

Ce TD a pour objectif de se familiariser avec les principales commandes (du logiciel SAS) en lien avec la statistique descriptive.

L'intégralité des questions porte sur le fichier "World\_cup.csv". Celui-ci contient des données sur l'ensemble des Coupes du monde de football comme les vainqueurs, les nombres de nations participantes ou encore les nombres de buts marqués.

**Questions :**

1. Créer une bibliothèque et importer le fichier "World\_cup.csv". (**LIBNAME** + **IMPORT**)
2. Afficher les données. (**PRINT**)
3. Combien y a-t-il de champs (variables) et d'observations (enregistrements) dans le fichier ? (**CONTENTS** ou **SUMMARY**)
4. Procédure **MEANS** et **UNIVARIATE** :
  - a. Avec la procédure **MEANS**, afficher les principales statistiques des variables numériques en fonction de la variable 'Winner'. D'abord avec l'option **CLASS** puis en utilisant l'option **BY** (nécessité de trier avant).
  - b. Utiliser la procédure **UNIVARIATE** sur la variable 'Attendance' et constater les différences d'affichage avec **MEANS**.
5. Quels pays (variable 'Country') ont accueilli le plus de CDM ? (**FREQ** avec **ORDER**)
6. Quelles sont les équipes ayant gagné à domicile ? (**FREQ** avec **WHERE**)
7. Diagramme à barres et camembert :
  - a. Afficher l'histogramme, trié du plus grand au plus petit, du nombre de CDM gagnées par pays. Quelles sont les 3 équipes à avoir remporté le plus de CDM ? (**SGPLOT** ou **CHART** avec **VBAR**)
  - b. Même question avec un camembert. (**CHART** avec **PIE** ou aidez-vous du menu « tâches et utilitaires » de SAS : Tâches / Graphique / Diagramme Circulaire)
8. Variable 'GoalsScored' :
  - a. Exporter dans une table annexe, le nombre de buts moyens selon le nombre de matchs joués (variable 'MatchesPlayed') ainsi que le nombre d'observations correspondantes. (**UNIVARIATE** avec **OUTPUT**)

	MatchesPlayed	Nobs	AvgGoal
1	17	1	70
2	18	2	77
3	22	1	88
4	26	1	140
5	32	3	91
6	35	1	126
7	38	2	99.5
8	52	4	133.5
9	64	6	160.66666667

- b. Tracer le nuage de points du nombre de buts moyens en fonction du nombre de matchs joués. (**SGPLOT** avec **SCATTER**)

9. Créer la variable 'GoalsPerMatch' : nombre de buts moyens par match pour chaque CDM. (DATA) Puis calculer le nombre moyen de but par match sur l'ensemble des CDM. (UNIVARIATE ou MEANS, les deux avec WEIGHT)
10. Créer la variable 'AttendancePerMatch' : nombre de spectateurs moyens par match pour chaque CDM. (DATA)  
Tracer l'histogramme de cette variable (SGPLOT avec HISTOGRAM) puis effectuer un test de normalité (UNIVARIATE avec NORMALTEST). Qu'en déduisez-vous ?
11. Tracer le graphique à bulle 'GoalsPerMatch' en fonction de 'AttendancePerMatch'. Identifier la taille des bulles selon les années (SGPLOT avec BUBBLE).
12. Recoder la variable 'AttendancePerMatch' en une variable qualitative comportant 3 modalités d'effectif égal. Pour cela :
  - a. Déterminer les terciles de la variable 'AttendancePerMatch' (UNIVARIATE avec PCTLPRE et PCTLPTS)
  - b. Utiliser les terciles déterminés précédemment pour créer un format de variable (FORMAT) puis appliquer ce format à la variable 'AttendancePerMatch' afin de créer la variable qualitative 'AttendanceModa' (PUT).
  - c. Vérifier les effectifs et fréquences des modalités de la variable qualitative ainsi créée (FREQ). Vous devriez obtenir le tableau suivant.

AttendanceModa	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
1 - Faible affluence	7	33.33	7	33.33
2 - Moyenne affluence	7	33.33	14	66.67
3 - Forte affluence	7	33.33	21	100.00

13. Tracer le boxplot conditionnel 'GoalsPerMatch' en fonction des modalités de 'AttendanceModa'. (SGPLOT avec VBOX et CATEGORY).
14. Centrer et réduire la variable 'GoalsMatch' (STDIZE). Donner la moyenne et l'écart type de la variable centrée réduite.