# TD de Statistiques - Séance N° 2

## 1 Travail sur des variables catégorisées avec Excel

### *1.1 Quelques commandes d'Excel utiles pour la saisie de données*

Saisie "assistée" ou non : utilisez le menu :Fichier > Options > Options avancées > Saisie semi-automatique des valeurs de cellule.

Saisissez par exemple, une dizaine de lignes d'un tableau tel que le suivant :

| Identifiant | Conso | SD residence | SD age categorise | SD sexe | SD etudes | SD CSP |
|---|---|---|---|---|---|---|
| Q1 | 15,5375 | Centre | - de 50 ans | Femme | Secondaire | Cadre supérieur |
| Q2 | 7,7605 | Centre | 50 ans à 60 ans | Homme | Primaire | Inactif |
| Q3 | 1,2815 | Centre | 60 ans à 65 ans | Homme | Secondaire | Retraité |
| Q4 | 7,9387 | Sud | 65 ans et plus | Femme | Supérieur | Retraité |
| Q5 | 38,4395 | Centre | 50 ans à 60 ans | Homme | CAP/BEP | Retraité |
| Q6 | 4,7260 | Sud | 50 ans à 60 ans | Homme | Secondaire | Cadre supérieur |
| Q7 | 6,9745 | Sud | 65 ans et plus | Homme | Primaire | Retraité |
| Q8 | 11,1575 | Nord | - de 50 ans | Femme | Secondaire | Employé |
| Q9 | 0,1614 | Autre | 50 ans à 60 ans | Femme | Supérieur | Cadre supérieur |
| Q10 | 7,2769 | Nord | 65 ans et plus | Homme | Secondaire | Retraité |

Remarques.
1) On peut utiliser les possibilités de recopie incrémentée pour générer les identifiants.
2) Les modalités des différentes variables catégorisées sont ici saisies sous forme de texte : très peu de traitements sont alors disponibles sous Excel sans recodage des données.
3) Voir et comprendre l'avertissement que l'on obtient lorsque l'on saisit le texte "- de 50 ans". La meilleure solution pour saisir un tel texte est de faire précéder le texte d'une apostrophe.
4) Pour enregistrer le document, on peut utiliser la barre d'outils Accès rapide. Pour l'enregistrer sous un nouveau nom, ou à un emplacement différent de celui de départ, utilisez l'onglet Fichier du ruban et l'item Enregistrer sous : Fichier > Enregistrer sous.

### *1.2 Tri à plat. Représentation graphique d'une variable catégorisée*

Ouvrez le fichier Conso-Crustaces.xls. Les données qui y sont enregistrées correspondent à la description suivante :

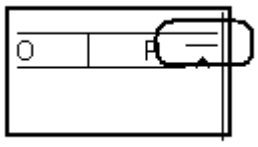
Une étude de consommation a été menée auprès des pêcheurs à pied en Bretagne (Cyndie Picot, thèse soutenue en 2011). 510 pêcheurs à pied fréquentant les plages bretonnes ont rempli un questionnaire relatif à leurs habitudes de consommation. A partir des données récoltées, on a évalué leur consommation alimentaire de crustacés, et on souhaite étudier la variation du niveau de consommation selon différents facteurs socio-économiques.

On s'intéresse ici aux variables suivantes :
- Conso : consommation de crustacés en grammes par personne et par jour ;

- SD Résidence : catégorisé de 1 à 4
  - 1 : nord de la zone étudiée
  - 2 : centre de la zone
  - 3 : sud de la zone
  - 4 : autre zone
- SD Age Catégorisé : catégorie d'âge, avec la catégorisation suivante :
  - 1 : moins de 50 ans
  - 2 : de 50 ans à moins de 60 ans
  - 3 : de 60 ans à moins de 65 ans
  - 4 : 65 ans et plus
- SD sexe : sexe de la personne interrogée
- SD Etudes : niveau d'études, catégorisé de 1 à 4
  - 1 : primaire ou sans diplôme
  - 2 : CAP/BEP
  - 3 : Etudes secondaires, Bac
  - 4 : études supérieures
- SD CSP : catégorie socio-professionnelle
  - 1 : agriculteur
  - 2 : petit patron
  - 3 : cadre supérieur
  - 4 : profession intermédiaire
  - 5 : employé
  - 6 : ouvrier qualifié
  - 7 : retraité
  - 8 : inactif

### 1.2.1 Consultation des données

Le fichier comporte 512 observations. Il peut être intéressant de couper la fenêtre pour afficher deux sous-fenêtres. Pour cela, déplacez la barre située au-dessus de l'ascenseur vertical.



**Exploration des données** : placez la sélection dans l'une des cellules contenant les intitulés des variables et utilisez le menu Données > Filtrer :

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Consc | SD residen | SD age categoris | SD sexe | SD etude | SD CSP |
| 2 | 5,2668 | 1 | 4 | 1 | 1 | 7 |

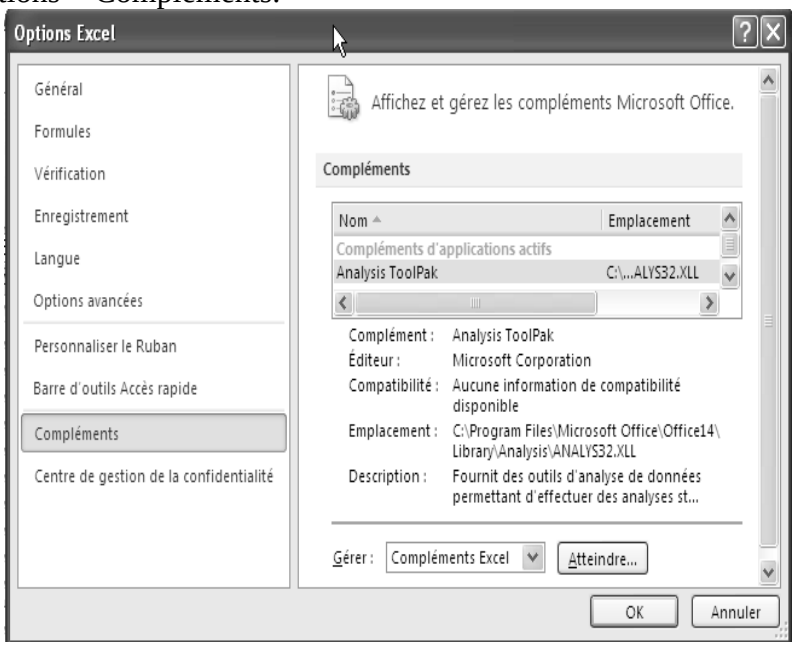
Par exemple, affichez les lignes pour lesquelles SD sexe = 1 et SD residence = 1.

Pour réafficher l'ensemble des données ou supprimer le filtre, utilisez de nouveau le menu Données > Filtrer.

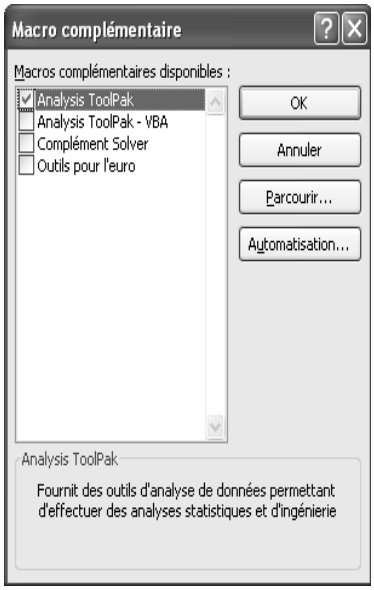
**Tri à plat** : on veut par exemple, obtenir un tri à plat selon les modalités (codées de 1 à 8) de la variable SD CSP.

Première solution : on utilise l'utilitaire d'analyse.

L'utilitaire d'analyse est un complément à Excel (un ensemble de macros), fourni avec le logiciel, mais qui n'est pas activé dans l'installation par défaut. Pour activer l'utilitaire d'analyse, utilisez le menu Fichier > Options > Compléments.

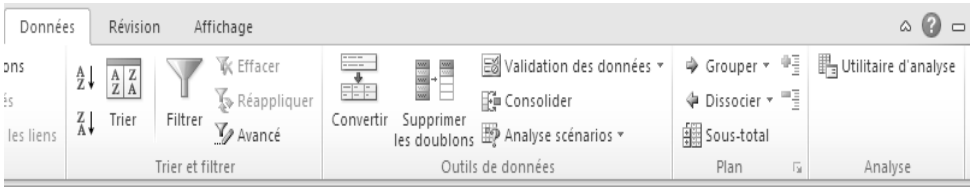



Cliquez sur le bouton "Atteindre". La fenêtre suivante s'affiche :




Sélectionnez ensuite l'item Analysis Toolpak.

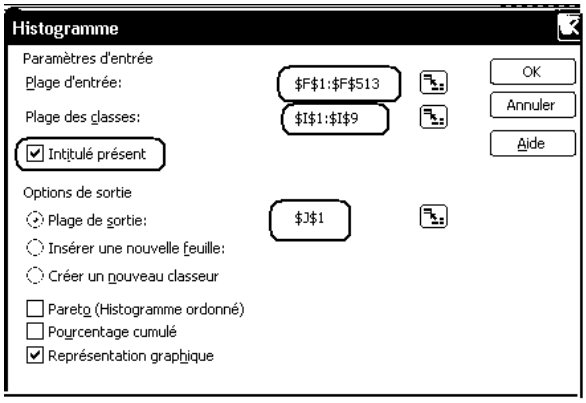
L'onglet Données du ruban comporte alors un item supplémentaire "Utilitaire d'analyse".




N.B. Excel modifie alors le fichier de préférences de votre configuration personnelle du logiciel, et l'utilitaire d'analyse devrait être présent lors des prochains chargements du logiciel sur l'un des appareils des salles de TD ou de la BU.

Complétez la feuille de données en indiquant les modalités de la variable CSP dans la plage de cellules I1 à I9, par exemple (en-tête "CSP" en I1, autres valeurs à laisser impérativement sous forme numérique).

Utilisez ensuite le menu Outils - Utilitaire d'analyse, puis l'item Histogramme. La fenêtre de dialogue pourra par exemple être complétée comme suit :




Vous devriez obtenir le tableau de valeurs suivant dans la plage de cellules J1:K10, accompagné éventuellement d'un "histogramme" (ce qu'Excel appelle un histogramme) :

| CSP | Fréquence |
|---|---|
| 1 | 5 |
| 2 | 7 |
| 3 | 46 |
| 4 | 23 |
| 5 | 69 |
| 6 | 33 |
| 7 | 289 |
| 8 | 40 |
| ou plus... | 0 |

N.B. Comme on peut l'observer sur l'exemple ci-dessus, les résultats produits par l'utilitaire d'analyse sont de simples valeurs numériques ou textuelles, et non des formules de calcul pour lesquelles le recalcul automatique de la feuille produit des mises à jour des résultats.

Deuxième solution : la fonction NB.SI :

Par exemple, entrez en cellule L2 : `=NB.SI($F$1:$F$513; "=1")`
La fonction NB.SI compte le nombre de cellules de la plage indiquée qui satisfont la condition placée en deuxième paramètre.

Troisième solution : la fonction FREQUENCE, que nous verrons ultérieurement.

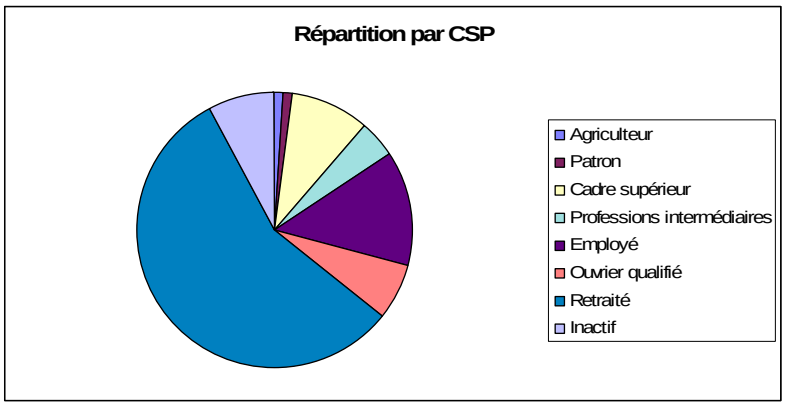
### 1.2.2 Représenter la distribution à l'aide d'un diagramme circulaire

Il faut partir du tri à plat de la variable à représenter. Mais le graphique sera plus explicite si les modalités sont indiquées par leurs libellés.

Constituer un tableau tel que le suivant :

| | |
|---|---|
| Agriculteur | 5 |
| Patron | 7 |
| Cadre supérieur | 46 |
| Professions intermédiaires | 23 |
| Employé | 69 |
| Ouvrier qualifié | 33 |
| Retraité | 289 |
| Inactif | 40 |

Sélectionnez ensuite la plage contenant les intitulés des CSP et les effectifs puis utilisez le menu Insertion > Secteurs pour obtenir le diagramme circulaire suivant :



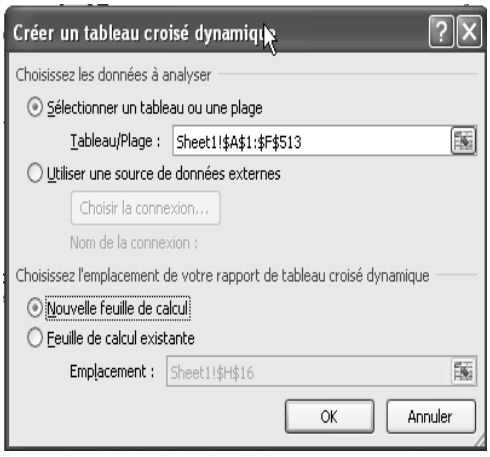

Ajustez au besoin la taille du graphique pour afficher la totalité de la légende.

### 1.2.3 Effectuer un tri croisé sur deux variables catégorisées

On veut, par exemple, réaliser un tableau de contingence en croisant les variables SD Etudes et SD CSP. On va utiliser l'outil fourni par le menu Données - Rapport de tableau croisé dynamique... dont les dialogues ne sont pas très explicites, c'est le moins que l'on puisse dire !)

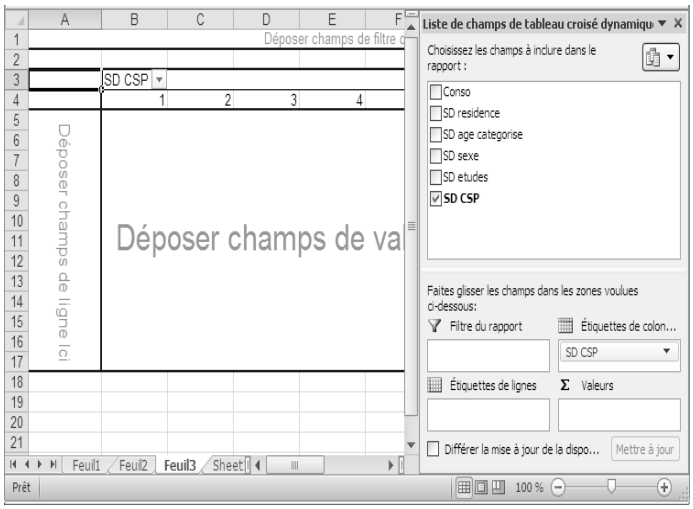
- Onglet Insertion > Tableaux > TblCroiséDynamique

Premier dialogue :
Indiquez A1:F513 comme plage de données et une nouvelle feuille comme emplacement pour le tableau croisé dynamique.

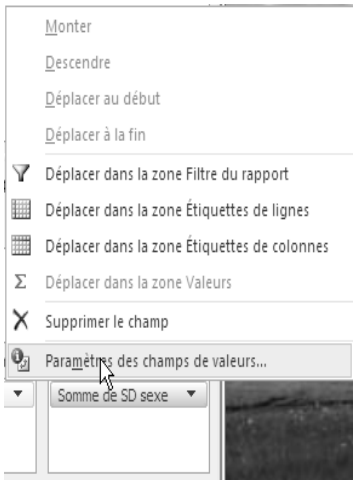



Second dialogue : faites glisser SD CSP dans la zone "Déposer champs de colonne ici", SD études dans la zone "Déposer champs de ligne ici" et l'une quelconque des variables dans la zone "Champs

de valeurs. Le choix a peu d'importance, mais il faut que la variable ne comporte pas de valeur manquante, faute de quoi le tableau de contingence sera incorrect. L'idéal est sans doute de prendre ici un identifiant des réponses :




Par défaut, Excel fait une somme des valeurs correspondantes, alors que nous volons un simple comptage. Pour cela, dans la zone étiquetée Σ Valeurs, faites un clic droit sur "Somme de SD xxx", sélectionner le menu local "Paramètres des champs de valeurs..." et, dans le dialogue "Paramètres des champs de valeurs" choisissez "Nombre" au lieu de "Somme".




On obtient le résultat suivant :

| Nombre de SD age categorise | SD CSP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SD etudes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| 1 | 1 | 5 | 2 | 2 | 14 | 5 | 101 | 14 | 144 |
| 2 | | 1 | 3 | 3 | 24 | 22 | 83 | 11 | 147 |
| 3 | 2 | 1 | 12 | 5 | 21 | 5 | 57 | 5 | 108 |
| 4 | 2 | | 29 | 13 | 10 | 1 | 48 | 10 | 113 |
| Total | 5 | 7 | 46 | 23 | 69 | 33 | 289 | 40 | 512 |

Remarques.

1) Les palettes d'outils flottantes permettent éventuellement de modifier le tableau : changement de champ ligne, de champ colonne ou de champ pivot, changement de fonction à appliquer au champ pivot, etc.
2) On peut également indiquer un champ "page" (par exemple, la variable SD sexe), de manière à afficher alternativement les tableaux de contingence obtenus sur les sous-populations correspondant aux deux modalités de cette variable.

3) Comment Excel traite-t-il les valeurs manquantes ?
Supprimez les valeurs de SD études ou de SD CSP dans une ou deux lignes du fichier, faites recalculer le tableau croisé et observez le résultat. De même, explorez l'effet d'une valeur manquante dans le champ pivot.

### 1.2.4 Affichage du résultat du tri croisé sous forme de fréquences lignes, fréquences colonnes, etc

On peut utiliser le menu Outils de tableau croisé dynamique > Calcul > Afficher les valeurs pour afficher le tableau à double entrée sous forme de fréquences lignes (item "% du total de la ligne") ou de fréquences colonnes (item "% du total de la colonne").

Exemple :

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | Nombre de SD residence | SD CSP | | | | | | | | |
| 4 | SD etudes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total général |
| 5 | 1 | 0,69% | 3,47% | 1,39% | 1,39% | 9,72% | 3,47% | 70,14% | 9,72% | 100,00% |
| 6 | 2 | 0,00% | 0,68% | 2,04% | 2,04% | 16,33% | 14,97% | 56,46% | 7,48% | 100,00% |
| 7 | 3 | 1,85% | 0,93% | 11,11% | 4,63% | 19,44% | 4,63% | 52,78% | 4,63% | 100,00% |
| 8 | 4 | 1,77% | 0,00% | 25,66% | 11,50% | 8,85% | 0,88% | 42,48% | 8,85% | 100,00% |
| 9 | Total général | 0,98% | 1,37% | 8,98% | 4,49% | 13,48% | 6,45% | 56,45% | 7,81% | 100,00% |
| 10 | | | | | | | | | | |

# 2 Statistiques descriptives sur une variable numérique avec Excel

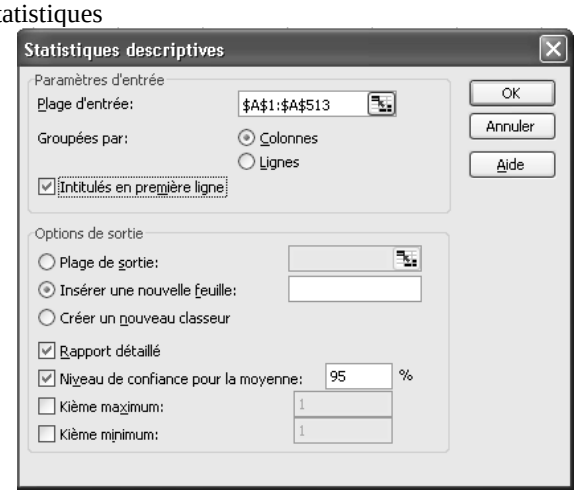
## *2.1 Moyenne, écart type, etc*

### 2.1.1 Moyenne, écart type, etc à l'aide de l'utilitaire d'analyse

On reprend le fichier de travail : Conso-crustaces.xls

On souhaite calculer le nombre d'observations, la moyenne, la médiane, la variance et l'écart type, le maximum, et le minimum de la consommation, tous groupes confondus.

L'utilitaire d'analyse comporte un item "Statistiques descriptives" permettant d'obtenir ces résultats sans avoir à saisir de formules :




| Conso | |
|---|---|
| | |
| Moyenne | 11,33 |
| Erreur-type | 0,78 |
| Médiane | 5,50 |
| Mode | 1,49 |
| Écart-type | 17,61 |
| Variance de l'échantillon | 310,09 |
| Kurstosis (Coefficient d'applatissement) | 29,65 |
| Coefficient d'assymétrie | 4,63 |
| Plage | 173,84 |
| Minimum | 0,08 |
| Maximum | 173,92 |
| Somme | 5777,54 |
| Nombre d'échantillons | 510,00 |
| Niveau de confiance(95,0%) | 1,53 |

### 2.1.2 Moyenne, écart type, etc à l'aide de fonctions Excel

Fonctions à utiliser :

```
= NB((<plage de cellules>)
= MOYENNE(<plage de cellules>)
= MEDIANE(<plage de cellules>)
= ECARTYPE(<plage de cellules>)
= ECARTYPEP(<plage de cellules>)
= VAR(<plage de cellules>)
= VAR.P(<plage de cellules>)
= MAX(<plage de cellules>)
= MIN(<plage de cellules>)
```

N.B. Pour désigner une plage de cellules : par exemple : `A2:A513`

On veut par exemple calculer les valeurs des paramètres nombre d'observations, moyenne, médiane, écart type corrigé, écart type, variance corrigée, variance, maximum et minimum pour la variable Conso en plaçant les résultats dans une nouvelle feuille de calcul.

Créez une nouvelle feuille en cliquant sur l'icône  présente dans le bas de la fenêtre.

Saisissez les légendes convenables en première colonne, puis les formules voulues en deuxième colonne. Vous devriez obtenir :

| | Conso |
|---|---|
| Nombre | 510 |
| Moyenne | 11,33 |
| Médiane | 5,50 |
| Ecart type corrigé | 17,61 |
| Ecart type | 17,59 |
| Variance corrigée | 310,09 |
| Variance | 309,48 |
| Maximum | 173,92 |
| Minimum | 0,08 |
| | |

## *2.2 Faire un regroupement en classes sur la variable Conso*

### 2.2.1 Utiliser l'utilitaire d'analyse

L'utilitaire d'analyse permet d'obtenir un résultat tel que le suivant :

| Classes Conso | Fréquence |
|---|---|
| 5 | 238 |
| 10 | 108 |
| 20 | 91 |
| 40 | 47 |
| 60 | 13 |
| 80 | 6 |
| ou plus... | 7 |

On commence par saisir un tableau contenant les bornes supérieures des classes souhaitées, puis on choisit l'item Histogramme de l'utilitaire d'analyse.

### 2.2.2 Utiliser la fonction FREQUENCE

L'utilitaire d'analyse se sert en fait de la fonction FREQUENCE. Contrairement aux fonctions vues jusqu'à présent, cette fonction renvoie non pas un résultat, mais une plage de résultats. Dans Excel, une telle fonction est appelée *fonction matricielle*. La saisie se fait comme suit :

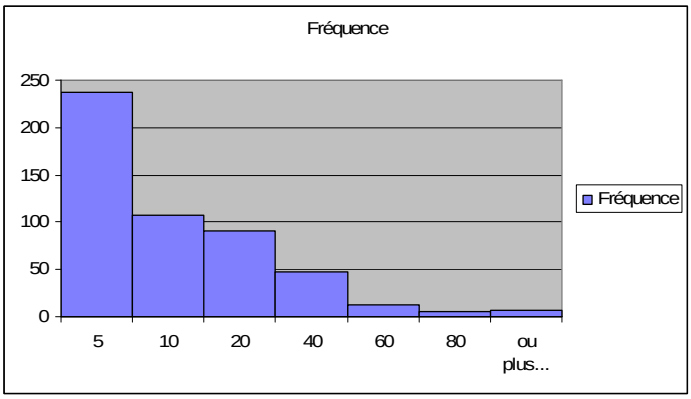
Comme précédemment, on saisit un tableau contenant les bornes supérieures des classes.
Par exemple, saisissons les bornes précédentes dans la plage J23 à J28.
On sélectionne ensuite une plage comportant une cellule de plus que la page des bornes de classes, et on saisit pour l'ensemble de ces cellules la formule :

```
=FREQUENCE(A2:A513;J23:J28)
```

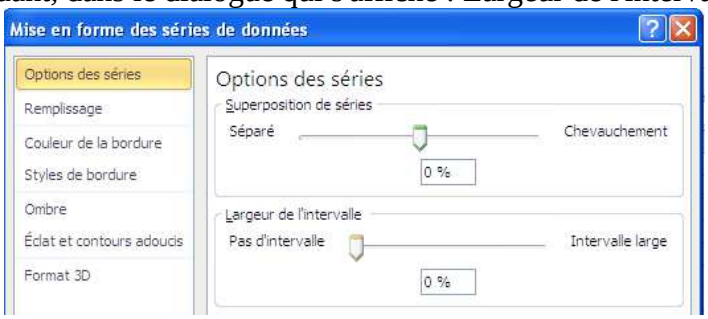
On valide ensuite la saisie en appuyant simultanément sur Maj + Ctrl + Return.

### 2.2.3 Construire une représentation graphique des données

On peut facilement construire un diagramme à bandes à partir du tableau d'effectifs. Attention cependant : il ne s'agit pas d'un véritable histogramme puisqu'il n'est pas tenu compte de l'amplitude des classes.




N.B. Sur le dessin ci-dessus, les rectangles ont été rendus adjacents en double-cliquant sur l'un d'entre eux et en indiquant, dans le dialogue qui s'affiche : Largeur de l'intervalle = 0% :




## *2.3 Croisement d'une variable numérique continue et d'une variable catégorielle (données réparties en groupes indépendants)*

### 2.3.1 Cas de données non triées

On veut, par exemple, obtenir la moyenne de la variable Conso, pour chacun des deux sexes.

Pour obtenir les moyennes dans les différents groupes définis par la variable catégorisée SD Sexe, par exemple, on peut utiliser les fonctions SOMME.SI et NB.SI dans des expressions telles que :

```
=SOMME.SI(D2:D513;"=1";A2:A513)/NB.SI(D2:D513;"=1")
```

On peut également former un tableau croisé en indiquant SD Sexe comme variable pour les champs de ligne, et Moyenne de Conso pour les champs de valeurs.

N.B. Les deux méthodes ne fournissent pas exactement les mêmes résultats. Essayez d'en trouver la raison (réfléchissez notamment au comportement des deux méthodes vis-à-vis des valeurs manquantes).

### 2.3.2 Cas de données triées par groupe (groupes superposés ou disposés sur deux colonnes d'une feuille de calcul

Faites une copie de la feuille de données de départ (Sheet1) et trier les données selon les valeurs de la colonne SD Sexe. Calculer ensuite la moyenne de la variable Conso pour chacun des deux sexes, à l'aide de la fonction MOYENNE()

Faites de nouveau une copie de la feuille de données de départ. Supprimez les colonnes autres que Conso et SD Sexe. Trier les données selon les valeurs de la colonne SD Sexe puis, à l'aide des commandes Copier-Coller ou Glisser-Déplacer, organisez les données en plaçant celles relatives à SD Sexe=1 dans une colonne et celles relatives à SD Sexe=2 dans la colonne suivante.
On peut alors utiliser l'item "Statistiques Descriptives" de l'utilitaire d'analyse pour obtenir des statistiques descriptives par groupe :

| Conso-Sexe=1 | | Conso-Sexe=2 | |
|---|---|---|---|
| | | | |
| Moyenne | 9,20 | Moyenne | 12,64 |
| Erreur-type | 1,05 | Erreur-type | 1,08 |
| Médiane | 4,75 | Médiane | 6,42 |
| Mode | 1,49 | Mode | 2,98 |
| Écart-type | 14,56 | Écart-type | 19,15 |
| Variance de l'échantillon | 212,02 | Variance de l'échantillon | 366,66 |
| Kurstosis (Coefficient d'applatissement) | 48,93 | Kurstosis (Coefficient d'applatissement) | 23,97 |
| Coefficient d'assymétrie | 5,81 | Coefficient d'assymétrie | 4,19 |
| Plage | 151,66 | Plage | 173,76 |
| Minimum | 0,08 | Minimum | 0,16 |
| Maximum | 151,74 | Maximum | 173,92 |
| Somme | 1784,68 | Somme | 3992,87 |
| Nombre d'échantillons | 194 | Nombre d'échantillons | 316 |
| Niveau de confiance(95,0%) | 2,06 | Niveau de confiance(95,0%) | 2,12 |

## *2.4 Calculer un paramètre descriptif d'une variable numérique continue pour chacun des groupes obtenus en croisant deux variables catégorielles.*

Exemple : On veut obtenir la moyenne de la variable Conso dans chacun des groupes obtenus en croisant les variables SD Sexe et SD Residence.

On peut pour cela construire un tableau croisé en indiquant l'une des variables catégorisées en colonne, l'autre en ligne et en indiquant Moyenne de Conso dans la zone champs de calcul. On obtient :

| Moyenne de Conso | SD residence | | | | |
|---|---|---|---|---|---|
| SD sexe | 1 | 2 | 3 | 4 | Total général |
| 1 | 12,66 | 7,65 | 7,49 | 7,15 | 9,20 |
| 2 | 19,61 | 12,14 | 11,09 | 5,65 | 12,64 |

| | | | | | |
|---|---|---|---|---|---|
| Total général | 16,46 | 10,53 | 9,81 | 6,15 | 11,33 |