir cette statistique de test comme le rapport de deux estimateurs de σ^2 : un qui est toujours bon, et l'autre seulement sous H_0.

3 Comparaison multiple

Définition: *Contraste*

Un contraste entre les paramètres $(\mu_i)_{i=1,...,p}$ est une combinaison linéaire des (μ_i) de la forme $\sum_{i=1}^{p} c_i \mu_i$ où les c_i sont des coefficients réels constants vérifiant la condition $\sum_{i=1}^{p} c_i = 0$.

Pour un contraste donné, nous allons tester l'hypothèse nulle

$$H_0 :\ll \psi = \sum_{i=1}^{p} c_i \mu_i = 0 \gg$$

contre l'alternative :

$$H_1 :\ll \psi \neq 0 \gg$$

Soit $\hat{\psi} = \sum_{i=1}^{p} c_i \hat{Y}_{i\bullet}$ l'estimateur sans biais du constraste $\sum_{i=1}^{p} c_i \mu_i$.

Théorème:

Dans le cadre du modèle complet d'ANOVA 1 :

$$Z = \frac{\sum_{i=1}^{p} c_i \bar{Y}_{i\bullet}}{\sqrt{\frac{SCR(M_p)}{n-p}\left(\sum_{i=1}^{p} \frac{c_i^2}{n_i}\right)}} \underset{H_0}{\sim} T_{n-p}$$

Proposition:

On construit un test sur cette statistique. La zone de rejet de H_0 au risque δ est alors de la forme :

$$\{|Z| > t_{n-p,\delta/2}\}$$

4 Estimation des paramètres

On va chercher à construire un intervalle de confiance pour chacun des μ_i :

Théorème:

Sous les hypothèses de normalité et d'indépendance des p échantillons, pour tout $i \in \{1, ..., p\}$, $\bar{Y}_{i\bullet}$ est un estimateur sans biais du paramètre μ_i et :

$$\bar{Y}_{i\bullet} \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{n_i}\right)$$

De plus, $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$ est un estimateur sans biais de σ_i^2 indépendant de $\bar{Y}_{\bullet\bullet}$ et on a :

$$\frac{(n_i - 1)}{\sigma_i^2} S_i^2 \sim \chi^2_{n_i - 1}$$

Corollaire:

Il est possible de bâtir des intervalles de confiance pour les paramètres μ_i, en prenant la statistique :

$$\frac{\sqrt{n_i}(\bar{Y}_{i\bullet} - \mu_i)}{S_i} \sim T_{n_i - 1}$$

On construit ainsi l'intervalle de confiance au niveau de confiance $(1 - \delta)$ de μ_i :

$$IC_{(1-\delta)}(\mu_i) = \left[\bar{y}_{i\bullet} \pm \frac{s_i t_{n_i - 1, \delta/2}}{\sqrt{n_i}}\right]$$