EXERCICES : GESTION de DONNÉES

G-1 : *ajout de variables, formule, assignation texte/numérique, tri, coloriage*

Ouvrir la feuille de données STATISTICA : *Agressivité.sta (3v par 50c).*
La feuille contient le sexe de l'individu et deux variables (AGR1 et AGR2). Ces variables mesurent, sur une échelle de 0 (doux) à 10 (violent), l'indice d'agressivité lorsque mis en face de deux situations susceptibles d'une réaction dans le cadre d'une simulation.

(a) Ajouter une nouvelle variable AGR = (AGR1 + AGR2)/2
(b) Ajouter une nouvelle variable CLAGR : agressivité de l'individu selon 4 niveaux :

AGR	CLAGR
(0.00, 0.25]	faible
(0.25, 0.50]	moyenne
(0.50, 0.75]	forte
(0.75, 1.00]	élevée

(c) Quel code numérique a été assigné à la variable CLAGR ?
Remplacer le code numérique par les 4 valeurs suivantes : 1 – 2 – 3 – 4.
(d) Trier les observations par SEXE et selon les valeurs croissantes de AGR.
(e) Colorer en jaune : les cellules de sexe féminin ayant une valeur de AGR > 0.76
(e) Changer la police de caractères par « Times New Roman » ; centrer toutes les colonnes.

G-2 : *importation, en-tête, format, tri, moyenne, écart type*

Ouvrir la feuille de données EXCEL : *Température.xls*
La feuille contient 4 variables : Date, Ville, Température (degrés F), Condition atmosphérique de 30 villes américaines en date du 22, 23 ou 24 Août 1998.

(a) Importer la feuille dans une feuille de données STATISTICA ; sauvegarder le fichier sous le nom *Température.sta*
(b) Ajouter un en-tête descriptif au fichier.
(c) Changer le format de la variable DATE pour le format : mois-année, (AUG-1998).
(d) Transformer la température en degrés Celsius.
(e) Trier les observations en ordre de température croissante. Quelle ville a la température la plus élevée ?
(f) Trouver la moyenne et l'écart type de la variable *température.*

G-3 : *création d'une feuille de données, saisie de données, assignation d'un nom pour les observations (cas), ajout/élimination variables/ observations statistiques de blocs*

(a) Créer une nouvelle feuille de données avec le nom *EX-G3.sta*
Saisir les 12 observations des 5 variables suivantes : ID1, ID2, X, Y, Z

ID1	ID2	X	Y	Z
1	AB	4	14	101.3
2	CD	7	-2	21.8
3	EF	9	7	122.1
4	GH	8	16	131.9
5	IJ	12	-22	99.2
6	KL	3	49	88.5
7	MN	5	-17	110.0
8	OP	7	6	87.4
9	QR	1	19	24.7
10	ST	2	-15	45.5
11	UV	1	-36	17.2
12	WX	4	0	99.6

le fichier *« nouveau »* contient toujours

10 variables (colonnes) par 10 observations. (lignes)

Il faudra ajouter 2 lignes

Utilisez le bouton *"Observations".*

(b) Ajouter un *"en tête de fichier"* intitulé « *fichier pour l'exercice G-3* ».
Sauvegarder la feuille avec la commande *"Enregistrez sous…"* .
(c) Nommer les observations avec la variable ID2. Employer "*Gestionnaire de Noms d'Observations*"
(d) Éliminer la colonne 10. Employez le bouton "*Variables*".
(e) Définir, dans les colonnes 6 à 9, les variables R (colonne 6), S (colonne 7) , T (colonne 8) U (colonne 9) . Les équations de définition sont :

$R = X + Y + Z$; $S = Log10(R^2)$; $T = R^2 + Rnd(100)$; $U = 100+RndNormal(10)$

Ouvrer la boîte de dialogue à l'aide d'un double clic sur le nom de la variable
Rnd (100) : données simulées selon loi uniforme sur (0,100)
RndNormal(10) : données simulées selon loi normale de moyenne 0 et d'écart type 10

Sauvegarder la feuille sous le même nom.
(f) Ajouter 5 rangées (cases) additionnelles après la rangée 12. Employez le bouton "Observations".
(g) Saisir le bloc formé par les colonnes ID2, X, Y, Z et les rangées 1, 2, 3, 4, 5 (le bloc est en noir) et copier ce bloc ("*Copier de Edition*") et collez le bloc ("*Coller de Edition*") dans les rangées 13 à 17 et les colonnes ID2, X, Y, Z.
(h) Déplacer les variables R, S, T, U après la colonne ID2. Faites une sauvegarde du fichier sous le nom *EXG3-A*. L'opérations suivante (i) sera exécutée sur ce fichier.
(i) Obtenir toutes les statistiques : MIN -MAX – MOYENNE - ÉCART TYPE des variables X, Y, Z avec la commande "*Statistiques de blocs*" du bouton *Statistiques* de la barre principale de STATISTICA. Sauvegardez le résultat sous le nom *EXG3-B*.

G-4 : *type de variables valeurs de date, saisie rapide des données par extrapolation*

(a) Créer une nouvelle feuille de données avec le nom *EX-G4.sta (5v par 100c)*.
Nommer les variables : ID JOUR MOIS AN MACHINE (dans cet ordre)
ID variable numérique avec une décimale.
JOUR, AN variables de type entier.
MOIS variable de type texte avec les codes numériques suivants :
janvier = 1 février = 2 mars = 3 avril = 4
MACHINE variable de type texte.
(a) Sauvegarder le fichier.
(b) Compléter la feuille de données selon les informations suivantes.

1. **ID commence à 10.0 et se poursuit avec 10.1, 10.2, …….**
2. **JOUR : valeur de 15 pour les observations 1 à 20**
   valeur de 16 pour les observations 21 à 40
   valeur de 17 pour les observations 41 à 60
   valeur de 18 pour les observations 61 à 80
   valeur de 19 pour les observations 81 à 100
3. **MOIS : janvier pour les cas 1 à 5**
   février pour les cas 6 à 10
   mars pour les cas 11 à 15
   avril pour les cas 16 à 20
   recommencer la structure précédente pour les cas 21 à 40,…., 81 à 100
4. **AN : commencer en 1901 suivie de 1902, 1903,…., 2000**
5. **MACHINE : cas 1 à 25 machine = A**
   cas 26 à 75 machine = B
   cas 76 à 100 machine = A

(c) Ajouter une sixième variable : DATE combinant les variables JOUR /MOIS /AN en une seule colonne dans un format d'affichage « date » de Statistica.

G - 5 : *ajout de variables, recodification, tri, statistiques de blocs, formules*

Ouvrir la feuille de données : ***Baseball.sta (7v par 40c)***

(a) Trier les données par année (YEAR) en ordre croissant et, pour chaque année, en ordre décroissant de la variable BA (moyenne au bâton).

(b) À quel numéro de cas correspond la meilleure moyenne au bâton en 1967?

(c) Introduire une nouvelle variable SCORE dans la feuille

SCORE = 1000* (RUNS + DP) / WALKS.

(d) Introduire une nouvelle variable TYPE de match dans la feuille :

Score < 1600	**TYPE = ennuyeux**
1600 ≤ Score < 1900	**TYPE = normal**
1900 ≤ Score	**TYPE = excitant**

Durant la période de 1965 à 1968, y a-t-il eu plus de matchs ennuyeux que de matchs excitants ?

EXERCICES : ANALYSE STATISTIQUE de BASE
Module Statistiques Élémentaires et fonctions Graphiques

Ouvrir la feuille de données : ***Expérience mémoire.sta (8v par 48c)***

Mettre tous les résultats de cet exercice dans une filière que l'on nommera : ***Expérience .mémoire.stw***

S -1 : *statistiques descriptives, décompositions , vérification normalité, test –t*

(a) variable STRESS - Calculer la moyenne l'écart type, le 5ième percentile, le 95^{ième} percentile

(b) variable STRESS - La variable est-elle normalement distribuée ?

(c) variable STRESS - Calculer la moyenne et la variance selon la variable SEXE.

(d) variable STRESS - Si on veut comparer les moyennes de STRESS pour les hommes et les femmes allez vous employer un test t (par groupe) ou un test non paramétrique?

S – 2 : *corrélations, tests de significatifs, p-level*

Ouvrir la feuille de données : ***Textile2.sta (5v par 27c)***

Il s'agit de données obtenues par l'exécution d'un plan expérimental avec 3 facteurs variant à 3 modalités.

(a) Trouver les coefficients de corrélations entre les variables LOAD, AMPLITUDE, LENGHT, LOG_CYCL. Les trois premières variables sont les facteurs contrôlés de l'expérience et LOG_CYCL représente la variable de réponse.

(b) Quels sont les coefficients qui sont statistiquement significatif au seuil de 0.05?

(c) Tracer le nuage de points de LENGTH et LOG_CYCL.

(d) Créer le graphique de réponse de LOG_CYCL.en fonction de LENGTH et LOAD. Quel est le comportement de la réponse lorsque LOAD décroît et que LENGTH croit?

(e) Tracer le graphique de normalité des variables CYCLES et LOG_CYCL. Les variables suivent –elles loi normale?

(f) Tracer le graphique Quantile-Quantile et le graphique Probabilité-Probabilité. Comparer les graphiques de la question (f) avec les graphiques de la question (e).

S – 3 : *test t pour échantillons indépendants, diagramme Boîte à Moustaches*

Ouvrir la feuille de données : ***Machine.sta (5v par 55c)***

(a) Comparer les 2 machines avec un test t pour 2 échantillons indépendants. Faites le test avec les variables des 2 premières colonnes seulement.
Les hypothèses de base pour exécuter d'un test t sont-elles vérifiées?

(b) Résumer le résultat du test t avec un diagramme boîte à moustaches.

(c) Les colonnes 3 et 4 contiennent les mêmes données que les colonnes 1 et 2 mais organisées différemment. Exécuter le test t de comparaison des machines en employant les colonnes 4 et 5.

Comparer avec le résultat obtenu avec le résultat obtenu en (a).

S – 4 : *décompositions, statistiques de groupes, ANOVA, comparaison a posteriori*

Ouvrir la feuille de données : ***Ventes GSC Inc.sta (13v par 130c)***

(a) Représenter avec un seul graphique, le volume (axe vertical) par période (axe horizontal) pour chaque région. Suggestion : employer les variables des colonnes 8 à 13.

(b) Employer la procédure *Décompositions & ANOVA à 1 facteur* du module *Statistiques Élémentaires* pour comparer le volume des ventes des 6 régions entre les années 1996 et 2000. Le volume moyen du volume des ventes est-il statistiquement différent au seuil de 0.05?

(c) Obtenir les différents graphiques qui permettent de visualiser les données et de vérifier si les données suivent une distribution normale.

(d) Obtenir le résultat du test de comparaison *Post Hoc (a posteriori)* HSD de Tukey permettant de comparer les régions 2 à 2.

S – 5 : *tableaux et tris croisé, test du khi deux, graphique d'interaction d'effectifs*

Ouvrir la feuille de données : ***Funmage.sta (2 v par 50 c)***

(a) Obtenir une table de fréquences pour la variable catégorie d'age.

(b) L'age est-elle liée à l'habitude de fumer? En d'autres termes existe t- il une relation entre la variable catégorie d'age et le fait d'être un fumeur? Suggestion : exécuter un test du khi deux.

(c) Obtenir un graphique d'interaction des fréquences.

S – 6 : *tests non paramétriques*

Ouvrir la feuille de données : ***Animaux.sta (2 v par 24 c)***

(a) Vérifier au moyen d'un graphique que la variable POIDS ne suit pas une distribution normale.

(b) Exécuter un test de Mann-Whitney pour comparer le groupe contrôle avec le groupe traitement. Fixer le seuil à 0.10.

(c) Représenter les données au moyen d'un diagramme *Boîte à Moustaches*.

S- 7 : *Test des signes, test de Wilcoxon*

Ouvrir la feuille de données : ***Accidents.sta (3v par 12 c)***

(a) Créer des noms d'observations à l'aide de la variable MOIS.

(b) Les données d'accidents de chaque mois constituent –elles 2 échantillons indépendants?

(c) Exécuter un test des signes sur les données.

(d) Exécuter un test de Wilcoxon pour les données appariées. La différence est-elle significative?

S – 8 : *création et exécution d'une analyse macro de session*

L'analyse macro s'appliquera sur une feuille de données dont les deux premières variables sont quantitatives.

Ouvrir une feuille de données, par exemple ***Diamètres.sta (3v par 100c)***

(a) Créer une analyse macro de session qui fera la séquence des opérations suivantes sur chacune des 2 premières variables de la feuille. Nommer la macro *MACROTEST.*

1. **le calcul de la moyenne et de l'écart type;**
2. **l'histogramme;**
3. **le diagramme Boîtes à Moustaches;**
4. **le calcul de la matrice de corrélation;**
5. **le test t de comparaison des moyennes en considérant les deux variables comme deux échantillons dépendants (appariés).**

Enregistrer la macro sous le nom de *MACROTEST.*

(b) Exécuter *MACROTEST sur la feuille de données IRIS.sta* (5v par 150c)

S-9 : *distributions de probabilité : z (Normale) t (Student) F (Fisher)*

<u>remarque</u> : repose sur une connaissance des lois de probabilités

Aller à la fonction « *Calculateur* » du module *STATISTIQUES ÉLÉMENTAIRES*. Compléter le tableau ici-bas.

DISTRIBUTION	PARAMÈTRES DE LA DISTRIBUTION	PROBABILITÉ (p)	VALEUR DU PERCENTILE
z (Normal)	**moyenne = 100** **écart type = 10**	**0.85**	**X = ?**
z (Normal)	**moyenne = 100** **écart type = 10**	**p = ?**	**X = 85**
t (Student)	**dl = 5**	**0.95**	**t = ?**
t (Student)	**dl = 10**	**p = ?**	**t = 2.00**
F (Fisher)	**df_1 = 3** **df_2 = 7**	**p = 0.80**	**F = ?**
F (Fisher)	**df_1 = 5** **df_2 = 15**	**p = ?**	**F = 1.55**