Licence de Psychologie - Semestre N° 5 - TD n° 2
Intervalles de confiance, lois de distribution classiques et tests paramétriques avec Statistica

10 Travail sur des données recensées : statistiques descriptives sur un tableau d'effectifs

10.1 Manipuler des données à partir de la console

On reprend les données de l'exercice 13 de la feuille de TD de statistiques : dans une grande entreprise (américaine), les salaires (en milliers de dollars) d'un échantillon aléatoire de 10 femmes possédant entre 3 et 5 ans d'expérience sont les suivants :

24 27 31 21 19 26 30 22 15 36.

Pour affecter cette série de données à une variable, on peut, à partir de la console, utiliser la fonction **c()** :

```
salaire <- c(24, 27, 31, 21, 19, 26, 30, 22, 15, 36)
```

Notez l'utilisation de **<-** comme opérateur d'affectation d'une valeur à une variable.
La série est ainsi représentée sous la forme d'un vecteur à 10 composantes, enregistré dans une variable appelée salaire.
La moyenne, l'écart type et la taille de l'échantillon peuvent être obtenus à l'aide des fonctions **mean(), sd(), length()** :

```
mean(salaire)
   [1] 25.1
sd(salaire)
   [1] 6.226288
length(salaire)
   [1] 10
```

On peut aussi utiliser la fonction c() avec des données de type texte, par exemple pour créer des identifiants des sujets de l'échantillon :

```
sujet <- c("s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", "s10")
```

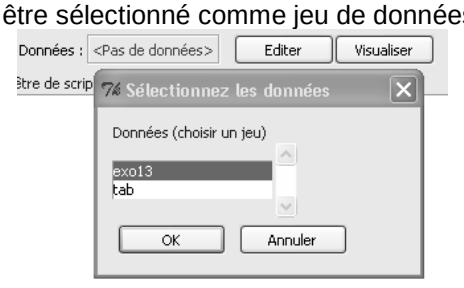
La liste de toutes les variables définies dans l'espace de travail peut être obtenue à l'aide de la fonction **ls()** :

```
ls()
    [1] "salaire"   "sujet"
```

Pour rendre ces données disponibles dans l'interface R Commander, il faut créer un jeu de données contenant ce vecteur, en utilisant la fonction **data.frame()** :

```
exo13 <- data.frame(sujet,salaire, row.names=NULL)
```

Le jeu de données exo13 peut alors être sélectionné comme jeu de données actif dans R Commander :



A partir de la console, on peut désigner les colonnes du jeu de données à l'aide de la syntaxe : exo13$sujet et exo13$salaire. Par exemple :

```
length(exo13$sujet)
   [1] 10
mean(exo13$salaire)
   [1] 25.1
```

On peut également utiliser l'opérateur **with()** qui indique à R que les variables qui suivent sont définies dans le jeu de données spécifié :

```
with(exo13, mean(salaire))
   [1] 25.1
```

N.B. Nos données existent alors en deux exemplaires dans l'espace de travail : les vecteurs sujet et salaire d'une part, les colonnes du jeu de données exo13$sujet et exo13$salaire d'autre part. Ces deux exemplaires sont indépendants, comme on pourra s'en rendre compte en modifiant l'un d'entre eux, par exemple .

10.2 La fonction rep()

La fonction rep() est un outil pratique pour saisir des données répétitives. Sa syntaxe élémentaire est rep(<valeur>, <nombre d'occurences>), mais on peut aussi utiliser la syntaxe rep(<vecteur de valeurs>, <vecteur de nombres d'occurences>). Ainsi rep(0.2, 5) permet de produire un vecteur dans lequel la valeur 0.2 est répétée 5 fois. Lorsque les arguments de la fonction sont eux-mêmes des vecteurs, l'effet est "distribué" de la manière suivante : rep(c("A","B"), c(2,3)) produit en résultat : "A" "A" "B" "B" "B"

Exemple : lors d'un sondage électoral, on interroge au hasard 1000 personnes. 535 personnes déclarent vouloir voter pour le candidat A pendant que 465 déclarent vouloir voter pour un autre candidat. On peut générer un vecteur correspondant au protocole ainsi décrit par :

```
vote <- c(rep("oui",535), rep("non",465))
```

ou

```
vote <- rep(c("oui","non"),c(535,465))
```

Placez ensuite cette variable dans un jeu de données appelé Candidat :

```
Candidat <- data.frame(vote)
```

Ce jeu de données qui sera utilisé dans le paragraphe 11.2. A l'aide de R Commander, enregistrez-le sous le nom Candidat.RData.

10.3 Utiliser des données recensées

On considère l'exemple suivant :
Dans le cadre d'une analyse médicale, deux méthodes de dosage peuvent être utilisées. A partir d'un même prélèvement, on répète 25 fois la méthode A et 30 fois avec la méthode B. Les résultats sont rassemblés dans les tableaux rassemblés dans les feuilles du classeur Excel Dosages.xls.

Ouvrez le classeur Dosages.xls et observez la façon dont les données ont été saisies dans les feuilles Méthode A, Méthode B et Ensemble.

Contrairement aux exemples traités précédemment, les données sont ici présentées sous la forme de tableaux recensés : les observations ont fait l'objet d'un tri à plat. Par exemple, la valeur 42 a été obtenue 7 fois comme résultat de mesure par la méthode A.

Pour pouvoir travailler sur ces données avec R Commander, il faut au préalable les réécrire sous la forme d'un tableau protocole. Pour cela, nous devons utiliser des instructions du langage R qui ne sont pas intégrées au package R Commander. Par exemple, pour les données de la feuille "Ensemble":

- Utilisez le menu Données - Importer des données - depuis Excel, Access ou dBase...
- Indiquez Dosages comme nom pour le jeu de données.
- Sélectionnez le fichier de données Dosages.xls puis la feuille "Ensemble".
- Visualisez les données importées et remarquez que les intitulés de colonnes "Valeur mesurée" et "Nombre de dosages" ont produit les noms de colonnes : Valeur.mesurée et Nombre.de.dosages.

On peut générer les données sous forme de protocole à l'aide des deux commandes suivantes :

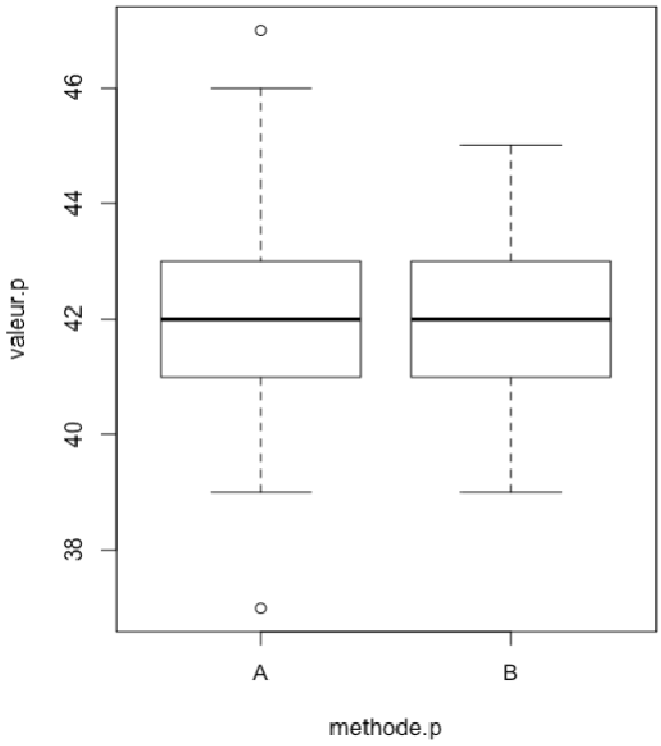
```
methode.p <- with(Dosages, rep(Methode, Nombre.de.dosages))
valeur.p <- with(Dosages, rep(Valeur.mesurée, Nombre.de.dosages))
```

On peut ensuite créer le jeu de données dosages.p rassemblant ces données par :

```
dosages.p <- data.frame(methode.p, valeur.p, row.names=NULL)
```

On peut alors revenir à la fenêtre de R Commander et activer le jeu de données dosages.p.
Sauvegardez ce jeu de données sous le nom dosages.p.RData.

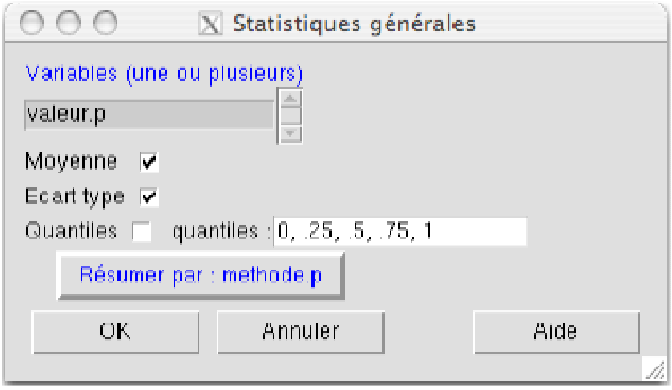
Réalisez, par exemple, le graphique suivant :



10.4 Calculer des paramètres de statistiques descriptives pour des données structurées "par groupe"

Nous souhaitons calculer la moyenne, la variance et l'écart type de la variable "Valeur mesurée" pour chacune des méthodes, et obtenir les résultats dans une même feuille de résultats.

- Utilisez le menu : Statistiques - Résumés - Statistiques descriptives...
- Complétez le dialogue comme suit :



Vous devriez obtenir :

```
     mean       sd  n
A 42.08 2.271563 25
B 42.10 1.398275 30
```

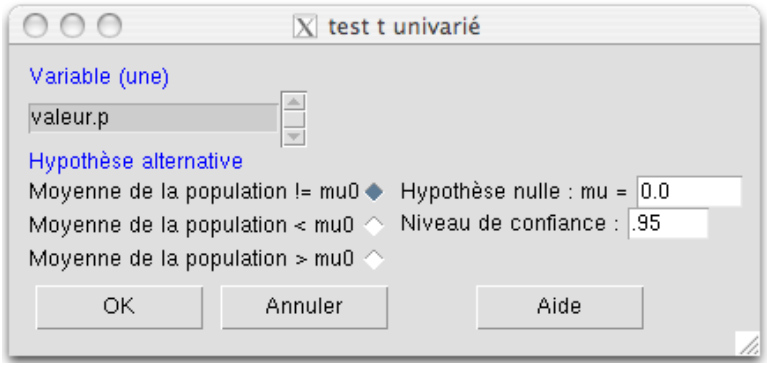
11 Intervalles de confiance

11.1 Intervalle de confiance pour une moyenne

R Commander ne fournit pas de menu explicite pour obtenir un intervalle de confiance. En revanche, des intervalles de confiance sont donnés comme résultats complémentaires de ceux relatifs aux tests de comparaison de moyennes.

Ainsi, à partir de l'échantillon de mesures réalisées par les deux méthodes, quel intervalle estimant la "vraie valeur" de la substance dosée peut-on donner avec un degré de confiance de 95% ?

- Utilisez le menu Statistiques - Moyennes - t-test univarié...
- Complétez la fenêtre de dialogue en indiquant la variable (valeur.p) et le degré de confiance (95%). Remarquez que cette dernière valeur doit être saisie sous la forme 0.95 ou .95 :



Seule la partie du résultat relative à l'intervalle de confiance nous intéresse ici :

```
        One Sample t-test

data:  dosages.p$valeur.p
t = 170.7156, df = 54, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 41.59659 42.58522
sample estimates:
mean of x
 42.09091
```

Autrement dit, on estime, avec un degré de confiance de 95%, que la vraie valeur de la quantité à doser est comprise entre 41,59 et 42,59.

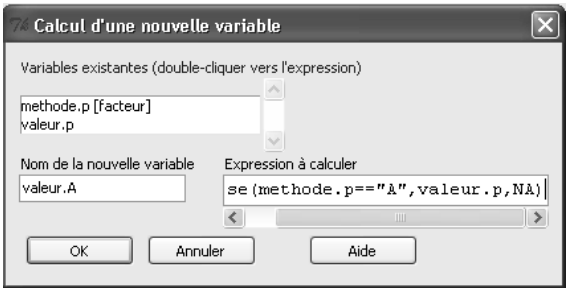
11.1.1 Intervalles de confiance pour les méthodes A et B

Il n'existe pas de menu permettant d'obtenir des intervalles de confiance "par groupe". En revanche, on peut générer des variables supplémentaires valeur.A et valeur.B de façon que valeur.A par exemple :

- soit égale à valeur.p si methode.p est égale à A
- soit NA (valeur manquante) si methode.p est égale à B.

Pour cela :

- Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Calculer une nouvelle variable...
- Complétez la fenêtre de dialogue de la façon suivante :



N.B. Dans le champ "Expression à calculer", saisissez :

```
ifelse(methode.p=="A",valeur.p,NA)
```

On obtient ensuite l'intervalle de confiance comme précédemment :

```
  One Sample t-test

data:  dosages.p$valeur.A
t = 92.6234, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 41.14234 43.01766
sample estimates:
mean of x
    42.08
```

Définissez de même une variable valeur.B et calculez l'intervalle de confiance. Vous devriez obtenir :

```
  One Sample t-test

data:  dosages.p$valeur.B
t = 164.9112, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 41.57788 42.62212
sample estimates:
mean of x
     42.1
```

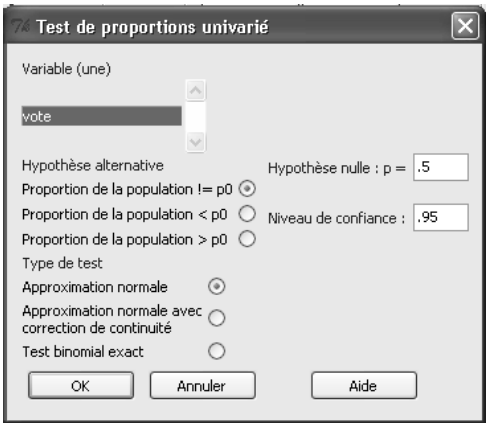
On voit que les intervalles de confiance obtenus pour chacune des méthodes, [41,14; 43,02] et [41,58; 42,62] se recouvrent largement. Un test de comparaison de ces moyennes devrait donc conduire à accepter leur égalité dans les populations parentes.

11.2 Intervalle de confiance pour une proportion

Lors d'un sondage électoral, on interroge au hasard 1000 personnes. 535 personnes déclarent vouloir voter pour le candidat A pendant que 465 déclarent vouloir voter pour un autre candidat. Quel intervalle de confiance, avec un degré de confiance de 95%, peut-on donner concernant le score du candidat A ?

Reprenez le jeu de données Candidat.RData défini au paragraphe 10.2, ou refaites les manipulations décrites dans ce paragraphe.

Utilisez ensuite le menu Statistiques - Proportions - Test de proportions univarié... et complétez la fenêtre de dialogue de la façon suivante :



Deux résultats sont produits :

- D'une part, le recensement de la variable :

```
vote
non oui
465 535
```

- D'autre part, un test donnant notamment un intervalle de confiance pour la modalité citée en premier dans le tableau précédent, c'est-à-dire la modalité "non" dans notre cas :

```
        1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.5
X-squared = 4.9, df = 1, p-value = 0.02686
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4342791 0.4959888
sample estimates:
    p
0.465
```

Au vu de l'intervalle trouvé, il semblerait que l'on puisse affirmer, avec un degré de confiance de 95%, que la modalité non ne peut dépasser 50% et donc que le candidat A sera élu...

Le menu utilisé (Statistiques - Proportions - Test de proportions univarié...) exige que notre colonne de données soit de type "Character" ou "facteur". On constate par exemple que ce menu est inactif si on travaille sur le jeu de données Candidat2 défini par :

```
vote2 <- rep(c(1,0), c(535,465))
Candidat2 <- data.frame(vote2)
```

Dans ce jeu de données, on a choisi de coder "oui" à l'aide du nombre 1 et "non" à l'aide du nombre 0.

Pour obtenir l'intervalle de confiance dans ce cas, deux solutions sont possibles :
- Définir un facteur à partir de la variable numérique vote2, en utilisant le menu - Gérer les variables dans le jeu de données actif - Convertir des variables numériques en facteurs...
- Utiliser le menu s'appliquant aux variables numériques, c'est-à-dire Statistiques - Moyennes - t-test univarié...

Dans ce dernier cas, on obtient comme résultat :

```
        One Sample t-test

data:  Candidat2$vote2
t = 2.2179, df = 999, p-value = 0.02678
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.5040333 0.5659667
sample estimates:
mean of x
    0.535
```

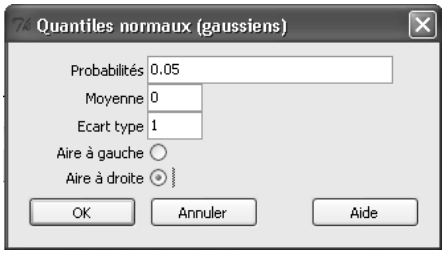
L'intervalle de confiance porte alors sur la moyenne de la série, autrement dit sur la proportion de 1 (c'est à dire de "oui") dans la population. Les deux bornes de l'intervalle sont supérieures à 0.50 et la conclusion est la même que précédemment.

12 Lois statistiques classiques

Le menu Statistiques - Calculateur de Probabilités - Distributions permet d'une part de trouver une valeur critique ou un niveau de significativité pour les lois statistiques continues usuelles, soit de réaliser des représentations graphiques de la densité ou de la fonction de répartition de ces lois.

12.1 La loi normale centrée réduite

Quelle est la valeur de Zcritique pour un test unilatéral à 5% ?
Utilisez le menu Distributions - Distributions continues - Distribution Normale - Quantiles normaux... et complétez le dialogue comme suit :



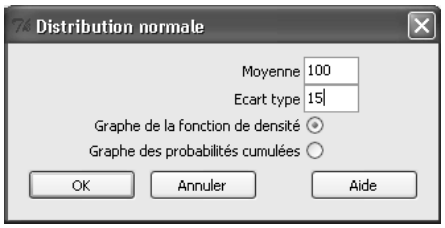
Remarquez que c'est le point qui sert de séparateur décimal et non la virgule, comme dans les logiciels usuels sur un système configuré pour la France.

La réponse est donnée dans la fenêtre de sortie :

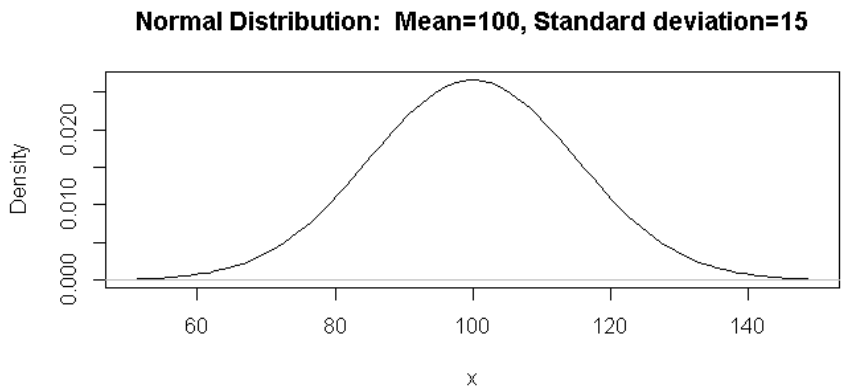
```
> qnorm(c(0.05), mean=0, sd=1, lower.tail=FALSE)
[1] 1.644854
```

12.2 Représenter graphiquement la densité d'une loi normale quelconque

On veut représenter la densité de la loi normale de paramètres m = 100 et s=15.
Utilisez le menu Distributions - Distributions Continues - Distribution Normale - Graphe de la distribution normale
Complétez la fenêtre de dialogue comme suit :



On obtient :

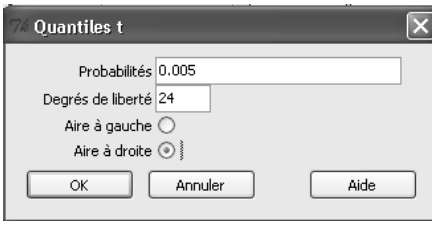


12.3 Loi de Student

Exemple 1.

Calculez la valeur critique de la loi de Student pour un test bilatéral avec ddl=24 et α=0,01.

Utilisez le menu Distributions - Distributions Continues - Distribution t - Quantiles t ... Complétez la fenêtre de dialogue en remarquant que le champ "Probabilités" doit contenir : $\frac{0,01}{2}$, c'est-à-dire 0,005 :



On obtient dans la fenêtre de sortie :

.

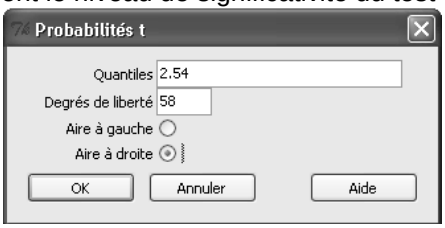
```
> qt(c(0.005), df=24, lower.tail=FALSE)
 [1] 2.79694
```

La valeur critique recherchée est donc 2,7969.

Exemple 2

On a réalisé un test de Student, avec un nombre de degrés de liberté égal à 58. La valeur observée de la statistique est t_{obs}=2,54. Quel est le niveau de significativité du résultat obtenu pour un test unilatéral ? pour un test bilatéral ?

Utilisez le menu Distributions - Distributions Continues - Distribution t - Probabilités t ... En remplissant le dialogue de la façon suivante, on obtient le niveau de significativité du test unilatéral :



La p-value du test bilatéral est le double de la précédente. Pour l'obtenir, on peut procéder de la manière suivante :

- Ajouter les caractères * 2 à la fin de la dernière ligne de la fenêtre de script.
- Sélectionner cette ligne (ou laisser le curseur d'insertion dans cette ligne) et cliquer sur le bouton "Soumettre".

Résultats obtenus :

```
> pt(c(2.54), df=58, lower.tail=FALSE)
[1] 0.006893142

> pt(c(2.54), df=58, lower.tail=FALSE) * 2
[1] 0.01378628
```

Réponses : p=1,38% pour un test bilatéral et p=0,69% pour un test unilatéral.

12.4 Loi du khi-2

Selon les habitudes américaines, cette loi est désignée par Chi-deux.
Déterminez la valeur critique du khi-2 pour un seuil de 5% et 6 ddl.
Vous devriez trouver : Khi-2 = 12,59.

Réalisez un graphique de la densité de la loi du khi-2 à 1 degré de liberté.

13 Tests paramétriques classiques

13.1 Test de comparaison d'une moyenne à une norme

Les données du fichier de données ADD.RData proviennent d'une étude de Howell et Huessy (1985). Ces auteurs ont rendu compte d'une étude portant sur 386 enfants qui avaient ou non manifesté, durant l'enfance, des symptômes liés à des troubles de l'attention. Ces données sont décrites dans l'appendice du livre.

Col 1 : ID = Numéro d'identification du sujet
Col 2 : ADDSC = Moyenne des scores de troubles de l'attention sur 3 ans
Col 3 : GENDER = Sexe. 1 = masculin; 2 = féminin
Col 4 : REPEAT = Nombre d'années scolaires que l'élève a doublées
Col 5 : IQ = QI calculé sur la base d'un test de QI administré au groupe
Col 6 : ENGL = Niveau d'anglais: 1 = niveau préparatoire à l'université; 2 = niveau moyen; 3 = rattrapage
Col 7 : ENGG = Résultats obtenus en anglais: 4 = très bon; 3 = bon; etc.
Col 8 : GPA = Moyenne des points obtenus en neuvième année
Col 9 : SOCPROB = Problèmes sociaux: 0 = non; 1 = oui
Col 10 : DROPOUT = 1 = l'élève a quitté l'école avant son terme; 0 = l'élève a terminé l'école

Chargez R, puis, si besoin, R Commander à l'aide de la commande :

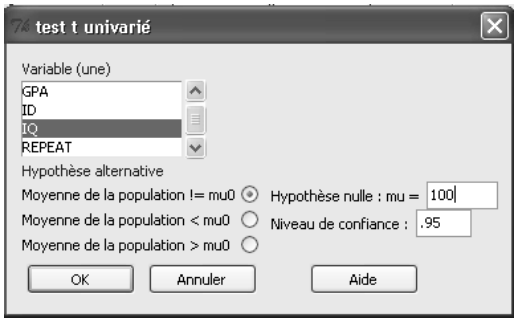
```
library(Rcmdr)
```

Utilisez le menu Données - Charger un jeu de données... et chargez le fichier ADD.RData

Visualisez les données à l'aide du bouton "Visualiser" de la fenêtre de R Commander.

On se pose la question suivante : la population étudiée diffère-t-elle significativement de la population générale (moyenne du QI égale 100) du point de vue du QI ?

- Utilisez le menu Statistiques - Moyennes - t-test univarié. .
- Compléter la fenêtre de dialogue en sélectionnant la variable IQ et en indiquant 100 comme valeur de référence :



L'hypothèse alternative est ici : Moyenne de la population != 100. Les deux caractères "!=" signifient "différent de". Autrement dit, on fait ici un test bilatéral. Vous devriez obtenir le résultat suivant :

```
        One Sample t-test

data:  ADD$IQ
t = 0.1888, df = 87, p-value = 0.8507
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  97.51011 103.01261
sample estimates:
mean of x
 100.2614
```

Lecture du résultat :
La moyenne observée sur l'échantillon de 88 sujets est de 100,26. Statistica compare cette valeur à la valeur de référence (100), à l'aide d'un test de Student, avec 87 degrés de liberté. La statistique de test vaut t=0,1888, ce

qui correspond à un niveau de significativité de 85%. Autrement dit, rien n'indique une différence entre la population étudiée et la population générale du point de vue du QI.

Autre résultat fourni : R nous donne également un intervalle de confiance pour la moyenne de la population parente, avec un degré de confiance de 95% : [97,51 ; 103,01]. On constate que la valeur 100 fait partie de cet intervalle, ce qui est en accord avec le résultat du test.

Remarque : la commande exécutée par R Commander est :

```
t.test(ADD$IQ, alternative='two.sided', mu=100, conf.level=.95)
```

On pourrait obtenir le même résultat en saisissant directement cette commande dans la console de R.

13.2 Test de comparaison de deux moyennes sur des groupes indépendants

13.2.1 Données saisies "par sujet"

Lorsque la saisie a été faite correctement, la feuille de données rassemble sur une même ligne les observations relative à un même individu statistique. Ainsi, la saisie des observations relatives à un plan S<A> comportera au moins 2 colonnes :

- Une colonne "Variable indépendante" ou "Groupe", "Condition expérimentale", avec, comme valeurs nominales, les différents niveaux du facteur A
- Une colonne "Variable dépendante".

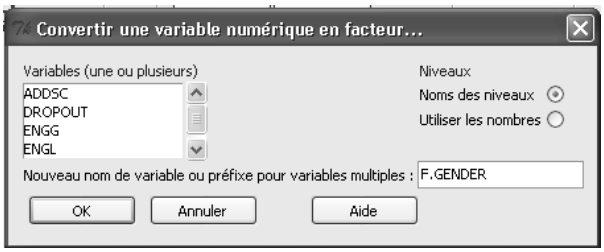
Mais, pour faire un test de comparaison de moyennes, R Commander exige que la variable indépendante soit un facteur et non une simple variable numérique ou texte à deux modalités.

Pour définir un facteur à partir d'une variable de la feuille de données, on peut utiliser le menu : Données - Gérer les variables dans le jeu de données actif - Convertir une variable numérique en facteur.

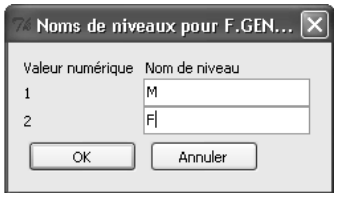
Exemple : On reprend les données du fichier ADD.RData.
On souhaite étudier si le score ADDSC est significativement différent pour les garçons et les filles dans la population étudiée.

On va donc tout d'abord transformer la variable GENDER en facteur. La nouvelle variable s'appellera F.GENDER.

- Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Convertir une variable numérique en facteur et complétez la fenêtre de dialogue comme suit :

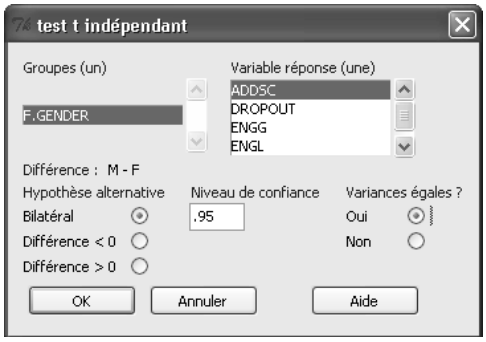


Indiquez ensuite les étiquettes des facteurs ainsi définis :



Le test de comparaison de moyennes peut alors être effectué en utilisant le menu : Statistiques - Moyennes - t-test indépendant.

Commençons par faire un test bilatéral. Complétez la fenêtre de dialogue de la façon suivante :



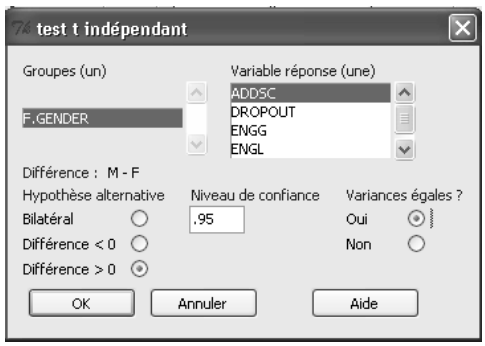
Le résultat fourni est le suivant :

```
        Two Sample t-test

data:  ADDSC by F.GENDER
t = 1.6629, df = 86, p-value = 0.09997
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8801887  9.8862493
sample estimates:
mean in group M mean in group F
       54.29091        49.78788
```

Lecture des résultats :
R fait un test t de Student. La valeur observée de la statistique t est t=1,6629. Le résultat du test, exprimé en termes de niveau de significativité, est p=0,09997, ce qui correspond à un niveau de significativité de presque 10%. Autrement dit, les différences ne sont pas significatives au seuil de 5% bilatéral.

Le résultat du test bilatéral (p-value légèrement inférieure à 10%) nous indique que nous aurions un résultat tout juste significatif en faisant un test unilatéral, avec l'hypothèse alternative : Moyenne pour Gender = M plus grande que Moyenne pour Gender = F.

Pour faire ce deuxième test, utilisez le même menu en complétant la fenêtre de dialogue comme ci-dessous :



On obtient comme résultat :

```
        Two Sample t-test

data:  ADDSC by F.GENDER
t = 1.6629, df = 86, p-value = 0.04999
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0003508505          Inf
sample estimates:
mean in group M mean in group F
       54.29091        49.78788
```

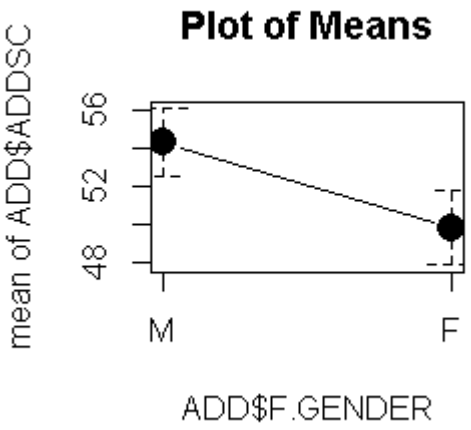
Autrement dit, les résultats sont les mêmes, à l'exception de la p-value, qui vaut maintenant 4,999%.

Les commandes R correspondant à ces deux tests sont :

```
t.test(ADDSC~F.GENDER, alternative='two.sided', conf.level=.95, var.equal=TRUE,
data=ADD)

t.test(ADDSC~F.GENDER, alternative='greater', conf.level=.95, var.equal=TRUE,
data=ADD)
```

Remarque

1. Le menu Graphes - Graphe des moyennes permet d'obtenir un graphe du type suivant, illustrant le résultat trouvé :



Enregistrement des résultats :

- On enregistre le jeu de données ADD mis à jour à l'aide du menu : Données - Jeu de données actif - Sauver le jeu de données actif... Indiquez ADD-traite.RData comme nouveau nom.

- On peut, si on le souhaite, enregistrer le graphique à l'aide du menu Graphes - Sauver le graphe - comme bitmap ou comme PDF/Postscript/EPS. Les formats .png (bitmap) ou .pdf (PDF/Postscript/EPS) sont les plus couramment utilisés.

- On peut enregistrer l'ensemble des commandes qui ont été exécutées à l'aide du menu Fichier - Sauver le script sous... L'extension par défaut pour un tel fichier est : .R.

- On peut enfin enregistrer les résultats des traitements à l'aide du menu Fichier - Sauver les sorties sous... L'enregistrement est fait au format "texte seul". Pour le relire, on peut utiliser le menu Fichier - Ouvrir un script... sous R Commander ou, de préférence, un logiciel de traitement de texte tel que WordPad, Word ou OpenOffice.

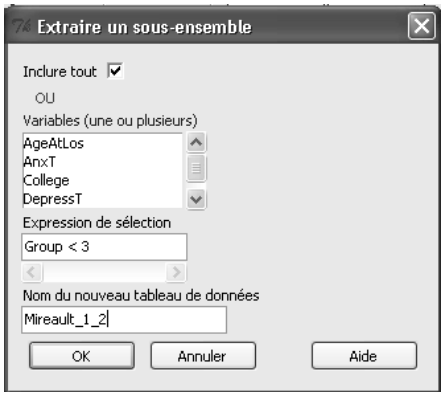
13.2.2 Données saisies "par sujet" : test sur une partie des observations

Lorsque le facteur A comporte plus de deux niveaux, il peut être utile de comparer les observations faites pour deux niveaux particuliers.

Reprenons, par exemple, les données Mireault. La variable Group comporte 3 modalités (codées 1, 2 et 3). Nous souhaitons comparer les scores de la variable PVLoss sur les groupes 1 et 2. Pour cela, nous allons définir un nouveau jeu de données, sous-ensemble du jeu complet.

- Chargez le jeu de données Mireault.RData.
- Utilisez le menu Données - Jeu de données actif - Sous-ensemble...

On inclut toutes les variables. L'expression de sélection est Group < 3 et le nouveau jeu de données ainsi produit est appelé Mireault_1_2.



Ensuite, rendez actif le nouveau jeu de données. Définissez un facteur F.Group à partir de la variable Group.

Remarque utile : le sous-ensemble a été défini à partir d'une expression de sélection portant sur une variable numérique, et le facteur a été défini sur le sous-ensemble, une fois éliminées les observations du groupe 3. Ceci garantit que l'on obtiendra bien un facteur à deux niveaux, nécessaire pour réaliser un test de Student. Si nous définissons le facteur F.Group avant de définir le sous-ensemble, il y a de grandes chances que R Commander "voit" un facteur à 3 niveaux dans le sous-ensemble, bien que le niveau "3" ne concerne aucune observation.

Réalisez ensuite un test de comparaison de moyennes sur PVLoss. Vous devriez obtenir :

```
        Two Sample t-test

data:  PVLoss by F.Group
t = 6.4432, df = 320, p-value = 4.294e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.076335 5.780808
sample estimates:
mean in group 1 mean in group 2
       22.21429        17.78571
```

Remarque : Le test de Fisher sur l'égalité des variances conclut ici sur une différence significative des deux variances. On pourra donc recommencer le test en cochant le bouton radio "Variances égales ? Non". Le résultat est peu différent du précédent :

```
        Welch Two Sample t-test

data:  PVLoss by F.Group
t = 6.3045, df = 270.967, p-value = 1.169e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.045627 5.811515
sample estimates:
mean in group 1 mean in group 2
       22.21429        17.78571.
```

13.2.3 Données saisies par variable

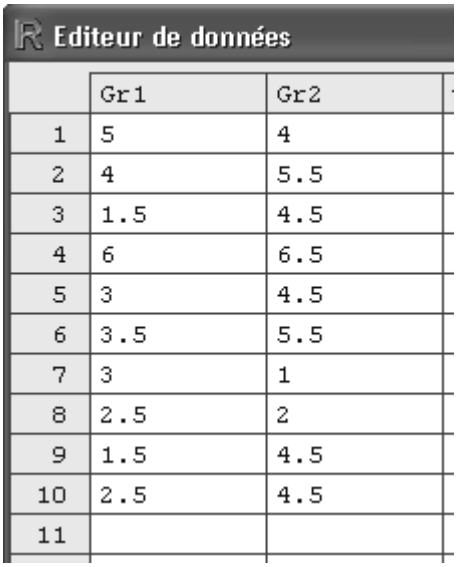
R permet également de réaliser un test de comparaison de moyennes sur deux groupes indépendants lorsque les observations de la VD ont été saisies dans deux colonnes différentes. Cependant, cette possibilité n'a pas été introduite dans R Commander, et nous devons alors saisir la commande en mode texte dans la console de R : RGui.

Exemple : On reprend l'énoncé suivant :
Lors d'une expérience pédagogique, on s'intéresse à l'effet comparé de deux pédagogies des mathématiques chez deux groupes de 10 sujets:

- pédagogie traditionnelle : Gr1
- pédagogie moderne : Gr2.

On note la performance à une épreuve de combinatoire.
Ces données expérimentales permettent-elles d'affirmer que la pédagogie a un effet sur les résultats à l'épreuve de combinatoire?

Chargez R, puis R Commander.
Utilisez le menu Données - Nouveau jeu de données et indiquez comme nom PEDA pour ce jeu de données.
Cliquez sur les têtes des deux premières colonnes, donnez-leur les noms Gr1 et Gr2, ainsi que le type "numeric".
Saisissez les données suivantes :



	Gr1	Gr2
1	5	4
2	4	5.5
3	1.5	4.5
4	6	6.5
5	3	4.5
6	3.5	5.5
7	3	1
8	2.5	2
9	1.5	4.5
10	2.5	4.5
11		

Refermez la fenêtre d'édition des données et enregistrez le jeu de données (menu Données - Jeu de données actif - Sauver le jeu de données actif...

Retournez ensuite à la console de R (fenêtre RGui) et saisissez une commande telle que :

```
t.test(PEDA$Gr1, PEDA$Gr2, var.equal=TRUE)
```

On obtient en résultat :

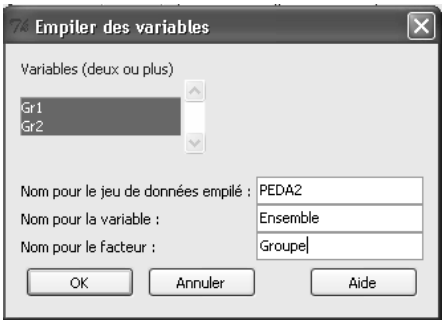
```
        Two Sample t-test

data:  PEDA$Gr1 and PEDA$Gr2
t = -1.451, df = 18, p-value = 0.164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4479606  0.4479606
sample estimates:
mean of x mean of y
     3.25      4.25
```

13.2.4 Passer de données saisies par variable à des données saisies par groupe : empiler les données

Il est assez simple de convertir les données PEDA en un jeu de données PEDA2 dans lequel les données sont saisies par groupe.

Utilisez le menu Données - Jeu de données actif - Empiler les variables dans le jeu de données actif...

- Sélectionnez les deux variables Gr1 et Gr2.
- Indiquez un nom pour le nouveau jeu de données. Par exemple : PEDA2.
- Indiquez un nom pour la nouvelle variable numérique. Par exemple : Ensemble.
- Indiquez un nom pour la variable indépendante (facteur). Par exemple : Groupe.



On dispose alors en mémoire de deux jeux de données, que l'on peut sélectionner en cliquant sur le nom du jeu de données actif, dans la fenêtre de R Commander.

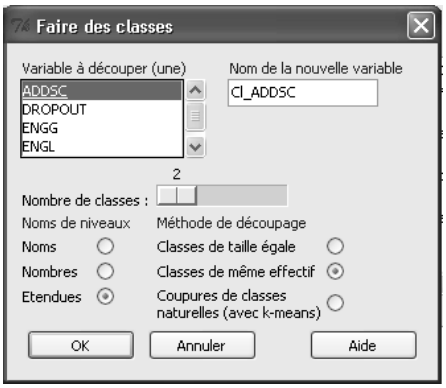
13.2.5 Construire une variable calculée pour définir les deux groupes

On reprend le jeu de données ADD.RData

La médiane de la variable ADDSC est égale à 50. On souhaiterait définir deux groupes en utilisant la position de l'observation par rapport à la médiane, et comparer ces deux groupes du point de vue de la variable GPA.

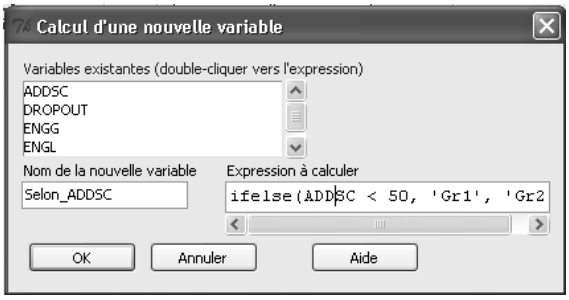
13.2.5.1 Utiliser un découpage en classes

Pour créer une variable définissant les deux groupes selon ce critère, on peut utiliser le menu Données - Gérer les variables dans le jeu de données actif - Découper une variable numérique en classes. On peut compléter la fenêtre de dialogue comme suit, après avoir remarqué qu'un découpage selon la médiane revient à faire deux classes de même effectif :



13.2.5.2 Utiliser une variable calculée

On peut aussi définir une variable calculée (menu Données - Gérer les variables dans le jeu de données actif - Calculer une nouvelle variable ...:) :



La formule de calcul est la suivante :

```
ifelse(ADDSC < 50, 'Gr1', 'Gr2')
```

La variable Selon_ADDSC ainsi créée est de type "character". Elle doit normalement apparaître automatiquement comme facteur possible pour la comparaison de groupes, mais il peut être nécessaire de forcer les choses, par exemple en ouvrant le jeu de données en édition (bouton Editer).

On remarquera que les deux méthodes ne fournissent pas exactement le même découpage en classes puisque les sujets pour lesquels ADDSC=50 sont rangés dans la première classe par le découpage en classes, et dans le groupe 2 par la formule de calcul.

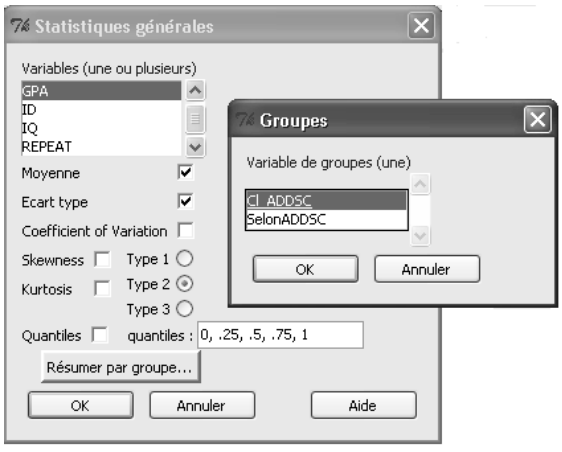
13.2.5.3 Réalisation du test

En utilisant comme facteur l'une ou l'autre des deux variables définies dans les deux paragraphes précédents, on obtient :

```
   Two Sample t-test

data:  GPA by Cl_ADDSC
t = 6.2686, df = 86, p-value = 1.404e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6566315 1.2665133
sample estimates:
mean in group [26,50] mean in group (50,85]
             2.904255              1.942683
```

Le test de Student conduit à une statistique t égale à 6,2686. Le niveau de significativité correspondant est inférieur à 1,4 x 10^{-8}; on conclut donc sur H_1 : il existe une différence très significative entre les scores au GPA dans les deux populations parentes.

- Utilisez le menu Statistiques - Résumés - Statistiques Descriptives.
- Cliquez sur le bouton Résumer par groupe... et sélectionnez la variable qui définit le partage en deux groupes
- Sélectionnez le calcul de la moyenne :



On obtient le résultat suivant, qui donne les valeurs des moyennes et écarts types dans les deux groupes, et montre que les deux groupes ne sont pas parfaitement équilibrés : 47 v/s 41. :

```
>   numSummary(ADD[,"GPA"],   groups=ADD$Cl_ADDSC,   statistics=c("mean",
"sd"), quantiles=c(0,.25,.5,.75,1))
             mean        sd data:n
[26,50] 2.904255 0.6532852     47
(50,85] 1.942683 0.7854999     41
```

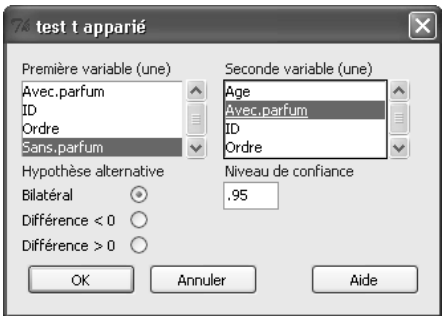
14 Test de comparaison de deux moyennes sur des groupes appareillés

14.1 Comparer les scores observés dans deux conditions

On reprend le jeu de données Parfums1.RData (données présentées dans le polycopié N° 1) et on souhaite étudier si les temps de parcours du labyrinthe sont significativement différents dans la condition "Avec.Parfum" et dans la condition "Sans.Parfum".

Il s'agit là de données appariées, puisqu'il s'agit des mêmes sujets, évalués dans les deux conditions, sur deux épreuves présentant le même niveau de difficulté.

- Utilisez le menu Statistiques - Moyennes - t-test apparié.
- Sélectionnez "Sans.parfum" comme première variable et "Avec.parfum" comme deuxième variable pour constituer la seconde.



Vous devriez obtenir le résultat suivant :

```
> t.test(Parfum1$Sans.parfum, Parfum1$Avec.parfum,
alternative='two.sided',
+   conf.level=.95, paired=TRUE)

  Paired t-test

data:  Parfum1$Sans.parfum and Parfum1$Avec.parfum
t = 0.3496, df = 20, p-value = 0.7303
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.753569  6.667855
sample estimates:
mean of the differences
              0.9571429
```

Lecture du résultat du test. La moyenne de la série des différences individuelles est de 0,957. La statistique de test suit une loi de Student et vaut 0,3496. Le niveau de significativité correspondant est p-value = 0,73 = 73%. On conclut donc sur H0 aux seuils traditionnels : on n'a pas mis en évidence de différence significative entre les scores dans les deux conditions.

14.2 Exercice : comparer deux à deux les scores observés dans plusieurs conditions

Exemple :

Ouvrez le jeu de données Performance-Cognitive.RData.

Dans une étude publiée en 2002, des chercheurs se sont intéressés à l'existence éventuelle d'un affaiblissement de la performance cognitive durant la grossesse. Dans cette étude, deux groupes de femmes ont été observés : 13 d'entre elles étaient enceintes, les 13 autres ne l'étaient pas et constituent le groupe "Contrôle" (variable : Grossesse - modalités : Oui / Non).

La performance cognitive est évaluée à l'aide du score au CFQ (Cognitive Failure Questionnaire). Il s'agit d'un questionnaire à items multiples demandant une auto-évaluation de la fréquence des troubles de la mémoire ou de la concentration, conduisant à une échelle de scores de 0 à 100. Chaque sujet est évalué à 3 reprises correspondant, pour les femmes enceintes au second et au troisième trimestre de la grossesse et au suivi 5 mois après l'accouchement (variables : CFQ.1 à CFQ.3).

On souhaite comparer le groupe des femmes enceintes et le groupe contrôle pour chacune des variables CFQ.1, CFQ.2 et CFQ.3.
On souhaite également comparer deux à deux les scores CFQ.1 à CFQ.3 d'une part chez les femmes enceintes et d'autre part dans le groupe contrôle.

Effectuer les commandes nécessaires. Vous serez conduits à définir deux nouveaux jeux de données : PerfCogOui.RData et PerfCogNon.RData et devriez obtenir les résultats suivants :

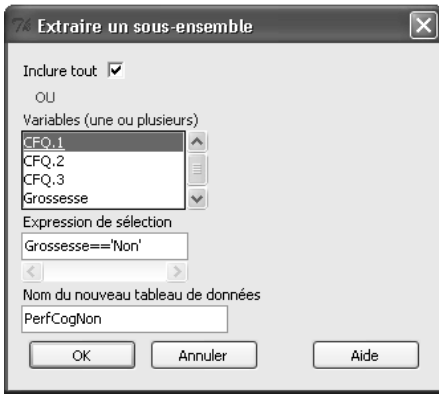
Comparaison entre les deux groupes, pour chacune des 3 variables dépendantes

```
> t.test(CFQ.1~Grossesse, alternative='two.sided', conf.level=.95,
+   var.equal=TRUE, data=Performances_Cognitives)

   Two Sample t-test

data:  CFQ.1 by Grossesse
t = -1.0819, df = 24, p-value = 0.29
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.787466   5.864389
sample estimates:
mean in group Non mean in group Oui
         42.61538          49.07692


> t.test(CFQ.2~Grossesse, alternative='two.sided', conf.level=.95,
+   var.equal=TRUE, data=Performances_Cognitives)

   Two Sample t-test

data:  CFQ.2 by Grossesse
t = -2.0249, df = 24, p-value = 0.05414
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.6099942   0.2253788
sample estimates:
mean in group Non mean in group Oui
         41.53846          53.23077


> t.test(CFQ.3~Grossesse, alternative='two.sided', conf.level=.95,
+   var.equal=TRUE, data=Performances_Cognitives)

   Two Sample t-test

data:  CFQ.3 by Grossesse
t = -3.0404, df = 24, p-value = 0.005637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.440721  -4.866971
sample estimates:
```

```
mean in group Non mean in group Oui
         33.23077          48.38462
```

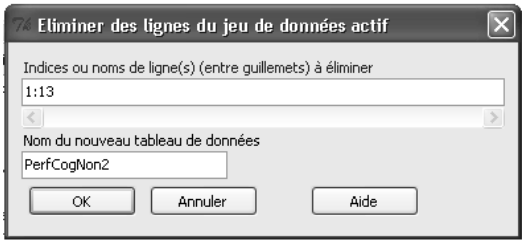
Comparaison entre les 3 conditions pour le groupe contrôle

Le jeu de données PerfCogNon peut être obtenu en utilisant le menu Données - Jeu de données actif - Sous-ensemble et en sélectionnant les lignes pour lesquelles "Grossesse" a la valeur "Non".



Plus simplement, on peut aussi former un nouveau jeu de données en éliminant les lignes 1 à 13 du jeu de données initial :

- Rendez actif le jeu de données Performances_Cognitives
- Utilisez le menu Données - Jeu de données actif - Eliminer une ou des ligne(s) du jeu de données actif...
- Renseignez la fenêtre de dialogue comme suit :



Remarquez l'utilisation de la notation 1:13 pour désigner la liste des nombres entiers de 1 à 13.

La comparaison des trois conditions aboutit au résultat suivant :

```
> t.test(PerfCogNon$CFQ.1, PerfCogNon$CFQ.2, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

  Paired t-test

data:  PerfCogNon$CFQ.1 and PerfCogNon$CFQ.2
t = 0.2825, df = 12, p-value = 0.7824
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.228773  9.382619
sample estimates:
mean of the differences
               1.076923


> t.test(PerfCogNon$CFQ.1, PerfCogNon$CFQ.3, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

  Paired t-test
```

```
data:  PerfCogNon$CFQ.1 and PerfCogNon$CFQ.3
t = 3.6936, df = 12, p-value = 0.003072
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.848706 14.920524
sample estimates:
mean of the differences
              9.384615


> t.test(PerfCogNon$CFQ.2, PerfCogNon$CFQ.3, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

   Paired t-test

data:  PerfCogNon$CFQ.2 and PerfCogNon$CFQ.3
t = 3.0248, df = 12, p-value = 0.01057
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.323531 14.291853
sample estimates:
mean of the differences
              8.307692
```

Comparaison entre les 3 conditions pour les femmes enceintes

En procédant de même pour le groupe des femmes enceintes, on obtient :

```
> t.test(PerfCogOui$CFQ.1, PerfCogOui$CFQ.2, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

   Paired t-test

data:  PerfCogOui$CFQ.1 and PerfCogOui$CFQ.2
t = -0.8067, df = 12, p-value = 0.4355
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.372267   7.064575
sample estimates:
mean of the differences
             -4.153846

> t.test(PerfCogOui$CFQ.1, PerfCogOui$CFQ.3, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

   Paired t-test

data:  PerfCogOui$CFQ.1 and PerfCogOui$CFQ.3
t = 0.0871, df = 12, p-value = 0.932
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.62853  18.01315
sample estimates:
mean of the differences
             0.6923077

> t.test(PerfCogOui$CFQ.2, PerfCogOui$CFQ.3, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

   Paired t-test

data:  PerfCogOui$CFQ.2 and PerfCogOui$CFQ.3
t = 0.7403, df = 12, p-value = 0.4734
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -9.417501 19.109809
sample estimates:
mean of the differences
              4.846154
```

15 Test de comparaison de deux proportions sur des groupes indépendants

15.1 Modalités codées à l'aide de deux niveaux d'une variable de type "Character"

Deux échantillons provenant de deux populations différentes ont passé un test commun.
Dans le premier groupe, d'effectif 150, le taux de succès a atteint 68%. Autrement dit, 102 sujets ont passé le test avec succès, et 48 ont échoué.
Dans le deuxième groupe, d'effectif 180, le taux de succès a atteint 55,5%. Autrement dit, 100 sujets ont passé le test avec succès et 80 ont échoué.

Peut-on dire que la seconde population réussit l'épreuve moins facilement que la première ?

La première étape consiste à définir un jeu de données comportant 330 lignes (le nombre de sujets) et deux colonnes : la première contient le groupe auquel appartient le sujet (Gr1 ou Gr2, par exemple), la deuxième contient le résultat au test. Pour définir des données conformes à l'énoncé, on peut procéder de la façon suivante :

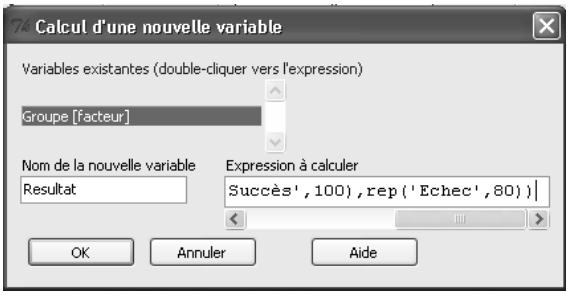
Activez la fenêtre de la console R (fenêtre RGui) et saisissez la commande :

```
Test_Commun <-data.frame(c(rep('Gr1',150),rep('Gr2',180)))
```

Cette commande a pour effet de définir un jeu de données nommé Test_Commun et de lui attribuer 330 lignes. Pour les 150 premières, la première colonne contiendra Gr1, pour les 180 dernières, la première colonne contiendra Gr2.

Dans la fenêtre de R Commander, sélectionnez le jeu de données Test_Commun et passez en mode édition.
Attribuez le nom Groupe à la première variable.

Utilisez le menu Données - Gérer les variables dans le jeu de données actif - Calculer une nouvelle variable. On peut compléter la fenêtre de dialogue comme suit :



N.B. La formule de calcul complète est :

```
c(rep('Succès',102),rep('Echec',48),rep('Succès',100),rep('Echec',80))
```

La formule précédente produit une variable de type "character" qui peut ne pas être reconnue directement comme facteur. Pour s'affranchir de ce problème, on peut utiliser la formule :

```
as.factor(c(rep('Succès',102),rep('Echec',48),rep('Succès',100),rep('Echec',80)))
```

Réalisez ensuite le test à l'aide du menu Statistiques - Proportions - Test de proportions bivarié...

On obtient :

```
        2-sample test for equality of proportions without continuity
        correction

data:  .Table
X-squared = 5.3366, df = 1, p-value = 0.02088
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.22857017 -0.02031872
sample estimates:
   prop 1    prop 2
0.3200000 0.4444444
```

Remarque. Il se peut que la colonne "Résultat" ne soit pas immédiatement reconnue comme étant de type "facteur", format requis pour la procédure demandée. Dans ce cas, le menu apparaît inactif. Pour y remédier, vous pouvez éditer le jeu de données et placer le curseur d'insertion dans l'une des cellules contenant l'une des modalités du résultat (Succès ou Echec) et valider l'édition sans changer la valeur de la cellule. Refermez ensuite la fenêtre d'édition du jeu de données.
Une alternative à la méthode précédente (en cas d'échec de la méthode, par exemple) est de définir une nouvelle variable calculée (Resultat2 par exemple) avec comme formule de calcul : `as.factor(Resultat)`.

Lecture du résultat. On retrouve ainsi les valeurs complémentaires des proportions observées (0,32 pour 68% observés et 0,4444 pour 55,55% observés). La valeur observée de la statistique de test est 5,3366 et la statistique suit une loi du khi-2. On obtient ici une p-value (colonne "p") égale à 0,020, c'est-à-dire à 2%. Au seuil de 5%, on conclut donc à une différence significative entre les deux groupes.

Notez qu'il est également possible de réaliser un test unilatéral. C'est toujours une statistique du khi-2 qui est calculée et le sens dans lequel est prise l'hypothèse alternative n'est pas très clair...

Rappel des formules du cours :

$$p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

$$Z = \frac{f_1 - f_2}{E} \text{ avec } E^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

15.2 Modalités codées à l'aide des nombres 1 et 0

On reprend le même énoncé, mais on choisit maintenant d'utiliser une variable numérique dichotomique pour coder les résultats. Les succès sont codés 1 et les échecs 0.

On peut définir le jeu de données à l'aide des commandes suivantes saisies sur la console :

```
Groupe <- c(rep('Gr1',150),rep('Gr2',180))
Resultat <- c(rep(1,102), rep(0,48), rep(1,100), rep(0,80))
```

```
Test_Commun2 <-data.frame(Groupe,Resultat)
```

Passez ensuite dans l'interface R Commander, rendez actif le jeu de données Test_Commun2 et réalisez le test à l'aide du menu Statistiques - Moyennes - t-test indépendant...

Vous devriez obtenir :

```
        Two Sample t-test

data:  Resultat by Groupe
t = 2.322, df = 328, p-value = 0.02085
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0190118 0.2298771
sample estimates:
mean in group Gr1 mean in group Gr2
        0.6800000         0.5555556
```

Lecture du résultat : les moyennes dans les deux groupes (0.68 et 0.56) correspondent cette fois aux fréquences de succès dans les deux groupes. R Commander teste la significativité de la différence entre ces deux fréquences en utilisant un test de Student, ce qui est peu différent du test vu en cours. La valeur de la statistique de test est 2.32 et la p-value correspondante est 2.08%, résultat pratiquement identique à celui obtenu précédemment.