ANOVA (ANALYSE DE LA VARIANCE)

Veillez à avoir lu les sections relatives à cette séance dans le document disponible sur le site « Introduction aux séances de TP de biostatistique » avant de vous présenter au TP.

Exercice 1

Lors du TP précédent, nous avons comparé deux moyennes, issues chacune d'un échantillon. Ces données étaient réparties selon un seul critère (facteur), càd le régime. Si on garde toujours un seul critère de classification des poids, à savoir le régime, mais que ce critère peut prendre 3 niveaux différents (Régime1 vs Régime2 vs Régime3), le test de t ne sait plus être utilisé. Il faut recourir à une ANOVA. Ce test est aussi nécessaire si le nombre de critères de classification des données est supérieur à 1 (habituellement à votre niveau 2, d'où le nom ANOVA2).

Dans R, vous allez créer des données classées selon un critère ayant 3 niveaux de valeur. Ces données doivent être issues d'une population dans laquelle le paramètre étudié suit une distribution normale.

Par exemple : on souhaite comparer le pH urinaire du chat européen suite à l'absorption d'un aliment spécialement étudié pour acidifier les urines et prévenir la formation de cristaux de struvite. Dans cette étude, 3 aliments (A, B et C) sont comparés. L'aliment C est considéré comme contrôle et est donc un aliment physiologique traditionnel sans composition spéciale pour diminuer le pH urinaire. Le pH urinaire félin de référence avec un aliment physiologique suit une distribution N(6,7 ; $0,15^2$), les moyennes du pH avec un aliment spécifique diminuent celui-ci de 0,2 à 0,5 unité, tout en conservant une variance identique. La taille des 3 groupes est laissée à votre appréciation et celle-ci peut différer d'un groupe à l'autre, seules les moyennes seront différentes…

Vous pouvez générer un premier vecteur contenant les groupes A, B ou C et ensuite un second vecteur contenant les valeurs de pH (l'ordre des données doit être conservé par rapport au premier vecteur). Combinez ensuite vos deux vecteurs pour créer un tableau dans lequel chaque ligne représente un individu et chaque colonne un paramètre pris sur celui-ci (variable ou facteur) et nommez ce nouvel objet « d ».

NB: les fonctions c(), rnorm(), round(), rep() et data.frame() vous seront utiles.

Vos données sont créées… Vous allez maintenant les traiter comme si une tierce personne vous déposait des données et en demandait un analyse afin de voir si les moyennes de pH sont différentes ou non entre les 3 marques d'alimentation.

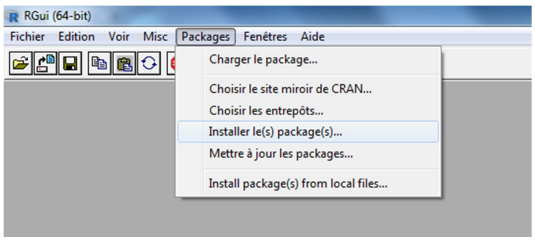
1. *Faire des statistiques descriptives des données afin de mieux se les représenter*

La fonction *summary()* va vous permettre de faire un résumé simple des données. Cependant la fonction *Summarize()* (package **FSA**) est beaucoup plus poussée et permet de tenir compte des différents groupes.
Par exemple, *Summarize(pH~Aliment,data=d,digits=3)* **OU** *Summarize(d$pH~d$Aliment,digits=3)*.
Comme toute fonction, vous pouvez l'enregistrer dans un objet, pour ensuite accéder à des éléments bien précis de cet objet. Par exemple, pour accéder aux moyennes :

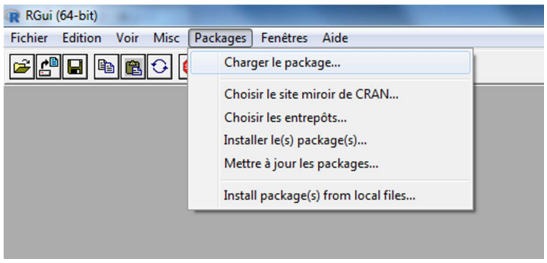
```
Summarize(pH~Aliment,data=d,digits=3)->resume
resume$mean
```

NB: dans R il existe un nombre très important de fonctions diverses. Les fonctions sont regroupées dans des « packages », dont certains sont installés dès l'installation de R (packages de base) et d'autres sont à installer en fonction des besoins de l'utilisateur.

Pour installer un package :



Pour charger le package dans R et pouvoir utiliser ses fonctions, tapez library(nom du package) ou bien, sélectionnez-le dans la liste ouverte via le menu, comme ci-dessous :



2. ***Choisir éventuellement un graphique***

Dans notre situation, nous pourrions afficher un histogramme, dont chaque barre représente la moyenne d'un groupe. La fonction *barplot()* peut vous y aider, en combinant avec le vecteur des moyennes créé au point 1. Vous pouvez paramétrer votre fonction *barplot()* afin d'agrémenter le graphique (nom, couleur, …). Tapez *?barplot* pour accéder à l'aide et ainsi voir les paramètres disponibles.

3. ***Etablir l'hypothèse nulle***

Comme pour tout test statistique, une hypothèse nulle doit être formulée. C'est cette dernière qui sera soit acceptée soit rejetée en fonction de la valeur du test et du seuil de décision (habituellement 5%).

4. ***Choix du test statistique et vérification des prérequis***

Ce choix s'opère en fonction de la question qui se pose au travers de l'hypothèse nulle. Dans notre cas, nous souhaitons mettre en évidence des différences de moyennes de pH, valeurs continues classées en 3 catégories. Notre choix se portera donc sur une ANOVA1 (pour 1 critère de classification, à savoir l'aliment). Toutefois, l'usage de ce test comporte certaines contraintes comme :

- La normalité des données
- L'égalité de variances des échantillons

Si les données s'écartaient fortement de ces contraintes, vous seriez probablement obligés d'appliquer une autre méthode statistique (cf. test non paramétrique de Kruskall-Wallis).

La normalité est testée par série mais moins il y a de données dans une série, plus la fiabilité de ce test sera naturellement faible. Plusieurs tests existent, les plus courants étant le test de Kolmogorov-Smirnov (*ks.test()*) et le test de Shapiro-Wilk (*shapiro.test()*). Ces fonctions testent la normalité de chaque série individuellement. Pour tester la normalité directement par groupe, utilisez la fonction *byf.shapiro(d$pH~d$Aliment)* du package **RVAideMemoire**. Outre ces tests de normalité, on peut aussi s'aider de graphiques afin d'apprécier si les données rentrent dans les limites de la normalité. La fonction *qqPlot()* du package **CAR** pour faire le graphique de normalité avec les intervalles d'une série et la fonction *byf.qqnorm()* du package **RVAideMemoire** pour faire simultanément les graphiques de plusieurs séries.

L'égalité de variance peut, elle, être testée soit 2 à 2 via un test de F (rapport de variance) via *var.test()* ou mieux dans la cadre d'une ANOVA, globalement via le test de Bartlett (fonction *bartlett.test()* du package **stats**.

Une fois ces prérequis réalisés, vous pouvez passer à l'étape de l'analyse statistique des données proprement dite. La fonction *aov()* réalise une comparaison de moyennes par analyse de variances telle que décrite dans le cours. Il faut enregistrer l'instruction dans un objet et ensuite éditer cet objet avec la fonction générique *summary()* ou bien la fonction *anova()*, afin de voir le tableau d'ANOVA classique et de conclure à l'absence ou non d'un effet.

Exercice 2

Nous allons maintenant réaliser les calculs détaillés de l'ANOVA1 dans Excel. Pour cela, il va falloir exporter le dataset créé dans l'exercice 1 vers Excel. La fonction *write.csv2()* va permettre d'exporter vos données vers le répertoire de travail de R. Vous devez fournir au minimum le nom du dataset et ensuite, et entre guillemets, le nom complet avec son extension ("nom.csv"). Pour les autres paramètres ou des exemples, n'hésitez pas à consulter l'aide de la fonction (?write.csv2).

Une fois vos données exportées, ouvrez le fichier CSV avec Excel. Vous devriez obtenir quelque chose de similaire :

	A	B	C	D
1		Aliment	pH	
2	1	A	6,82	
3	2	A	6,64	
4	3	A	6,63	
5	4	A	6,98	
6	5	A	6,8	
7	6	A	6,71	
8	7	A	6,84	
9	8	A	6,79	
10	9	A	6,69	
11	10	A	6,83	
12	11	A	6,7	
13	12	A	6,91	
14	13	B	6,03	
15	14	B	6,42	
16	15	B	6,25	
17	16	B	6,12	

La colonne A, qui numérote simplement les individus, peut éventuellement être effacée. Les données peuvent aussi être présentées différemment pour faciliter vos calculs dans Excel :

	A	B	C	D
1	A	B	C	
2	6,82	6,03	6,38	
3	6,64	6,42	6,9	
4	6,63	6,25	6,59	
5	6,98	6,12	6,63	
6	6,8	6,22	6,61	
7	6,71	6,13	6,51	
8	6,84	6,06	6,59	
9	6,79	6,24	6,46	
10	6,69		6,44	
11	6,83		6,59	
12	6,7			
13	6,91			
14				

Pour rappel, l'ANOVA compare la variance des données calculée de deux manières différentes. La première manière de calculer la variance est de calculer la variance entre les moyennes des lots (en tenant compte de la taille de chaque lot). L'autre manière de calculer la variance consiste à calculer la variance dans les lots, et à combiner ces variances (par une moyenne pondérée) pour obtenir une estimation de la variance des données. Cette estimation est indépendante du fait qu'il y ait une

différence entre les moyennes des lots, et fournit donc une estimation non biaisée de la variance recherchée. Par conséquent, si on calcule le rapport F="variance entre lots"/"variance dans les lots", on obtiendra une valeur proche de 1 si l'hypothèse nulle est vraie, et supérieure à 1 dans le cas contraire. Une valeur de F inférieure ou égale à 1 sera donc indicatrice du fait que l'hypothèse nulle peut être acceptée, et une valeur anormalement supérieure à 1 indiquera que H0 doit être rejetée. Comme d'habitude, "anormalement supérieure à 1" signifie que la probabilité d'avoir une valeur de F aussi élevée ou plus élevée est faible (typiquement, inférieure à α = 5%). Comme pour la distribution de Student, Excel permet d'éviter de devoir consulter une table de F en utilisant la fonction *LOI.F(F;ddl1;ddl2)*, dans laquelle ddl1 désigne les degrés liberté associés au numérateur (le nombre de lots - 1), et ddl2 désigne le nombre de degrés de liberté associés au dénominateur (nombre total de données - nombre de lots).

Une manière simple de présenter les résultats est de recourir à une **table de l'analyse de la variance**. On calcule successivement les sommes de carrés due à l'effet, due à l'erreur et totale, ainsi que les degrés de liberté correspondants, et on les dispose comme expliqué au cours. Les calculs des variances "entre les lots"et "dans les lots" et de la statistique F est alors très simple.

Dans ces calculs de sommes de carrés, **chaque donnée contribue à chacune des 3 sommes**…

1. ***Somme des carrés due à l'effet***: cette somme s'obtient en prenant pour chaque donnée la moyenne de son groupe, en soustrayant de celle-ci la moyenne générale, et en élevant le résultat au carré. Les contributions de chaque donnée sont alors additionnées pour donner la "somme des carrés due à l'effet".
2. ***Somme des carrés due à l'erreur***: on la calcule en soustrayant de chaque donnée la moyenne de son groupe et en élevant le résultat obtenu au carré. Les contributions sont alors sommées pour donner la "somme des carrés de l'erreur".
3. ***Somme des carrés totaux***: cette somme s'obtient en soustrayant de chaque donnée la moyenne générale, en élevant le résultat obtenu au carré et en sommant les contributions ainsi obtenues.

La table de l'analyse de la variance s'obtient alors sans peine:

Source	Somme ²	Degré de lib.	Carrés moy.	F	p>F
Effet					
Erreur					
Total					

RAPPEL : les carrés moyens s'obtiennent en divisant la somme des carrés par le nombre de degrés de liberté lui correspondant.

Dans tous les cas, la somme Total doit être égale à l'addition de la somme Effet (aliment dans notre cas) et de la somme Erreur.

Vous devriez alors obtenir les mêmes résultats (F et p-value) que dans R et donc les conclusions (acceptation ou rejet de l'hypothèse nulle au seuil choisi) doivent être similaires.

Exercice facultatif pour le brainstorming

L'ANOVA2 se caractérise par le fait que les données sont classées selon deux critères. Dans notre exemple sur l'alimentation des chats, nous pourrions ajouter une variable sexe (mâle >< femelle) à nos données, en essayant de respecter un ratio 50:50 dans les sexes.

L'analyse de ces nouvelles données est un peu différente par la présence de cette deuxième variable, mais peut à la fois être réalisée dans Excel et dans R. Vous aurez une conclusion à tirer pour chacun des effets (+ l'interaction éventuelle en plus).