TD de Statistiques - Séance N°5

Introduction à R et à R Commander

1 R et R Commander : où les télécharger, comment les installer ?

1.1 Installer R et R Commander sur votre machine personnelle (sous Windows)

A la différence de la plupart des logiciels de traitements statistiques, R est un logiciel libre. Vous pouvez donc l'installer sur votre machine et utiliser une version similaire à celle de nos salles de TD. Les manipulations décrites ci-dessous s'appliquent à R pour Windows, mais le logiciel existe également pour Mac OS et Linux et les manipulations seront tout à fait similaires si vous utilisez l'un de ces systèmes d'exploitation.

N.B. La version courante de R au moment où ce paragraphe est rédigé est la version 2.15.0. Il est vraisemblable qu'une version plus récente sera disponible lorsque vous installerez R, mais seuls les noms de fichiers devraient être légèrement modifiés.

A l'aide d'un navigateur, affichez le site :
http://www.r-project.org
Dans la partie gauche de l'écran, allez sur la rubrique **Download, packages** et cliquez sur le lien CRAN.
Choisissez ensuite l'un des sites français de téléchargement de R et de ses packages, par exemple :
http://cran.univ-lyon1.fr/

La page d'accueil de ce dernier site s'affiche alors.

N.B. Vous pouvez également vous rendre directement sur la page de l'Université de Lyon, http://cran.univ-lyon1.fr/ sans passer par la page www.r-project.org

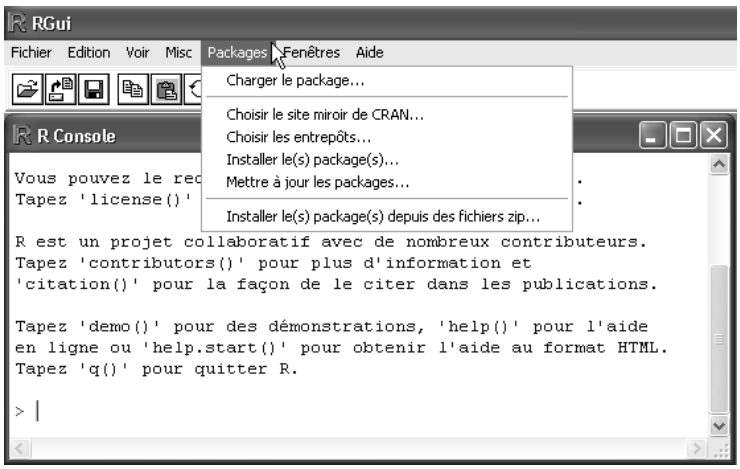
Cliquez sur le lien : Download R for Windows
puis sur le lien : Install R for the first time
et enfin sur le lien : Download R 2.15.1 for Windows

Une fois le téléchargement effectué, allez dans le répertoire où le fichier a été enregistré et exécutez (double-clic) le programme d'installation : **R-2.15.1-win.exe**

Sélectionnez le français comme langue pour l'installateur et les menus de R. Acceptez les choix par défaut de l'installateur pour les autres options.

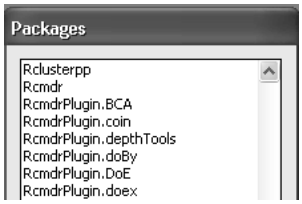
Une fois l'installation terminée, exécutez le logiciel en cliquant sur l'icône **R-i386-2.15.0**

La console de R (fenêtre **Rgui**) est alors chargée.
Utilisez le menu **Packages - Installer les packages**.



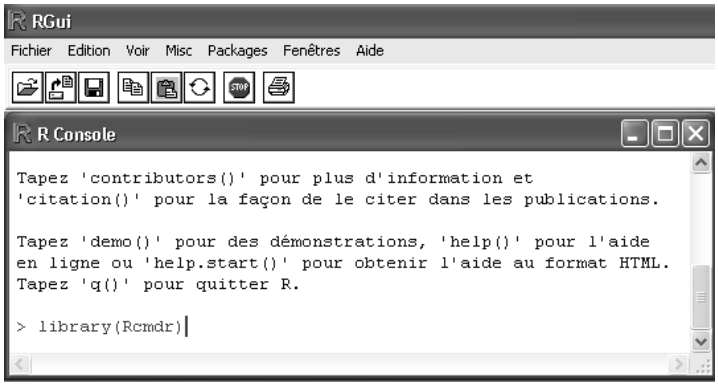
Le logiciel demande de choisir dans une liste un site de téléchargement. Sélectionnez : **France Lyon 1**.

La (longue) liste des packages s'affiche alors. Sélectionnez le package **Rcmdr**.



Une fois le package installé, essayez de le lancer en saisissant dans la console de R la ligne de commande suivante (en respectant majuscules et minuscules) :

```
library(Rcmdr)
```

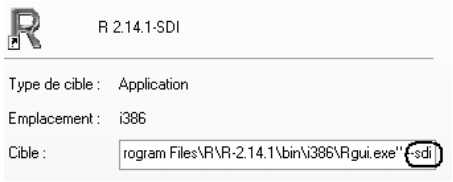


Le logiciel affiche alors une fenêtre d'avertissement indiquant que le fonctionnement correct de Rcmdr exige l'installation de 14 autres packages (car, sem, rgl, relimp, multcomp, lmtest, leaps, Hmisc, e1071, effects, colorspace, aplpack, abind, RODBC) et propose de les installer.

Cliquez sur **Installer**, puis sur **Installer depuis CRAN**.

Une fois l'installation de ces packages effectuée, l'interface Rcmdr devrait s'afficher en plus de celle de R.

N.B. On pourra faciliter le chargement de R en créant un raccourci vers le programme Rgui.exe sur le bureau. On peut alors améliorer le fonctionnement de R Commander en modifiant la ligne de commande associée à ce raccourci.
- Faites un clic droit sur le raccourci et sélectionnez l'item de menu "Propriétés".
- Ajoutez **`--sdi`** à la fin de la ligne de commande, dans le champ "Cible" :



1.2 Installer R et R Commander sur votre machine personnelle (Linux ou Mac OSX)

Dans les grandes lignes, l'installation se déroule de la même façon que sous Windows. On notera cependant que, pour faire fonctionner R Commander sous Mac OSX, il est nécessaire que X Windows soit chargé en mémoire avant R. Ce logiciel est fourni en standard avec le système sur les CD/DVD Apple, mais n'est pas installé par défaut. La première étape consiste donc à l'installer sur votre machine.

1.3 Installer R Commander sur votre compte dans les salles de TD

On peut exécuter R en sélectionnant le programme dans le menu Tous les programmes - Utilitaires - R (32 bits). On aura cependant un fonctionnement amélioré en copiant sur le bureau le raccourci Ri386-2.15.1-sdi depuis la page Web donnant accès aux fichiers du TD.

Dans ce raccourci, la ligne de commande **`"C:\Program Files\R\R-2.15.1\bin\i386\Rgui.exe"`** a été complétée par :

```
--sdi http_proxy=http://proxy.univ-brest.fr:3128
```

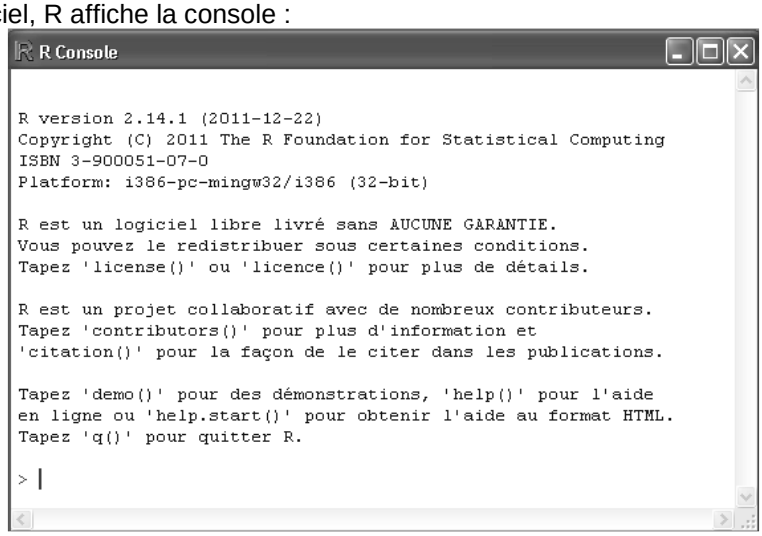
La dernière partie de la ligne de commande permet d'avoir accès, via le proxy de l'UBO, aux sites d'installation des packages.

2 R et R Commander : l'interface utilisateur

La plupart des copies d'écran indiquées ci-dessous ont été faites avec R version 2.14.1 et R Commander Version 1.7-3

2.1 Les fenêtres de travail

Au chargement du logiciel, R affiche la console :



On dispose alors d'un outil permettant de faire des calculs en mode "ligne de commande".
Essayez par exemple :

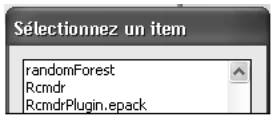
```
5 + 3
mean(c(1,4,8))
```

Pour faciliter l'utilisation de cet outil, nous utilisons une surcouche logicielle fournissant une interface utilisateur plus conviviale : R Commander.

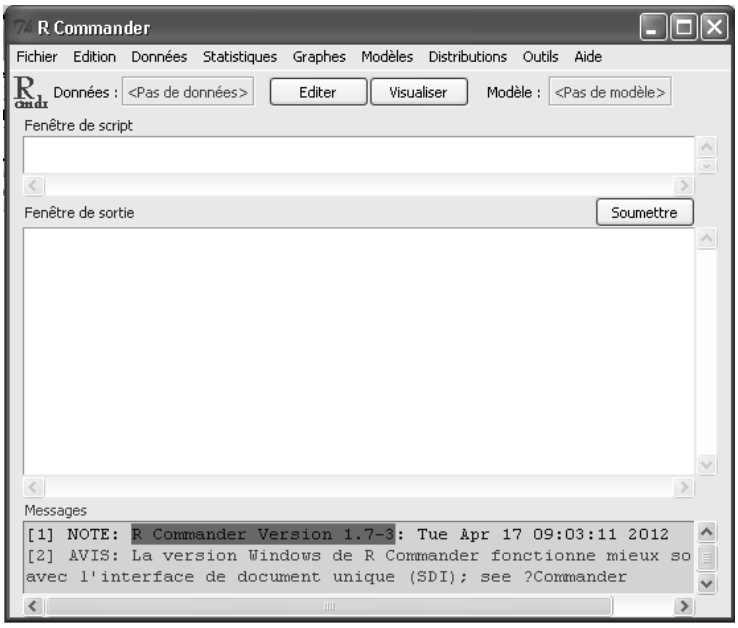
Pour activer R Commander, saisissez dans la console la commande :

```
library(Rcmdr)
```

ou, de manière alternative, utilisez le menu Packages > Charger le package et sélectionnez Rcmdr dans la liste qui s'affiche :



S'affiche alors une fenêtre disposant d'une barre de menu, d'une barre d'outils, d'une fenêtre de script, d'une fenêtre de sortie et d'une fenêtre de messages.



2.2 L'aide en ligne et ses particularités de fonctionnement dans nos salles de TD

Sous Windows, pour afficher les pages d'aide (malheureusement en anglais), R utilise un serveur Web et le navigateur défini par défaut dans le système.

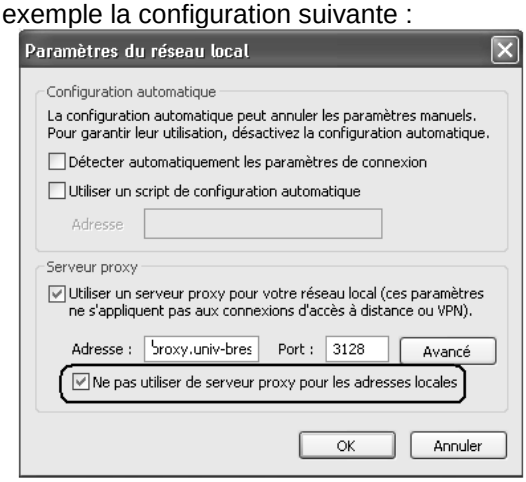
Essayez par exemple de saisir dans la console :

```
help(mean)
```

Si tout se passe bien, une page s'affiche dans Internet Explorer ou Firefox.

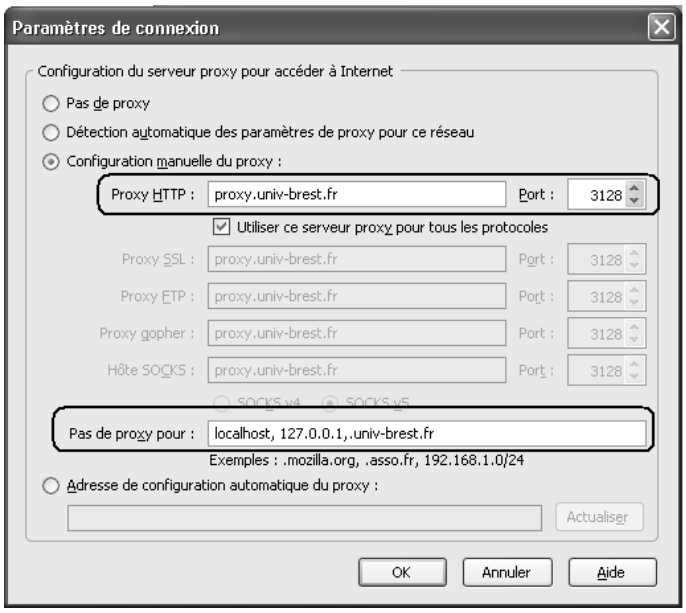
Mais, il se peut que la page refuse de s'afficher, car le serveur Web activé par R n'est pas reconnu par le proxy activé automatiquement dans nos salles. Dans ce cas :

- Avec Internet Explorer, utilisez le menu Outils - Options Internet puis l'onglet Connexion et le bouton Paramètres Réseau. Utilisez par exemple la configuration suivante :



N.B. L'adresse complète du serveur Proxy est : proxy.univ-brest.fr

- Avec Firefox, utilisez le menu Outils - Options puis l'icône Avancé, l'onglet Réseau et le bouton Paramètres. Activez alors le bouton "Pas de proxy" ou utilisez la configuration suivante :



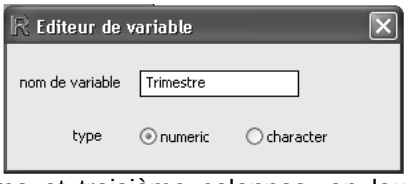
3 Travailler avec R Commander : saisie de données

3.1.1 Définition d'un nouveau jeu de données. Saisie des données

Les données utilisées par R Commander se trouvent enregistrées dans des **jeux de données** (dataframes).

Utilisez le menu Données - Nouveau jeu de données pour créer un nouveau jeu, nommé Manage.

La fenêtre de l'éditeur de données s'affiche. Cliquez sur la tête de la première colonne. Attribuez le nom Trimestre à cette colonne, et le type "numeric".



Procédez de même pour les deuxième et troisième colonnes, en leur attribuant les noms : H_Incident et D_Incident.

Remarquez que certains caractères, tels que le tiret, l'espace, les signes d'opérations (+, /, %) ne sont pas autorisés dans les noms des jeux de données et les noms de variables.

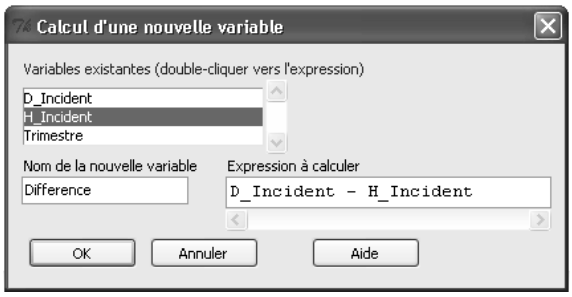
Saisissez dans cette feuille les observations des variables Trimestre, H_Incident et D_Incident ci-dessous.

Trimestre	H_Incident	D_Incident
1	11	8
2	11	13
3	14	12
4	21	17
5	12	14
6	10	9
7	15	10
8	15	12
9	17	13
10	9	10
11	12	8
12	12	13
13	15	12
14	12	9
15	11	9

Refermez ensuite la fenêtre de saisie. On peut contrôler la saisie à l'aide du bouton "Visualiser".

3.1.2 Variables calculées : Calcul du protocole dérivé des différences individuelles

On veut calculer la différence D_Incident - H_Incident, pour chacun des 15 trimestres.
Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Calculer une nouvelle variable...
Complétez la fenêtre de dialogue en indiquant le nom de la nouvelle variable (Difference) et la formule de calcul : D_Incident - H_Incident :



Cliquez ensuite sur le bouton OK.
Vous pouvez ensuite visualiser le jeu de données dans son état actuel.

Remarque : Les valeurs de la variable Difference ont été générées grâce à un calcul. Mais, cette formule de calcul n'est pas mémorisée comme propriété de la colonne considérée et il n'y a aucune remise à jour de cette colonne si les colonnes D_Incident ou H_Incident sont modifiées.

3.1.3 Enregistrement du jeu de données dans un fichier

On utilise le menu Données - Jeu de données actif - Sauver le jeu de données actif... pour enregistrer le jeu de données dans un fichier sur disque. Un tel fichier a par défaut l'extension .RData et R Commander nous propose le nom Manage.RData (ou Manage.RDA sur les versions plus anciennes).

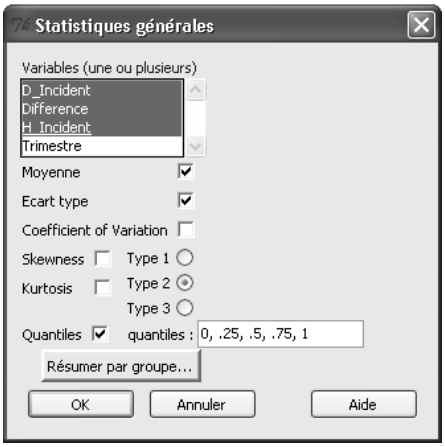
N.B. Pour les versions plus anciennes de R et R Commander, c'est l'extension .RDA ou .rda qui est utilisée. En conséquence, dans les fenêtre de dialogue d'ouverture de fichiers, il faut parfois activer l'option "tous les fichiers" pour que les fichiers d'extension .RData s'affichent.
On notera également que le format du fichier ne dépend en principe pas du système d'exploitation de la machine : un fichier de données enregistré avec R sous Windows est en principe lisible par R sous Mac OSX. Cependant, avec les versions de R un peu anciennes, se posent des problèmes d'encodage des caractères (notamment au niveau des noms de variables). Par exemple, le fichier Manage.RData n'est pas lisible sous R 2.15 pour Mac OSX si la variable calculée précédente a été appelée Différence (avec un "é"), mais devrait l'être si la variable a été appelée Difference (sans accent).

3.2 Calcul des paramètres de statistiques descriptives

On veut calculer les paramètres descriptifs relatifs aux variables H_Incident, D_Incident et Différence :

Utilisez le menu Statistiques - Résumés - Statistiques descriptives...

Sélectionnez les variables choisies (Majuscule+Clic pour une sélection continue, Ctrl+Clic pour une sélection discontinue) et cliquez sur OK.



Le résultat s'affiche (en bleu) dans la fenêtre de sortie :

```
                 mean       sd 0%  25% 50% 75% 100%  n
D_Incident 11.266667 2.548576  8  9.0  12  13   17 15
Difference -1.866667 2.325838 -5 -3.5  -3   0    2 15
H_Incident 13.133333 3.090693  9 11.0  12  15   21 15
```

4 Comment sauvegarder son travail

Dans le paragraphe précédent, nous avons vu que R sauvegardait les jeux de données dans des fichiers d'extension .RData. Lors d'une séance de travail ultérieure, le jeu de données peut être rechargé en mémoire à l'aide du menu Données - Charger un jeu de données...

R et R Commander permettent d'autres sauvegardes :

- Le contenu de la fenêtre de script de R Commander peut être sauvegardé (menu Fichier - Sauver le script sous...). On a ainsi une trace (pas très explicite) de l'ensemble des commandes qui ont été effectuées. Le fichier de script a généralement une extension .R. C'est en fait un fichier "texte seul" qui peut être ouvert aussi bien par R que par un éditeur ou un traitement de textes.

Lors d'une séance de travail ultérieure, on peut utiliser le menu Fichier - Ouvrir un script... On retrouve alors dans la fenêtre de scripts les commandes qui avaient été enregistrées et on peut ré-exécuter certaines d'entre elles en les sélectionnant et en cliquant sur le bouton "Soumettre".

- Le contenu de la fenêtre de sortie peut également être sauvegardé, dans un fichier "texte seul", avec l'extension .txt. Ces fichiers peuvent ensuite être ouverts dans un éditeur ou un traitement de textes.

- Enfin R propose également de sauvegarder l'ensemble de l'environnement de travail (fichier .RData). Lorsqu'on recharge un environnement de travail préalablement sauvegardé, on restaure l'ensemble des données, variables, scripts etc qui y ont été définis.

5 Importer un jeu de données

R Commander permet de définir un nouveau jeu de données, et l'éditeur permet alors de saisir les valeurs voulues. Cependant, les fonctionnalités de cet éditeur restent limitées et, si le volume de données est plus important, il est préférable de saisir les données dans un autre logiciel et de les importer ensuite sous forme de jeu de données dans R Commander.

Exemple :

Les parfums agréables aident-ils un étudiant à apprendre plus efficacement ? Hirsch et Johnston pensent que la présence d'un parfum floral peut améliorer la capacité d'apprentissage d'un sujet dans certaines situations. Dans leur expérience, 21 sujets devaient effectuer à 6 reprises un parcours dans un labyrinthe tracé sur une feuille de papier. Pour chacune des 6 épreuves, les sujets portaient un masque. Pour 3 épreuves, le masque

était parfumé. Pour les 3 autres épreuves, le masque ne l'était pas. Pour une moitié des sujets, le masque parfumé correspondait aux 3 premières épreuves, pour l'autre moitié, aux trois dernières. A chaque essai, les expérimentateurs notaient le temps nécessaire au parcours du labyrinthe par le sujet. On demandait également à chaque sujet s'il trouvait le parfum agréable ou non.

Les variables rassemblées dans le fichier de données sont les suivantes :

ID : identifiant du sujet
Genre : sexe du sujet
Fumeur : le sujet est-il fumeur (Y = oui, N = non)
Opinion : opinion du sujet sur le parfum (positive, négative, indifférente)
Age : âge du sujet en années
Ordre : 1 = le masque parfumé est porté aux épreuves 4 à 6 ; 2 = le masque parfumé est porté aux épreuves 1 à 3.
Sans parfum : moyenne des temps mis par le sujet aux trois épreuves avec masque sans parfum
Avec parfum : moyenne des temps mis par le sujet aux trois épreuves avec masque parfumé.

Les données observées sont les suivantes :

	ID	Genre	Fumeur	Opinion	Age	Ordre	Sans parfum	Avec parfum
1	1	M	N	pos	23	1	30,6	38,0
2	2	F	Y	neg	43	2	48,4	51,6
3	3	M	N	pos	43	1	60,8	56,7
4	4	M	N	neg	32	2	36,1	40,5
5	5	M	N	neg	15	1	68,5	49,0
6	6	F	Y	pos	37	2	32,4	43,2
7	7	F	N	pos	26	1	43,7	44,6
8	8	F	N	pos	35	2	37,1	28,4
9	9	M	N	pos	26	1	31,2	28,2
10	10	F	N	indiff	31	2	51,2	68,5
11	11	F	Y	pos	35	1	65,4	51,1
12	12	F	Y	indiff	55	2	58,9	83,5
13	13	F	Y	pos	25	1	54,5	38,3
14	14	M	Y	indiff	39	2	43,5	51,4
15	15	M	N	indiff	25	1	37,9	29,3
16	16	M	N	pos	26	2	43,5	54,3
17	17	M	Y	neg	33	1	87,7	62,7
18	18	M	N	neg	62	2	53,5	58,0
19	19	F	Y	pos	54	1	64,3	52,4
20	20	F	N	neg	38	2	47,4	53,6
21	21	M	N	neg	65	1	53,7	47,0

Ouvrez le classeur Excel Parfums1.xls. Les données y ont été saisies dans la feuille "Données". La première ligne indique les noms des variables.

- Utilisez le menu Données - Importer des données - depuis Excel, Access ou dBase...
- Indiquez Parfums comme nom pour le jeu de données.
- Sélectionnez le fichier Parfums1.xls.

Utilisez ensuite le bouton "Visualiser". On constate que l'importation s'est déroulée correctement. Les variables "Sans parfum" et "Avec parfum", dont le nom n'est pas accepté par R (car il comporte un espace) ont été renommées Sans.parfum et Avec.parfum.

Remarque 1. Il peut paraître étonnant que R Commander accepte de lire le fichier Parfums1.xls alors que ce fichier est déjà ouvert sous Excel. Mais R Commander utilise une instruction qui fait une requête de base de données sur le fichier. Ces instructions sont justement conçues pour permettre des accès concurrents à des fichiers de données. Notez le message d'avertissement dans la fenêtre de R : "Le chargement a nécessité le package : RODBC". ODBC, Open Data Base Connectivity est une interface standardisée de connexion à une base de données.

Remarque 2. L'importation de fichiers Excel fonctionne, même si Microsoft Office n'est pas installé sur la machine. En revanche, cet item de menu ne semble pas présent sur les versions Mac OS de R / Rcmdr. Dans ce cas, on travaillera à partir d'un fichier de données R (fichier d'extension .RData) ou on convertira le fichier Excel en fichier "texte" ou "csv".

Remarque 3. Il est tout à fait essentiel que le fichier Excel soit convenablement structuré pour que l'importation soit satisfaisante.

Par exemple, ouvrez le fichier Parfums2.xls. Les données y sont enregistrées sous trois formes différentes, dans trois feuilles du classeur :
- la feuille "Données" identique à celle que nous venons d'importer ;
- la feuille "Données mises en forme" dans laquelle un effort de présentation a été fait (nombre de décimales, cadres et bordures, centrage pour les variables catégorielles, etc) ;
- la feuille "Données avec titre" dans laquelle les trois premières lignes forment un titre présentant les données.

Refaites une importation. Indiquez Parfums2 comme nom pour le jeu de données.

Comme le classeur Excel comporte plusieurs feuilles, R Commander nous demande de sélectionner la feuille contenant les données à importer à l'aide d'une fenêtre de dialogue telle que :

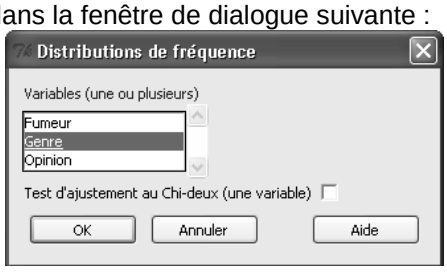
Sélectionnez tout d'abord la feuille "Données mises en forme". Vous devriez obtenir un résultat identique à celui de l'importation précédente.

Refaites alors une troisième importation (Parfums3), en sélectionnant la feuille "Données avec titre".
Visualisez le résultat : vous constatez que le titre a servi de nom pour la première variable, que les autres variables ont été nommées "par défaut" F2, F3, etc., que les 3 premières lignes ont été importées comme les autres, de sorte que toutes les colonnes sont de type "character".

Reprenez Parfums comme jeu de données actif et enregistrez-le sous le nom Parfums.RData.

5.1 Recensement (tri à plat) sur une variable nominale

Pour faire un tri à plat des observations selon les valeurs de la variable "Genre" :
- Utilisez le menu Statistiques - Résumés - Distributions de fréquences...
- Sélectionnez la variable Genre dans la fenêtre de dialogue suivante :



Vous devriez obtenir le résultat suivant :

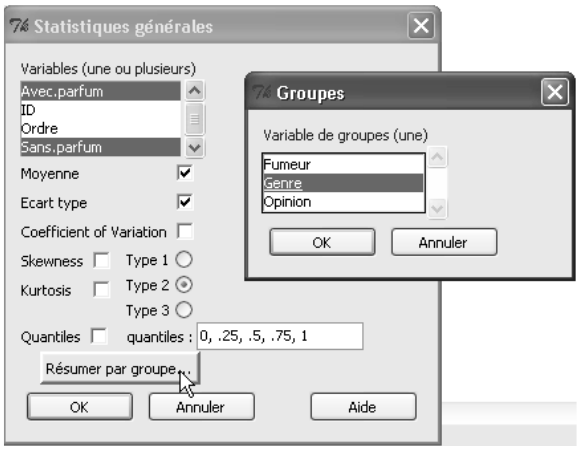
```
> .Table <- table(Parfums$Genre)
> .Table  # counts for Genre
F  M
10 11
> round(100*.Table/sum(.Table), 2)  # percentages for Genre
    F     M
47.62 52.38
```

Autrement dit, l'échantillon étudié comporte 10 femmes et 11 hommes. La proportion de femmes est de 47,62% et celle des hommes de 52,38%.

5.2 Calculer des paramètres descriptifs (moyennes, variances, etc) "par groupe"

La variable "Genre" permet de définir deux échantillons indépendants de sujets. On veut calculer, par exemple, la moyenne, la variance et l'écart type des variables dépendantes "Avec parfum" et "Sans parfum" dans chacun des deux groupes.

- Utilisez le menu Statistiques - Résumés - Statistiques descriptives...
- Sélectionnez les deux variables Avec.parfum et Sans.parfum et cochez Moyenne et Ecart type.
- Cliquez sur le bouton "Résumer par groupe..." et sélectionnez la variable Genre.



- Cliquez ensuite sur OK dans chacune des deux fenêtres.

Vous devriez obtenir :

```
Variable: Avec.parfum
      mean       sd  n
F 51.51667 15.42599 10
M 46.82121 11.51559 11

Variable: Sans.parfum
      mean       sd  n
F 50.33667 10.88255 10
M 49.72121 17.48762 11
```

6 Représentations graphiques d'une variable nominale

6.1 Le cas Fast-food

Vous menez une enquête sur les préférences de consommation des jeunes adultes. Plus particulièrement, vous vous intéressez :

1) à leurs préférences pour différents types de fast-food
2) à leurs préférences pour différents types de voitures
3) à leur comportement (déclaré) par rapport à certains restaurants de fast-food.

De plus, vous notez également le sexe du sujet.

Les variables observées sont décrites ci-dessous :

SEXE (variable nominale simple). Le sexe du sujet interrogé est entré comme variable catégorisée dans le fichier de données (i.e., HOMME, FEMME).

Fast-food favori (variable à réponses multiples). Dans le questionnaire utilisé pour cette étude, on demande aux sujets de sélectionnez leurs trois types de fast-food préférés dans une liste comportant 8 types. Les 8 types de fast-food proposés sont :

(1) Hamburger
(2) Sandwiches
(3) Poulet
(4) Pizza
(5) Fast-food Mexicain
(6) Fast-food Chinois
(7) Fruits de mer
(8) Autres

Les trois choix de chaque sujet sont saisis dans le fichier comme variable à réponse multiple, c'est-à-dire que leur premier choix est saisi dans la variable FASTF_1, leur second choix dans la variable FASTF_2 et leur troisième choix dans la variable FASTF_3.

Voiture favorite (variable à réponses multiples). Dans ce cas, nous avons demandé à chaque individu de renseigner les voitures (marque et modèle) qu'ils aimeraient posséder (l'argent n'étant pas un problème). Ces réponses (particulièrement les marques et les modèles) ont été codées en quatre catégories :

(1) Voiture de sport, nationale
(2) Voiture de tourisme, nationale
(3) Voiture de sport, étrangère
(4) Voiture de tourisme, étrangère

Comme pour la variable fast food favori (voir ci-dessus), cette variable a été enregistrée comme une variable à réponses multiples, c'est-à-dire que les préférences des personnes interrogées ont été renseignées dans les variables Voitur_1 à Voitur_3. Remarque : dans ce cas, les individus peuvent répéter la même réponse trois fois (par exemple, ils peuvent citer 3 voitures de sports comme leurs voitures les plus convoitées). Dans le cas des fast foods ci-dessus, les réponses multiples identiques n'étaient pas autorisées (c'est-à-dire, ignorées).

Enfin, nous avons demandé aux personnes interrogées d'indiquer les différents restaurants rapides (locaux) qu'elles avaient eu l'occasion de visiter au cours des deux semaines précédant l'enquête, parmi les quatre existants. Les données ont été enregistrées sous la forme d'une variable pour chaque restaurant particulier. Les quatre variables, Burger_1 à Burger_4, représentant les quatre restaurants locaux :

(1) Burger Meister
(2) Bill's Best Burgers
(3) Hamburger Heaven
(4) Bigger Burger

Si une personne interrogée a déclaré avoir mangé dans au moins un des ces restaurants récemment, un 1 a été entré dans la colonne respective ; sinon, la colonne respective reste vierge. Par conséquent, il s'agit d'une dichotomie multiple, et nous souhaitons tabuler le nombre (ou la proportion) de personnes interrogées ayant déclaré avoir mangé dans chacun des quatre restaurants locaux.

6.2 Examen des variables

Chargez le jeu de données Fastfood.RData et visualisez les données.

Utilisez ensuite le menu Statistiques - Résumés - Jeu de données actif. Vous obtenez en résultat :

```
> summary(Fastfood)
    SEXE         FASTF_1       FASTF_2       FASTF_3       VOITUR_1
 FEMME: 36   PIZZA   :68   PIZZA   :44   HAMBURGR:27   ETR_SPRT:90
 HOMME:164   HAMBURGR:47   HAMBURGR:40   PIZZA   :26   ETR_TRSM:19
             SANDWICH:18   AUTRE   :18   SANDWICH:17   NAT_SPRT:60
             CHINOIS :17   FRUI_MER:18   AUTRE   :15   NAT_TRSM:31
             POULET  :16   POULET  :17   FRUI_MER:15
             FRUI_MER:13   (Other) :45   (Other) :40
             (Other) :21   NA's    :18   NA's    :60
    VOITUR_2      VOITUR_3  BURGER_1   BURGER_2   BURGER_3   BURGER_4
 ETR_SPRT:98   ETR_SPRT:83   OUI : 60   OUI : 68   OUI : 61   OUI : 59
 ETR_TRSM:27   ETR_TRSM:25   NA's:140   NA's:132   NA's:139   NA's:141
 NAT_SPRT:48   NAT_SPRT:63
 NAT_TRSM:27   NAT_TRSM:29
```

Notez que le nombre de modalités représentées dans les réponses du logiciel est limité à 5 ou 6, les autres modalités étant rassemblées dans la catégorie (Other). Pour les colonnes FASTF_2, FAST_3 et BURGER_1 à 4, on voit apparaître une catégorie "NA's". Pour R, NA signifie "valeur manquante". En particulier, pour les 4 dernières variables, seules les réponses "oui" ont été saisies. La cellule est laissée vide dans le cas contraire.

6.3 Tri à plat sur variables nominales

La commande précédente permet d'avoir un recensement de l'échantillon en termes d'effectifs. On peut également obtenir des résultats en termes de fréquences en utilisant le menu Statistiques - Résumés - Distributions de fréquences...

Par exemple, pour la variable FASTF_1 :

```
> .Table  # counts for FASTF_1
   AUTRE  CHINOIS FRUI_MER HAMBURGR MEXICAIN    PIZZA   POULET SANDWICH
       9       17       13       47       12       68       16       18

> round(100*.Table/sum(.Table), 2)  # percentages for FASTF_1
   AUTRE  CHINOIS FRUI_MER HAMBURGR MEXICAIN    PIZZA   POULET SANDWICH
     4.5      8.5      6.5     23.5      6.0     34.0      8.0      9.0
```

Et, pour la variable FASTF_2 :

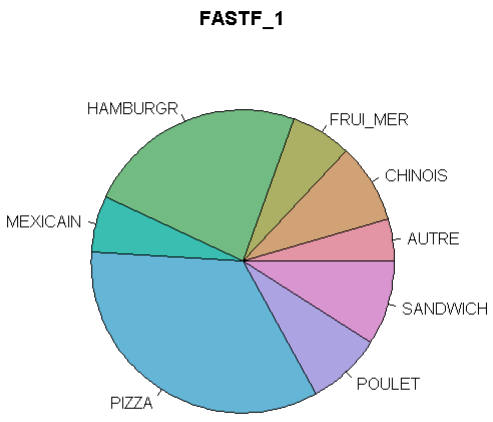
```
> .Table  # counts for FASTF_2
   AUTRE  CHINOIS FRUI_MER HAMBURGR MEXICAIN    PIZZA   POULET SANDWICH
      18       16       18       40       15       44       17       14

> round(100*.Table/sum(.Table), 2)  # percentages for FASTF_2
   AUTRE  CHINOIS FRUI_MER HAMBURGR MEXICAIN    PIZZA   POULET SANDWICH
    9.89     8.79     9.89    21.98     8.24    24.18     9.34     7.69
```

Remarquez que, pour le calcul des fréquences (pourcentages), les valeurs manquantes sont neutralisées.

6.4 Diagrammes circulaires

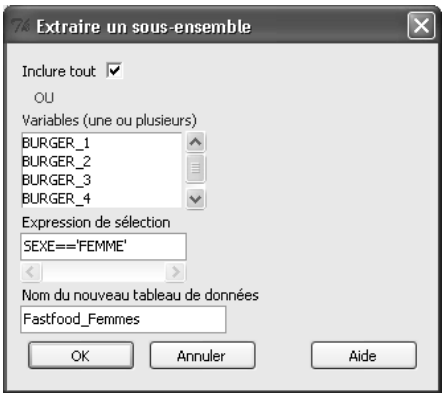
On veut représenter les distributions de la variable FASTF_1 à l'aide d'un diagramme circulaire.
Utilisez le menu Graphes - Graphe en camembert...
et sélectionnez la variable FASTF_1. Vous devriez obtenir le résultat suivant :



6.5 Graphiques sur une partie des données

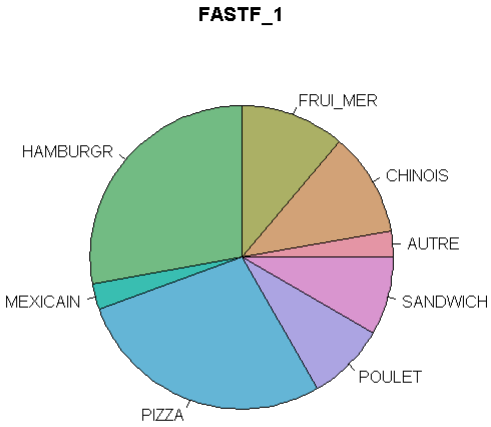
Nous voudrions faire un graphique circulaire du type précédent, en nous limitant aux données concernant les femmes.
La solution la plus simple à mettre en oeuvre consiste à définir un nouveau jeu de données, limité aux seules observations pour lesquelles SEXE est égal à 'FEMME'.

Utilisez le menu Données - Jeu de données actif - Sous-ensemble...
Complétez la fenêtre de dialogue de la façon suivante :



Cliquez sur OK.

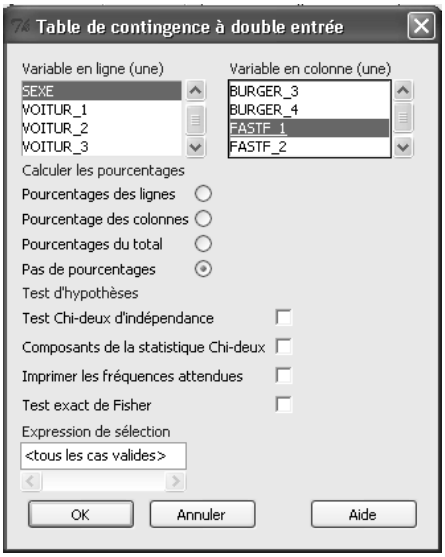
Le jeu de données Fastfood_Femmes devient alors le jeu de données actif. Une commande analogue à celle exécutée précédemment produit le graphique :



6.6 Produire un tableau de contingence

On revient au jeu de données Fastfood et on veut constituer un tableau de contingence à partir des modalités des variables SEXE et FASTF_1.

Utilisez le menu Statistiques - Tables de contingence - Tableau à double entrée.
Indiquez les noms des variables ligne et colonne :

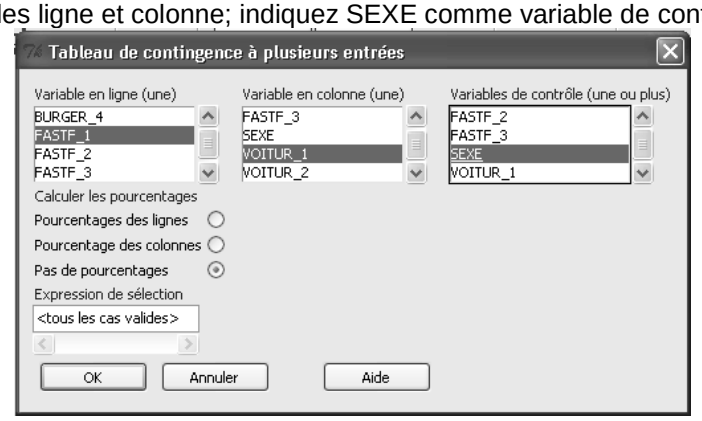


On obtient le résultat suivant :

```
        FASTF_1
SEXE    AUTRE CHINOIS FRUI_MER HAMBURGR MEXICAIN PIZZA POULET SANDWICH
  FEMME     1       4        4        10        1    10      3        3
  HOMME     8      13        9        37       11    58     13       15
```

On peut également introduire une troisième variable et produire autant de tableaux de contingence que de modalités de cette variable. Par exemple, on souhaite croiser le premier choix de fastfood et le premier choix de type de voiture, en séparant le tableau relatif aux hommes et celui relatif aux femmes.

Utilisez le menu Statistiques - Tables de contingence - Tableau à plusieurs entrées. Indiquez FASTF_1 et VOITUR_1 comme variables ligne et colonne; indiquez SEXE comme variable de contrôle.



Vous devriez obtenir :

```
, , SEXE = FEMME

          VOITUR_1
FASTF_1    ETR_SPRT ETR_TRSM NAT_SPRT NAT_TRSM
  AUTRE           0        0        1        0
  CHINOIS         2        0        1        1
  FRUI_MER        1        1        1        1
  HAMBURGR        4        1        4        1
  MEXICAIN        1        0        0        0
  PIZZA           4        0        4        2
  POULET          3        0        0        0
  SANDWICH        1        0        1        1

, , SEXE = HOMME

          VOITUR_1
FASTF_1    ETR_SPRT ETR_TRSM NAT_SPRT NAT_TRSM
```

AUTRE	1	1	4	2
CHINOIS	7	2	2	2
FRUI_MER	7	0	2	0
HAMBURGR	14	2	15	6
MEXICAIN	4	2	3	2
PIZZA	23	7	18	10
POULET	9	0	3	1
SANDWICH	9	3	1	2

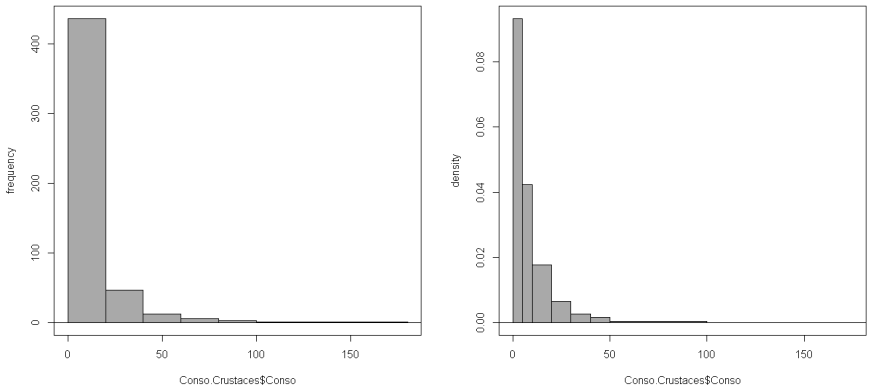
7 Graphiques relatifs à des variables numériques

On reprend les données "Conso-Crustacés" utilisées dans le TD N° 2. On peut, par exemple, importer le contenu du fichier Excel Conso-crustaces.xls.

7.1 Histogrammes

On veut réaliser un histogramme représentant la distribution de la variable Conso.

Utilisez le menu Graphes - Histogramme...
Indiquez Conso comme variable à représenter et gardez les valeurs par défaut pour les autres paramètres.
Vous devriez obtenir le graphique figurant à gauche ci-dessous. Le découpage en classes pour ce graphique n'est guère satisfaisant. On peut soit augmenter le nombre de classes, soit remplacer dans la ligne de commande *scale="frequency", breaks="Sturges*" par : *scale="density", breaks=c(0,5,10,20,30,40,50,100,175)*

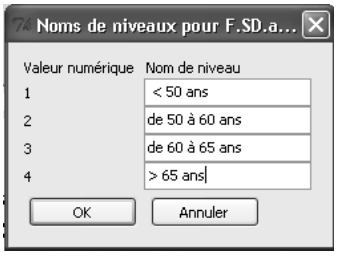


7.2 Graphiques à barres pour une variable ordinale.

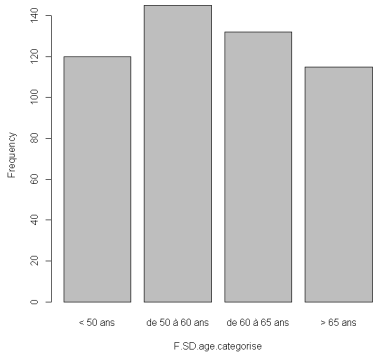
La variable SD.age.categorise peut être considérée comme une variable ordinale. Pour représenter la distribution de cette variable dans la population étudiée, on souhaite utiliser un graphique à barres. Cependant, R Commander ne permet pas de réaliser un tel graphique sur une variable numérique, et nous devons au préalable transformer cette variable en facteur.

Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Convertir des variables numériques en facteurs... Sélectionnez SD.age.categorise et indiquez F.SD.age.categorise comme nom pour la nouvelle variable qui contiendra le facteur associé à SD.age.categorise .

Attribuez ensuite des noms aux niveaux :



Utilisez ensuite le menu Graphes - Graphique en barres... et sélectionnez F.SD.age.categorise comme facteur à représenter. Vous obtenez :



7.3 Graphiques de type "boîte à moustache"

La variable Conso n'a pas une distribution suffisamment régulière pour fournir un bon exemple de telles représentations. Nous allons donc utiliser un autre jeu de données : hdv2003.Rdata.

Ce jeu de données est fourni avec le package rgrs. Sa description est la suivante :

hdv2003 est un extrait comportant 2000 individus et 20 variables provenant de l'enquête Histoire de Vie réalisée par l'INSEE en 2003. L'extrait est tiré du fichier détail mis à disposition librement (ainsi que de nombreux autres) par l'INSEE à l'adresse suivante :
http://www.insee.fr/fr/themes/detail.asp?ref_id=fd-HDV03
Les variables retenues ont été parfois partiellement recodées. La liste des variables est la suivante :

Variable	Description
id	Identifiant (numéro de ligne)
poids	Variable de pondération 3
age	Âge
sexe	Sexe
nivetud	Niveau d'études atteint
occup	Occupation actuelle
qualif	Qualification de l'emploi actuel
freres.soeurs	Nombre total de frères, soeurs, demi-frères et demi-soeurs
clso	Sentiment d'appartenance à une classe sociale
relig	Pratique et croyance religieuse
trav.imp	Importance accordée au travail
trav.satisf	Satisfaction ou insatisfaction au travail
hard.rock	Ecoute du Hard rock ou assimilés
lecture.bd	Lecture de bandes dessinées
peche.chasse	Pêche ou chasse pour le plaisir au cours des 12 derniers mois
cuisine	Cuisine pour le plaisir au cours des 12 derniers mois
bricol	Bricolage ou mécanique pour le plaisir au cours des 12 derniers mois
cinema	Cinéma au cours des 12 derniers mois
sport	Sport ou activité physique pour le plaisir au cours des 12 derniers mois
heures.tv	Nombre moyen d'heures passées à regarder la télévision par jour

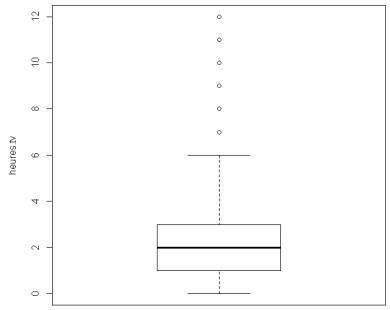
Pour des variables dont la distribution n'est pas régulière, il peut être intéressant de les représenter à l'aide d'une boîte à moustaches (box and whiskers).

Considérons, par exemple, la variable heures.tv.

- Calculez d'abord ses valeurs extrêmes, sa médiane et ses quartiles.

```
     mean       sd 0% 25% 50% 75% 100%    n NA
 2.246566 1.775853  0   1   2   3   12 1995  5
```

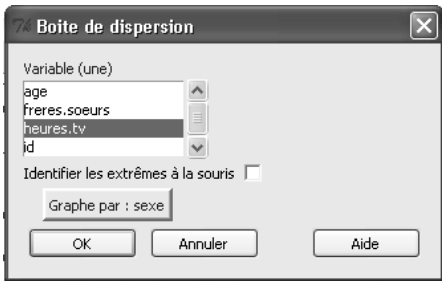
- Construisez ensuite un graphique de type "boîte à moustaches", à l'aide du menu Graphes - Boîte de dispersion... Vous devriez obtenir un résultat du type suivant :



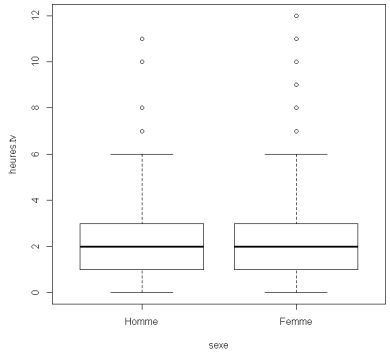
Règles de construction du graphique : les bases de la boîte représentent les premier et troisième quartiles. Ainsi, la boîte représente les 50% d'observations centrales. On calcule ensuite l'écart inter-quartile (3-1=2) et on le multiplie par 1,5 : 2 * 1,5 = 3. Pour les valeurs extérieures à l'intervalle délimité par les quartiles, les valeurs qui s'écartent de plus de 3 du quartile correspondant sont considérées comme atypiques.
Sur notre exemple, le minimum est de 0. La distance au premier quartile ne dépasse pas 1-3, c'est-à-dire -1. Toutes les valeurs inférieures au premier quartile sont donc typiques et la moustache correspondante va du premier quartile au minimum. En revanche, le maximum est de 12, et la distance de ce maximum au troisième quartile est de 12-3, c'est-à-dire 9. La moustache correspondante s'arrête donc sur la valeur 6, et la valeur 12 est représentée comme valeur atypique.

7.4 Boîtes à moustaches comparant deux groupes indépendants

Ces représentations en boîtes à moustaches sont notamment intéressantes pour comparer les résultats observés sur deux groupes. On veut, par exemple, réaliser, par exemple, des boîtes à moustaches concernant la variable heures.tv pour les deux groupes définis par la variable sexe.



Vous devriez obtenir le résultat suivant :



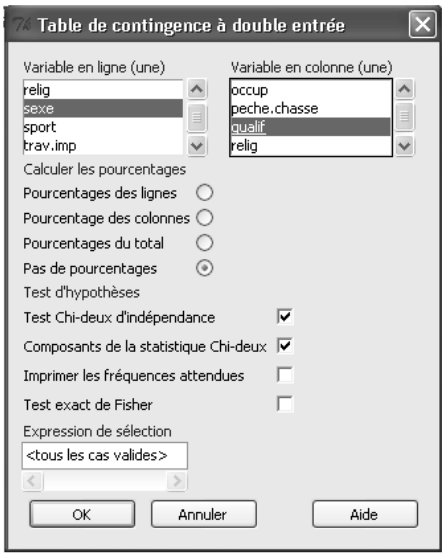
8 Travail sur un tableau de contingence - Test du khi-2

Nous pouvons être amenés à réaliser un test du khi-2 sur des données structurées de différentes façons : tableau de contingence (c'est généralement le cas lorsque les données sont issues d'un exercice de TD), ou tableau du protocole (par exemple, vous avez saisi les réponses que vous avez recueillies à un questionnaire). Nous allons donc étudier comment réaliser un test du khi-2 dans chacun de ces deux cas.

8.1 Le test du khi-2 à partir d'un tableau protocole

Reprenons le jeu de données hdv2003. On veut étudier s'il existe un lien entre le sexe et l'emploi actuel (variable qualif).

Utilisez le menu Statistiques - Tables de Contingence - Tableau à double entrée. S'affiche alors la fenêtre de dialogue suivante :



L'application de la méthode produit plusieurs résultats.

- R construit d'abord un tableau de contingence à partir des données fournies :

```
> .Table <- xtabs(~sexe+qualif, data=hdv2003)

> .Table
       qualif
sexe    Ouvrier specialise Ouvrier qualifie Technicien Profession intermediaire
  Homme                 96              229         66                        88
  Femme                107               63         20                        72
       qualif
sexe    Cadre Employe Autre
```

```
  Homme   145      96    21
  Femme   115     498    37
```

- R calcule ensuite la statistique du khi-2 sur ce tableau de contingence et nous renvoie la valeur du khi-2 (χ^2 = 5,5631) ainsi que son niveau de significativité (p=0,1349):

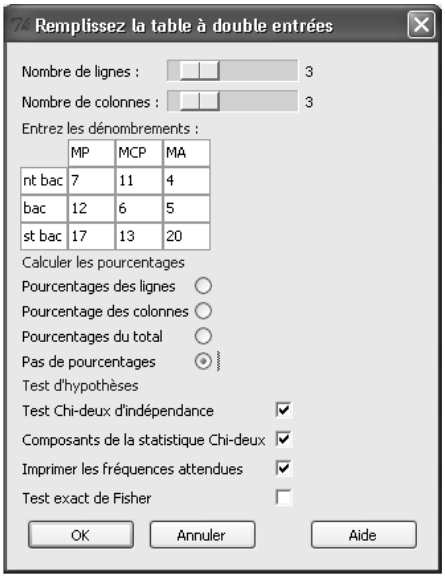
```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
  Pearson's Chi-squared test
data:  .Table
X-squared = 387.5644, df = 6, p-value < 2.2e-16
```

Lecture du résultat. Le niveau de significativité ($p=2.2\ 10^{-16}$) indique qu'il existe un lien très significatif entre le sexe et l'emploi occupé..

8.2 Le test du khi-2 à partir d'un tableau de contingence

On peut aussi fournir à R Commander un tableau de contingence. Pour cela, on utilise le menu Statistiques - Tables de Contingence - Remplir et analyser un tableau à double entrée...

On spécifie tout d'abord les dimensions du tableau, puis on indique les effectifs correspondant aux combinaisons de modalités. L'ordre dans lequel sont prises les différentes modalités est sans importance. Nous allons ici reprendre l'exemple vu en cours (type de professionnalisation et niveau d'études pour un échantillon de musiciens) :



On obtient les résultats suivants :

```
> .Table <- matrix(c(7,11,4,12,6,5,17,13,20), 3, 3, byrow=TRUE)
> rownames(.Table) <- c('avant bac', 'bac', 'post bac')
> colnames(.Table) <- c('MP', 'MCP', 'MA')
```

R écrit le tableau de contingence :

```
> .Table  # Counts
          MP MCP MA
avant bac  7  11  4
bac       12   6  5
post bac  17  13 20
```

On obtient ensuite le résultat du test du khi-2 :

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
  Pearson's Chi-squared test
data:  .Table
X-squared = 7.8786, df = 4, p-value = 0.09613
```

Enfin, comme nous avions cochez la boîte "Composants de la statistique du Chi-deux" et "Imprimer les fréquences attendues", R nous indique les effectifs théoriques et les contributions des différentes cases du tableau au khi-deux :

```
> .Test$expected # Expected Counts
                  MP       MCP        MA
avant bac  8.336842  6.947368  6.715789
bac        8.715789  7.263158  7.021053
post bac  18.947368 15.789474 15.263158

> round(.Test$residuals^2, 2) # Chi-square Components
            MP  MCP   MA
avant bac 0.21 2.36 1.10
bac       1.24 0.22 0.58
post bac  0.20 0.49 1.47
```

On peut remarquer que les données saisies sous forme de tableau de contingence ne sont pas mémorisées par R, mais seulement utilisées pour le calcul immédiat du khi-deux.

9 Comparer deux proportions sur deux (ou plusieurs) groupes indépendants

On a observé une variable dichotomique (Oui/Non, Succès/Echec, etc) sur deux (ou plusieurs) échantillons de sujets format des groupes indépendants. On souhaite étudier si la proportion de "1" (ou de "Oui", de "Succès") est significativement différente dans les populations parentes des groupes.

Cette situation se ramène en fait à un tableau de contingence du type suivant :

	Groupe 1	Groupe 2
1 (ou Oui, Succès)		
0 (ou Non, Echec)		

Par exemple, dans le jeu de données hdv2003, nous souhaitons étudier si, pour la question relative au bricolage, le pourcentage de "Oui" est significativement différent selon le sexe.

Au niveau descriptif, les pourcentages observés sur les échantillons de chaque sexe sont les suivants :

```
bricol
sexe     Non  Oui Total Count
  Homme 42.7 57.3   100   899
  Femme 69.3 30.7   100  1101
```

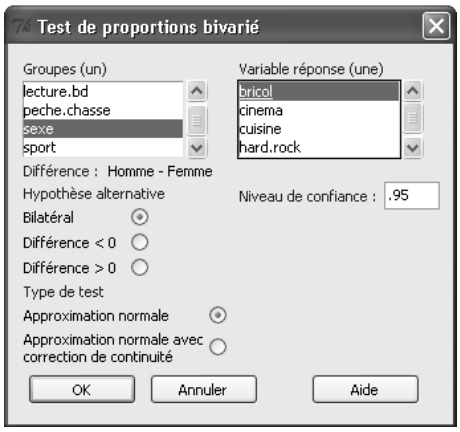
Autrement dit, 57,3% des hommes ont déclaré s'être adonnés au bricolage, alors que ce n'était le cas que pour 30,7% des femmes.

A l'aide du menu Statistiques - Tables de Contingence - Tableau à double entrée, on peut réaliser un test du khi-2 qui fournit les résultats suivants :

```
> .Test
   Pearson's Chi-squared test
data:  .Table
X-squared = 143.0168, df = 1, p-value < 2.2e-16
```

Autrement dit, les réponses sont très significativement différentes selon le sexe.

On retrouve ce résultat en utilisant le menu Statistiques > Proportions > Test de proportions bivarié :



R Commander fournit le résultat suivant :

```
> prop.test(.Table, alternative='two.sided', conf.level=.95,
correct=FALSE)

  2-sample test for equality of proportions without continuity
  correction
data:  .Table
X-squared = 143.0168, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.3081483 -0.2235819
sample estimates:
   prop 1    prop 2
0.4271413 0.6930064
```

Le calcul du khi-2 conduit au même résultat que précédemment, mais nous aurions pu ici faire une correction de continuité (correction de Yates). Nous aurions alors obtenu X-squared = 141.932. Remarquez que les deux proportions indiquées (0,427 et 0,693) sont en fait les proportions de "Non" dans les deux groupes. En effet, ces proportions sont calculées sur la modalité classée première par ordre alphabétique.

Remarquez également que les deux menus utilisés exigent que la variable "Groupe", aussi bien que la variable dichotomique, soient déclarées comme variables de type "facteur".

Exercice. Comparer de même chez les hommes et chez les femmes, les proportions de réponses "Oui" sur la question relative au cinéma. Ces proportions sont-elles significativement différentes ?