Chapitre 9

Analyse de variance à un facteur

9.1 Introduction

L'analyse de variance à un facteur est une généralisation du test basé sur la loi de *Student* pour comparer deux moyennes : ici on compare k moyennes, $k \geq 2$. Nous illustrons le problème à l'aide des données suivantes [Source anonyme. Rapporté dans McNeil, D. R. (1977) *Interactive Data Analysis*. New York: Wiley.]

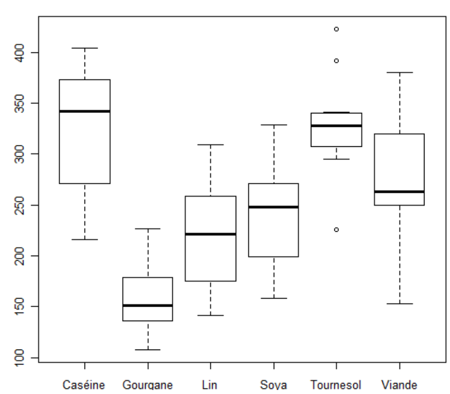
Une expérience est menée sur 71 poussins afin de comparer les effets sur la croissance de six régimes alimentaires. Les poussins ont été répartis au hasard, immédiatement après leur éclosion, en six groupes et leur poids après six semaines, a été noté. Le tableau 9.1 présente les résultats.

Tableau 9.1

Poids en grammes à 6 *semaines de* 71 *poussins*
répartis selon le type de grain consommé

	Type de grain					
	Caséine	Gourgane	Lin	Viande	Soya	Tournesol
	368 ; 390 379 ; 260 404 ; 318 352 ; 359 216 ; 222 283 ; 332	179 ;160 136 ; 227 217 ; 168 108 ; 124 143 ; 140	309 ; 229 181 ; 141 260 ; 203 148 ; 169 213 ; 257 244 ; 271	325 ; 257 303 ; 315 380 ; 153 263 ; 242 206 ; 344 258 ;	243 ; 230 248 ; 327 329 ; 250 193 ; 271 316 ; 267 199 ; 171 158 ; 248	423 ; 340 392 ; 339 341 ; 226 320 ; 295 334 ; 322 297 ; 318
n_i	12	10	12	11	14	12
$\bar{y}_i$	323,58	160,20	218,75	276,91	246,43	328,92
S_i	64,43	38,63	52,24	64,90	54,13	48,84

La question est de savoir si la croissance d'un poussin dépend du type de grain consommé. Une analyse graphique suggère une différence entre les grains :



À première vue, la caséine et le tournesol semblent particulièrement efficaces et la gourgane beaucoup moins. Ces différences sont-elles significatives ? L'hypothèse nulle est

> H_o : Il n'y a pas de différence entre les différents grains (quand à leur effet sur la croissance des poussins)

Nous devons formuler le modèle, pour pouvoir ensuite exprimer H_o en termes des paramètres du modèle.

9.2 Décomposition d'une somme de carrés

Les données prennent la forme suivante :

T_1	T_2	$\ldots$	T_k	
y_{11}	y_{21}	$\ldots$	y_{k1}	
y_{12}	y_{22}	$\ldots$	y_{k2}	
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
y_{1n_1}	y_{2n_2}	$\ldots$	y_{kn_k}	
$\bar{y}_1$	$\bar{y}_2$	$\ldots$	$\bar{y}_k$	$\bar{y}$

Dans le langage habituel, on dit que l'expérience comporte k groupes soumis à k *traitements* (k = 6 ici), chaque groupe recevant un traitement distinct. Une mesure Y est prise sur chaque individu. Soit

n_i : le nombre d'individus dans le groupe i, i =1, ..., k ;

$n = \sum_{i=1}^{k} n_i$ le nombre total d'individus ;

y_{ij} : la j^e observation (poids du j^e poussin) du *groupe i* ;

$\bar{y}_i$: la moyenne des mesures du groupe i.

Les échantillons présentent des différences entre les moyennes $\bar{y}_1, \ldots, \bar{y}_k$. On dira de ces différences qu'elles sont *significatives* si on conclut qu'elles reflètent des différences d'efficacité réelles entre les différents grains — qu'elles ne sont pas le fruit du hasard. Bien qu'à première vue, certaines des différences semblent assez importantes, on ne peut pas vraiment en juger sans avoir une idée de la variabilité naturelle des poids. Il faudra mettre en rapport la dispersion des moyennes $\bar{y}_i$ avec une mesure de la variabilité normale des poids. Des indices de cette variabilité naturelle sont donnés par les écarts-types S_i des poids à l'intérieur des groupes, des indices que nous devrons combiner en une seule mesure.

Les moyennes $\bar{y}_i$ sont

$$\bar{y}_i = (1/n_i)\sum_{j=1}^{n_i} y_{ij} \, , \qquad (9.2.1)$$

et la moyenne globale (des k groupes) est

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_i \qquad (9.2.2)$$

Ce sont les écarts entre les $\bar{y}_i$ et $\bar{y}$ qui nous diront si H_o devrait être rejetée ou non. Une mesure de la dispersion entre les $\bar{y}_i$ est donnée par la *somme des carrés expliquée*, désignée par SCE :

$$\textit{Somme des carrés expliquée} = \text{SCE} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y})^2 \, . \qquad (9.2.3)$$

On notera que les écarts au carré $(\bar{y}_i - \bar{y})^2$ sont *pondérés* par les tailles n_i : l'écart d'un groupe i par rapport à la moyenne globale a d'autant plus d'importance que le groupe est grand.

C'est une dispersion « expliquée » dans la mesure où elle peut être attribuée—en partie—à des différences dans l'effet des traitements.

Alors que la dispersion des données à l'intérieur d'un groupe ne peut être expliquée de cette manière puisque toutes les données d'un groupe ont été soumises à un même traitement. Cette dispersion interne est mesurée par la *somme des carrés résiduelle* SCR, définie par

$$\textit{Somme des carrés résiduelle} = \text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \qquad (9.2.4)$$

Les sommes de carrés SCE et SCR constituent une décomposition de la *somme des carrés totale* SCT :

$$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 \qquad (9.2.5)$$

C'est la dispersion de l'ensemble des données par rapport à la moyenne globale.

On peut montrer que

$$\text{SCT} = \text{SCE} + \text{SCR}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \qquad (9.2.6)$$

Dans l'exemple, SCE est une mesure des différences *entre* les poids moyens des différents groupes, donc attribuable en partie aux effets différentiels des grains et SCR est la partie ***résiduelle***, la dispersion entre les poids des poussins *d'un même groupe*, donc attribuable simplement à la variabilité naturelle des poids.

Le jugement concernant l'effet des traitements dépendra de l'importance relative de SCE par rapport à SCR.

9.3 Modèle et hypothèses

Nous supposons que les poids y_{ij} des poussins du groupe i sont des variables aléatoires de loi normale, de moyenne μ_i et de variance σ^2. On peut donc écrire

$$y_{ij} = \mu_i + \varepsilon_{ij},\ j = 1, 2, \ldots, n_i \qquad (9.3.1)$$

où les ε_{ij} (les écarts des y_{ij} par rapport à leur espérance) sont indépendantes et

$$\varepsilon_{ij} \sim \mathcal{N}(0\,;\sigma^2) \qquad (9.3.2)$$

Les paramètres inconnus du modèle sont $\mu_1, \ldots, \mu_k$ et σ^2.

Remarque *Le paramètre σ^2 représente la dispersion entre les poids de poussins soumis à un même régime. Notre modèle suppose que la variance σ^2 des ε_{ij} est la même pour tous les groupes. Cette hypothèse n'est pas toujours réaliste, mais un modèle sans cette hypothèse est mathématiquement difficile à traiter.* ■

L'hypothèse nulle peut maintenant être exprimée en termes des paramètres du modèle :

$$\text{H}_o : \quad \mu_1 = \mu_2 = \ldots = \mu_k \qquad (9.3.3)$$

L'alternative est la négation de H_o :

$$\text{H}_1 : \textit{au moins une des égalités } \mu_i = \mu_j\ (i \neq j) \textit{ n'est pas vérifiée} \qquad (9.3.4)$$

Lorsque les μ_i sont égaux, les moyennes $\bar{y}_i$ devraient être assez rapprochées les unes des autres et SCE devrait être petite. Donc nous devrions rejeter H_o si SCE est grand par rapport à SCR.

Le test pourrait donc être basé sur le rapport SCE/SCR. Mais une certaine normalisation est nécessaire si l'on veut rendre ces sommes sont plus comparables, car elles varient avec le nombre de termes qu'elles comprennent. On divisera donc chaque somme par le « nombre de degrés de liberté » pour obtenir une sorte de moyenne. Les *moyennes de carrés* sont définies comme ceci :

$$\text{MCE} = \frac{\text{SCE}}{k-1} \qquad \text{MCR} = \frac{\text{SCR}}{n-k} \tag{9.3.5}$$

La statistique de test est le rapport

$$F = \frac{\text{MCE}}{\text{MCR}} \tag{9.3.6}$$

Sous H_o la statistique F suit une loi de Fisher à k-1 et n-k degrés de liberté :

$$\text{Sous } H_o,\ F \sim \mathcal{F}_{k-1;\ n-k} \tag{9.3.7}$$

où $\mathcal{F}_{\nu_1;\nu_2}$ désigne une variable de loi de Fisher à ν_1 et ν_2 degrés de liberté. On rejette H_o si

$$F > \mathcal{F}_{k-1;\ n-k;\alpha}. \tag{9.3.8}$$

$\mathcal{F}_{k-1;\ n-k;\alpha}$. désigne le point critiques, soit le point tel que $P(\mathcal{F}_{k-1;n-k} > \mathcal{F}_{k-1;\ n-k;\alpha}) = \alpha$. La coutume est de présenter le tout sous la forme d'une *table d'analyse de variance*.

9.4 Table d'analyse de variance

On présente, dans la table d'analyse de variance, les sommes de carrés, le nombre de degrés de de liberté, et le rapport F. (Les logiciels statistiques ajoutent à cela le seuil expérimental.) Nous présentons en outre l'espérance de chaque somme de carrés.

Source	Somme de carrés	Degrés de liberté	Moyenne des carrés	Espérances des moyennes des carrés
Expliquée	SCE	k-1	$\text{MCE} = \frac{\text{SCE}}{k-1}$	$\sigma^2 + \frac{\sum_i n_i(\mu_i - \bar{\mu})^2}{k-1}$
Résiduelle	SCR	n-k	$\text{MCR} = \frac{\text{SCR}}{n-k}$	σ^2
Total	SCT	n-1	$\text{MCT} = \frac{\text{SCT}}{n-1}$	$\sigma^2 + \frac{\sum_i n_i(\mu_i - \bar{\mu})^2}{n-1}$

Remarque MCT *est simplement la variance S^2 de toutes les données de l'échantillon, sans distinction de groupe.* ■

Les espérances ci-dessus justifient le choix du rapport F = MCE/MCR, puisque MCE et MCR ont la même espérance si et seulement si H_o est vraie; si H_o est fausse, MCE a tendance à être plus grand que MCR et le rapport F sera en conséquence grand. La distribution de F est connue sous H_o.

Exemple 9.4.1 *Décomposition de la somme des carrés totale*

Dans l'exemple présenté au début du chapitre, nous avons k = 6, et $\bar{y} = \frac{\sum_{i=1}^{6} n_i \bar{y}_i}{n} = \frac{18553}{71} =$ 261,3099.

SCE = 231 129; SCR= $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}(n_i-1)S_i^2$ = 195 556.

MCE = SCE/5 = 46 226; MCR = SCR/(71-6) = 3009.

F = MCE/MCR = 15,36.

Le point critique à 5 % est $\mathcal{F}_{k\text{-}1;\ n\text{-}k;\alpha} = \mathcal{F}_{5;\ 65;0,05}$ = 2,35. Donc on rejette H_o pour conclure que les moyennes ne sont pas toutes égales. La valeur p, calculée à l'aide d'un logiciel, est $p = 5{,}93642\times10^{-10}$.

Voici ce que produit le logiciel statistique R :

```
Analysis of Variance Table
Response: Poids
          Df Sum Sq Mean Sq F value    Pr(>F)
Grain      5 231129   46226  15.365 5.936e-10 ***
Residuals 65 195556    3009
```

La valeur p est habituellement présentée dans la table d'analyse de variance. ■

9.5 Estimation des paramètres

Estimation des moyennes

Les μ_i sont des paramètres qu'on peut également estimer. Les estimateurs les plus naturels sont les moyennes des groupes

$$\hat{\mu}_i = \bar{y}_i \tag{9.5.1}$$

En plus de leur simple attrait intuitif, ces estimateurs se justifient par le principe des moindres carrés, un principe qui choisit pour estimateurs les valeurs qui minimisent les écarts entre les observations et leur espérance. Spécifiquement, es $\hat{\mu}_i$ sont les valeurs de μ_i qui minimisent la somme des carrés

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij}-\mathrm{E}(y_{ij})]^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 \tag{9.5.2}$$

On peut montrer que Q est minimisée lorsque $\mu_i = \bar{y}_i$. La somme des carrés résiduelle est la valeur de Q minimisée.

Estimation de σ^2

MCT n'est pas ici un estimateur sans biais de la variance σ^2 car les observations de l'échantillon n'ont pas toutes la même espérance. Elle le serait si H_o était vraie, si les μ_i étaient tous égaux. Par contre, chaque S_i^2 est un estimateur sans biais de σ^2. Pour estimer σ^2, on prend une moyenne des S_i^2, pondérée par le nombre de degrés de liberté :

$$\hat{\sigma}^2 = \sum_{i=1}^{k}\frac{(n_i-1)S_i^2}{n-k}. \tag{9.5.3}$$

On constate de cet estimateur est exactement MCR :

$$\text{MCR} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{n-k} = \sum_{i=1}^{k}\frac{(n_i-1)S_i^2}{n-k}. \tag{9.5.4}$$

C'est un estimateur sans biais de σ^2.

Remarque *La statistique F illustre un développement, caractéristique en statistique, qu'il est instructif de suivre ici. Chaque* $\bar{y}_i$ *est de variance* σ^2/n_i *et donc* $\frac{\bar{y}_i-\mu}{\sigma/\sqrt{n_i}} \sim N(0\ ;\ 1)$, *où* μ *est la moyenne, commune sous* H_o. *Donc* $\frac{(\bar{y}_i-\mu)^2}{\sigma^2/n_i} = \frac{n_i(\bar{y}_i-\mu)^2}{\sigma^2} \sim \chi_1^2$ et $\sum_{i=1}^{k}\frac{n_i(\bar{y}_i-\mu)^2}{\sigma^2} \sim \chi_k^2$. *Lorsqu'on remplace* μ *par* $\bar{y}$, *on perd un degré de liberté et alors* $\sum_{i=1}^{k}\frac{n_i(\bar{y}_i-\bar{y})^2}{\sigma^2} \sim \chi_{k-1}^2$. *Si* σ^2 *était connue, la statistique naturelle aurait été* $\frac{\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2/\sigma^2}{(k-1)} = \frac{\text{MCE}}{\sigma^2}$. *Mais puisque* σ^2 *n'est pas connue, on remplace* $\frac{\text{MCE}}{\sigma^2}$ *par* $\frac{\text{MCE}}{\hat{\sigma}^2}$. *Cette quantité est précisément la statistique F.* ■

9.6 Une interprétation de la statistique F

Nous pouvons considérer toute somme de carrés expliquée comme une différence de sommes de carrés résiduelles. La somme des carrés résiduelle est, par définition,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \qquad (9.6.1)$$

Mais la somme des carrés totale est *elle aussi* une somme de carrés résiduelle: c'est la somme des carrés résiduelle dans le modèle $y_{ij} = \mu + \varepsilon_{ij}$, le modèle selon lequel les moyennes des groupes sont toutes égales. Or ce modèle est précisément le modèle stipulé par l'hypothèse nulle. On peut donc écrire

$$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu})^2 = \text{SCR}_o. \qquad (9.6.2)$$

où la notation SCR_o signifie qu'il s'agit d'une somme de carrés résiduelle sous un ***modèle réduit***, réduit par les contraintes imposées par H_o. La somme des carrés expliquée, qui est la différence SCT-SCR ou

$$\text{SCE} = \text{SCR}_O - \text{SCR} \qquad (9.6.3)$$

Si ν est le nombre de degrés de liberté de SCR et ν_o est le nombre de degrés de liberté de SCR_o, alors le rapport *F* devient

$$F = \frac{[\text{SCR}_O\text{-SCR}]/(\nu_O-\nu)}{\text{SCR}/\nu} \qquad (9.6.4)$$

Cette formule est assez générale : le rapport *F* est toujours de cette forme. La différence SCR_o - SCR, la somme des carrés expliquée, représente la *réduction* d'erreur due à l'introduction du modèle plus complexe. Le nombre de degrés de liberté du numérateur, $\nu_o - \nu$, est égal à la réduction du nombre de paramètres imposée par H_o.

9.7 Combinaisons linéaires des moyennes

Ayant rejeté l'hypothèse que toutes les moyennes sont égales, on voudra normalement procéder à certaines comparaisons deux à deux. Dans l'exemple considéré dans ce chapitre, supposons qu'on veuille tester l'hypothèse

$$H_o : \mu_1 = \mu_2 \qquad (9.7.1)$$

La statistique habituelle est $T = \dfrac{\bar{y}_1 - \bar{y}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$. Habituellement, dans un test d'égalité de deux moyennes, $\hat{\sigma}$ est une estimation commune basée sur les deux échantillons. Mais ici, l'estimation de σ exploite l'information contenue dans tous les échantillons, c'est-à-dire, $\hat{\sigma}$ est la racine carrée de MCR. Sous H$_o$,

$$T = \frac{\bar{y}_1 - \bar{y}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t_{n-k}. \qquad (9.7.2)$$

Exemple 9.7.1 *Égalité de deux des moyennes*

Testons l'hypothèse H$_o$: $\mu_1 = \mu_2$. $\bar{y}_1$ = 323,5833; $\bar{y}_2$ = 160,2, n_1 = 12, n_2 = 10; $\hat{\sigma}$ = 54,85029.

$T = \dfrac{323{,}5833-160{,}2000}{54{,}85029\sqrt{\dfrac{1}{12}+\dfrac{1}{10}}}$ = 6,96, ce qui est hautement significatif.

On conclut avec confiance que la caséine est plus efficace que la gourgane. ■

Plus généralement, on peut estimer toute combinaison linéaire des μ_i et tester des hypothèses à propos de ces combinaisons linéaires. Soit $\varphi = \sum_i c_i\mu_i$ une combinaison linéaire avec coefficients fixes c_i. Un estimateur sans biais $\hat{\varphi}$ de φ est $\sum_i c_i\hat{\mu}_i = \sum_i c_i\bar{y}_i$. Sous les hypothèses du modèle, $\hat{\varphi}$ est normale, de moyenne φ et de variance $\sum \dfrac{c_i^2\sigma^2}{n_i} = \sigma^2 \sum \dfrac{c_i^2}{n_i}$. La variable

$$T = \frac{\hat{\varphi}-\varphi}{\hat{\sigma}\sqrt{\sum c_i^2 / n_i}} \sim t_{n-k} \qquad (9.7.3)$$

Ceci nous permet de déterminer un intervalle de confiance pour φ et de tester des hypothèses du genre H_o: $\varphi = \varphi_o$. Ce qui inclut des hypothèses sur de la différence entre 2 des moyennes, ou encore sur certaines moyennes particulières.

Exemple 9.7.2 *Estimation d'une combinaison linéaire de moyennes*

Estimer le poids moyen φ des poussins à six semaines si 20 % sont nourries à la caséine ; 10 % à la gourgane ; 10 % au lin; 20 % à la farine de viande ; 30 % au soya ; et 10 % au tournesol. Déterminer un intervalle de confiance à 95 % pour φ.

Solution $\varphi = (0{,}2)\mu_1 + (0{,}1)\mu_2 + (0{,}1)\mu_3 + (0{,}2)\mu_4 + (0{,}3)\mu_5 + (0{,}1)\mu_6$,

$\hat{\varphi} = (0{,}2)\bar{y}_1 + (0{,}1)\bar{y}_2 + (0{,}1)\bar{y}_3 + (0{,}2)\bar{y}_4 + (0{,}2)\bar{y}_5 + (0{,}1)\bar{y}_6$ = 264,8137;

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\dfrac{(0,2)^2}{12}+\dfrac{(0,1)^2}{10}+\dfrac{(0,1)^2}{12}+\dfrac{(0,2)^2}{11}+\dfrac{(0,3)^2}{14}+\dfrac{(0,1)^2}{12}}$ = 6,952138.

Le point critique pour un intervalle de confiance à 95 % est 1.996. L'intervalle de confiance est donc

$$\hat{\varphi} - 1{,}996\,\hat{\sigma}_{\hat{\varphi}} \le \varphi \le \hat{\varphi} + 1{,}996\,\hat{\sigma}_{\hat{\varphi}},$$

$$251 \le \varphi \le 278.$$ ■

Exercices

9.1 Montrez que le test t pour comparer deux moyennes est équivalent à une analyse de variance. [Utilisez le fait que si T est de loi t_ν, alors T^2 est de loi $F_{1;\nu}$.]

9.2 Supposons que M. Martin peut se rendre chez lui le soir par trois routes différentes. Il essaye chacune d'elles 5 fois en prenant note du temps à chaque fois. Voici les résultats, en minutes :

Route 1: 22, 26, 25, 25, 3l
Route 2: 25, 27, 28, 26, 29
Route 3: 26, 29, 33, 30, 33.

Testez à 5% l'hypothèse que les trois routes sont comparables.

9.3 [*Exemple tiré de Snedecor et Cochran*] Une expérience est faite pour comparer 4 traitements sur la culture de la betterave à sucre. Chaque traitement a été appliqué à 5 champs, et la récolte moyenne par arpent a été notée. Voici les résultats en centaines de livres.

		Engrais appliqué…		
	Pas d'engrais	en janvier par labourage	en janvier à la volée	en avril à la volée
Moyenne	38,7	48,7	48,8	45,0

Les calculs ont donné : $\hat{\sigma}^2 = 7{,}443$. Testez chacune des hypothèses suivantes :

a) L'engrais n'a aucun effet.

b) En moyenne, l'engrais appliqué en janvier n'a ni plus ni moins d'effet que lorsqu'il est appliqué en avril.

c) L'engrais appliqué à la volée en janvier a le même effet que lorsqu'il est labouré.

Dans les cas b) et c), déterminez un intervalle de confiance pour l'effet étudié.

9.4 Dans une grande classe de statistique, les élèves proviennent de 4 groupes distincts, définis comme suit:

Groupe 1: Les élèves n'ayant suivi aucun cours de mathématiques au Cégep
Groupe 2: Les élèves ayant suivi des cours de mathématiques et de statistique
Groupe 3: Les élèves ayant suivi des cours de mathématiques mais pas de statistique
Groupe 4: Les élèves ayant terminé un programme de science au Cégep.

Les résultats au cours de statistique sont présentés au tableau 9.2 :

a) Dressez une table d'analyse de variance

b) Testez l'hypothèse que la moyenne du premier groupe est différente de celle des trois autres.

9.5 Lors d'une étude sur le « concept de soi » des adolescents nigérians, un chercheur [Jegede, R. Olukayode, *The Journal of Psychology* 110, 249-261 (1982)] a administré le test *Piers-Harris Self-Concept Scale* à 1300 élèves de niveau secondaire à Ibadan.

a) Le score moyen a été de 58,19 avec un écart-type de 10,06. Dans une étude faite antérieurement auprès de 1183 adolescents américains on avait trouvé une moyenne de 51,84 avec un écart-type de 13,87. La différence entre le Nigérians et les Américains est-elle significative ? Supposez que la variance σ^2 des deux groupes est la même et estimez σ^2.

Tableau 9.2

Notes en statistique de quatre groupes d'étudiants

Groupe 1		Groupe 2		Groupe 3		Groupe 4	
66	69	63	58	36	48	45	83
35	87	51	47	72	78	91	88
61	82	29	38	32	45	56	67
74	80	53	50	91	47	82	67
47	84	63	68	75	50	131	84
72	54	45	66	63	53	88	83
57	66	33	37	85	21	68	54
48	35	84	62	54		51	81
55	40	83	68			84	57
		59	74			60	56
		83	60			83	44
						75	

b) L'échantillon de Jegede était composé de 552 filles et 828 garçons. Les filles avaient une moyenne de 56,82 avec un écart-type de 9,96; et les garçons une moyenne de 59,11 avec un écart-type de 10,01. La différence entre les garçons et les filles est-elle significative ? Supposez que la variance σ^2 des deux groupes est la même et estimez σ^2.

c) Maintenant traitez simultanément les trois groupes : filles nigérianes, garçons nigérians, et Américains. Testez l'hypothèse que les moyennes de trois groupes (μ_1, μ_2 et μ_3, respectivement) sont égales. Supposez que la variance σ^2 des trois groupes est la même et estimez σ^2.

d) Dans le cadre du modèle en c) testez l'hypothèse que les filles et les garçons nigérians ont la même moyenne.

e) Dans le cadre du modèle en c) testez l'hypothèse que la moyenne des Américains est égale à la moyenne des Nigérians en supposant qu'il y a autant de filles que de garçons au Nigéria (en d'autres termes, vous devez tester l'hypothèse que $\mu_3 = (\mu_1+\mu_2)/2$.)

9.6 Dans une étude sur la relation entre certains traits de personnalité et des facteurs astrologiques, des chercheurs [Sakofske, Kelly et McKerracher, *The Journal of Psychology* 110, 275-80, 1982] ont fait compléter un questionnaire (le *Eysenck Personality Questionnaire*] à 241 étudiants néo-zélandais. L'hypothèse (avancée antérieurement par des astrologues) que ces chercheurs se proposer de vérifier est que les personnes nées sous un signe positif (Bélier, Balance, Gémeaux, Lion, Verseau, Sagittaire) sont moins introverties que les personnes nées sous un signe négatif (Cancer, Capricorne, Poisson, Scorpion, Taureau, Vierge). Sur l'échelle introversion-extraversion du test, les extravertis ont un score élevé. L'échantillon était composé d'hommes et de femmes. Voici les moyennes, les écarts-types et le tailles des 4 groupes ainsi que la désignation des moyennes des populations :

		Hommes				Femmes			
		$\bar{y}$	S	n	μ	$\bar{y}$	S	n	μ
Signe du zodiac	Positif	13,50	4,38	38	μ_1	13,17	4,57	79	μ_3
	Négatif	15,52	4,21	38	μ_2	13,73	4,39	86	μ_4

Dans ce qui suit, considérer qu'il s'agit de quatre groupes issus de quatre populations de même variance σ^2.

a) Tester l'hypothèse que les hommes et les femmes ont la même moyenne, c'est-à-dire, $\mu_1+\mu_2 = \mu_3+\mu_3$.

b) Tester l'hypothèse que les personnes nées sous un signe positif ont la même moyenne que ceux nés sous un signe négatif, c'est-à-dire, $\mu_1+\mu_3 = \mu_2+\mu_4$.

c) Tester l'hypothèse que la différence entre les positifs et les négatifs et la même chez les hommes et les femmes.

d) Tester l'hypothèse qu'il n'y a pas de différence entre les positifs et les négatifs chez les hommes.

9.7 Dans une étude sur la sexualité des jeunes en Australie, un chercheur [Hong, Sung-Mook, The *Journal of Psychology* 115, 17-22 (1983)] a fait remplir un questionnaire à 560 étudiants d'université. Le questionnaire rempli permet de calculer un score qui indique dans quelle mesure l'attitude du répondant est permissive (un score élevé dénote une attitude permissive). L'objectif est de déterminer si le niveau de pratique religieuse affecte l'attitude concernant les comportements sexuels. Voici les moyennes, les écarts-types et les effectifs de trois sous groupes.

Vont à l'église	$\bar{y}$	S	n	μ
Régulièrement	3,31	1,54	128	μ_1
De temps en temps	4,73	1,10	230	μ_2
Jamais	5,24	0,79	202	μ_3

a) Dresser une table d'analyse de variance. Expliquez votre conclusion. Estimer la variance σ^2.

b) Tester, dans le cadre du modèle en a) l'hypothèse que ceux qui ne vont jamais à l'église sont différents de ceux qui y vont, régulièrement ou de temps en temps, c'est-à-dire, tester l'hypothèse $\mu_3 = (\mu_1+\mu_2)/2$.

c) Tester l'hypothèse qu'il n'y a pas de différence entre ceux qui vont régulièrement à l'église et ceux qui y vont de temps en temps.

d) Tester l'hypothèse qu'il n'y a pas de différence entre ceux qui vont à l'église de temps en temps et ceux qui n'y vont jamais.

e) Amalgamez les deux derniers groupes et testez l'hypothèse que ceux qui vont régulièrement à l'Église ont la même moyenne que les autres. Supposez une même variance σ^2 et estimez σ^2.

9.8 Lors d'un projet d'étude des problèmes de racisme dans une force policière, on prélève un échantillon de 32 policiers, dont 16 ont 11 ans de scolarité ou moins et 11 on plus de 11 ans de scolarité. Chacun des deux groupes est divisé en deux sous-groupes de 8. L'un des deux sous-groupes suit un cours de sensibilisation aux cultures ethniques, l'autre pas. Le tableau suivant donne les résultats à un test d'hostilité aux groupes ethniques.

		Scolarité	
		≤ 11 ans	> 11 ans
Cours de sensibilisation	Suivi	60 58 56 54 52 50 48 46	36 36 33 32 30 29 26 26
	Pas suivi	50 48 46 44 42 40 38 36	30 28 26 24 22 20 28 26

a) Dresser une table d'analyse de variance. Estimer la variance σ^2

b) Tester l'hypothèse que la moyenne de ceux qui ont suivi le cours est la même que celle de ceux qui ne l'on pas suivi.

c) Testez l'hypothèse que le cours a le même effet chez ceux de 11 ans de scolarité et ceux de plus de 11 ans de scolarité (l'« effet » du cours est mesuré par la différence des deux moyennes).

d) Supposez qu'on n'a pas tenus compte des niveaux de scolarité. Vous avez donc deux groupes. Tester l'hypothèse que la moyenne de ceux qui ont suivi le cours est la même que celle de ceux qui ne l'on pas suivi. Estimez σ^2.

9.9 Deux groupes de 11 enfants de troisième année du cycle primaire ont complété le test psychologique IAR (*Intelligence Achievement Responsibility*) avant et après une période de quatre mois et demi d'expérimentation avec l'un ou l'autre de deux langages informatiques : LOGO et Delta Drawing. Contrairement au LOGO, le langage Delta Drawing n'attache pas une grande importance à la décomposition d'un problème complexe ou à l'apprentissage par la correction des erreurs. Le test IAR mesure la propension du sujet à se sentir maître de ses apprentissages et de son succès intellectuel. Les chercheurs ont voulu montrer que l'exercice du langage LOGO augmente cette propension. Voici les résultats obtenus :

Tableau 9.3

Comparaison des langages Logo et Delta Drawing

LOGO			Delta		
Sexe	Score		Sexe	Score	
	Avant	Après		Avant	Après
F	16	29	F	15	21
F	20	24	M	18	22
M	21	23	F	21	21
M	22	21	F	21	19
M	22	26	F	22	20
F	23	30	F	22	20
F	24	26	F	23	23
F	24	23	F	23	30
F	25	32	M	26	21
M	27	34	M	27	25
M	28	29	M	30	27

a) Soit Y la différence Après-Avant. Dresser une table d'analyse de variance considérant que l'échantillon est composé de quatre groupes : Filles*LOGO (μ_1) ; Filles*Delta (μ_2) ; Garçons*LOGO (μ_3) ; et Garçons*Delta (μ_3).

b) Tester l'hypothèse qu'il n'y a pas de différence entre LOGO et Delta.

c) Tester l'hypothèse qu'il n'y a pas de différence entre filles et garçons.

d) Tester l'hypothèse que la différence entre les deux sexes est la même pour LOGO que pour Delta.

e) Tester l'hypothèse que la différence entre LOGO et Delta est la même pour les filles que pour les garçons.

f) Supposons que la distinction entre filles et garçons n'a simplement pas été retenue pour l'expérience. Dans ce modèle à deux groupes, tester l'hypothèse qu'il n'y a pas de différence entre LOGO et Delta. Estimer la variance σ^2.

9.10 Voici les prix d'un échantillon de maisons vendus dans la région de Montréal, classées en six catégories, selon le nombre de chambres à coucher :

Chambres à coucher	Secteur		
	Centre	Nord	Sud
1 ou 2	98 249 289 299	97 157 170 184 385 145	249 184 239 314 69 69 80 85 85 269 60 89 89 89 89 142 142
3	369 499 700 495 499 499	142 175 349 435 89 201 226 239 479	259 269 269 299 219 349 119 195 199 329 339 365 153 158 200
4 ou plus	169 319 540	140 200 339 339 359 429 699 289 469 180 246 399 439 450 539 86 87 90	312 319 549 775 1385 142

Moyennes

Chambres à coucher	Secteur		
	Centre	Nord	Sud
1 ou 2	233,7500	189,6667	137,8235
3	510,1667	259,4444	248,0667
4 ou plus	342,6667	321,0556	580,3333

Sommes des carrés : les sommes $\sum(y_i - \bar{y})^2$ dans chaque case

Chambres à coucher	Secteur		
	Centre	Nord	Sud
1 ou 2	25970,75	50223,33	106998,47
3	56568,83	141972,22	84456,93
4 ou plus	69660,67	499990,94	1018799,33

a) Sans tenir compte du nombre de chambres à coucher, tester l'hypothèse que les prix moyens dans les trois secteurs sont les mêmes.

b) Pour chacune des trois classes de chambres à coucher, déterminer s'il y a une différence significative entre les trois secteurs (trois tests distincts).

c) Tester l'hypothèse qu'il n'y a pas de différences entre les six classes de maisons.

d) Dans le cadre de l'analyse faite en c), tester l'hypothèse que le prix moyen des maisons du secteur Centre est égal au prix moyen des maisons du secteur Sud

9.11 Démontrez la décomposition SCT = SCE + SCR, soit

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

9.12 Démontrez la formule de calcul

$$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y})^2 = \sum_{i=1}^{k} n_i \bar{y}_i^2 - n\bar{y}^2$$

9.13 Considérons le cas où $n_1 = n_2 = ... = n_k = m$. Montrez que SCE/$\sigma^2$: χ^2_{k-1} si H_o est vraie. Sans supposer que H_0 est vraie, déterminez l'espérance de SCE. Montrez comment votre réponse justifie une région critique de la forme $F > F_{k-1;k(m-1);\alpha}$.