

Licence de Psychologie - Semestre N° 5 - TD n° 1

Deuxième partie : Introduction à R et à R Commander

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

2.1 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 2.15.0 for Windows](#)

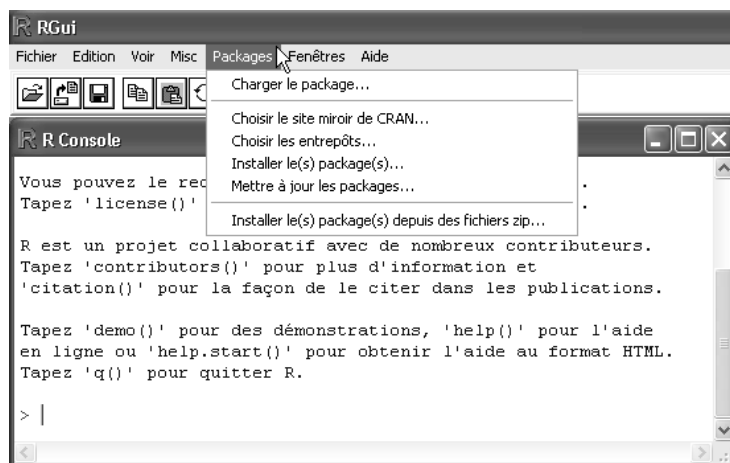
Une fois le téléchargement effectué, allez dans le répertoire où le fichier a été enregistré et exécutez (double-clic) le programme d'installation : **R-2.15.0-win.exe**

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.

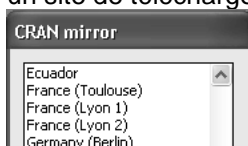
Une fois l'installation terminée, exécutez le logiciel en cliquant sur l'icône **R-i386-2.15.0**

La console de R (fenêtre **Rgui**) est alors chargée.

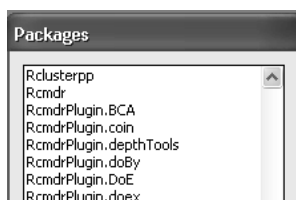
Utilisez le menu **Packages - Installer les packages**.



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

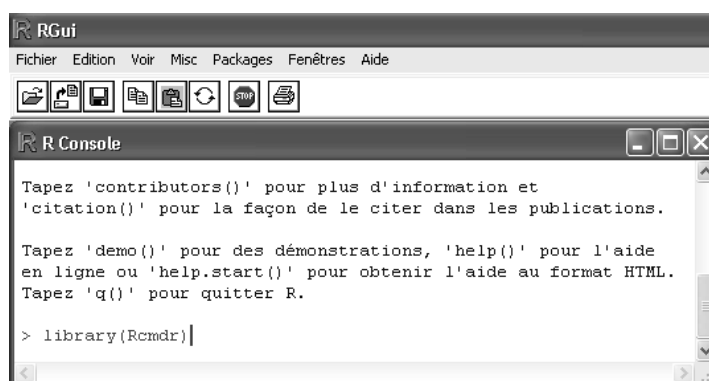


La (longue) liste des packages s'affiche alors. Sélectionnez le package **Rcmdr**.



Une fois le package installé, essayez de le lancer en utilisant le menu **Packages > Charger le package...** de la console de R. On peut aussi saisir dans la console de R la ligne de commande suivante (en respectant majuscules et minuscules) :

```
library(Rcmdr)
```



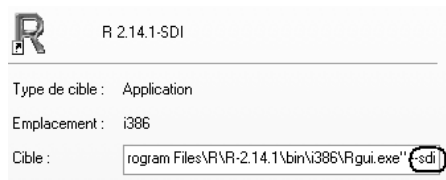
Le logiciel affiche alors une fenêtre d'avertissement indiquant que le fonctionnement correct de Rcmdr exige l'installation de 14 autres packages (car, sem, rgl, relimp, multcomp, lmtest, leaps, Hmisc, e1071, effects, colorspace, aplpack, abind, RODBC) et propose de les installer.

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" :



2.2 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.

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

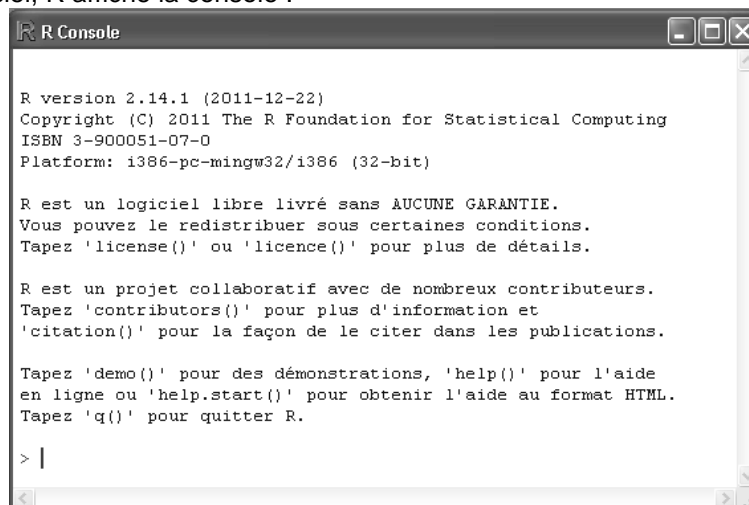
Actuellement (juillet 2013), la version de R installée dans nos salles est la version 2.15. On pourra commencer par copier sur le bureau le raccourci vers le programme disponible sur le volume tdlettres, dans le dossier PSY3. Si vous choisissez de charger R à partir des items du menu Programmes, chargez de préférence la version 32 bits. En effet, certaines fonctionnalités (la lecture de fichiers Excel notamment) ne sont pas actives sur la version 64 bits.

3 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

3.1 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, deux méthodes sont possible :

- Avec le clavier, saisissez dans la console la commande :

```
library(Rcmdr)
```

- Ou, avec la souris, utilisez le menu Packages > Charger le package... de la console et sélectionnez l'item Rcmdr.

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.



3.2 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.

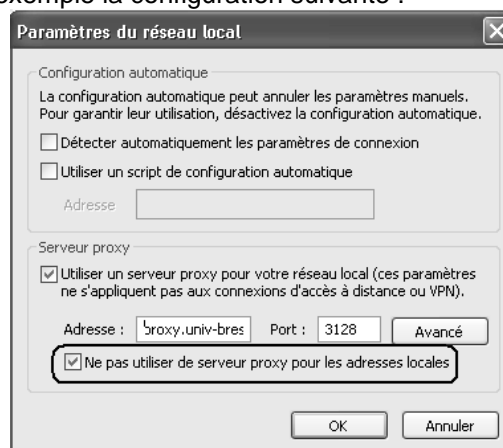
Essayez par exemple de saisir dans la console :

```
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 :



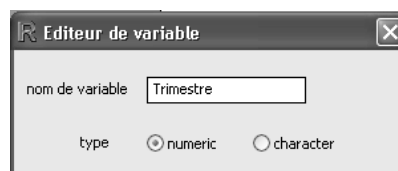
4 Travailler avec R Commander : saisie de données

4.1.1 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

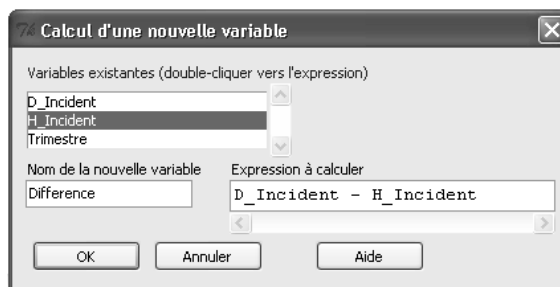
Refermez ensuite la fenêtre de saisie. On peut contrôler la saisie à l'aide du bouton "Visualiser".

4.1.2 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.

4.1.3 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.

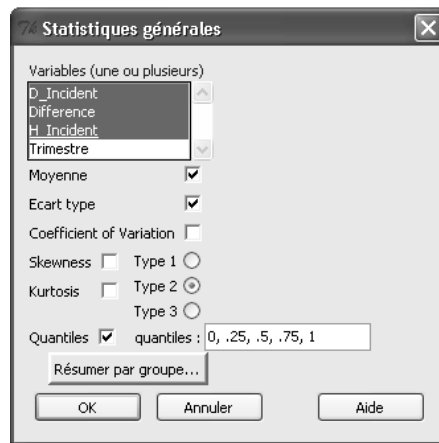
On notera également que le format du fichier ne dépend en principe pas du système d'exploitation de la machine : un fichier de données enregistré avec R sous Windows est en principe lisible par R sous Mac OSX. Cependant, avec les versions de R un peu anciennes, se posent des problèmes d'encodage des 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).

4.2 Calcul des paramètres de statistiques descriptives

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

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

5 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.

6 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 ni avec la version "64 bits" de R. 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".

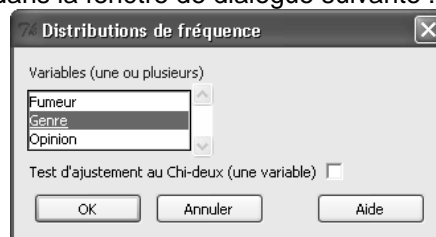
Visualisez le résultat : vous constatez que le titre a servi de nom pour la première variable, que les autres variables ont été nommées "par défaut" F2, F3, etc., que les 3 premières lignes ont été importées comme les autres, de sorte que toutes les colonnes sont de type "character".

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

6.1 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 :

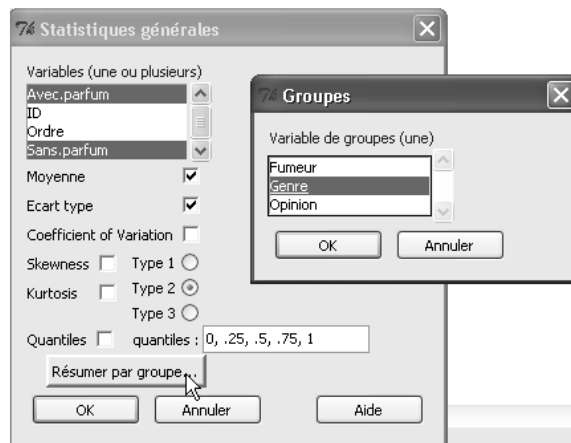
```
> .Table <- table(Parfums$Genre)
> .Table # counts for Genre
 F  M
10 11
> round(100*.Table/sum(.Table), 2) # percentages for Genre
      F      M
47.62 52.38
```

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%.

6.2 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
```

7 Représentations graphiques d'une variable nominale

7.1 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 sélectionner 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.

7.2 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, FASTF_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.

7.3 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
```

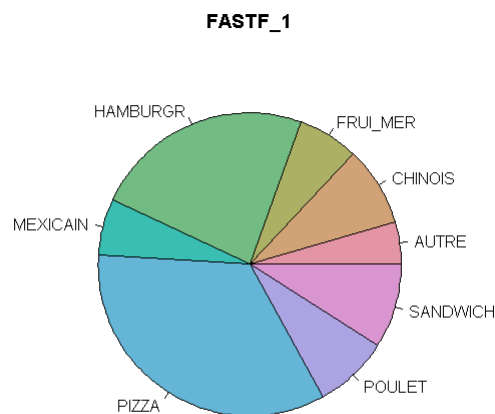
Remarquez que, pour le calcul des fréquences (pourcentages), les valeurs manquantes sont neutralisées.

7.4 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 :



7.5 Graphiques sur une partie des données

Nous voudrions faire un graphique circulaire du type précédent, en nous limitant aux données concernant les femmes.

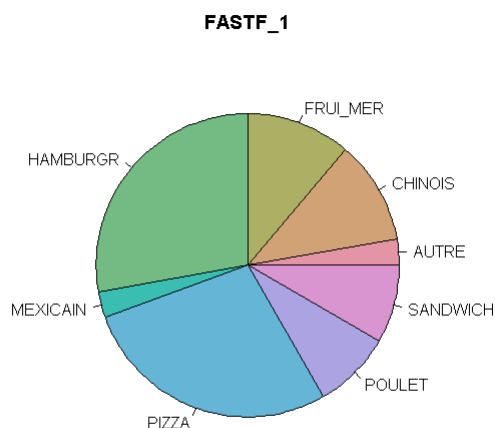
La solution la plus simple à mettre en oeuvre consiste à définir un nouveau jeu de données, limité aux seules observations pour lesquelles SEXE est égal à 'FEMME'.

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



Cliquez sur OK.

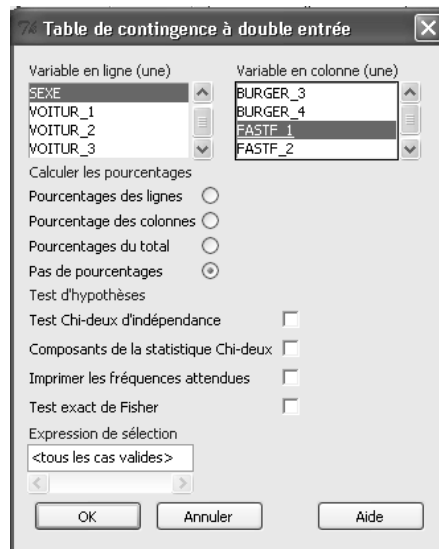
Le jeu de données Fastfood_Femmes devient alors le jeu de données actif. Une commande analogue à celle exécutée précédemment produit le graphique :



7.6 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 :

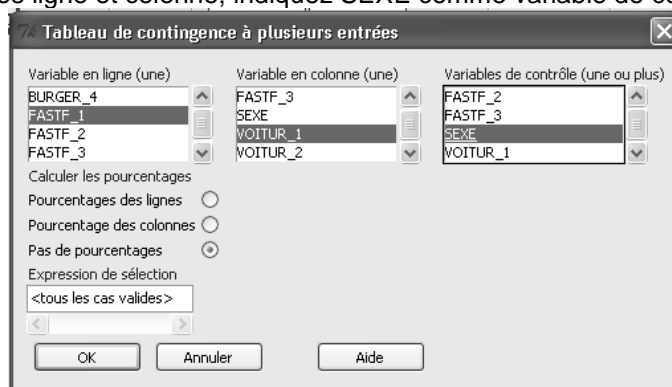


On obtient le résultat suivant :

	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

8 Graphiques relatifs à des variables numériques

On considère le cas suivant :

Données de Mireault (1990) à propos de l'impact de la mort d'un parent.

Mireault a recueilli des données auprès de 381 étudiants universitaires dont certains avaient perdu un de leurs parents pendant leur enfance, suite à un décès. Les étudiants se répartissaient en trois groupes. Le groupe 1 se composait des sujets dont un des parents étaient décédé. Le groupe 2 comprenait les sujets dont les parents vivaient toujours et étaient restés mariés. Le groupe 3 regroupait les étudiants dont les parents avaient divorcé. Mireault voulait observer les effets de la perte d'un parent sur le niveau actuel de symptomatologie mesuré chez la personne par l'inventaire condensé des symptômes et sur la perception par l'individu de sa propre vulnérabilité vis-à-vis de toute perte future.

Pour toutes les mesures, un score élevé indique que le concept mesuré est davantage présent.

Col 1	ID	= Numéro d'identification du sujet
Col 2	Group	= 1 = décès; 2 = marié; 3 = divorcé
Col 3	Gender	= Sexe. 1 = masculin; 2 = féminin
Col 4	YearColl	= Année d'étude. 1 = première année; 2 = deuxième année; etc.
Col 5	College	= Type d'études. 1 = lettres et sciences; 2 = santé; 3 = ingénieur; 4 = sciences économiques et/ou commerciales; 5 = agriculture
Col 6	GPA	= Moyenne des points obtenus. 4 = très bon; 3 = bon; 2 = satisfaisant; 1 = passable; 0 = insuffisant
Col 7	LostPGen	= Sexe du parent décédé
Col 8	AgeAtLos	= Age de l'enfant lors du décès du parent
Col 9	SomT	= Score T de somatisation
Col 10	ObsessT	= Score T de comportements obsessionnels et compulsifs
Col 11	SensitT	= Score T de sensibilité interpersonnelle
Col 12	DepressT	= Score T de dépression
Col 13	AnxT	= Score T d'anxiété
Col 14	HostT	= Score T d'hostilité
Col 15	PhobT	= Score T d'anxiété phobique
Col 16	ParT	= Score T d'idées paranoïaques
Col 17	PsyT	= Score T de comportements psychotiques
Col 18	GSIT	= Score T sur l'index global des symptômes
Col 19	PVTotal	= Score total de vulnérabilité perçue
Col 20	PVLoss	= Score de vulnérabilité perçue aux pertes ultérieures
Col 21	SuppTotl	= Score T de soutien social

Les données correspondantes se trouvent dans le fichier [Mireault.RData](#).

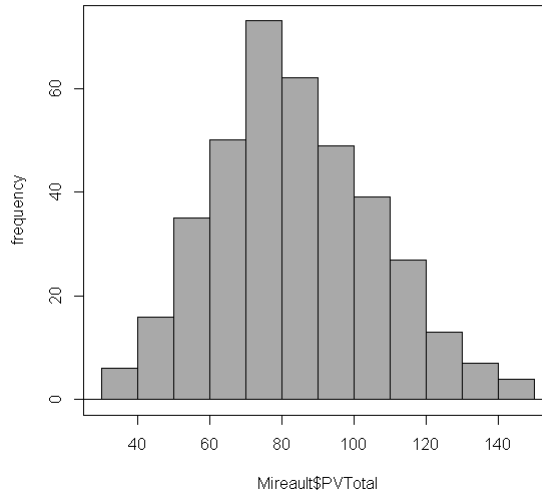
8.1 Histogrammes

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

Utilisez le menu Graphes - Histogramme...

Indiquez PVTotal comme variable à représenter et gardez les valeurs par défaut pour les autres paramètres.

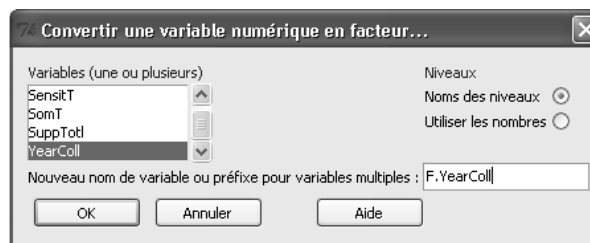
Vous devriez obtenir un graphique du type suivant :



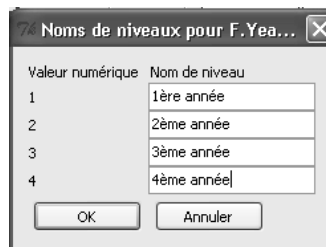
8.2 Graphiques à barres pour une variable ordinale.

La variable YearColl 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.

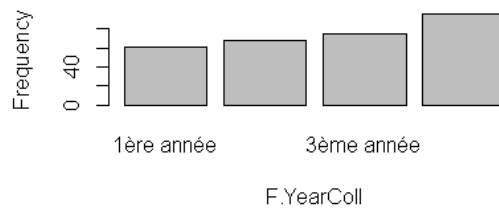
Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Convertir des variables numériques en facteurs... Sélectionnez YearColl et indiquez F.YearColl comme nom pour la nouvelle variable qui contiendra le facteur associé à YearColl.



Attribuez ensuite des noms aux niveaux :



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



8.3 Graphiques de type "boîte à moustache"

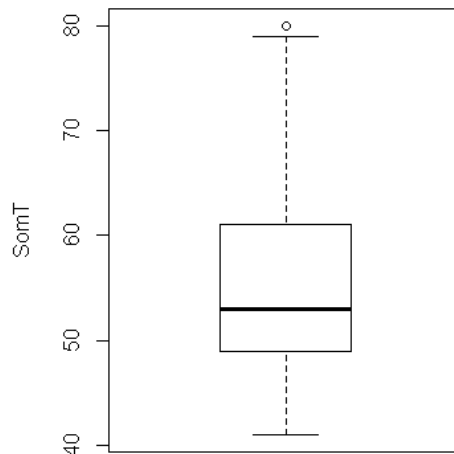
Pour des variables dont la distribution n'est pas aussi régulière que celle de PVTtotal, il peut être intéressant de les représenter à l'aide d'une boîte à moustaches (box and whiskers).

Considérons, par exemple, la variable SomT.

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

mean	sd	0%	25%	50%	75%	100%	n	NA
54.62133	10.07143	41	49	53	61	80	375	6

- 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 :



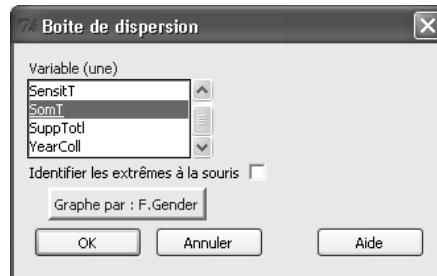
Règles de construction du graphique : les bases de la boîte représentent les premier et troisième quartiles. Ainsi, la boîte représente les 50% d'observations centrales. On calcule ensuite l'écart inter-quartile ($61-49=12$) et on le multiplie par 1,5 : $12 * 1,5 = 18$. Pour les valeurs extérieures à l'intervalle délimité par les quartiles, les valeurs qui s'écartent de plus de 18 du quartile correspondant sont considérées comme atypiques.

Sur notre exemple, le minimum est de 41. La distance au premier quartile ne dépasse pas $49-41$, c'est-à-dire 8. 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 80, et la distance de ce maximum au troisième quartile est de $80-61$, c'est-à-dire 19. La moustache correspondante s'arrête donc sur la valeur 79, et la valeur 80 est représentée comme valeur atypique.

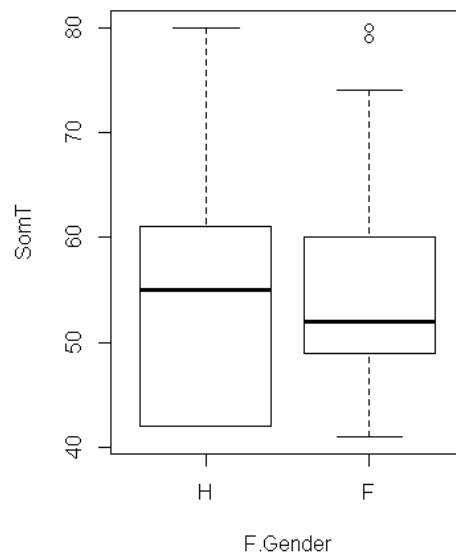
8.4 Boîtes à moustaches comparant deux groupes indépendants

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 SomT pour les deux groupes définis par la variable SEXE.

- Convertissez la variable Gender en facteur, en plaçant le résultat dans une nouvelle variable nommée F.Gender.
- Utilisez le menu Graphes - Boîte de dispersion...
- Indiquez SomT comme variable à représenter et F.Gender comme variable de classement.

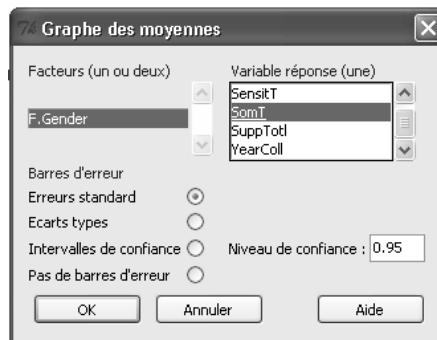


Vous devriez obtenir le résultat suivant :

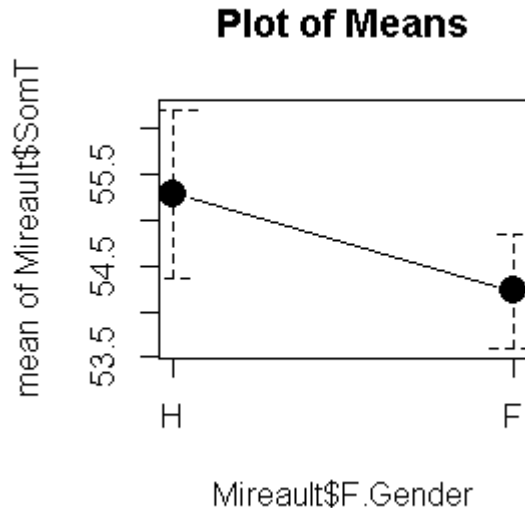


8.5 Graphe comparant deux moyennes

On obtient également une représentation graphique intéressante à l'aide du menu Graphes > Graphe des moyennes.



En indiquant comme précédemment F.Gender comme facteur et SomT comme variable numérique, on produit le graphe suivant :



Les tracés en pointillés sont appelés "barres d'erreur". Par défaut, ils représentent l'erreur type dans chacun des deux groupes (écart type divisé par la racine carrée de la taille de l'échantillon). Mais d'autres paramétrages sont possibles : écarts types, intervalles de confiance pour le degré de confiance indiqué, etc.

9 Nuages de points

On mène une enquête sur les variations circadiennes de la charge mentale induite par une tâche simple et répétitive. On considère un échantillon homogène de sujets et on relève, à différents moments de la journée :

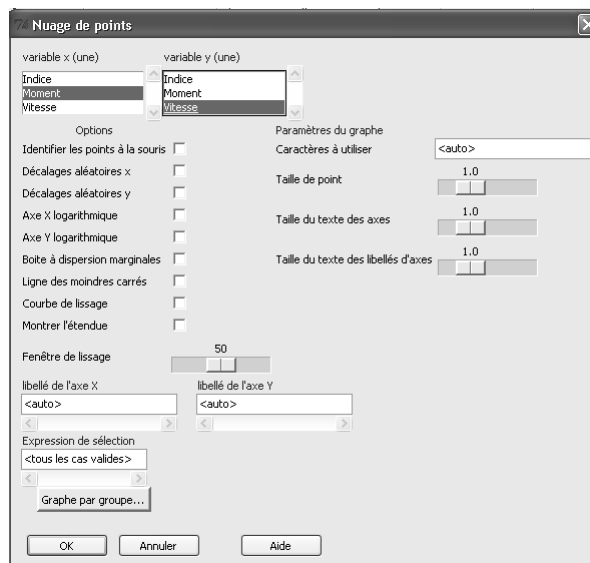
- la vitesse d'exécution d'une tâche répétitive simple (nombre d'appuis sur un bouton par minute)
- l'indice de charge mentale induite (mesuré à partir du temps de réaction à un stimulus auditif simple).

Les résultats (moyennes obtenues sur l'ensemble des sujets observés) sont donnés dans la feuille de données.

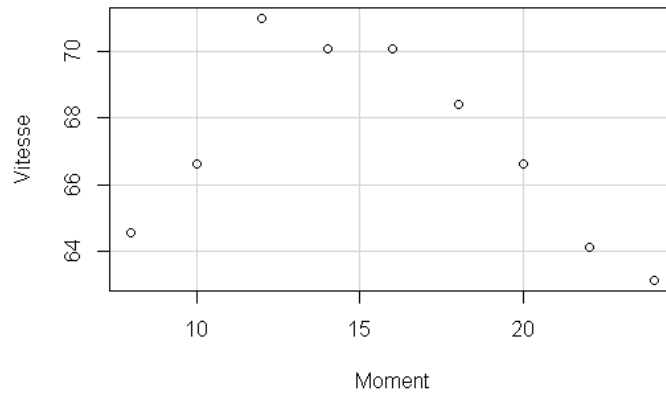
Importez comme nouveau jeu de données, le contenu de la feuille "Données" du classeur [Variations-Circadiennes.xls](#).

On veut réaliser une première représentation graphique, de type nuage de points, avec en abscisse le moment de la journée et en ordonnée, la vitesse.

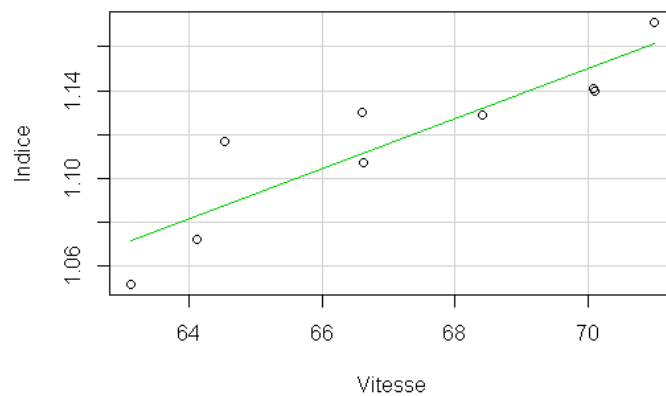
- Utilisez le menu Graphes - Nuages de points...
- Spécifiez Moment comme variable X et vitesse comme variable Y.
- Désactivez les différentes boîtes à cocher.
- Cliquez sur le bouton OK.



On obtient le graphique suivant :



Réalisez un graphique analogue avec les variables Moment et Indice. Enfin, réalisez un graphique avec les variables Vitesse et Indice, en activant la boîte à cocher "Ligne des moindres carrés". On obtient alors :



Le coefficient de corrélation entre les trois variables Moment, Vitesse et Indice peut être obtenu à l'aide du menu Statistiques - Résumés - Matrice de corrélations.

On obtient ici :

	Indice	Moment	Vitesse
Indice	1.0000000	-0.7158028	0.8990227
Moment	-0.7158028	1.0000000	-0.3732135
Vitesse	0.8990227	-0.3732135	1.0000000

On constate que le coefficient de corrélation entre Vitesse et Indice est $r=0,90$, c'est-à-dire une valeur élevée.