R-Commander : Notions du chapitre 1

Statistiques descriptives

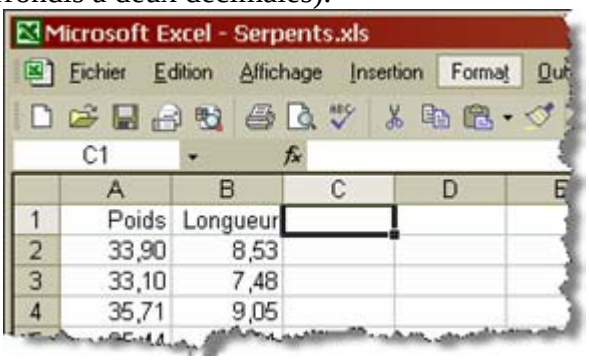
1) Lecture des données

La première étape est de lire le jeu de données qui nous intéresse. R-Commander lit des fichiers de plusieurs formats :

- Excel, Access, dBase
- Texte
- Enregistré sur le presse-papier (opérations « sélectionner », « copier »)
- SPSS, Stata, Minitab

a) Exemple de lecture d'un fichier Excel : *serpents.xls*

- Le fichier serpents.xls contient les mesures de poids et de longueur sur 147 serpents à la naissance (ici, arrondis à deux décimales).



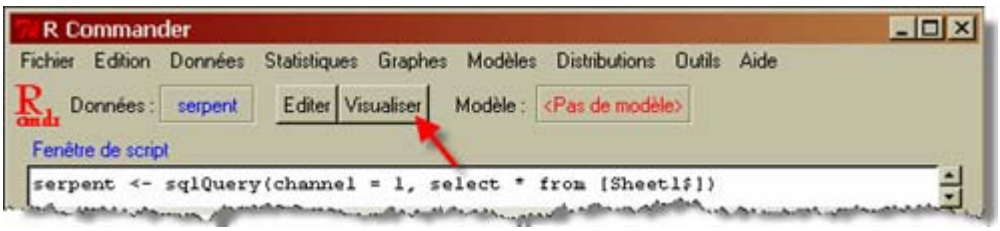
- Télécharger d'abord le fichier sur votre ordinateur (il est disponible sur le site web du cours). Ouvrir *R*, et charger le package *Rcmdr*.

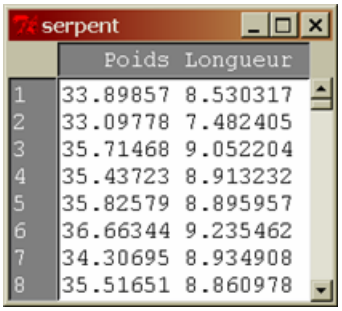
- Cliquer sur les commandes suivantes dans le menu *Données* de R-Commander.

- On peut donner n'importe quel nom au fichier, mais on suggère de donner un nom évocateur de son contenu.

- Identifier la localisation du fichier sur l'ordinateur et la feuille excel dans le fichier qui contient les données.

- Le jeu de données est saisi, on peut le constater avec l'option *Visualiser*. On voit ici les 8 premières lignes de données entrées. Le logiciel interprète la première ligne comme le nom des variables.





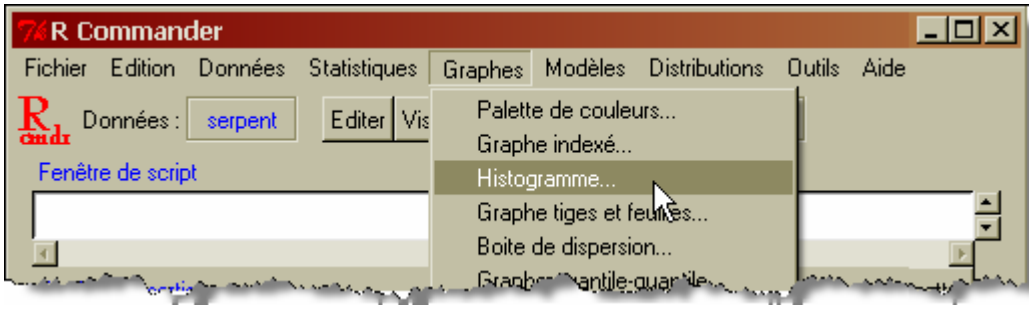
- On peut ensuite aménager les données : ajouter ou modifier des valeurs, renommer les variables, créer de nouvelles variables à partir des variables existantes, placer les valeurs d'une variable en ordre croissant, etc.

2) Représentations graphiques

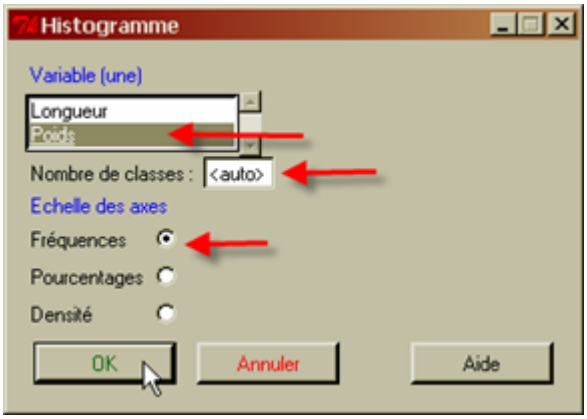
a) Données univariées

- **Histogramme** (variables quantitatives continues)

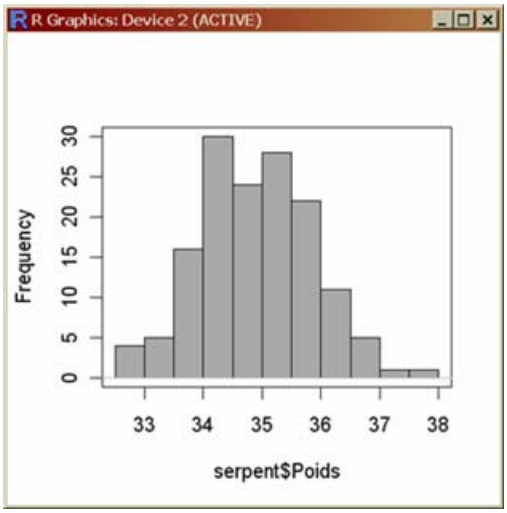
On sélectionne l'histogramme dans le menu *Graphes*.



Il suffit de choisir la variable pour laquelle on veut un graphique, le nombre de classes (si on veut) et le type d'histogramme.

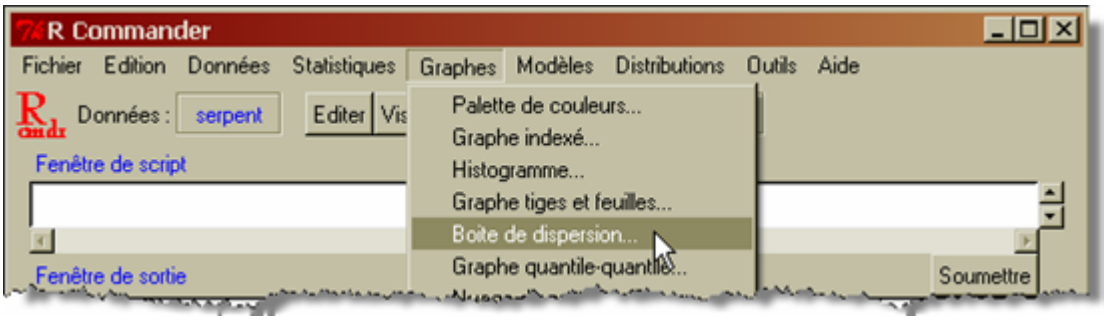


Une nouvelle fenêtre s'ouvre, dont on peut modifier les dimensions pour changer l'allure du graphique. (Il faut parfois cliquer sur l'onglet RGui en bas de l'écran pour voir la fenêtre graphique.) En cliquant sur le bouton droit de la souris, on peut *copier comme vectoriel* et coller dans un fichier .doc. On peut sauvegarder le graphique sous différents formats (pdf, ps, eps, jpg, png) en retournant dans la fenêtre principale de R-Commander dans le menu *Graphes-Sauver le graphe*.

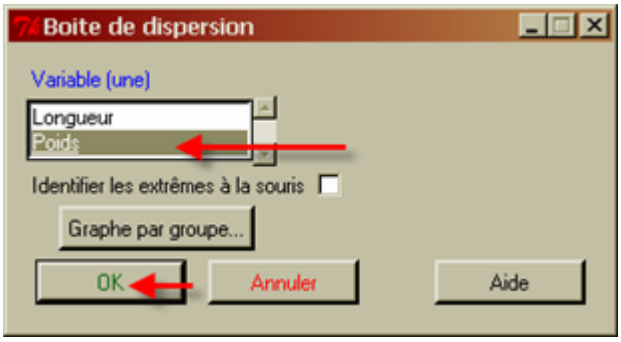


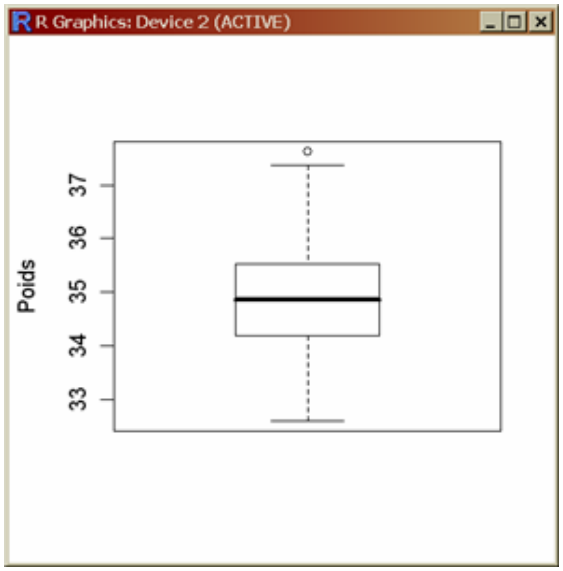
- **Diagramme en boîte** (variables quantitatives continues et discrètes)

On sélectionne la boîte de dispersion dans le menu *Graphes*.



Il suffit de choisir la variable pour laquelle on veut un graphique.

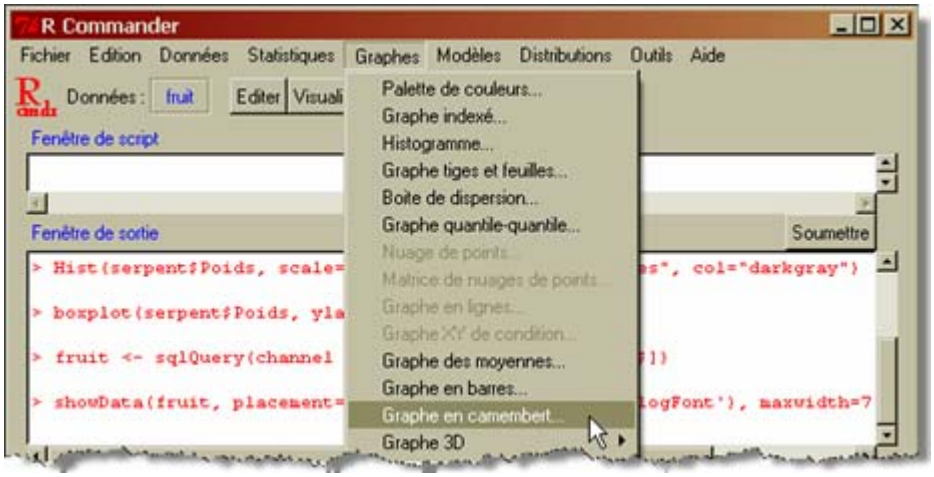




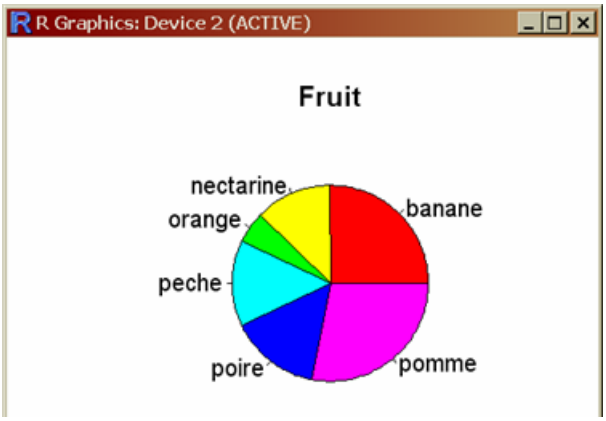
- **Diagramme en pointes de tarte** (variables qualitatives)

On considère un jeu de 275 données, soient les fruits préférés d'écoliers parmi les suivants : banane, orange, nectarine pêche, poire, pomme. La variable *Fruit* est qualitative.

On sélectionne le *Graphe en camembert* dans le menu *Graphes*.

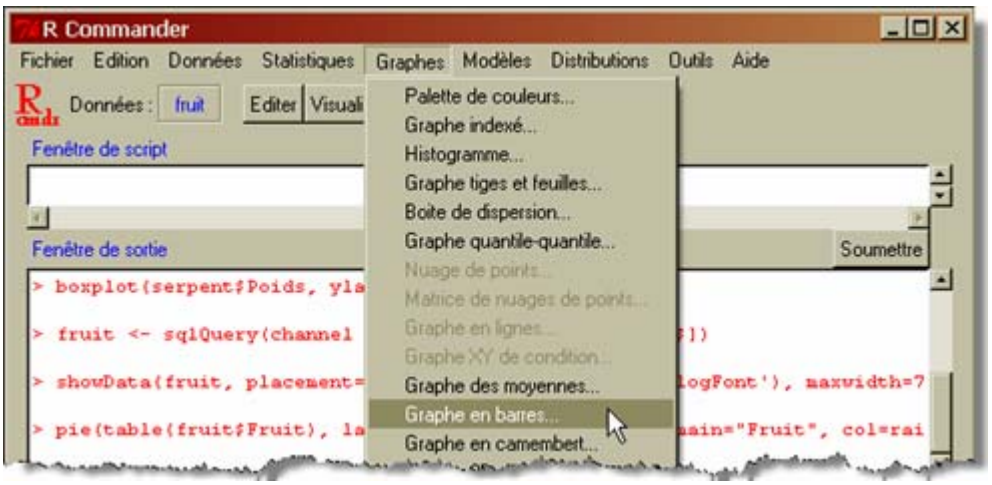


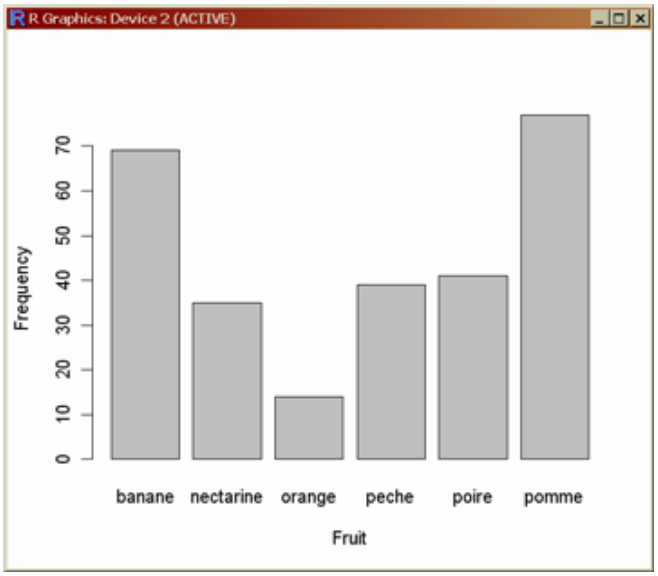
Il suffit de choisir la variable pour laquelle on veut un graphique, en l'occurrence la variable *Fruit*.



- **Diagramme en bâtons** (variables qualitatives et quantitatives discrètes)

Ce diagramme est aussi appelé *Graphe en barres*. Il suffit de choisir la variable pour laquelle on veut un graphique.



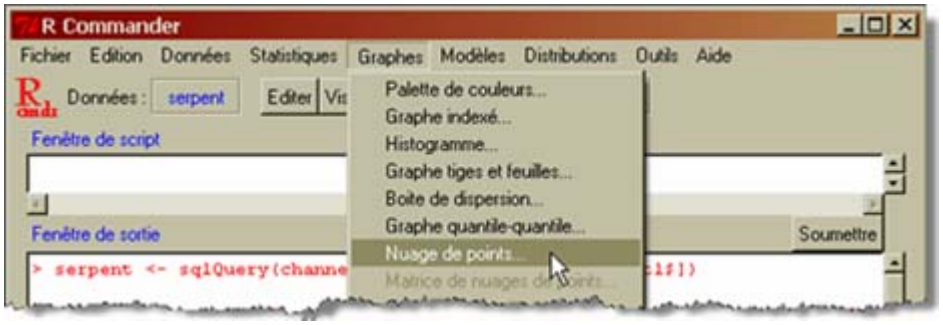


b) Données bivariées

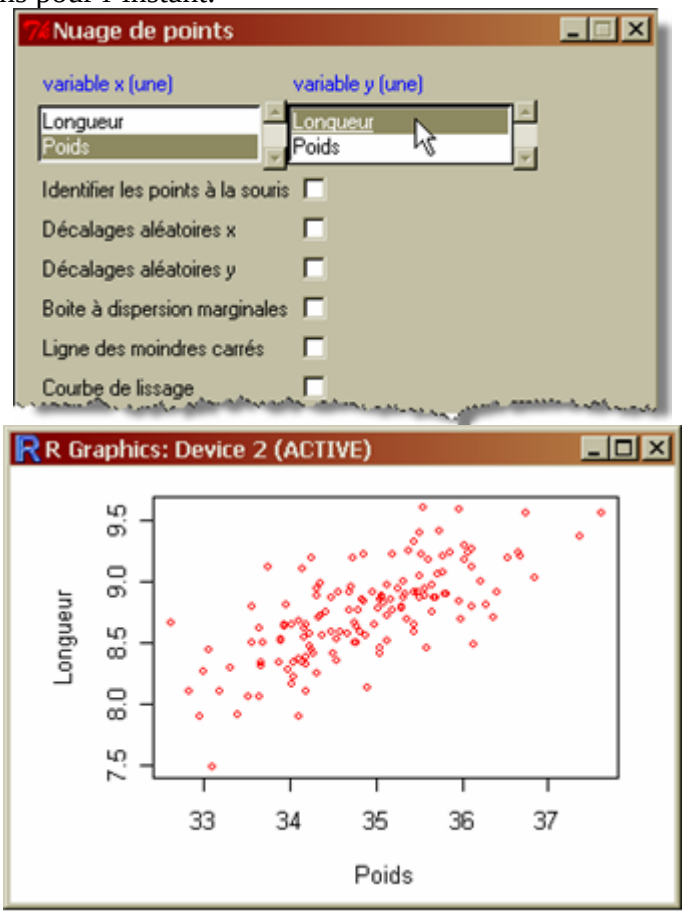
- **Diagramme de dispersion** (deux variables quantitatives)

Note : Lorsque plusieurs jeux de données ont été importés au cours d'une même session de travail, on peut passer de l'un à l'autre facilement. Pour travailler de nouveau sur le jeu de données *serpent* :

Pour le diagramme de dispersion, on choisit le *nuage de points* dans le menu *Graphes*.



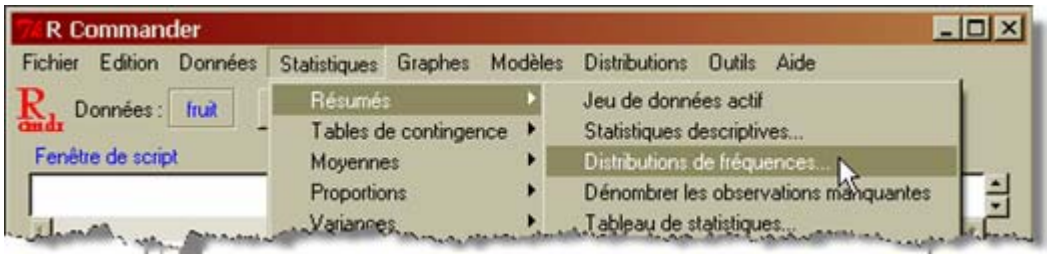
Il suffit de sélectionner la variable sur l'axe des x et la variable sur l'axe des y. Laissons tomber les options pour l'instant.

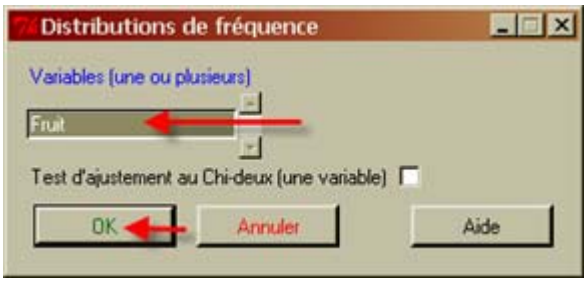


3) Description numérique

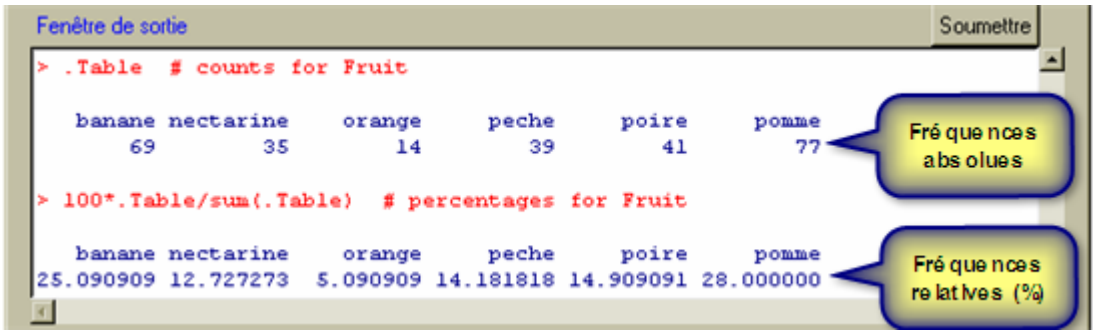
- **Tableau de fréquences** (disponible pour les variables qualitatives seulement)

  Pour connaître le nombre de fois où chaque fruit est cité dans le jeu de données, on construit le tableau de fréquences comme suit.



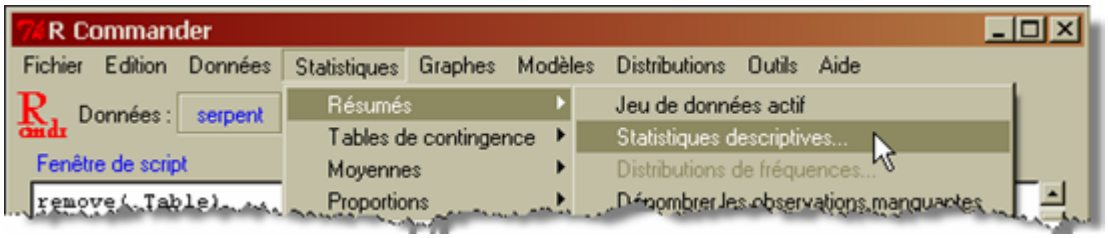


Le résultat apparaît dans la fenêtre de sortie.

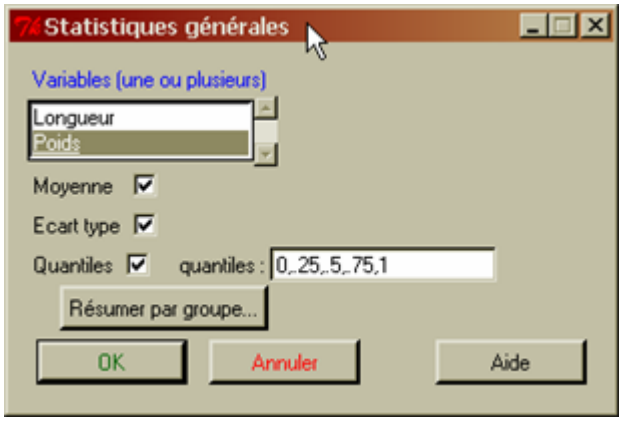


- **Moyenne, écart-type et résumé à cinq nombres** (variables quantitatives)

  Ces quantités sont obtenues à partir du menu *Statistiques – Résumés – Statistiques descriptives*.

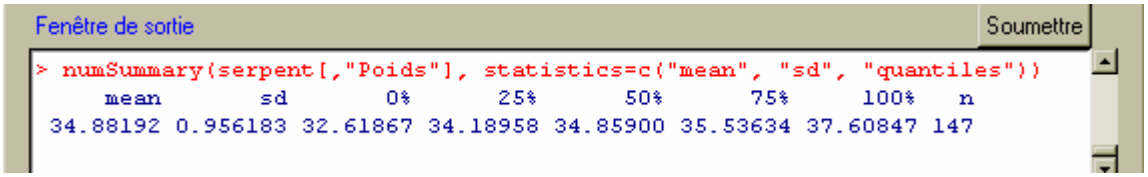


  On sélectionne la variable sur laquelle on veut des statistiques ainsi que les statistiques souhaitées.



  Les résultats apparaissent dans la fenêtre de sortie dans l'ordre suivant :

$$\bar{x}, \quad s, \quad x_{(1)}, \quad q_{0.25}, \quad q_{0.5}, \quad q_{0.75}, \quad x_{(n)}, \quad n$$

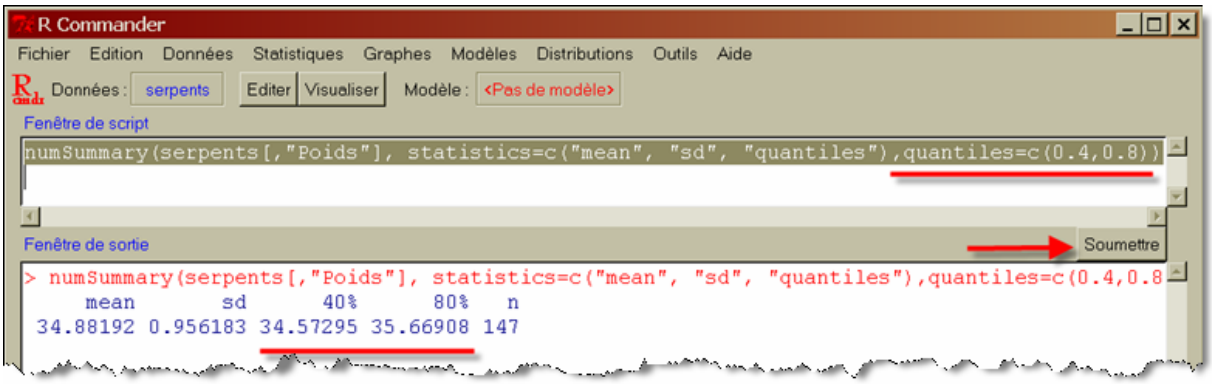


- **Quantiles**

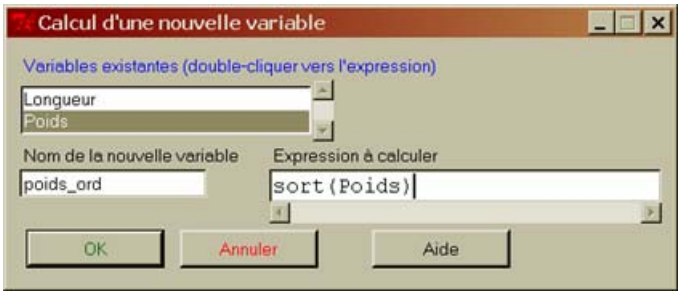
➢ Seuls les quantiles d'ordre 0.25, 0.5 et 0.75 ainsi que le minimum et le maximum apparaissent dans les statistiques descriptives par défaut. Pour obtenir d'autres quantiles, il faut modifier la commande dans la fenêtre de script.

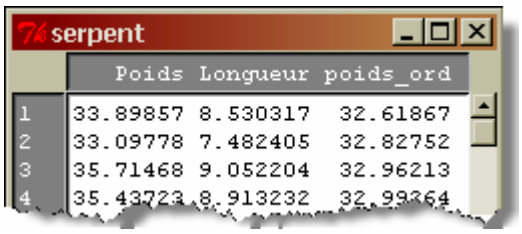
1) Calculer les statistiques descriptives par défaut (voir page précédente) ;
2) Dans la fenêtre de script, ajouter les instructions $quantiles = c(\gamma_1, \gamma_2, \gamma_3, ...)$ à la fin de la commande *numSummary* pour obtenir les quantiles d'ordre γ_1 , γ_2 , γ_3, etc.
3) Sélectionner la ligne *numSummary* et cliquer sur *Soumettre*.
4) Les quantiles apparaîtront dans la fenêtre de sortie.

Voici un exemple pour les quantiles d'ordre 0.4 et 0.8.



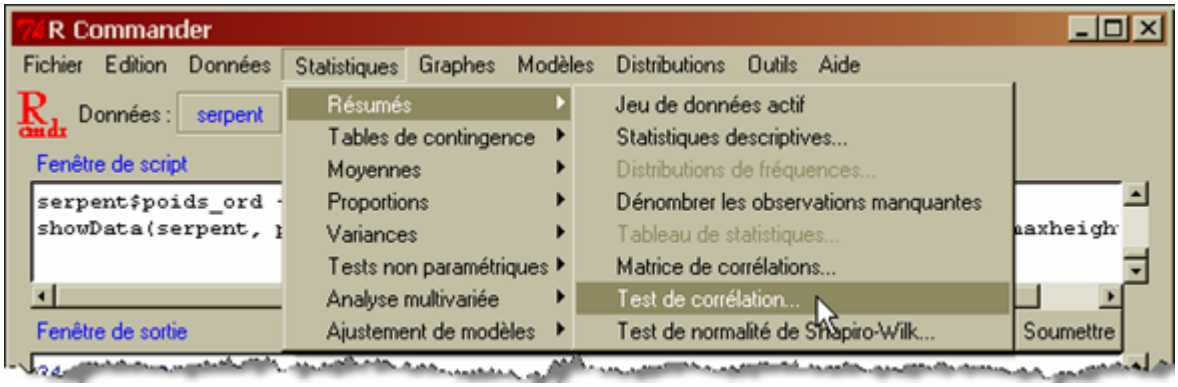
➢ On peut aussi placer les valeurs d'une variable en ordre croissant. Il faut alors créer une nouvelle variable, disons poids_ord, à l'aide de la commande *sort(Poids)* :



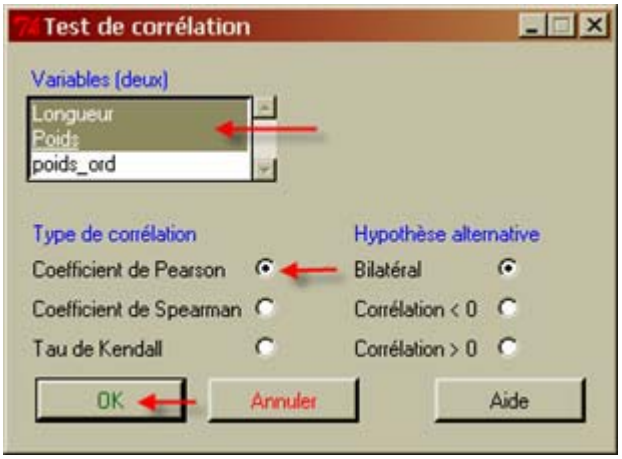


- **Coefficient de corrélation**

Cette quantité est obtenue à partir du menu *Statistiques – Résumés – Test de corrélation*.



On sélectionne deux variables dans la liste (utiliser la touche « Shift » ou « Ctrl »). On choisit le *Coefficient de Pearson*, et on clique sur OK. Pour le moment, il n'est pas nécessaire de tenir compte de la colonne sur l'hypothèse alternative.



Le résultat apparaît dans la fenêtre de sortie, identifié par *sample estimates : cor*.