Université d'Evry Val d'Essonne **18-20 janvier 2017**
Bases statistiques pour la biologie

Probabilités

Exercice 1

Dans une collection de roches, 40% sont de type basalte et 60% sont de type granite. Cinq roches sont choisies au hasard (sans remise) à des fins d'analyses chimiques. On note X le nombre de roches de type basalte dans l'échantillon.

1. Préciser la loi de probabilité de X et ses paramètres.
2. Calculer la probabilité que l'échantillon contienne seulement des roches de même type.

Exercice 2

Dans une population de patients présentant une douleur abdominale, 30% des individus ont une appendicite aiguë. Parmi ces derniers, 70% ont une température corporelle supérieure à $37,5°C$ alors que chez les patients sans appendicite, une température supérieure à $37,5°C$ est retrouvée dans 40% des cas.

1. Quelle est la population d'étude ?
2. Dans cette population, quelle est la probabilité d'avoir une température corporelle supérieure à $37,5°C$?
3. Pour un patient dont la température est supérieure à $37,5°C$, quelle est la probabilité d'avoir une appendicite aiguë ?

Exercice 3

Notons X la variable aléatoire correspondant au niveau d'expression (normalisé) d'un gène G. De nombreuses expériences ont permis d'établir que $X \sim N(0.2, 1.3)$.

1. Représenter graphiquement la distribution de X.
2. Déterminer les probabibilités suivantes : $P(X > 0.6)$, $P(X < -0.2)$, $P(0.1 < X < 0.7)$, $P(0 \leq X < 0.5)$.
3. Déterminer les valeurs de a telles que : $P(X \leq a) = 0.45$, $Pr(X > a) = 0.62$

Une modification des réglages du scaner de la plateforme conduit à multipluier par 1.2 toutes les intensités X. Quelles est la loi de la nouvelle variable aléatoire Y correspondant au niveau d'expression du gène G ?

Exercice 4

On suppose que la glycémie est distribuée normalement dans la population, avec une moyenne de 1 g/l et un écart-type de 0,03 g/l. On mesure la glycémie chez un individu.

1. Calculer la probabilité pour que sa glycémie soit :
   (a) inférieure à 1,06
   (b) supérieure à 0,9985
   (c) comprise entre 0,94 et 1,08
2. On mesure la glycémie chez 1 000 individus. Donner le nombre moyen d'individus dont la glycémie est supérieure à 0,99.

Estimation

Exercice 5

Pour estimer la densité bactérienne d une suspension, on ensemence avec le même volume v 10 boîtes de Pétri sur lesquelles on compte les nombres suivants de colonies (qui sont aussi les nombres de bactéries présentes dans chacun des volumes v) :

$$47, 47, 55, 47, 56, 56, 38, 42, 48, 45$$

On note N le nombre de bactéries présentes dans un volume v.

1. Estimer l'espérance μ de N.
2. Estimer la variance σ^2 de N.
3. On suppose que N suit une distribution de Poisson. Donner une autre estimation de σ^2 que celle obtenue en 2.

Exercice 6

Pour déterminer la concentration en glucose d un échantillon sanguin, on effectue des dosages à l aide d une technique expérimentale donnée. On considère que le résultat de chaque dosage est une variable aléatoire normale. On effectue 10 dosages indépendants, qui donnent les résultats suivants (en g/l) :

$$0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08$$

Calculer un interval de confiance de cette concentration de niveau 95%.

Tests d'hypothèses

Exercice 7

Le temps de réaction moyen des souris d'un certain élevage à un test déterminé est de 19 minutes. On désire expérimenter un produit pharmaceutique sur ces souris. On administre à 8 d'entre elles une dose de ce produit et l'on observe les temps de réaction suivants (en minutes) :

15 14 21 12 17 12 19 18

On suppose les temps de réaction normalement distribués. Au niveau $\alpha = 5\%$, l'action du produit est-elle significative ?

Exercice 8

La quinine est une molécule utilisée dans le traitement du paludisme. Des médecins ont constaté que les patients qui suivent un traitement à base de quinine semblent présenter des réactions allergiques au soleil plus fréquentes.

1. Pour étudier ce phénomène, une étude préliminaire portant sur 10 patients suivant un traitement à base de quinine a été mise en place. Des études antérieures on permis d'établir que le pourcentage d'individus dans la population générale qui présente une réaction allergique au soleil est de 20%. Sur les 10 patients traités, 3 ont eu une réaction allergique. Proposez un test statistique pour vérifier l'hypothèse des médecins et conclure.
2. Une plus grande étude portant sur 1000 patients suivant un traitement à base de quinine a été mise en place. Sur les 1000 patients traités, 357 ont eu une réaction allergique. En utilisant l'approximation gaussienne, proposez un nouveau test statistique pour vérifier l'hypothèse des médecins et conclure.

Exercice 9

En population générale, la proportion d'enfants dont la maturation osseuse atteint un retard de un an ou plus (par rapport à une certaine norme) est $p = 20\%$. Dans le cadre d'une étude portant sur les conséquences éventuelles d'une exposition modérée au fluor sur la santé des enfants, on prévoit d'observer 15 enfants habitant à proximité d'une source de fluor.

1. Construire un test statistique de niveau $\alpha = 5\%$ permettant de déterminer si une exposition au fluor augmente significativement le risque d'avoir un retard de la maturation osseuse.
2. Quelle est la puissance du test si le risque d'avoir un retard de la maturation osseuse pour un enfant exposé au fluor est en réalité de 30% ?
3. Sur les 15 enfants observés, 5 présentent un retard. Que peut-on conclure ?

Exercice 10

Un laboratoire de recherche étudie, sur une nouvelle espèce de vers (C. marginalus), les gènes pouvant être impliqués dans la mort programmée. La distribution de la durée de vie de ces vers est inconnue mais des études antérieures ont montré que l'espérance de vie de ces vers est $\mu = 250$ heures et que l'écart-type de leur durée de vie est $\sigma = 24$ heures. Après modification d'un gène supposé intervenir dans la mort programmée, on a relevé sur un échantillon de 80 vers une durée de vie moyenne de 256 heures.

1. Peut-on conclure que la modification de ce gène induit une augmentation significative (au niveau 5%) de l'espérance de vie des vers ?

2. On suppose que la modification du gène induit une réelle augmentation de l'espérance de vie de 7 heures. Quelle est la puissance du test mis en oeuvre ?
3. Combien de vers faudrait-il inclure dans l'étude pour avoir une puissance de 90% ?

Exercice 11

On envisage d'ajouter un adjuvant au traitement usuel d'un certain type de rhumatisme. Sans adjuvant, la durée séparant deux crises de récurrence rhumatismale peut être modélisée par une variable aléatoire suivant une distribution normale d'espérance $\mu = 560$ (exprimée en jours). On administre le traitement avec adjuvant à 10 sujets. Les durées de récurrence observées sont les suivantes :

$$646, 573, 485, 752, 742, 636, 607, 665, 506, 575.$$

Au niveau $\alpha = 5\%$, l'adjuvant modifie-t-il significativement la durée moyenne de récurrence ?

Exercice 12

Un producteur de lait souhaite comparer le rendement moyen des vaches normandes et hollandaises de son unité de production. Pour ce faire, il a relevé la production de lait (exprimée en kg) de 10 vaches prises au hasard dans chaque groupe. On suppose que la production dans chaque groupe suit une distribution normale.

Normandes	552	464	423	506	497	544	486	531	496	501
Hollandaises	487	489	470	482	494	500	504	567	482	526

Conclure au vu de ces données.

Exercice 13

Un laboratoire pharmaceutique produit des tubes de pommade dont les poids suivent une distribution normale. On dispose de deux échantillons issus de 2 sites de production différents. Les poids sont donnés dans le tableau suivant :

										$\sum x_i$	$\sum x_i^2$
Echantillon 1	56.4	57.5	55.8	54.3	58.9	56.9	54.8	54.2	58.1	506.9	28572.45
Echantillon 2	54.6	58.2	60.3	59.5	61.1	58.7	59.8	57.5		469.7	27605.93

1. Les variances des 2 échantillons sont-elles significativement différentes ?
2. Le poids des tubes est-il significativement différent d'un site de production à l'autre ?

Exercice 14

On veut étudier l'effet secondaire d'un certain médicament sur le taux de cholestérol. On administre le médicament à 10 personnes et on relève pour chacune de ces personnes le taux de cholestérol. On obtient pour cet échantillon un taux moyen de 2.4 g/l avec une variance empirique corrigée de 12 $(\mathrm{g/l})^2$. Sur un autre échantillon de 8 personnes n'ayant pas été traitées avec ce médicament, on relève un taux moyen de 2 g/l avec une variance empirique corrigée de 9 $(\mathrm{g/l})^2$. On suppose que le taux de cholestérol d'un individu suit une distribution normale dont les paramètres sont μ_A et σ_A^2 pour une personne ayant pris le médicament, μ_B et σ_B^2 pour une personne n'ayant pas pris le médicament. On suppose en outre que $\sigma_A^2 = \sigma_B^2$. Comparer μ_A et μ_B à l'aide d'un test de niveau 5%.

Exercice 15

On fait une numération globulaire à un groupe de 20 personnes à deux périodes différentes de l'année. Pour chaque sujet, on note les résultats des deux numérations (à multiplier par 10^5) :

Sujet	1	2	3	4	5	6	7	8	9	10
Janvier	46	38	42	47	48	40	40	43	42	49
Septembre	48	47	44	45	51	44	47	48	47	57

On suppose que les sujets sont mutuellement indépendants et suivent une loi gaussienne. Tester au niveau 0.05 l'hypothèse selon laquele les résultats de la numération sont les mêmes aux deux périodes.

Exercice 16

La quantité de bactéries par cm^3 de lait provenant de 8 vaches différentes est estimée juste après la traite et 24h plus tard. La distribution des résultats obtenus est supposée normale. Au niveau $\alpha = 5\%$, existe-t-il un accroissement significatif du nombre de bactéries par cm^3 de lait au cours du temps ?

Vache	1	2	3	4	5	6	7	8
Juste après la traite	12000	13000	21500	17000	15000	22000	11000	21000
24h après la traite	14000	20000	31000	28000	26000	30000	16000	29000

Exercice 17

Le tableau suivant donne la répartition (en pourcentages) des quatre groupes sanguins pour l'ensemble de l'Europe :

O	A	B	AB
40%	43%	12%	5%

Pour un échantillon de 100 individus prélevés au hasard dans la population d'un région montagneuse (et isolé) de l'Europe, on a relevé les effectifs suivants :

O	A	B	AB
35	35	20	10

Y a-t-il conformité entre ces observations et la répartition pour l'ensemble de l'Europe au seuil $\alpha = 5\%$?

Exercice 18

Une boîte de Petri a été photographiée au microscope. La photographie est divisée en carrés de surfaces égales. Le dénombrement dans chaque carré des colonies de bactéries donne le tableau suivant :

nombre de colonies par carré	0	1	2	3	4	5
nombre de carrés	10	24	34	23	6	3

1. Estimer le nombre moyen de colonies par carré.
2. Peut-on accepter l'hypothèse selon laquelle le nombre de colonies par carré est distribué suivant une loi de Poisson ?

Exercice 19

Après de nombreuses années d'études cliniques, on a constaté que pour les malades atteints d'un cancer anaplasique bronchopulmonaire primitif, la survie sans traitement, une fois le diagnostic posé, se distribue de la façon suivante :

Survie (en mois)	<6	6 à 12	12 à 24	>24
Fréquence des survies	0.45	0.35	0.15	0.05

Pour 60 patients soumis à un traitement T associant une polychimoithérapie première suivie d'une radiothérapie on a observé les résultats suivants :

Survie (en mois)	<6	6 à 12	12 à 24	>24
Nombre de patients	6	24	12	18

Au vu de ces résultats, peut-on conclure (au niveau 5%) que le traitement a un effet significatif sur la survie ?

Exercice 20

On étudie, chez les enfants asthmatiques, le lien éventuel entre intensité de l'asthme et présence d'eczéma (pendant l'obervation ou antérieurement à celle-ci). L'étude de 200 enfants asthmatiques a fourni les résultats suivants :

eczéma \ asthme	fort	moyen	léger
présent	24	6	5
passé	30	30	10
jamais	18	54	23

1. Sous l'hypothèse d'indépendance des deux caractères asthme et eczéma, calculer les effectifs théoriques des 9 classes.
2. Au seuil $\alpha = 5\%$ peut-on conclure à l'indépendance des deux caractères ?

Exercice 21

Dans une population P d'hommes qui a été suivie pendant une période de 4 ans, on a sélectionné par tirage au sort 100 sujets qui avaient maigri au cours des 4 ans (poids final inférieur au poids initial de plus de 1kg), 100 sujets dont le poids n'avaient pas varié de plus de 1kg et 100 sujets qui avaient grossi. La répartition des 300 sujets selon l'évolution de leur cholestérolémie est donnée dans le tableau suivant :

Poids \ Cholestérolémie	a diminué	a augmenté
a diminué	52	48
n'a pas varié	45	55
a augmenté	32	68

Au niveau $\alpha = 5\%$, peut-on conclure qu'il existe une relation significative entre les modifications de poids et les modifications de cholestérolémie ?

Exercice 22

Deux lots de souris doivent sortir d'un labyrinthe et disposent de 8 sorties correspondant aux 8 directions de la rose des vents. Le premier lot est formé de souris de laboratoire, le second de souris sauvages capturées au Nord-Est du laboratoire.

Direction de fuite	N	NO	O	SO	S	SE	E	NE	Total
Souris de laboratoire	17	25	13	28	19	20	22	16	160
Souris sauvages	26	17	9	2	3	16	33	54	160

Les directions de fuite sont-elles réparties de la même façon dans les deux groupes ?

Exercice 23

Une nouvelle technique de dosage du glucose sanguin vient d'être mise au point. Pour un même échantillon de sang, sept dosages effectués à l'aide de cette nouvelle technique ont donné les résultats suivants (en g/l) :

1.17 1.16 1.16 1.19 1.21 1.19 1.18

On admet que les sept mesures sont des variables aléatoires indépendantes et identiquement distribuées de loi $N(\mu, \sigma^2)$ où σ^2 caractérise la précision du procédé. La technique utilisée jusqu'alors était caractérisée par un écart-type de 0.05 mg/l. Peut-on dire que la nouvelle technique est plus précise que l'ancienne ?

Exercice 24

Le degré d'acidité a été mesuré dans deux types de solutions chimiques A et B. Dans la solution A, six mesures ont été faites, avec un pH moyen de 7,52 et une variance estimée $S_A^{*2} = 0,000576$. Dans la solution B, cinq mesures ont été faites, avec un pH moyen de 7,49 et une variance estimée $S_B^{*2} = 0,001024$. On suppose les observations normalement distribuées.

1. Peut-on supposer que la précision des mesures est identique dans les deux groupes (tester l'égalité des variances) ?
2. Les deux solutions ont-elles des pH significativement différents (tester l'égalité des espérances) ?

Exercice 25

On sait que la concentration plasmatique du calcium d'un sujet sain suit une distribution normale d'espérance μ inconnue et de variance $\sigma^2 = 1.21 \mu mol/ml$. Chez 18 personnes, on a trouvé une moyenne $m = 3.2 \mu mol/ml$.

1. Peut-on conclure que la calcémie moyenne de ces 18 personnes est significativement augmentée ?
2. Proposer un intervalle de confiance bilatéral de niveau 95% pour l'espérance μ.

Exercice 26

Lors d'une étude médicale, on a déterminé le génotype de $n = 1000$ personnes. Les observations sont les suivantes :

	AA	Aa	aa
Effectifs	652	310	38

Proposer un test permettant de savoir si la population est sous l'équilibre d'Hardy-Weinberg.

Exercice 27

La notice d'un sirop contre la toux indique comme valeur de référence pour la moyenne m_0 de l'agent actif $40g$/litre. Le contrôleur de la fabrication décidera d'arrêter provisoirement la production si la moyenne m inconnue est strictement infèrieure à cette valeur de référence. Il souhaite ne prendre qu'un risque minime c'est-à-dire $\alpha = 0.01$ en décidant d'arrêter à tort la production.

Le contrôleur de la fabrication prélève de manière indépendantes 9 bouteilles au hasard dans la production et mesure la quantité d'agent actif. Les résultats pour ces 9 dosages indépendants sont les suivants (en g/litre) :

38.7, 39.6, 37.9, 40.6, 40.5, 37.7, 41.2, 37.5, 39.1.

On suppose que la quantité d'agent actif conditionnée dans une bouteille de sirop est une variable normale, centrée sur la vraie valeur m (absence de biais). Sous `R`,

1. proposer un test au niveau 1% permettant de savoir quelle décision prendre ;
2. déterminer un intervalle de confiance à 99% pour m ;
3. faire une étude de puissance.

Exercice 28

Un échantillon de 40 poissons de la même espèce a fourni les poids suivant (en g) :

61, 82, 92, 97, 101, 104, 109, 118, 131, 155, 69, 82, 93, 97, 101, 104, 110, 120, 133, 145, 105, 110, 121, 138, 166, 74, 85, 93, 99, 102, 106, 110, 125, 140, 171, 79, 87, 94, 99, 102

1. Présenter une synthèse de ce tableau (graphiques et paramètres).
2. La distribution de cette variable peut-elle être considérée comme normale ?
3. Déterminer un intervalle de confiance à 5% de la moyenne.
4. La moyenne est-elle significativement différente de 100 avec un risque de 5% ? de 1% ?

Exercice 29

On veut étudier l'effet secondaire d'un certain médicament sur le taux de cholestérol. On administre le médicament à 10 personnes et on relève pour chacune de ces personnes le taux de cholestérol. On obtient pour cet échantillon un taux moyen de 2.4 g/l avec une variance empirique de 12 $(g/l)^2$. Sur un autre échantillon de 8 personnes, n'ayant pas été traitées avec ce médicament, on relève un taux moyen de 2 g/l avec une variance empirique de 9 $(g/l)^2$. On admet que l'on peut supposer que le taux de cholestérol d'un individu est une variable aléatoire de loi gaussienne d'espérance μ_A et de variance σ_A^2 pour une personne ayant pris le médicament et μ_B et σ_B^2 pour une personne n'ayant pas pris le médicament.

1. Peut-on conclure à partir des données observées à une différence significative entre les variances σ_A^2 et σ_B^2 au niveau 5% ?
2. Comparer μ_A et μ_B à l'aide d'un test au niveau 5%.

Exercice 30

Plusieurs sujets sont choisis au hasard dans une population et, parmi ceux-ci, certains sont tirés au sort pour recevoir un traitement (Groupe A), les autres devant servir de témoins (Groupe B).
Le traitement est censé modifier le résultat d'un dosage biologique. Les résultats, exprimés en mg/l, sont les suivants :

Groupe A	6,50	5,50	8,00	7,00	6,00		
Groupe B	7,00	8,50	8,00	7,50	9,00	7,20	8,20

1. Quel test choisir ?
2. Préciser les hypothèses (H_0) et (H_1).
3. Rappeler les conditions d'application du test utilisé.
4. Peut-on admettre ($\alpha = 5\%$) que le traitement modifie le paramètre biologique ?

Exercice 31

On souhaite étudier l'effet d'une nouvelle stratégie de traitement du diabète sur la glycémie. On dose la glycémie chez 15 sujets avant le début du nouveau protocole (série A) et 3 mois après (série B) :

A	2,47	3,09	2,14	2,47	3,06	2,72	2,29	1,90	2,34	2,75	2,67	2,80	2,51	2,23	2,20
B	2,30	2,96	2,23	2,34	2,84	2,59	2,15	1,88	2,32	2,65	2,68	2,58	2,43	2,02	2,17

Le nouveau protocole est-il efficace ?

Exercice 32

Cinq rats sont entraînés à imiter un rat leader dans un labyrinthe en T, pour atteindre une source de nourriture. Puis ces rats sont ensuite transférés dans une situation où par imitation d'un rat leader, ils apprennent à éviter un choc électrique. Leur comportement dans cette situation est comparé à celui de rats n'ayant pas été entraînés à suivre un leader. La comparaison se fait en terme de nombre d'essais nécessaire à chaque rat pour obtenir 10 réponses d'évitement lors de 10 essais.

Exp	78	64	75	45	82
Témoins	110	70	53	51	

Les 5 rats préalablement conditionnés à imiter un congénère réussissent-ils rapidement que les autres à éviter les chocs ?

Exercice 33

On a mesuré sur Dunaliella Marina, la quantité d'azote protéique par cellule, à la même date et dans des conditions expérimentales identiques, sur une culture témoin et sur une culture préalablement irradiée. On pense que l'irradiation favorise un développement anormal des cellules.

Culture témoin	1.65	2.00	1.69	2.20	2.13	1.66	2.30	1.87	1.74	1.97
Culture irradiée	2.29	2.57	2.66	2.45	2.97	2.27	1.76	2.74	2.36	

Interpréter les résultats.

Exercice 34

On souhaite expérimenter l'action d'un produit pharmaceutique sur des souris. Le temps de réaction X des souris d'un certain élevage à un test déterminé est supposé suivre une distribution normale, le temps de réaction moyen μ étant de 19 minutes dans des conditions normales. On administre à dix souris de cet élevage une dose du produit pharmaceutique, puis on leur fait passer le test. On obtient les temps de réaction suivants :

15 mn, 14 mn, 21 mn, 12 mn, 17 mn, 12 mn, 19 mn, 18 mn, 20 mn, 21 mn.

Remarque : Dans ce qui suit, on pourra utiliser la fonction `t.test` de `R` pour vérifier les résultats. La fonction `qt` permet le calcul des fractiles d'une loi de Student et `pt` le calcul de sa fonction de répartition.

1. Saisir les données sous la forme d'un vecteur `donnees`. Calculer $\bar{x}$ et $s^\star$ à l'aide des fonctions `sum` et `length` puis comparer avec les fonctions `mean` et `var` de `R`.
2. (a) Écrire une fonction `student.test` qui revoie la p-valeur associé à un test de Student de comparaison de deux écahantillons gaussiens de même variance. Elle prend en arguments
   - un vecteur de données `x` et un vecteur de donées `y`
   - (optionnel) une chaîne de caractère `alternative` pouvant prendre les valeurs `"two.sided"`, `"less"` ou `"greater"` selon la forme désirée.
3. On s'attend à ce que le produit ait un effet stimulant et donc que le temps de réaction des souris ayant reçu une dose soit en moyenne inférieur à celui des souris n'en ayant pas reçu. A l'aide de la fonction construite précédemment, tester au niveau 5% l'effet du produit. Vérifier le résultat à l'aide de la fonction `t.test`. Calculer la valeur du seuil correspondant au niveau choisi.
4. Calculer la valeur de la puissance π pour diverses valeurs de μ_1, le paramètre sous l'hypothèse alternative. Tracer la courbe de puissance.

Exercice 35

On souhaite comparer trois traitements notés A, B, C contre l'asthme : le traitement B est un nouveau traitement, que l'on souhaite mettre en compétition avec les traitements classiques A et C. On répartit par tirage au sort les patients et on mesure sur chacun la durée en jours avant la prochaine crise d'asthme.

1. Visualisation des données.
   (a) Stocker les données `asthme.dat` dans une variable de votre choix à l'aide de la fonction `read.table` (mettre l'option `header` à `TRUE` – pourquoi ?). La table ainsi créée a deux colonnes : l'une contenant le délai observé avant la prochaine crise d'asthme, l'autre le type de traitement reçu. Vous pourrez utiliser les commandes `attach` / `detach` pour travailler dans le contexte des données `asthme`.
   (b) Faire un résumé numérique des données à l'aide de la commande `summary`. À l'aide de la commande `tapply`, faire un résumé numérique par traitement. Représenter graphiquement ces résultats à l'aide de boîtes à moustaches (fonction `boxplot`). Que peut-on en conclure ?
2. Analyse de la variance.
   (a) Créer une fonctions `somme.carres` qui prend en argument un vecteur d'observation `donnees` et un vecteur de facteur `groupe`, et qui renvoie dans une liste
      — la somme des carrés totale SCT et le nombre de degrés de liberté associé,
      — la somme des carrés résiduelle SCR et le nombre de degrés de liberté associé,
      — la somme des carrés des facteurs SCF et le nombre de degrés de liberté associé.
   (b) Calculer la valeur observée de la statistique de test de l'anova 1, la valeur du seuil de rejet au niveau α de votre choix et la valeur de la p-valeur.
   (c) Comparer les résultats avec ceux de la fonction `anova` de `R`.
3. Étude de contrastes.
   (a) On note μ_A, μ_B et μ_C les délais moyens obtenus avec les traitements A, B et C. Tester l'hypothèse selon laquelle $\mu_A = \mu_C$, puis $\mu_A = \mu_B$. On utilisera comme estimateur de la variance la valeur des CMR obtenu à la question précédente. Conclure.
   (b) On souhaite déterminer l'apport du traitement B par rapport au traitement A. À cet effet, nous allons étudier le contraste C_{AB}. Construire un intervalle bilatère puis unilatère à 95% pour C_{AB}. Conclusion ?

Exercice 36

On mesure les rendements en blé sur 21 parcelles de même aire représentant 6 niveaux :

niveau	1	2	3	4	5	6
	53.4	76.9	55.1	71.6	90.1	76.4
	64	62.4	72	80.4	89.6	77
	68.3	70.8	67	67.4	81.6	77.8
	65.5	70			82.2	

Table 1 – données blé

1. On note $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ les espérances de rendements pour les divers niveaux. On considère le modèle $H_6 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$.
   (a) Créer un vecteur `donnees` contenant les observations et un vecteur `niveaux` contenant les facteurs associés au modèle H_6.
   (b) Représenter les boîtes à moustaches et les dotplot pour les différents groupes à l'aide des fonctions `boxplot` et `dotchart`.
   (c) Tester $H_1 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ contre H_6 à l'aide d'une anova 1.
2. Les niveaux 1,2,3 correspondent à une région "Nord", les niveaux 4,5,6 à une région "Sud". Même question qu'en 1 en testant H_1 contre $H_2 : \mu_{Nord} \neq \mu_{Sud}$. Le vecteur des facteurs est noté `regions`.
3. Les niveaux 1 et 4 correspondent à une variété A d'espérance μ_A, ; les niveaux 2 et 5 correspondent à une variété B d'espérance μ_B, les niveaux 3 et 6 correspondent à une variété C d'espérance μ_C. Même question qu'en 1 en testant H_1 contre $H_3 : \mu_A \neq \mu_B \neq \mu_C$. Le vecteur des facteurs est noté `varietes`.
4. Refaire le problème en Anova 2 avec comme facteurs régions et variétés. Utiliser la fonction `interaction.plot` pour constater que les interactions peuvent être négligées.

Exercice 37

On souhaite étudier l'effet du niveau de fertilisation et de la rotation de culture sur le poids des grains de colza. On compare pour cela 2 niveaux de fertilisation (notés 1 pour faible et 2 pour fort) et 3 types de rotation de culture maïs / blé / colza / blé : A (sans enfouissement de paille), B (avec enfouissement de paille) et C (avec quatre années de prairie temporaire entre chaque succession sans enfouissement de paille).

1. Questions préliminaires
   (a) Charger le fichier de données `colza.dat` à l'aide de la fonction `read.table`, contenant le poids moyen mesuré dans chacune des 60 parcelles ainsi que les conditions de fertilisation et de rotation associées.
   (b) Tracer les boîtes à moustaches pour les différentes niveaux des facteurs (fonction `boxplot`).
   (c) Tracer le graphe des interactions entre entre les deux facteurs (fonction `interaction.plot`).
2. Analyse de la variance
   (a) Calculer les sommes des carrées utiles pour le modèle complet (avec interaction), c'est-à-dire SCT, SCR, SCF_1, SCF_2 et SCF_1F_2, où F_i représente le facteur i.
   (b) Tester l'interaction entre les facteurs, l'effet du facteur fertilisation et l'effet du facteur rotation. Enfin, tester l'intérêt du modèle.
   (c) Comparer vos résultats avec ceux de la fonction `lm`. Étudier le modèle avec interactions et le modèle additif. Étudier également le modèle à un seul facteur, celui de fertilisation (anova 1).

Exercice 38

On s'intéresse aux performances sportives d'enfants de 12 ans. Chaque enfant passe une dizaine d'épreuves (courses, sauts, lancers, etc.), et les résultats sont synthétisés dans un indice global, noté Y. On cherche à mesurer l'incidence sur ces performances de deux variables : la capacité thoracique X_1 et la force musculaire X_2. Ces trois quantités, Y, X_1 et X_2, sont repérées par rapport à une valeur de référence, notée à chaque fois 0, les valeurs positives étant associées aux « bonnes » performances.

Les mesures associées à un échantillon de 60 enfants sont stockés dans le vecteur `data`, dont vous disposerez sous `R` une fois chargé le fichier `perf.dat`.

On adopte, au moins dans un premier temps, le modèle

$$Y = a_1\ X_1 + a_2\ X_2 + \varepsilon,$$

où ε est un résidu non expliqué par le modèle : les ε_i associés aux différents individus seront modélisés par des $\mathcal{N}(0, \sigma^2)$ indépendantes (Notons que le « calage » des données autour de zéro se traduit par le fait que, quand $X_1 = X_2 = 0$, alors $\mathbb{E}(Y) = 0$).

1. Représenter le nuages de points à l'aide de la fonction `plot`.
2. Estimer les paramètres a_1 et a_2. Vérifiez vos résultats avec la fonction `lm` de `R`.
3. Tester H_2 contre H_0 : conclusion ?
4. On adopte maintenant le modèle $Y = a\ X_1 + b$. Estimer a et b, et représenter les données et la droite de régression associée. Observer également les résidus du modèle. Enfin, vous testerez H_1 contre H_0.

Exercice 39

On mesure le taux de leucocytes T_4 chez le chat X jours après avoir inoculé à l'animal le virus FeLV, analogue du HIV. On appelle Y le logarithme de ce taux. Le tableau suivant donne les mesures faites sur $n_1 = 17$ chats mâles et $n_2 = 15$ chattes.

mâles		femelles	
X	Y	X	Y
44	4.66	84	3.45
317	3.08	47	3.89
292	1.28	20	3.79
179	3.17	209	3.79
39	5.59	106	3.81
257	2.88	343	0.61
354	1.60	325	2.04
349	3.48	346	0.41
195	3.39	151	2.67
245	3.47	267	0.89
270	3.20	80	4.39
166	2.90	249	2.56
57	4.83	341	0.28
198	2.96	189	2.43
20	5.17	50	3.85
187	3.44		
270	3.18		

Table 2 – données chat

1. On définit le modèle H_4 comme celui où, pour chaque sexe, Y varie linéairement en fonction de X :
   — pour les mâles, $Y = a_1 X + b_1 + \varepsilon$
   — pour les femelles, $Y = a_2 X + b_2 + \varepsilon$.

   Pour chaque groupe, ajuster la droite de régression. Tester l'égalité des variances des résidus.
2. On définit le modèle H_2 comme celui où une droite de régression commune explique les mesures des deux sexes : $Y = aX + b + \varepsilon$. Tester H_2 contre H_4.
3. Comparer séparément le modèle retenu aux modèles suivants :
   — H_a : « $a_1 = a_2$ », b_1 et b_2 quelconques, i.e.,
   $$Y_i = a\ X + b_1 \mathbb{1}_{\{\text{mâle}\}} + b_2 \mathbb{1}_{\{\text{femelle}\}} + \varepsilon_i.$$
   — H_b : « $b_1 = b_2$ », a_1 et a_2 quelconques , i.e.,
   $$Y_i = a_1 \mathbb{1}_{\{\text{mâle}\}}\ X + a_2 \mathbb{1}_{\{\text{femelle}\}}\ X + b + \varepsilon_i.$$
4. Calculer $\mathbb{E}(\hat{a}_1)$, $\mathbb{E}(\hat{b}_1)$, $\mathrm{Var}(\hat{a}_1)$, $\mathrm{Var}(\hat{b}_1)$, $\mathrm{Cov}(\hat{a}_1, \hat{b}_1)$. Faire de même pour les estimateurs des paramètres dans le modèle “femelle”.
5. Donner des intervalles de confiance pour a_1, b_1, a_2, b_2. Les intervalles de confiance pour a_1 et a_2 se coupent-ils ? Ceux pour b_1 et b_2 ?