

Mathematical statsites, Wackerly, Mendenhall, Schaeffer (*théorie v. visionneuse*)

Traduction google A-F des énoncés, corrections sommaires éventuels. Développements des résolutions, Master Sc, non mathématicien, reprends les statistiques dans un but d'approfondissement de connaissance dans le domaine.

méthodologie : J'ai tenté de résoudre les problèmes en faisant abstraction de connaissances déjà acquises en statistique, et en suivant les apports théoriques successifs du manuel dans l'ordre des chapitres. La formule de la loi normale de distribution n'a été intentionnellement employée qu'à partir de l'exercice 1.13 (via Excel) (A partir de l'exercice 1.6 énoncés E/F en 2 colonnes)

1.1 For each of the following situations, identify the population of interest, the inferential objective, and how you might go about collecting a sample.

a The National Highway Safety Council wants to estimate the proportion of automobile tires with unsafe tread among all tires manufactured by a specific company during the current production year.

b A political scientist wants to determine whether a majority of adult residents of a state favor a unicameral legislature.

c A medical scientist wants to estimate the average length of time until the recurrence of a certain disease.

d An electrical engineer wants to determine whether the average length of life of transistors of a certain type is greater than 500 hours.

e A university researcher wants to estimate the proportion of U.S. citizens from "Generation X" who are interested in starting their own businesses.

f For more than a century, normal body temperature for humans has been accepted to be 98.6° Fahrenheit. Is it really? Researchers want to estimate the average temperature of healthy adults in the United States.

g A city engineer wants to estimate the average weekly water consumption for single-family dwelling units in the city.

1.1 Pour chacune des situations suivantes, identifiez la population d'intérêt, l'objectif inférentiel, et comment procéder pour prélever un échantillon.

a Le Conseil national de la sécurité routière souhaite estimer la proportion de pneus automobiles présentant une bande de roulement non sécuritaire parmi tous les pneus fabriqués par une entreprise spécifique au cours de l'année de production actuelle.

b Un politologue souhaite déterminer si une majorité de résidents adultes d'un État sont favorables à une législature monocamérale.

c Un chercheur en sciences médicales souhaite estimer la durée moyenne avant la récurrence d'une certaine maladie.

d Un ingénieur électricien souhaite déterminer si la durée de vie moyenne des transistors d'un certain type est supérieure à 500 heures.

e Un chercheur universitaire souhaite estimer la proportion de citoyens américains de la « génération X » intéressés par la création de leur propre entreprise.

f Depuis plus d'un siècle, la température corporelle normale chez l'être humain est considérée comme étant de 37 ° C (98,6 ° F). Est-ce vraiment le cas ? Des chercheurs souhaitent estimer la température moyenne des adultes en bonne santé aux États-Unis.

g Un ingénieur municipal souhaite estimer la consommation d'eau hebdomadaire moyenne des logements unifamiliaux de la ville.

Solution

	Population d'intérêt	Objectif inférentiel	procédé d'échantillonnage
a	pneus fabriqués par l'entreprise X	fraction de bandes de roulement non sécuritaires	méthode de sélection aléatoire pour des tests techniques
b	électeurs résidents d'un Etat	choix d'une législature monicamérale	méthodes d'enquête publique de préférence , échantions représentatives
c	population présentant la pathologie cible	durée moyenne précédant une récurrence	méthode d'enquête ciblée de potentiels malades
d	transistors de type cible	durée de vie moyenne supérieure à 500 h	méthode de sélection aléatoire et tests de laboratoires
e	population de la génération X	intérêt par la création d'entreprise propre	méthode d'enquête publique
f	température corporelle de référence chez l'homme adulte	réévaluation de la validité de la température de référence (37°C)	méthode de sélection de candidat pour une estimation en laboratoire
g	logement unifamiliaux d'une ville	estimation de la consommation d'eau hebdomadaire	méthode de sélection aléatoire et représentative de la population , relevé de compteurs

Dans tous les cas de figures, le paramètre clé du procédé d'échantillonnage reste la dimension de l'échantillon représentative de la population, c'est-à-dire la limite au-delà de laquelle aucune précision n'est apportée. Des tests préalables permettent de définir cette limite, dans le cas contraire on emploie des outils statistiques.

1.2. Are some cities more windy than others? Does Chicago deserve to be nicknamed “The Windy City”? Given below are the average wind speeds (in miles per hour) for 45 selected U.S. cities:

a Construct a relative frequency histogram for these data. (Choose the class boundaries without including the value 35.1 in the range of values.)

b The value 35.1 was recorded at Mt. Washington, New Hampshire. Does the geography of that city explain the magnitude of its average wind speed?

c The average wind speed for Chicago is 10.3 miles per hour. What percentage of the cities have average wind speeds in excess of Chicago’s? d Do you think that Chicago is unusually windy?

8.9	12.4	8.6	11.3	9.2	8.8	35.1	6.2	7.0
7.1	11.8	10.7	7.6	9.1	9.2	8.2	9.0	8.7
9.1	10.9	10.3	9.6	7.8	11.5	9.3	7.9	8.8
8.8	12.7	8.4	7.8	5.7	10.5	10.5	9.6	8.9
10.2	10.3	7.7	10.6	8.3	8.8	9.5	8.8	9.4

Source: The World Almanac and Book of Facts 2004

1.2 Certaines villes sont-elles plus venteuses que d’autres ? Chicago mérite-t-elle son surnom de ville "venteuse" ? Vous trouverez ci-dessous les vitesses moyennes du vent (en miles par heure) pour 45 villes américaines sélectionnées :

a) Construisez un histogramme de fréquences relatives pour ces données. (Choisissez les limites de classe.) sans inclure la valeur 35,1 dans la plage de valeurs.) b La valeur 35,1 a été enregistrée à Mount Washington, dans le New Hampshire. La géographie de cette ville peut-elle expliquer l’ampleur de sa vitesse moyenne du vent ? c La vitesse moyenne du vent à Chicago est de 10,3 miles par heure. Quel pourcentage des villes ont des vitesses de vent moyennes supérieures à celles de Chicago ? d Pensez-vous que Chicago soit particulièrement venteuse ?

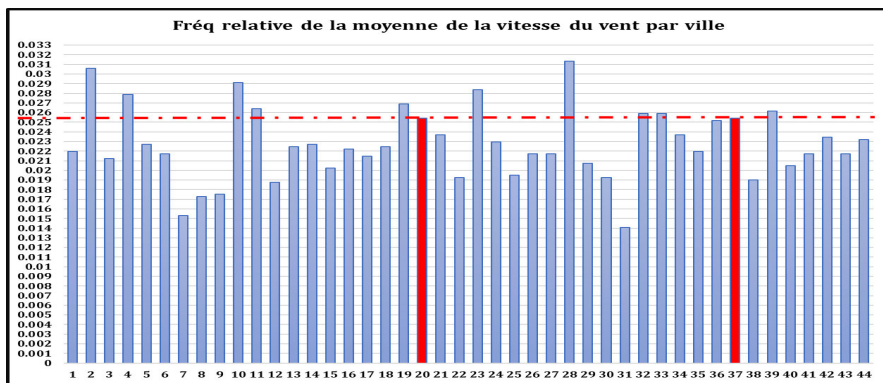
8.9	12.4	8.6	11.3	9.2	8.8	35.1	6.2	7.0
7.1	11.8	10.7	7.6	9.1	9.2	8.2	9.0	8.7
9.1	10.9	10.3	9.6	7.8	11.5	9.3	7.9	8.8
8.8	12.7	8.4	7.8	5.7	10.5	10.5	9.6	8.9
10.2	10.3	7.7	10.6	8.3	8.8	9.5	8.8	9.4

Source: The World Almanac and Book of Facts 2004

Solution En considérant les données brutes et faisant une analyse grossière juste par observation des valeurs, tentons de répondre d’ores et déjà à la question principale. L’énoncé en c nous donne la température de Chicago. On constate que les valeurs se distribuent autour de 10 m/s avec plus ou moins 2 à 3 m/s d’écarts maximaux ¹ et qu’environ moins d’une dizaine de ville ont une valeur supérieure à 10.3. Cette température se trouve également dans un petit intervalle de 1 jusqu’au pallier de 10, parmi moins d’une dizaine d’autres villes. Le nombre total d’échantillons étant de 44 en faisant abstraction de la valeur extrême de 35.1, on peut à priori dire que Chicago ne mérite pas absolument le nom de ville venteux, si on entend par là qu’elle se trouve "au dessus du lot"; elle doit se trouver parmi les seconds plus venteux et au dessus du mode de la distribution, environ 5 villes dépassent largement 10, de 20% cette valeur.

Remarque: les fréquences relatives de ces données sont proportionnelles au valeurs de la variable vitesse, on peut indifféremment inférer sur l’un ou l’autre ensemble, puisque celles-ci n’ont pas été regroupées en classes.

a La construction de l’histogramme des fréquences relatives (f.r.) sans regroupement des intervalles de villes en classe, nous donne déjà des informations plus claires (Chicago en rouge): Seules 10 villes ont une f.r. plus élevée que celle de Chicago, dont 2 à peine moins de 5% plus élevée; Chicago se trouve approximativement en dessous de 20 % des villes et parmi 6 à 8 villes à vit. moyenne de vent très proche.

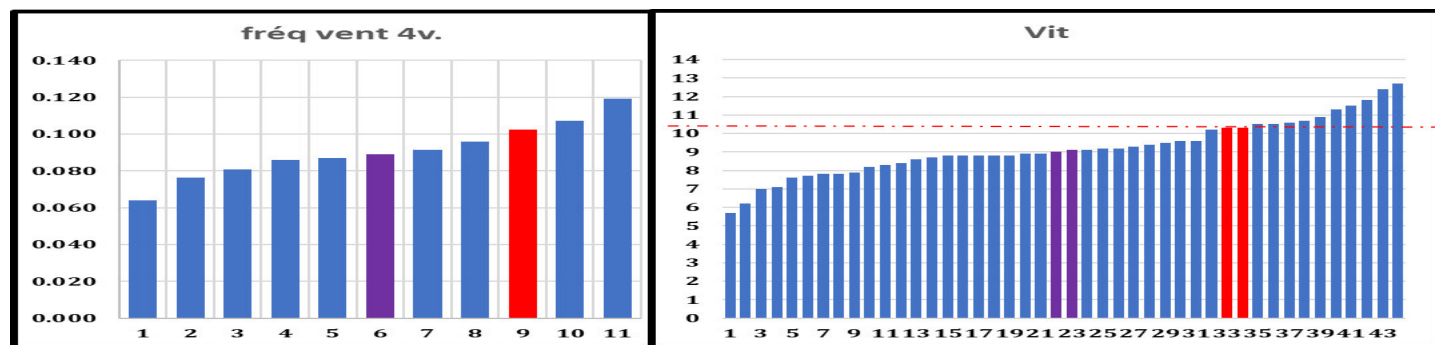


¹ maximums (adj.), non recommandé par l’Académie!

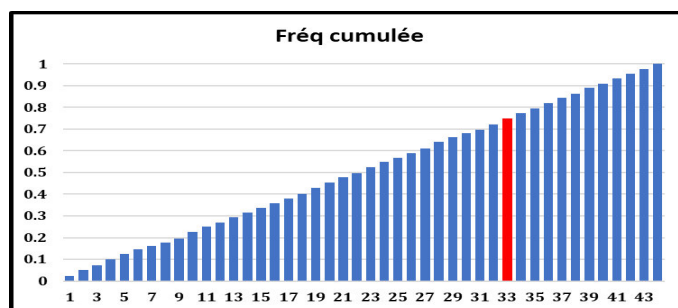
On peut donc confirmer que Chicago se trouve au pallier inférieur approximatif du 20 % des villes les plus venteux et qu'il peut mériter le nom de ville venteux sans se trouver parmi les plus venteux, si l'on considère que les données sont précises (on s'est aussi autorisé à exclure la valeur excessive de 35.1)

b/ La valeur pour Washington étant de 35.1 et la somme de toutes les valeurs 441 (405.5 sans W.), sa fréquence est de 0.8 , càd plus du double de la fréquence la plus élevée. Dans nos données statistiques nous n'avons pas d'informations chiffrées relatives à la géophysique de Washington quelque-elles fussent (pressions atmosphériques, courants , dénivelés..etc) . On ne peut donc supputer que 3 hypothèses; les mesures la concernant sont fausses et dans ce cas là on suppose aussi que cette erreur s'est répétée dans toute l'étendue de l'intervalle de mesures, puisqu'il s'agit d'une moyenne, qu'il y a eu un phénomène extra-ordinaire géophysique durant ce délai de mesures, ou qu'effectivement la géographie de Washington est particulière et qu'elle peut expliquer cette amplitude. On ne peut poser aucune statistique sur les deux premières hypothèses, ne disposant pas de données pouvant les étayer, par contre on peut dire que si les deux premières doivent être exclues, la dernière est dans ce cas-là la seule probable, et pour la confirmer il faudrait effectuer de nouvelles mesures d'ensemble en excluant une probabilité non nulle d'erreur sur Washington.

c/ d/Le pourcentage de villes à vitesse moyenne supérieure à celle de Chicago se calcule très aisément, 10 villes sur 44 ont une valeur supérieure et en incluant la vitesse de 10.2 juste au-dessous, 13 (12) sur 44 ont une valeur supérieur ou (quasi-)égale, càd respectivement 23%, (27)29.5 % Comme on a vu, 23 pourcents ont strictement une vit. supérieure et 29.5 % quasi-égale à supér. , donc on consolide nos considérations sur les données brutes, Chicago se trouve dans un quartier des valeurs supérieurs mais au bas de cet intervalle. On peut créer un nouveau résumé, afin de voir si l'on peut tirer davantage d'informations. Prenons pour commencer avec de petites étendues compte tenu du petit nombre de valeur, comme le recommande l'auteur, avec de petits intervalles de 4 éléments , ce qui semble judicieux pour l'ensemble de 44 valeurs;on a un rapport entier donc un nombre entier d'intervalles avec de plus une classe centrale (6). Pour cela, on trie les valeurs dans un tableur et somme celles-ci selon le nombre de valeurs individuelles par classe (4) et recalcule les fréquences et reproduit le graphe. Cette distribution indique que stricto sensu 22 % des effectifs se retrouvent au-dessus de la classe de Chicago.



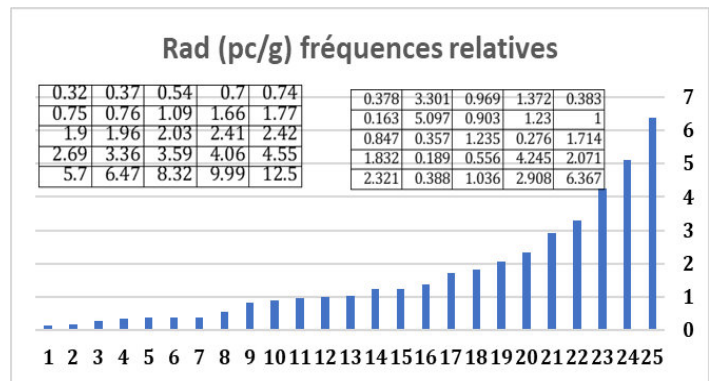
L'histogramme des fréquences cumulées permet également de chiffrer directement le seuil de Chicago qui se trouve dans la colonne 33, à la limite supérieur des 75 % cumulés. Donc 22 à 25 % des villes ont une vit. moyenne de vent supérieurs ou égale à supérieur à Chicago et elle est donc venteuse mais non de manière marquante.



1.3. Of great importance to residents of central Florida is the amount of radioactive material present ²³⁸ in the soil of reclaimed phosphate mining areas. Measurements of the amount of U in 25 soil samples were as follows (measurements in picocuries per gram): Construct a relative frequency histogram for these data.

1.3 La quantité de matières radioactives présentes revêt une grande importance pour les résidents du centre de la Floride. ²³⁸ dans les sols des zones minières de phosphate réhabilitées. Mesures de la quantité d'uranium dans 25 sols Les échantillons étaient les suivants (mesures en picocuries par gramme) : Construisez un histogramme de fréquence relative pour ces données.

Solution On trie les données dans un tableur, calcule les fréq. relat. Ci-ctr. graph. et 2 tabl. des valeurs brutes et fréq. triées.



1.4 The top 40 stocks on the over-the-counter (OTC) market, ranked by percentage of outstanding shares traded on one day last year are as follows:

a Construct a relative frequency histogram to describe these data.

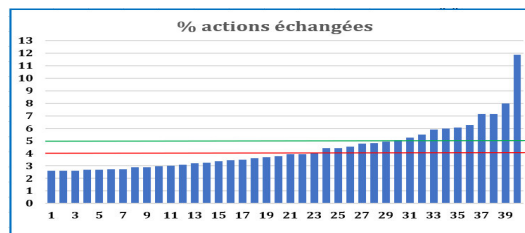
b What proportion of these top 40 stocks traded more than 4 % of the outstanding shares?

c If one of the stocks is selected at random from the 40 for which the preceding data were taken, what is the probability that it will have traded fewer than 5 % of its outstanding shares?

1.4 Les 40 principales actions cotées sur le marché hors cote (OTC), classées par pourcentage d'actions en circulation. Voici le nombre d'actions échangées en une journée l'année dernière :

a Construisez un histogramme de fréquence relative pour décrire ces données. b Quelle proportion de ces 40 actions les plus importantes a représenté plus de 4 % des actions en circulation ? c Si l'une des actions est sélectionnée au hasard parmi les 40 pour lesquelles les données précédentes ont été recueillies, étant donné que cette société a échangé moins de 5 % de ses actions en circulation, quelle est la probabilité qu'elle ait fait l'objet de telles transactions ?

11.88	6.27	5.49	4.81	4.40	3.78	3.44	3.11	2.88	2.68
7.99	6.07	5.26	4.79	4.05	3.69	3.36	3.03	2.74	2.63
7.15	5.98	5.07	4.55	3.94	3.62	3.26	2.99	2.74	2.62
7.13	5.91	4.94	4.43	3.93	3.48	3.20	2.89	2.69	2.61



Solution On procède comme dans l'ex. précédent, on peut trier et employer des formules excel ou code VBA* (Les 1ères macros (2 versions) permettent de créer 1 liste en colonne triée pour la sélection de données > graphique (les variables à caractériser (N°) sont à saisir manuellement). Les secondes (id.) permettent de sélectionner les valeurs au-dessus ou dessous d'un seuil et d'en déterminer le pourcentage de l'effectif total. voir instructions et codes ci-dessous)

b/ La proportion de ces actions échangées dépassant de 4 % se calcule simplement; on fait un simple rapport du nombre dépassant ce seuil au nombre total: $\frac{18}{40} = 0.45$

c/ Il n'est pas nécessaire de d'appliquer l'analyse combinatoire dans ce cas, (v. Probabilité chap II), la sélection d'une action quelle qu'elle soit est équiprobable, donc c'est également le rapport du nombre d'actions échangé à moins de 5 % sur le nombre totale que l'on cherche: $\frac{29}{40} = 0.725$ (OU $\frac{29}{C_{40}^1}$)

* Le 2 premiers codes VBA suivants permettent de copier le contenu d'un table de valeur excel dans une colonne, d'effectuer une copie dans une autre colonne et à y faire le tri, il y a 2 versions; la première à coller dans un module (developpeur>visual basic>insertion> module) s'applique à toutes les feuilles du classeur , la seconde à la feuille spécifique ("ex1.4" à remplacer) à intégrer dans les fichiers Feuil (developpeur>visual basic>insertion>Feuil"X" au choix). Pour appeler ces macros personnellement, j'affecte une macro à une forme placée dans une cellule (insertion>forme). Pour ne pas charger le code il n'y a qu' une annulation seulement pour un faux nom de feuille fourni dans la verison 1. Dans les deux versions les numéros de la première ligne du tableau , de la dernière , de la première colonne et dernière du tableau ainsi que les numéros de la première ligne de la liste en colonne et le numéro de la colonne doivent être saisis via des input box, la macro colle alors la série tableau dans la colonne indiquée et une copie est placée dans une plage à deux colonnes à droite de la première pour un tri automatique. (code pas prévu pour un tableau incomplet v. ex 1.8) Les 2 seconds codes sont également applicables pour toutes les feuilles ou la feuille à spécifier, ils permettent de calculer le rapport du nombre d'effectifs au-dessus ou dessous d'un seuil déterminé et aussi leur pourcentage retourné par la boîte de message, pour récupérer ce rapport dans une cellule ou ailleurs, il faut adapter le code. Je préfère utiliser les coordonnées worksheets.cells(x,y), pour d'autres options coordonnées Range("A",y), ou même conversion des lettres A , B , C..etc en indice numérique 1, 2, 3 , pour la sélection manuelle de la plage du tableau à traiter, demander à l'IA de modifier mes codes le plus simplement sans commentaire peut-être sauf aux lignes de changement pour les identifier, mes codes ont l'avantage d'être des plus simples.

Rem: malheureusement je ne parviens pas encore en formater le document pdf de manière à ce que le code copié puisse être recollé conserve la mise en forme sans concaténation (bug avec Lulatex.)

TABLEAU COPIÉ EN COLONNE DE VALEURS , SECONDE COPIE COLLÉE DANS COLONNE + 2 ET TRI
1/ VERSION APPLICABLE A TOUTE LES FEUILLES 2/ A LA FEUILLE SPECIFIQUE AFFICHAGE DE LA
PROPORTION D'EFFECTIFS EN-DESSUS OU DESSOUS D'UN SEUIL 1'/ VERSION APPLICABLE A TOUTE
LES FEUILLES 2'/ A LA FEUILLE SPECIFIQUE

TABLEAU COPIÉ EN COLONNE DE VALEURS , SECONDE COPIE COLLÉE DANS COLONNE + 2 ET TRI 1/
VERSION APPLICABLE A TOUTE LES FEUILLES

```
Sub reporttableencolettridansautrecol()
NomFeuille = Application.InputBox("Entrez le nom exact de la feuille de calcul :", "Cible Feuille", Type:=2)
If NomFeuille = "" Or NomFeuille = "False" Then Exit Sub
Set ws = ThisWorkbook.Worksheets(NomFeuille)
lgni = Application.InputBox("Entrer la ligne initiale:", Title:="ligne initiale", Type:=1)
lgnf = Application.InputBox("Entrer la ligne finale:", Title:="ligne finale", Type:=1)
coli = Application.InputBox("Entrer la colonne initiale:", Title:="colonne initiale", Type:=1)
colf = Application.InputBox("Entrer la colonne finale:", Title:="colonne finale", Type:=1)
lgndestf = Application.InputBox("Entrer la première ligne de destination finale:", Title:="1ère ligne destina")
coldestf = Application.InputBox("Entrer la colonne de destination finale:", Title:="colonne de destination fi")
premligncol = lgndestf
For j = lgni To lgnf
    For k = coli To colf
        ws.Cells(lgndestf, coldestf) = ws.Cells(j, k)
        dernligncol = lgndestf
        lgndestf = lgndestf + 1
    Next k
Next j
For j = premligncol To dernligncol
    ws.Cells(j, coldestf + 2) = ws.Cells(j, coldestf)
Next j
Set PlageTri = ws.Range(ws.Cells(premligncol, coldestf + 2), ws.Cells(dernligncol, coldestf + 2))
Set CleTri = ws.Cells(premligncol, coldestf + 2)
Application.ScreenUpdating = False
```

```

Application.DisplayAlerts = False
ws.Sort.SortFields.Clear
ws.Sort.SortFields.Add Key:=CleTri, SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortTextAsNumbers
With ws.Sort
    .SetRange PlageTri
    .Header = xlNo
    .MatchCase = False
    .Orientation = xlTopToBottom
    .SortMethod = xlPinYin
    .Apply
End With
Application.DisplayAlerts = True
Application.ScreenUpdating = True
End Sub

```

TABLEAU COPIÉ EN COLONNE DE VALEURS , SECONDE COPIE COLLÉE DANS COLONNE + 2 ET TRI 2/ A LA FEUILLE SPECIFIQUE

```

Sub tableetlisteencolfeuilX()
lgni = Application.InputBox("Entrer la ligne initiale:", Title:="ligne initiale", Type:=1)
lgnf = Application.InputBox("Entrer la ligne finale:", Title:="ligne finale", Type:=1)
coli = Application.InputBox("Entrer la colonne initiale:", Title:="colonne initiale", Type:=1)
colf = Application.InputBox("Entrer la colonne finale:", Title:="colonne finale", Type:=1)
lgndestf = Application.InputBox("Entrer la première ligne de destination finale:", Title:="1ère ligne destination", Type:=1)
coldestf = Application.InputBox("Entrer la colonne de destination finale:", Title:="colonne de destination finale", Type:=1)
premligncol = lgndestf
For j = lgni To lgnf
    For k = coli To colf
        Worksheets("ex1.4").Cells(lgndestf, coldestf) = Worksheets("ex1.4").Cells(j, k)
        dernligncol = lgndestf
        lgndestf = lgndestf + 1
    Next k
Next j
For j = premligncol To dernligncol
    Worksheets("ex1.4").Cells(j, coldestf + 2) = Worksheets("ex1.4").Cells(j, coldestf)
Next j
Set ws = Worksheets("ex1.4")
Set PlageTri = ws.Range(ws.Cells(premligncol, coldestf + 2), ws.Cells(dernligncol, coldestf + 2))
Set CleTri = ws.Cells(premligncol, coldestf + 2)
Application.ScreenUpdating = False
Application.DisplayAlerts = False
ws.Sort.SortFields.Clear
ws.Sort.SortFields.Add Key:=CleTri, SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortTextAsNumbers
With ws.Sort
    .SetRange PlageTri
    .Header = xlNo
    .MatchCase = False
    .Orientation = xlTopToBottom
    .SortMethod = xlPinYin
    .Apply
End With
Application.DisplayAlerts = True
Application.ScreenUpdating = True
End Sub

```

AFFICHAGE DE LA PROPORTION D’EFFECTIFS EN-DESSUS OU DESSOUS D’UN SEUIL 1’/ VERSION AP- PLICABLE A TOUTE LES FEUILLES

```

Sub nbrvaldepseuil()
If MsgBox("ce macro détermine le pourcentage d’un effectif en-dessus ou dessous d’une valeur, faire le calcul", vbYesNo) = vbYes Then
    NomFeuille = Application.InputBox("Entrez le nom exact de la feuille de calcul :", "Cible Feuille", Type:=2)
    If NomFeuille = "" Or NomFeuille = "False" Then Exit Sub
    Set ws = ThisWorkbook.Worksheets(NomFeuille)
    premligne = Application.InputBox("Entrer la ligne initiale de la colonne de valeurs:", Title:="ligne initiale", Type:=1)
    dernligne = Application.InputBox("Entrer la ligne finale de la colonne de valeurs:", Title:="ligne finale", Type:=1)
End If

```

```

colonne = Application.InputBox("Entrer la colonne de valeurs:", Title:="colonne de la plage de valeurs", Type:=xlTypeText)
vallim = Application.InputBox("Entrer la valeur limite:", Title:="pourcentage de l'effectif en dessus ou dessous du seuil", Type:=xlTypeText)
If MsgBox("Déterminer le pourcentage de valeurs au-dessus (Yes) ou en dessous (No) d'une limite donnée?", vbYesNo) = vbYes Then
    n = 0
    For j = premligne To dernligne
        If ws.Cells(j, colonne) > vallim Then
            n = n + 1
        End If
    Next j
    MessagePrompt = "nombre et pourcent de valeurs au-dessus de " & vallim & " : " & n & "/" & dernligne - premligne + 1 & " %"
    MsgBox MessagePrompt, vbInformation, "nbr et pc valeurs au-dessus du seuil"
Else
    n = 0
    For j = premligne To dernligne
        If ws.Cells(j, colonne) < vallim Then
            n = n + 1
        End If
    Next j
    MessagePrompt = "nombre et pourcent de valeurs au-dessous de " & vallim & " : " & n & "/" & dernligne - premligne + 1 & " %"
    MsgBox MessagePrompt, vbInformation, "nbr et pc valeurs au-dessous du seuil"
End If
Exit Sub
End If
End Sub

```

AFFICHAGE DE LA PROPORTION D'EFFECTIFS EN-DESSUS OU DESSOUS D'UN SEUIL 2' / A LA FEUILLE SPECIFIQUE

```

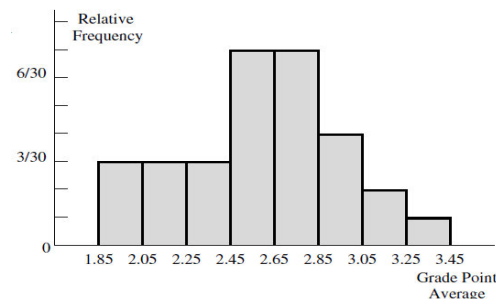
Sub nombrvaldepseuilfeuilX()
If MsgBox("ce macro détermine le pourcentage d'un effectif en-dessus ou dessous d'une valeur, faire le calcul", vbYesNo) = vbYes Then
    premligne = Application.InputBox("Entrer la ligne initiale de la colonne de valeurs:", Title:="ligne initiale", Type:=xlTypeText)
    dernligne = Application.InputBox("Entrer la ligne finale de la colonne de valeurs:", Title:="ligne finale", Type:=xlTypeText)
    colonne = Application.InputBox("Entrer la colonne de valeurs:", Title:="colonne de la plage de valeurs", Type:=xlTypeText)
    vallim = Application.InputBox("Entrer la valeur limite:", Title:="pourcentage de l'effectif en dessus ou dessous du seuil", Type:=xlTypeText)
    If MsgBox("Déterminer le pourcentage de valeurs au-dessus (Yes) ou en dessous (No) d'une limite donnée?", vbYesNo) = vbYes Then
        n = 0
        For j = premligne To dernligne
            If Worksheets("ex1.4").Cells(j, colonne) > vallim Then
                n = n + 1
            End If
        Next j
        MessagePrompt = "nombre et pourcent de valeurs au-dessus de " & vallim & " : " & n & "/" & dernligne - premligne + 1 & " %"
        MsgBox MessagePrompt, vbInformation, "nbr et pc valeurs au-dessus du seuil"
    Else
        n = 0
        For j = premligne To dernligne
            If Worksheets("ex1.4").Cells(j, colonne) < vallim Then
                n = n + 1
            End If
        Next j
        MessagePrompt = "nombre et pourcent de valeurs au-dessous de " & vallim & " : " & n & "/" & dernligne - premligne + 1 & " %"
        MsgBox MessagePrompt, vbInformation, "nbr et pc valeurs au-dessous du seuil"
    End If
End If
Exit Sub
End If
End Sub

```

1.5 Given here is the relative frequency histogram associated with grade point averages (GPAs) of a sample of 30 students: a Which of the GPA categories identified on the horizontal axis are associated with the largest proportion of students? b What proportion of students had GPAs in each of the categories that you identified? c What proportion of the students had GPAs less than 2.65?

1.5 Voici l'histogramme des fréquences relatives associées aux moyennes pondérées cumulatives (MPC) d'un échantillon de 30 étudiants :

a) Parmi les catégories de MPC identifiées sur l'axe horizontal, lesquelles sont associées à la plus grande proportion d'étudiants ? b) Quelle proportion d'étudiants avait une moyenne pondérée cumulative dans chacune des catégories que vous avez identifiées ? c) Quelle proportion des étudiants avait une moyenne pondérée cumulative inférieure à 2,65 ?



Solution a/ Les deux plages de moyennes allant des notes de 2.45 à 2.65 et 2.65 à 2.85 contiennent les plus grandes proportions d'étudiants.

b/ ? Les proportions d'étudiants pour ces 2 catégories de plus grandes proportions sont de 7/30. ?? Les différentes classes de moyennes (MPC) ont des fréquences suivantes : { 1/30 , 2/30 , 3/30 , 4/30 , 7/30 } dont 3 classes à 3 30ème et 2 à 7 30ème, càd 9/30 pour la 1ère cl 14/30 pour la 2nd. ?? Rép. de l'auteur 7/30 !? la moyenne cumulative se trouve dans une classe 7/30 ; $3 \times 3/30 + 1 \times 7/30 = 0.53$, question non compréhensible!

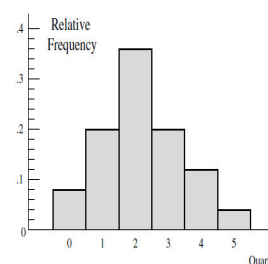
c/ la variable unitaire 2.65 est à la borne supérieur de la classe 2.25-2.65 et à la borne inférieure de la classe supérieure. Si l'on effectue la somme des 4 premiers intervalles on inclut théoriquement aussi les fréquences relatives à la limite de 2.65 si cette valeur est comprise dans la 4ème classe. Il faut choisir un critère qui stipule quelle valeur individuelle est considérée comme inférieure à 2.65. Comme les valeurs décimales sont d'ordre deux, le seuil de différenciation individuelle des notes est de 0.01, or les intervalles ont des étendues de 0.2, quel critère créer? On peut supposer que la limite supérieur de la 4ème, ne comprend en fait juste pas 2.65, comme si elle était la borne d l'intervalle ouvert, dans ce cas là la réponse est 16/30 somme des fréquences des 4 premières classes. Dans le cas contraire, la fréquence recherchée devrait être plus petite, mais dans ce cas là la classe supérieure ne devrait pas contenir 2.65. Mais si 2.65 est sensu stricto à la borne inférieur de la classe suivante, cette valeur ne pouvant mathématiquement pas être compris dans les deux intervalles, la réponse est bien 16/30. A partir de ces arguments on peut logiquement poser pour réponse la plus probable 16/30. On peut aussi supposer que les fréquences relatives liées au point 2.65 sont également réparties entre les deux intervalles et choisir arbitrairement comme seuil de différenciation 0.1, dans ce cas là on peut diviser la classe 4 en quatre parts, 1 affectée à un demi intervalle supérieur de 2.45, 2 à 2.55 et 1 à 2.65 , dans ce cas on ne prends que le 3/4 de la quatrième classe. Si l'on choisit un seuil de 0.2 on aurait arbitrairement la moitié de la classe "égale" à 2.45 et l'autre à 2.65, dont 1/2 de chacune "égale" à 2.55 puisque nécessairement ces demi-classes devraient contenir 2.55, donc finalement on aboutit au même résultat. les 3/4 de la classe seront inférieurs à 2.65. Et ainsi de suite.

intervalles classe 3& 4	Seuil de différenciation	$\sum F(X < 2.45)$	$\sum F(2.45 < X < 2.65)$	$\sum F(X < 2.65)$
[2.45:2.65[[2.65:2.85]	0	9/30	7/30	16/30
[2.45:2.65] [2.65:2.85]	0.2	9/30	$3/4 \times 7/30$	57/120
[2.45:2.65] [2.65:2.85]	0.1	9/30	$3/4 \times 7/30$	57/120
[2.45:2.65] [2.65:2.85]	0.01	9/30	$3/4 \times 7/30$	57/120

1.6 The relative frequency histogram given next was constructed from data obtained from a random sample of 25 families. Each was asked the number of quarts of milk that had been purchased the previous week. a Use this relative frequency histogram to determine the number of quarts of milk purchased by the largest proportion of the 25 families. The category associated with the largest relative frequency is called the modal category . b What proportion of the 25 families purchased more than 2 quarts of milk? c What proportion purchased more than 0 but fewer than 5 quarts?

1.6 L'histogramme de fréquence relative présenté ci-après a été construit à partir de données obtenues auprès d'un échantillon aléatoire de 25 familles. On a demandé à chacune d'elles le nombre de litres de lait achetés la semaine précédente. a Utilisez cet histogramme de fréquences relatives pour déterminer le nombre de litres de lait achetés par la plus grande proportion des 25 familles. La catégorie associée à la fréquence relative la plus élevée est appelée la catégorie modale. b Quelle proportion des 25 familles a acheté plus de 2 litres de lait ? c Quelle proportion a acheté plus de 0 mais moins de 5 litres ?

a/ La fréquence relative ou proportion la plus élevée sur l'ensemble de l'effectif total de familles (25) d'une classe donnée est 36 % et correspond à la consommation de 2 litres qui est la classe modale. (1/4 de gallon). Le nombre de famille consommant cette quantité est donc de $0.36 \times 25 = 9$ Rappel, la classe modale étant la classe qui contient la variable unitaire qui partage la distribution en deux parts égales. Sans déterminer cette valeur par un calcul, si l'on présuppose que le mode se trouve dans la classe 2 approximativement au milieu de l'intervalle central, la valeur modale partage la distribution à ce point entre 47 % et 53 % des effectifs. La variable volume de lait étant continue, pour déterminer la valeur précise du mode, on doit connaître précisément les valeurs des bornes des intervalles, p.ex pour la classe x, on aurait $[x-0.5:x+0.5]$, mais ici ce rayon ne serait pas applicable à 0 si les intervalles sont admis avec étendues égales, on aurait nécessairement $x \rightarrow [x:x+1[$ à moins de prendre 1/2 intervalle pour la classe 0 (et 5) .



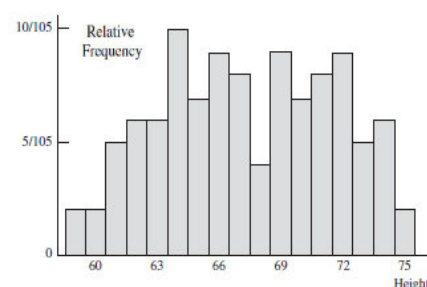
b/ La proportion de famille ayant acheté plus de 2 L est donnée par la somme des proportions des classes supérieures { 3, 4, 5 } est donc 36 % (pour emprunter la formalisation statistique v. Ch II) $(F(X \leq 2) + F(X > 2) = 0.64 + 0.36 = 1)$

c/ La proportion achetant plus de 0 et moins de 5 : $F(0 < X < 5) = F(\{ X=2 \} \cup \{ X=3 \} \cup \{ X=4 \}) = 0.56$

1.7 The self-reported heights of 105 students in a bio-statistics class were used to construct the histogram given below. a Describe the shape of the histogram. b Does this histogram have an unusual feature? c Can you think of an explanation for the two peaks in the histogram? Is there some consideration other than height that results in the two separate peaks? What is it?

1.7 Les tailles déclarées par 105 étudiants d'un cours de biostatistiques ont été utilisées pour construire le histogramme ci-dessous. a) Décrivez la forme de l'histogramme. b) Cet histogramme présente-t-il une caractéristique inhabituelle ? c) Pouvez-vous expliquer la présence de deux pics dans l'histogramme ? Y a-t-il un autre facteur que la hauteur qui explique ces deux pics distincts ? Lequel ?

a/ b/ La distribution est en forme de dôme voire de double cloche avec un intervalle quasi central de fréquence moyenne, on peut supposer que ce résumé est une combinaison de deux distributions qui mesurent donc deux tendances de deux variables statistiques dans la seule variable taille des étudiants, en imaginant deux courbes de variable continue épousant et représentant au mieux cet histogramme (de somme d'intégrales = 1), on aurait deux pics placés approximativement en 64 (64.5) et 71 (71.5) , correspondant à (1 pc = 2.54cm) environ 160 cm et 180 cm



c/ Partant de la connaissance de la différence statistique de la valeur centrale de la taille entre les hommes et les femmes , on peut proposer que la variable sexe explique cette forme de distribution, avec une taille centrale autour de 160cm pour les femmes et 180 cm pour les hommes. On peut comparer ces valeurs avec des données des valeurs modales des tailles de jeunes étudiants du pays d'échantillonnage. La question que l'on peut se poser est la suivante, existe-t-il des distributions de taille dans la nature de forme similaire associée à une seule variable? Si on a la certitude que non, pour démontrer que notre hypothèse est la seule permettant d'inférer que le sexe explique ces distribution, il faut reconduire l'échantillonnage en distinguant les deux sexes.

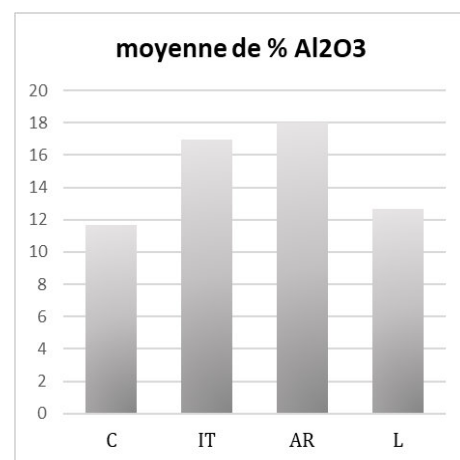
1.8 An article in Archaeometry presented an analysis of 26 samples of Romano-British pottery, found at four different kiln sites in the United Kingdom. The percentage of aluminum oxide in each of the 26 samples is given below: a Construct a relative frequency histogram to describe the aluminum oxide content of all 26 pottery samples. b What unusual feature do you see in this histogram? Looking at the data, can you think of an explanation for this unusual feature?

Llanedryn	Caldicot	Island Thorns	Ashley Rails
14.4	11.6	11.8	18.3
13.8	11.1	11.6	15.8
14.6	13.4	18.0	16.7
11.5	12.4	18.0	14.8
13.8	13.1	20.8	19.1
10.9	12.7		
10.1	12.5		

Source: A. Tubb, A. J. Parker, and G. Nickless, "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry," *Archaeometry* 22 (1980): 153.

Solution

a/ b/ L'histogramme est quasiment plat, ce résumé n'est pas bien représentatif de la distribution du niveau d'oxyde relatif à la variable site; il y a un étalement des fréquences (ou des valeurs d'oxyde puisque il n'y pas de regroupement de la variable indépendante, les fréquences sont proportionnelles aux teneurs individuelles); plusieurs fréquences représentent des valeurs individuelles pour caractériser chaque site et ceci en redondance diverses, ainsi les pondérations par site sont différentes, sur cette base statistique et sans regroupement imaginaire ou méthodique, on ne peut pas tirer d'inférence sur les différences entre les sites. On pourrait au mieux caractériser une distribution pour l'ensemble du RU selon un critère spatial si le nombre d'échantillons analysés par site est proportionnel à la taille du site, et si ce critère trouve un sens statistique !? . On doit donc faire une moyenne d'ensemble pour de chaque site, caractéristique qui permet des inférences ultérieures en archéométrie (correlation possible entre la teneur porterie-sol environnant intact, -sol ératique ...).



Les moyennes par site donnent un sens statistique; on a une distribution plus classique et répondant aux lois statistiques avec deux tendances une centrale qui caractérise les sites IT et AR de moyennes proches et l'autre qui caractérise les sites C et L également de moyennes proches. On pourrait dans ce cas élaborer des hypothèses de corrélation, tirer des inférences intéressant la recherche archéologique.

1.9 Resting breathing rates for college-age students are approximately normally distributed with mean 12 and standard deviation 2.3 breaths per minute. What fraction of all college-age students have breathing rates in the following intervals?

- a 9.7 to 14.3 breaths per minute**
- b 7.4 to 16.6 breaths per minute**
- c 9.7 to 16.6 breaths per minute**
- d Less than 5.1 or more than 18.9 breaths per minute**

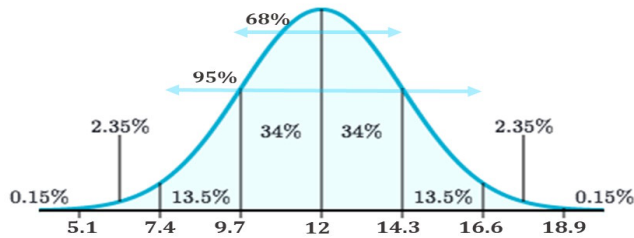
1.9 La fréquence respiratoire au repos des étudiants est approximativement distribuée selon une loi normale de moyenne 12 et d'écart type 2,3 respirations par minute. Quelle proportion de ces étudiants présente des fréquences respiratoires comprises dans les intervalles suivants ?

- a) 9,7 à 14,3 respirations par minute
- b) 7,4 à 16,6 respirations par minute
- c) 9,7 à 16,6 respirations par minute
- d) Moins de 5,1 ou plus de 18,9 respirations par minute

Solution Calculons les limites des fréquences caractéristiques statistiques : $\bar{y} \pm 1\sigma = 9.7 - 14.3(68\%)$ $\bar{y} \pm 2\sigma = 7.4 - 16.6(95\%)$ $\bar{y} \pm 3\sigma = 5.1 - 18.9(> 95\%)$

On obtient les limites que l'on peut représenter sur un graphique canonique caractéristique de dispersion normale, on considère que la distribution est symétrique. On peut directement répondre aux 2 premières questions. La plage de c comporte un intervalle d'écart d'une part et 2 de l'autre. Pour la d on considère qu'elle contient la différence entre la fréquence totale et celle comportant deux écarts-type. L'opérateur ou dans l'intitulé est synonyme de et on recherche la proportion totale des fréquences en dessous de 5,1 et en dessus de 18.9.

9,7-14,3	7,4-16,6	9,7-16,6	< 5,1 ou > 18,9
68 %	95 %	81.5 %	< 5 %
		68 % + 13.5 %	< 99.9- 95 %
		34% + 0.5* 95 %	



1.10 It has been projected that the average and standard deviation of the amount of time spent online using the Internet are, respectively, 14 and 17 hours per person per year (many do not use the Internet at all!).

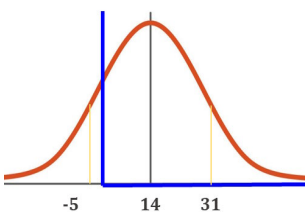
- a What value is exactly 1 standard deviation below the mean?**
- b If the amount of time spent online using the Internet is approximately normally distributed, what proportion of the users spend an amount of time online that is less than the value you found in part (a)?**
- c Is the amount of time spent online using the Internet approximately normally distributed? Why?**

1.10 On estime que la moyenne et l'écart type du temps passé en ligne sur Internet sont respectivement de 14 et 17 heures par personne et par an (beaucoup n'utilisent pas du tout Internet !).

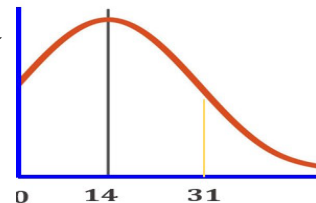
- a Quelle valeur est exactement inférieure à 1 écart type de la moyenne ?
- b Si le temps passé en ligne sur Internet suit approximativement une loi normale, quelle proportion des utilisateurs passe un temps en ligne inférieur à la valeur trouvée dans la partie (a) ?
- c Le temps passé en ligne sur Internet suit-il approximativement une distribution normale ? pourquoi?

Solution Calculons la 1ère plage de fréquence caractéristique et représentons la distribution normale sur la base des données statistiques : $\bar{y} \pm 1\sigma = 17 \pm 14 = \pm - 5/31(68\%)$

a/ b/ La limite d'un écart-type en dessous de la moyenne est de 14-17 = -5, donc toute valeur inférieure à cette valeur théorique est celle recherchée. La proportion des utilisateurs qui passe moins d'un écart-type sur IN peut être calculé de la manière suivante; les utilisateurs dans l'intervalle d'un écart-type représentent 68 % de l'effectif, donc la différence entre l'effectif total et celui-ci représente la fréquence des plages en-deçà et au-dessus d'un écart-type.



Comme nous cherchons uniquement ceux en-dessous d'un écart-type, on doit donc prendre la moitié de cette proportion, donc $\frac{1 - 0.68}{2} = 16\%$. c/ Oui la distribution virtuelle est normale, la distribution réelle ne l'est pas; si on devait la représenter elle serait tronquée à la limite de 0, puisque les valeurs en dessous de zéro ne correspondent à une réalité de temps d'utilisation.



On peut recalculer le point b/ sur cette base d'information ; $P(x \leq 0) = P(0) = P(\bar{x}) - \sum P(\bar{x} + 2\sigma), P(\bar{x} - s\sigma)) = 1 - 0.5 - 0.34 = 0.16$

1.11 The following results on summations will help us in calculating the sample variance s^2 . For any constant.

$$a/ \sum_{i=0}^n c = nc \quad b/ \sum_{i=0}^n cy_i = c \sum_{i=0}^n y_i \quad c/ \sum_{i=0}^n (x_i + y_i) = \sum_{i=0}^n x_i + \sum_{i=0}^n y_i$$

Use (a), (b), and (c) to show that:

1.11 Les résultats suivants sur les sommes nous aideront à calculer la variance de l'échantillon s^2 . toute constante c ,

Utilisez (a), (b) et (c) pour montrer que:

$$s^2 = \frac{1}{n-1} \sum_{i=0}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 - \frac{1}{n} \left(\sum_{i=0}^n y_i \right)^2 \right]$$

Solution Il suffit de développer le membre du milieu de l'équation de la variance;

$$s^2 = \frac{1}{n-1} \sum_{i=0}^n (y_i - \bar{y})^2 = \frac{1}{n-1} (y_1^2 + \bar{y}^2 - 2y_1\bar{y} + y_2^2 + \bar{y}^2 - 2y_2\bar{y} + \dots + y_n^2 + \bar{y}^2 - 2y_n\bar{y}) =$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 + n\bar{y}^2 - 2\bar{y} \sum_{i=0}^n y_i \right] = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 + n \left(\frac{\sum_{i=0}^n y_i}{n} \right)^2 - 2 \left(\frac{\sum_{i=0}^n y_i}{n} \right) \sum_{i=0}^n y_i \right] =$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 + n \frac{(\sum_{i=0}^n y_i)^2}{n^2} - 2 \frac{(\sum_{i=0}^n y_i)^2}{n} \right] = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 - \frac{1}{n} \left(\sum_{i=0}^n y_i \right)^2 \right] \text{ form. équivalente :}$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 + n\bar{y}^2 - 2\bar{y} \sum_{i=0}^n y_i \right] = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 + n\bar{y}^2 - 2n\bar{y}\bar{y} \right] = \frac{1}{n-1} \left[\sum_{i=0}^n y_i^2 - n\bar{y}^2 \right]$$

1.12 Use the result of Exercise 1.11 to calculate s for the $n = 6$ sample measurements 1, 4, 2, 1, 3, and 3.
Solution

1.12 Utilisez le résultat de l'exercice 1.11 pour calculer s pour les $n = 6$ mesures d'échantillon 1, 4, 2, 1, 3, et 3

On saisit les valeurs dans un tableur, calcule la moyenne, la somme des carrés, puis la variance et l'écart-type .

$$n = 6 \quad \sum_{i=0}^n y_i = 14 \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i = 2.33 \quad \sum_{i=0}^n y_i^2 = 40 \quad s^2 = \frac{1}{5} (40 - \frac{1}{6} (14)^2) = 1.466 \quad s = 1.211$$

1.13 Refer to Exercise 1.2.

- a Calculate $\bar{y}$ and s for the data given.
 b Calculate the interval $\bar{y} \pm ks$ for $k = 1, 2$, and 3 .
 Count the number of measurements that fall within each interval and compare this result with the number that you would expect according to the empirical rule.

Solution On reporte les valeurs triées (vitesses moyennes du vent dans les cités américaines)

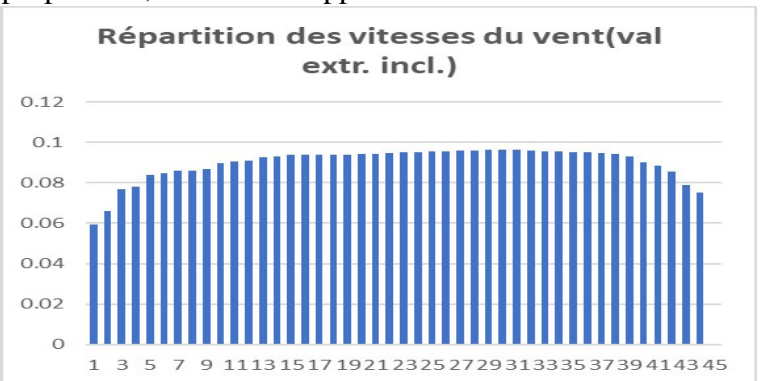
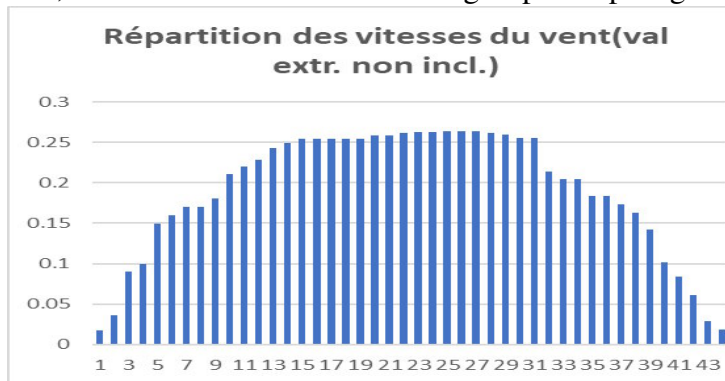
[5.7,6.2,7,7.1,7.6,7.7,7.8,7.8,7.9,8.2,8.3,8.4,8.6,8.7,8.8,8.8,8.8,8.8,8.8,8.9,8.9,9,9.1,9.1,9.2,9.2,9.3,9.4,9.5,9.6,9.6,10.2,10.3,10.3,10.5,10.5,10.6,10.7,10.9,11.3,11.5,11.8,12.4,12.7,(35.1)]

a/ b/ On saisit les valeurs dans un tableur, calcule la moyenne, la somme des carrés, puis la variance et l'écart-type.

$$n = 45 \quad \sum_{i=0}^n y_i = 440.6 \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i = 9.791 \quad \sum_{i=0}^n y_i^2 = 5067.38 \quad s^2 = \frac{1}{44} (5067.38 - \frac{1}{45} (440.6)^2) =$$

$$17.23 \quad s = 4.138 \quad \text{intervalles} > : [5.65; 13.93] \quad [1.52; 18.07] \quad [-2.62; 22.21] \quad f(y) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{(y-\bar{y})^2}{2s^2}}$$

On pose des formules conditionnelles (Excel) permettant d'identifier les valeurs situées dans les limites des 3 intervalles, en excluant la valeur extrême de 35,1, on trouve respectivement ; 76 % 96 % et 100 % . , contre les 68 , 95 et 100 % attendues. Ce qui montre que la distribution effective est proche de la théorique. On avait en effet des redondances de valeurs (5X 8.8) et des séries de valeurs très proches (7X 8.6-8.8) qu'il a fallu regrouper pour un résumé exploitable et une meilleure interprétation des résultats, ce que l'on fait afin d'approximer une répartition normale. Cette redondance de valeurs explique l'aplatissement de la forme de la courbe réelle traduisant un rapprochement du 1er intervalle du sec., en effet les valeurs centrales regroupent 1 plus grande proportion , la courbe s'apparenterait à 1 bosse étalée.



1.14 Refer to Exercise 1.3 and repeat parts (a) and (b) of Exercise 1.13.

Solution

Ci-dessous la liste triée des valeurs (radioactivité U (pc/g) sur 25 sites): [0.32,0.37,0.54,0.7,0.74,0.75,0.76,1.09,1.66,1.77,1.9,1.96,2.03,2.41,2.42,2.69,3.36,3.59,4.06,4.55,5.7,6.47,8.32,9.99,12.4]

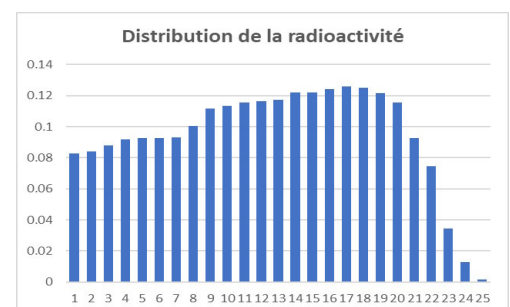
On saisit les valeurs dans un tableur, calcule la moyenne, la somme des carrés,

$$\text{puis la variance et l'écart-type. } n = 25 \quad \sum_{i=0}^n y_i = 80.63 \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i =$$

$$3.22 \quad \sum_{i=0}^n y_i^2 = 500.75 \quad s^2 = \frac{1}{24} (500.746 - \frac{1}{25} (80.63)^2) = 10 \quad s = 3.16$$

Les intervalles sont donc : [0.06 : 6.39] [-3.11 : 9.56] [-6.28 : 12.73]

1.14 Reportez-vous à l'exercice 1.3 et répétez les parties (a) et (b) de l'exercice 1.13.



On pose des formules conditionnelles dans les cellules excel permettant d'identifier les valeurs qui se situent dans les limites des 3 intervalles, on trouve respectivement ; 84 % 92 % et 100 % . , contre les 68 , 95 et 100 % attendues. On a également comme dans l'exercice précédent , une répartition de la distribution large et une même interprétation analogue, ici seules 2 valeurs différencient les demi-intervalles droits successifs sur un ensemble de 25 valeurs, la courbe devrait ressembler à une montagne égueulée avec un flanc droit abrupte.

1.15 Refer to Exercise 1.4 and repeat parts (a) and (b) of Exercise 1.13. **1.15 Reportez-vous à l'exercice 1.4 et répétez les parties (a) et (b) de l'exercice 1.13.**

Solution Ci-dessous la liste triée des valeurs (actions cotées):

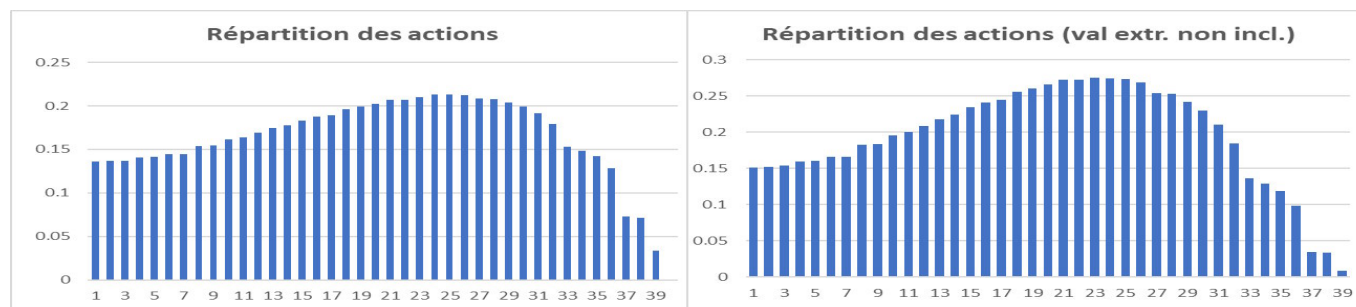
[2.61,2.62,2.63,2.68,2.69,2.74,2.74,2.88,2.89,2.99,3.03,3.11,3.2,3.26,3.36,3.44,3.48,3.62,3.69,3.78,3.93,3.94,4.05,4.4,4.43,4.55,4.79,4.81,4.94,5.07,5.26,5.49,5.91,5.98,6.07,6.27,7.13,7.15,7.99,11.8]

On saisit les valeurs dans un tableur, calcule la moyenne, la somme des carrés, puis la variance et l'écart-type .

$$n = 40 \sum_{i=0}^n y_i = 175.48 \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i = 4.387 \quad \sum_{i=0}^n y_i^2 = 906.4118 \quad s^2 = \frac{1}{39} (906.4 - \frac{1}{40} (175.5)^2) = 3.5 \quad s = 1.87$$

Les intervalles sont donc : [2.52 : 6.26] [0.64 : 8.13] [-1.23 : 10]

On pose des formules conditionnelles dans les cellules Excel permettant d'identifier les valeurs qui se situent dans les limites des 3 intervalles, on trouve respectivement ; 87.5 % 97.5 % et 100 % . , contre les 68 , 95 et 100 % attendues. On a également comme dans l'exercice précédent , une répartition de la distribution étalée et une même interprétation analogue, ici seules 4 puis 1 valeurs différencient les demi-intervalles droits successifs sur un ensemble de 40 valeurs, la courbe devrait ressembler à une montagne égueulée avec un flanc droit abrupte. (Graphe de droite v. ex suiv.)



1.16 In Exercise 1.4, there is one extremely large value (11.88). Eliminate this value and calculate y and s for the remaining 39 observations. Also, calculate the intervals $y \pm ks$ for $k = 1, 2$, and 3 ; count the number of measurements in each; then compare these results with those predicted by the empirical rule. Compare the answers here to those found in Exercise 1.15. Note the effect of a single large observation on y and s . **1.16 Dans l'exercice 1.4, une valeur est extrêmement élevée (11.88). Éliminez cette valeur et calculez y et s pour les 39 observations restantes. Calculez également les intervalles $y \pm ks$ pour $k = 1, 2$ et 3 ; comptez le nombre de mesures dans chaque intervalle ; puis comparez ces résultats avec ceux prédits par la règle empirique. Comparez les réponses obtenues ici avec celles de l'exercice 1.15. Notez l'effet d'une seule observation importante sur y et s.**

Solution Ci-dessous la liste triée des valeurs (actions cotées):

[2.61,2.62,2.63,2.68,2.69,2.74,2.74,2.88,2.89,2.99,3.03,3.11,3.2,3.26,3.36,3.44,3.48,3.62,3.69,3.78,3.93,3.94,4.05,4.4,4.43,4.55,4.79,4.81,4.94,5.07,5.26,5.49,5.91,5.98,6.07,6.27,7.13,7.15,7.99]

On saisit les valeurs dans un tableur, calcule la moyenne, la somme des carrés, puis la variance et l'écart-type .

$$n = 39 \quad \sum_{i=0}^n y_i = 163.6 \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i = 4.19 \quad \sum_{i=0}^n y_i^2 = 765.27 \quad s^2 = \frac{1}{38} (765.27 - \frac{1}{39} (163.6)^2) = 2.079 \quad s = 1.44$$

Les intervalles sont donc : [2.75 : 5.64] [1.31 : 7.08] [−0.13 : 8.52]

On pose des formules conditionnelles dans les cellules excel permettant d’identifier les valeurs qui se situent dans les limites des 3 intervalles et trouve respectivement ; 64 % 92 % et 100 % de répartition contre les 68 , 95 et 100 % attendus. En éliminant la valeur extrême, on se rapproche d’une distribution cannique avec un décalage systématique de 3 à 4 % pour les premiers intervalles, la courbe a une forme en cloche légèrement étalée, légèrement mieux distribuée, comme le montre le graphique de droite.

1.17 The range of a set of measurements is the difference between the largest and the smallest values. The empirical rule suggests that the standard deviation of a set of measurements may be roughly approximated by one-fourth of the range (t.i., range/4). Calculate this approximation to s for the data sets in Exercises 1.2, 1.3, and 1.4. Compare the result in each case to the actual, calculated value of s.

1.17 L’étendue d’un ensemble de mesures est la différence entre la valeur la plus élevée et la valeur la plus basse. La règle empirique suggère que l’écart type d’un ensemble de mesures peut être approximativement estimé par un quart de l’étendue (étendue/4). Calculez cette approximation de s pour les ensembles de données des exercices 1.2, 1.3 & 1.4. Comparez le résultat obtenu dans chaque cas à la valeur calculée de s.

Solution On reporte en tableau les références et variables mesurées, les moyennes, les intervalles, les étendues, les 1/4 d’étendues, s , l’écart relatif du quart d’étendue par rapport à s, l’écart relatif de s par rapport au quart d’étendue. Ces deux valeurs offrent un moyen de comparaison par cas.

	Ex 1.2 ss val extr.	Ex 1.2 av val extr.	ex 1.3	Ex 1.4 av val extr.	Ex 1.4 ss val extr.
	Vitesse du vent 1	Vitesse du vent 2	Radioactivité	cotes d’actions 1	cotes d’actions 2
n	44	45	25	40	39
moyennes	9.22	9.79	3.22	4.38	4.19
limites	[5.6:12.7]	[5.6:35.1]	[0.32:12.4]	[2.61:11.8]	[2.61:7.99]
étendue	7.1	29.5	12.08	9.19	5.38
1/4 étendue	1.77	7.37	3.02	2.29	1.34
s	1.51	4.14	3.16	1.87	1.44
éca.relat 1/4E / s	17 %	77 %	-4 %	22%	- 7%
éca.relat. s/1/4E	-14.7 %	-43.8 %	4 %	-18%	7.5%

On remarque en effet que l’écart relatif du quart d’étendu à l’écart-type est au maximum 17 % lorsque les valeurs extrêmes sont écartées (ex 2 et 4). La prise en compte de ces valeurs rend une déviation importante surtout pour le calcul de la vitesse du vent. On a même un bonne approximation de s pour la répartition de la radioactivité et des cotes boursières dont la valeur extrême a été exclue.

1.18 The College Board's verbal and mathematics Scholastic Aptitude Tests are scored on a scale of 200 to 800. It seems reasonable to assume that the distribution of test scores are approximately normally distributed for both tests. Use the result from Exercise 1.17 to approximate the standard deviation for scores on the verbal test.

Solution

On reporte le tableau en incluant l'intervalle de la variable note de test, moy. théorique, étendue et 1/4 d'étendue.

	1.2 ss val extr.	1.2 av val extr.	1.3	1.4 av val extr.	1.4 ss val extr.	1.18
	Vit. vent 1	Vit. vent 2	Radioactiv.	actions 1	actions 2	note verbale
moyennes	9.22	9.79	3.22	4.38	4.19	500 (théor.)
limites	[5.6:12.7]	[5.6:35.1]	[0.32:12.4]	[2.61:11.8]	[2.61:7.99]	[200:800]
étendue	7.1	29.5	12.08	9.19	5.38	600
1/2 étendue	1.77	7.37	3.02	2.29	1.34	150
s	1.51	4.14	3.16	1.87	1.44	128 (156)
ER 1/4E/ s	17 %	133 %	-4 %	22%	- 7%	17 % (-4 %)
ER s/1/4E	-14.7 %	-43.8 %	4 %	-18%	7.5%	

Si l'on se base sur le maximum 17 % d'écart obtenu entre le quart d'étendue et l'écart-type des différentes dispersions des références où l'on a exclu les valeurs extrêmes, on peut évaluer s pour le score du test verbal en se basant sur cette valeur; pour l'estimation de s on pose l'équation de l'écart relatif du quart d'étendu à s que l'on égale au maximum

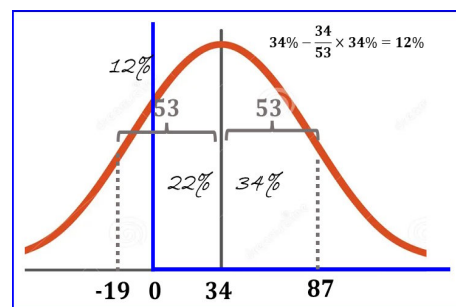
trouvé 17 % (min 4 %) , ce qui donne :
$$\frac{1/4E - s}{s} = 0.17(-0.04) \Rightarrow s = \frac{1/4E}{1.17(0.96)} = \frac{150}{1.17(0.96)} = 128(156)$$

1.19 According to the Environmental Protection Agency, chloroform, which in its gaseous form is suspected to be a cancer-causing agent, is present in small quantities in all the country's 240,000 public water sources. If the mean and standard deviation of the amounts of chloroform present in water sources are 34 and 53 micrograms per liter (mcg/L), respectively, explain why chloroform amounts do not have a normal distribution.

L'écart-type dépasse largement la moyenne, différence est négative; -19, ce qui veut dire qu'à l'instar de la distribution du temps d'utilisation d'IN , moins d'un écart-type (12 %) de la moyenne a pour valeur effective nulle. La distribution virtuelle est normale, mais effectivement elle est tronquée, les valeurs en dessous de zéro n'ont dans cette étude pas une signification réelle. On peut évaluer la proportion d'effectif entre la moyenne et la limite zéro par simple rapport de proportionnalité, le résultat montre que la distribution n'est pas normale!

1.18 Les tests d'aptitude scolaire (SSAT) de raisonnement verbal et mathématique du College Board sont notés sur une échelle de 200 à 800. Il semble raisonnable de supposer que la distribution des scores suit approximativement une loi normale pour les deux tests. Utilisez le résultat de l'exercice 1.17 pour estimer l'écart type des scores au test de raisonnement verbal.

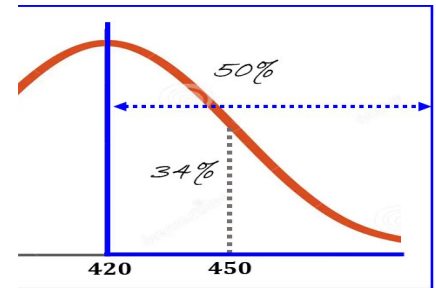
1.19 Selon l'Agence de protection de l'environnement, le chloroforme, suspecté d'être cancérigène à l'état gazeux, est présent en faibles quantités dans les 240 000 sources d'eau potable publiques du pays. Si la moyenne et l'écart type des concentrations de chloroforme dans ces sources d'eau sont respectivement de 34 et 53 microgrammes par litre (mcg/ L), expliquez pourquoi ces concentrations ne suivent pas une distribution normale.



1.20 Weekly maintenance costs for a factory, recorded over a long period of time and adjusted for inflation, tend to have an approximately normal distribution with an average of \$420 and a standard deviation of \$30 . If \$450 is budgeted for next week, what is an approximate probability that this budgeted figure will be exceeded?

Déterminons les intervalles correspondants aux proportions statistiques caractéristiques de la loi normale et situons la valeur de référence dans ces intervalles, on a respectivement pour 1 à 3 écart-types, $\bar{y} = 420$ $\sigma = 30$; [390:450] comportant 68 % des effectifs, [360:480] 95 % des effectifs, [330:510] 100 % Le premier demi-intervalle comprend les valeurs inférieures ou égales à 450 d'un écart-type de 30, donc 34 % des effectifs, la probabilité de dépassement se trouve au-delà de cette valeur. Comme 50 % des valeurs sont au-dessus de la moyenne, la probabilité de dépassement est la différence des deux , donc 16 %.

1.20 Les coûts hebdomadaires d'entretien d'une usine, enregistrés sur une longue période et ajustés en fonction de l'inflation, suivent généralement une loi normale, avec une moyenne de 420\$ et un écart type de 30\$. Si un budget de 450 \$ est prévu pour la semaine prochaine, quelle est la probabilité approximative que ce montant soit dépassé ?



1.21 The manufacturer of a new food additive for beef cattle claims that 80 % of the animals fed a diet including this additive should have monthly weight gains in excess of 20 pounds. A large sample of measurements on weight gains for cattle fed this diet exhibits an approximately normal distribution with mean 22 pounds and standard deviation 2 pounds. Do you think the sample information contradicts the manufacturer's claim? (Calculate the probability of a weight gain exceeding 20 pounds.)

Déterminons les intervalles correspondants aux proportions statistiques caractéristiques de la loi normale et situons la valeur de référence dans ces intervalles, on a respectivement pour 1 à 3 écart-types, $\bar{y} = 22$ $\sigma = 2$; [20:24] comportant 68 % des effectifs, [18:26] 95 % des effectifs, [16:28] 100 % Le premier demi-intervalle comporte les valeurs entre 20 et 22, donc 34 % des effectifs est supérieur à 20 jusqu'à la limite de 22 , au-delà de 22 on a la moitié supérieur à la moyenne. Donc en tout 34 + 50 = 84 %. l'affirmation du fabricant est donc statistiquement juste en se basant sur les données de mesures de prise de poids.

1.21 Le fabricant d'un nouvel additif alimentaire pour bovins affirme que 80 % des animaux nourris avec un régime contenant cet additif devraient présenter une prise de poids mensuelle supérieure à 9 kg (20 livres). Un large échantillon de mesures de prise de poids chez des bovins nourris avec ce régime présente une distribution approximativement normale, avec une moyenne de 10 kg (22 livres) et un écart type de 0,9 kg (2 livres). Pensez-vous que les données de l'échantillon contredisent l'affirmation du fabricant ? (Calculez la probabilité d'une prise de poids supérieure à 9 kg (20 livres).)

1.22 Prove that the sum of the deviations of a set of measurements about their mean is equal to zero; that

$$\sum_{i=0}^n (y_i - \bar{y}) = y_1 + y_2 + ..y_n - \bar{y} - \bar{y}.... = \sum_{i=0}^n y_i - n\bar{y} = \sum_{i=0}^n y_i - n * \frac{\sum_{i=0}^n y_i}{n} = 0$$

1.22 Démontrer que la somme des écarts d'un ensemble de mesures autour de leur moyenne est égale à zéro; c'est-à-dire

1.23 The mean duration of television commercials is 75 seconds with standard deviation 20 seconds. Assume that the durations are approximately normally distributed to answer the following. a What percentage of commercials last longer than 95 seconds? b What percentage of the commercials last between 35 and 115 seconds? c Would you expect commercial to last longer than 2 minutes? Why or why not?

1.23 La durée moyenne des publicités télévisées est de 75 secondes avec un écart type de 20 secondes. Supposons que les durées suivent approximativement une loi normale pour répondre aux questions suivantes. a) Quel pourcentage des publicités durent plus de 95 secondes ? b) Quel pourcentage des publicités durent entre 35 et 115 secondes ? c) Pensez-vous qu'une publicité puisse durer plus de 2 minutes ? Pourquoi ?

Solution On procède comme dans les exercices précédents , ici on va représenter les données sous forme de tableau

% statistique	intervalle1 68 %	intervalle2 95 %	intervalle3 100 %
moyenne	75	75	75
limites	[55:95]	[35:115]	[15:135]
s	20	40	60
estimation	$F(> 95) = F(> \bar{y}) - F(\bar{y} + s)$	$F(< 115, > 35) = F(\bar{y} \pm 2s)$	$F(>120)= F([120:135])$
résult. stat.	50-34 % = 16 %	$F(<115,>35)$	$<50-47.5 = 2.5 \%$

a) Le pourcentage au-delà de 95 secondes est donné par la différence de toutes fréquences au-dessus de la moyenne au premier demi-intervalle au-dessus de la moyenne , c'est à dire $50 - 68/2 = 16 \%$. b) Le pourcentage entre 35 et 115s correspond exactement à la fréquence du second intervalle 95 %. c) Le troisième intervalle comprends le délai au-delà de 120 secondes puisque sa limite est de 135s, mais la proportion est très faible et moins 2.5 %. Un écart-type constitue la proportion de durée comprise entre 115 et 135 sec, ce écart comporte la différence entre 50 % et 47.5 % des effectifs, 120 étant au-dessus de 115s, la proportion au-delà de 120s est nécessairement inférieure à cette valeur donc $< 2.5 \%$.

1.24 Aqua running has been suggested as a method of cardiovascular conditioning for injured athletes and others who desire a low-impact aerobics program. In a study to investigate the relationship between exercise cadence and heart rate,¹ the heart rates of 20 healthy volunteers were measured at a cadence of 48 cycles per minute (a cycle consisted of two steps). The data are as follows:

1.24 L'aquagym est proposée comme méthode de conditionnement cardiovasculaire pour les athlètes blessés et toute personne souhaitant un programme d'aérobic à faible impact. Dans une étude visant à examiner la relation entre la cadence d'exercice et la fréquence cardiaque¹, la fréquence cardiaque de 20 volontaires sains a été mesurée à une cadence de 48 cycles par minute (un cycle étant composé de deux foulées). Les résultats sont les suivants :

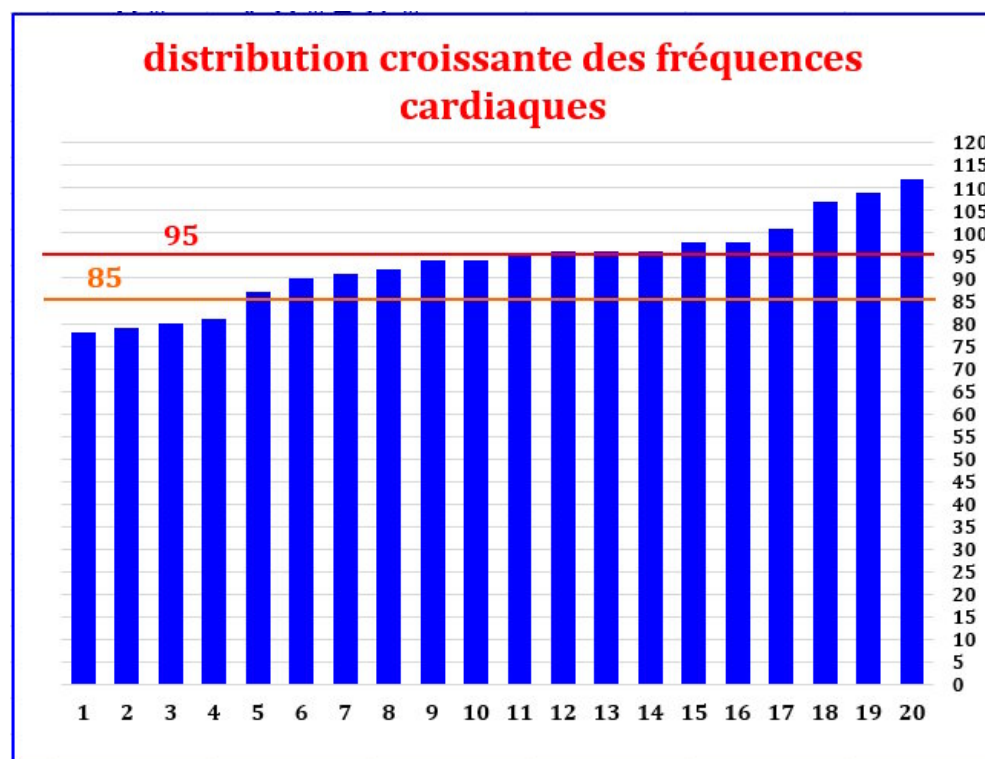
98	96	92	90	95	96	80	79	109	87	96	81	107	94	98	94	112	78	91	101
----	----	----	----	----	----	----	----	-----	----	----	----	-----	----	----	----	-----	----	----	-----

a Use the range of the measurements to obtain an estimate of the standard deviation. b Construct a frequency histogram for the data. Use the histogram to obtain a visual approximation to y and s. c Calculate y and s. Compare these results with the calculation checks provided by parts (a) and (b). d Construct the intervals $y \pm ks$, $k = 1, 2$, and 3 , and count the number of measurements falling in each interval. Compare the fractions falling in the intervals with the fractions that you would expect according to the empirical rule.

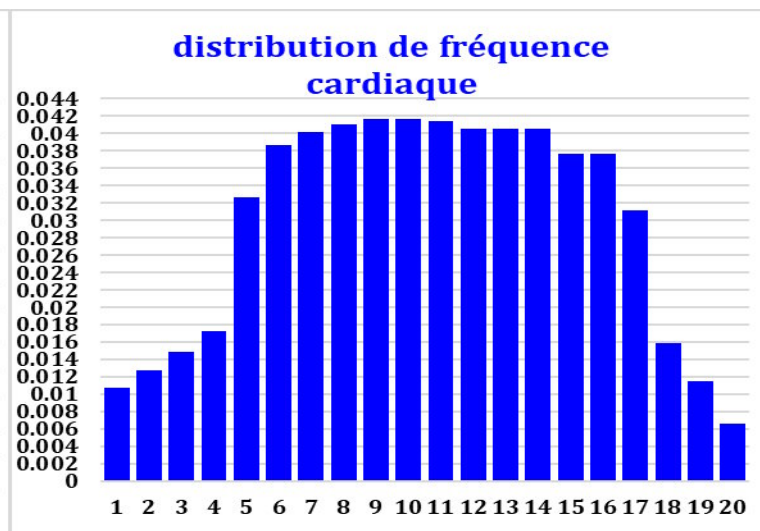
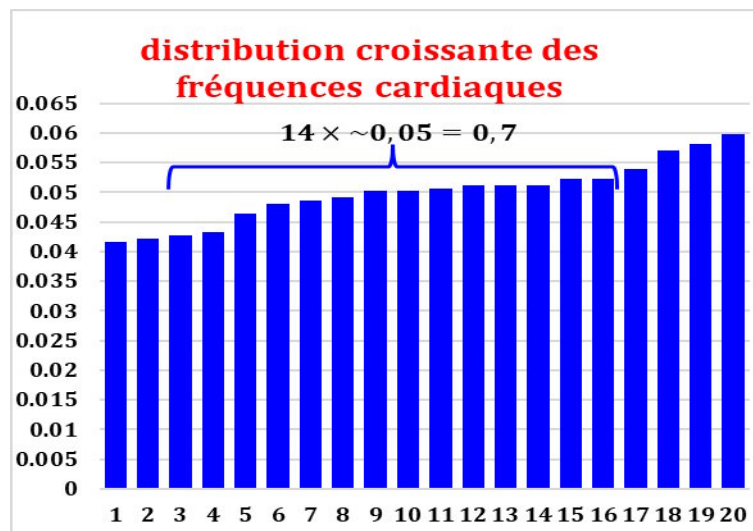
a) Utilisez l'étendue des mesures pour estimer l'écart type. b) Construisez un histogramme de fréquences pour les données. Utilisez cet histogramme pour obtenir une approximation visuelle de y et s. c) Calculez y et s. Comparez ces résultats avec les vérifications de calcul fournies dans la partie (a). d. Construisez les intervalles $y \pm ks$, $k = 1, 2$ et 3 , et comptez le nombre de mesures appartenant à chaque intervalle. Comparez les fractions de mesures appartenant aux intervalles avec les fractions attendues selon la règle empirique.

Solution a) b) On procède comme dans un des exercices précédents, on calcule le quart de l'étendue qui peut approximer l'écart-type. On effectue un tri des valeurs, détermine moyenne et écart-type visuellement et par formule de la distribution $f(x)$ au différentes valeurs afin de répartir les fréquences dans la forme normale.

	vis./calc.	calcul		vis./calc.	calcul		vis./calc.	calcul
étend.	34	34	I1 (1/4e)	[86.5:103.5] (13, 65 %)	[85.2:102.2] (13, 65 %)	I1 (s)	[88:102] (12, 60 %)	[84.15:103.25] (13, 60 %)
1/4 d'ét.	8.5	8.5	I2 (1/4e)	[78:112] (20, 100 %)	[76.7:110.7] (19, 95 %)	I2 (s)	[81:109] (16, 80 %)	[74.59:112.81] (19, 95 %)
moy.	95	93.7	I3 (1/4e)	[69.5:120.5] (20, 100 %)	[68.2:119.2] (20, 100 %)	I3 (s)	[74:116] (20,100 %)	[65.04:122.36] (20,100 %)
s	7 (14/2)	9.55						



Les calculs de moyenne et d'écart-type conduisent à des distributions de fréquences ; 60 %, 95 % et 100%. Le premier intervalle s'écarte de la valeur canonique de 68 %, l'écart-type approximé par le 1/4 d'étendue par rapport à la moyenne visuelle ou de calcul donne une valeur plus proche 65 %. Certaines valeurs sont proches des limites, le nombre de variable est faible avec des une tendance centrale forte, ce qui explique cet écart important pour le premier intervalle. la combinaison d'évaluation visuelle avec le calcul peut donner des valeurs peu cohérentes (I2 (s) 80 %) en raison d'un cumul d'écarts.



1.25 The following data give the lengths of time to failure for n = 88 radio transmitter-receivers: a Use the range to approximate s for the n = 88 lengths of time to failure. b Construct a frequency histogram for the data. [Notice the tendency of the distribution to tail outward (skew) to the right.] c Use a calculator (or computer) to calculate y and s. (Hand calculation is much too tedious for this exercise.) d Calculate the intervals $y \pm ks$, $k = 1, 2$, and 3 , and count the number of measurements falling in each interval. Compare your results with the empirical rule results. Note that the empirical rule provides a rather good description of these data, even though the distribution is highly skewed.

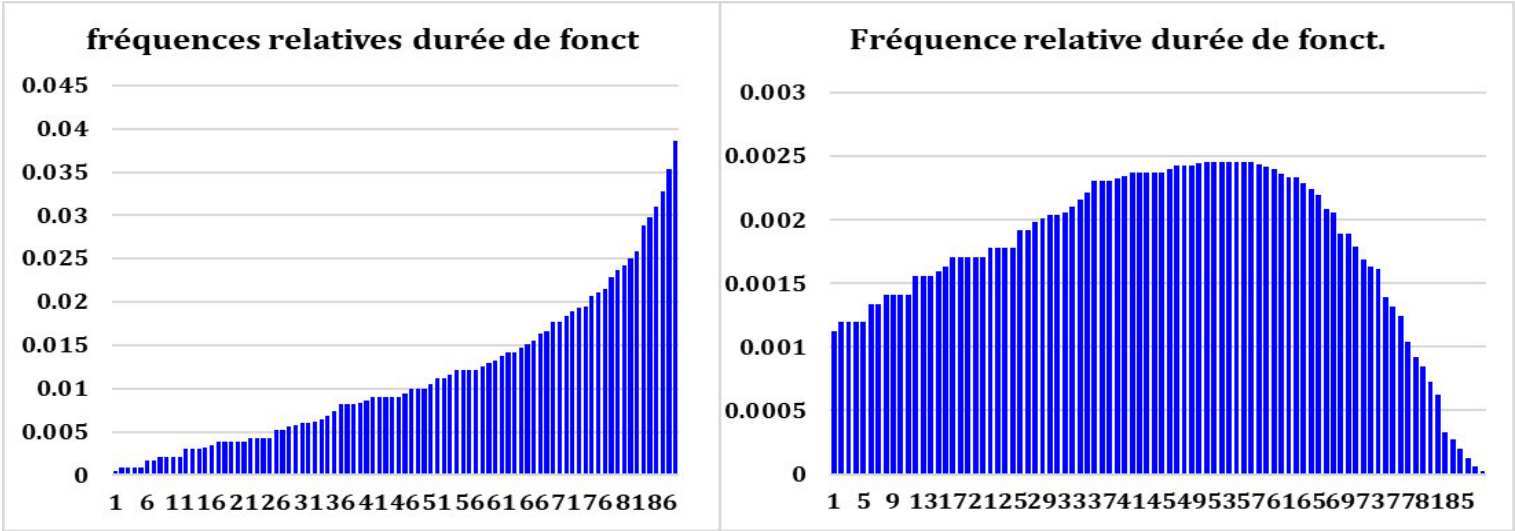
1.25 Les données suivantes indiquent les durées de fonctionnement avant défaillance pour n = 88 émetteurs-récepteurs radio : a) Utilisez l'étendue pour estimer s pour les n = 88 durées de temps avant défaillance. b) Construisez un histogramme de fréquences pour ces données. [Remarquez la tendance de la distribution à queue vers l'extérieur (oblique) vers la droite.] c Utilisez une calculatrice (ou un ordinateur) pour calculer y et s. (Le calcul manuel est beaucoup trop fastidieux.) (pour cet exercice.) d) Calculez les intervalles $y \pm ks$, $k = 1, 2$ et 3 , et comptez le nombre de mesures appartenant à chaque intervalle. Comparez vos résultats avec ceux obtenus par la règle empirique. Notez que cette dernière décrit assez bien ces données, malgré une distribution fortement asymétrique.

16	224	16	80	96	536	400	80
392	576	128	56	656	224	40	32
358	384	256	246	328	464	448	716
304	16	72	8	80	72	56	608
108	194	136	224	80	16	424	264
156	216	168	184	552	72	184	240
438	120	308	32	272	152	328	480
60	208	340	104	72	168	40	152
360	232	40	112	112	288	168	352
56	72	64	40	184	264	96	224
168	168	114	280	152	208	160	176

Solution a) Les valeurs extrêmes sont 716 et 8 , donc le quart d'étendue approximant $s = 177$. b) excel donne pour y et s :210.8 et 162.17

moy	approx 1/4 d'ét	calcul s	ER ./	canon
210.8	177	162.17		
I1	[48.62:372.97] 71.5 %	[33.8:387.8] 77.2 %	5 %	14 %
I2	[-113.55:535.14] 93.2 %	[-143.2:564.8] 95.5 %	2 %	0 %
I2	[-275.72:697.31] 99 %	[-320.2:741.8] 100 %	0 %	1 %

a à nouveau dans ce cas, un premier intervalle un peu éloigné des 68 % canoniques, ce qui se justifie par la forme asymétrique de la distribution, mais l'écart selon le calcul n'est que de 5 %.



1.26 Compare the ratio of the range to s for the three sample sizes ($n = 6, 20$, and 88) for Exercises 1.12, 1.24, and 1.25. Note that the ratio tends to increase as the amount of data increases. The greater the amount of data, the greater will be their tendency to contain a few extreme values that will inflate the range and have relatively little effect on s . We ignored this phenomenon and suggested that you use 4 as the ratio for finding a guessed value of s in checking calculations.

Solution En effet, le modèle d'approximation du quart d'étendue comme approximation de s repose sur une approximation numérique l'intervalle fait de deux fois deux écarts-types autour de la moyenne comporte la presque totalité des valeurs (95 %) , donc il paraît logique que s équivaut à un peu moins d' 1/4 de l'étendue des valeurs. Mais effectivement, on peut avoir un ou deux valeurs extrêmes qui faussent cette approximation d'autant plus que le nombre de données est faible.

1.26 Comparez le rapport entre l'étendue et l'écart-type pour les trois tailles d'échantillon ($n = 6, 20$ et 88) des exercices 1.12, 1.24 et 1.25. Notez que ce rapport tend à augmenter avec la quantité de données. Plus les données sont nombreuses, plus elles auront tendance à contenir quelques valeurs extrêmes qui augmenteront l'étendue et auront relativement peu d'influence sur l'écart-type. Nous avons négligé ce phénomène et suggéré d'utiliser 4 comme rapport pour estimer l'écart-type lors de la vérification des calculs.

n	6	20	88
1/4 Et.	0.75	8.5	177
s	1.2	9.55	162.17
ET/s	0.625	0.89	1.09
ER Et/s	-38 %	-11 %	9 %
ER s/ET	60 %	12 %	-8 %

1.27 A set of 340 examination scores exhibiting a bell-shaped relative frequency distribution has a mean of $\mu = 72$ and a standard deviation of $\sigma = 8$. Approximately how many of the scores would you expect to fall in the interval from 64 to 80? The interval from 56 to 88?

Solution Créons les intervalles théoriques. Les intervalles investigués correspondent exactement aux 2 premiers intervalles canoniques. Le nombre de note est donné par le produit des pourcentages par le nombre total, donc respectivement de : 231 et 323.

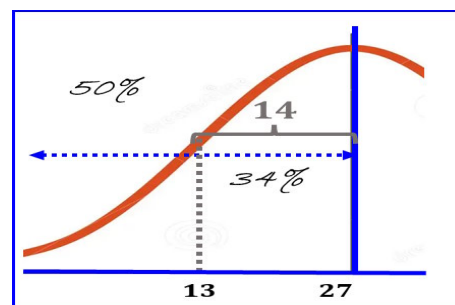
1.27 Un ensemble de 340 notes d'examen présentant une distribution de fréquences relatives en forme de cloche a une moyenne de $\mu = 72$ et un écart type de $\sigma = 8$. Combien de notes environ se situeraient dans l'intervalle de 64 à 80 ? Et dans l'intervalle de 56 à 88 ?

	II1	II2	II3
théor	[64;80]	[56;88]	[48;96]
investig.	[64;80]	[56;88]	
%	68 %	95 %	99 %

1.28 The discharge of suspended solids from a phosphate mine is normally distributed with mean daily discharge 27 milligrams per liter (mg/L) and standard deviation 14 mg/L. In what proportion of the days will the daily discharge be less than 13 mg/L?

Solution L'intervalle calculé à partir de la moyenne de 27 et d'écart-type 14 comportant le 68 % de l'effectif est [13;41]. 50 % des valeurs se trouvent en-dessous de la moyenne de 27, donc l'effectif en dessous de 13 est la différence entre 50 % et 1/2 68 % = 16 %.

1.28 Les rejets de matières en suspension provenant d'une mine de phosphate suivent une loi normale, avec un rejet journalier moyen de 27 milligrammes par litre (mg/L) et un écart type de 14 mg/L. Dans quelle proportion des jours le rejet journalier sera-t-il inférieur à 13 mg/L ?



1.29 A machine produces bearings with mean diameter 3.00 inches and standard deviation 0.01 inch. Bearings with diameters in excess of 3.02 inches or less than 2.98 inches will fail to meet quality specifications. a) Approximately what fraction of this machine's production will fail to meet specifications? b) What assumptions did you make concerning the distribution of bearing diameters in order to answer this question?

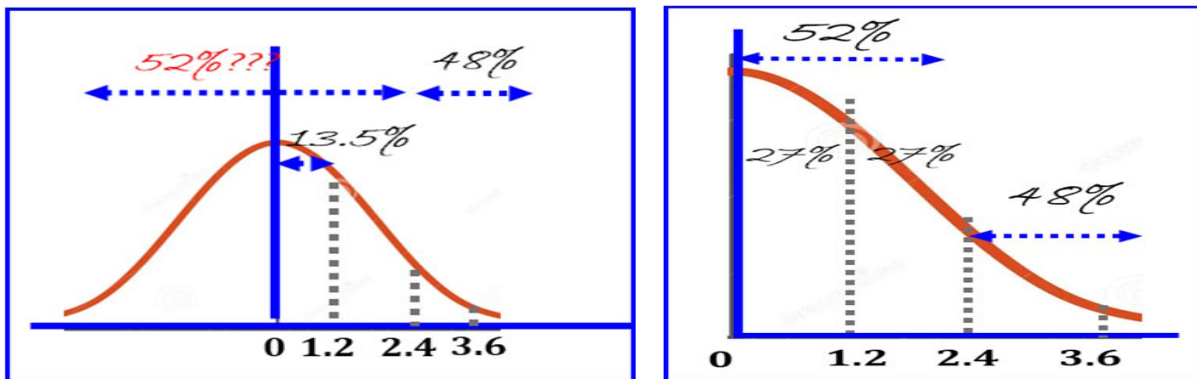
1.29 Une machine produit des roulements avec un diamètre moyen de 3,00 pouces et un écart type de 0,01 pouce. Les roulements dont le diamètre est supérieur à 3,02 pouces ou inférieur à 2,98 pouces ne répondront pas aux spécifications de qualité. a) Quelle proportion approximative de la production de cette machine ne sera pas conforme aux spécifications ? b) Quelles hypothèses avez-vous formulées concernant la distribution des diamètres de roulements pour répondre à cette question ?

Solution a) b) Seuls les roulements de diamètres limites inclus dans l'intervalle [2.98;3.02] répondent aux spécifications de qualité, intervalle formé par la moyenne de rayon ± 2 écarts-types représentant donc 95 % de l'ensemble. Ce qui signifie que la différence du total de l'effectif de l'échantillonnage ne répond pas au critère de qualité, donc 5 % ne seront pas conformes. Comme cette estimation repose sur la conformité d'une dispersion réelle à celle théorique d'une distribution normale, on ne peut pas avoir une précision de cette proportion que dans la mesure où cette conformité soit mesurable. On a vu dans l'ex. précédent que l'écart-type se rapproche du quart d'étendue lorsque la taille de l'échantillon augmente et l'écart des valeurs extrêmes a peu d'influence sur s . Si on compare l'écart relatif entre deux écarts-types comportant ou non le dernier échantillon, prenons la distribution de la durée de fonctionnement des émetteurs à 88 échantillons de l'exercice 1.25 (5.5 %, différence 60, étendue = 708 s = 162.1, 153.2) contre la distribution des fréquences cardiaques de l'exercice 1.24 à 20 échantillons, (9 % différence 3, étendue 34 étendue s = 9.55, 8.76) , on constate que cet écart relatif passe de 5 à 9, pour deux séries dont le nombre d'échantillons diminue de plus d'un quart et dont le rapport entre la différence des deux derniers échantillon et l'étendue est quasiment similaire ($60/160 = 8.5\%$, $3/34 = 8.8\%$). Il est évident que l'on ne peut pas comparer 2 séries dont l'une comporte 1 écart extrême.

1.30 Compared to their stay-at-home peers, women employed outside the home have higher levels of high-density lipoproteins (HDL), the "good" cholesterol associated with lower risk for heart attacks. A study of cholesterol levels in 2000 women, aged 25–64, living in Augsburg, Germany, was conducted by Ursula Haertel, Ulrich Keil, and colleagues² at the GSF-Medis Institut in Munich. Of these 2000 women, the 48 % who worked outside the home had HDL levels that were between 2.5 and 3.6 milligrams per deciliter (mg/dL) higher than the HDL levels of their stay-at-home counterparts. Suppose that the difference in HDL levels is normally distributed, with mean 0 (indicating no difference between the two groups of women) and standard deviation 1.2 mg/dL. If you were to select an employed woman and a stay-at-home counterpart at random, what is the probability that the difference in their HDL levels would be between 1.2 and 2.4?

1.30 Comparées aux femmes au foyer, les femmes qui travaillent à l'extérieur de leur domicile présentent des taux plus élevés de lipoprotéines de haute densité (HDL), le « bon » cholestérol associé à un risque moindre de crise cardiaque. Une étude portant sur les taux de cholestérol de 2 000 femmes âgées de 25 à 64 ans vivant à Augsburg, en Allemagne, a été menée par Ursula Haertel, Ulrich Keil et leurs collègues² à l'Institut GSF-Medis. Munich. Parmi ces 2 000 femmes, les 48 % qui travaillaient à l'extérieur de leur domicile présentaient des taux de HDL supérieurs de 2,5 à 3,6 mg/dL à ceux des femmes au foyer. Supposons que la différence de taux de HDL suive une loi normale, de moyenne 0 (indiquant l'absence de différence entre les deux groupes de femmes) et d'écart-type 1,2 mg/dL. Si l'on choisissait au hasard une femme active et une femme au foyer, quelle est la probabilité que la différence de leurs taux de HDL soit comprise entre 1,2 et 2,4 mg/dL ?

Solution Avec la supposition d'une distribution normale de moyenne 0 pour la différence entre les deux catégories de femme et un écart-type de 1.2 mg/dL, on a théoriquement les deux intervalles $[-1.2, 1.2]$ et $[-2.4, 2.4]$ avec les proportions théoriques de 68 % et 95 % des effectifs. Ici les valeurs négatives ont un sens, les femmes au foyer peuvent théoriquement avoir un taux de HDL plus faible. La différence de ces deux plages est de $95 - 68 = 27$ %. Toutefois tenant compte de l'hypothèse que les femmes au foyers peuvent avoir un taux supérieur, on ne doit donc prendre la moitié de cette valeur correspondant aux différences positives, donc 13.5 %. Mais si l'on présume que le taux de HDL des femmes au foyer ne peut être plus élevé que ceux de celles travaillant à l'extérieur, et que cette hypothèse est strictement validée par des mesures, on doit imaginer que la distribution de la différence débute avec la valeur 0 jusqu'à une valeur extrême par exemple 3.6 et qu'elle est tronquée à gauche. En effet, on a pour information que 48 % des femmes travaillant à l'extérieur ont un taux > 2.5 , cela signifie que 52 % ont un taux inférieur à 2.5 dont 27 % supérieur à 1.2, donc 27 % sont dans l'intervalle de $[0; 1.2]$. On peut poser deux cas de figures, une distribution centrée sur 0 comme le propose l'énoncé de l'exercice et celle-ci admettant une symétrie ou non. Une symétrie infirmerait toutefois l'affirmation que le taux des femmes de l'extérieur est plus élevé, puisqu'on aurait théoriquement la même différence inverse et ceci contredit les résultats de l'étude. De plus d'après les données expérimentales 52 % des femmes de l'extérieur ont un taux inférieur à 2.5, donc il est hautement improbable qu'au plus 50% aient un taux inférieur à ceux des femmes au foyer, cela signifierait que 5 écarts-type englobe cet intervalle et qu'un écart serait de 10 %, ce qui contredit le résultat trouvé et la distribution ne serait pas normale mais toute différente. Donc une asymétrie semble plus plausible.



1.31 Over the past year, a fertilizer production process has shown an average daily yield of 60 tons with a variance in daily yields of 100. If the yield should fall to less than 40 tons tomorrow, should this result cause you to suspect an abnormality in the process? (Calculate the probability of obtaining less than 40 tons.) What assumptions did you make concerning the distribution of yields?

Solution La variance étant le carré de l'écart-type, on a les intervalles suivant correspondant à 86, 95 et 99 %; $[50; 70]$, $[40; 80]$, $[30; 90]$. Un rendement de moins de 40 tonnes signifie une différence de plus de deux écarts-types en dessous de la moyenne, donc inférieur à la moitié de 100 - 95 %, puisque l'on ne prend que l'effectif inférieur à 40 qui équivaut à celui de plus de 80, donc < 2.5 %. Cette valeur suppose une parfaite symétrie entre les valeurs supérieures et inférieures, à priori cette distribution est statistiquement possible puisqu'on peut même constituer un intervalle de 4 écarts-types avec les valeurs limites supérieures à 0, répondant à la logique quantitative de la production, $[0; 100]$. Il ne demeure plus que la distribution soit canonique, ce qu'un échantillonnage sur une longue durée permettra de valider.

1.31 Au cours de l'année écoulée, un procédé de production d'engrais a affiché un rendement journalier moyen de 60 tonnes, avec une variance de 100 tonnes. Si le rendement devait chuter à moins de 40 tonnes demain, devriez-vous suspecter une anomalie dans le procédé ? (Calculez la probabilité d'obtenir moins de 40 tonnes.) Quelles hypothèses avez-vous formulées concernant la distribution des rendements ?

1.32 Let $k \geq 1$. Show that, for any set of n measurements, the fraction included in the interval $y - ks$ to $y + ks$ is at least $(1 - 1/k^2)$.

[Hint: $s^2 = \frac{1}{n-1} \sum_{i=0}^n (y_i - \bar{y})^2$

In this expression, replace all deviations for which $|y_i - \bar{y}| \geq ks$ with ks . Simplify.] This result is known as Tchebysheff's theorem.

Solution La probabilité qu'une variable aléatoire tombe dans un intervalle donné égale l'intégrale de la densité de probabilité ou si l'on préfère la somme de fréquences de probabilité sur l'ensemble d'intervalle de valeurs lorsque ceux-ci tendent vers zéro. Il suffit donc de remplacer les bornes par celles de nos différents intervalles et d'obtenir les fractions de distribution en fonction des autres paramètres que sont x, k après avoir intégré la fonction.

$$\begin{aligned} P(\alpha < x < \beta) &= \int_{\alpha}^{\beta} f(x) dx = \int_{\bar{x}-k\sigma}^{\bar{x}+k\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx \stackrel{t=\frac{x-\bar{x}}{\sqrt{2}\sigma}}{=} \int_{-\frac{k}{\sqrt{2}}}^{\frac{k}{\sqrt{2}}} \sqrt{2}\sigma \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2} dt \\ &= \frac{1}{\sqrt{\pi}} \int_{-\frac{k}{\sqrt{2}}}^{\frac{k}{\sqrt{2}}} e^{-t^2} dt = \frac{1}{\sqrt{\pi}} \int_{-\frac{k}{\sqrt{2}}}^{\frac{k}{\sqrt{2}}} (1 - t^2 + \frac{t^4}{2} - \frac{t^6}{6} + \dots) dt = \frac{1}{\sqrt{\pi}} (t - \frac{t^3}{3} + \frac{t^5}{10} - \frac{t^7}{42} + \dots) \Big|_{-\frac{k}{\sqrt{2}}}^{\frac{k}{\sqrt{2}}} \\ &= \frac{1}{\sqrt{\pi}} \left(\frac{k}{\sqrt{2}} - \frac{k^3}{6\sqrt{2}} + \frac{k^5}{40\sqrt{2}} - \frac{k^7}{336\sqrt{2}} + \dots - \left(-\frac{k}{\sqrt{2}} + \frac{k^3}{6\sqrt{2}} - \frac{k^5}{40\sqrt{2}} + \frac{k^7}{336\sqrt{2}} - \dots \right) \right) \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \left(k - \frac{k^3}{6} + \frac{k^5}{40} - \frac{k^7}{336} + \frac{k^9}{3456} \dots \right) \begin{cases} \stackrel{k=1}{=} \sqrt{\frac{2}{\pi}} \left(1 - \frac{1}{6} + \frac{1}{40} - \frac{2}{336} + \frac{1}{3456} \dots \right) = 0.68 \\ \stackrel{k=2}{=} \sqrt{\frac{2}{\pi}} \left(2 - \frac{8}{6} + \frac{32}{40} - \frac{128}{336} + \frac{512}{3456} \dots \right) = 0.98 \\ \stackrel{k=3}{=} \sqrt{\frac{2}{\pi}} \left(3 - \frac{27}{6} + \frac{243}{40} - \frac{2187}{336} + \frac{19683}{3456} \dots \right) = 3 \end{cases} \end{aligned}$$

On a pour les seconds et troisièmes intervalles des valeurs inexactes, il faudrait intégrer plus de termes. On tente toute de même de montrer que pour les différentes valeurs de $k = \frac{\sqrt{2}}{\sqrt{\pi}} \left(k - \frac{k^3}{6} + \frac{k^5}{40} \dots \right) \geq 1 - \frac{1}{k^2}$ en admettant que l'on a retrouvé les proportions canoniques, mais l'inégalité mathématique n'est cependant pas rigoureusement démontrée.

$$\begin{cases} 68\% = \sqrt{\frac{2}{\pi}} \left(1 - \frac{1}{6} + \frac{1}{40} - \frac{2}{336} + \frac{1}{3456} \dots \right) \geq 0 \\ 95\% = \sqrt{\frac{2}{\pi}} \left(2 - \frac{8}{6} + \frac{32}{40} - \frac{128}{336} + \frac{512}{3456} \dots \right) \geq \frac{3}{4} = 0.75 \\ 99\% = \sqrt{\frac{2}{\pi}} \left(2 - \frac{8}{6} + \frac{32}{40} - \frac{128}{336} + \frac{512}{3456} \dots \right) \geq \frac{8}{9} = 0.88 \end{cases}$$

Considérons une série de valeurs triées et regroupées en 3 parties, de sorte que l'intervalle central formé de la différence absolue de la variable à l'écart-type soit égale à ks , les intervalles extérieurs auront nécessairement une différence inférieure. On peut donc écrire en considérant la suggestion de l'auteur, p, q, n étant les effectifs :

$$s^2 = \frac{1}{n-1} \left(\sum_{l=0}^a (y_l - \bar{y})^2 + \sum_{m=a}^b (y_m - \bar{y})^2 + \sum_{p=b}^c (y_p - \bar{y})^2 \right) \stackrel{\substack{|y_m - \bar{y}| = ks \\ |y_{\neq m} - \bar{y}| \leq ks}}{=} \frac{(pk^2 s^2 + qk^2 s^2)}{n-1} \stackrel{p+q=n}{=} \frac{n}{n-1} k^2 s^2$$

1.32 ? (actuellement irrésolu) Soit $k \geq 1$. Montrer que, pour tout ensemble de n mesures, la fraction incluse dans l'intervalle $y - ks$ à $y + ks$ est au moins $(1 - 1/k^2)$.

[Indice : $s^2 = \frac{1}{n-1} \sum_{i=0}^n (y_i - \bar{y})^2$

Dans cette expression, remplacez tous les écarts pour lesquels $|y_i - \bar{y}| \geq ks$ par ks . Simplifiez.] Ce résultat est connu sous le nom de **théorème de Tchebychev***.

$$\text{Or } f = \sum_{m=a}^b (p_m) \approx \frac{p}{n-1} \Rightarrow s^2 \leq k^2 s^2 (f + 1 - f) = k^2 s^2 \Rightarrow k > 1$$

Et l'on doit aboutir à

$$f = P(\bar{x} - k\sigma < x < \bar{x} + k\sigma) \geq 1 - \frac{1}{k^2} \quad \text{OU} \quad k^2 \leq \frac{1}{1-f} \text{????}$$

L'inégalité de Markov stipule que pour 1 variable aléatoire positive X et 1 seuil $a > 0$, $P(X \geq a) \leq \frac{E(X)}{a}$.

Démonstration basée sur le fait que $X \geq a \cdot 1_{X \geq a}$, ($1_{X \geq a}$ fonction indicatrice de l'événement $\{X \geq a\}$).

Soit $X > 0$ et $E(X)$. Pour tout $a > 0$, considérons la variable aléatoire $Y = a \cdot 1_{X \geq a}$.

Si $X \geq a$, alors $X \geq a$ et $1_{X \geq a} = 1$, donc $Y = a$. On a $X \geq a = Y$.

Si $X < a$, alors $1_{X \geq a} = 0$, donc $Y = 0$. On a $X \geq 0 = Y$.

Dans tous les cas, on a $X \geq Y = a \cdot 1_{X \geq a}$.

Puisque $X \geq Y$, on a $E(X) \geq E(Y)$.

$E(Y) = E(a \cdot 1_{X \geq a}) = a \cdot E(1_{X \geq a})$. L'espérance d'une variable indicatrice est égale à la probabilité de l'événement qu'elle représente, donc $E(1_{X \geq a}) = P(X \geq a)$. Ainsi, $E(Y) = a \cdot P(X \geq a)$.

$E(X) \geq E(Y) = a \cdot P(X \geq a)$. Comme $a > 0$, on peut diviser par a pour obtenir l'inégalité de Markov : $P(X \geq a) \leq \frac{E(X)}{a}$

La démonstration du théorème de Tchebychev (inégalité de Bienaymé-Tchebychev) part de l'inégalité de Markov qui stipule que pour toute $Y > 0$ et toute constante $b > 0$, on a $P(Y \geq b) \leq \frac{E(Y)}{b}$. On applique cette inégalité à la variable aléatoire positive $Y = (X - E(X))^2$ et à la constante $b = a^2$ (où $a > 0$).

L'espérance de la variable Y est l'espérance de $(X - E(X))^2$, qui est par définition la variance de X , soit $V(X)$. En substituant dans l'inégalité de Markov, on obtient : $P((X - E(X))^2 \geq a^2) \leq \frac{V(X)}{a^2}$.

Puisque $|X - E(X)| \geq a$ est équivalent à $(X - E(X))^2 \geq a^2$ (car $a > 0$), on peut remplacer le terme de gauche par $P(|X - E(X)| \geq a)$ pour obtenir l'inégalité finale : $P(|X - E(X)| \geq a) \leq \frac{V(X)}{a^2}$

1.33 A personnel manager for a certain industry has records of the number of employees absent per day. The average number absent is 5.5, and the standard deviation is 2.5. Because there are many days with zero, one, or two absent and only a few with more than ten absent, the frequency distribution is highly skewed. The manager wants to publish an interval in which at least 75 % of these values lie. Use the result in Exercise 1.32 to find such an interval.

Solution Avec 1s de 2.5 et une moyenne de 5.5, on a des intervalles larges, [3:8] qui constituent 86 % de probabilité d'absence et [0.5:10.5] 95 %. L'inégalité proposé dans l'ex. précédent permet de déterminer l'intervalle recherché; $f = 0.75 = \frac{3}{4} \geq 1 - \frac{1}{k^2} \Rightarrow k \leq 2$ l'intervalle de confiance est donc [0.5 : 10.5] à 95 % du maximum ; le résultat obtenu est conforme aux attentes que l'intervalle se situe à 2 écart de la moyenne au plus.

1.34 For the data discussed in Exercise 1.33, give an upper bound to the fraction of days when there are more than 13 absentees.

Solution L'intervalle de 3 écart-type est donné par la moyenne 5.5 plus ou moins 3s (7.5), on a théoriquement [-2.5:13] , effectivement [0:13] puisque les valeurs négatives n'ont pas ici de sens. Cet intervalle représente 99 % de probabilité, et 13 est la borne limite au-delà de laquelle le nombre d'absence dépasse 13. L'effectif dépassant 13 absence serait inclus dans l'intervalle théorique qui comprendrait 4s et représente moins de 1 % de probabilité, c'est à dire < 0.5 % puisque seul le dépassement de la borne supérieur caractérise cet effectif.

1.35 A pharmaceutical company wants to know whether an experimental drug has an effect on systolic blood pressure. Fifteen randomly selected subjects were given the drug and, after sufficient time for the drug to have an impact, their systolic blood pressures were recorded.

172	140	123	130	115	148	108	129	137	161	123	152	133	128	142
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

a Approximate the value of s using the range approximation. b Calculate the values of y and s for the 15 blood pressure readings. c Use Tchebysheff's theorem (Exercise 1.32) to find values a and b such that at least 75 % of the blood pressure measurements lie between a and b. d Did Tchebysheff's theorem work? That is, use the data to find the actual percent of blood pressure readings that are between the values a and b you found in part (c). Is this actual percentage greater than 75 %?

1.33 Un responsable des ressources humaines d'une entreprise dispose des données relatives au nombre d'employés absents chaque jour. La moyenne d'absences est de 5,5, avec un écart type de 2,5. La distribution des fréquences est fortement asymétrique, car de nombreux jours comptent zéro, un ou deux employés absents, tandis que seuls quelques-uns en comptent plus de dix. Le responsable souhaite définir un intervalle de confiance contenant au moins 75 % de ces valeurs. Utilisez le résultat de l'exercice 1.32 pour déterminer cet intervalle.

1.34 Pour les données présentées dans l'exercice 1.33, donnez une borne supérieure à la fraction de jours où il y a plus de 13 absents.

1.35 Une entreprise pharmaceutique souhaite déterminer si un médicament expérimental a un effet sur la pression artérielle systolique. Quinze sujets sélectionnés au hasard ont reçu le médicament et, après un délai suffisant pour que celui-ci agisse, leur pression artérielle systolique a été mesurée.

a) Estimez la valeur de s par approximation de l'étendue. b) Calculez les valeurs de y et s pour les 15 mesures de pression artérielle. c) Utilisez le théorème de Tchebychev (exercice 1.32) pour trouver les valeurs de a et b telles qu'au moins 75 % des mesures de pression artérielle se situent entre a et b. d) Le théorème de Tchebychev s'est-il vérifié ? Autrement dit, utilisez les données pour déterminer le pourcentage réel de mesures de pression artérielle comprises entre les valeurs a et b trouvées en c). Ce pourcentage est-il supérieur à 75 %

Solution

a) b) c) le quart de l'étendue est $64/4 = 16$, la moyenne est de 136 et $s=17.1$

On procède comme dans l'exercice 1.33, on détermine k en fonction de la fraction se situant à au moins de 75 %, donc également :

$$f = 0.75 = \frac{3}{4} \geq 1 - \frac{1}{k^2} \Rightarrow k \leq 2 \text{ l'intervalle de confiance est donc } [101.86 : 170.27]$$

d) On trouve par ailleurs que 14/15 valeurs se trouvent dans cet intervalle, c'est à dire 93.3 %, le pourcentage est supérieur à 75 % et à peine inférieur à 95 % correspondant à un intervalle à 2 écart-type. Le théorème est bien vérifié . Ici, on peut tout aussi bien prendre pour écart-type le quart d'étendue , on obtient le même résultat.

1.36 A random sample of 100 foxes was examined by a team of veterinarians to determine the prevalence of a specific parasite. Counting the number of parasites of this specific type, the veterinarians found that 69 foxes had no parasites of the type of interest, 17 had one parasite of the type under study, and so on. A summary of their results is given in the following table:

Nbr de parasites	0	1	2	3	4	5	6	7	8
Nbr de renards	69	17	6	3	1	2	1	0	1

a Construct the relative frequency histogram for the number of parasites per fox. b Calculate y and s for the data given. c What fraction of the parasite counts falls within 2 standard deviations of the mean? Within 3 standard deviations? Do your results agree with Tchebysheff's theorem (Exercise 1.32) and/or the empirical rule?

Solution

a) b) c) $n = 9$ $y = 0.2233$ $s = 0.1111$, il n'y a qu'une valeur 0 en dessous de 2 écarts-types $[0.0011:44]$, ce qui représente 11 % de fréquence, mais aucune à moins de 3 s.

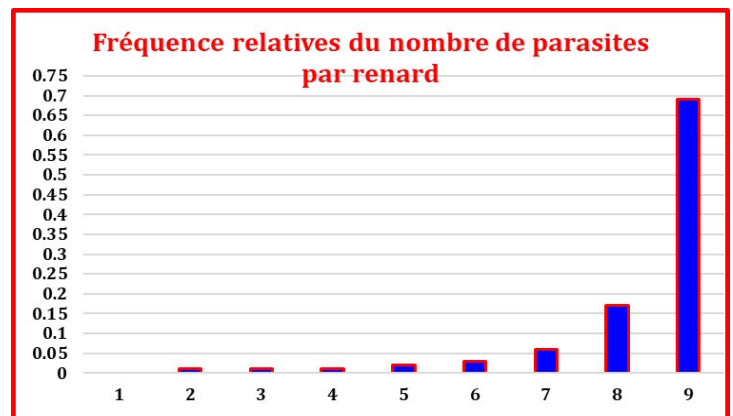
Si l'on applique la formule de Tchebychev à l'intervalle à 2s, on a effectivement:

$$f(k = 2) = 0.95 \geq 1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75$$

Par contre 11 % de l'effectif est en dessous de l'intervalle, alors que l'on devrait avoir moins de 2.5 % en dessous cet intervalle.

1.36 Un échantillon aléatoire de 100 renards a été examiné par une équipe de vétérinaires afin de déterminer la prévalence d'un parasite spécifique. En comptant le nombre de parasites de ce type précis, les vétérinaires ont constaté que 69 renards n'en présentaient aucun, tandis que 17 en présentaient un type étudié, etc. Un résumé de leurs résultats est présenté dans le tableau suivant :

a) Construisez l'histogramme des fréquences relatives du nombre de parasites par renard. b) Calculez y et s pour les données fournies. c) Quelle proportion du nombre de parasites se situe à moins de 2 écarts-types de la moyenne ? À moins de 3 écarts-types ? Vos résultats sont-ils en accord avec le théorème de Tchebychev (exercice 1.32) et/ou la règle empirique ?

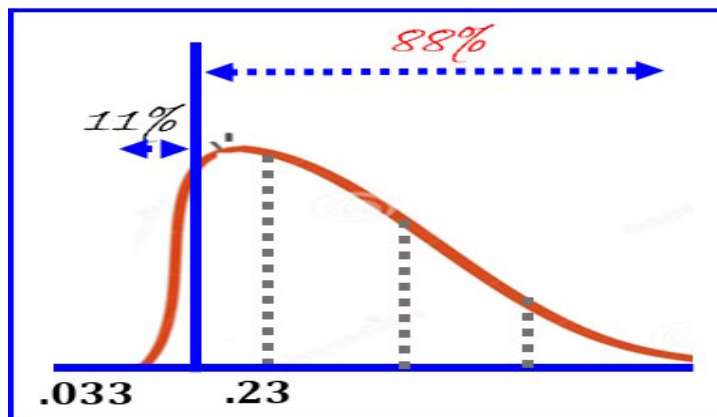


1.37 Studies indicate that drinking water supplied by some old lead-lined city piping systems may contain harmful levels of lead. Based on data presented by Karalekas and colleagues,⁴ it appears that the distribution of lead content readings for individual water specimens has mean .033 mg/L and standard deviation .10 mg/L. Explain why it is obvious that the lead content readings are not normally distributed.

1.37 Des études indiquent que l'eau potable distribuée par certains anciens réseaux de canalisations urbaines revêtus de plomb peut contenir des niveaux de plomb nocifs. D'après les données présentées par Karalekas et ses collègues, la distribution des concentrations de plomb dans les échantillons d'eau individuels présente une moyenne de 0,033 mg/L et un écart type de 0,10 mg/L. Expliquez pourquoi il est évident que les concentrations de plomb ne suivent pas une distribution normale.

Solution

L'écart-type est trois fois la valeur de la moyenne, donc théoriquement 1/3 de 34 % se situe en dessous de la moyenne, puisque les valeurs négatives n'ont pas de sens ici. La différence sur 100 c'est à dire environ 88 % se situe dans un intervalle au-delà de la moyenne, la distribution aurait donc une forme égaulée avec un pente abrupte à gauche de la moyenne.



1.38 In Exercise 1.19, the mean and standard deviation of the amount of chloroform present in water sources were given to be 34 and 53, respectively. You argued that the amounts of chloroform could therefore not be normally distributed. Use Tchebysheff's theorem (Exercise 1.32) to describe the distribution of chloroform amounts in water sources.

1.38 Dans l'exercice 1.19, la moyenne et l'écart type de la quantité de chloroforme présente dans les sources d'eau étaient respectivement de 34 et 53. Vous avez avancé que les quantités de chloroforme ne pouvaient donc pas suivre une loi normale. Utilisez le théorème de Tchebychev (exercice 1.32) pour décrire la distribution des quantités de chloroforme dans les sources d'eau..

Solution

$\mu = 34$ $\sigma = 53$, les intervalles théoriques sont donc $[(-19)0;87]$, $[0;140]$, $[0;193]$, les valeurs négatives n'ont pas de sens ici. En considérant que le dernier intervalle comporte 99 % des effectifs, si l'on applique la formule de Tchebychev à l'intervalle à 3s, on a effectivement: $f(k = 3) = 0.99 \geq 1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 0.88$

On a sur le plus grand intervalle comportant 99 % 34/154 des valeurs qui sont situés en dessous de la moyenne, donc linéairement 22 %, on peut alors représenter la distribution comme dans l'exercice précédent avec moins du quart de l'effectif à gauche de la moyenne.

