Bases statistiques pour la biologie

Formation doctorale

Cyril Dalmasso

20 - 22 juin 2022

Statistiques descriptives

Exercice 1

Pour étudier l'effet d'un somnifère, on mesure chez 20 patients le nombre d'heures de sommeil supplémentaires par rapport à la durée moyenne de leur nuit sans traitement. On obtient les résultats suivants:

```
extra <- c(-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0,1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
```

1. Faire un résumé numérique.
2. Tracer un histogramme des données.
3. Ces données sont en fait issues de deux groupes d'individus : apposer une variable indiquant le groupe associé à l'observation de la variable `extra` sachant que les 10 premiers individus sont issus du groupe 1 et les 10 suivants du groupe 2 (utiliser, par exemple, la commande `data.frame`). Faire un résumé statistique pour chaque groupe et tracer alors les boîtes à moustaches des observations selon les groupes. Qu'en pensez-vous ?

Probabilités

Exercice 2

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 multiplier 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 3

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 4

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) :

```
nbbact <- c(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 5

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) :

```
dosages <- c(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 6

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) :

```
tpsreact <- c(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 ?

1. (Facultatif) Ecrire une fonction 'my.t.test' prenant en entrée un vecteur d'observations et une valeur μ_0 et retournant la p-value du test. Utiliser la fonction pour tester l'action du produit.

2. Faire le test en utilisant la fonciton 't.test'

Exercice 7

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, 237 ont eu une réaction allergique. En utilisant l'approximation gaussienne, proposez un nouveau test statistique pour vérifier l'hypothèse des médecins et conclure.

Exercice 8

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.

Sur les 15 enfants observés, 5 présentent un retard. Que peut-on conclure (réaliser un test de niveau $\alpha = 5\%$) ?

Exercice 9

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 :

```
adjv <- c(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 10

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 :

```
##       Echantillon 1 Echantillon 2
##  [1,]          56.4          54.6
##  [2,]          57.5          58.2
##  [3,]          55.8          60.3
##  [4,]          54.3          59.5
##  [5,]          58.9          61.1
##  [6,]          56.9          58.7
##  [7,]          54.8          59.8
##  [8,]          54.2          57.5
##  [9,]          58.1            NA
```

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 11

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 Hollandaises
##  [1,]       552          487
##  [2,]       464          489
##  [3,]       423          470
##  [4,]       506          482
##  [5,]       497          494
##  [6,]       544          500
##  [7,]       486          504
##  [8,]       531          567
##  [9,]       496          482
## [10,]       501          526
```

Conclure au vu de ces données.

Exercice 12

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

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

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

Exercice 13

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 Juste après la\ntraite 24h après la traite
## [1,]     1                 12000                14000
## [2,]     2                 13000                20000
## [3,]     3                 21500                31000
## [4,]     4                 17000                28000
## [5,]     5                 15000                26000
## [6,]     6                 22000                30000
## [7,]     7                 11000                16000
## [8,]     8                 21000                29000
```

Exercice 14

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

```
##    O    A    B   AB
## 0.40 0.43 0.12 0.05
```

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 15

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:

```
##                              [,1] [,2] [,3] [,4] [,5] [,6]
## 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 16

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 :

```
##                        [,1]   [,2]     [,3]      [,4]
## 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 :

```
##                    [,1] [,2]     [,3]      [,4]
## 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 17

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:

```
##         fort moyen léger
## présent   24     6     5
## passé     30    30    10
## jamais    18    54    23
```

Au seuil $\alpha = 5\%$ peut-on conclure à l'indépendance des deux caractères ?

Exercice 18

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 :

```
##                                .         ..
## PoidsxCholestérolémie 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 19

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.

```
##   DirectionDeFuite SourisDeLAboratoire SourisSauvages
## 1                N                  17             26
## 2               NO                  25             17
## 3                O                  13              9
## 4               SO                  28              2
## 5                S                  19              3
## 6               SE                  20             16
## 7                E                  22             33
## 8               NE                  16             54
```

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

Exercice 20

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 de Hardy-Weinberg (c'est à dire que, pour un locus donné dont la fréquence de l'allèle A est p, alors : $P(AA) = p^2$, $P(Aa) = 2p(1-p)$ et $P(aa) = (1-p)^2$).

Exercice 21

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).

1. Proposer un test au niveau 1% permettant de savoir quelle décision prendre ;
2. Déterminer un intervalle de confiance à 99% pour m ;

Exercice 22

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

```
poids <- c(61, 82, 92, 97, 101, 104, 109, 118, 131, 155, 69, 82, 93, 97, 101, 104, 110, 120, 133, 145,
```

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 23

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

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 24

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 25

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 o'u 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 26

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 27

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`. 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.

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 : Tester l'égalité des espérances pour les trois traitements à l'aide d'une analyse de la variance (fonctions `lm` et `anova`).

Exercice 28

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 :

Tester l'interaction entre les facteurs, l'effet du facteur fertilisation et l'effet du facteur rotation. Enfin, tester l'intérêt du modèle.

Exercice 29

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 H_2

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

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 $E(Y) = 0$).

1. Représenter le nuages de points à l'aide de la fonction `plot`.
2. Donner une estimation des paramètres a_1 et a_2.
3. Tester H_2 contre H_0 : conclusion ?
4. On adopte maintenant le modèle H_1 $Y = a\ X_1 + b$. 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.