**Introduction à Rcommander**

***Pauline Scherdel***

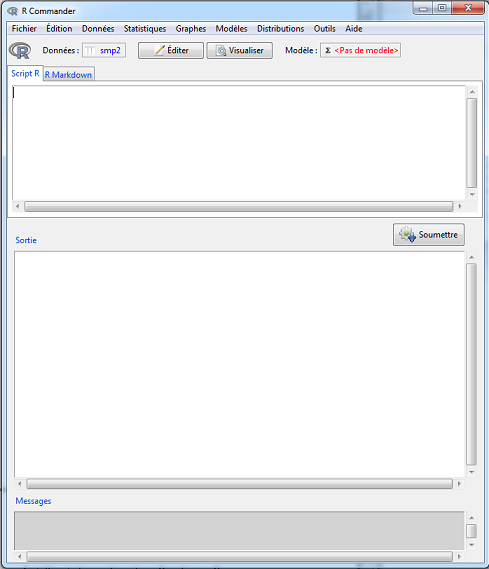
***Septembre 2014***

* [Introduction à Rcmdr sous R](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#introduction-a-rcmdr-sous-r)
* [Interagir avec R](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#interagir-avec-r)
* [Installer et charger le package Rcmdr sous R](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#installer-et-charger-le-package-rcmdr-sous-r)
* [Importation des données](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#importation-des-donnees)
  + [Importation à partir d’un fichier Excel](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#importation-a-partir-dun-fichier-excel)
  + [Importation à partir d’un fichier CSV](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#importation-a-partir-dun-fichier-csv)
* [Manipulation des données](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#manipulation-des-donnees)
  + [Visualisation brève au jeu de données](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#visualisation-breve-au-jeu-de-donnees)
  + [Conversion des données quantitatives en qualitatives](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#conversion-des-donnees-quantitatives-en-qualitatives)
  + [Recodage des données quantitatives en qualitatives](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#recodage-des-donnees-quantitatives-en-qualitatives)
  + [Création de nouvelles variables](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#creation-de-nouvelles-variables)
* [Description des données](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#description-des-donnees)
  + [Distribution des variables quantitatives et qualitatives](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#distribution-des-variables-quantitatives-et-qualitatives)
  + [Représentation des variables quantitatives et qualitatives](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#representation-des-variables-quantitatives-et-qualitatives)
  + [Représentation de variables quantitatives en fonction d’une variable qualitative](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#representation-de-variables-quantitatives-en-fonction-dune-variable-qualitative)
* [Tests statistiques](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#tests-statistiques)
  + [Comparaison de moyennes d’une variable quantitative entre deux groupes](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#comparaison-de-moyennes-dune-variable-quantitative-entre-deux-groupes)
  + [*Comparaison de proportions d’une variable qualitative entre deux groupes*](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#comparaison-de-proportions-dune-variable-qualitative-entre-deux-groupes)
* [Modèles statistiques](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#modeles-statistiques)
  + [Modèles linéaires](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#modeles-lineaires)
  + [Modèles logistiques](https://www.france-universite-numerique-mooc.fr/c4x/UPSUD/42001S03/asset/introRcmdr.html#modeles-logistiques)

**Introduction à Rcmdr sous R**

Ce document constitue une présentation succincte du package Rcommander (Rcmdr), une sur-couche du logiciel R. Il s’agit d’une interface graphique qui facilite l’interactivité avec le logiciel R. En particulier, on s’intéressera à l’importation et la manipulation des données quantitatives et qualitatives, à la description d’un jeu de données et à l’analyse statistique.

L’interface du package Rcmdr est assez rudimentaire. Elle est composée d’un menu avec des listes déroulantes afin de remplacer les fonctions R à taper dans un script, d’une fenêtre “script R” avec les commandes R, d’une fenêtre “Sortie” relative aux résultats et d’une fenêtre “Message” relative aux messages d’erreur. Les commandes exécutées par le menu sont traduites en script R dans la fenêtre script. Il est possible de taper des commandes R directement dans cette fenêtre.



*Aperçu de l’interface de Rcmdr*

**Interagir avec R**

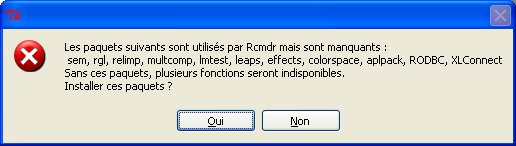
**Démarrer avec R.** Quelque soit le système d’exploitation utilisé (Windows, Mac, Linux), R fonctionne comme tout autre logiciel : il suffit généralement de double-cliquer sur l’icône de l’application pour démarrer R. On dispose ensuite d’une console interactive dans laquelle on peut commencer à saisir des commandes après l’invite R >. Les résultats seront affichés aussitôt dans la console.

**Installer et charger le package Rcmdr sous R**

Il faut installer le package Rcmdr grâce à la commande install.packages() :

**install.packages**("Rcmdr")

On obtient la fenêtre suivante :



*Message lors de l’installation du package Rcmdr*

Cliquer sur Oui pour installer les dépendances.

On installera aussi les packages epicalc, epitools et prettyR :

**install.packages**("epicalc")

**install.packages**("epitools")

**install.packages**("prettyR")

Les packages sont installés définitivement (tant qu’on ne les désinstalle pas).

Ensuite, il faut “charger les packages” à chaque session de R pour avoir accès aux fonctions qui les composent. On utilise pour cela la commande library() :

**library**(Rcmdr)

**library**(epicalc)

**library**(epitools)

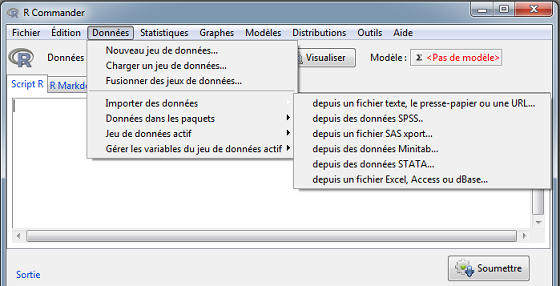
**library**(prettyR)

**Importation des données**

Il est possible d’importer un jeu de données à partir d’un fichier Excel, Access, dBase ou texte mais également à partir d’autres formats comme SAS, SPSS ou STATA. Sous MAC, il est impossible d’importer un jeu de données à partir d’un fichier Excel, Access ou dBase.

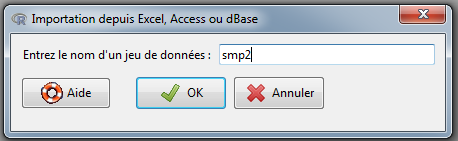
**Importation à partir d’un fichier Excel**

Pour importer un jeu de données à partir d’un fichier Excel, Access ou dBase : **Données** > **Importer des données** >**Depuis un fichier Excel, Access ou dBase**



*Importation d’un fichier Excel - Etape 1*

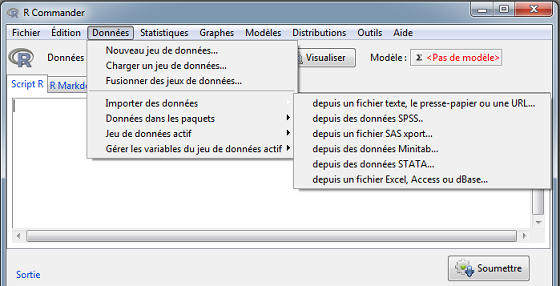
Il faut nommer le fichier qui vient d’être importé, par exemple en smp2, et parcourir vos documents pour chercher le jeu de données smp2.



*Importation d’un fichier Excel - Etape 2*

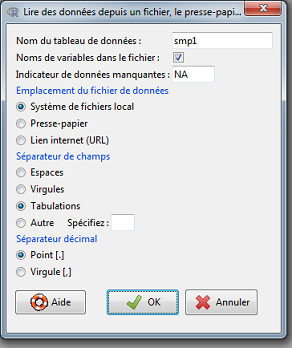
**Importation à partir d’un fichier CSV**

Pour importer un jeu de données à partir d’un fichier csv : **Données** > **Importer des données** > **Depuis un fichier texte, le presse-papiers ou URL**



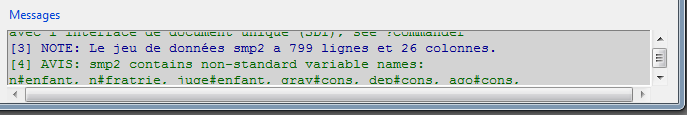
*Importation d’un fichier CSV - Etape 1*

Il faut nommer le fichier qui vient d’être importé, par exemple en smp2, et parcourir vos documents pour chercher le jeu de données smp2.

**

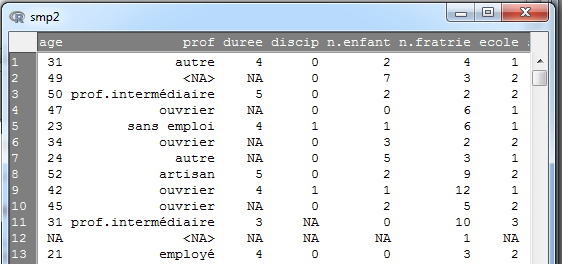
*Importation d’un fichier CSV - Etape 2*

Après l’importation du jeu de données dans Rcmdr, il est important de vérifier le nombre d’observations et de variables, afin de savoir s’il est bien adéquat avec le fichier initial.



*Vérification du nombre d’observations et de variables*

En cliquant sur Visualiser, il est possible d’apercevoir le jeu de données smp2 :



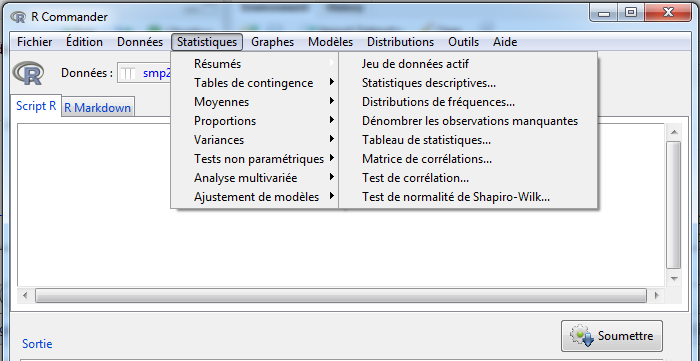
*Visualisation des variables et observations de la table smp2*

**Manipulation des données**

**Visualisation brève au jeu de données**

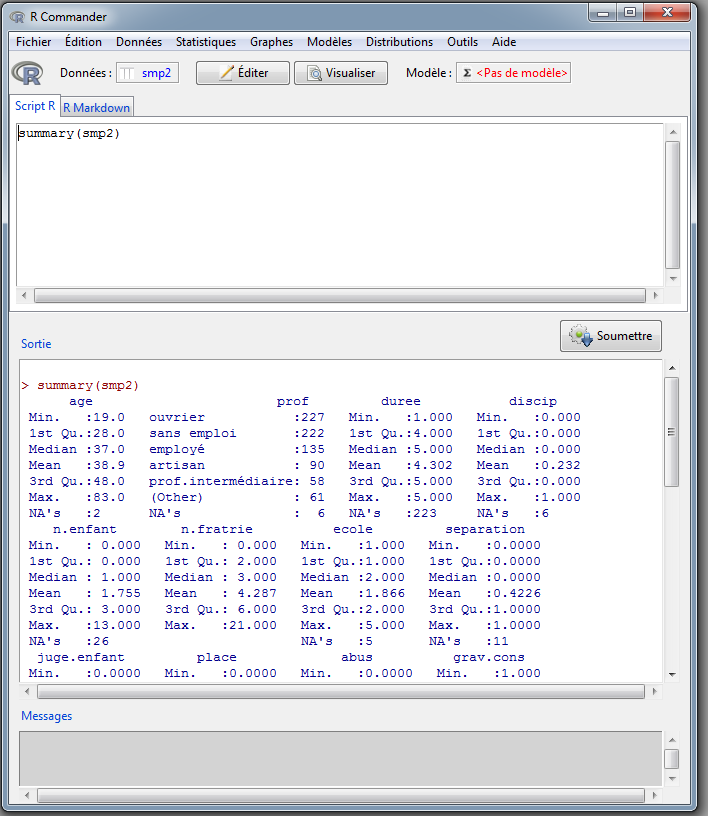
Une description brève du jeu de données importé peut être obtenue.

**Statistiques** > **Résumés** > **Jeu de données actif**



*Description de l’ensemble des variables du jeu de données smp2*

Pour chacune des variables du jeu de données, nous disposons d’indicateurs de positions (moyenne, médiane, quartiles). Attention, toutes les variables du jeu de données sont par défaut de type quantitatif. Nous verrons donc dans la partie suivante comment convertir ces variables en variables qualitatives.

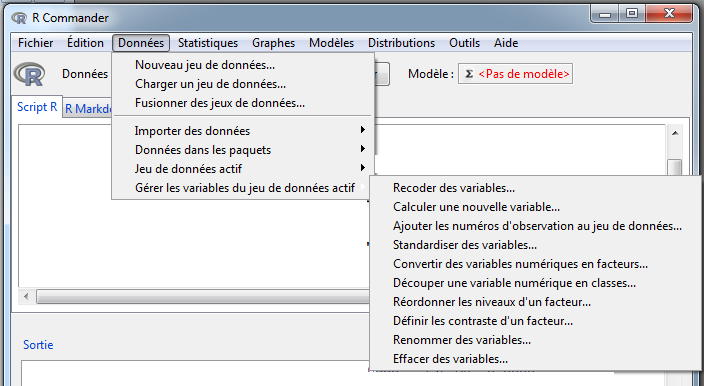


*Résultats: description du jeu de données smp2*

**Conversion des données quantitatives en qualitatives**

L’ensemble des variables issues du jeu de données importé sont de type quantitatif par défaut. Avant d’analyser le jeu de données, il faut donc convertir les variables quantitatives, qui sont supposées être qualitatives, en variables qualitatives.

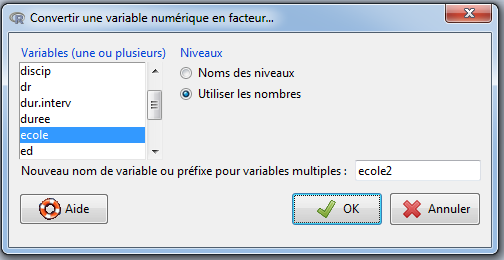
**Données** > **Gérer les variables du jeu de données actifs** > **Convertir des variables numériques en facteurs**



*Conversion de variables quantitatives en facteurs*

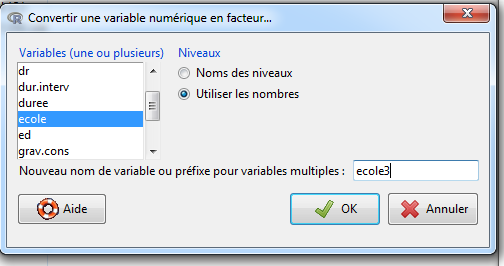
Par exemple, la variable “ecole” (niveau de formation actuel) est quantitative par défaut. Nous allons donc la convertir en variable qualitative en 5 classes.

La première possibilité est de transformer cette variable avec des modalités en chiffre “1”, “2”, “3”, “4”, “5” :

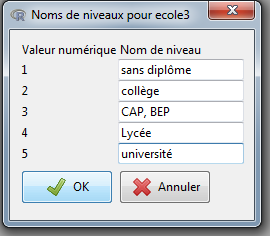


*Conversion de la variable ecole en facteurs - Etape 1*

La seconde possibilité est de transformer cette variable avec des modalités en texte : “sans diplôme”, “collège”, “CAP, BEP”, “Lycée”, “université” :



*Conversion de la variable ecole en facteurs - Etape 2*

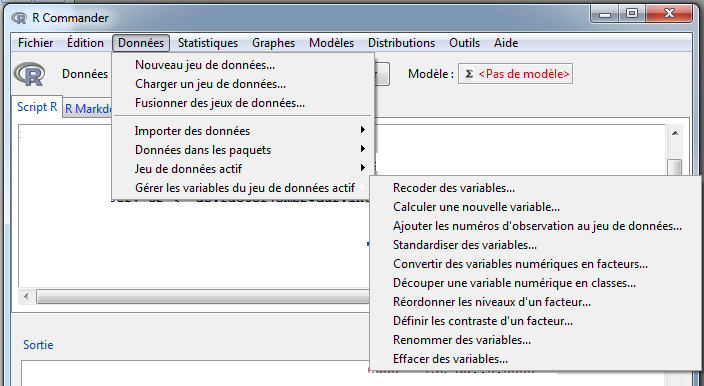


*Conversion de la variable ecole en facteurs - Etape 3*

**Recodage des données quantitatives en qualitatives**

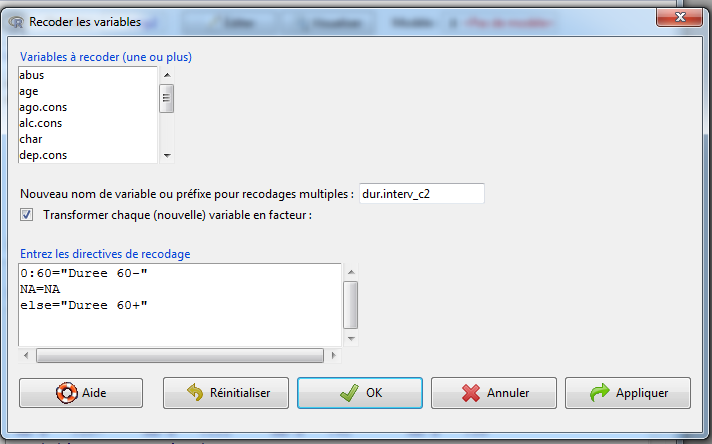
Certaines analyses demandent de recoder des variables quantitatives en variables qualitatives, à 2 ou plusieurs catégories. Lors du recodage, il faut faire attention aux données manquantes.

**Données** > **Gérer les variables du jeu de données actifs** > **Recoder des variables**



*Recodage des variables*

Dans cet exemple, nous allons recoder la variable quantitative durée d’interview (dur.interv) en variable qualitative à 3 classes (dur.interv\_c2). Si dur.interv est compris entre 0 et 60 minutes alors dur.interv\_c2=“Duree 60-”, si dur.interv est vide alors dur.interv\_c2=NA, sinon dur.interv\_c2 = “Duree 60+”.

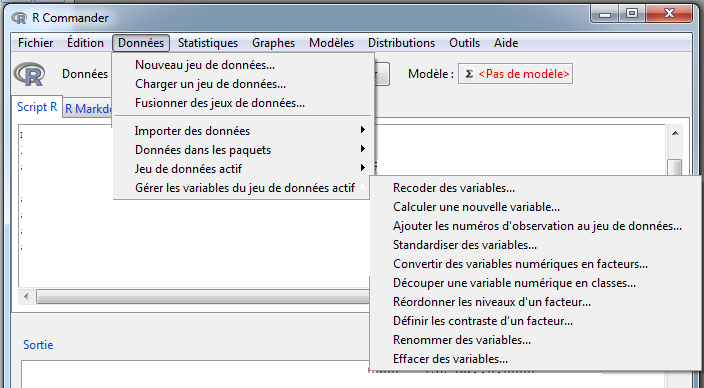


*Résultats: recodage de la variable dur.interv*

**Création de nouvelles variables**

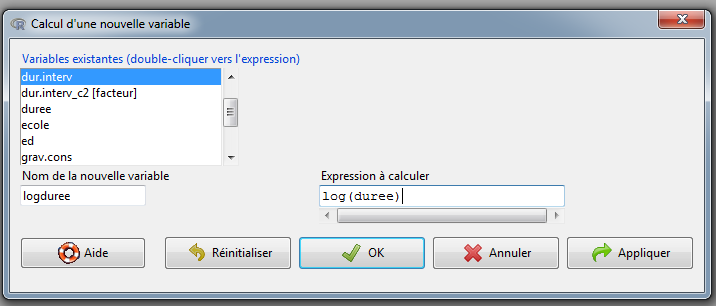
De nouvelles variables peuvent être créées à partir d’autres variables à l’aide de fonctions mathématiques : des opérateurs (+, -, \*, /, ^…) ou des fonctions (log, exp, sin, cos, tan…).

**Données** > **Gérer les variables du jeu de données actifs** > **Calculer une nouvelle variable**



*Création de variables*

Pour exemple, nous allons créer la variable log(duree), qui représente le logarithme de la durée d’interview :



*Résultats: création de la variable log(duree)*

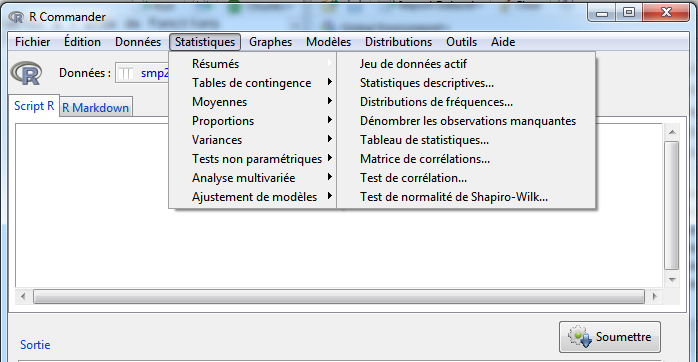
**Description des données**

**Distribution des variables quantitatives et qualitatives**

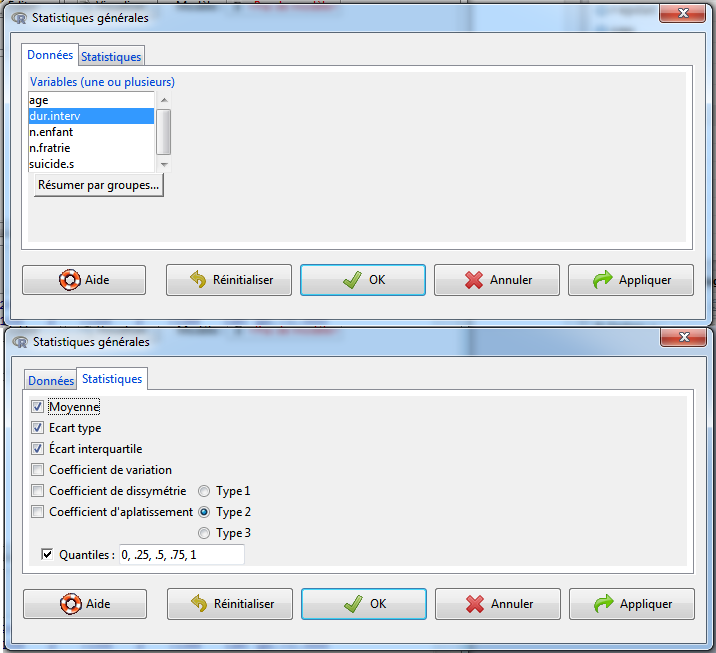
Dans une étude, il est important de décrire les variables de son jeu de données.

Pour les variables quantitatives, il est intéressant d’obtenir des moyennes, écart-types, médiane…

**Statistiques** > **Résumés** > **Statistiques descriptives**

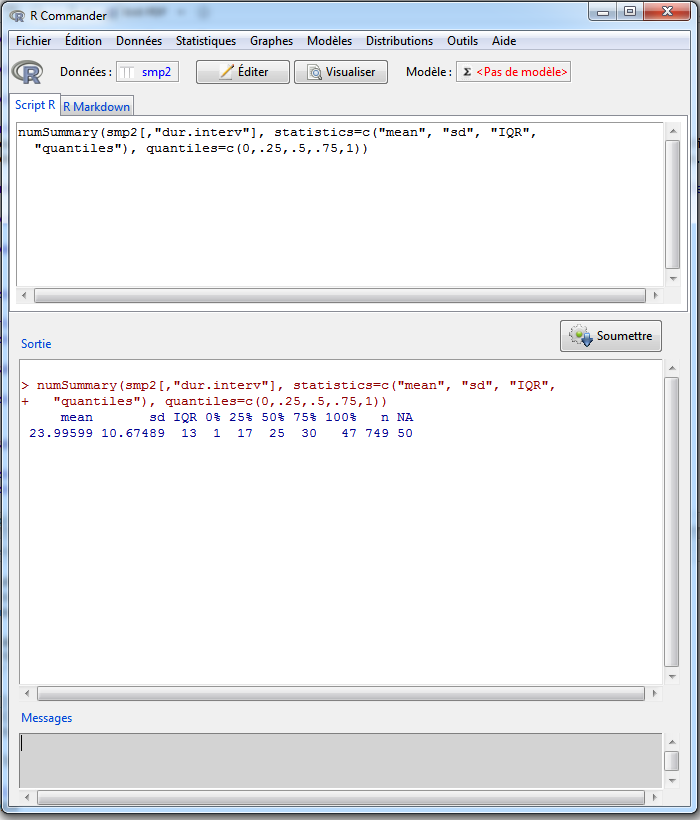


*Description des variables quantitatives du jeu de données smp2 - Etape 1*



*Description des variables quantitatives du jeu de données smp2 - Etape 2*

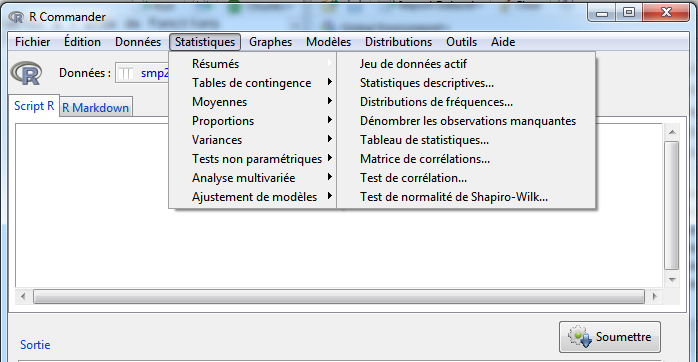
En moyenne, la durée d’interview est de 23.99 minutes (+/- 10 écart-types). La médiane de la durée d’interview est de 25 minutes, c’est à dire que la moitié de la population a une durée d’interview de 25 minutes.



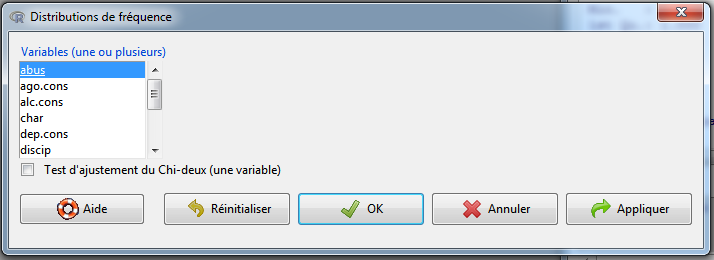
*Résultats: description de la variable dur.interv issue du jeu de données smp2*

Pour les variables qualitatives, il est intéressant d’obtenir des proportions et des intervalles de confiance.

**Statistiques** > **Résumés** > **Distribution de fréquence**

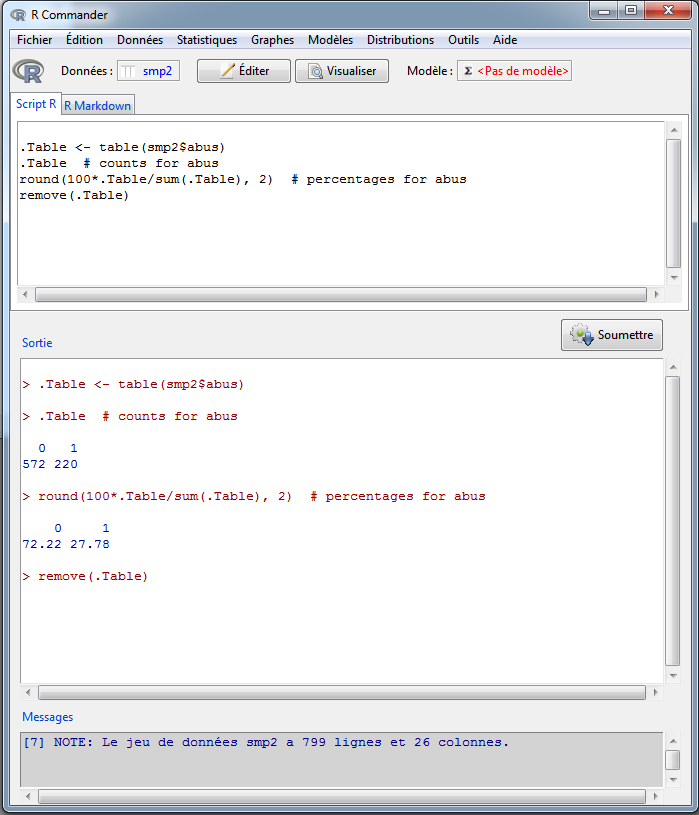


*Description des variables qualitatives du jeu de données smp2 - Etape 1*



*Description des variables qualitatives du jeu de données smp2 - Etape 2*

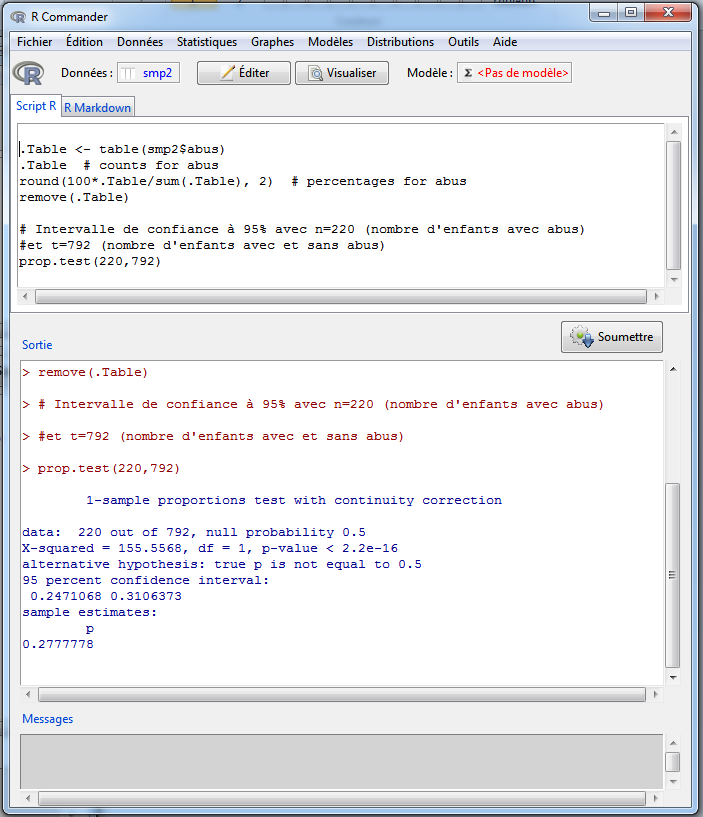
La proportion de détenus ayant subi des maltraitances pendant l’enfance est de 27,78% (220).



*Résultats: description de la variables abus issue du jeu de données smp2 (proportion)*

Pour obtenir l’intervalle de confiance d’une proportion, il n’existe pas de commande sous Rcmdr. Il faut taper et soumettre la procédure suivante dans le script: prop.test(n,t) où n est le nombre de cas et t est le nombre total d’individus concernés par la variable testée.

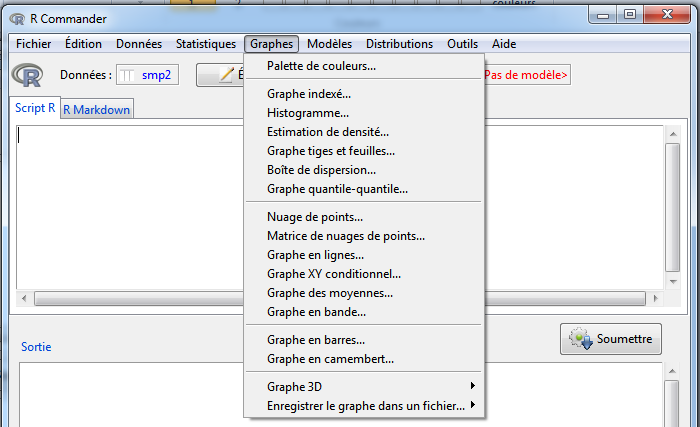
… avec un IC95% [24,71% - 31.06%]



*Résultats: description de la variable abus issue du jeu de données smp2 (IC)*

**Représentation des variables quantitatives et qualitatives**

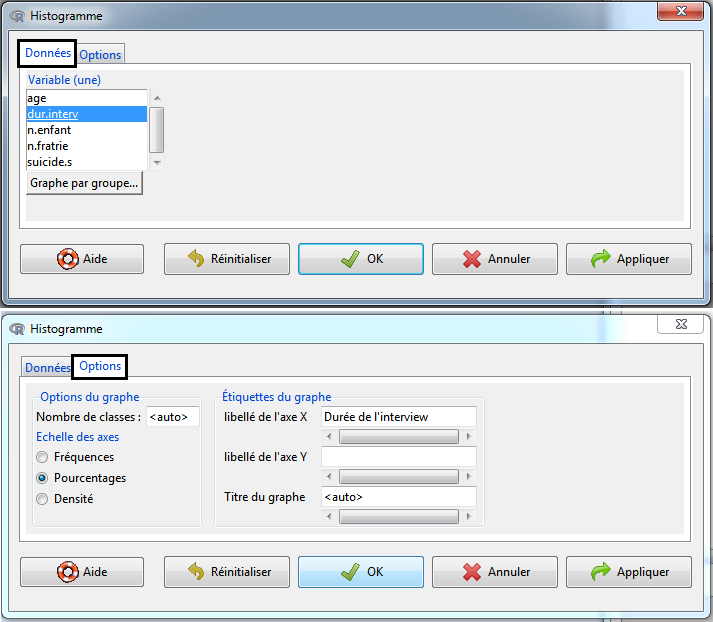
La distribution des variables quantitatives va être illustrée par des histogrammes ou des boxplots et celle des variables qualitatives par des diagrammes.



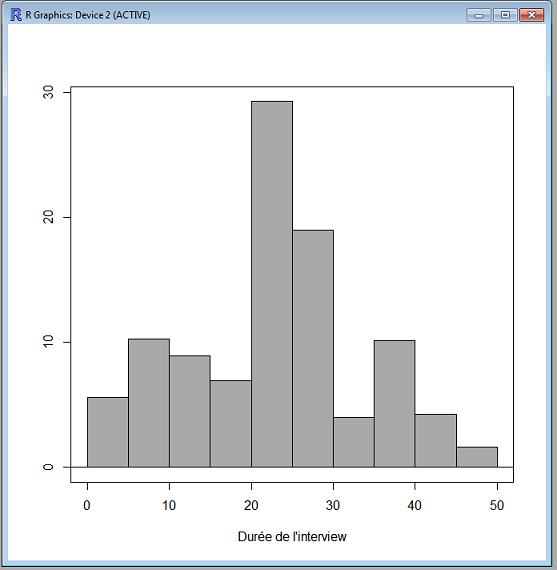
*Réalisation de graphiques de différents types de variables*

**Graphes** > **Histogramme**

Par exemple, la distribution de la durée d’interview (dur.interv) est représentée par un histogramme. Dans les options, il est possible de choisir quel type de données nous intéresse (effectifs, pourcentages, densités), de renommer les libellés des axes et de donner un titre au graphique.



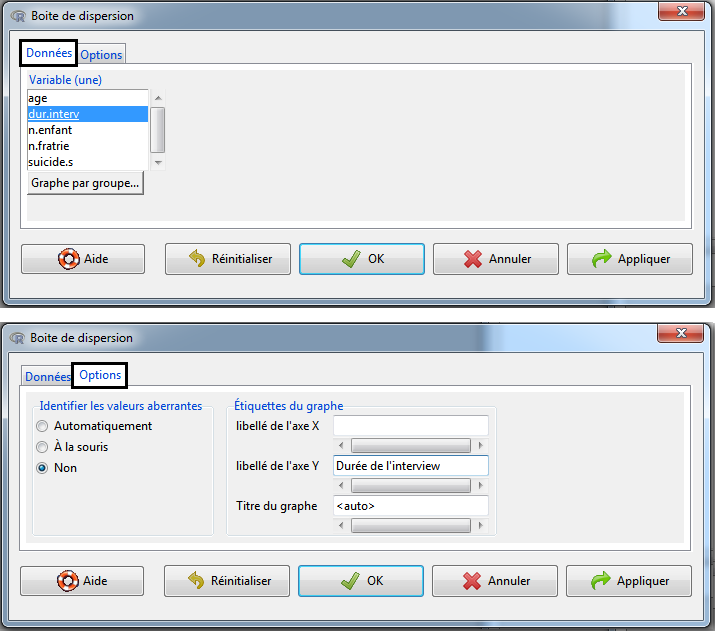
*Réalisation d’un histogramme d’une variable quantitative*



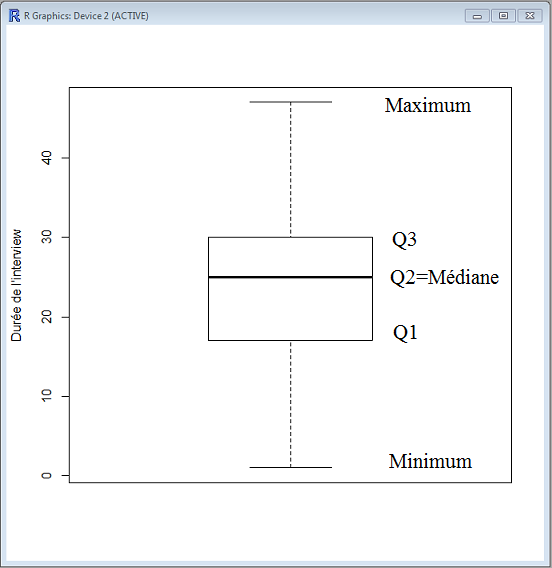
*Résultats : histogramme de la variable dur.interv*

**Graphes** > **Boite de dispersion**

La distribution de la durée d’interview peut être représentée également par un boxplot.



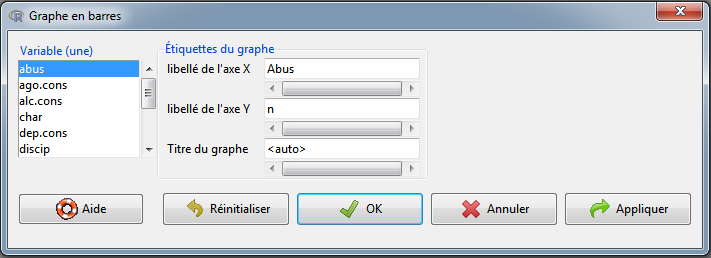
*Réalisation d’un boxplot d’une variable quantitative*



*Résultats : boxplot de la variable dur.interv*

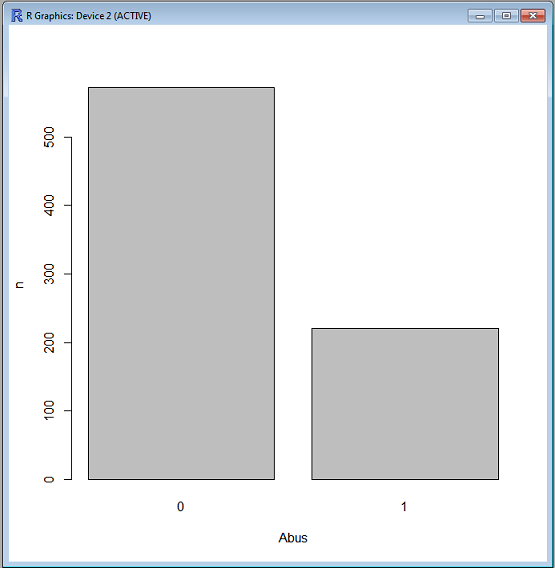
**Graphes** > **Graphes en barres**

Nous allons représenter la variable abus (abus) par un diagramme en barres. Il est possible de renommer les libellés des axes et de donner un titre au graphique.



*Réalisation d’un diagramme d’une variable qualitative*

Sur cet exemple, nous constatons que les détenus ayant subi des maltraitances pendant l’enfance sont moins nombreux.



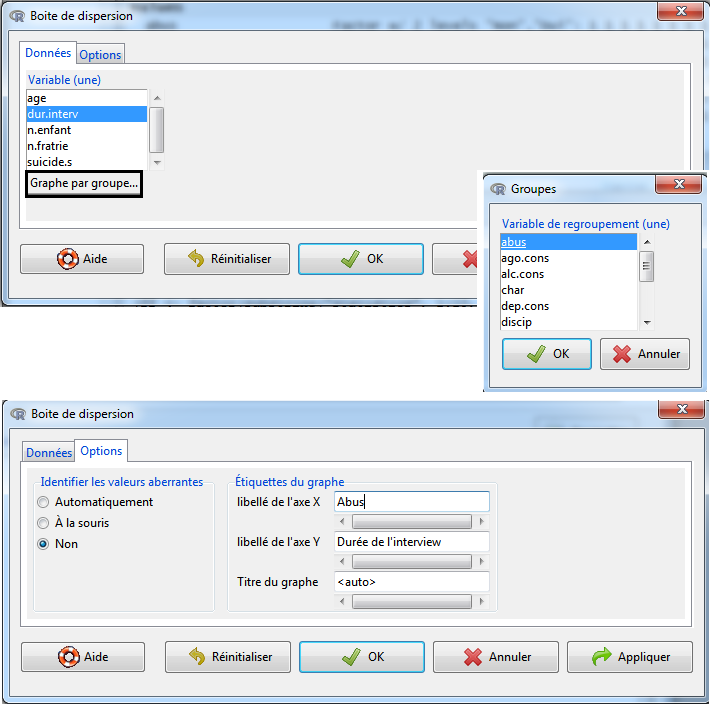
*Résultats : diagramme de la variable abus*

Pour représenter le diagramme en barre en pourcentage, il faut préalablement créer une variable abus en pourcentage.

**Représentation de variables quantitatives en fonction d’une variable qualitative**

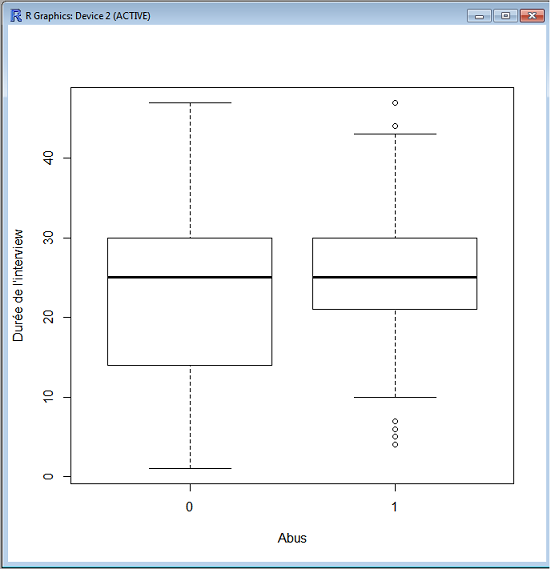
Nous pouvons illustrer par un boxplot si le fait d’avoir subi des maltraitances pendant l’enfance pouvait entraîner des écarts dans la durée de l’interview.

**Graphes** > **Boite de dispersion**



*Réalisation d’un diagramme d’une variable quantitative en fonction d’une variable qualitative*

La médiane entre les deux groupes semble identique.



*Résultats - diagramme de la variable dur.interv en fonction de la variable abus*

**Tests statistiques**

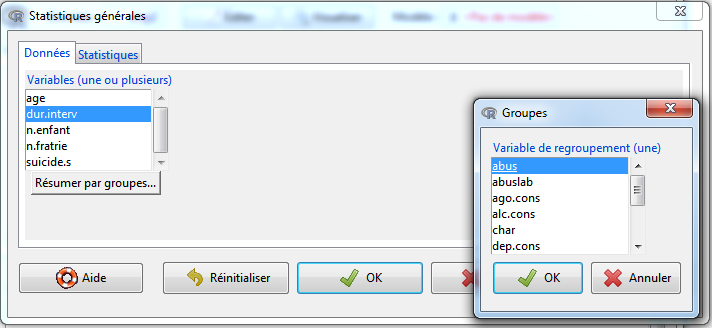
**Comparaison de moyennes d’une variable quantitative entre deux groupes**

Nous allons comparer statistiquement la différence de durée d’interview entre les détenus ayant subi des maltraitances pendant l’enfance et ceux n’ayant jamais subi de maltraitance pendant l’enfance.

1. Effectifs par groupe (n1 et n2)

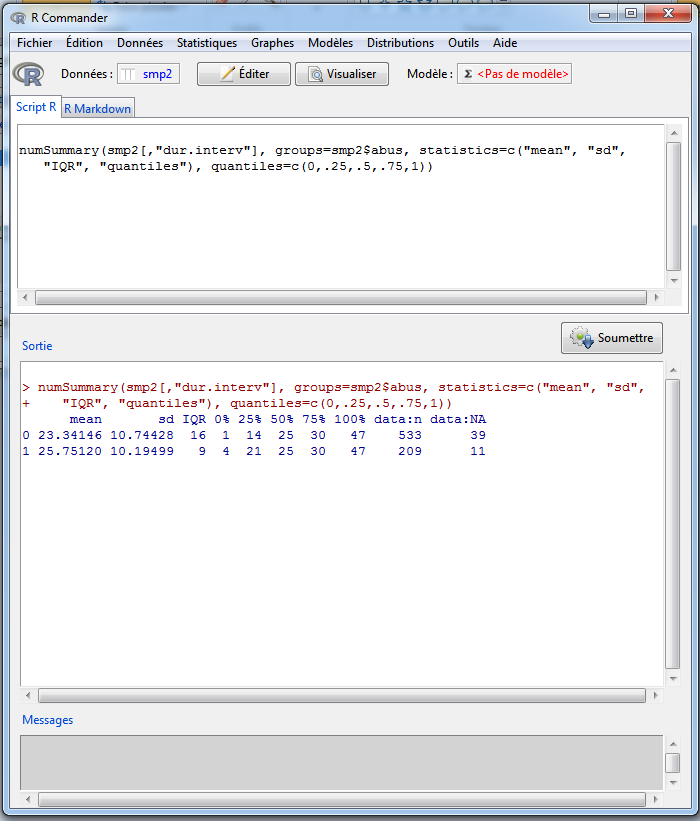
Nous allons regarder la moyenne de durée d’interview entre les deux groupes ainsi que les effectifs par groupe

**Statistiques** > **Résumés** > **Statistiques descriptives**



*Calcul des moyennes par groupe*

Les effectifs des deux groupes sont supérieurs à 30. Nous montrons que la médiane dans les deux groupes est de 25 minutes, et avec une moyenne de 23.34 minutes (+/- 10.7) dans le groupe “non abus” et de 25.75 minutes (+/- 10.2) dans le groupe “abus”.

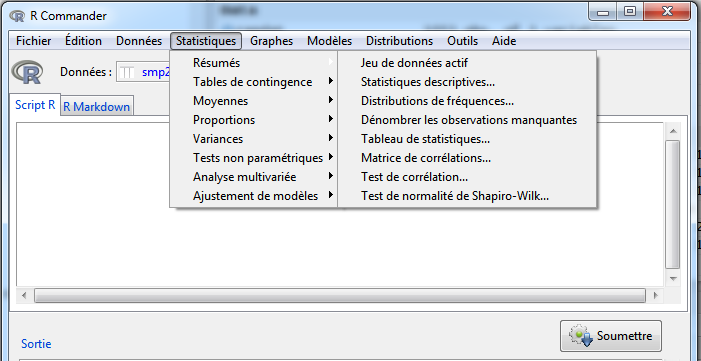


*Résultats: moyenne de la variable dur.interv chez les détenus ayant subi des maltraitances et ceux qui n’en ont pas subi*

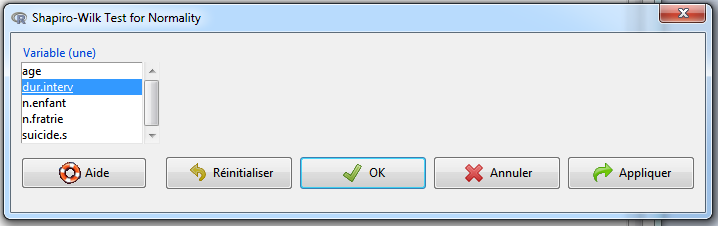
1. Tester la normalité de la distribution de la variable quantitative (si n1 et n2 <= 30)

Nous allons tester la normalité de la distribution de la durée d’interview.

**Statistiques** > **Résumés** > **Tests de normalité de Shapiro-Wilk**

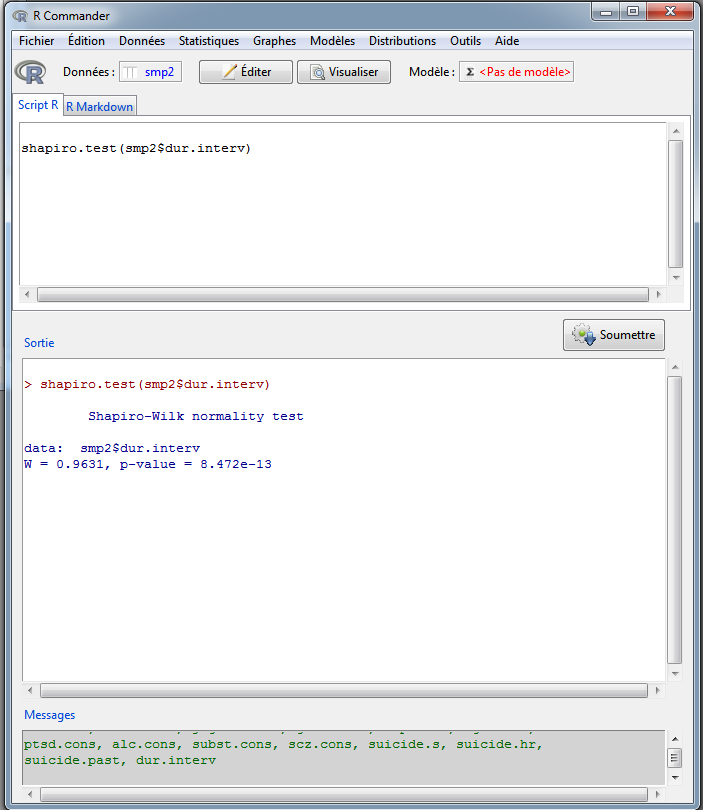


*Réalisation d’un test de normalité d’une variable quantitative - Etape 1*



*Réalisation d’un test de normalité d’une variable quantitative - Etape 2*

Le test de Shapiro est significatif (p<0.05). Cela signifie que la variable ne suit pas une loi normale.



*Résultats: test de normalité de la variable dur.interv*

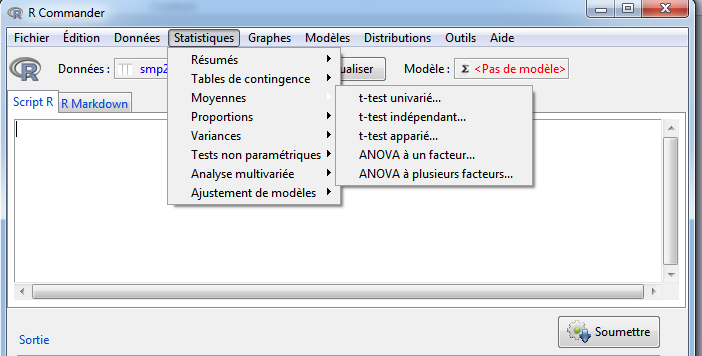
1. Comparaison de moyennes d’une variable quantitative entre deux groupes

Dans notre exemple, les effectifs dans chaque groupe sont supérieurs à 30, donc nous pouvons réaliser un test de Student.

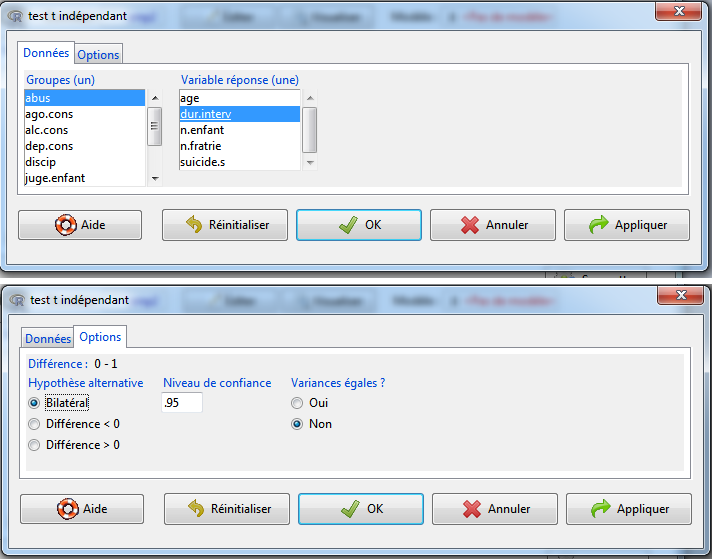
Si les effectifs étaient inférieurs à 30 et que la durée d’interview suivait une loi normale, alors le test de Student serait toujours valable (seuls les degrés de liberté et la loi seraient modifiés). Si les effectifs étaient inférieurs à 30 et que la durée d’interview ne suivait pas suivi une loi normale, alors le test non paramétrique de Mann-Whitney serait recommandé.

a- Test de Student

**Statistiques** > **Moyennes** > **t-test indépendant**

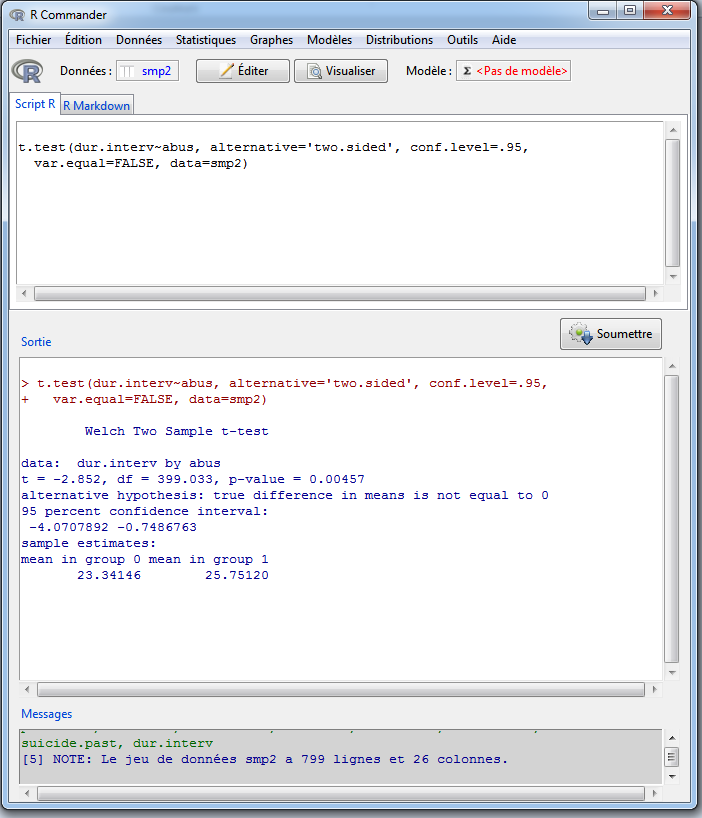


*Réalisation du test de Student pour comparer les moyennes entre deux groupes - Etape 1*



*Réalisation du test de Student pour comparer les moyennes entre deux groupes - Etape 2*

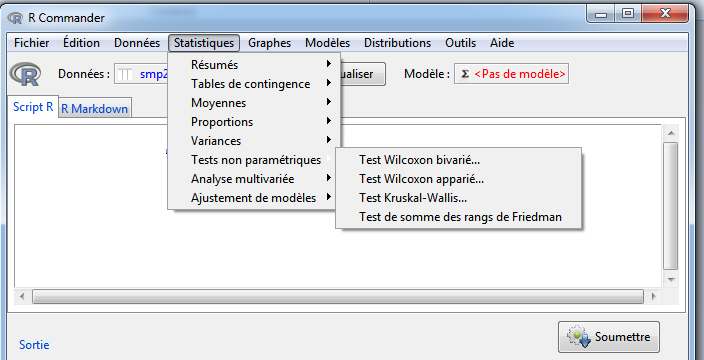
Le test indique qu’il y a une différence significative (p<0.05) de la durée d’interview entre les deux groupes.



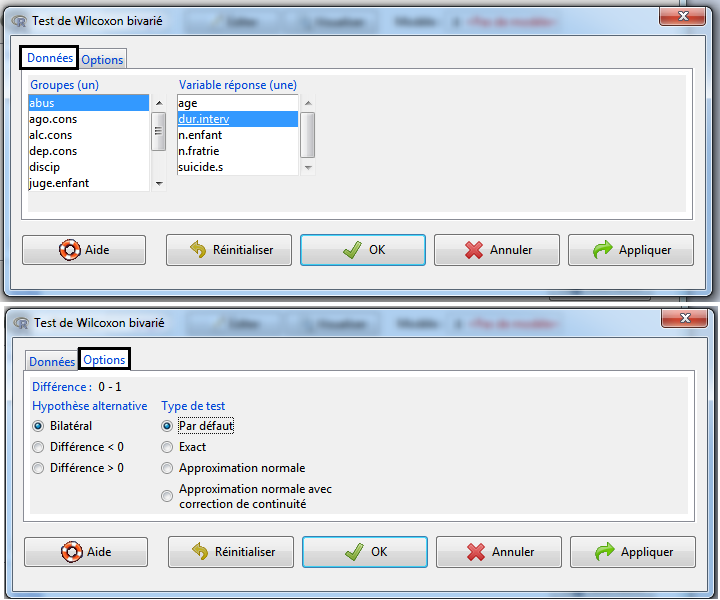
*Résultats : test de Student pour comparer les moyennes de la variable dur.interv entre les détenus ayant subi des maltraitances et ceux qui n’en ont pas subi*

b- Test non paramétrique de Wilcoxon

**Statistiques** > **Tests non paramétriques** > **Test de Wilcoxon bivarié**



*Réalisation du test de Wilcoxon pour comparer les moyennes entre deux groupes - Etape 1*



*Réalisation du test de Wilcoxon pour comparer les moyennes entre deux groupes - Etape 2*

Le test indique qu’il y a une différence significative (p<0.05) de la durée d’interview entre les deux groupes.

(Ce test est réalisé juste pour l’exemple et n’est pas recevable)

***Comparaison de proportions d’une variable qualitative entre deux groupes***

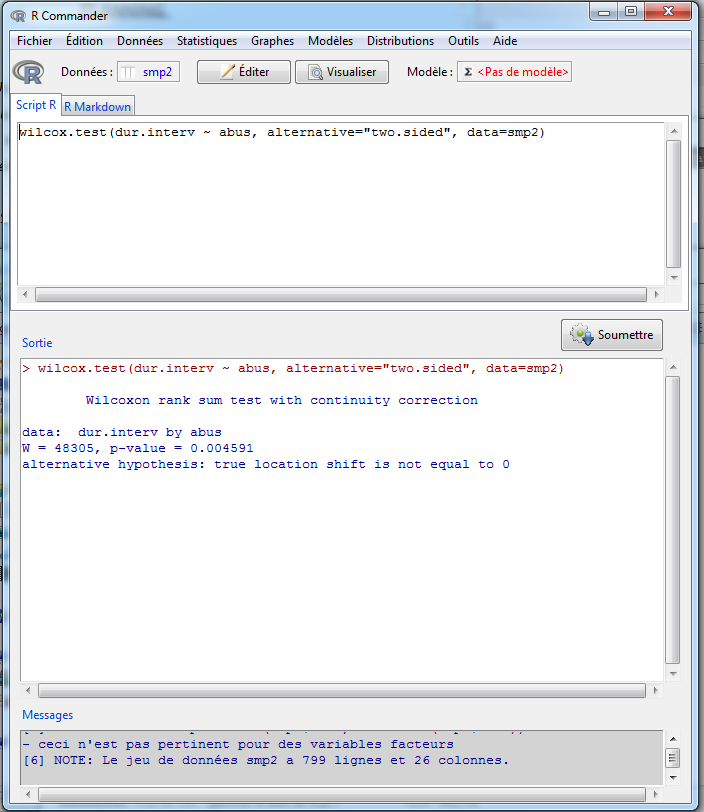
Nous allons comparer statistiquement l’association entre le fait de subir des maltraitances pendant l’enfance et l’existence d’un trouble dépressif.

1. Comparaison de proportions d’une variable qualitative entre deux groupes

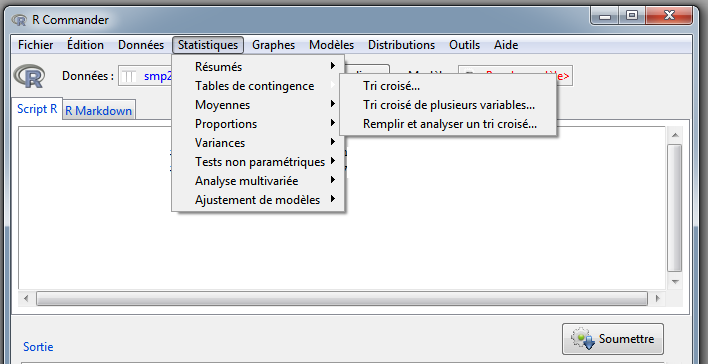
Dans notre exemple, le test du khi² est faisable si et seulement si les effectifs espérés sont supérieurs à 5, sinon le test de Fisher est recommandé.

* Test du Khi²

**Statistiques** > **Tables de contingences** > **Tri croisé**



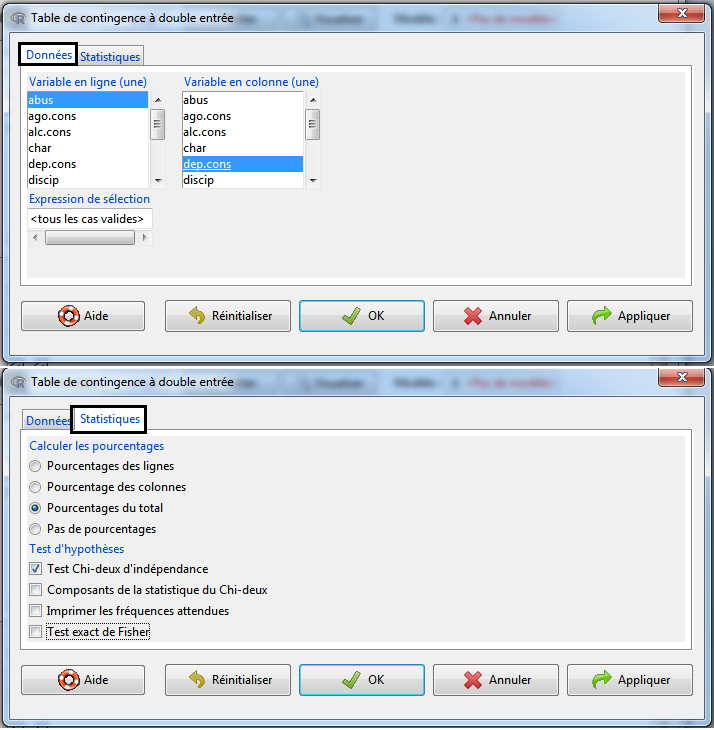
*Réalisation du test du Khi² pour comparer les proportions entre deux variables quantitatives - Etape 1*



*Réalisation du test du Khi² pour comparer les proportions entre deux variables quantitatives - Etape 2*

Les effectifs théoriques sont bien supérieurs à 5, le test du Khi² est donc recevable.

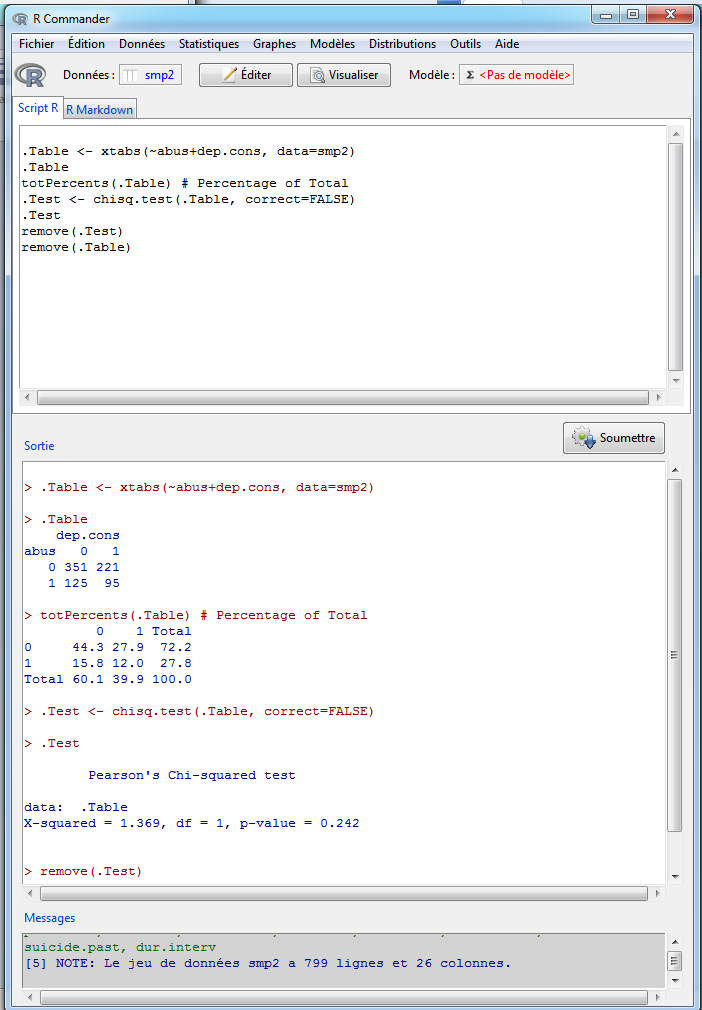
Ce test indique qu’il n’y a pas d’association entre le fait de subir des maltraitances pendant l’enfance et l’existence d’un trouble dépressif (p>0.05).



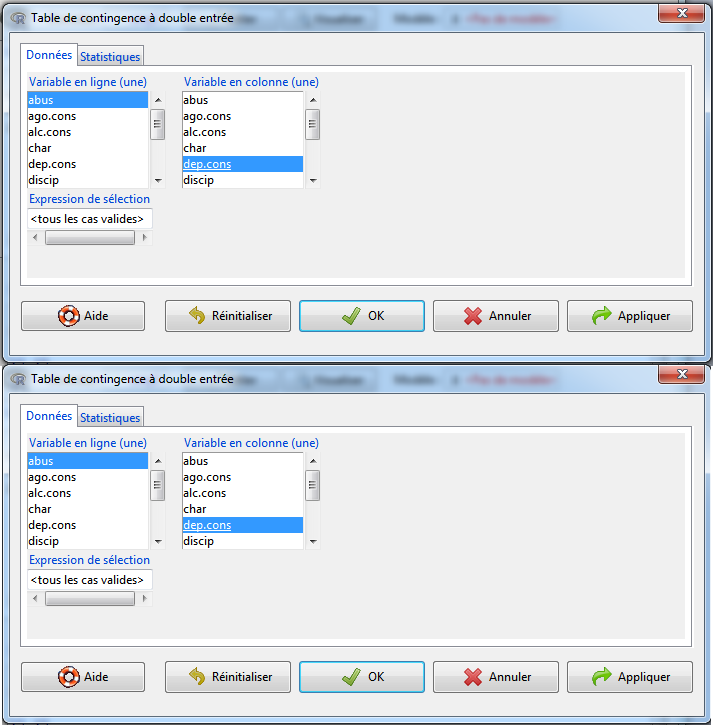
*Résultats: test du Khi² pour comparer les proportions entre la variable abus et dep.cons*

* Test non paramétrique de Fisher

**Statistiques** > **Tables de contingences** > **Tri croisé**

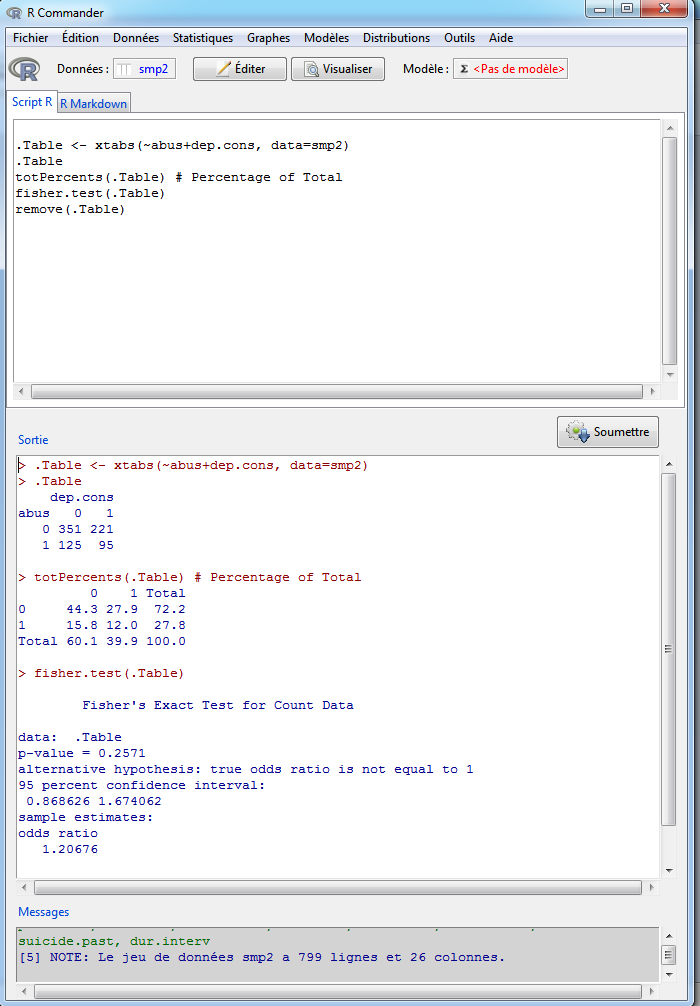


*Réalisation du test de Fisher pour comparer les proportions entre deux variables quantitatives - Etape 1*



*Réalisation du test de Fisher pour comparer les proportions entre deux variables quantitatives - Etape 1 - Etape 2*

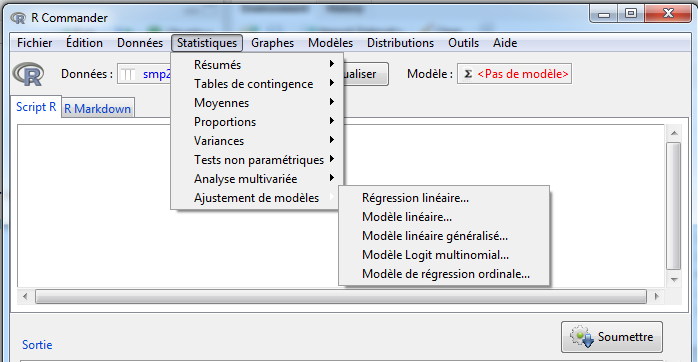
Ce test indique qu’il n’y a pas d’association entre le fait de subir des maltraitances pendant l’enfance et l’existence d’un trouble dépressif (p>0.05).



(Ce test est réalisé juste pour l’exemple et n’est pas recevable)

**Modèles statistiques**

**Statistiques** > **Ajustement de modèles**



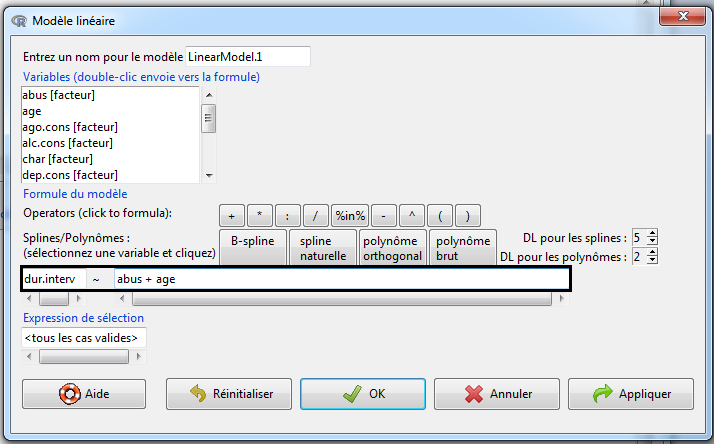
*Modèles linéaires et logistiques*

**Modèles linéaires**

Nous allons étudier l’association entre la durée de l’interview et l’abus, ajusté sur l’âge du détenu

**Statistiques** > **Ajustement de modèles** > **Modèle linéaire**

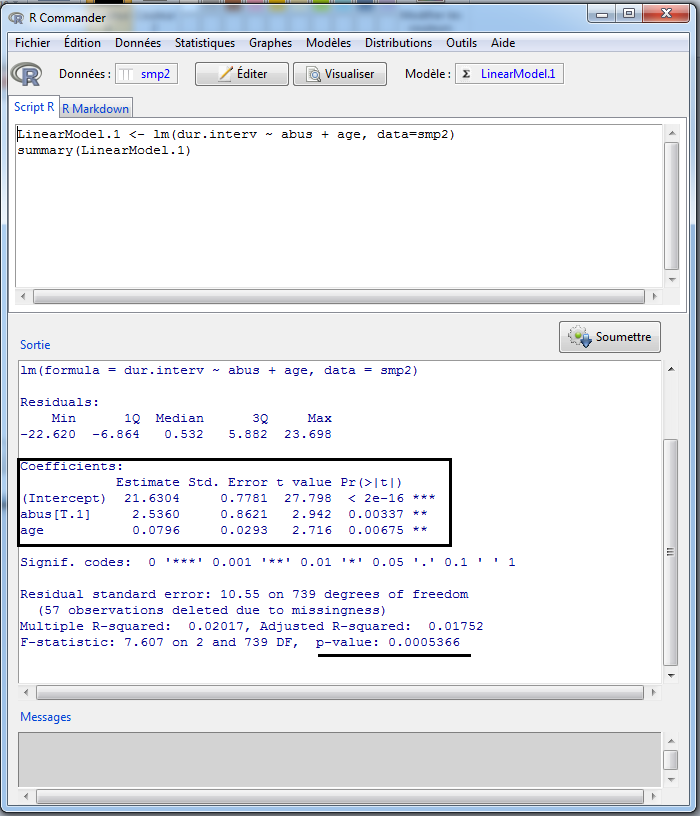
Il faut double-cliquer sur les variables afin qu’elles s’affichent dans la formule du modèle.



*Réalisation d’un modèle linéaire entre une variable quantitative et une variable qualitative*

Le p-value global du modèle est de 0,0005366. Cela signifie qu’une des variables (abus ou âge) est significative.

Il y a une association significative entre la durée d’interview et abus (p=0,003), après ajustement sur l’âge du détenu.



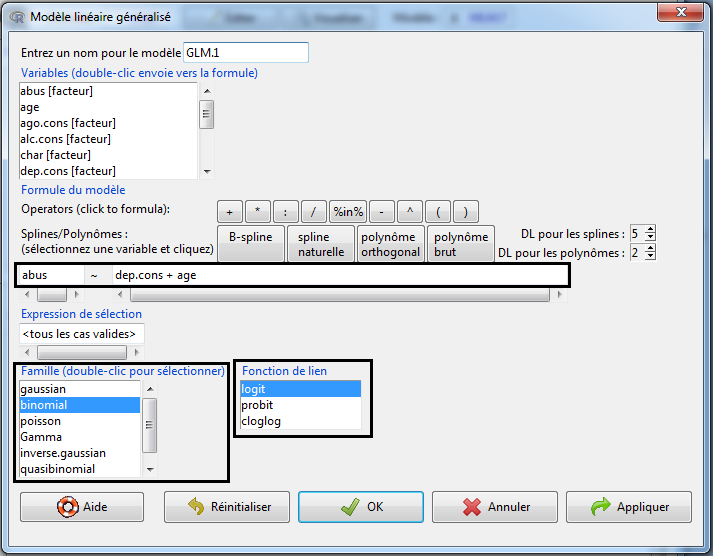
*Résultats: modèle linéaire entre la variable dur.interv et abus*

**Modèles logistiques**

Nous allons étudier entre le fait de subir des maltraitances pendant l’enfance et l’existence d’un trouble dépressif, ajusté sur l’âge du détenu

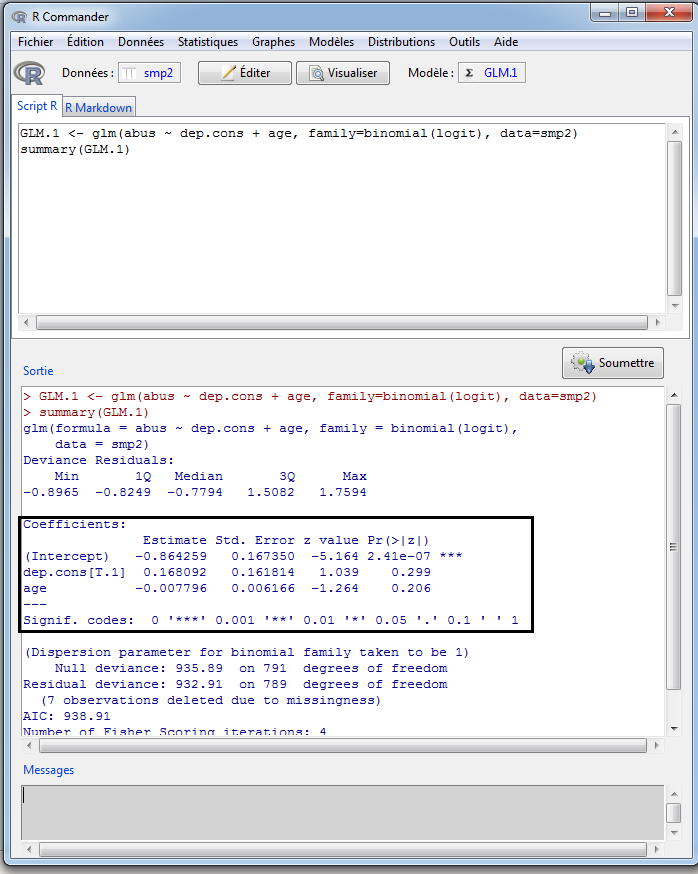
**Statistiques** > **Ajustement de modèles** > **Modèle linéaire généralisé**

Il faut double-cliquer sur les variables afin qu’elles s’affichent dans la formule du modèle. Pour indiquer qu’il s’agit d’un modèle logistique, il faut indiquer que la famille est binomiale et la fonction de lien logit.



*Réalisation d’un modèle linéaire entre deux variables qualitatives*

Il n’y a pas association significative entre le fait de subir des maltraitances pendant l’enfance et l’existence d’un trouble dépressif, après ajustement sur l’âge du détenu.



*Résultats: modèle logistique entre la variable abus et dep.cons*