TD07 - STATISTIQUE DESCRIPTIVE - Calcul des indicateurs statistiques (1)

Les indicateurs synthétiques sont des descripteurs numériques qui condensent dans une valeur unique une certaine caractéristique de toute une distribution de valeurs.
A l'aide des indicateurs synthétiques sont évalués :

I. la tendance centrale,
II. la variabilité
III. La forme de la distribution de fréquence.
Les indicateurs de la tendance centrale sont des valeurs typiques, représentatives, qui décrivent la distribution toute entière;
Les indicateurs de la variabilité sont des valeurs qui décrivent la caractéristique de la dispersion de la distribution.
Les indicateurs de la forme de la distribution sont des valeurs qui se rapportent à la forme de la courbe de représentation graphique de la distribution, en comparaison avec une courbe normale (oblique, plat)

I. Les indicateurs de la tendance centrale
- valeur moyenne*;*
- valeur médiane ;
- valeur dominante (module) .

I. Les indicateurs de la variabilité :
- L'amplitude (rang)
- Dispersion (variante) (variance)
- Déviation standard (déviation standard)
- Coefficient de variation
- Intervalle interquartilique

I. Les indicateurs de la forme de la distribution
- Indice d'asymétrie (coefficient.asymétrie)
- Voûte (kurtosis)

Objectifs :
1. Le calcul à l'aide des **formules de calcul** de tous les indicateurs synthétiques
2. **L'interprétation des** indicateurs de centralité, dispersion et localisation
3. L'obtention des indicateurs synthétiques suite à l'appellation des fonctions statistiques implémentées en Excel.
4. L'obtention des indicateurs synthétiques suite à l'appellation de l'option Statistiques descriptives de Analyses des données.

Problème 1

Dans le tableau ci-dessous sont présentées les valeurs des concentrations de cholestérol dans le sang, mesurées en mg/dl, pour un échantillon de 50 patients.

250	200	240	210	180	160	210	170	240	140
160	220	150	260	150	180	170	140	180	190
145	220	150	170	210	220	210	230	140	220
230	180	250	230	230	240	170	260	240	200
190	160	180	250	180	160	190	220	260	200

- Calculez la moyenne, la médiane et la valeur modale :
  1. A base des formules de calcul,
  2. A l'aide des fonctions comprises dans Excel (F_x)

3. A l'aide de l'option Statistiques descriptives de Analyses de données

- Calculez l'amplitude, la variance, la déviation standard, le coefficient de variation :
  1. Se basant sur les formules de calcul,
  2. A l'aide des fonctions contenues en Excel (F_x)
  3. A l'aide de l'option Statistiques descriptive de Analyses de données

- Calculez l'indice d'asymétrie et de voûte à l'aide des fonctions introduites en Excel (Fx)
- Déterminez le quartiles Q1, Q2 et Q3 qui divisent le set de données en quatre sections à nombre égal de valeurs.
- Tracez le diagramme BoxPlot

Indications :

- On introduira dans une feuille de calcul EXCEL les 50 valeurs sur une colonne
  a. **Le calcul à base des formules de calcul ou de la définition** :
  - Pour le calcul de la **moyenne arithmétique**, on va additionner à l'aide de la fonction SOMME ou AutoSomme les 50 valeurs.
  - On introduit la formule de calcul pour la division de la somme 50.
  - Si la série d'observations est donnée sous forme de répartition de fréquence, le calcul de la valeur moyenne se fait tenant compte de la fréquence d'apparition de chaque valeur ($m = \frac{\sum x_i n_i}{\sum n_i}$)
  - **La médiane de** l'onde de la série statistique ordonnée est la valeur qui divise la ligne ordonnée des valeurs de la variable en deux parties, chaque partie contenant le même nombre de valeurs. On note avec Me. Pour trouver cette valeur, on procède conformément à la définition : on écrit les valeurs sur une colonne (en Excel), on ordonne de manière croissante, on compte le nombre des valeurs (50), dans notre cas, nombre paire, la 25-ème et la 26-ème valeur sont 200 respectivement 200, donc, dans ce cas, Me =200 (moyenne arithmétique de la 25-ème respectivement la 26-ème valeur de la ligne d'observations). Si le nombre d'observations avait été impaire, par exemple 51, la valeur de la médiane aurait été la 26-ème valeur des observations ordonnées de manière croissante.
  - Pour calculer la valeur modale il est nécessaire de réaliser la répartition de fréquence, la **valeur modale** est la valeur de la caractéristique à plus haute fréquence d'apparition.
  - Pour le calcul de l'**amplitude** il est nécessaire de trouver la valeur maximale et minimale. On peut les trouver en ordonnant de manière croissante la série de valeurs. L'amplitude est la différence entre la valeur maximale et la valeur minimale.
  - Pour le calcul de la **variance** à l'aide de la formule $S^2 = \frac{\sum(xi-m)^2}{N-1}$, il est nécessaire de calculer m (moyenne arithmétique), puis il est nécessaire de calculer chaque valeur séparement : (A2-198,7)^2 ; (A3-198,7)^2 … on peut tirer les cellules pour affecter la formule aux 50 valeurs. Ensuite à l'aide de SOMME, on additionne les valeurs obtenues que l'on divise par N-1 (=SOMME(B2:B51)/49).
  - **La déviation standard** est caculée comme racine carrée de la variance.
  - **Le coefficient de variance** est calculé à l'aide de la formule:

    $CV\ [\%] = \frac{deviation\ standard}{moyenne\ arithmétique} \cdot 100$ déviation standard/moyenne arithmétique

b. **Pour le calcul des exigences à l'aide des fonctions intégrées en Excel on procédera ainsi** :

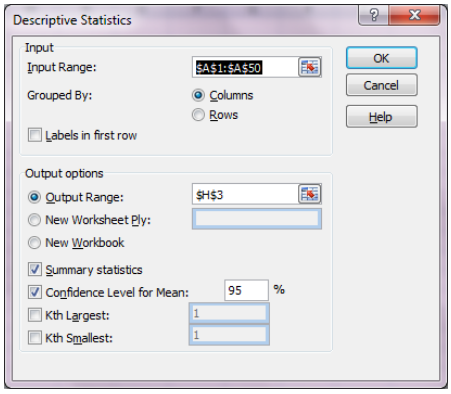
- Sur une colonne on va écrire, comme étiquettes, les exigences du problème.
- En face, on sélectionne la cellule voisine et on appelle les fonctions qui calculent les exigences : **AVERAGEA, MEDIANE, MODE, VAR, STDEVA, QUARTILE, COEFFICIENT.ASYMETRIE, KURTOSIS**, etc.,
- Toutes ces fonctions ont comme argument le domaine de la fiche de calcul qui contient le tableau avec les observations (tableau d'évidence primaire ou tableau d'effectifs)

moyenne	=AVERAGEA(A1:A50)	198,7
médiane	=MEDIANE(A1:A50)	200
module	=MODE(A1:A50)	180
variance	=VAR(A1:A50)	1319,193878
déviation standard	=STDEVA(A1:A50)	36,32070866
Coefficient de variation	=(B5/B1) % (Déviation standard/moyenne)	0,181603543
Asymétrie	=COEFFICIENT.ASYMETRIE (A1:A50)	0,034976528
Voûte	=KURTOSIS(A1:A50)	-1,165363744
Q1	=QUARTILE(A1:A50,1)	170
Q2	=QUARTILE(A1:A50,2)	200
Q3	=QUARTILE(A1:A50,3)	230

c. Tous les indicateurs peuvent être calculés par simple appellation de la fonction Statistiques Descriptive d'Analyse de données (menu Données).

L'Installation de l'option Analyse de données a été présentée dans le travail précédent.

De la fenêtre de dialogue qui apparait, on sélectionne **Statistiques Descriptives**. Clic sur **OK**.



Voilà une description des champs de la fenêtre de dialogue ci-jointe.

- **Input Range (Plage de données)**: Introduisez les références du domaine où se trouve la variable pour laquelle on désire calculer les paramètres de la statistique descriptive. Pour l'introduction, cliquez dans le champ **Plage de données**, ensuite click dans la cellule de la fiche de calcul où se trouvent les données pour lesquelles on fait le calcul.
- **Grouped by (Grouper par)**: on va sélectionner **Colonnes** si chaque variable est introduite dans une colonne où **Lignes** si chaque variable est introduite dans une ligne. Dans notre cas, nous allons cocher **Colonnes**.
- **Labels in first row**. L'en-tête de colonne ou la ligne peut être sélectionné ou il peut manquer. Si l'on sélectionne l'en-tête de colonne aussi, alors dans la page de résultats apparaitra cet en-tête, c'est-à-dire, le nom de la variable. Dans ce cas, il faut cocher **Labels in first row (Intitulés)**. Si on ne coche pas la fonction va retourner l'erreur : "Input range contents non numeric data", parce que on considère l'en-tête de colonne comme étant l'une des valeurs de la variable. Dans

le cas où on ne sélectionne pas l'en-tête de colonne, il faudrait ne pas cocher Labels **in first row**. Si on coche **Labels in first row** alors la première valeur de la variable va être prise comme en-tête de colonne et les résultats seront erronnés. Dans notre cas, on sélectionne **labels in first row**.

- **Les options Output (Plage de sortie)** se rapportent à l'endroit de l'emplacement du tableau de fréquence. On va sélectionner l'option **New Worksheet Ply**. Le tableau de fréquence sera affiché à une nouvelle page dont le nom doit être introduit dans le champ **New Worksheet Ply**. **Output Range** est, dans le cas où on désire que le résultat soit affiché sur la même page que le tableau, commençant par une certaine cellule qui doit être introduite dans le champ **Output Range**. **New Woorkbook** est coché dans le cas où on désire que le résultat soit affiché dans un autre fichier.
  Au moins une de ces options doit être sélectionnée.
- On coche **Summary statistic (Rapport détaillé)** pour calculer les principaux paramètres statistiques**.**
- On coche **Confidence Level for Mean (niveau de confiance pour la moyenne)** pour calculer aussi l'intervalle de confidence pour la moyenne arithmétique. Le niveau de confidence pour la moyenne est 95% ; on peut le changer avec un autre nombre de 1 à 99,9.
- **Kième Maximale** retourne le nombre k-ème plus petit que le nombre le plus grand. K est une constante qui est introduite dans le champ correspondant. Cliquez **Kième Maximale**. Introduisez le nombre 2 dans le champ **Kième Maximale**.
- **Kième Minimale** retourne le nombre k-ème plus grand que le nombre le plus petit. K est une constante qu'on introduit dans le champ correspondant. Il peut être différent de K du point 10. Cliquez **Kième Minimale**. Introduisez le nombre 2 dans le champ **Kième Minimale**.
- Cliquez **OK**.

Les résultats apparaissent dans le tableau sous la forme:

moyenne	Moyenne	198,7
Erreur standard de la moyenne	Erreur standard	5,136524
médiane	Médiane	200
module	Mode	180
déviation standard	Deviation standard	36,32071
variance	Variance	1319,194
voûte	Kurtosis	-1,16536
asymétrie	Coefficient.asymétrie	0,034977
amplitude	Rang	120
	Minimum	140
	Maximum	260
	Somme	9935
	Nombre d'échantillon	50
	Confidence Level (95,0%)	10,32223

Interprétation des résultats

- **Mean** – Moyenne arithmétique. On peut calculer aussi à l'aide de la fonction **AVERAGEA (ou MOYENNE)**. Si la variable est normalement distribuée, la moyenne nous indique le milieu de l'intervalle entre le minimum et le maximum (intervalle de distribution des données). Toujours dans le cas de la distribution normale, autour de la moyenne (plus

précisément dans l'intervalle moyenne-déviation standard, moyenne+déviation standard) se trouve la majorité des données.

- **Standard Error** – Erreur standard. L'erreur standard est impliquée dans le calcul de l'intervalle de confidence de 95% autour de la moyenne (seulement pour une variable à distribution normale), elle est impliquée aussi dans l'inférence statistique.
- **Médiane** – La médiane est une valeur de la série ainsi que la moitié d'observations a des valeurs plus petites (ou égales) et l'autre moitié a des valeurs plus grandes (ou égales). On peut calculer aussi avec la fonction **MEDIANE**. Dans le cas de la distribution normale, la moyenne et la médiane sont égales. Ainsi que, la médiane et la moyenne arithmétique deviennent des indicateurs pour la distribution normale, plus elles ont les valeurs plus proches plus il est probable que la variable soit distribuée normalement. Le terme "plus proches" est estimé en fonction de la dimension de l'erreur standard.
- **Mode** – Le module représente la valeur qui a la plus haute fréquence de la série. Dans le cas du module, apparait une situation dans laquelle la série n'a pas de module, c'est-à-dire toutes les valeurs apparaissent une seule fois. A ce moment, on affiche la valeur #N/A. Une autre situation possible : la série soit bimodale ou trimodale. Alors on n'affiche que la première valeur dans l'ordre de leur apparition dans la série. Dans ce cas, pour déterminer toutes les valeurs du module, on peut faire un tableau de fréquence. On peut calculer aussi avec la fonction MODE. Le module est utile dans le cas d'une variable qualitative ordonnée, mais aussi dans le cas d'autres types de variables, par exemple, dans le cas de la variable continue à distribution normale, il est probable que le module ait une valeur proche de la moyenne.
- **Déviation Standard** – La déviation standard peut être calculée aussi avec STDEV ou pour la déviation standard de population STDEVA. L'écart standard nous montre quelle est l'écart carré moyen de la moyenne arithmétique des valeurs de la variable. S'il a une valeur réduite, les données varient un peu autour de la moyenne. Dans le cas où la distribution est représentée par la courbe de Gauss, les paramètres qui mesurent la tendance centrale (moyenne arithmétique, la médiane, la valeur modale et la valeur centrale) ont les mêmes valeurs. Dans ce cas, (de la distribution normale) ont lieu les répartitions des données suivantes :
  - L'intervalle $\bar{X} \pm 1 \cdot s$contient environ 68.3 % observations
  - L'intervalle $\bar{X} \pm 2 \cdot s$contient environ 95.5 % observations
  - L'intervalle $\bar{X} \pm 3 \cdot s$contient environ 99.7 % observations

- **Variance** – La variation peut être calculée aussi avec VAR ou pour la variation de population VAR
- **Voute** – L'excès ou la voûte mesure la hauteur de l'aplatissement ou de la voûte d'une distribution par comparaison à une distribution normale.

L'excès est zéro pour une série de données ayant une distribution normale, il est positif pour une série de données ayant la traine plus haute que celle d'une distribution normale (à moyenne $\overline{X}$ et la variation S^2) et il est négatif pour une série de données dont la traine est plus basse que celle d'une distribution normale. On peut calculer aussi avec la fonction KURTOSIS.

- **COEFFICIENT.ASYMETRIE** – L'asymétrie mesure l'écart de l'aspect symétrique et la direction de l'asymétrie (positive ou négative) par rapport à la courbe normale.

L'asymétrie est 0 pour une série de données ayant une distribution normale, elle est négative pour une série de données asymétrique vers la gauche (la série a plusieurs valeurs plus réduites), elle est positive pour une série de données asymétrique vers la droite (la série a plusieurs valeurs plus grandes). On peut calculer aussi avec la fonction SKEW.

- **Rang** – L'amplitude est la différence Maximum-Minimum de la série de données.
- **Minimum** – Le minimum est la plus petite valeur de la série. On peut le calculer aussi avec la fonction MIN.

- **Maximum** – Le Maximum est la plus grande valeur de la série. On peut le calculer aussi avec la fonction MAX
- **Somme** – La somme ou le Total des valeurs de la série peut être calculé aussi avec la fonction SUM.
- **NB** – Le nombre d'observations n=20 ou le volume de l'échantillon. On peut le calculer aussi avec la fonction NB.
- **Les Quartiles** et les **percentiles** sont semblables à la médiane. Ainsi, le premier quartile este une valeur ayant la propriété que 25% de données de la série sont plus petites ou égales à elle, et 75% plus grandes ou égales au premier quartile. Le deuxième quartile est représenté par la médiane. Le troisième quartile est une valeur ayant la propriété que 75% de données de la série sont plus petites ou égales à elle, et 25% plus grandes ou égales au troisième quartile.

  Le percentile d'ordre *a* est une valeur ayant la propriété suivante : une proportion égale à *a* de données sont plus petites ou égales, tandis que les autres sont plus grandes.

- **CV=STDEVA/AVERAGEA** – Le coefficient de variation : on peut utiliser les règles empiriques suivantes pour l'interprétation :
  - si CV est en-dessous de 10%, la population peut être considérée homogène;
  - si CV est entre 10%-20%, la population peut être considérée relativement homogène;
  - si CV est entre 20%-30%, la population peut être considérée relativement hétérogène;
  - si CV dépasse 30% , la population peut être considérée hétérogène.

Pour tracer le diagramme boxplot on recommande l'utilisation du programme online de l'adresse: http://www.alcula.com/calculators/statistics/box-plot/

Problème 2.

On réalise une étude sur un lot formé de 50 patients. On recueille des données sur les paramètres médicaux suivants : tension artérielle diastolique (TAD) (mmHg), la tension artérielle systolique (TAS) (mmHg), l'âge (jours), la taille (cm), le poids (grammes), le sexe. Les données sont présentées dans le fichier table_TD07-2. Sauvez le fichier dans votre dossier et réalisez les transformations statistiques suivantes dans ce fichier.

Exigences

On poursuit le calcul de :

- Indicateurs de la tendance centrale (moyenne arithmétique, la médiane, le module, la moyenne géométrique, la moyenne harmonique, la valeur centrale),
- Dispersion (amplitude, moyenne de la déviation, variation, écart standard, coefficient de variation, erreur standard, asymétrie, voûte) et
- Localisation (premier quartile (minimum), le deuxième quartile, le troisième quartile (médiane), le quatrième quartile, le cinquième quartile (maximum), pour les variables quantitatives.
- Introduisez le tableau ci-dessous sur la page Feuille1 du fichier Tableau Excel SD. Calculez les indicateurs requis dans ce tableau. Pour calculer les indicateurs, utilisez les fonctions statistiques de Excel ou des formules – voir instructions. Interprétez du point de vue statistique les résultats obtenus (se basant sur la moyenne arithmétique, la médiane, la voûte, l'asymétrie, l'erreur standard, la déviation standard, le minimum et le maximum, appréciez si la distribution des variables suivent la distribution normale de Gauss). (Les données se trouvent dans la feuille de calcul dans le domaine D2:H51)

Indicateurs de la tendance centrale	Moyenne arithmétique		=AVERAGEA(D2:D51)	=AVERAGE(E2:E51)	=AVERAGE(F2:F51)	=AVERAGE(G2:G51)	=AVERAGE(H2:H51)
	médiane		=MEDIANE(D2:D51)	=MEDIAN(E2:E51)	=MEDIAN(F2:F51)	=MEDIAN(G2:G51)	=MEDIAN(H2:H51)
	modul		=MODE(D2:D51)	=MODE.SNGL(E2:E51)	=MODE.SNGL(F2:F51)	=MODE.SNGL(G2:G51)	=MODE.SNGL(H2:H51)
	Moyenne géométrique		=GEOMEAN(D2:D51)	=GEOMEAN(E2:E51)	=GEOMEAN(F2:F51)	=GEOMEAN(G2:G51)	=GEOMEAN(H2:H51)
	Moyenne harmonique		=HARMEAN(D2:D51)	=HARMEAN(E2:E51)	=HARMEAN(F2:F51)	=HARMEAN(G2:G51)	=HARMEAN(H2:H51)
Indicateurs de la dispersion	amplitude		=MAX(D2:D51)-MIN(D2:D51)	=MAX(E2:E51)-MIN(E2:E51)	=MAX(F2:F51)-MIN(F2:F51)	=MAX(G2:G51)-MIN(G2:G51)	=MAX(H2:H51)-MIN(H2:H51)
	Déviation standard		=STDEV.S(D2:D51)	=STDEV.S(E2:E51)	=STDEV.S(F2:F51)	=STDEV.S(G2:G51)	=STDEV.S(H2:H51)
	Variance		=VAR.S(D2:D51)	=VAR.S(E2:E51)	=VAR.S(F2:F51)	=VAR.S(G2:G51)	=VAR.S(H2:H51)
	Coefficient de variation		=D58/D52	=E58/E52	=F58/F52	=G58/G52	=H58/H52
	Intervalles	m-s	=D52-D58	=E52-E58	=F52-F58	=G52-G58	=H52-H58
		m+s	=D52+D58	=E52+E58	=F52+F58	=G52+G58	=H52+H58
		m-2s	=D52-2*D58	=E52-2*E58	=F52-2*F58	=G52-2*G58	=H52-2*H58
		m+2s	=D52+2*D58	=E52+2*E58	=F52+2*F58	=G52+2*G58	=H52+2*H58
		m-3s	=D52-3*D58	=E52-3*E58	=F52-3*F58	=G52-3*G58	=H52-3*H58
		m+3s	=D52+3*D58	=E52+3*E58	=F52+3*F58	=G52+3*G58	=H52+3*H58
	Erreur standard		=D58/SQRT(50)	=E58/SQRT(50)	=F58/SQRT(50)	=G58/SQRT(50)	=H58/SQRT(50)
Indicateurs de la	asymétrie		=SKEW(D2:D51)	=SKEW(E2:E51)	=SKEW(F2:F51)	=SKEW(G2:G51)	=SKEW(H2:H51)
	voûte		=KURT(D2:D51)	=KURT(E2:E51)	=KURT(F2:F51)	=KURT(G2:G51)	=KURT(H2:H51)

forme et de localisation	minimum	=MIN(D2:D51)	=MIN(E2:E51)	=MIN(F2:F51)	=MIN(G2:G51)	=MIN(H2:H51)
	Q1	=QUARTILE(D2:D51;1)	=QUARTILE(E2:E51;1)	=QUARTILE(F2:F51;1)	=QUARTILE(G2:G51;1)	=QUARTILE(H2:H51;1)
	Q2	=QUARTILE(D2:D51;2)	=QUARTILE(E2:E51;2)	=QUARTILE(F2:F51;2)	=QUARTILE(G2:G51;2)	=QUARTILE(H2:H51;2)
	Q3	=QUARTILE(D2:D51;3)	=QUARTILE(E2:E51;3)	=QUARTILE(F2:F51;3)	=QUARTILE(G2:G51;3)	=QUARTILE(H2:H51;3)
	maximum	=MAX(D2:D51)	=MAX(E2:E51)	=MAX(F2:F51)	=MAX(G2:G51)	=MAX(H2:H51)

Inscrivez le niveau d'homogénéité de la variable dans la cellule correspondante au tableau.

	TAS	TAD	Age	Taille	Poids
Niveau d'homogénéité					

On peut utiliser les règles empiriques suivantes pour l'interprétation statistique du coefficient de variation CV :

- si CV est en-dessous de 10%, la population peut être considérée homogène ;
- si CV est entre 10%-20%, la population peut être considérée relativement homogène ;
- si CV est entre 20%-30%, la population peut être considérée relativement hétérogène ;
- si CV dépasse 30%, la population peut être considérée hétérogène.

1. Calculez les statistiques descriptives (en utilisant Analyse des données, Statistiques Descriptives) pour les 5 variables et interprétez les résultats obtenus.

2. Formatez les résultats obtenus : réduisez le nombre de décimales, effacez les colonnes aux dénominations qui sont en plus, introduisez la dénomination des indicateurs en français (mettre les formules du tableau en français).

Diagramme BOXPLOT

Appellant le lien http://www.alcula.com/calculators/statistics/box-plot/, pour les observations de la variable Age, on obtient :