T. D. n^{o} 3
Tests de permutations et tests exacts

Exercice 1. *Mangez des pommes !*
On souhaite comparer la teneur en vitamine C de cinq variétés de pommes notées V_1, V_2, V_3, V_4 et V_5. Pour chaque variété la teneur en vitamine C, exprimée en $mg/(100g)$ a été mesurée dans cinq pommes prises au hasard. On obtient les données suivantes regroupées dans le tableau ci-dessous :

V_1	V_2	V_3	V_4	V_5
93,6	95,3	94,5	98,8	94,6
95,3	96,9	97,0	98,2	97,8
96,0	95,8	97,8	97,8	98,0
93,7	97,3	97,0	97,2	95,0
96,2	97,7	98,3	97,9	98,9

Qu'est-il possible de conclure ? Pour répondre à cette question, on construira le tableau de l'analyse de la variance.

Exercice 2. Génotype.
Scheffé (1959, pp. 140–141) a reproduit les données d'une expérience destinée à étudier les variations dans la masse, en g, de rats femelles hybrides en fonction du génotype de la mère nourricière de la portée et du génotype de la portée. Il s'agit des masses moyennes des individus femelles de 61 portées en fonction du génotype de la mère nourricière de la portée et du génotype de la portée. Précisons que les portées proviennent de mères distinctes et n'ont pas eu les mêmes mères nourricières.

1. Télécharger le fichier `foster` du package `HSAUR`.
2. Afficher-le à l'écran.
3. Combien de lignes ? Combien de colonnes ? Quel est le type de chaque colonne ?
4. Vous remarquez qu'il y a deux facteurs dans ce data.frame. On souhaite étudier le gain de masse en fonction de ces deux facteurs à l'aide d'un modèle de l'analyse de la variance. Il se pose alors la question naturelle de savoir si le plan est équilibré ou déséquilibré ? Quel type d'analyse de la variance peut-on faire sous réserve que les conditions soient vérifiées ? À deux facteurs avec interaction ? À deux facteurs sans interaction ?
5. Exécuter la ligne de commande suivante :

```
> plot.design(foster)
```

Remarque : Faire un help de `plot.design`.

6. La figure que vous venez d'obtenir indique que les différences dans les poids pour les quatre niveaux du génotype de la mère sont substancielles. Les différences correspondantes pour le génotype de la lignée sont plus petites.
Nous allons appliquer une analyse de la variance en utilisant la fontion `aov`, mais dans ce cas présent une complication apparait du au fait que le plan n'est pas équilibré. Ici, le nombre d'observations est différent dans chaque groupe. Par conséquent, il n'est plus possible de partitionner la variation que l'on observe dans le jeu de données en des sommes de carrés orthogonales représentant les effets principaux et les interactions.
Dans un plan déséquilibré d'analyse de la variance à deux facteurs notés A et B, il y a une proportion de la variance de la variable réponse qui peut être attribuée soit au facteur A, soit au facteur B. La conséquence est que les facteurs A et B ensemble expliquent moins de variation de la variable dépendante que la somme de chacun explique à elle seule.
Nous allons maintenant envisager deux types d'analyse de la variance. Exécuter à la suite les lignes de commande suivantes :

```
> summary(aov(weight~litgen*motgen, data=foster))
```

et

```
> summary(aov(weight~motgen*litgen, data=foster))
```

Qu'observez-vous ? Il y a une petite différence dans la somme des carrés pour les deux effets principaux, et par conséquent, dans la statistique de Fisher et dans la p-valeur. En fait, cette différence qui apparait est due au type de la somme des carrés qui est calculé. Ici la fonction *aov* calcule une somme de carrés de type I.
Vous pouvez lire cet extrait suivant :
« Le tableau Type I SS est construit en ajoutant les variables une à une dans le modèle, et en évaluant l'impact sur la somme des carrés du modèle. De ce fait, l'ordre dans lequel les variables sont entrées dans le modèle influe sur les résultats obtenus. Le tableau Type III SS est calculé en enlevant ponctuellement chacune des variables du modèle, toutes les autres étant présentes, afin d'évaluer l'impact de la variable supprimée sur le modèle. Ainsi, les valeurs obtenues dans le tableau Type III SS sont indépendantes de l'ordre dans lequel sont sélectionnées les variables. Le tableau Type III SS est souvent préféré pour l'analyse des résultats d'un modèle avec interactions. »
Il ne reste plus qu'à savoir comment calculer les sommes de carrés de Type III. Il est nécessaire d'utiliser la fonction *Anova* (attention avec un A majuscule du package `car`). La ligne de commande à exécuter est la suivante :

```
> Anova(aov(weight~motgen*litgen, data=foster), type="III")
```

7. La représentation graphique classique pour une analyse de la variance se fait à l'aide des lignes de commande suivantes :

```
> layout(matrix((1:4), nrow=2, ncol=2, byrow=T)
> plot(aov(weight~motgen*litgen, data=foster))
```

8. Il est évident qu'il faut absolument vérifier les hypothèses classiques d'une analyse de la variance paramétrique. Pour cela, exécuter les lignes de commande suivantes :

```
> residus<-residuals((aov(weight~litgen*motgen, data=foster)))
> qqnorm(residus)
> qqline(residus)
> shapiro.test(residus)
> library("car")
> bartlett.test(residus~litgen*motgen, data=foster)
```

Remarque : Le test de Bartlett est un test d'égalité des variances construit pour comparer au moins 3 échantillons. Indiquons également que le test de Fisher d'égalité des variances a pour commande « var.test (échantillon 1, échantillon 2) ».

9. Enfin, nous pouvons calculer les estimations des coefficients du modèle en exécutant la ligne de commande suivante :

```
> coefficients(aov(weight~motgen*litgen, data=foster))
```

10. Nous pouvons aussi calculer les valeurs ajustées en exécutant la ligne de commande

```
> fitted.values(aov(weight~motgen*litgen, data=foster))
```

11. Les « anovas » conduisent à dire qu'il y a un effet principal du facteur « motgen », qui signifie que le génotype de la mère ou encore que le facteur « génotype de la mère » est significatif au seuil $\alpha = 5\%$. Maintenant, nous souhaitons savoir quelles sont les différences entre les effets des différents niveaux du facteur. Pour cela, nous allons avoir recours à des tests de comparaison multiples. Choisissons un des modèles d'analyse de la variance.

```
> foster_aov <- aov(weight~litgen*motgen, data=foster)
```

Puis exécuter les lignes de commande suivantes :

```
> foster_hsd <- TukeyHSD(foster_aov, "motgen")
> foster_hsd
```

Qu'observez-vous ? Que concluez-vous ?

Exercice 3. Défaillance d'un élément mécanique

On s'intéresse ici à l'étude de la pièce ayant causé l'explosion de la navette spatiale Challenger en 1986. (Données de Dalal et *al.*, 1986). L'étanchéité du moteur de la navette spatiale est assurée par six pièces identiques appelées « O-ring ». L'explosion de la navette Challenger est due à la défaillance d'au moins l'une de ces pièces. On a reporté dans le tableau ci-dessous des données collectées au cours des 24 vols précédents d'une navette spatiale : la variable $Température$ est la température au moment du lancement, exprimée en °F, et la variable $Défaillance$ est une variable qui vaut 0 si aucun des « O-ring » n'a été endommagé au cours du lancement et 1 si au moins l'un d'entre eux a été endommagé.

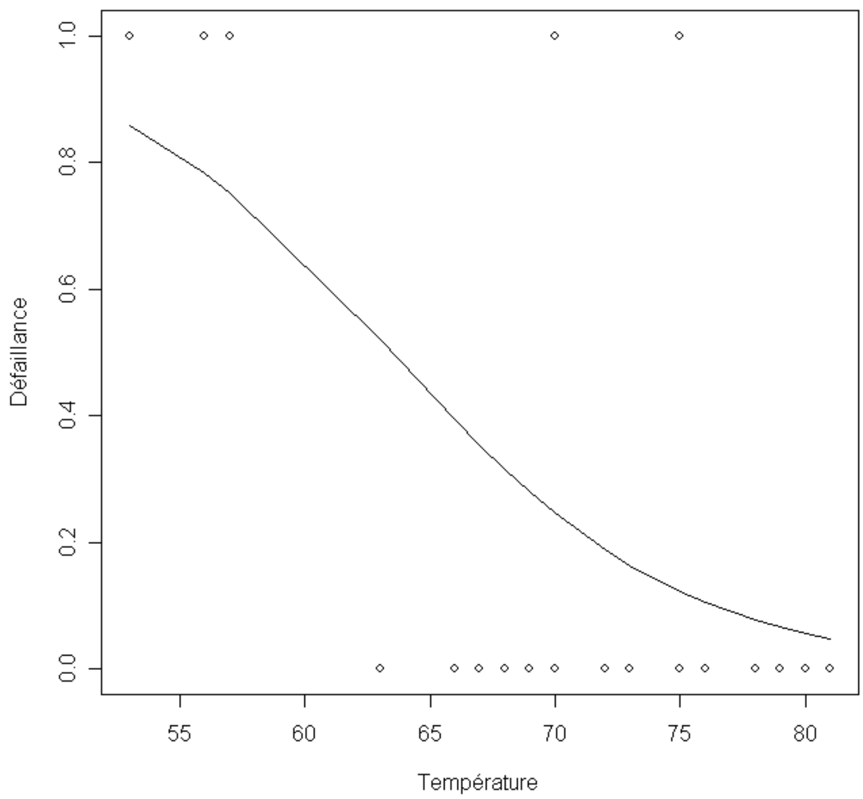
Température	*Défaillance*	*Température*	*Défaillance*
53	1	70	1
56	1	70	1
57	1	72	0
63	0	73	0
66	0	75	0
67	0	75	1
67	0	76	0
67	0	76	0
68	0	78	0
69	0	79	0
70	0	80	0
70	1	81	0

1. Préciser la nature de la réponse observée ainsi que celle du facteur explicatif considéré. Quelle modélisation peut-on utiliser ?

2. Représenter graphiquement à l'aide d'un nuage de points les données du tableau ci-dessus. Que conclure quant à une éventuelle relation entre la température et l'apparition d'une défaillance ?

3. En utilisant une régression logistique, déterminer si le facteur *Température* a une influence sur l'apparition d'une *Défaillance*. Représenter les probabilités prédites par le modèle sur le nuage de points construit à la question **2.**. On comparera le résultat obtenu avec le graphique au verso.

4. Ces données étaient à la disposition des ingénieurs avant le lancement de la navette Challenger. Sachant que le lancement de la navette Challenger a eu lieu à une température de 31 °F, quelle était la probabilité prédite à l'aide du modèle de l'apparition d'au moins une défaillance ? Bien que les valeurs avec lesquelles on a établi le modèle soient éloignées de la valeur de 31 °F, et que de ce fait la probabilité de l'apparition d'au moins une défaillance est modélisée avec une grande incertitude, auriez-vous tout de même autorisé le décollage ?



Exercice 4. *Groupe sanguin et Rhésus*
Le tableau suivant donne la répartition de 10 000 personnes en fonction de leur groupe sanguin et de leur facteur Rhésus.

	O	A	B	AB	Total
Rh_+	3570	3825	935	170	8500
Rh_-	630	675	165	30	1500
Total	4200	4500	1100	200	10000

Les deux caractères, groupe sanguin et Rhésus, sont-ils indépendants ?

Exercice 5. *Campagne publicitaire*
Un publicitaire décide de lancer une campagne sur le thème « la publicité fait vendre ». Dans ce but, il fait prélever au hasard 80 dossiers parmi ceux de ses clients. Il obtient la statistique suivante dans laquelle X désigne le budget publicitaire exprimé en milliers de francs et Y désigne le chiffre d'affaires exprimé en millions de francs.

X/Y	$[0;5[$	$[5;20[$	$[20;100[$	$[100;500[$	$[500;1000[$	Total
$[0;0,2[$	30	1	0	0	0	31
$[0,2;0,8[$	13	10	2	0	0	25
$[0,8;1,0[$	5	4	4	1	0	14
$[1,0;3,0[$	0	1	2	2	1	6
$[3,0;10,0[$	0	0	0	1	3	4
Total	48	16	8	4	4	80

1. Calculer le coefficient de corrélation linéaire entre X et Y.
2. Calculer les rapports de corrélation $\eta^2_{X,Y}$ et $\eta^2_{Y,X}$.