



Université Lumière Lyon 2

M2-IDS – SISE

Inférence statistique

Analyse de la variance & covariance

Travaux dirigés

Année Universitaire 2022-2023

R. Abdesselam

rafik.abdesselam@univ-lyon2.fr

Web : <http://perso.univ-lyon2.fr/~rabdesse/fr/>

Polycopiés : <http://perso.univ-lyon2.fr/~rabdesse/Documents/>

Support informatique - Logiciels d'application :



SPAD 7.4 - Inférence statistique

Projet Edition Affichage Diagramme Dessin Outils ?

Inférence statistique

- Diagrammes de traitements
 - Tests paramétriques
 - ANOVA & ANCOVA

Méthodes

- Imports / Exports de données
- Data Management
- Statistiques descriptives
 - Statistiques de base
 - Générateur automatique de graphiques
 - Tests statistiques
 - Caractérisation d'une variable qualitative
 - Caractérisation d'une variable quantitative
 - Tableaux croisés
 - Analyse bivariée
 - Marquage sémantique des modalités
- Analyses factorielles
- Classifications - Typologies
- Amado - Graphiques de Bertin
- Scoring et Modélisation
- Arbres de décision - Segmentations
- Règles d'association
- Tableaux multiples
- Text Mining
- Modèles structurels
- Industrialisation - Archiving
- Statistiques avec R
- Process

Exécutions

Niveau	Elément	Stop	Statut	Indicateur	Début	Temps	Résultats
0	voitures		Terminé (ok)		30/08/2012	00:00,16	
1	Statistiques de base		Terminé (ok)		08/12/2012	00:00,171	
1	Tests statistiques		Terminé (ok)		08/12/2012	00:00,0	
1	Regression Anova		Terminé (ok)		08/12/2012	00:00,234	

Modélisation (MLG) : ANOVA - ANCOVA

Test de normalité & Intervalles de confiance
Tests paramétriques et non-paramétriques
Comparaisons de moyennes & Variances



SAS

File Edit View Tools Run Solutions Window Help

Results

- Print: The SAS System
- Univariate: The SAS System
- Ttest: The SAS System
- ANOVA: The SAS System
- GLM: The SAS System

Log - (Untitled)

```
NOTE: PROCEDURE TTEST used (Total process time):
  real time    0.06 seconds
  cpu time     0.01 seconds

47 /* Modélisation - Modèle Linéaire Général */
48 proc anova; class sexe; model NPRO = SEXE; run;
49 !                               quit;

NOTE: PROCEDURE ANOVA used (Total process time):
  real time    0.10 seconds
  cpu time     0.09 seconds

49 proc GLM; class sexe nres; model NPRO = SEXE NRES TAPT; run;
49 !                               quit;
```

Inference-ANOVA-ANCOVA.sas

```
9 50 121 F S
10 45 108 M J
11 46 112 M I
12 49 116 F I
13 48 110 F J
14 50 115 F I
15 52 125 F S
16 46 108 M I
17 48 122 M S
18 45 107 M J
19 47 115 M J
20 51 120 F I
21 48 111 F S
22 48 108 F I
23 51 123 F S
24 49 118 F I
25 47 113 M J
26 46 111 M J
27 51 124 F S

run;
proc print label; run;
proc sort; by sexe; run;
/*Test de normalité */
proc univariate normal; by sexe; var npro; run;
/* Test de comparaison de variances & Test de comparaison de moyennes */
proc ttest; class sexe; var npro; run;
/* Modélisation - Modèle Linéaire Général */
proc anova; class sexe; model NPRO = SEXE; run; quit;
proc GLM; class sexe nres; model NPRO = SEXE NRES TAPT; run; quit;
```

NOTE: 49 Lines Submitted.

C:\Users\Rafik Abdesselam Ln 1, Col 1



Fiche N°1 - Exercices d'application SAS
(Initiation aux procédures SAS : Gestion des données - Statistique descriptive)

Exercice 1 : _____

Analyse de 38 marques françaises d'eau en bouteilles décrites par 10 variables homogènes explicitées dans le dictionnaire des variables. Les données (composition en ions (mg/l)) proviennent des informations fournies sur les étiquettes des bouteilles.

DICTIONNAIRE DES VARIABLES - Base de données : Marques-eaux.xls

1. Nom - Nom complet de l'eau inscrit sur l'étiquette	(IDENTIFICATEUR)
2. TE - Type d'eau (Minérale / Source)	(2 MODALITES)
EM - Eau Minérale	ES - Eau Source
3. NE - Nature de l'eau (Plate / Gazeuse)	(2 MODALITES)
EP - Eau plate	EG - Eau gazeuse
4. CA - ions Calcium en mg/litre	(CONTINUE)
5. MG - ions Magnésium en mg/litre	(CONTINUE)
6. NA - ions Sodium en mg/litre	(CONTINUE)
7. K - ions Potassium en mg/litre	(CONTINUE)
8. SULF - ions Sulfates en mg/litre	(CONTINUE)
9. NO3 - ions Nitrates en mg/litre	(CONTINUE)
10. HCO3 - ions Carbonates en mg/litre	(CONTINUE)
11. CL - ions Chlorures en mg/litre	(CONTINUE)

- 1- Création d'une table SAS : Ecrire le programme qui permet de créer la table SAS : EAUX avec les instructions : DATA, INPUT, CARDS et RUN et vérifier avec les procédures : CONTENTS et PRINT le bon déroulement de la création de la table SAS : EAUX.
 - 2- Pour une meilleure identification des variables : améliorer à l'aide des instructions LABEL et FORMAT, la présentation des données de la table EAUX en complétant les libellés des variables ainsi que les modalités des variables alphanumériques.
 - 3- A l'aide des options NODATE et TITLE, supprimer la date et ajouter un titre au tableau.
 - 4- Présenter à l'aide de la procédure SORT, les données avec regroupement : par la composition en nitrates (ordre croissant), par type d'eau puis par nature de l'eau et composition en calcium.
 - 5- Etablir les statistiques sommaires des variables quantitatives à l'aide de la procédure MEANS et une description plus complète des variables calcium, sulfates et chlorures à l'aide de la procédure UNIVARIATE.
 - 6- Utilisation de la procédure SORT et de la procédure MEANS : calculer les statistiques sommaires des trois variables quantitatives composition en ions (calcium, sulfates et chlorures) par modalité de la variable nominale : nature de l'eau.
 - 7- Donner la répartition c'est-à-dire le tri à plat du type d'eau et de la nature de l'eau, ainsi que le tableau croisé des deux variables à l'aide de la procédure FREQ.
 - 8- Dépouillement de variables quantitatives et d'une variable qualitative : déterminer les distributions de fréquences absolues des variables potassium, carbonates ainsi les distributions de fréquences absolues de ces variables selon la nature de l'eau.
-

Exercice 2 :

- 1- Création d'une table SAS EMPLOYES à partir d'un fichier : importer les données de la base « employes.xls » ou encore de la base SPAD « employes.sba » .
- 2- Ecrire le programme qui permet de créer une nouvelle table SAS en complétant les libellés des variables ainsi que les modalités des variables alphanumériques.

DICTIONNAIRE DES VARIABLES - Base de données : Employes.xls

1. Sexe	(2 MODALITES)	
FEMM - Femme	HOMM - Homme	
2. Niveau d'éducation (années passées à l'école)	(3 MODALITES)	
<12 - moins de 12 ans	1216 - [12 ans -16 ans[>=16 - 16 ans et plus
3. Statut professionnel	(3 MODALITES)	
EBUR - Employé de bureau	ASEC - Agent de sécurité	MANA - Manager
4. Nationalité française	(2 MODALITES)	
OUI - Française	NON - Autre	
5. SCOU - Salaire courant (euros)	(CONTINUE)	
6. SEMB - Salaire d'embauche (euros)	(CONTINUE)	
7. ANCI - Ancienneté (nombre de mois)	(CONTINUE)	
8. EXPE - Expérience passée (nombre de mois)	(CONTINUE)	

- 3- Modifier le programme précédent afin de créer une nouvelle variable 'recodage' CSAL 'Classes de salaires' à trois modalités : Petit - < 20 M€) ; Moyen - 20 - 25 M€ ; Grand - > 25 M€..
- 4- Si tous les employés doivent payer un impôt de 30% sur le salaire courant, écrivez le programme qui permet d'imprimer la liste ordonnée par sexe montrant le salaire courant, le montant de l'impôt et qui indique le montant total de l'impôt.
- 5- Modifier le programme précédent pour que les résultats soient imprimés par ordre décroissant du montant de l'impôt à payer.
- 6- Ecrire le programme SAS qui permet d'obtenir la moyenne et l'écart-type de toutes les variables.
Modifier le programme précédent pour obtenir ces statistiques par sexe.
Représenter la variable sexe par un diagramme circulaire en pourcentage du salaire courant moyen.
Représenter le salaire courant par sexe et par niveau d'éducation.
- 7- Donner par sexe, la description statistique complète de la variable salaire courant. (indices statistiques, diagrammes en boîte, test de normalité).
- 8- Quels sont la moyenne et l'écart-type du salaire courant selon le sexe ? Selon le niveau d'éducation ?
- 9- Représenter à l'aide d'un stéréogramme le salaire courant moyen selon le sexe et le niveau d'éducation.
- 10- Quel est le coefficient de corrélation linéaire entre le salaire courant et le salaire d'embauche ?
Représenter à l'aide d'un diagramme de dispersion le salaire courant en fonction du salaire d'embauche selon le sexe.
- 11- Donner les répartitions des variables qualitatives.
Donner la répartition de l'ensemble des employé(e)s selon le sexe et le statut professionnel.
- 12- Donner la répartition des employées de nationalité française selon le niveau d'éducation et le statut professionnel (filtre - sous population).
- 13- Standardiser les variables continues en vue d'une analyse de régression.

Fiche N°2 - Exercices d'application SAS
(SI : Statistique inférentielle)

Exercice SI1 : _____

Un laboratoire indépendant a vérifié pour le compte de l'organisme de la protection du consommateur, la résistance à l'éclatement (en kg/cm^2) d'un réservoir à essence d'un certain fabricant. Les résultats des essais en laboratoire sur un échantillon de 10 réservoirs sont les suivants :

Résistance (kg/cm^2)	215	223	215	221	217	218	222	219	216	224
---------------------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

- Déterminer les statistiques sommaires de la résistance à l'éclatement. Peut-on considérer, avec un risque d'erreur $\alpha = 5\%$, que la résistance à l'éclatement d'un réservoir est normalement distribuée ?
- Estimer par intervalle de confiance la résistance moyenne à l'éclatement de ce type de réservoir avec un niveau de confiance de 95%.
- Pour le même niveau de confiance, déterminer le nombre d'essais requis de sorte que la marge d'erreur dans l'estimation de la résistance moyenne à l'éclatement n'excède pas 1 kg/cm^2 .

Exercice SI2 : _____

Une unité de fabrication de fuel à basse teneur en soufre (en %) doit être réglée de façon à respecter une spécification de teneur en soufre normalement distribuée. La stabilité du réglage de cette unité n'a pu être déterminé qu'à partir d'un échantillon de 12 mesures qui ont permis d'estimer l'écart-type du processus à $s = 0,08$.

- Déterminer un intervalle de confiance de niveau 95% de l'écart-type σ du processus.
- Quel est le risque d'erreur que l'on attribue à cet intervalle de confiance, bilatéral symétrique, de la teneur moyenne en soufre : $m \in [0.1969 ; 0.3031]$ obtenue à partir de cette série de 12 mesures ?
- Quelle est la valeur moyenne m^* de teneur en soufre que l'unité doit avoir pour respecter la spécification c'est-à-dire une teneur en soufre inférieure ou égale à 0,30 avec une probabilité au moins égale à 95% ?

Exercice SI3 : _____

Lors d'une étude statistique portant sur un échantillon de 600 employés d'une entreprise, 450 employés sont favorable à la réduction du temps de travail avec réduction de salaire.

- Déterminer un intervalle de confiance pour la proportion des employés de l'entreprise qui sont favorables à cette réforme, avec un niveau de confiance de 95%, 98% et 99%.
- A quel risque d'erreur correspond l'intervalle $[72\% ; 78\%]$?
- Quelle taille d'échantillon aurait-il fallu choisir pour réduire cet intervalle de moitié ?

Exercice SI4 : _____

Dans un sondage, on accorde a priori 50% de voix à un candidat à une élection.

- Combien faut-il interroger d'électeurs pour confirmer cette hypothèse à 1% près avec un risque d'erreur de 5% ? de 1% ?
- Estimer, par intervalle de confiance de niveau 95%, la proportion de voix d'un candidat à partir d'un sondage portant sur 38416 électeurs.

Exercice SI5 :

Pour de bonnes conditions de vieillissement, une cave à vin doit impérativement être bien isolée pour éviter des variations trop importantes de température préjudiciables à la qualité du vin. Il est donc essentiel de contrôler la variabilité de la température, la référence idéale serait un écart-type de 1.50 °C (degré Celsius). On considère que la température est une variable aléatoire normalement distribuée de moyenne connue $m = 12.00^{\circ}\text{C}$. Afin de contrôler la variabilité de la température, on a relevé 21 fois la température sur une période de deux mois et on a observé un écart-type de la température de la cave de 1.62°C.

- Entre quelles limites peut se situer l'écart-type de la température de cette cave à vin si on utilise un niveau de confiance de 95% ?
- Peut-on considérer avec un seuil de signification 5%, que la variabilité observée des températures est différente de la référence idéale indiquée ? Justifier votre réponse.

Exercice SI6 :

Quel placement choisir ? A partir de deux échantillons de rendements annuels (%), on veut comparer les rendements annuels d'une banque et d'une banque mutualiste.

Rendements										Moyenne	Ecart-type (N)	
Banque	2	1,5	4,2	6,3	9,6	4,3	10,2	10,2		6,04	3,37	
Mutuelle	4	6	8	11,3	5,4	2,6	5,7	10	9,3	7,3	6,96	2,60

Peut-on conclure, au seuil $\alpha = 5\%$, qu'il y a une différence significative entre les performances des deux types de banques ?

- Quel type de test devrait-on utiliser ?
- Effectuer le test approprié. Que peut-on conclure ?

Exercice SI7 :

Résultats de deux méthodes d'apprentissage.

Apprentissage	1	2	3	4	5	6	7	8	Moyenne	Ecart-type (N)
Méthode 1	14	13,5	11	10	9	14	11	15,5	12,25	2,15
Méthode 2	12	10,5	10,5	8	9	12,5	7,5	10	10,00	1,66

Peut-on conclure, au seuil $\alpha = 5\%$, qu'il y a une différence significative entre les deux méthodes d'apprentissage ?

Exercice SI8 : (Comparaison de deux proportions)

Deux types de publicités sont envisagées par une entreprise pour lancer un nouveau produit. Après avoir visionné les deux types de publicités mis au point par des spécialistes en communication, la direction émet l'hypothèse selon laquelle la publicité du type A sera plus efficace que celle du type B. Deux régions, considérées comme marché-test (possédant sensiblement les mêmes caractéristiques de consommation) sont choisies pour évaluer l'efficacité des deux types de publicité. Les résultats du sondage sont résumés dans le tableau suivant :

	Publicité A	Publicité B
se sont procuré le nouveau produit	44	32
ne se sont pas procuré le nouveau produit	81	68
Nombre total de personnes	125	100

- Avec un risque d'erreur $\alpha = 5\%$, déterminer puis comparer les intervalles de confiance des proportions : p_A "personnes influencées par la publicité A", p_B "personnes influencées par la publicité B".
- Donner un intervalle de confiance de niveau 95% pour la proportion ($p_A - p_B$). Que peut-on conclure ? Justifiez votre réponse.
- Est-ce que les résultats de ce sondage permettent de confirmer, au seuil de signification $\alpha = 5\%$, l'hypothèse émise par la direction ? Etablir puis tester les hypothèses statistiques relatives à l'affirmation de la direction.

Exercice SI9 : _____

Les résultats des élections municipales du nouveau maire, dans deux bureaux de vote ou quartiers de la ville, se résument comme suit.

Bureaux de vote	1	2
Intentions de vote	57%	48%
Taille de l'échantillon	200	200

a) Quel est le niveau de confiance que l'on peut attribuer à l'intervalle bilatéral symétrique, de la différence des intentions de vote suivant : $(p_1 - p_2) \in [0.30\% ; 17.70\%]$?

Ecrire les hypothèses statistiques et vérifier, avec un seuil de signification $\alpha = 5\%$, chacune des affirmations suivantes :

- b) Les intentions de vote pour le nouveau maire sont identiques dans les deux bureaux.
- c) On a plus voté pour le nouveau maire dans le bureau 1 que dans le bureau 2.

Exercice SI10 : _____

Pour obtenir une estimation de la population d'hyperglycémiques des personnes âgées de plus de 60 ans d'une ville de Rhône-Alpes, on a choisi au hasard 170 personnes et on a constaté que parmi celles-ci, 34 sont hyperglycémiques.

- a) Déterminer un intervalle de confiance de niveau 95% de la proportion p de personnes hyperglycémiques âgées de plus de 60 ans..
- b) Si l'on effectuait 200 fois le tirage au sort de 170 personnes dans A , on pourrait construire 200 intervalles de confiance du type précédent. Parmi ces 200 intervalles, combien, en moyenne, contiendraient la valeur vraie p ?
- c) Peut-on conclure, avec un risque d'erreur de 5%, que la proportion de personnes hyperglycémiques âgées de plus de 60 ans est inférieure à 25,40% ?

Exercice SI11 : _____

On veut tester les hypothèses suivantes : $H_0 : m = 2500$ contre $H_1 : m < 2500$. Sur la base d'un échantillon de taille $n = 100$, prélevé au hasard d'une population dont l'écart-type $\sigma = 300$, on adopte la règle de décision suivante :

Rejeter l'hypothèse nulle H_0 si $\bar{X}_{100} < 2450,65$ sinon, ne pas rejeter H_0 .

- a) Déterminer la probabilité de commettre une erreur de première espèce avec ce test. Que représente cette probabilité ?
- b) Quel est le risque de deuxième espèce selon l'hypothèse alternative $H_1 : m = 2400$ que l'on suppose vraie ? Quelle est la puissance de ce test ?

Exercice SI2 : _____

On a relevé les prix de deux titres du CAC40 du secteur bancaire des dix dernières années. Les prix de ces titres - actions exprimés en euros (€), sont supposés normalement distribués. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Prix des actions du CAC40											Moyenne	Ecart-type(N)
SG	29,8	29,6	30,2	29,6	29,6	29,7	29,9	30,0	29,8	30,1	29,83	0,21
BNP	26,7	27,2	27,1	28,5	27,5	26,9	26,5	26,6	26,2	26,5	26,97	0,63

- a) Les échantillons sont-ils indépendants ?
- b) Peut-on considérer comme vraisemblable au seuil de signification $\alpha = 5\%$, l'hypothèse selon laquelle les variances des prix de ces deux titres sont différentes ?
- c) Peut-on affirmer avec un risque d'erreur $\alpha = 5\%$, que les prix moyens des deux titres sont différents ?

Indications et résultats : Statistique Inférentielle

SI1 : b) $219 \pm 2,384$ c) $n' \geq 43$ réservoirs (remarque sorties SAS : $\text{ecart-type SAS} = s^* = 3.3333$)

SI2 : a) $[0.0592 ; 0.1418]$ b) 5% c) $m^* \leq 0.255$

SI3 : a) $[71,5\% ; 78,5\%]$, $[70,9\% ; 79,1\%]$, $[70,4\% ; 79,6\%]$ b) 8,92% c) $n = 4 \times 600 = 2400$ employés.

SI4 : a) $n \geq 9604$; $n \geq 16642$ électeurs b) $[49,5\% - 50,5\%]$

SI5 : a) $\sigma \in [1.246^\circ\text{C} ; 2.315^\circ\text{C}]$ b) $\sigma_0 = 1.50^\circ\text{C} \in [1.246^\circ\text{C} ; 2.315^\circ\text{C}]$.

SI6 : Echantillons indépendants de petites tailles. Vérifier que les 2 échantillons sont normalement distribués. Sous H_0 : Egalité des variances : $f_1 = 0.207 < f_0 = 1.73 < f_2 = 4.20$ (Fisher $F_{(7,9)}$ d.d.l.), non rejet de l'hypothèse nulle H_0 . On peut donc considérer, avec un risque d'erreur $\alpha = 5\%$ que les variances de ces rendements sont égales.

Sous H_0 : Egalité des moyennes : $t_1 = -2.12 < t_0 = -0.62 < t_2 = 2.12$ (Student 16 d.d.l.), non rejet de l'hypothèse nulle H_0 . On peut donc considérer, avec un risque d'erreur $\alpha = 5\%$ que les rendements ne sont pas significativement différents.

SI7 : Echantillons appariés de taille $n = 8$ (petite). Vérifier que l'échantillon différence des 2 méthodes d'apprentissage D est bien normalement distribué. Sous H_0 : $m_D = 0$, $t_0 = 3.63 > t_{\alpha/2 = 2.5\%} = 2.365$ (Student à $n - 1 = 7$ d.d.l.), rejet de l'hypothèse nulle H_0 . On peut donc conclure, avec un risque d'erreur $\alpha = 5\%$ que les méthodes d'apprentissage sont différentes.

SI8 : a) 0.352 ± 0.084 et 0.320 ± 0.091 b) 0.032 ± 0.124 ce sondage ne confirme pas l'hypothèse émise par la direction : les 2 types de publicité semblent avoir la même efficacité.

SI9 : a) $1 - \alpha = 92\%$ b) Non rejet de H_0 , pas de différence significative ($\alpha = 5\%$) entre les proportions de vote dans les deux bureaux. c) Rejet de H_0 , on peut donc affirmer ($\alpha = 5\%$) qu'on a plus voté pour le nouveau maire dans le bureau A que dans le bureau B.

SI10 : a) $14\% \leq p \leq 26\%$ b) 190 c) NON

SI11 : a) 5% b) 4,55% 95,45%

SI12 : a) Les échantillons sont indépendants
b) Les variances des prix sont différentes.
c) Les prix moyens des deux titres sont différents - Student à $v = 10.90$ d.d.l.

Fiche N°3 - Exercices d'application SAS - SPAD (ANOVA & ANCOVA)

Exercice AV1 : _____

Un test a été soumis à des étudiants universitaires de niveau licence. On veut vérifier si les résultats du test sont identiques selon que l'étudiant est en première année L1, deuxième année L2 ou troisième année L3. Les résultats obtenus sont les suivants :

Année de licence	L2	L1	L3	L2	L1	L3	L2	L1	L1	L1	L3
Résultat	36	22	27	29	21	36	31	20	22	26	29
Année de licence	L1	L3	L2	L2	L2	L2	L1	L3	L3	L2	L3
Résultat	17	26	25	24	25	26	22	26	20	21	26

Peut-on affirmer, au seuil $\alpha = 5\%$, qu'il y a une différence significative entre les résultats au test selon que l'étudiant est en L1, L2 ou L3 ?

Tester : $H_0 : m_{L1} = m_{L2} = m_{L3}$ versus H_1 : au moins deux moyennes différentes.

Vérifier les conditions d'application du modèle d'analyse de variance :

H_1 : Normalité - les données de chaque population doivent suivre une loi normale.

H_2 : Homoscédasticité - les données de chaque population doivent posséder la même variance.

Exercice AV2 : _____

Pendant la cuisson, les croissants absorbent de la graisse en quantité variable. Nous cherchons à voir si la quantité absorbée dépend du type de graisse. Les quantités de graisse absorbée par croissant, pour quatre types de graisses, sont présentées dans le tableau ci-dessous.

G1	G2	G3	G4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

- Calculer la moyenne pour chaque groupe et la moyenne générale.
- Tester : $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ versus H_1 : les variances σ_k^2 ne sont pas toutes identiques.
- Calculer le tableau d'analyse de variance.
- Peut-on affirmer, au seuil $\alpha = 5\%$, qu'il y a un effet du type de graisse sur les quantités absorbées ?

Exercice AV3 : _____

Table SAS : employés.

- Peut-on considérer, avec un risque d'erreur $\alpha = 5\%$, que le salaire courant, le salaire d'embauche, l'ancienneté et l'expérience sont normalement distribués ?
- Même question selon le niveau d'éducation.

Tester, au seuil de signification $\alpha = 5\%$, l'effet du niveau d'éducation :

- sur le salaire courant des employé(e)s.
- sur le salaire d'embauche des employées

Exercice AV4 :

Une entreprise cherche à tester quatre modèles de machines à écrire. Pour faire ce test, elle demande à cinq secrétaires professionnelles de taper un texte pendant 15 minutes. A la fin du test, on compte le nombre moyen de mots tapés par minute. On répète l'expérience le lendemain. Les résultats (nombre moyen de mots/minute) sont présentés dans le tableau ci-dessous.

		Secrétaires				
		S1	S2	S3	S4	S5
Machines à écrire	M1	33	31	34	34	31
		36	31	36	33	31
	M2	32	37	40	33	35
		35	35	36	36	36
	M3	37	35	34	31	37
		39	35	37	35	40
	M4	29	31	33	31	33
		31	33	34	27	33

Soit le modèle : $y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$ avec $i = 1, 2$ $j = 1 \text{ à } 4$ et $k = 1 \text{ à } 5$.

Où, α correspond à l'influence de la machine à écrire, β l'influence de la secrétaire et γ l'interaction entre la machine j et la secrétaire k .

a) Tester les 3 hypothèses nulles suivantes :

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ versus $H_1 : \text{au moins un } \alpha_j \text{ est différent de } \alpha_{j'}, \alpha_j \neq \alpha_{j'}$.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ versus $H_1 : \text{au moins un } \beta_k \text{ est différent de } \beta_{k'}, \beta_k \neq \beta_{k'}$.

$H_0 : \gamma_{11} = \gamma_{12} = \dots = \gamma_{45}$ versus $H_1 : \text{au moins un } \gamma_{jk} \text{ est différent}$.

La première hypothèse nulle test l'effet séparé du facteur machine, la deuxième l'effet séparé du facteur secrétaire et la troisième les effets séparés avec interaction des deux effets.

Exercice AV5 :

Un fabricant de jus d'orange veut analyser les résultats d'une étude effectuée par l'Organisme de la Défense du Consommateur (ODC). L'étude portait sur 22 marques de jus d'orange selon sept caractéristiques. Le fabricant tente de déterminer quelles sont les principales caractéristiques qui influencent le prix de vente.

DICTIONNAIRE DES VARIABLES- Base de données : Jus-orange.xls

1. CONT - Contenant-commercialisation	(3 MODALITES)
BVER - Bouteille verre	PACK - Pack carton
	BPLA - Bouteille plastique
2. PRIX - Prix de vente moyen (€/litre)	(CONTINUE)
3. PULP - pulpe d'orange (1 : contient / 0 : ne contient pas)	(CONTINUE)
4. DILU - Dilution (%)	(CONTINUE)
5. VITA - Vitamine C (%)	(CONTINUE)
6. FOLA - Folacine (%)	(CONTINUE)
7. RODC - Recommandation ODC	(2 MODALITES)
RECO - jus Recommandé	NREC - jus non recommandé

- a) Créer la table SAS correspondant aux données puis vérifier le bon déroulement de l'insertion ou l'importation de ces données (fichier Excel : jusorange.xls)
- b) Vérifier, avec un risque $\alpha = 5\%$, les hypothèses suivantes :
- le prix de vente moyen (PRIX) est-il normalement distribué selon le contenant ? selon la recommandation ODC ?
 - Peut-on considérer l'hocédasticité du prix de vente selon les 3 contenants ? Selon les recommandations ODC ?
- c) Peut-on considérer, avec un risque $\alpha = 5\%$, qu'il y a un effet du contenant et de la recommandation ODC sur le prix de vente ?
- d) Peut-on considérer, avec un risque $\alpha = 5\%$, qu'il y a un effet du contenant et de la recommandation ODC et de la teneur en vitamine sur le prix de vente ?

Exercice AV6 :

Une entreprise désire lancer sur le marché un nouveau produit. Ce produit pourrait être disponible en 3 couleurs : bleu, blanc et rouge. Quant à l'emballage du produit, l'entreprise considère les 3 options suivantes : sans emballage, emballage plastique et emballage carton. Afin de tester la réaction du marché, l'entreprise a mis en vente ce produit dans 25 magasins jugés homogènes. Dans chaque magasin, le produit n'était que d'une seule couleur et sous une seule forme d'emballage. Le prix de vente variait aussi d'un magasin à l'autre. Les résultats des ventes d'une semaine sont compilés et consignés dans la table SAS produits.

DICTIONNAIRE DES VARIABLES - Base de données : Produits.xls

1. NBUV - Y : Nombre d'unités vendues	(CONTINUE)
2. PVEN - X_1 : Prix de vente unitaire en euros	(CONTINUE)
3. CBLA - $X_2 = 1$ si la couleur est blanche, $X_2 = 0$ autrement	(CONTINUE)
4. CROU - $X_3 = 1$ si la couleur est rouge, $X_3 = 0$ autrement	(CONTINUE)
5. EMBA - $X_4 = 1$ s'il y a un emballage, $X_4 = 0$ sinon	(CONTINUE)
6. EPLA - $X_5 = 1$ produit emballé dans un sac en plastique, $X_5 = 0$ sinon	(CONTINUE)

- Créer la table SAS (importer les données du fichier Excel : produit.xls). Vérifier le bon déroulement de la phase de lecture des données et de la création de la table.
- Créer, à partir des variables binaires CBLA et CROU, une nouvelle variable nominale notée COUL 'Couleur' à trois modalités : Bleu, Blanc, Rouge.
- Créer, à partir des variables binaires EMBA et EPLA, une nouvelle variable nominale notée TEMB 'Type d'emballage' à trois modalités : Sans, plastique, carton.
- Etablir les statistiques sommaires des variables NBUV et PVEN, selon le type d'emballage TEMB et selon la couleur du produit.
- Etablir la répartition (tableau) croisée en effectif selon le type d'emballage TEMB et la couleur COUL.
- Etablir la répartition (tableau) croisée en nombre moyen d'unité vendues selon le type d'emballage TEMB et la couleur COUL.
- Peut-on considérer, avec un risque $\alpha = 5\%$, qu'il y a un effet du type d'emballage sur le nombre d'unités vendues NBUV ? De la couleur sur le nombre d'unités vendues NBUV ? Du type d'emballage et de la couleur sur le nombre d'unités vendues NBUV ?

Exercice AV7 :

Un fabricant de jus d'orange veut analyser les résultats d'une étude effectuée par l'Organisme de la Défense du Consommateur (ODC). L'étude portait sur 22 marques de jus d'orange selon sept caractéristiques. Le fabricant tente de déterminer quelles sont les principales caractéristiques qui influencent le prix de vente.

DICTIONNAIRE DES VARIABLES - Base de données : Jus-orange.xls

1. CONT - Contenant-commercialisation	(3 MODALITES)	
BVER - Bouteille verre	PACK - Pack carton	BPLA - Bouteille plastique
2. PRIX - Prix de vente moyen (€/litre)	(CONTINUE)	
3. PULP - pulpe d'orange (1 : contient / 0 : ne contient pas)	(CONTINUE)	
4. DILU - Dilution (%)	(CONTINUE)	
5. VITA - Vitamine C (%)	(CONTINUE)	
6. FOLA - Folacine (%)	(CONTINUE)	
7. RODC - Recommandation ODC	(2 MODALITES)	
RECO - jus Recommandé	NREC - jus non recommandé	

- Créer la table SAS correspondant aux données puis vérifier le bon déroulement de l'insertion ou l'importation de ces données (fichier Excel : jusorange.xls)
- Peut-on considérer, avec un risque $\alpha = 5\%$, qu'il y a un effet du pourcentage en vitamine C et de la recommandation ODC sur le prix de vente ?
- Etablir puis interpréter le modèle complet de l'effet du pourcentage en vitamine C, du contenant et de la recommandation ODC sur le prix de vente ?

Exercice AV8 :

Cette base de données a été extraite de la volumineuse base sur les constructeurs mondiaux d'automobiles. Elle est constituée de seize caractéristiques-variables mesurées sur quarante-cinq voitures-individus, berlines des trois principaux constructeurs français. Deux thèmes d'étude ont été retenus : le premier se réfère à la gamme et le second porte sur les caractéristiques des voitures. L'objectif poursuivi est de donner une description d'ensemble de ces voitures berlines françaises.

DICTIONNAIRE DES VARIABLES : Base de données : Berlines-FR.xls

1 . Type de carburant DIES - diesel	ESSE - essence			(2 MODALITES)
2 . Marque du constructeur CITR - Citroen	PEUG - Peugeot	RENA - Renault		(3 MODALITES)
3 . Type de véhicule CITA - Citadine	COMP - Compacte	FAMI - Familiale	RSPO - Routière-Sportive	
4 . Puissance Fiscale 4-5 - 4-5 CV	6-7 - 6-7 CV	8-9 - 8-9 CV	1011 - 10-11 CV	
5 . Nombre de portes 3POR - 3 portes	4POR - 4 portes	5POR - 5 portes		(3 MODALITES)
6 . PRIX - Prix du véhicule (Milliers €)				(CONTINUE)
7 . CYLI - Cylindrée				(CONTINUE)
8 . LONG - Longueur				(CONTINUE)
9 . RESE - Volume du réservoir - Autonomie				(CONTINUE)
10 . POID - Poids du véhicule				(CONTINUE)
11 . VITE - Vitesse moyenne				(CONTINUE)
12 . DARR - Démarrage à l'arrêt				(CONTINUE)
13 . REPR - Reprise				(CONTINUE)
14 . FREI - Freinage				(CONTINUE)
15 . BRUI - Bruit				(CONTINUE)
16 . VOLU - Volume du coffre				(CONTINUE)

Peut-on considérer, avec un risque $\alpha = 5\%$, qu'il y a un effet de la cylindrée, du type de carburant et de la marque du constructeur sur le prix du véhicule ?

Indications et résultats : ANOVA - ANCOVA (cf. programmes SAS & SPAD)

AV1 :
AV2 :
AV3 :
AV4 :
AV5 :
AV6 :
AV7 :
AV8 :