STT1000

CHAPITRE 9 ANALYSE DE VARIANCE

SOLUTIONS

2017-12-30

9.1 Montrez que le test de *Student* pour comparer deux moyennes est équivalent à une analyse de variance.

Test *de Student*: on rejette H_o ($\mu_1 - \mu_2 = 0$) si $T = \dfrac{|\bar{y}_1 - \bar{y}_2|}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} > t_{n_1+n_2-2;\alpha/2}$

$\Leftrightarrow T^2 = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{\hat{\sigma}^2(1/n_1 + 1/n_2)} > (t_{n_1+n_2-2;\alpha/2})^2$.

Test F : on rejette H_o si $\dfrac{MCE}{MCR} > F_{n_1+n_2-2;\alpha}$. On vérifie aisément que MCR = $\hat{\sigma}^2$, alors que

$$MCE = \frac{n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2}{2-1} = n_1[\bar{y}_1 - (n_1\bar{y}_1 + n_2\bar{y}_2)/n]^2 + n_2[\bar{y}_2 - (n_1\bar{y}_1 + n_2\bar{y}_2)/n]^2$$

$$= n_1[(n_2\bar{y}_1 - n_2\bar{y}_2)/n]^2 + n_2[(n_1\bar{y}_2 - n_1\bar{y}_1)/n]^2 = \frac{n_1n_2^2}{n^2}(\bar{y}_1 - \bar{y}_2)^2 + \frac{n_2n_1^2}{n^2}(\bar{y}_1 - \bar{y}_2)^2 = \frac{n_1n_2}{n}(\bar{y}_1 - \bar{y}_2)^2$$

$= \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{1/n_1 + 1/n_2}$. Donc $\dfrac{MCE}{MCR} = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{\hat{\sigma}^2(1/n_1 + 1/n_2)}$.

Il faut maintenant montrer que $(t_{n_1+n_2-2;\alpha/2})^2 = F_{n_1+n_2-2;\alpha}$. Ceci découle du fait que si $T \sim t_\nu$, alors $T^2 \sim F_{1;\nu}$. Pour tout $a > 0$, $P(|T| > a) = P(T^2 > a^2) = P(F_{1;\nu} > a^2)$, ce qui entraîne que $(t_{n_1+n_2-2;\alpha/2})^2 = F_{n_1+n_2-2;\alpha}$.

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.

MCE=25,867 ; MCR = 7,3 ; F = 3,54 ; point critique : $F_{2;12;.05}$ = 3,89. On ne peut pas rejeter H_o à 5 %. La valeur p est 0,062.

9.3 [*Exemple tiré de Snedecor et Cochran*] Une expérience est menée afin de 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.
On teste l'hypothèse que les quatre moyennes sont égales.
SCE = 337,3 ; MCE = 112,433 ; MCR = 7,443 ; F = 15,105 ; $F_{3;16}$ = 3,23. On rejette H_o. La valeur p est 0,000063.

b) En moyenne, l'engrais appliqué en janvier n'a ni plus ni moins d'effet que lorsqu'il est appliqué en avril.
Si les quatre moyennes sont désignées, dans l'ordre du tableau, par μ_1, μ_2, μ_3, et μ_4, on interprétera cette hypothèse comme H_o : $\varphi = 0$, où $\varphi = (\mu_2 + \mu_3)/2 - \mu_4$ [on reconnais, cependant, que d'autres interprétations sont possibles, comme $\mu_2 = \mu_4$ et $\mu_3 = \mu_4$]. $\hat{\varphi} = (\bar{y}_2 + \bar{y}_3)/2 - \bar{y}_4$ et $\hat{\sigma}_\varphi = \hat{\sigma}\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2} + \frac{1}{n_3}}$.
La statistique de test est $T = \dfrac{\hat{\varphi}}{\hat{\sigma}_\varphi} = \dfrac{(48{,}7 + 48{,}8)/2 - 45{,}0}{\sqrt{7{,}443}\sqrt{\frac{1}{20} + \frac{1}{20} + \frac{1}{5}}} = 2{,}510$. Le point critique est $t_{16;0{,}025} = 2{,}12$. On rejette H_o à 5 %.

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

Ici on teste l'hypothèse $H_0 : \mu_2 = \mu_3$. La statistique est $T = \dfrac{\bar{y}_2 - \bar{y}_3}{\hat{\sigma}\sqrt{1/5 + 1/5}} = -0{,}06$, ce qui ne peut certainement pas justifier le rejet de l'hypothèse.

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

Intervalle de confiance pour $\varphi = (\mu_2 + \mu_3)/2 - \mu_4$: $[\hat{\varphi} - t_{16;0,025}\hat{\sigma}_{\hat{\varphi}} ; \hat{\varphi} + t_{16;0,025}\hat{\sigma}_{\hat{\varphi}}]$. $\hat{\varphi} = (\bar{y}_2 + \bar{y}_3)/2 - \bar{y}_4 = 3{,}74$, $t_{16;0,025} = 2{,}12$; $\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{20} + \frac{1}{20} + \frac{1}{5}} = 1{,}494$. L'intervalle de confiance est donc [3,74-2,12(1,494) ; 3,74+2,12(1,494)] = [0,582 ; 6,918].

Intervalle de confiance pour $\eta = \mu_2 - \mu_3$: $[\hat{\eta} - t_{16;0,025}\hat{\sigma}_{\hat{\eta}} ; \hat{\eta} + t_{16;0,025}\hat{\sigma}_{\hat{\eta}}]$. $\hat{\eta} = \bar{y}_2 - \bar{y}_3 = -0{,}01$, $t_{16;0,025} = 2{,}12$; $\hat{\sigma}_{\hat{\eta}} = \hat{\sigma}\sqrt{\frac{1}{5} + \frac{1}{5}} = 1{,}725$.

L'intervalle de confiance est donc [-0,1 - 2,12(1,725) ; -0,1 + 2,12(1,725)] = [-3,76 ; 3,56].

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 étudiants des groupes 1, 2 et 3 suivaient un programme autre que sciences. Les résultats au cours de statistique sont pr1ésentés dans le tableau 9.2.

Tableau 9.2

Notes en statistique de quatre groupes d'étudiants

Groupe 1 Ni math ni stat		Groupe 2 Math et stat		Groupe 3 Math sans stat		Groupe 4 Sciences	
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	

a) Dressez une table d'analyse de variance

Données

	Groupe 1	Groupe 2	Groupe 3	Groupe 4	
n_i	18	22	15	23	$\bar{y} = 63$
$\bar{y}_i$	1112/18	1274/22	850/15	1678/23	SCE = 3478,8
$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	4659,111	5251,818	5725,333	8398,957	SCR = 24035,2

Table d'analyse de variance (produit par un logiciel ; la valeur p ne peut être calculée sans logiciel)

```
          Dl Somme des carés Moyenne des carrés        F        vp
Facteur    3       3478.781          1159.5936 3.765113 0.0142159
Résiduel  74      22790.800           307.9838
Total     77      26269.580
```

On peut rejeter (avec $\alpha = 0{,}02$) l'hypothèse que les 4 moyennes sont toutes égales.

b) Testez l'hypothèse que la moyenne du groupe 4 est égale à la moyenne des trois autres moyennes.

On interprétera cette hypothèse comme H_o : $\mu_4 = (\mu_1 + \mu_2 + \mu_3)/3$.

Soit $\varphi = (\mu_1 + \mu_2 + \mu_3)/3 - \mu_4$; $\hat{\varphi} = \frac{1}{3}(\bar{y}_1 + \bar{y}_2 + \bar{y}_3) - \bar{y}_4 = -14{,}172$; $\hat{\sigma} = 17{,}54947$;

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{9n_1} + \frac{1}{9n_1} + \frac{1}{9n_3} + \frac{1}{n_4}} = 4{,}3736$; et la statistique de test est $T = \frac{-14{,}172}{4{,}3736} = -3{,}240$ (à 74 degrés de liberté); *vp* = 0,0018.

On conclut donc que le groupe 4 se démarque de l'ensemble des groupes 1, 2 et 3.

c) Testez l'hypothèse que la moyenne du groupe 2 est égale à la moyenne du groupe 3.

Soit $\varphi = \mu_2 - \mu_3$. Nous testons l'hypothèse H_o : $\varphi = 0$. $\hat{\varphi} = \bar{y}_2 - \bar{y}_3 = 1{,}242424$;

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}} = 18{,}02222\sqrt{\frac{1}{22} + \frac{1}{15}} = 6{,}03465$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = 0{,}20588$, tout-à-fait non significatif (*vp* = 0,837).

d) Selon les dires des enseignants de statistique, la performance d'un étudiant dans un cours universitaire de statistique dépend surtout de sa compétence en mathématiques (et non en statistique). Cette information a priori, ajoutée au résultat en c), suggère qu'il est raisonnable de considérer les groupes 2 et 3 comme étant comparables. Donc réunissez ces deux groupes en un seul et reprenez l'analyse.

Statistiques descriptives :

	Groupe 1	Groupes 2 et 3	Groupe 4	
n_i	18	37	23	$\bar{y} = 63$
$\bar{y}_i$	1112/18	2124/37	1678/23	SCE = 3465,013
$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	4659,111	10 990,919	8398,957	SCR = 24 049

Table d'analyse de variance :

	Dl	Somme des carrés	Moyenne des carrés	F	vp
Facteur	2	3465.013	1732.5067	5.697894	0.004969652
Résiduel	75	22804.567	304.0609		
Total		26269.580			

On peut donc conclure qu'il y a bel et bien des différences entre les moyennes.

e) Désignons les trois moyennes du modèle développé en d) par μ_1, μ_{23}, et μ_4. Dans ce modèle, tester l'hypothèse que la moyenne des étudiants en sciences est égale à la moyenne des deux autres moyennes, c'est-à-dire, $\mu_4 = (\mu_1 + \mu_{23})/2$.

Soit $\varphi = \mu_4 - (\mu_1 + \mu_{23})/2$. $\hat{\varphi} = \bar{y}_4 - \frac{1}{2}(\bar{y}_1 + \bar{y}_{23}) = -13{,}36493$; $\hat{\sigma} = 17{,}90679$.

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{n_4} + \frac{1}{4n_1} + \frac{1}{4n_{23}}} = 4{,}414606$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = \frac{\bar{y}_3 - \frac{1}{2}(\bar{y}_1 + \bar{y}_2)}{\hat{\sigma}\sqrt{\frac{1}{n_4} + \frac{1}{4n_1} + \frac{1}{4n_{23}}}} = 3{,}027$ (75 degrés de liberté). *vp* = 0,00338 (test bilatéral).

f) Dans le modèle en d) tester l'hypothèse que la moyenne du groupe 1 est égale à la moyenne des groupes 2 et 3 réunis (H_o : $\mu_1 = \mu_{23}$ dans la notation en e)).

$\hat{\varphi} = \bar{y}_1 - \bar{y}_{23} = 4{,}3723$; $\hat{\sigma} = 17{,}437$; $\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_{23}}} = 17{,}90679\sqrt{\frac{1}{18} + \frac{1}{37}} = 5{,}011$

$T = \frac{\bar{y}_1 - \bar{y}_{23}}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_{23}}}} = 0{,}8726$, ce qui est manifestement non significatif (*vp* = 0,386).

g) Combinez les 3 premiers groupes et tester l'hypothèse que la moyenne du groupe reconstitué (réunissant les groupes 1, 2 et 3) est égale à la moyenne du groupe 4.
La différence entre les deux moyennes est de 14,12, la statistique T = 3,27 (76 degrés de liberté); *vp* = 0,002.

h) Résumer l'ensemble de vos conclusions
Il semblerait que le seul facteur qui contribue au succès en statistique est le programme suivi au Cégep: ceux qui ont suivi le programme de sciences ont un avantage sur les autres. Nous n'avons pas pu démontrer que le seul fait d'avoir suivi un cours de mathématiques ou un cours de statistique aide à mieux performer en statistique à l'université. NOTE : On ne peut pas sanctionner cette conclusion globale par un niveau de confiance, étant donné le grand nombre de décisions prises pour y arriver. Cette démarche, qui vise à construire un modèle, peut être qualifiée d' « exploratoire » plutôt que « confirmatoire », donc provisoire. Mais le niveau de confiance, bien que difficilement quantifiable à ce stade, est néanmoins assez élevé du fait

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 1380 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.
Différence entre les moyennes : 6,35; S = 11,96526; T = 13,40 (2561 degrés de liberté); *vp* ≈ 0 (test bilatéral).

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.
Différence entre les moyennes : 2,29; S = 9,99; $|T|$= 4,17, 1378 degrés de liberté; *vp* = 0,000032 (test bilatéral).

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.

	Sommes	D.L.	Moyennes	F
Groupe	27453	1	13727	96,3
Erreur	364915	2560	142,5	
Total	392368			

d) Dans le cadre du modèle en c) testez l'hypothèse que les filles et les garçons nigérians ont la même moyenne. On rejette l'hypothèse ici comme on la rejette en b). Qu'est-ce qui explique que la statistique *t* est légèrement plus petite ici qu'en c)?
$\hat{\varphi}$ = -2,23; $\hat{\sigma}$ = 11,93922; $\hat{\sigma}_{\hat{\varphi}}$ = 5,00,656; T = 3,49; vp = 0,0005. On peut rejeter l'hypothèse avec confiance.

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 $\mu_3 = (\mu_1 + \mu_2)/2$.)
Soit $\varphi = \mu_3 - (\mu_1 + \mu_2)/2$. $\hat{\varphi} = |\bar{y}_3 - (1/2)(\bar{y}_1 + \bar{y}_2)|$ = 6,125; $\hat{\sigma}$ = 11,93922; écart-type de $\hat{\sigma}_{\hat{\varphi}}$ = 0.4775896; T = 12,82482, 2560 degrés de liberté. La différence est certainement significative.

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 le degré d'introversion ne dépend ni du sexe ni du signe du zodiac.

	Degrés de liberté	Somme des carrés	Moyenne des carrés	F	vp
Facteur	3	147,,60	49,20060	2,516972	0,0589
Résiduel	237	4632,77	19,54753		
Total	239	4780,37			

On affirmera qu'il y a des différences entre les moyennes si on accepte un risque de 6 %.

b) Tester l'hypothèse que les hommes et les femmes ont la même moyenne, c'est-à-dire, $\mu_1 + \mu_2 = \mu_3 + \mu_4$.

Soit $\varphi = \mu_1 + \mu_2 - \mu_3 - \mu_4$ et $\hat{\varphi} = \bar{y}_1 + \bar{y}_2 - \bar{y}_3 - \bar{y}_4 = 2,12$; $\hat{\sigma}_{\hat{\varphi}} = 1,226194$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = 1,728927$ à 237 degrés de liberté; $vp = 0,08512435$. On affirmera qu'il y a une différence entre les femmes et les hommes si on accepte un risque de 10 %.

c) 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, $\mu_1 + \mu_3 = \mu_2 + \mu_4$.

Soit $\varphi = \mu_1 + \mu_3 - \mu_2 - \mu_4$ et $\hat{\varphi} = \bar{y}_1 + \bar{y}_3 - \bar{y}_2 - \bar{y}_4 = -2,58$; $\hat{\sigma}_{\hat{\varphi}} = 1.226194$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = -2,104072$, à 237 degrés de liberté; $vp = 0,03642605$. On affirmera qu'il y a une différence entre les positifs et les négatifs si on accepte un risque de 5 %.

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

Soit $\varphi = \mu_1 - \mu_2 - \mu_3 + \mu_4$ et $\hat{\varphi} = \bar{y}_1 - \bar{y}_2 - \bar{y}_3 + \bar{y}_4 = -1,46$; $\hat{\sigma}_{\hat{\varphi}} = 1,226194$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = -1,190676$; à 237 degrés de liberté; $vp = 0,2349719$.

e) Tester chacune des hypothèses suivantes :

i) Le degré d'introversion ne dépend pas du signe de Zodiac chez les hommes;

Soit $\varphi = \mu_1 - \mu_2$. $\hat{\varphi} = \bar{y}_1 - \bar{y}_2 = -2,02$; $\hat{\sigma}_{\hat{\varphi}} = 1,014306$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = -1,991509$ à 237 degrés de liberté; $vp = 0,04757294$.

ii) Le degré d'introversion ne dépend pas du signe de Zodiac chez les femmes;

Soit $\varphi = \mu_3 - \mu_3$. $\hat{\varphi} = \bar{y}_3 - \bar{y}_4 = -0,56$; $\hat{\sigma}_{\hat{\varphi}} = 0,6890094$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = -0,812761$ à 237 degrés de liberté; $vp = 0,4171708$

iii) Le degré d'introversion ne dépend pas du signe de Zodiac—ni chez les femmes ni chez les hommes, c'est-à-dire, l'hypothèse H_o : $\mu_1 = \mu_2$ *et* $\mu_3 = \mu_4$.

Le modèle restreint est un modèle dans lequel les groupes 1 et 2 sont réunis en un seul de même que les groupes 3 et 4. Dans ce modèle, les effectifs sont 76 et 165; les moyennes sont 14,51000 et 13,46188; et les écarts-types sont 4,386558 et 4,472179. Nous obtenons la table d'analyse de variance suivante :

	Degrés de liberté	Somme des carés	Moyenne des carrés	F	vp
Facteur	1	57,16	57,16149	2,892441	0,09029668
Résiduel	239	4723,21	19,76237		
Total	240	4780,37			

Donc SCR_0 = 4723,20552 à 239 degrés de liberté alors que dans le modèle initial, SCR = 4632,765 à 237 degrés de liberté.

La statistique de test est $F = \dfrac{(4723,20552 - 4632,765)/2}{(4632,765)/237}$ = 2,313, à 2 et 237 degrés de liberté; *vp* = 0,1012.

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.

Table d'analyse de variance

	Sommes	D.L.	Moyennes	F
Église	299,57	2	149,78400	118,554
Erreur	703,73	557	1,2634242	
Total	1003,26	559	1,7948037	

Le point critique pour *F* est (à 5 %) est 3,01. On peut certainement rejeter H_0 (selon laquelle la permissivité sexuelle est liée à la fréquentation de l'Église.)

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

Soit $\varphi = \mu_3 - (\mu_1 + \mu_2)/2$; $\hat{\varphi} = \bar{y}_3 - (\bar{y}_1 + \bar{y}_2)/2 = 1,22$; $\hat{\sigma} = 1,124021$; $T = \dfrac{1,22}{1,124021\sqrt{\frac{1}{202} + \frac{1}{4(128)} + \frac{1}{4(230)}}}$ = 12,142. Cette statistique est de loi de Student à 557 degrés de liberté. On peut certainement rejeter l'hypothèse. On conclut que ceux qui ne vont jamais à l'Église sont plus permissifs que ceux qui y vont.

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.

Soit $\varphi = \mu_1 - \mu_2$; $\hat{\varphi} = \bar{y}_1 - \bar{y}_2$ = -1,42, $T = \dfrac{-1,42}{1,124021\sqrt{\frac{1}{128} + \frac{1}{230}}}$ = -11,45621. Encore une fois, la différence est fortement significative. Ceux qui vont à l'Église de temps en temps sont plus permissifs que ceux qui y vont régulièrement.

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.

Soit $\varphi = \mu_1 - \mu_3$; $\hat{\varphi} = \bar{y}_1 - \bar{y}_3$ = --0,51, $T = \dfrac{-0,51}{1,124021\sqrt{\frac{1}{230} + \frac{1}{202}}}$ = -4,705. La différence est significative.

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.

Il n'est pas évident qu'on veuille amalgamer les deux derniers groupes, mais nous le ferons quand même. La moyenne des groupes 2 et 3 combinés est 4,968472 et la différence entre le premier groupe et la moyenne des deux autres est -1,658472. L'écart-type *S* de l'échantillon combiné peut être calculée à partir des variances et des moyennes et donne

S = 2,15. On obtient alors une nouvelle estimation de σ, soit $\sqrt{\frac{(128-1)(1,54)^2+(432-1)(2,15)^2}{128+432-2}}$ = 2,027. La statistique devient $T = \frac{-1,6658472}{2,027\sqrt{\frac{1}{128}+\frac{1}{432}}}$ = -8,13, ce qui est encore significatif.

Remarque 1 On aurait de bonnes raisons de s'objecter à cette combinaison des deux derniers groupes. Le seul fait que la différence entre les deux est significative devrait suffire à proscrire cette opération. Remarquez aussi qu'on obtient une estimation plus pessimiste de σ, qui passe de 1,12 à 2,03, ce qui est dû au fait que la dispersion dans le groupe combiné est élevée, précisément parce que les groupes ne sont pas de même moyenne.

Remarque 2 Supposons que deux échantillons de tailles n_1 et n_2, de moyennes $\bar{y}_1$ et $\bar{y}_2$, et d'écarts-types S_1 et S_2, voici comment calculer l'écart-type S de l'échantillon combiné. Soit $n = n_1 + n_2$, $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n}$ la moyenne de l'échantillon combiné. La variance S^2 est $\frac{(n_1-1)S_1^2+(n_2-1)S_2^2+n_1\bar{y}_1^2+n_2\bar{y}_2^2-n\bar{y}^2}{n_1+n_2-1}$.

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 ont 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.

Table d'analyse de variance :

```
Source      Df    Sum_Sq     Mean_Sq    F-value    Pr(>F)
catégorie   3     3641.5     1213.83    64.861     1.012e-12    ***
Residuals   28    524.0      18.71
```

La probabilité `Pr(>F)` est la valeur p. Elle est extrêmement faible, ce qui signifie qu'il y a des différences entre les quatre catégories. L'estimation de σ est $\hat{\sigma} = \sqrt{MCR} = \sqrt{18,71} = 4,326$.

b) Tester l'hypothèse que la moyenne de ceux qui ont suivi le cours est la même que celle de ceux qui ne l'ont pas suivi. Si on désigne les quatre moyennes des populations à l'aide de la notation matricielle usuelle, μ_{11}, μ_{12}, μ_{21}, μ_{22}, l'hypothèse à tester est H_o : $\varphi = (\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2 = 0$.
L'estimation de φ est $\hat{\varphi} = (\bar{y}_{11}+\bar{y}_{12})/2-(\bar{y}_{21}+\bar{y}_{22})/2 = 7,75$ où $\bar{y}_{11}$, $\bar{y}_{12}$, $\bar{y}_{21}$ et $\bar{y}_{22}$ sont les moyennes échantillonnales. L'écart-type de cet estimateur est $\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{4n_{11}}+\frac{1}{4n_{12}}+\frac{1}{4n_{21}}+\frac{1}{4n_{22}}} = 1,529$. La statistique de test est $\frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = 5,06$, une statistique de loi de Student (sous H_o) à 32 - 4 = 28 degrés de liberté. Le point critique est 2,048 (*vp* = 0,000023). On peut donc rejeter H_o.

c) Testez l'hypothèse que le cours a le même effet chez ceux de 11 ans de scolarité que chez ceux de plus de 11 ans de scolarité (l'« effet » du cours est mesuré par la différence des deux moyennes).
Il s'agit de l'hypothèse H_o : φ = 0, où $\varphi = \mu_{11} - \mu_{21} - (\mu_{12} - \mu_{22})$.
L'estimation de φ est $\hat{\varphi} = (\bar{y}_{11}-\bar{y}_{12})-(\bar{y}_{21}-\bar{y}_{22}) = 2,25$.
L'écart-type de cet estimateur est $\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} = 1,529472$.

La statistique de test est $\frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = \frac{(\bar{y}_{11}-\bar{y}_{21})-(\bar{y}_{12}-\bar{y}_{22})}{\hat{\sigma}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} = 1{,}47$, une statistique de loi de Student (sous H_o) à 32 - 4 = 28 degrés de liberté. Le point critique est 2,048 (*vp* = 0,1524). On ne peut donc par rejeter H_o.

d) Supposons qu'on ait décidé plutôt de ne pas tenir compte des niveaux de scolarité. On aurait alors un modèle à deux groupes de taille 16—ceux qui ont suivi le cours et ceux qui ne l'ont pas suivi.

i) Tester dans ce modèle l'hypothèse que la moyenne la moyenne des deux groupes est la même.
Il s'agit d'un test d'égalité de deux moyennes. Les moyennes sont 42 (pour ceux qui ont suivi le cours) et 34,25 pour ceux qui ne l'ont pas suivi. L'écart-type estimé est

$$\hat{\sigma} = \sqrt{\frac{(16-1)S_1^2+(16-1)S_2^2}{16+16-2}} = \sqrt{\frac{(16-1)(12{,}149074)^2+(16-1)(9{,}902862)^2}{16+16-2}} = 11{,}08.$$

La statistique *t* est $T = \frac{\bar{y}_1-\bar{y}_2}{\hat{\sigma}\sqrt{\frac{1}{16}+\frac{1}{16}}} = -1{,}978$, une statistique de loi de Student (sous H_o) à 32 - 2 = 30 degrés de liberté. Le point critique ($\alpha = 0{,}05$) est 2,042. La différence n'est donc pas significative.

ii) Expliquer pourquoi une différence qui était significative en b) ne l'est plus ici [comparer l'estimation de σ ici et en b)].
Examinons d'abord l'estimation de σ dans les deux modèles. Dans le modèle initial (à 4 groupes) on trouve $\hat{\sigma} = 4{,}326$ alors que dans le modèle à deux groupes considéré ici, $\hat{\sigma} = 11{,}08$. Les deux écarts-types mesurent la dispersion à l'intérieur des groupes, mais le deuxième est plus grand car les deux groupes sont moins homogènes, étant constitués de personnes avec des scolarités plus variables. Quand l'écart-type est grand, la probabilité de rejeter l'hypothèse est plus faible, et c'est pourquoi il n'a pas été possible de rejeter H_o ici.

iii) Justifier les affirmations suivantes :

- Si, afin de constituer chacun des deux groupes, on avait délibérément choisi 8 personnes parmi ceux qui ont 11 ans et moins de scolarité et 8 parmi ceux qui en ont plus, alors le modèle décrit dans ce numéro n'est pas correct (dans le sens que les hypothèses du modèle ne sont nettement pas vérifiées).
- Le modèle exige que les observations dans un même groupe soient de même moyenne (et même variance). Ce ne serait pas le cas si on procédait de la façon décrite car la moyenne de la population ayant une scolarité de 11 ans ou moins est possiblement différente ce celle dont la scolarité est supérieure à 11 ans.

- Si on avait laissé le choix des sujets au hasard (sans tenir compte de la scolarité), le modèle aurait été correct mais moins efficace (les tests auraient été moins puissant).
- Dans ce cas, les 16 observations de chaque groupe auraient été faites dans les mêmes conditions. Elles auraient donc la même moyenne et la même variance. Considérons le groupe expérimental. Chaque observation est de moyenne $p\mu_{11} + (1 - p)\mu_{12}$ et de variance $\sigma^2 + p(1 - p)(\mu_{11} - \mu_{12})^2$, où σ^2 est la variance pour une classe de scolarité donnée (≤ 1 ou > 11) supposée commune aux 4 groupes définis au début; et *p* est la proportion des individus de la population de scolarité ≤ 11. Le modèle est donc adéquat (dans la mesure où les hypothèses sur σ sont à peu près vérifiées).

iv) Soit H_o l'hypothèse que le cours de sensibilisation est inutile et supposons que les sujets ont été choisis au hasard, sans tenir compte de la scolarité. Montrer que l'hypothèse de l'égalité des deux moyennes n'est pas l'hypothèse testée en b), à moins que *p*, la proportion de la population avec 11 ans de scolarité ou moins, soit égale à ½.
L'hypothèse testée ici est $p\mu_{11} + (1 - p)\mu_{12} = p\mu_{21} + (1 - p)\mu_{22}$ alors qu'en b) l'hypothèse testée est $\mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$. Les deux hypothèses sont identiques si et seulement si $p = ½$.

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

Quelques statistiques :

Moyennes théoriques (Notation)

Sexe	DELTA	LOGO
F	μ_{11}	μ_{12}
M	μ_{21}	μ_{22}

Effectifs

Sexe	DELTA	LOGO
F	7	6
M	4	5

Moyennes

Sexe	DELTA	LOGO
F	1,0	5,333
M	-1,5	2,600

Sommes des carrés (Écarts par rapport à la moyenne)

Sexe	DELTA	LOGO
F	90	117,333
M	45	37,200

Écarts-types

Sexe	DELTA	LOGO
F	3,87298	4,84424
M	3,87298	3,04959

a) Soit *Y* la différence Après-Avant. Dresser une table d'analyse de variance considérant que l'échantillon est composé de quatre groupes : Filles*Delta (μ_{11}) ; Filles*Logo (μ_{12}) ; Garçons*Delta (μ_{21}) ; et Garçons*LOGO (μ_{22}).

Table d'analyse de variance :

```
          Df Sum Sq Mean Sq F value  Pr(>F)
Groupe     3 124.28  41.428  2.5756 0.08587 .
Residuals 18 289.53  16.085
```

On pourrait rejeter l'hypothèse que toutes les moyennes sont égales si on acceptait un risque de 10 % de tirer une fausse conclusion.

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

On teste $\varphi = 0$, où $\varphi = \mu_{11} + \mu_{21} - \mu_{12} - \mu_{22} = 0$. $\hat{\varphi} = \bar{y}_{11} + \bar{y}_{21} - (\bar{y}_{12} + \bar{y}_{22}) = -8{,}433$; $\hat{\sigma} = 4{,}0106$;

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{7}+\frac{1}{4}+\frac{1}{6}+\frac{1}{5}} = 3{,}49$; $T = \dfrac{-8{,}4333}{4{,}0106\sqrt{\frac{1}{7}+\frac{1}{4}+\frac{1}{6}+\frac{1}{5}}} = -2{,}412$. Le point critique est 2,10; on rejette H_0 à 5 %. La valeur *p* est 0,0267. On rejette l'hypothèse si $\alpha = 0{,}05$.

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

On teste $\varphi = 0$, où $\varphi = \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} = 0$. $\hat{\varphi} = \bar{y}_{11} + \bar{y}_{12} - \bar{y}_{21} - \bar{y}_{22} = 5{,}233$; $\hat{\sigma} = 4{,}0106$;

$\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{7}+\frac{1}{6}+\frac{1}{4}+\frac{1}{5}} = 3.495$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = \frac{5{,}23333}{4{,}0106\sqrt{\frac{1}{7}+\frac{1}{6}+\frac{1}{4}+\frac{1}{5}}} = 1{,}497$. Le point critique à 5 % est 2,101. Aucune raison de rejeter H_o. La valeur *p* est 0,152. On ne rejette pas l'hypothèse à moins d'être disposé à prendre un grand risque de tirer une fausse conclusion.

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

On teste $\varphi = 0$, où $\varphi = \mu_{11}-\mu_{21}-(\mu_{12}-\mu_{22})$. $\hat{\varphi} = \bar{y}_{11} - \bar{y}_{12} - (\bar{y}_{12} - \bar{y}_{22}) = -0{,}23333$; $\hat{\sigma} = 4{,}0106$; $\hat{\sigma}_{\hat{\varphi}} = \hat{\sigma}\sqrt{\frac{1}{7}+\frac{1}{4}+\frac{1}{6}+\frac{1}{5}} = 3{,}49$; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}} = \frac{-0{,}23333}{4{,}0106\sqrt{\frac{1}{7}+\frac{1}{6}+\frac{1}{4}+\frac{1}{5}}} = -0{,}067$. Aucune raison de rejeter H_o.

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

On teste $\varphi = 0$, où $\varphi = \mu_{11} - \mu_{12} - (\mu_{21} - \mu_{22})$. Cette hypothèse est identique à la précédente.

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.

$\hat{\sigma} = 4{,}0362$; $T = -2{,}32$. Le point critique à 5 % est 2,101. On peut rejeter H_o à 5 %. *vp* = 0,0308.

9.10 Voici les prix d'un échantillon de maisons vendues dans la région de Montréal, classées en six catégories, selon la taille (nombre de chambres à coucher) et le secteur :

Données brutes

Taille	Secteur		
	Centre	Nord	Sud
Petites (1 ou 2 chambres à coucher)	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
Moyennes (3 chambres à coucher	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
Grandes (4 chambres à coucher 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 des populations et des échantillons :

Taille	Secteur		
	Centre	Nord	Sud
Petites (1 ou 2 chambres à coucher)	μ_{11} 233,7500 ($\bar{y}_{11}$)	μ_{12} 189,6667 ($\bar{y}_{12}$)	μ_{13} 137,8235 ($\bar{y}_{13}$)
Moyennes (3 chambres à coucher	μ_{21} 510,1667 ($\bar{y}_{21}$)	μ_{22} 259,4444 ($\bar{y}_{22}$)	μ_{23} 248,0667 ($\bar{y}_{23}$)
Grandes (4 chambres à coucher ou plus)	μ_{31} 342,6667 ($\bar{y}_{31}$)	μ_{32} 320,5556 ($\bar{y}_{32}$)	μ_{33} 580,3333 ($\bar{y}_{33}$)

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

Taille	Secteur		
	Centre	Nord	Sud
Petites (1 ou 2 chambres à coucher)	25970,75	50223,33	106998,47
Moyennes (3 chambres à coucher	56568,83	141972,22	84456,93
Grandes (4 chambres à coucher ou plus)	69660,67	499990,94	1018799,33

a) Dresser une table d'analyse de variance (9 groupes).

Source	Degrés de liberté	Somme des carrés	Moyenne des carrés	*F*	*vp*
Facteur	8	1 319 682	164 960,3	6,026 973	5,718e-06
Résiduels	75	2 052 775	27 370,33		
Total	83	3 372 457			

$\bar{y}$ = 283,4881, $\hat{\sigma}$ = 165,4398

b) Dans le modèle traité en a) tester les trois hypothèses suivantes en utilisant l'approche décrite à la section 9.6. La somme des carrés résiduelle dans le modèle est SCR = 2 052 775.

i) Le prix des maisons d'une ou deux chambres à coucher ne dépendent pas du secteur, en d'autres termes, l'hypothèse que les trois moyennes $\mu_{11} = \mu_{12} = \mu_{13}$ sont égales. Modifier le modèle de façon à supposer une même moyenne pour toutes les maisons d'une ou deux chambres à coucher, déterminer une analyse de variance dans ce modèle réduit, calculer la somme des carrés résiduelle SCR_o dans le modèle réduit, et utiliser la statistique 9.8.4. (SCR_o = 2 087 831).

Voici l'analyse de variance dans le modèle réduit :

```
> Analyse1$A
                Dl     Somme des carrés     Moyenne des carrés         F           vp
Facteur          6          1284626           214104,31        7,896248    1,21e-06
Résiduel        77          2087831            27114,69
Total                       3372457
```

Cette analyse nous donne SCR_o = 20878319, à ν_o = 77 degrés de liberté. Dans le modèle complet, on a SCR = 2052775 à ν = 75 degrés de liberté. Donc SCR_o – SCR = 35056,1; $F = \dfrac{(2087831 - 2052775)/(77-75)}{2052775/75}$ = 0,6404032; *vp* = 0,5299374. On ne peut pas rejeter l'hypothèse que les prix des petites maisons sont les mêmes (en moyenne) dans les trois secteurs.

La table d'analyse de variance montre, par contre, qu'il y a des différences quelque part entre les 7 moyennes de ce modèle.

ii) Les prix des maisons de trois chambres à coucher ne dépendent pas du secteur.

L'analyse de variance dans le modèle réduit donne ceci:

```
> bii$Reduced
$A
          Dl Somme des carés Moyenne des carrés        F           vp
Facteur    6        999859.3          166643.21 5.408219 0.0001045188
Résiduel  77       2372597.7           30812.96
Total              3372457.0
```

SCR_o = 2 372 598; SCRo - SCR = 319 822,7; $F = \dfrac{(2372598 - 2052775)/2}{2052775/75}$ = 5,842506; *vp* = 0,004384264.

iii) Les prix des maisons de plus de trois chambres à coucher ne dépendent pas du secteur.

L'analyse de variance dans le modèle réduit donne ceci:

```
          Dl Somme des carés Moyenne des carrés        F           vp
Facteur    6         1011109          168518.21 5.495126 8.884328e-05
```

```
Résiduel 77         2361348           30666.85 0.000000 0.000000e+00
          0         3372457               0.00 0.000000 0.000000e+00
```

SCR_o = 2 372 598; SCRo - SCR = 319 822,7; $F = \frac{(2361348 - 2052775)/2}{2052775/75}$ = 5,637; *vp* = 0,00524.

On rejette fermement l'hypothèse. On conclut donc que le prix varie selon le secteur.

c) Utiliser les résultats en c) pour tester l'hypothèse qu'il n'y a pas de différence entre les secteurs dans les prix des maisons de même taille (en d'autres termes $\mu_{11} = \mu_{12} = \mu_{13}$ *et* $\mu_{21} = \mu_{22} = \mu_{23}$ *et* $\mu_{31} = \mu_{32} = \mu_{33}$ pour i = 1, 2 et 3).

Le modèle restreint est un modèle dans lequel intervient seulement la taille de la maison. Voici la table d'analyse de variance pour ce modèle :

```
Analysis of Variance Table

Response: E[, 1]
          Df  Sum Sq Mean Sq F value    Pr(>F)
E[, 2]     2  656230  328115  9,7847 0,0001562 ***
Residuals 81 2716227   33534
```

Donc la somme des carrés résiduelle est SCRo = 2716227, à 81 degrés de liberté. Dans le modèle complet, SCR = 2052775, à 75 degrés de liberté. La statistique $F = \frac{(2716227 - 2052775)/6}{2052775/75}$ = 4,04. La valeur *p* est *vp* = 0,00146.

On peut rejeter l'hypothèse que le secteur n'influence pas (le contraire aurait contredit les résultats en b-ii) et b-iii))).

d) Ignorer maintenant les tailles des maisons et déterminer une analyse de variance avec seul le Secteur comme facteur. Vous verrez que la valeur *p* est de 0,1115, et donc que vous pouvez difficilement conclure à une différence de prix entre les secteurs. Comment expliquer la contradiction avec les conclusions précédentes?

La table d'analyse de variance donne ceci :

```
           Df  Sum Sq Mean Sq F value Pr(>F)
Facteur     2  177817   88908  2,2543 0,1115
Residuals  81 3194640   39440

> anova(lm(E[,1]~E[,3]))
Analysis of Variance Table
Response: E[, 1]
```

e) La distribution des tailles des maisons est donnée dans le tableau suivant :

	Centre	Nord	Sud
Petite	33%	19	24
Moyennes	42	25	53
Grandes	25	56	23
	100%	100%	100%

Tester chacune des hypothèses suivantes :

i) Le prix moyen des maisons du secteur centre est égal au prix moyen des maisons du secteur sud. [Notez bien que les moyennes en question doivent être pondérées. Par exemple, le prix moyen des maisons du centre est $0{,}33\mu_{11}+0{,}42\mu_{21}+0{,}25\mu_{31}$].

La différence entre les deux moyennes est

$|\hat{\varphi}| = |0{,}33\bar{y}_{11} + 0{,}42\bar{y}_{21} + 0{,}25\bar{y}_{31} - 0{,}24\bar{y}_{13} + 0{,}53\bar{y}_{23} + 0{,}23\bar{y}_{33}|$ = 79,04452; $\hat{\sigma}_{\hat{\varphi}}$ = 54.46713; $T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}}$ = 1,4512, à 75 degrés de liberté; *vp* = 0,1508849.

ii) Le prix moyen des maisons du secteur nord est égal au prix moyen est égal au prix moyen des maisons du secteur sud.

$|\hat{\varphi}| = |0{,}33\bar{y}_{12} + 0{,}42\bar{y}_{22} + 0{,}25\bar{y}_{32} - 0{,}24\bar{y}_{13} + 0{,}53\bar{y}_{23} + 0{,}23\bar{y}_{33}|$ = 17,62076; $\hat{\sigma}_{\hat{\varphi}}$ = 40,96595

$T = \frac{\hat{\varphi}}{\hat{\sigma}_{\hat{\varphi}}}$ = 0,4301318 à 75 degrés de liberté, *vp* = 0,668.

f) Résumer vos conclusions concernant les différences entre les secteurs.
Il existe, en fait, des différences entre les secteurs, comme on le voit lorsqu'on compare des maisons de mêmes tailles : leurs prix diffèrent selon le secteur. Mai la taille et le secteur interagissent : l'effet de l'un dépend de l'autre. Par exemple, les petites et moyennes maisons se vendent moins cher dans le secteur sud que dans les deux autres secteurs; alors que les grandes maisons se vendent plus cher dans le secteur sud que dans les deux autres—et ainsi compensent le faible prix des plus petites maisons.

9.11 Démontrez la décomposition SCT = SCE + SCR, soit

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

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

9.14 Considérer l'exercice 9.10-b-i). L'hypothèse à tester est $H_o : \mu_{11} = \mu_{12} = \mu_{13}$.

a) Vérifier numériquement que SCR_o – SCR peut être calculé par la formule $\sum_{j=1}^{3}(\bar{y}_{1j}-\bar{y}_1)^2$ où

$\bar{y}_1 = \frac{1}{n_1}\sum_{j=1}^{3} n_{1j}\bar{y}_{1j}$ et $n_1 = n_{11} + n_{12} + n_{13}$.

b) Montrer que $E\left(\sum_{j=1}^{3}(\bar{y}_{1j}-\bar{y}_1)^2\right) = (3-1)\sigma^2 + \sum_{j=1}^{3} n_{1j}(\mu_{1j}-\bar{\mu})^2$, où $\bar{\mu} = \frac{1}{n_1}\sum_{j=1}^{3} n_{1j}$ et donc que

$$E\left(\frac{\sum_{j=1}^{3}(\bar{y}_{1j}-\bar{y}_1)^2}{3-1}\right) = \sigma^2 + \frac{\sum_{j=1}^{3} n_{1j}(\mu_{1j}-\bar{\mu})^2}{3-1}.$$

c) Justifier l'utilisation de la statistique *F* à l'aide de cette forme d'écriture (c'est-à-dire, montrer que si H_o est fausse, la statistique *F* tend à prendre une valeur élevée).

d) Si la population est normale, le numérateur et le dénominateur soient indépendants de la statistique *F* sont indépendants. Justifier cette affirmation (cette condition est nécessaire pour que *F* suive une loi de Fisher).

9.15 [*Données du tableau* A09] Faites une analyse de variance pour comparer les trois méthodes d'enseignement

a) En vous basant sur les scores A1 et B1.

i) Commencez pas vérifier qu'il n'y a pas de différence par rapport à A1.
On détermine une analyse de variance sur les scores au test A1 (produite par le logiciel R):

```
            Df Sum Sq Mean Sq F value Pr(>F)
traitement   2  20,58 10,2879  1,1322 0,3288
Residuals   63 572,45  9,0866
```

La différence n'est pas significative au départ.
On considère la variable y_1 = B1-A1 comme mesure de l'effet du traitement

ii) Maintenant faites une analyse basée sur les différences B1 – A1.

	Traitement1	Traitement2	Traitement3
Moyennes ($\bar{y}_i$)	-3,81818182	0,04545455	-1,36363636
Écart-type (s_i)	2,872470	2,170742	2,646569

Table d'analyse de variance :

```
            Df Sum Sq Mean Sq F value   Pr(>F)
traitement   2 168,21  84,106  12,636 2,43e-05 ***
Residuals   63 419,32   6,656
```

b) En vous basant sur les scores A2 et B2.

i) Commencez pas vérifier qu'il n'y a pas de différence par rapport à A2.

On détermine une analyse de variance sur les scores au test A1 :

```
            Df Sum Sq Mean Sq F value Pr(>F)
traitement   2   1,12  0,5606  0,1114 0,8948
Residuals   63 317,14  5,0339
```

ii) Maintenant faites une analyse basée sur les différences B1 – A1

On considère la variable y_2 = B2 - A2 comme mesure de l'effet du traitement

	Traitement 1	Traitement 2	Traitement 3
Moyennes ($\bar{y}_i$)	0,2727273	1,1363636	3,4090909
Écart-type (s_i)	2,585340	2,587432	3,216906

Table d'analyse de variance :

```
            Df Sum Sq Mean Sq F value   Pr(>F)
traitement   2 115,48  57,742  7,3008 0,001407 **
Residuals   63 498,27   7,909
```

La différence est significative entre le pré-test et le post-test.

c) En vous basant B3,

Un test basé sur B3 seul a l'inconvénient de ne pas tenir compte des capacités initiales des sujets. Cela ne crée pas nécessairement de biais dans l'analyse, à condition que les sujets soient répartis entre les trois groupes au *hasard*.

	Traitement1	Traitement2	Traitement3
Moyennes ($\bar{y}_i$)	41,04545	46,72727	44,27273
Écart-type (s_i)	5,635578	7,388420	5,766750

Table d'analyse de variance :

```
            Df Sum Sq Mean Sq F value  Pr(>F)
traitement   2  357,3 178,652  4,4811 0,01515 *
Residuals   63 2511,7  39,868
```

Id	T	A1	A2	B1	B2	B3	Id	T	A1	A2	B1	B2	B3
1	1	4	3	5	4	41	34	2	6	2	7	0	55
2	1	6	5	9	5	41	35	2	8	4	10	6	57
3	1	9	4	5	3	43	36	2	9	6	8	6	53
4	1	12	6	8	5	46	37	2	9	4	8	7	37
5	1	16	5	10	9	46	38	2	8	4	10	11	50
6	1	15	13	9	8	45	39	2	9	5	12	6	54
7	1	14	8	12	5	45	40	2	13	6	10	6	41
8	1	12	7	5	5	32	41	2	10	2	11	6	49
9	1	12	3	8	7	33	42	2	8	6	7	8	47
10	1	8	8	7	7	39	43	2	8	5	8	8	49
11	1	13	7	12	4	42	44	2	10	6	12	6	49
12	1	9	2	4	4	45	45	3	11	7	11	12	53
13	1	12	5	4	6	39	46	3	7	6	4	8	47
14	1	12	2	8	8	44	47	3	4	6	4	10	41
15	1	12	2	6	4	36	48	3	7	2	4	4	49
16	1	10	10	9	10	49	49	3	7	6	3	9	43
17	1	8	5	3	3	40	50	3	6	5	8	5	45
18	1	12	5	5	5	35	51	3	11	5	12	8	50
19	1	11	3	4	5	36	52	3	14	6	14	12	48
20	1	8	4	2	3	40	53	3	13	6	12	11	49
21	1	7	3	5	4	54	54	3	9	5	7	11	42
22	1	9	6	7	8	32	55	3	12	3	5	10	38
23	2	7	2	7	6	31	56	3	13	9	9	9	42
24	2	7	6	5	6	40	57	3	4	6	1	10	34
25	2	12	4	13	3	48	58	3	13	8	13	1	48
26	2	10	1	5	7	30	59	3	6	4	7	9	51
27	2	16	8	14	7	42	60	3	12	3	5	13	33
28	2	15	7	14	6	48	61	3	6	6	7	9	44
29	2	9	6	10	9	49	62	3	11	4	11	7	48
30	2	8	7	13	5	53	63	3	14	4	15	7	49
31	2	13	7	12	7	48	64	3	8	2	9	5	33
32	2	12	8	11	6	43	65	3	5	3	6	8	45
33	2	7	6	8	5	55	66	3	8	3	4	6	42

>