Inférence statistique / Séance 1

Concept d'inférence statistique

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy *

17 mars 2025

Annotations utilisées dans ces feuilles d'exercices

- Lorsque le numéro d'une question est indiqué entre parenthèses [**(1).**], cela signifie que la question peut être sautée sans nuire à la continuité de l'exercice.
- Une question affectée d'une fleur en exposant de son numéro [**1.**$^{❀}$] est une simple question intermédiaire, posée uniquement pour détailler un raisonnement : elle est donc censée être résolue sans difficulté particulière.
- Une question avec une étoile en exposant [**1.**$^{★}$] est une question un peu plus difficile que les autres.
- Une question avec une double étoile [**(1).**$^{★★}$] est une question de difficulté hors-programme, mentionnée uniquement par intérêt culturel.

EXERCICE 1 — Vocabulaire

Dans tout cet exercice, on considère le modèle du pédagogue (dans le cadre explicatif ou prédictif selon le contexte).

Première partie : Qui est qui ?

Pour les questions 1 à 10 de cet exercice, la consigne est toujours la même : pour la quantité décrite par l'énoncé :

- *Dire si elle dépend du paramètre caché et/ou de l'observation (passée) et/ou de l'observation future ; et dans le cas où elle ne dépend d'aucun des trois, dire si sa valeur dépend des paramètres intervenant dans la description le modèle ;*
- *En déduire s'il s'agit d'une statistique, d'une quantité d'intérêt explicative, d'une quantité d'intérêt prédictive, d'un paramètre du modèle, ou d'autre chose.*

1.$^{★}$ L'écart-type (sur le long terme) de la distribution des notes avec l'ancienne méthode. (Cet écart-type est appelé $\sigma_{\text{réf}}$ dans le polycopié).

2. Le nombre total d'élèves dans les deux premières promotions à expérimenter la nouvelle méthode.

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

3. Le score moyen obtenu par la première promotion à expérimenter la nouvelle méthode.

4. La “surperformance” de la première promotion à expérimenter la nouvelle méthode : j’entends par là, la différence entre la moyenne *réelle* de cette promotion et l’espérance de cette moyenne connaissant la distribution des résultats de la nouvelle méthode (laquelle espérance étant ce que nous avons appelé $\mu_{\checkmark}$ dans le polycopié).

5. L’identité de l’élève ayant eu le meilleur résultat de la première promotion.

6. La pire note obtenue au sein de la seconde promotion à expérimenter la nouvelle méthode.

7. Le fait que la seconde promotion ait eu, ou pas, un résultat moyen supérieur à celui de la première promotion.

8.$^{\bigstar}$ L’espérance de la variable aléatoire « écart-type (empirique) des résultats de la seconde promotion[*] », du point de vue de quelqu’un connaissant la distribution générale des résultats avec la nouvelle méthode.

9. La proportion d’élèves, sur le long terme, atteignant le score $\mu_{\text{réf}}$ (ou mieux) avec la nouvelle méthode.

10. Le nombre d’or : $\varphi \coloneqq \left(1+\sqrt{5}\right)/2$. On répondra plus spécifiquement à la question suivante : *peut-on* le considérer resp. une statistique, une quantité d’intérêt explicative, une quantité d’intérêt prédictive ?...

Seconde partie : La foire à tout

Les trois questions de cette partie sont toutes bâties sur le même modèle, et se composent chacune de quatre sous-questions similaires : au final, cette partie comporte donc douze items du même modèle. Vous n’êtes pas obligés de traiter la totalité de ces items : votre caïman(e) pourra en choisir certains arbitrairement pour les traiter en classe. Rappelons qu’on se place toujours dans le modèle du pédagogue.

11. Donnez des exemples pertinents (et, autant que possible, variés) de statistiques qui soient respectivement :

(i) À valeurs réelles ;

(ii) À valeurs booléennes ;

(iii) À valeurs « intervalle de $\mathbb{R}$ » ;

(iv) À valeurs « distribution de probabilité sur $\mathbb{R}$ ».

[*]. L’écart-type empirique des résultats de la seconde promotion est un indicateur de dispersion des résultats de la seconde promotion qu’on peut, par exemple, définir ainsi : c’est la racine carré de la différence entre la moyenne des carrés des résultats de la seconde promotion et le carré de la moyenne des résultats. Par exemple, si $m = 4$ et que les résultats de la seconde promotion sont $(112, 96, 132, 100)$, alors la moyenne de ces résultats vaut $(112+96+132+100)/4 = 110$ et la moyenne de leurs carrés vaut $(112^2+96^2+132^2+100^2)/4 = 12\,296$, de sorte que l’écart-type empirique vaut $\sqrt{12\,296-110^2} = 14$.

12. Même question avec des quantités d'intérêt explicatives : donner des exemples pertinents et variés de quantités d'intérêt explicatives dont les réalisations correspondrons respectivement à un nombre réel, un booléen, un intervalle, et une distribution de probabilité sur $\mathbb{R}$.

13. Même question avec des quantités d'intérêt prédictives.

Troisième partie : Traitement de données

Une statisticienne regarde la liste des notes obtenues par les élèves de la première promotion "cobaye", et procède aux calculs suivants :

(1°) On calcule la moyenne du jeu de données, qu'on stocke dans une variable `m`.

(2°) On retire les deux meilleures notes et les deux pires notes du jeu de données.

*(3°) On calcule ce qu'on appelle l'*écart-type empirique *du jeu de données ainsi tronqué : il s'agit de la racine carrée de la différence entre la moyenne des carrés des données (du jeu tronqué) et le carré de leur moyenne. (Cette différence étant ici positive, on peut bien prendre sa racine carrée). On le stocke dans une variable* `s`.

(4°) On définit la variable `muref` *comme valant* 100.

(5°) Enfin, on calcule `(m - muref) / s`.

La valeur trouvée à la fin de ce calcul par notre statisticienne est +0,370 5.

14. Dans le jargon de la statistique, à quoi vient de procéder notre statisticienne ? A-t-elle calculé un paramètre caché, une quantité d'intérêt (explicative ? prédictive ?), une statistique, ... ?

EXERCICE 2 — Methodenbeschreibung für die Schätzung von epidemiologischen Parametern des COVID19 Ausbruchs

L'intitulé de cet exercice, signifiant « Description de la méthode d'estimation des paramètres épidémiologiques de propagation de la Covid-19 », est le titre d'un article scientifique publié en avril 2020, durant la première vague épidémique de Covid-19 ayant touché l'Autriche. Face à une telle épidémie en effet, il était essentiel de comprendre la dynamique de propagation de la maladie, afin d'éclairer au mieux les pouvoirs publics[†] *sur les mesures les plus adéquates à prendre.*

Je cite ci-dessous (une traduction de) les extraits essentiels de cet article, avec les notations d'origine :

> *Soit y_t le nombre de nouveaux cas au jour t. On suppose que y_t suit une loi de Poisson de paramètre λ_t. Soit R_{eff} le taux de reproduction effectif de la maladie, i.e. le nombre moyen de contaminations directes générées par un infecté. En l'absence d'information supplémentaire sur R_{eff}, on suppose que ce taux provient*

[†]. Notez qu'il y a une pertinence à ce que chaque pays procède à sa propre analyse relativement à son sol, car les paramètres épidémiologiques sont susceptibles de dépendre des habitudes de vie d'un pays, et le mode de collecte des données également. En l'occurrence, l'étude scientifique considérée était destinée à éclairer les décideurs autrichiens, d'où le fait que l'article fût écrit en allemand.

d'une loi à priori de forme gamma, avec pour paramètres de forme $a = 1$ et $b = 5$, dont la densité est

$$p(R) = \mathbf{1}_{R>0} \frac{R^{a-1} \mathrm{e}^{-R/b}}{b^a \Gamma(a)}.$$

Nous supposons maintenant que, pour t dans la plage de temps $\{1, \dots, \tau\}$ étudiée (avec $\tau := 13$), les infectés d'un jour $(t-s)$ contribuent au nombre d'infectés du jour t avec un taux $R_{\text{eff}} \times w_s$:

$$\lambda_t = R_{\text{eff}} \sum_{s=1}^{t} y_{t-s} w_s.$$

Ici w_s correspond à la probabilité, lorsqu'une personne en contamine une autre, que l'infection "fille" ait lieu s jours après l'infection "mère". Nous faisons l'hypothèse que les w_s sont la fonction de masse d'une certaine loi gamma discrétisée, à savoir

$$w_s = \frac{p(s; k, \theta)}{\sum_{s=1}^{\infty} p(s; k, \theta)},$$

où

$$p(s; k, \theta) := \frac{s^{k-1} \mathrm{e}^{-s/\theta}}{\theta^k \Gamma(k)}$$

($\Gamma(\bullet)$ désignant ici la fonction gamma d'Euler), pour $k = 2{,}88$, $\theta = 1{,}55$ d [‡] : ces paramètres ayant été déterminés d'après d'autres études sur la maladie.

1.★ Récapituler le modèle décrit par l'article sous forme d'un schéma, en montrant qui dépend de quoi, et selon quelle loi.

☛ *Attention, vérifiez le caractère correct de votre réponse à la première question avant de passer à la suite de l'exercice, sinon vous risquez de bâtir tout votre édifice de réflexion de travers… !*

2. Quelle est l'observation (passée) ?

3. Au vu du titre et de la tournure de l'article, diriez-vous qu'on va considérer un modèle de nature explicative et prédictive ? Dans le second cas, dire quelle est l'observation future.

4. Quel est le statut de R_{eff} dans le modèle statistique que nous allons considérer : est-ce le paramètre caché, l'observation, l'observation future, une quantité d'intérêt (explicative ? prédicitive ?), une statistique... ?

Dans la suite de l'article, les auteurs expliquent qu'un de leurs buts principaux est de se faire une idée du taux de doublement *de l'épidémie : ce « taux de doublement » de l'épidémie correspondant, comme son nom l'indique, à l'intervalle de temps sur laquel le nombre de cas double à supposer qu'on soit dans un régime où le nombre de contaminations est suffisamment*

[‡]. 'd' est le symbole du jour (du latin *dies*), vu comme unité de temps.

grand pour lisser les effets statistiques. Plus précisément, si on pose r *comme étant l'unique solution dans* $\mathbb{R}_+^*$ *de l'équation*

$$\sum_{s=1}^{\infty}(\mathrm{e}^{-sr} \times R_{\mathrm{eff}} \times w_r) = 1,$$

le taux de doublement est défini comme

$$T \coloneqq \ln 2 \,/\, r.$$

5. De quel type de quantité relève T : est-ce le paramètre caché, l'observation, l'observation future, une quantité d'intérêt (explicative ? prédicitive ?), une statistique... ?

Dans le modèle introduit par les chercheurs autrichiens, il y a des quantités λ_t *qui ne sont pas directement observables : on se dit donc qu'elles vont devoir apparaitre dans le modèle comme fonctions du paramètre caché et/ou de l'observation future. Pourtant, ce n'est pas ainsi qu'on formaliserait correctement le modèle statistique associé à cet article (du moins, si on suit les conventions du cours) : en fait, dans ce modèle statistique, les* λ_t *n'interviendront pas* du tout*... !*

6.★ Expliquer comment on peut se dispenser complètement de λ_t pour écrire le modèle statistique, tout en étant bien capable d'écrire un loi de l'observation (éventuellement complétée) sachant le paramètre caché.

7. Pourquoi sommes-nous ici dans un cadre bayésien ? Quelle est la priore ?

8. L'énoncé introduit quantre quantités littérales auxquelles il donne une valeur connue : a, b, τ, k et θ. Parmi ces quantités, lesquels sont des paramètres du modèle, et lesquelles n'en sont pas ? (Remarque : « lesquelles » ne présume en rien qu'il y en ait au moins deux : c'est un pluriel générique qui inclurait aussi les cas de 0 ou 1 quantité !).

(9).★ Identifier d'autres paramètres du modèle. (Il y en a !).

EXERCICE 3 — Le modèle du classement Elo

Imaginons que nous souhaitions comparer la force d'un grand nombre J *de joueurs d'échecs (penser à* $J = 5\,776$*), par exemple lors d'un tournoi. Le plus simple serait évidemment de demander à chaque joueur de disputer une partie contre chaque autre, mais ce serait logistiquement impossible... On pourrait également décider d'organiser, par exemple,* 7 *tours de jeu en appariant à chaque fois des joueurs au hasard, et considérer qu'un joueur est d'autant plus fort qu'il a remporté un grand nombre de ses parties ; mais le résultat serait obtenu par l'aléa d'être tombé sur des adversaires plus ou moins forts... Une idée un peu plus astucieuse est de ne faire s'affronter, à un moment donné, que des joueurs ayant eu le même profil de résultats jusque-là (par exemple, si vous avez perdu au premier tour, au second tour vous affronterez nécessairement un(e) adversaire ayant également perdu son premier tour) ; mais dans quel sens comparer, disons, un joueur qui a perdu son premier match avant de gagner tous les suivants (il est donc essentiellement le plus fort* parmi ceux ayant perdu leur

premier match*) et une joueuse ayant gagné ses deux premiers matchs puis perdu les cinq suivants (elle est donc essentiellement la moins forte* parmi ceux ayant gagné leurs deux premiers matchs*) ? Sans compter que résultats d'une partie ne permettent pas de comparer la force des adversaires avec certitude : parfois, le plus faible l'emporte sur une partie…*

Pour résoudre ce problème, l'Américain d'origine hongroise Arpad Elo *proposa en 1960 un système encore utilisé de nos jours, s'appuyant sur des considérations statistiques. Le modèle que nous allons présenter ci-dessous n'est pas exactement celui d'Elo, mais il en est très proche dans l'esprit.*

Ici nous considérons pour simplifier que les résultats d'une partie se terminent nécessairement par la victoire d'un des deux adversaires (pas de partie nulle), et qu'il n'y a pas d'avantage à jouer avec les pièces blanches plutôt qu'avec les noires[§]*. Par ailleurs (et c'est un point essentiel des systèmes Elo), il n'y a pas de notion de victoire "large" ou "serrée" : tout ce qui compte dans le résultat d'une partie, c'est* qui *l'a gagnée, et pas* comment *!*

Le modèle d'Elo repose sur l'hypothèse suivante : chaque joueur j possède un niveau intrinsèque $E_j \in \mathbb{R}$, constant au cours du tournoi ; et lorsqu'un joueur A *affronte un joueur* B*, la probabilité que ce soit* A *qui gagne vaut* $10^{E_A} / (10^{E_A} + 10^{E_B})$.

Pour simplifier, plaçons-nous dans une situation où les organisateurs du tournoi ont décidé à l'avance *qui allait affronter qui au cours des différentes rondes, via un tirage au sort. En supposant que J est pair et qu'il y a R rondes dans le tournoi, pour tout $r \in [\![0, R[\![$, pour tout $t \in [\![0, J/2[\![$, on désigne par $b(r,t)$ et $n(r,t)$ les joueurs ayant respectivement les pièces blanches et les pièces noires à la table numéro t lors de la ronde numéro r. Le résultat de la partie correspondante, quant à lui, est désignés par $V(r,t) \in \{0,1\}$, où* 0 *signifie que c'est le joueur ayant les Noirs qui a gagné, et* 1 *que c'est celui ayant les Blancs qui a gagné.*

1. Donner un exemple de quantité d'intérêt (explicative) à laquelle nous pourrions nous intéresser dans le modèle.

(2).❀ Expliquer pourquoi J et R constituent des paramètres du modèle.

(3).★ Bien que les $b(r,t)$ et les $n(r,t)$ aient été tirés au sort par les organisateurs, nous allons choisir de les voir comme des *paramètres* de notre modèle. Pourquoi cela, à votre avis ?

4. Quel est le paramètre caché de notre modèle ? Quelle est l'observation ? Quelle est la loi[¶] de l'observation sachant le paramètre caché ?

5. Expliquer comment on pourrait transformer ce modèle en modèle prédictif, et ce que pourrait être une quantité d'intérêt prédictive dans ce contexte.

[§]. Pour les amateurs d'échecs : considérez que c'est un tournoi où toutes les parties se jouent en mode Armageddon ! ◡̈

[¶]. Remarque importante : Dans ce cours, sauf mention explicite du contraire, quand on dit « quelle est la loi », on ne demande pas nécessairement de reconnaitre une loi classique possédant un nom, ni d'écrire une formule à partir de telles lois classiques : on demande simplement une description parfaitement univoque et précise de cette loi, peu importe qu'elle s'exprime *in fine* par une formule ou non. Par exemple, « la loi qu'on obtient quand on fait la somme d'une v.a. X de loi Normale$(0,1)$ et d'une v.a. Y de loi $\text{Unif}^{\text{me}}(-1,1)$ qui soit indépendante de X » serait une réponse tout à fait valable : on pourrait aussi l'écrire, en termes plus formels, comme « la mesure-image de la mesure-produit Normale$(0,1) \otimes \text{Unif}^{\text{me}}(-1,1)$ par l'application "somme" de $\mathbb{R} \times \mathbb{R}$ dans $\mathbb{R}$ » ; cependant cette transcription serait assez délicate et relèverait purement du formalisme mathématique technique des probabilités, ce qui n'est pas l'objet sur lequel ce cours se focalise.

6. Dans quel cas pourrait-on avoir une priore intéressante sur le paramètre caché ? À quoi ressemblerait cette priore le cas échéant ? Quel serait l'intérêt d'une approche bayésienne dans un tel contexte ?

EXERCICE 4 — Monty Hall devient statisticien

Le problème dit « de Monty Hall » est un immense classique des probabilités, que la plupart d'entre vous avez certainement déjà rencontré. J'en rappelle le contexte :

Définition (Jeu de Monty Hall, version standard)**.** Un présentateur de jeu vous montre trois portes. Derrière l'une de ces portes (mais vous ignorez laquelle) il a caché une récompense ; et derrière les deux autres, il n'y a rien. Le jeu fonctionne ainsi. Dans un premier temps, vous désignez une des trois portes. Dans un second temps, le présentateur ouvre une porte perdante parmi les deux que vous n'avez pas choisies [‖]. Dans un troisième temps, il vous demande si vous souhaitez changer votre choix initial. C'est en fonction de votre choix final de porte que vous gagnerez, ou pas, la récompense.

Le « problème de Monty Hall » consiste alors à savoir si, à la troisième étape, il est dans notre intérêt ou pas (ou neutre) de changer notre choix initial. Comme ceux d'entre vous qui ont déjà vu le problème s'en souviennent probablement : la solution correcte, mais paradoxale, est qu'il est clairement plus intéressant de changer de porte que de rester sur notre choix initial !

(1).❀ Montrer que rester sur son choix initial conduit à gagner une fois sur trois, tandis que changer de porte conduit à gagner deux fois sur trois.

Ici notre but n'est pas tant de résoudre le problème de Monty Hall, que de voir en quoi est-ce qu'on peut le traiter comme une question de statistique.

(2).★ Justifier qu'on peut considérer que la porte que vous choisissez initialement est déterministe (disons que c'est la porte A), permettant ainsi une simplification en termes de modélisation.

3.★ Formaliser le jeu de Monty Hall comme un modèle de statistique.

Indication : Je vous suggère de réfléchir dans l'ordre suivant :
- À quel moment du jeu se place-t-on pour notre modélisation ?
- Quelle est notre quantité d'intérêt ? Dans quel espace vit-elle ?
- Quelle est l'observation (passée) ? Dans quel espace vit-elle ?
- Quel est le paramètre caché (dans quel espace vit-il), et quelle est la loi de l'observation (passée) sachant le paramètre caché ?
- Y a-t-il une observation future, et si oui, laquelle ? (et dans quel espace vit-elle ?). Le cas échéant, quelle est la loi de l'observation *globale* sachant le paramètre caché ?

Indication : Assurez-vous d'avoir répondu correctement à cette question avant de passer aux suivantes, en demandant à votre caïman(e) (ou en regardant le corrigé si vous êtes chez vous).

(4).★ Argüer qu'il est préférable de voir le modèle de Monty Hall comme un modèle *explicatif* plutôt que comme un modèle *prédictif*.

[‖]. Notez en effet qu'il y a forcément au moins une porte perdante parmi elles.

5.❀ Argüer qu'on dispose clairement (j'entends par là, de façon non ambigüe) d'une loi de probabilité à priori sur le paramètre caché en l'occurrence, et dire quelle est cette loi.

(6). Plaçons-nous, pour fixer les idées, dans le cas où la réalisation de l'observation est « porte B ». Calculer alors la loi du paramètre caché *conditionnellement à cette observation.* (Nous verrons dans la prochaine séance que cette loi conditionnelle est appelée *postériore*). En conclure qu'on a effectivement intérêt à changer de porte.

En fait, dans le problème de Monty Hall tel qu'il est énoncé habituellement, il n'est pas clair que le présentateur soit dans l'obligation de nous proposer de changer de porte... Du coup, on pourrait par exemple considérer la variante suivante :

Définition (Jeu de Monty Hall, version mesquine)**.** Dans cette version du jeu de Monty Hall, la deuxième étape se passe comme précédemment ; néanmoins, le présentateur n'est *pas obligé* de nous proposer de changer de porte : cette fois-ci, il tire au sort la décision de nous proposer ou pas un changement de porte ; et ce, de la façon suivante :

— Lorsque le choix initial de porte était perdant, on ne propose un changement qu'une fois sur cinq (au sens de « avec une probabilité de 1/5 ») ;
— En revanche, lorsque le choix initial était gagnant, on propose un changement quatre fois sur cinq.

7. Qu'est-ce que cela change en termes de modélisation ? (y compris concernant la priore). (Il est attendu une réponse suffisamment précise pour que le nouveau modèle soit décrit sans ambigüité).

(8). Calculer la postériore dans le cadre de ce nouveau modèle ; et en déduire que cette fois-ci, il ne faut surtout pas changer de porte ! À votre avis, quelle leçon est-ce que le compositeur de cet exercice souhaite que vous en tiriez ?

On va généraliser encore le problème :

Définition (Jeu de Monty Hall, version statistique)**.** Dans cette version, de même que dans la version mesquine, le présentateur est libre de décider s'il nous propose de changer de porte ou pas ; et, à nouveau comme dans la version mesquine, en pratique, il tire au sort sa décision de nous proposer de changer de porte, et ce, potentiellement avec des probabilités différentes selon que notre choix initial était gagnant ou perdant. Par contre, les probabilités de proposer un changement dans le cas d'un choix initial gagnant, resp. d'un choix initial perdant, ne sont cette fois-ci pas connues ! Ces probabilités sont néanmoins supposées fixes au cours de l'histoire du jeu (et donc entre les candidats précédents et vous) : par conséquent, l'histoire du jeu est susceptible de nous aider à prendre notre décision.

9.★ Décrire le nouveau modèle correspondant.

10. Argüer que, cette fois-ci, il n'y a plus de choix évident pour la priore, et que par conséquent nous aurons besoin de notions de statistique fréquentiste pour prendre notre décision...

Inférence statistique / Séance 2

Théorème de Bayes

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy*

24 mars 2025

EXERCICE 1 — Le golfeur

Chaque jour, un golfeur va s'entrainer sur le practice. Il essaie d'envoyer la balle dans une cible tracée au sol, jusqu'à y arriver, et arrête son entrainement une fois cet objectif atteint. Pour chacun des $7 =: J$ *derniers jours, notre golfeur a noté quotidiennement le nombre d'essais infructueux dont il a eu besoin avant d'atteindre son objectif : respectivement 49, 191, 97, 61, 9, 145 et 167. Le golfeur suppose qu'à chaque tentative, de façon indépendante, il a une certaine probabilité* $\pi_{✓}$*, toujours la même, de réussir son coup. Il se demande quelle information est-ce que les résultats qu'il a notés peuvent lui donner concernant la valeur de cette probabilité, et ce qu'il peut en déduire sur le nombre d'essais infructueux dont il aura besoin lors de son prochain entrainement.*

1. Modéliser ce problème avec le vocabulaire de la statistique : quels sont les éventuels paramètres du modèle ; que représente l'observation (passée), dans quel espace vit-elle et quelle est sa valeur effective ; y a-t-il une observation future, que représente-t-elle le cas échéant, et dans quel espace vit-elle ; quel est le paramètre caché et l'espace dans lequel il vit ; quelle est la loi de l'observation (complétée, s'il y a lieu) sachant la valeur du paramètre caché ? À mesure que vous répondrez à ces questions, proposez des notations appropriées pour traiter le problème. (Dans les questions suivantes, nous supposerons qu'on a pris les choix de notations les plus standards dans le contexte).

(2).❀ Pour $\pi \in]0,1[$, $k \in \mathbb{N}$, calculer la probabilité qu'une loi géométrique de paramètre π (qui décrit le nombre d'échecs [*] avant le premier succès d'une série d'expériences ayant individuellement une probabilité de succès π) vaille k. Attention, ici on demande de *démontrer* le résultat et pas de ressortir une formule toute faite !

3. Calculer la fonction de vraisemblance du paramètre caché pour notre observation effective. (Pour ceux préférant manipuler des valeurs littérales, on pourra introduire la notation $s_{✓} := \sum_{j=0}^{J-1} x_{j✓} \underset{\text{déf}}{=} 719$).

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

[*]. Attention, ce n'est pas la même convention que celle que vous avez vue en classes préparatoires, où on considérait que la loi géométrique compte le nombre total d'*essais*, incluant le succès terminal... !

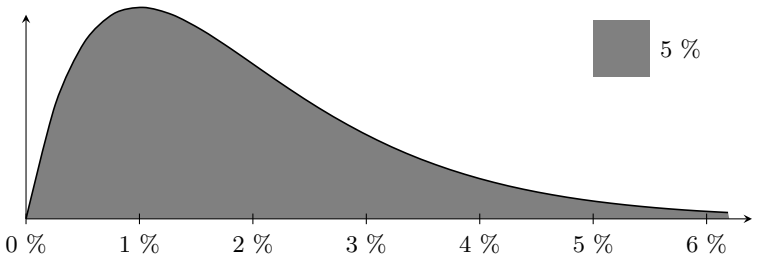


FIGURE 1 – Tracé de la loi Bêta(2, 100).

(4).❀ Déterminer grossièrement le comportement de la fonction de vraisemblance calculée à la question précédente [†], et observer que les valeurs de $\boldsymbol{\pi}$ pour lesquelles cette fonction prend les plus grandes valeurs semblent effectivement être plus « vraisemblables » que les autres, au moins au sens intuitif.

Si on avait demandé au golfeur, avant *l'expérience, de parier sur la valeur de sa probabilité de succès, il aurait considéré que les probabilités (de son point de vue) que sa probabilité de succès vaille tant ou tant étaient les mêmes que celles d'une loi* Bêta$(2, \beta)$ *avec* $\beta \coloneqq 100$*, où* Bêta$(2, \beta)$ *est une certaine loi de probabilité sur* $]0, 1[$ *dont vous trouverez un tracé de la densité en figure 1. Cette réponse revient essentiellement à dire que, de base, le golfeur estime que sa probabilité de succès à chaque coup est probablement entre* 0,5 % *et* 4 %*, mais qu'il ne serait pas complètement étonné pour autant que ce soit un peu moins ou un peu plus.*

5. Traduire ce que je viens d'écrire dans le vocabulaire de la statistique inférentielle.

Pour la suite de l'exercice, on donne la formule suivante : pour $t \in]0, 1[$*,* $\mathrm{d}t$ *un voisinage infinitésimal de* t*,*

$$\mathbb{P}\big(\text{Bêta}(2, \beta) \in \mathrm{d}t\big) = \beta(\beta + 1)t(1 - t)^{\beta - 1} \operatorname{vol}_1(\mathrm{d}t).$$

6. Calculer que

$$\text{Loi}(\boldsymbol{\pi} \mid X = x_{\checkmark}) = \text{Bêta}(9, 819),$$

où la loi Bêta$(9, 819)$ est la distribution de probabilité sur $]0, 1[$ définie par

$$\mathbb{P}\big(\text{Bêta}(9, 819) \in \mathrm{d}t\big) \coloneqq 4\,295\,945\,446\,555\,453\,215\,525 \times t^8(1 - t)^{818} \operatorname{vol}_1(\mathrm{d}t).$$

(Nous admettrons que cette mesure est bien une mesure de probabilité).

Une représentation de la loi Bêta$(9, 819)$ *est donnée en figure 2.*

(7). En déduire, visuellement [‡], un intervalle dans lequel, au vu de ses résultats de la semaine, le golfeur peut maintenant affirmer que sa probabilité de succès à chaque coup a environ 80 % de chances de se situer.

[†]. On pourra, notamment, établir le tableau de variations de cete fonction ; et comparer les valeurs numériques qu'elle prend en quelques points judicieusement choisis.

[‡]. Dès la prochaine séance, nous apprendrons à interpréter ce genre de résultats plus rigoureusement, à l'aide de "fonctions de tables".

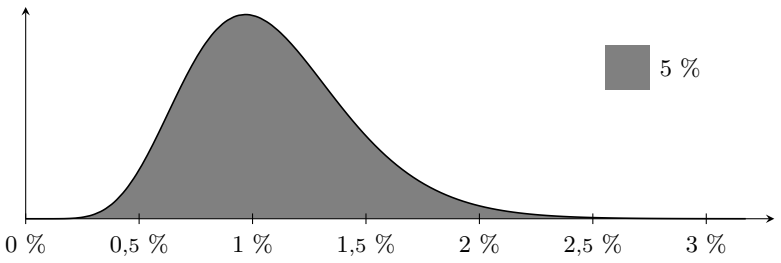


FIGURE 2 – Tracé de la loi Bêta(9, 819).

Maintenant, notre golfeur se demande quelle est la probabilité que, lors de son prochain entrainement, n'ait toujours pas atteint la cible au bout de k essais ?

(8). Expliquer pourquoi k n'est pas un paramètre du modèle !

9. Montrer que, sous le véritable contexte probabiliste (autrement dit, du point de vue de quelqu'un connaissant la vraie valeur $\pi_{\checkmark}$), on a :

$$\mathbb{P}_{\checkmark}(Y \geqslant k \mid X = x_{\checkmark}) = (1 - \pi_{\checkmark})^k.$$

Indication : Attention, la question exige de prendre en compte le conditionnement... !

10.★ En déduire que

$$\mathbb{P}_{\text{post}}(Y \geqslant k) = \mathbb{E}\big(\big(1 - \text{Bêta}(9, 819)\big)^k\big).$$

(11).❀ En s'appuyant sur les résultats de l'indication, calculer numériquement la probabilité à postériori que le golfeur réussisse son prochain entrainement en k essais ou moins, pour resp. $k = 3, k = 50, k = 360$.

Indication : Pour $\alpha, \beta \in \mathbb{R}_+^*, k \in \mathbb{N}$, on donne :

$$1 - \text{Bêta}(\alpha, \beta) = \text{Bêta}(\beta, \alpha)\,;$$

$$\mathbb{E}\big(\text{Bêta}(\alpha, \beta)^k\big) = \prod_{i=0}^{k-1} \frac{\alpha + i}{\alpha + \beta + i}.$$

(12).★ À la question précédente, on trouve des probabilités respectives de 3,22 %, 41,19 % et 96,18 %. Observer que ces valeurs ne correspondent clairement pas à celles qu'on pourrait obtenir avec une loi géométrique... Comment l'expliquez-vous ?

EXERCICE 2 — Tubes électroniques

Un tube électronique est une pièce sans usure, mais susceptible de subir une défaillance à tout moment : sa durée de vie suit donc une loi exponentielle. On cherche ici à connaitre le paramètre $\lambda_{\checkmark}$ de cette loi exponentielle pour un certain type de tubes produits par un fabricant. Pour ce faire, on prend un échantillon de $10 =: n$ tubes de ce type dont on mesure des durées de fonctionnement, en arrêtant l'expérience au bout de 96 h $=: t_{\text{stop}}$ s'il reste encore des tubes en fonctionnement à ce moment-là. Les durées mesurées, en heures, sont

$$(60, 31, 34, 40, t_{\text{stop}}, t_{\text{stop}}, 26, 72, 49, 85) =: (t_{0\checkmark}, \dots, t_{9\checkmark}) =: \vec{t}_{\checkmark},$$

où, quand le temps mesuré est « t_{stop} », cela signifie en fait que le tube était encore en fonctionnement après 96 heures.

1. Identifier dans ce modèle :
— Les paramètres du modèle (il y en a deux) ;
— Le paramètre caché ;
— L'observation (passée).

Pour $\lambda \in \mathbb{R}_+^*$, rappelons que, par définition, la loi exponentielle de paramètre λ décrit la durée de vie d'un individu qui, sur chaque intervalle de temps infinitésimal $\mathrm{d}t$, indépendamment du passé, a (à supposer qu'il soit encore en vie) une probabilité $\lambda \operatorname{vol}_1(\mathrm{d}t)$ de mourir.

(2). Pour $\lambda \in \mathbb{R}_+^*$, à partir de la définition rappelée ci-dessus, calculer $\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \geqslant t)$ pour $t \in \mathbb{R}_+^*$, puis en déduire que, pour $\mathrm{d}t$ un voisinage infinitésimal de t, on a :

$$\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \in \mathrm{d}t) = \lambda \mathrm{e}^{-\lambda t} \operatorname{vol}_1(\mathrm{d}t).$$

Indication : Pour la première partie de la question, on pourra commencer par découper l'intervalle de temps $[0, t[$ en un grand nombre de pas infinitésimaux de même taille. Pour sa seconde partie, on pourra considérer le cas d'un pas de temps infinitésimal de la forme $[t, t + \delta t[$, où δt représente une durée infinitésimalement petite fixé arbitrairement.

3. Décrire, pour une valeur donnée quelconque λ du paramètre caché, la loi (sous $\mathbb{P}_\lambda$) de la variable aléatoire T_i [§] à valeurs dans $]0, t_{\text{stop}}]$ dont $t_{i\checkmark}$ est une réalisation.

Indication : La loi recherchée ayant une partie à densité (sur $]0, t_{\text{stop}}[$) et une partie discrète (en t_{stop}), on l'exprimera en donnant respectivement la valeur de $\mathbb{P}_\lambda(T_i \in \mathrm{d}t)$ pour $t \in]0, t_{\text{stop}}[$ et la valeur de $\mathbb{P}_\lambda(T_i = t_{\text{stop}})$.

4.* Dans cette question, on considère que l'observation concerne seulement la variable aléatoire T_i. Calculer la fonction de vraisemblance dans ce cas, si on suppose que la valeur effectivement observée de T_i est t. (On distinguera les cas $t \in]0, t_{\text{stop}}[$ et $t = t_{\text{stop}}$).

Indication : Pour cette question, exceptionnellement, il y aura besoin de revenir à la définition fondamentale de la vraisemblance...

5. En déduire, pour l'expérience globale, la fonction de vraisemblance

$$\lambda \mapsto \mathscr{L}(\boldsymbol{\lambda} = \lambda \mid \vec{T} = \vec{t}_{\checkmark}).$$

On se propose maintenant d'utiliser une méthode bayésienne pour inférer $\boldsymbol{\lambda}$, en utilisant la priore suivante (sur $\mathbb{R}_+^*$) :

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\lambda} \in \mathrm{d}\lambda) \propto \lambda^{-1} \operatorname{vol}_1(\mathrm{d}\lambda).$$

6. S'agit-il d'une priore propre ou impropre ? (Justifiez votre réponse).

(7). Pour $\lambda_1, \lambda_2 \in \mathbb{R}_+^*$ deux valeurs quelconques, et $r > 1$ fixé, comparer $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\lambda} \in [\lambda_1, r\lambda_1])$ et $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\lambda} \in [\lambda_2, r\lambda_2])$. En déduire que la priore choisie "ne privilégie aucun ordre de grandeur pour $\boldsymbol{\lambda}$ par rapport à un autre".

[§]. Cette loi étant bien entendu la même pour tout i.

8. Calculer la distribution à postériori de $\boldsymbol{\lambda}$, qu'on reconnaitra comme une loi gamma (voir l'indication) dont on donnera les paramètres.

Indication : Pour $k > 0$ et $\beta > 0$, la loi Gamma(k, β), où k est appelé « paramètre de forme » et β « paramètre de taux », est la loi à densité sur $\mathbb{R}_+^*$ donnée par

$$\mathbb{P}\big(\mathrm{Gamma}(k,\beta) \in \mathrm{d}x\big) = \frac{\beta^k}{\Gamma(k)} x^{k-1} \exp(-\beta x)\, \mathrm{vol}_1(\mathrm{d}x),$$

où le facteur $\Gamma(k)$ qui intervient correspond à ce qu'on appelle la « fonction gamma d'Euler » (qui est implémentée dans tous les bons logiciels de calcul numérique ☺).

En pratique, on ne s'intéresse pas au paramètre $\lambda_{✓}$ en lui-même, mais à ce qu'on appelle la « demi-vie » des tubes électroniques, notée $\tau_{½✓}$, liée à $\lambda_{✓}$ par la relation

$$\tau_{½✓} \coloneqq \frac{\ln 2}{\lambda_{✓}}.$$

9. À l'aide d'un changement de variable, déterminer la densité de la loi à postériori de la quantité d'intérêt $\boldsymbol{\tau}_{½}$.

Les lois gamma étant des distributions de probabilité très classiques, de nombreux logiciels disposent de fonction pré-implémentées permettant de calculer leur fonction de répartition. Par exemple, dans le langage **R** *(très utilisé en analyse de données), l'évaluation en* `x` *de la loi gamma de paramètre de forme* `k` *et de paramètre de taux* `b` *s'obtient en exécutant la commande «* `pgamma(x, k, b)` *».*

10. L'entreprise fabriquant les tubes aimerait savoir si sa méthode de fabrication lui permet d'atteindre le standard imposé par une certaine norme, qui demande à ce qu'on ait $\tau_{½✓} \geqslant 36\ \mathrm{h} =: \tau_{\mathrm{ref}}$. Quelle ligne de calcul faudrait-il écrire en langage **R** pour que le fabricant obtienne la probabilité, au vu des données dont il dispose, que le dispositif actuel de fabrication des tubes satisfasse la norme ?

(11). Le calcul évoqué à la question précédente a donné une valeur de 0,877 477 1. Conclure qualitativement sur ce qu'il convient d'en penser du point de vue industriel.

Imaginons maintenant que la personne ayant fait les tests n'ait pas pensé à noter quel tube défaille à quel moment, mais ait juste relevé le multi-ensemble[¶] *des durées observées, soit*

$$\{\{26, 31, 34, 40, 49, 60, 72, 85, t_{\mathrm{stop}}, t_{\mathrm{stop}}\}\}\ :$$

l'espace de l'observation $\mathcal{X}$ n'est alors plus $]0, t_{\mathrm{stop}}]^n$, mais l'ensemble des multi-ensembles de cardinal n dont les éléments appartiennent à $]0, t_{\mathrm{stop}}]$.

(12).★ Calculer la fonction de vraisemblance pour ce nouveau modèle.

Indication : Conformément à ce que l'intuition suggère (à savoir : comme les tubes sont tous identiques, peu importe lequel défaille plus tôt ou plus tard...), la conclusion sera que la fonction de

[¶]. Un *multi-ensemble* (ou « sac »), qu'on note par des accolades doubles, est une collection d'objets où l'ordre d'écriture ne compte pas (comme pour un ensemble), mais où les répétitions sont prises en compte (contrairement au cas des ensembles) : par exemple, les multi-ensembles $\{\{1, 2, 2\}\}$ et $\{\{2, 1, 2\}\}$ sont identiques, mais ils sont par contre distincts du multi-ensemble $\{\{1, 2\}\}$! Voir aussi en p. 11 du polycopié.

vraisemblance est la même que pour l'ancien modèle. L'enjeu est de s'en assurer mathématiquement... !

EXERCICE 3 — Dépouillement

Dans un village de Lorraine, il vient d'y avoir lieu une élection municipale pour qui sera la ou le prochain maire, parmi deux candidats appelés respectivement Katia *et* Nadir. *À l'issue de l'élection,* $444 =: B_{\mathrm{T}}$ *('*T*' comme « total ») bulletins ont été déposés dans l'urne. Vous êtes membre du comité de contrôle de l'intégrité du processus électoral, chargé(e) de l'expertise statistique. Comme on craint des fraudes électorales, on vous a demandé de vérifier que l'évolution des résultats au cours du dépouillement ne présente pas d'anomalie. En particulier, on sait qu'avant que le dépouillement ne commence, les* B_{T} *bulletins de l'urne ont été parfaitement mélangés, donc "normalement" les nombres de bulletins pour resp. Katia et Nadir devraient évoluer de façon "à peu près" régulière. Mais quantifier précisément quels niveaux d'« à peu près » sont crédibles ou pas n'est pas une mince affaire, et on compte donc sur vos compétences d'ingénieur(e) sur ce point !...*

Nous supposerons ici pour simplifier qu'il ne peut y avoir aucun bulletin blanc ni nul. Les bulletins sont dépouillés en trois étapes : on regarde d'abord un premier résultat intermédiaire au bout de $100 =: B_0$ *bulletins dépouillés, puis un second résultat intermédiaire au bout de* $100 =: B_1$ *bulletins dépouillés supplémentaires, avant de finir le dépouillement des* $244 =: B_2$ *bulletins restants. Vos instructions sont plus spécifiquement de vérifier que les résultats obtenus à l'issue de la deuxième étape sont cohérents avec ceux obtenus à l'issue de la première étape, afin qu'on interrompe immédiatement le processus dans le cas contraire. Dans cette perspective, ce problème va se focaliser sur l'enjeu suivant :*

> **Comment prédire les résultats qu'on devrait avoir à la seconde étape (en l'absence de fraude, s'entend) à partir des résultats obtenus à la première étape ?**

On note $K_{\mathrm{T}}, N_{\mathrm{T}}$ *les nombres totaux de bulletins pour respectivement Katia et Nadir contenus dans l'urne,* K_0, N_0 *les nombres de bulletins obtenus par eux à l'issue de la première étape de dépouillement,* K_1, N_1 *les nombres de bulletins obtenus lors de la seconde étape de dépouillement (à l'issue de laquelle des scores deviennent donc resp.* $K_0 + K_1$ *et* $N_0 + N_1$*), et* K_2, N_2 *les nombres obtenus lors de la dernière étape.*

Dans la mesure où le résultat de Katia est automatiquement lié à celui de Nadir, on raisonnera toujours en termes du score de Nadir, la notation 'K_1*' (et de même pour toutes les notations apparentées) devant alors être considérée comme un simple raccourci pour « *$B_1 - N_1$ *». Par ailleurs, en ce qui concerne les nombres de bulletins recouvrant plusieurs étapes de dépouillement, on se permettra de noter par exemple '*B_{01}*' pour «* $B_0 + B_1$ *», ou '*K_{12}*' pour «* $K_1 + K_2$ *».*

Un outil qui nous sera utile pour résoudre ce problème est la loi hypergéométrique*, définie de la façon suivante :*

Définition. Pour $N, K, n \in \mathbb{N}$ avec $K, n \leqslant N$, la *loi hypergéométrique* de paramètres N (taille de la population totale), K (taille de la population d'intérêt) et n (taille de l'échantillon), notée $\mathrm{Hgéom}(N, K, n)$, décrit la loi du nombre de boules noires qu'on obtient quand on tire simultanément (donc *sans* remise) n boules au sein d'un récipient contenant une population totale de N boules indiscernables parmi lesquelles K sont noires.

(1). Démontrer que pour tout entier naturel $k \in [\![0, n]\!]$:

$$\mathbb{P}\big(\mathrm{Hgéom}(N, K, n) = k\big) = \mathbf{1}_{k \in [\![K+n-N, K]\!]} \frac{K!(N-K)!n!(N-n)!}{k!(n-k)!(K-k)!(N-K-n+k)!N!}.$$

Indication : On "rappelle" que, pour $n \in \mathbb{N}, p \in [\![0, n]\!]$:

$$\binom{n}{p} = \frac{n!}{p!(n-p)!}.$$

On formalise la situation que l'on doit observer en l'absence de fraude par un modèle de statistique prédictive, où N_T, N_0 et N_1 constituent resp. le paramètre caché, l'observation passée et l'observation future, à valeurs dans les espaces respectifs $[\![0, B_T]\!]$, $[\![0, B_0]\!]$ et $[\![0, B_1]\!]$.

2. Montrer que, dans le contexte probabiliste sachant le paramètre caché, la fonction de masse de la loi de l'observation complétée est donnée par : (pour $n_0 \in [\![0, B_0]\!], n_1 \in [\![0, B_1]\!]$),

$$\mathbb{P}(N_0 = n_0 \text{ et } N_1 = n_1 \mid N_T = n_T) = \mathbf{1}_{n_0 \in [\![n_T - B_{12}, n_T]\!], n_1 \in [\![n_T - n_0 - B_2, n_T - n_0]\!]} \frac{B_0!B_1!B_2!n_T!k_T!}{n_0!n_1!n_2!k_0!k_1!k_2!B_T!}$$

(où 'n_2' doit être considéré comme un simple raccourci pour « $n_T - n_0 - n_1$ »).

(On notera au passage qu'en remplaçant B_1 par 0 et B_2 par B_{12} — auquel cas N_1 vaut nécessairement 0 —, la formule ci-dessus donne aussi la fonction de masse de la loi de N_0 sachant le paramètre caché).

3.* Dans ce modèle, les observations passée et future sont-elles indépendantes sous le véritable contexte probabiliste [‖] ? (autrement dit, le contexte sachant la véritable valeur du paramètre caché). Et sous le contexte loi à priori ? Et sous le contexte à postériori ? Et dans les cas où N_0 et N_1 ne sont pas indépendantes, s'attend-on plutôt à ce que la corrélation entre ces variables soit *positive* (à savoir, que plus N_0 est grande, plus N_1 a tendance à être grande) ou *négative* (à savoir, plus N_0 est grande, plus N_1 a tendance à être petite) ? Justifier vos réponses [**].

4. On note $k_{0\checkmark}$ et $n_{0\checkmark}$ les scores effectifs obtenus resp. par Katia et de Nadir à l'issue de la première partie du dépouillement. Montrer qu'une fonction de vraisemblance pour le paramètre caché est alors donnée par

$$\mathscr{L}(N_T = n_T \mid N_0 = n_{0\checkmark}) = \mathbf{1}_{n_T \in [\![n_{0\checkmark}, n_{0\checkmark} + B_{12}]\!]} \frac{n_T!(B_T - n_T)!}{(n_T - n_{0\checkmark})!(B_{12} - n_T + n_{0\checkmark})!}.$$

On considère à présent un modèle bayésien dans lequel on a posé une certaine priore sur N_T.

[‖]. Il est entendu que, pour chacune des trois questions posées, on répondra « non » dès lors qu'on n'a pas *systématiquement* indépendance : on ne répondra « oui » que si l'indépendance est vraie pour *toutes* les valeurs possibles du paramètre caché et/ou pour tous les choix de priore possibles.

[**]. Ici on ne demande pas d'être mathématiquement rigoureux, mais juste que les explications soient suffisamment convaincantes pour emporter l'adhésion.

5. Montrer qu'on aura alors

$$\mathbb{P}_{\text{post}}(N_1 = n_1) \propto \sum_{n_T=0}^{B_T} \mathscr{L}(N_T = n_T \mid N_0 = n_{0\checkmark})\, \mathbb{P}_{\text{pr}}(N_T = n_T)\, \mathbb{P}(N_1 = n_1 \mid N_0 = n_{0\checkmark} \text{ et } N_T = n_T).$$

6. Concernant le dernier facteur intervenant dans la somme ci-dessus, expliquer pourquoi

$$\text{Loi}(N_1 \mid N_0 = n_{0\checkmark} \text{ et } N_T = n_T).$$

est une loi hypergéométrique, et préciser quels sont ses paramètres.

À la question précédente, vous avez dû trouver un résultat ne faisant pas intervenir la priore. Cela était en fait prévisible !…

7. Plus généralement, à quelle condition [††] sur le modèle a-t-on la garantie que la loi de l'observation future sachant la valeur effective de l'observation passée et une certaine valeur du paramètre caché (autrement dit, avec les notations génériques, $\text{Loi}(Y \mid X = x_{\checkmark} \text{ et } \theta = \theta)$ peut être déterminée sans avoir à connaitre la priore ?

On introduit maintenant une priore sur N_T. Rappelons que cette priore a été choisie de façon à exprimer au mieux les attentes qu'il était légitime d'avoir, avant *de commencer le dépouillement, quant au résultat de l'élection. Après une longue réflexion (que nous n'expliquerons pas), vous décidez d'opter pour la priore suivante :*

$$\mathbb{P}_{\text{pr}}(N_T = n_T) \propto (n_T + 1)^2 (B_T - n_T + 1)^2 \exp\left(2 \frac{2\pi_{\text{att}} - 1}{\pi_{\text{att}}(1 - \pi_{\text{att}})} \frac{n_T}{B_T}\right),$$

où π_{att} est un paramètre que correspondant à une proportion, que vous prenez égal à 40 %. L'allure graphique de cette priore est donnée en figure 3.

8. Dans le cours, il est expliqué que « la priore décrit l'information (ou la croyance) dont on dispose sur le paramètre caché, lorsqu'on se place avant de collecter les données ». En l'occurrence, au vu de la fonction de répartition de la priore, comment décririez-vous (approximativement), à l'aide du langage ordinaire, l'information qu'elle décrit ? (Autrement dit : Si, au lieu de décrire la priore avec une formule, on essayait juste de dire grosso-modo à quelle croyance sur N_T elle correspond, comment décrirait-on cette « croyance » ?). Quelle notion informelle la valeur π_{att} exprime-t-elle, à votre avis ?

Comme notre formule pour déterminer la loi à postériori ne vaut qu'à proportionnalité près, et que de toutes façons notre priore n'est elle-même définie qu'à proportionnalité près, les résultats "bruts" que nous obtenons pour la postériore sur N_1 ne sont pas les bons ! Par exemple, si nous prenons (juste pour la question suivante !) $B_0 = 10$ et $B_1 = 5$ (en gardant $B_T = 444$), avec $n_{0\checkmark} = 7$, on obtient que

$$\mathbb{P}_{\text{post}}(N_1 = n_1) \propto \begin{cases} 5{,}258 \times 10^{37} & \text{pour } n_1 = 0\,; \\ 2{,}448 \times 10^{38} & \text{pour } n_1 = 1\,; \\ 5{,}462 \times 10^{38} & \text{pour } n_1 = 2\,; \end{cases} \quad \begin{cases} 7{,}320 \times 10^{38} & \text{pour } n_1 = 3\,; \\ 5{,}931 \times 10^{38} & \text{pour } n_1 = 4\,; \\ 2{,}347 \times 10^{38} & \text{pour } n_1 = 5. \end{cases}$$

[††]. Il est ici attendu de donner une condition aussi faible que possible, bien entendu ☊

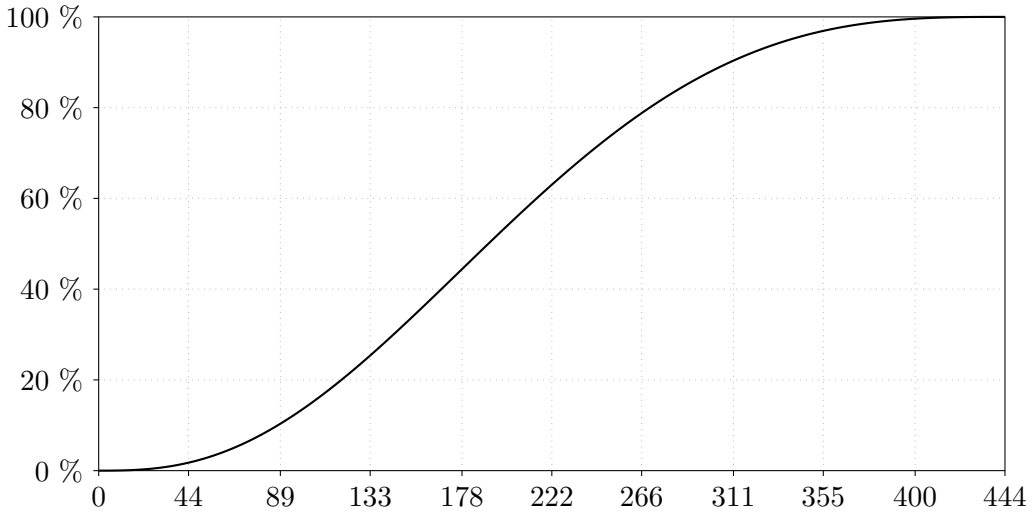


FIGURE 3 – Fonction de répartition de la priore choisie pour N_T. Les pointillés servent juste à faciliter la lecture graphique ; les graduations choisies en abscisse correspondant à resp. 10 %, 20 % &c. du nombre total B_T de bulletins (modulo arrondi).

9. Calculer les véritables valeurs des $\mathbb{P}_{\text{post}}(N_1 = n_1)$ (sans facteur de proportionnalité).

10.* Expliquer en quoi les valeurs numériques trouvées corroborent l'idée que la postériore réalise une sorte de "synthèse" entre l'information fournie par la priore et celle fournie par l'observation.

(11).* Observer que cette postériore est "plus étalée" qu'une loi binomiale. En quoi était-ce attendu ?...

Les autres membres du Comité de Contrôle, à qui vous avez expliqué votre démarche statistique, soulèvent une objection : certes, votre modèle permet de prédire ce qui devrait *se passer en l'absence de fraude... Mais pour autant, il ne semble pas capable de conclure sur la probabilité qu'un comportement inattendu corresponde effectivement à une fraude !*

(12).** L'objection de vos collègues est-elle justifiée ?
— Si non, expliquer dans les grandes lignes comment on pourrait récupérer, à partir de notre modèle, la probabilité qu'un comportement donné pour la valeur (B_0, B_1) corresponde à une fraude.
— Si oui, expliquer dans les grandes lignes comment on pourrait essayer d'améliorer notre modèle pour répondre à l'objection de vos collègues.

EXERCICE 4 — Calibration d'un tour automatique

☛ *Pour cet exercice, on "rappelle" le résultat suivant :*

Théorème. *Pour $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $x \in \mathbb{R}$, $\mathrm{d}x$ un voisinage infinitésimal de x :*

$$\mathbb{P}\big(\text{Normale}(\mu, \sigma^2) \in \mathrm{d}x\big) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}x).$$

Dans une usine, un tour automatique fabrique des axes cylindriques de diamètre nominal $3\,600$ µm $=: \mu_{\text{ref}}$. *En pratique, cependant, pour un réglage donné du tour, le diamètre effectivement obtenu pour les axes est aléatoire, le diamètre de chaque axe suivant, indépendamment, une loi* Normale$(\mu_{\checkmark}, \sigma_1^2)$, *où* $\mu_{\checkmark}$ *dépend du réglage de la machine et* σ_1 *est connu, égal à* 20 µm. *D'une série de production sur l'autre, le réglage du tour est susceptible de varier pour différentes raisons : l'expérience montre que, pour une série donnée,* $\mu_{\checkmark}$ *se comporte comme la réalisation d'une loi* Normale$(\mu_{\text{ref}}, \sigma_0^2)$, *où* σ_0 *est connu et vaut* 12 µm.

Lors d'un contrôle de qualité, on prend $8 =: n$ *axes issus de la même série de production et on mesure leurs diamètres effectifs : on trouve la série de valeurs (en micromètres) :*

$$3\,612, \quad 3\,574, \quad 3\,584, \quad 3\,578, \quad 3\,588, \quad 3\,542, \quad 3\,589, \quad 3\,557,$$

que nous noterons $d_{0\checkmark}, d_{1\checkmark}, \dots, d_{(n-1)\checkmark}$. *On se demande, à partir de cette information, quelle probabilité est-ce que* $\mu_{\checkmark}$ *a de valoir tant ou tant, et ce qu'on peut en déduire sur la production future si on laisse le réglage de la machine tel quel.*

1. Modéliser ce problème avec le vocabulaire de la statistique bayésienne : y a-t-il des paramètres du modèle (et si oui lesquels) ; quel est le paramètre caché, l'espace dans lequel il vit, et sa loi à priori ; que représente l'observation et quelle est sa valeur effective, dans quel espace vit l'observation et quelle est sa loi sachant la valeur du paramètre caché ?

2. Déterminer la fonction de vraisemblance pour $\boldsymbol{\mu}$ dans ce modèle, pour l'observation effective décrite ci-dessus.

Indication : Pour alléger les notations, vous pouvez utiliser les notations suivantes : $m_{\checkmark} := (d_{0\checkmark} + \dots + d_{(n-1)\checkmark})/n \underset{\text{déf}}{=} 3\,578$ µm ; $\sigma_2 = (\sigma_0^{-2} + n\sigma_1^{-2})^{-1/2} \underset{\text{déf}}{=} 6{,}092$ µm.

3. Déterminer et identifier la loi à postériori de $\boldsymbol{\mu}$. (On calculera aussi les applications numériques).

(4). Par rapport à la question précédente, supposons que la valeur de σ_0 change et qu'on s'autorise à faire tendre σ_0 vers l'infini. Dans ces conditions, que devient, à la limite, la loi à priori de $\boldsymbol{\mu}$? Et sa loi à postériori ?... Vérifier que ce comportement asymptotique coïncide avec ce qu'on aurait obtenu par la technique de priore impropre.

Nous revenons à présent à la véritable valeur de σ_0. *La question que se posent les ingénieurs de l'usine, c'est de savoir s'ils doivent laisser la production se poursuivre avec le réglage actuel de la machine. En pratique, ce qui compte pour les axes fabriqués, c'est que leur diamètre soit compris dans l'intervalle de tolérance* $[\mu_{\text{ref}} \pm \varepsilon_{\text{tol}}]$, *avec* $\varepsilon_{\text{tol}} := 50$ µm.

On introduit dans le modèle une observation future $\boldsymbol{D'}$, *correspondant au diamètre du prochain axe (en conservant le réglage actuel).*

5.$^{\bigstar}$ Justifier que, conditionnellement à l'évènement $\{\vec{D} = \vec{d}_{\checkmark}\}$, les v. a. $\boldsymbol{\mu}$ et $\boldsymbol{D'} - \boldsymbol{\mu}$ sont indépendantes, la seconde suivant une loi Normale$(0, \sigma_1^2)$.

6. En déduire la loi à postériori de $\boldsymbol{D'}$.

(7). En déduire la probabilité (à postériori) que le prochain axe produit soit défectueux. On supposera pour cette question qu'on dispose de la fonction de répartition de la loi normale standard, notée $\Phi(\bullet)$ et implémentée, par exemple, par la fonction `pnorm` sous **R**. Si votre matériel de calcul le permet, donner l'application numérique.

Inférence statistique / Séance 3

Analyses bayésiennes

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy *

31 mars 2025

EXERCICE 1 — Test de QI

Le quotient intellectuel *(QI) est un outil psychométrique visant à quantifier l'« intelligence générale » d'un individu. Bien que ce qu'on entende précisément par « intelligence générale » demeure assez ambigu, il y a consensus au sein des chercheurs du domaine pour dire que cela correspond bel et bien à un concept pertinent, et que c'est bien ce concept-ci que les tests de QI mesurent.*

Néanmoins, la « mesure » du QI d'un individu donné se fait via un test d'environ deux heures, lors duquel la "forme intellectuelle" du jour de la personne testée peut se trouver être meilleure ou moins bonne que d'habitude ; en outre, avec seulement une grosse centaine d'items pour l'ensemble du test, le simple facteur chance peut faire que quelqu'un de plus doué est tombé sur des items où il a eu plus de mal (par rapport à d'autres items de difficulté équivalente dans l'absolu, s'entend), et se retrouver avec un moins bon score qu'une personne un peu moins douée mais plus chanceuse…

En fait, on sait mesurer l'ampleur des fluctuations mentionnées à l'alinéa précédent : des expériences ont montré que, lorsqu'une personne donnée passait un test, le résultat de son test suivant une distribution aléatoire Normale$(\gamma_{\checkmark}, \sigma_{\text{fl}}^2)$*, où $\gamma_{\checkmark}$ est l'« intelligence générale véritable » de la personne en question, qui ne dépend que des qualités cognitives intrinsèques de cette personne (et pas de sa forme le jour du test, ni de sa chance avec les items qui lui seront soumis) et σ_{fl} est une valeur connue, la même pour tous les gens, égale à* 7 pt *(où* pt *désigne le « point de QI », l'unité utilisée par les psychométriciens dans ce genre de cadre.*

On s'intéresse ici à une personne, dont on ne sait rien à priori, qui vient de passer un test de QI et qui y a obtenu un score $q_{\checkmark}$. On ce demande ce que cette performance nous permet effectivement *de savoir sur l'intelligence générale, $\gamma_{\checkmark}$, de la personne en question.*

1.❀ Reformuler ce qui précède comme un modèle statistique.

On se propose de suivre une approche bayésienne pour étudier ce problème. Il se trouve que les tests de quotient intellectuels sont calibrés de sorte que, par construction, dans la situation qui nous intéresse, la loi à priori de $\boldsymbol{Q}$ [la v.a. dont $q_{\checkmark}$ est la réalisation] soit Normale$(\mu_{\text{st}}, \sigma_{\text{st}}^2)$*, avec $\mu_{\text{st}} \coloneqq 100$ pt et $\sigma_{\text{st}} = 15$ pt.*

(2).★ Justifier que, en ce qui concerne la loi à priori de $\boldsymbol{\gamma}$ [la v.a. dont $\gamma_{\checkmark}$ est la réalisation], ici il convient de prendre la loi Normale$(\mu_{\text{st}}, \sigma_{\text{st}}^2 - \sigma_{\text{fl}}^2)$.

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

3. Montrer que la fonction de ln-vraisemblance associée à notre expérience statistique est

$$\gamma \mapsto -(\gamma - q_{\checkmark})^2 / 2\sigma_{\mathrm{fl}}^2.$$

4. En déduire que la distribution à postériori pour $\boldsymbol{\gamma}$ suit la loi

$$\mathrm{Normale}\big(q_{\checkmark} + r^2(\mu_{\mathrm{st}} - q_{\checkmark}), (1 - r^2)\sigma_{\mathrm{fl}}^2\big),$$

où on a posé $r \coloneqq \sigma_{\mathrm{fl}} / \sigma_{\mathrm{st}}$.

5.★ En déduire que, si on notre but est d'estimer la véritable intelligence $\boldsymbol{\gamma}$ de l'individu à partir de la valeur Q obtenue au test, et qu'on se fixe une fonction de perte $\ell(\gamma, \hat{\gamma}) \coloneqq |\hat{\gamma} - \gamma|^k$ pour $k \in]1, \infty[$, alors l'estimateur bayésien optimal ne dépend pas de la valeur de k et vaut en l'occurrence

$$\hat{\boldsymbol{\gamma}}^{\mathrm{Bay}} \coloneqq Q + r^2(\mu_{\mathrm{st}} - Q).$$

Indication : Éventuellement, on pourra se contenter de démontrer le résultat pour $k = 2$.

Indication : Pour le cas général $k \in]1, \infty[$, on pourra commencer par redémontrer, ou éventuellement admettre, le résultat suivant : si $f : \mathbb{R} \to \mathbb{R}$ est une fonction convexe dérivable et X une v. a. vérifiant $\mathbb{E}(f'(X)) = 0$, alors pour tout $a \in \mathbb{R}$, $\mathbb{E}\big(f(X + a)\big) \geqslant \mathbb{E}(f(X))$.

6.★ Montrer que, si un individu a passé un test de QI lui ayant donné un résultat $q_{1\checkmark}$, et qu'il repasse indépendamment un autre test de QI, le résultat Q_2 qu'il obtiendra à celui-ci suivra (du point de vue de l'observateur ne connaissant rien de l'individu à part de résultat de son premier test) la loi

$$\mathrm{Normale}\big(q_{1\checkmark} + r^2(\mu_{\mathrm{st}} - q_{1\checkmark}), (2 - r^2)\sigma_{\mathrm{fl}}^2\big).$$

7. Application numérique. Traditionnellement, on considère une personne comme « surdouée » lorsque son test de QI a donné une valeur supérieure ou égale $130 =: q_{\mathrm{HP}}$. Dans le cas où $q_{1\checkmark} = 131$, quelle est la probabilité que $Q_2 \geqslant q_{\mathrm{HP}}$? Commenter.

Indication : On pourra utiliser les fonctions de tables de la loi normale standard, que vous trouverez en annexe dans la version de cet énoncé déposée sur *Arche*.

EXERCICE 2 — Comment compter des chars d'assaut

☛ *Cet exercice est inspiré d'une histoire authentique : pendant la seconde guerre mondiale, les Alliés avaient observé que les chars allemands qu'ils capturaient étaient munis de numéros de série manifestement attribués de façon régulière, et s'étaient alors servi de cette observation pour estimer, par une approche statistique similaire à celle que nous allons présenter, la production ennemie — ce qui était une information stratégique importante ! Après la guerre, l'examen des archives allemandes a montré que les estimations obtenues par l'analyse statistique étaient bien plus précises que celles obtenues par les missions de renseignement "classiques" !*

Nous proposons ici une version grossièrement simplifiée de ce problème, où nous avons notamment approximé la situation discrète par une situation continue. Du coup, le parallèle avec l'histoire des chars d'assaut est quasiment impossible à comprendre tel quel : pour ceux que les explications sur ce parallèle intéressent, voir la dernière question.

Dans cet exercice, on considère X_1, X_2, X_3 trois variables aléatoires tirées indépendamment selon la loi $\mathrm{Unif}^{\mathrm{me}}(0, \theta_{\checkmark})$*, où $\theta_{\checkmark}$ est inconnue, et vue comme une réalisation d'une variable aléatoire $\boldsymbol{\theta}$. Le but du problème est de dire des choses intelligentes sur la valeur de $\boldsymbol{\theta}$ à partir de la connaissance des réalisations de X_1, X_2, X_3, en utilisant une méthode d'analyse*

bayésienne. Pour ce faire, diverses considérations théoriques nous amènent à vouloir utiliser la distribution à priori

$$\mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = Z^{-1}\mathbf{1}_{\theta>0}\theta^{-1}\,\mathrm{vol}_1(\mathrm{d}\theta),$$

où Z est, formellement, la constante de normalisation qui permet de faire de $\mathbb{P}_{\text{pr}}$ une distribution de probabilité.

(1).❀ Traduire ce qui précède dans le langage de la statistique bayésienne :
— Y a-t-il des paramètres du modèle ; si oui lesquels ?
— Quel est le paramètre caché ; l'espace dans lequel il vit ?
— Quelle est l'observation (passée), resp. l'observation future (s'il y en a une) ; dans quels espaces respectifs vivent-elles ? Quelle est la loi de l'observation (complétée le cas échéant) sachant le paramètre caché ?
— Quelle est la priore proposée sur le paramètre caché ; s'agit-il d'une priore propre ou impropre ?

En pratique, on observe les réalisations $x_{1\checkmark} = 2\,597$, $x_{2\checkmark} = 4\,202$ et $x_{3\checkmark} = 3\,589$.

2. Calculer la fonction de vraisemblance pour $\boldsymbol{\theta}$.

3. En déduire la distribution de probabilité à postériori pour $\boldsymbol{\theta}$, qu'on décrira en donnant sa densité.

4. En déduire le mode à postériori de $\boldsymbol{\theta}$.

Indication : Pour les besoins de cette question, il pourra éventuellement être nécessaire d'étendre la définition du « mode » d'une fonction f à certains cas où f n'admet pas de maximum, de la façon suivante : s'il existe une valeur x_{mod} (nécessairement unique le cas échéant) telle qu'on ait l'implication « $f(x_n) \overset{n\to\infty}{\to} \sup f \implies x_n \overset{n\to\infty}{\to} x_{\text{mod}}$ », alors on dit que le mode de f est x_{mod}.

5. Que vaut l'espérance à postériori de $\boldsymbol{\theta}$?

6. Calculer la fonction de répartition de la loi à postériori de $\boldsymbol{\theta}$.

7. En déduire :
— Notre crédence (probabilité) à postériori sur le fait que $\boldsymbol{\theta}$ soit compris entre 5 000 et 6 000 ;
— À quel niveau de risque nous pouvons certifier l'hypothèse que $\boldsymbol{\theta}$ est inférieur à 12 000. (Commenter le résultat obtenu en le reformulant qualitativement).

8. Pour $p \in]0,1[$, calculer le quantile de niveau p pour la distribution à postériori de $\boldsymbol{\theta}$.

9. En déduire (pour ceux disposant d'une calculatrice, on calculera les valeurs numériques) :
— La médiane à postériori de $\boldsymbol{\theta}$;
— L'intervalle de confiance bayésien pour $\boldsymbol{\theta}$, au niveau de confiance 92 % ;
— Un majorant (à postériori) de $\boldsymbol{\theta}$, au niveau de risque 5 %.

Maintenant, supposons qu'on cherche à prédire une observation future $\boldsymbol{X}_4$, suivant elle aussi la loi $\mathrm{Unif}^{\mathrm{me}}(0,\theta_{\checkmark})$*, indépendamment des autres $\boldsymbol{X}_i$.*

(10).❀ Préciser ce que cela change en termes de modélisation.

(11). Démontrer que, pour tout $\theta \in \mathbb{R}_+^*$, pour tous $x_1, x_2, x_3 \in]0, \theta[$:

$$\mathrm{Loi}(X_4 \mid \boldsymbol{\theta} = \theta \text{ et } \vec{\boldsymbol{X}}_{[\![1,3]\!]} = \vec{x}_{[\![1,3]\!]}) = \mathrm{Loi}_\theta(X_4).$$

Expliquer en quoi ce résultat est intuitivement évident.

(12). Démontrer que, pour tout sous-ensemble infinitésimal $\mathrm{d}x_4 \subseteq]0, \infty[$,

$$\mathbb{P}_{\mathrm{post}}(X_4 \in \mathrm{d}x_4) = \int_{\theta \in \mathbb{R}_+^*} \mathbb{P}_\theta(X_4 \in \mathrm{d}x_4)\, \mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta).$$

13. En déduire la loi à postériori de X_4 au vu des observations. (On donnera la densité de cette loi).

14.★ Déterminer la médiane et l'espérance à postériori de X_4. Déterminer la probabilité à postériori que $X_4 \in [2\,597, 4\,202]$; un intervalle de prévision à 92 % pour la valeur de X_4 ; un majorant de X_4 à 5 % de risque ; et le niveau de confiance qu'on peut avoir pour rejeter l'hypothèse que $x_{4\checkmark} > 12\,000$.

(15).★★ Reprendre toute la première partie de l'exercice (celle concernant l'inférence explicative) en supposant, cette fois-ci, que $\boldsymbol{\theta}$ et les $\boldsymbol{X}_i$ sont à valeurs entières, les quatre valeurs $\boldsymbol{X}_i$ étant tirées uniformément et *sans remise* dans $[\![0, \theta_\checkmark[\![$, en supposant qu'on a pris à priori $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} = \theta) \propto \mathbf{1}_{\theta \geqslant 100} \theta^{-1}$ pour tout θ entier. Constater que, "au premier ordre", tous les calculs se déroulent exactement de la même façon... En déduire en quoi on peut effectivement relier cette exercice à la problématique des chars d'assaut présentée en début d'énoncé.

EXERCICE 3 — Les prédictions de Nate Silver

L'exercice qui suit est inspiré de la méthode utilisée par la statisticien Nate Silver *pour estimer les probabilités de victoire des différents candidats lors des élections présidentielles étasuniennes (confer* `www.natesilver.net`*).*

Dans un pays ayant une tradition politique de bipartisme, deux partis s'affrontent à chaque élection, tous les quatre ans : les Bleus *et les* Rouges*. Au préalable de chaque élection, un collectif d'instituts de sondage tente de prédire le résultat de l'élection, calculé sous la forme de la différence entre les point de pourcentages de voix respectifs reçus par le (ou la) candidat(e) rouge et le (ou la) candidat(e) bleu(e). Nous notons $s_{i\checkmark}$ la valeur prédite par les sondeurs lors de l'année i, resp. $\rho_{i\checkmark}$ le véritable résultat de l'élection. Pour simplifier les notations, l'origine des temps est prise à l'an 1970 (de sorte que l'an 2024, par exemple, correspond à l'année 54).*

Nous faisons la modélisation suivante : les différents $\boldsymbol{S}_i$ dont les $s_{i\checkmark}$ sont les réalisations sont indépendants et suivent la loi Normale$(\rho_{i\checkmark}, \sigma_\checkmark^2)$*, où $\sigma_\checkmark$ est une quantité inconnue, indépendante de i. Nous connaissons toutes les valeurs de $\rho_{i\checkmark}$ et $s_{i\checkmark}$ pour i entre 2 et 50, ainsi que la valeur de $s_{54\checkmark}$: voir table 1.*

1.★ Récapituler les informations ci-dessus sous forme d'un modèle statistique. On précisera, comme d'habitude :

— Y a-t-il des paramètres du modèle, et si oui lesquels ?
— Quel est le paramètre caché, et son espace ; que représente-t-il ?
— Quelle est l'observation (passée), l'espace dans laquel elle vit, et la valeur qu'elle a effectivement prise ; que représente cette observation ?

i	02	06	10	14	18	22	26
$\rho_{i\checkmark}$	+23	−2	+10	+18	+8	−6	−9
$s_{i\checkmark}$	+24	−1	+2	+18	+10	−7	−13

i	30	34	38	42	46	50	54
$\rho_{i\checkmark}$	−1	+2	−7	−4	−2	−4	?
$s_{i\checkmark}$	+4	+2	−7	0	−4	−8	−1

TABLE 1 – Résultat des élections ($\rho_{i\checkmark}$) de 1972 à 2020, comparés aux chiffres des sondages ($s_{i\checkmark}$), incluant pour ces derniers le chiffre pour l'année 2024.

— Quelle est, le cas échéant, l'observation future, l'espace dans laquel elle vit ; et que représente-t-elle ?
— Quelle est la loi de l'observation (complétée le cas échéant) sachant le paramètre caché ?

Indication : Attention, les réponses à cette question sont un peu contre-intuitives... ! Pour vous aider, sachez qu'il n'y a pas plus de 4 paramètres cachés [†], et au moins 10 paramètres du modèle.

☛ *Pour la suite de cet exercice, on donne le résultat suivant :*

Théorème. *que pour $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $x \in \mathbb{R}$, $\mathrm{d}x$ un voisinage infinitésimal de x :*

$$\mathbb{P}\big(\mathrm{Normale}(\mu,\sigma^2) \in \mathrm{d}x\big) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}x). \qquad \diamond$$

2. Déterminer la fonction de vraisemblance du paramètre caché (pour les valeurs effectives des observations).

Dans la suite de l'énoncé, on suppose que, à priori, $\boldsymbol{\sigma}$ et $\boldsymbol{\rho}_{50}$ étaient indépendants, de lois (impropres) respectives

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\sigma} \in \mathrm{d}\sigma) \propto \sigma^{-1}\,\mathrm{vol}_1(\mathrm{d}\sigma)\,;$$
$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\rho}_{50} \in \mathrm{d}\rho_{50}) \propto \mathrm{vol}_1(\mathrm{d}\rho_{50}).$$

3. Déterminer la loi à postériori du paramètre caché, sous la forme d'une densité exprimée à constante multiplicative près. Vérifier que cette loi à postériori est propre. Les variables $\boldsymbol{\sigma}$ et $\boldsymbol{\rho}_{50}$ sont-elles indépendantes à postériori ?

4.* Montrer que la loi à postériori de $\boldsymbol{\rho}_{50}$ est une loi de la forme $a\,\mathrm{T}_{\mathrm{St}}(\nu)+b$ (concernant la loi $\mathrm{T}_{\mathrm{St}}(\nu)$, voir la première indication), pour des valeurs de a, ν et b qu'on identifiera.

Indication : La loi de Student à ν degré de liberté, notée $\mathrm{T}_{\mathrm{St}}(\nu)$, est la loi à densité sur $\mathbb{R}$ caractérisée par

$$\mathbb{P}(\mathrm{T}_{\mathrm{St}}(\nu) \in \mathrm{d}x) \propto \left(\frac{1}{\nu+x^2}\right)^{(\nu+1)/2} \mathrm{vol}_1(\mathrm{d}x).$$

[†]. Évidemment, on peut toujours décider de regrouper plusieurs paramètres sous forme d'un unique paramètre vectoriel... Ici, lorsque je compte le nombre de paramètres (cachés, resp. du modèle), c'est du nombre de paramètres *scalaires* que je parle.

Indication : Une première étape utile est de calculer la densité de la loi $a\,\mathrm{T}_{\mathrm{St}}(\nu) + b$. (On pourra se restreindre au cas $a > 0$, vu que de toutes façons la loi $\mathrm{T}_{\mathrm{St}}(\nu)$ est symétrique).

Pour la fin de cet exercice, on suppose qu'on dispose d'une logiciel qui nous permet de calculer la fonction de répartition et les quantiles des lois de Student, via des fonctions que nous appellerons respectivement `répartStudent` *et* `qtileStudent` *: on suppose que*

$$\texttt{répartStudent}(x, \nu) \coloneqq \mathbb{P}(\mathrm{T}_{\mathrm{St}}(\nu) \leqslant x)$$

et que $\texttt{qtileStudent}(p, \nu)$ *est le quantile de niveau* p *de la loi* $\mathrm{T}_{\mathrm{St}}(\nu)$ [‡].

5. En déduire, à l'aide de ces fonctions (on fera si possible les applications numériques) :
— Quelle est notre intervalle de croyance à 80 % sur le score qu'obtiendra le candidat rouge par rapport à la candidate bleue ;
— À quel niveau de risque nous pouvons rejeter l'hypothèse que la candidate bleue obtienne un meilleur résultat que le candidat bleu de l'élection précédente (commenter qualitativement ce niveau de risque) ;
— Quelle est, au vu des sondages, la probabilité que le candidat rouge obtienne plus de voix que la candidate bleue.

[‡]. Sous *R*, de telles fonctions sont implémentées sous les noms respectifs de `pt` et `qt`.

Inférence statistique / Séance 4

Questions de modélisation

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy *

7 avril 2025

EXERCICE 1 — Effets imprévus

Les bidulines *sont une classe de molécules thérapeutiques (imaginées pour les besoins de cet exercice) ayant des effets médicaux très intéressants (et variables d'une biduline à l'autre), mais aussi des risques d'effets secondaires graves, dont la fréquence est variable d'une biduline à l'autre... À ce jour, quelques dizaines de molécules de la famille des bidulines ont déjà été expérimentées ; pour toutes, la proportion d'effets secondaires graves était comprise entre 1 % et 5 %. Aujourd'hui, on s'apprête à tester une nouvelle molécule de cette classe, la* biduline QP*. Soit $\pi_{\checkmark}$ la fréquence d'effets secondaires graves que la biduline QP engendre. On veut être capable de donner une estimation de $\pi_{\checkmark}$ dès les tout premiers essais, grâce à une méthode bayésienne.*

*Deux groupes de statisticiens, que nous appellerons resp. **A** et **B**, travaillent sur le problème indépendamment. (À noter cependant qu'on suppose ici qu'ils s'appuient exactement sur les mêmes informations, notamment en ce qui concerne le choix de la priore).*

1. Au vu du contexte expliqué par l'énoncé, quel type de démarche va-t-il falloir suivre ici pour choisir la priore : sera-t-on plutôt dans un cas il y a une priore clairement définie de façon exacte, dans un cas où il s'agira de traduire un certaine expertise, ou dans un cas où il conviendra de proposer une priore non informative ; et pour déterminer notre priore, utilisera-t-on les résultats d'une expérience sur $\boldsymbol{\pi}$ posée dans un contexte statistique rigoureux, ou s'agira-t-il plutôt de connaissances plus informelles ?...

En l'occurrence, concrètement, quelles contraintes devra satisfaire notre priore pour pouvoir être considérée comme un choix raisonnable ?

*Les groupes **A** et **B** utilisent des choix de priore quelque peu différents sur le paramètre caché $\boldsymbol{\pi}$ (lequel est pris comme vivant dans $]0,1[$) :*

— *Le premier groupe propose de modéliser la priore sur $\boldsymbol{\pi}$ comme une loi bêta de paramètres 5 et 161. La densité correspondante est caractérisée par*

$$\mathbb{P}_{\mathrm{pr}(A)}(\boldsymbol{\pi} \in \mathrm{d}\pi) \propto x^4(1-x)^{160}\,\mathrm{vol}_1(\mathrm{d}\pi).$$

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Anouk Rago, Pierre-Adrien Tahay.

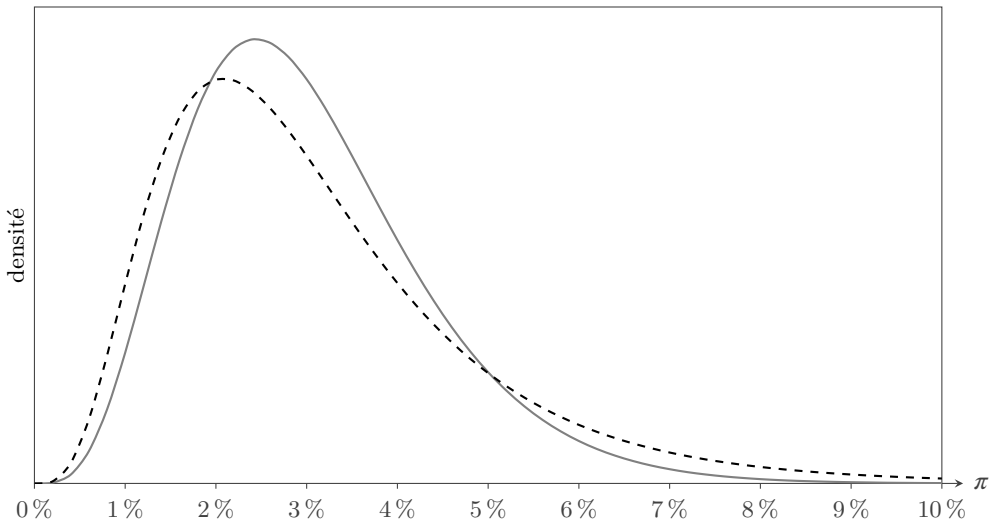


FIGURE 1 – Densités utilisées par les goupes $\boldsymbol{A}$ (en gris plein) et $\boldsymbol{B}$ (en pointillés noirs), tracées à la même échelle.

— *Le second groupe, quant à lui, n'aime pas faire intervenir des facteurs d'exposant trop grand : il utilise donc, à la place, une densité en forme de fraction rationnelle :*

$$\mathbb{P}_{\mathrm{pr}(\boldsymbol{B})}(\boldsymbol{\pi} \in \mathrm{d}\pi) \propto \frac{x^4}{(16x+1)^{16}} \,\mathrm{vol}_1(\mathrm{d}\pi).$$

2. Au vu de l'allure de ces priores, donnée en figure 1, diriez-vous que les priores proposées par les groupes respectifs $\boldsymbol{A}$ et $\boldsymbol{B}$ sont pertinentes ? (Autrement dit, respectent-elles bien les contraintes que vous avez mises en valeur à la question précédente ?). À première vue, les différences de choix de priore entre les deux groupes sont-elles subtantielles ou minimes ?

Maintenant, lors des premiers essais cliniques, il se passe quelque chose de tout à fait inattendu : sur les 20 *premiers patients testés,* 18 *montrent des effets indésirables graves !...*

(3). Donner la fonction de vraisemblance pour $\boldsymbol{\pi}$ correspondant à l'observation de ces vingt patients.

(4). En déduire, à constante multiplicative près, l'expression des lois à postériori relatives de $\boldsymbol{\pi}$ que vont trouver les groupes respectifs $\boldsymbol{A}$ et $\boldsymbol{B}$.

5. La figure 2 représente les densités des postériores pour $\boldsymbol{\pi}$ trouvées resp. par les groupes $\boldsymbol{A}$ et $\boldsymbol{B}$. Expliquer comment, à partir de la réponse à la question précédente, on peut tracer ce diagramme.

Indication : Sans parler de la partie purement « tracer le graphique avec un logiciel », il y a une opérations mathématiques à faire pour tracer le graphique en question : voyez-vous laquelle... ?

6.★ Expliquer en termes informels d'où vient la différence radicale qu'on observe entre les conclusions des groupes $\boldsymbol{A}$ et $\boldsymbol{B}$.

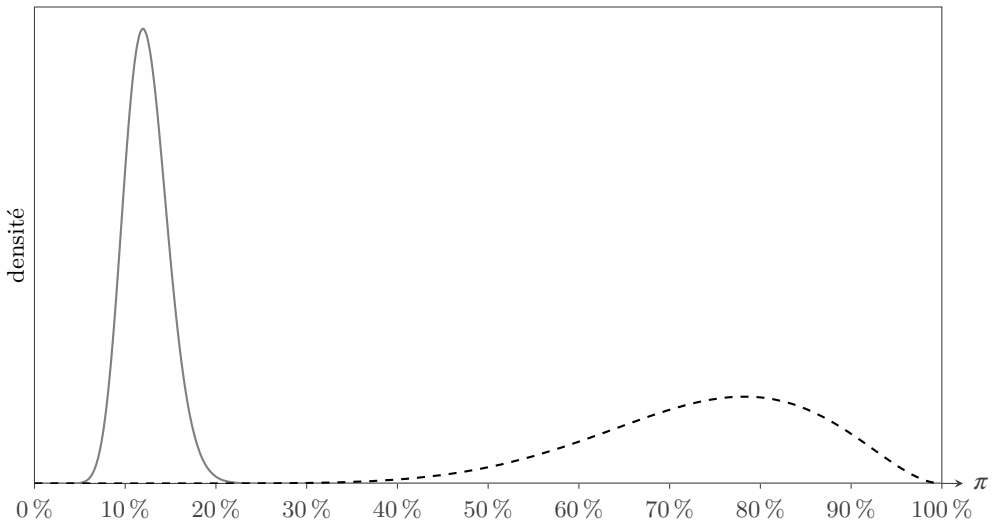


FIGURE 2 – Densités à postériori trouvées resp. par le groupe ***A*** (trait gris plein) et par le groupe ***B*** (trait pointillé noir). Les valeurs numériques des intervalles de fluctuation (à 95 % de confiance) correspondant à ces lois sont resp. [8,0 %, 17,5 %] et [46,9 %, 93,5 %].

7.* Quel est le groupe dont les conclusions vous semblent le plus raisonnables, et pourquoi ?

8. Quel passage du cours cet exercice a-t-il permis d'illustrer ?...

(9).* En quoi peut-on dire qu'en fait, aucun des deux groupes n'a pris une priore très raisonnable ? Quel choix aurait-il été meilleur ?...

EXERCICE 2 — Des réflexes de Jedi !

On s'intéresse à un test[*] *très simple destiné à mesurer les temps de réaction des gens. Le test est fait de sorte qu'il est essentiellement impossible de s'améliorer en s'entrainant : de la sorte, quand on mesure le niveau d'un individu à ce test, ce sont bien les caractéristiques intrinsèques de l'individu qu'on mesure, et pas de son entrainement.*

Après avoir fait passer le test à 65 *cobayes supposés tirés complètement au hasard dans la population mondiale (adulte), on constate que la variabilité des temps de réaction suit très manifestement une courbe en cloche, avec des queues abruptes : voir la figure 3. Pour cette raison, il parait raisonnable de supposer que le temps de réaction d'un individu pris au hasard dans la population mondiale est décrit par une loi normale (d'espérance* $\mu_{\checkmark}$ *et de variance* $\sigma^2_{\checkmark}$*, à priori inconnues).*

Ici on choisit l'inférer les valeurs de $\boldsymbol{\mu}$ *et* $\boldsymbol{\sigma}$ *à l'aide d'une priore non informative.*

1. Outre l'aspect visuellement satisfaisant de la modélisation gaussienne, en quoi est-un un modèle raisonnable du point de vue "physique" ?

[*]. Au sens courant du terme : une épruve, quoi ! ☺

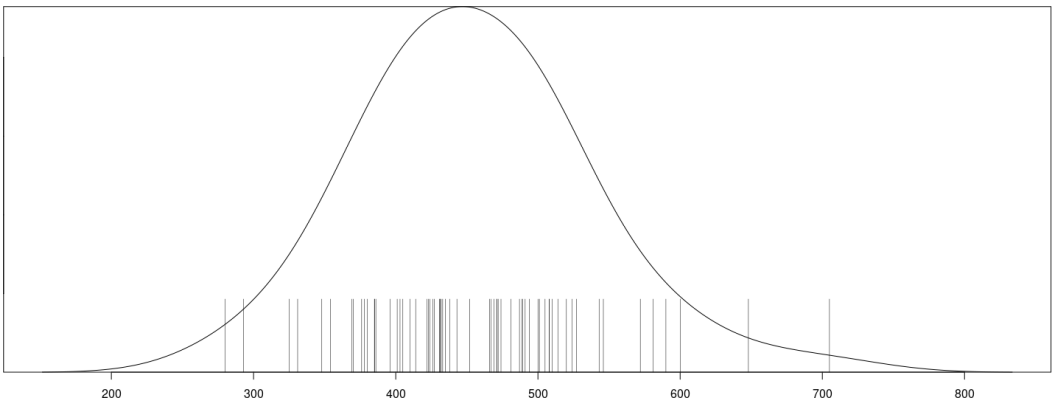


FIGURE 3 – Les barres verticales tracées sur cette figure correspondent aux temps de réaction des différents cobayes testés lors de notre expérience. La courbe continue est une estimation automatique (non paramétrique) de la densité dont ces données semblent être issues.

2. Rappeler dans quel cas il est recommandé de recourir à une priore non informative. Sommes-nous dans cette situation ici ? Pourquoi cela peut-il être intéressant malgré tout du point de vue de la démarche scientifique ? Argüer que le fait qu'on dispose d'un nombre de cobayes assez important rend en fait assez innocent, voire presque pertinent, le choix d'une priore non informative. En outre, il y a un avantage supplémentaire à utiliser une priore non informative pour un modèle très classique comme celui-ci (à savoir, inférer les paramètres d'une loi normale dans un modèle d'échantillonnage) : lequel ?...

Une approche pour construire une priore non informative est l'approche de Haar [confer remarque (`GS'`*) du polycopié], que nous explicitons ici. On commence par observer qu'on dispose d'un groupe de transformations naturelles sur l'espace* $\mathbb{R}\times\mathbb{R}_+^*$ *du paramètre caché*[†]*, consistant en les tranformations* $T_{a,b}$ *(pour* $a\in\mathbb{R}_+^*, b\in\mathbb{R}$*) définies par*

$$T_{a,b}(\mu,\sigma) \coloneqq (a\mu+b, a\sigma).$$

3. En quoi s'agit-il effectivement d'un groupe de transformations *naturelles* ?

Le but est alors de chercher s'il existe une distribution de probabilité (éventuellement) impropre qui soit invariante par (mesure-image par) ce groupe de transformations. Le cas échéant, cette distribution est nécessairement unique, sous réserve que l'action de notre groupe de transformations en questions soit transitive*, autrement dit, qu'il existe toujours un* $T_{a,b}$ *permettant d'envoyer un* (μ,σ) *sur un* (μ',σ') *donnés — ce qui est bien le cas ici. Le cas échéant, on peut effectivement considérer que cette distribution de probabilité, qu'on qualifie de « mesure de Haar », est une priore non informative pour* $(\boldsymbol{\mu},\boldsymbol{\sigma})$*, dans la mesure où elle donne la même masse à toutes les zones de* $\mathbb{R}\times\mathbb{R}_+^*$ *qui se déduisent les unes des autres par une transformation naturelle.*

[†]. Comme on est dans un cadre complètement non informatif, on néglige ici le fait que $\boldsymbol{\mu}$ ne peut en réalité pas être négatif en l'occurrence. Par ailleurs, exclure le cas $\boldsymbol{\sigma}=0$ simplifie les calculs et est une hypothèse tout à fait raisonnable en pratique : les humains ne sont pas des robots, il y a donc forcément des variations de l'un à l'autre... !

J'affirme qu'en l'occurrence, il existe effectivement une priore de Haar : à savoir, la priore impropre sur $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ *définie formellement par*

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\mu} \in \mathrm{d}\mu \text{ et } \boldsymbol{\sigma} \in \mathrm{d}\sigma) \propto \sigma^{-2} \operatorname{vol}_1(\mathrm{d}\mu) \operatorname{vol}_1(\mathrm{d}\sigma)$$

4.* Démontrer que la loi à priori de $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ est effectivement invariante par les transformations $\boldsymbol{T}_{a,b}$. (La formule de changement de variables multidimensionnelle est rappelée en indication).

Indication : On "rappelle" que si φ est une application C^1 de $\mathbb{R}^d$ dans $\mathbb{R}^d$, x un point de $\mathbb{R}^d$ et $\mathrm{d}x$ un voisinage infinitémisal de x dans $\mathbb{R}^d$, on a

$$\operatorname{vol}_d(\varphi(\mathrm{d}x)) = |\det \boldsymbol{J}_\varphi(x)| \operatorname{vol}_d(\mathrm{d}x)$$

(sous réserve que $\det \boldsymbol{J}_\varphi(x)$ soit non nul), où $\boldsymbol{J}_\varphi(x)$ est la matrice jacobienne de φ (en x), autrement dit la matrice de ses dérivées partielles (évaluées en x).

(5).* Calculer la postériore pour $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, à constante multiplicative près. On pourra utiliser la formule de la densité des lois normales unidimensionnelles rappelée en indication.

Indication : On a, pour $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*, x \in \mathbb{R}$:

$$\mathbb{P}\big(\mathrm{Normale}(\mu, \sigma^2) \in \mathrm{d}x\big) = \frac{\operatorname{vol}_1(\mathrm{d}x)}{\sqrt{2\pi}\,\sigma \exp\big((x-\mu)^2 / 2\sigma^2\big)}.$$

(6).* Observer que, avec les données numériques du fichier `reaction.tsv`, la postériore est fortement concentrée autour du couple de valeurs $(\hat{\mu}_\checkmark, \hat{\sigma}_\checkmark) \coloneqq (445\ \mathrm{ms}, 81\ \mathrm{ms})$, au sens où la probabilité à postériori qu'on ait $\{\boldsymbol{\mu} \in [455 \pm 25]\ \mathrm{ms}$ et $\boldsymbol{\sigma} \in [81 \pm 13]\ \mathrm{ms}\}$ est $\geqslant 85\ \%$.

7. Expliquer (sans forcément rentrer dans les détails techniques) quelle approximation est-ce que l'observation de la question ci-dessus permettra de faire quand il s'agira déterminer la loi à postériori d'une observation *future*.

On s'intéresse maintenant à la question de savoir : quel est le temps de réaction de la personne la plus rapide du monde ? Nous faisons ici la modélisation que les temps de réactions de tous les adultes du monde sont i.i.d. $\mathrm{Normale}(\mu_\checkmark, \sigma_\checkmark^2)$*, et qu'il y a en tout* $6 \times 10^9 =: \boldsymbol{N}$ *adultes dans le monde : on se retrouve alors avec un modèle d'échantillonnage dans le cadre prédictif, où notre quantité d'intérêt est l'infimum des* $\boldsymbol{N}$ *observations futures.*

(8). Quelle légère approximation avons-nous faite ci-dessus ? Argüer qu'elle ne prêtera pas à conséquence en pratique.

Dans la suite, soit $\boldsymbol{S}$ *la loi normale standard,* $\boldsymbol{P}$ *la loi* $\mathrm{Normale}(\mu_\checkmark, \sigma_\checkmark^2)$ *et* $\boldsymbol{P'}$ *la loi de la quantité d'intérêt (qui, rappelons-le, est le temps de réaction le plus court parmi les* $\boldsymbol{N}$ *personnes de la population mondiale) sous la véritable distribution* $\mathbb{P}_\checkmark$ *; notons* $\boldsymbol{F_S}, \boldsymbol{F_P}, \boldsymbol{F_{P'}}$*, resp.* $\boldsymbol{Q_S}, \boldsymbol{Q_P}, \boldsymbol{Q_{P'}}$ *leurs fonctions de répartition et de quantiles respectives.*

9. Exprimer $\boldsymbol{F_{P'}}$ en fonction de $\boldsymbol{F_P}$, puis en fonction de $\boldsymbol{F_S}$ (et de $\mu_\checkmark$ et $\sigma_\checkmark$). En déduire l'expression de $\boldsymbol{Q_{P'}}$ en fonction de $\boldsymbol{Q_S}$.

10. En utilisant les approximations appropriées (dont la formule donnée en indication), en déduire quelle est (à peu près) la médiane à postériori de notre quantité d'intérêt. Commenter l'absurdité du résultat obtenu.

Indication : Pour p de l'ordre de grandeur de 10^{-10}, on a $Q_S(p) \approx -\ln(1/p)^{1/2} - 1{,}56$ avec une précision de l'ordre de 10^{-1}.

Suite à la question précédente, on se demande si le souci ne vient pas de l'hypothèse de normalité du modèle. Il se trouve qu'il existe un outil très connu pour vérifier la gaussianité d'un jeu de données, appelé droite de Henry *(ou « diagramme Q/Q ») : lorsqu'un jeu de données est gaussien, les points tracés sur les axes du diagramme de Henry doivent s'aligner selon une droite, aux inévitables fluctuations aléatoires près. En l'occurrence, ce diagramme quantile-quantile est tracé en figure 4.*

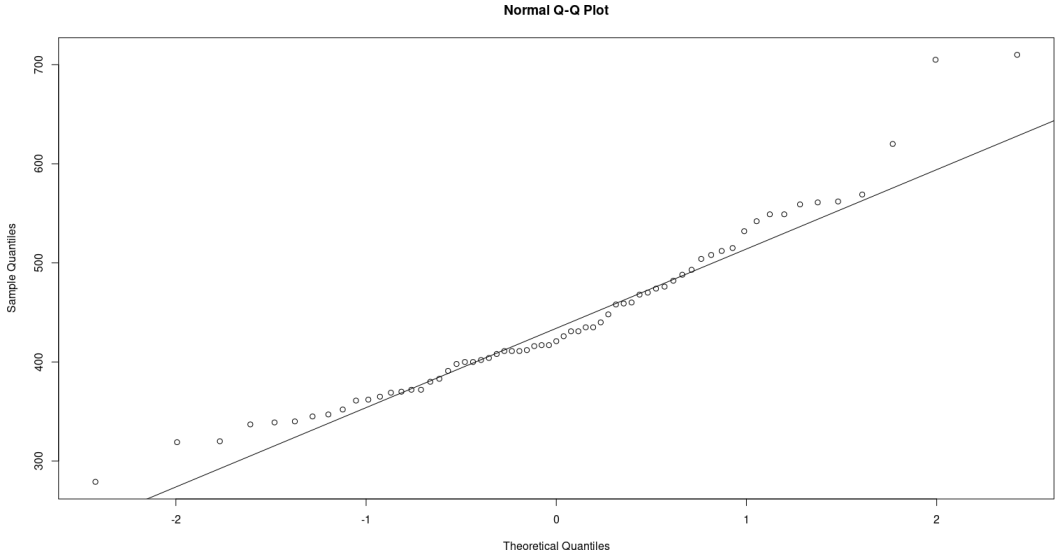


FIGURE 4 – Diagramme quantile-quantile pour le jeu de données des temps de réaction des cobayes, sur lequel est tracé la « droite de Henry » estimant comment "devraient" s'aligner les points en cas de normalité s'il n'y avait pas de fluctuations. Attention, il est connu que les fluctuations de ce genre de diagramme ont tendance à être particulièrement "violentes" au niveau des points extrêmes, de sorte qu'on recommande plutôt, en général, de se focaliser sur l'alignement des points en dehors de ces extrêmes.

11.* Le diagramme Q/Q contredit-il manifestement la gaussianité des données ? Comment expliquer le paradoxe de la question précédente au vu de celui-ci ?

12. Quel autre modèle des mêmes données pourrait-on proposer, toujours en utilisant une modélisation normale, mais qui nous fera des prédictions plus raisonnables pour le temps de réaction du record du monde ?...

EXERCICE 3 — Ondes gravitationnelles

Un détecteur d'ondes gravitationnelles a été mis en service pour détecter les coalescences de trous noirs. Une première fenêtre d'observations, durant 30 d [‡] $=: T_0$*, a permis d'observer* $4 =: n_{\checkmark}$ *coalescences. Les scientifiques opérant le détecteur souhaitent à présent mener une nouvelle fenêtre d'observations, dans les mêmes conditions observationnelles, qui durerait cette fois-ci* 90 d $=: T_1$*. Cependant, l'expérience étant particulièrement onéreuse, leurs instances administratives leur demandent de remplir un dossier pour dire quelle sera la probabilité d'observer tant ou tant de coalescences dans la nouvelle fenêtre.*

Les astrophysiciens considèrent que les coalescences de trous noirs observables se produisent avec un certain taux inconnu $\lambda_{\checkmark}$ *; et que le nombre de coalescences se produisant sur un intervalle de temps de longueur* T *suit une loi* Poisson$(\lambda_{\checkmark} T)$*, indépendamment entre deux intervalles de temps disjoints.*

1. Expliquer le choix d'avoir opéré une modélisation par loi de Poisson, et discuter ses faiblesses éventuelles. L'un dans l'autre, à quel point cette modélisation vous parait-elle approcher la réalité de façon fiable ?

2. ❀ Modéliser la situation comme un problème de prévision statistique : on donnera le paramètre caché, l'observation passée et l'observation future (à chaque fois, en donnant la signification, l'espace de valeurs, et la valeur effective le cas échéant), et la loi des observations sachant le paramètre caché. Noter qu'il y a aussi deux paramètres du modèle : lesquels ?

Au moment où le détecteur a été construit, les technologies précédentes (certes plus frustres) n'avaient encore jamais permis de détecter la moindre onde gravitationnelle, de sorte qu'on était d'ores et déjà à peu près sûr que, pour le nouveau détecteur, $\lambda_{\checkmark}$ *ne serait pas plus grand que* $0{,}15\ \mathrm{d}^{-1} =: \lambda_{\text{réf}}$*. À vrai dire, les échecs des détecteurs précédents avaient même amené certains astrophysiciens à douter que les ondes gravitationnelles existassent réellement... C'est pourquoi, avant le début de l'expérience (celle de durée* T_0*, je veux dire), quand les ingénieurs statisticiens ont mis au point leur protocole d'analyse bayésienne, ils ont proposé la priore suivante pour* $\boldsymbol{\lambda}$ *:*

$$\mathbb{P}_{\text{pr}}(\boldsymbol{\lambda} \in \mathrm{d}\lambda) \propto \begin{cases} \lambda_{\text{réf}} / 8 & \text{pour } \lambda = 0 \text{ ;} \\ \exp(-2\lambda / \lambda_{\text{réf}}) \, \mathrm{vol}(\mathrm{d}\lambda) & \text{pour } \lambda > 0. \end{cases}$$

3. ❀ Cette priore est-elle propre ou impropre ?

4. Expliquer au mieux les idées sous-tendant ce choix de priore.

5. Calculer la fonction de vraisemblance du modèle (pour l'observation passée effectivement réalisée). On veillera à écrire cette fonction de vraisemblance sous la forme la plus simple possible.

Indication : On donne, pour $k \in \mathbb{N}$, $\theta \in \mathbb{R}_+$: $\mathbb{P}(\text{Poisson}(\theta) = k) = \mathrm{e}^{-\theta} \theta^k / k!$ (en prenant $0^0 = 1$ pour $(\theta, k) = (0, 0)$).

[‡]. 1 d $:= 24 \times 60^2$ s correspond à « un jour », vu comme unité de temps (le symbole provient du latin *dies*).

6. En déduire la loi à postériori de $\boldsymbol{\lambda}$. Y reconnaitre une loi gamma (confer l'indication) dont on précisera les paramètres.

Indication : La loi gamma de paramètre de forme $k \in \mathbb{R}_+^*$ et de paramètre de taux $\lambda \in \mathbb{R}_+^*$ est la loi sur $\mathbb{R}_+^*$ suivante :

$$\mathbb{P}\big(\mathrm{Gamma}(k,\lambda) \in \mathrm{d}x\big) = \frac{\lambda^k}{(k-1)!} x^{k-1} \mathrm{e}^{-\lambda x} \,\mathrm{vol}_1(\mathrm{d}x),$$

où la notation « $(k-1)!$ » désigne implicitement la valeur $\Gamma(k)$ de la fonction d'Euler si on a $k \notin \mathbb{N}^*$.

(7).* Calculer la probabilité (à postériori) que le nombre d'observations dans l'expérience future soit égal à n, pour tout $n \in \mathbb{N}$. On pourra mener les calculs à constante près. Identifier dans le résultat une loi *binomiale-négative* dont on précisera les paramètres (voir l'indication).

Indication : On rappelle que, pour tout $n \in \mathbb{N}$,

$$\int_0^\infty x^n \mathrm{e}^{-x} \,\mathrm{d}x = n!.$$

Par ailleurs, cette formule est utilisée pour étendre la définition de la fonction factorielle pour tout $n \in]-1, \infty[$.

Indication : La loi binomiale-négative de paramètres $n \in \mathbb{R}_+$ et $p \in [0,1]$ est la loi à valeurs dans $\mathbb{N}$ caractérisée, pour $k \in \mathbb{N}$, par

$$\mathbb{P}\big(\mathrm{NégBin}(n,p) = k\big) \propto \frac{(k+n-1)!}{k!}(1-p)^k.$$

EXERCICE 4 — Coïncidence ?...

Un meurtre vient d'avoir lieu dans un petit village de Lorraine. Le seul témoin est un petit garçon qui a vu que le coupable était un homme atteint de calvitie conduisant un utilitaire jaune de modèle Kangoo. *L'inspectrice Lise* Meyer, *en charge de l'enquête, fait passer en revue tous les habitants du village possédant un véhicule, et s'avère qu'il y a bien un homme chauve possédant un kangoo jaune parmi eux : ce suspect, appelé* Gérald, *est en outre la seule personne du village à cocher toutes ces cases !*

Mais l'inspectrice est perplexe : en effet, même après vérification approfondie, le suspect n'a aucune raison connue d'en vouloir à la victime... L'inspectrice fait donc appel au cabinet de détectives consultants fondé par Sherlock Holmes, expert en science de la déduction, pour savoir si cela constitue une raison suffisante d'inculper le suspect. Vous êtes ingénieur·e détective dans ce cabinet : à vous de jouer ! ◡̈

Nous allons nous placer dans une situation idéalisée où :

- *Le témoignage de petit garçon est absolument fiable. En outre, on a la certitude absolue qu'on n'a pas cherché à aiguiller la police sur une fausse piste, et que le coupable est* réellement *un homme chauve possédant un kangoo jaune.*
- *Les notions de calvitie, de posséder un kangoo jaune, d'habiter le village, d'avoir un mobile, etc. sont définies sans aucune ambigüité : pour chaque personne, chaque caractéristique est vérifiée ou n'est pas vérifiée, de façon complètement binaire (il n'y a pas d'histoire de confusion entre les couleurs ou les modèles d'utilitaires, de « calvitie partielle », etc.).*

— *Il n'est pas possible d'établir à l'avance la liste des personnes possédant un mobile : car ce n'est généralement qu'après enquête approfondie sur un suspect qu'on découvre s'il a un mobile ou non !*

Vous avez choisi d'utiliser l'inférence bayésienne pour attaquer ce problème. Afin de vous préparer le terrain, l'inspectrice vous a fourni les informations suivantes :

— *Le village comporte* $640 =: h$ *habitants (victime et témoin exclus). Même si on ne sait pas lesquels de ces habitants possédaient un mobile, on peut supposer que cela ne concernait qu'une petite fraction d'entre eux.*

— *D'après l'expérience de l'inspectrice, dans ce genre d'affaire, dans une proportion* $30\ \% := p$ *des cas, le coupable habite le village, et dans une proportion* $90\ \% := q$ *des cas, il a un mobile.*

— *La probabilité qu'une personne prise au hasard soit un homme chauve conduisant un kangoo jaune, probabilité que nous noterons* ε*, est d'une sur* $2\,000$*.*

1. Selon vous, quelle probabilité à postériori sur la culpabilité du suspect justifiera son inculpation par l'inspectrice Meyer, resp. sa condamnation par le juge ? (en supposant, le cas échéant, que l'instruction judiciaire n'aura pas permis de mettre en lumière d'éléments supplémentaires pertinents).

Vous décidez de considérer que, dans votre raisonnement, la priore intègre déjà *le fait que vous sachiez qu'il y a eu un assassinat.*

2. Expliquer en quoi ce choix est pertinent.

Une fois n'est pas coutume, nous n'allons pas raisonner à coup de vraisemblance et de théorème de Bayes pour estimer la probabilité à postériori que Gérald soit coupable, mais plutôt travailler en termes de probabilités conditionnelles "classiques" : l'enjeu étant donc de savoir, sachant qu'une seule personne du village correspond aux critères énoncés par le témoin, mais qu'il n'a pas de mobile connu, quelle est la probabilité que ce soit lui l'assassin ?

Pour construire votre modèle probabiliste, vous décidez de considérer que, sous le contexte à priori, les évènements « l'assassin est du village » et « l'assassin possède un mobile » sont indépendantes.

3. À quel point ce choix vous semble-t-il justifié ? Expliquer dans quelle mesure il vous était difficile de ne pas faire ce choix dans tous les cas.

Vous décidez également de considérer que, sous le contexte à priori, le fait qu'une personne soit un homme chauve conduisant un kangoo jaune est indépendant du fait que cette personne habite le village et de ce qu'elle possède un mobile.

(4). Même question que précédemment : à quel point ce choix vous semble-t-il justifié ? Expliquer dans quelle mesure il vous était difficile de ne pas faire ce choix dans tous les cas.

5. Montrer que la probabilité à priori que, parmi l'ensemble des habitants du village, aucun ne soit un homme chauve possédant un kangoo jaune, resp. exactement un soit un homme chauve possédant un kangoo jaune, vaut approximativement resp. $e^{-\varepsilon h}$ et $\varepsilon h e^{-\varepsilon h}$.

(6). Quelles approximations avons-nous faites à la question précédente ? À quel point ces approximations vous semblent-elles raisonnables ?

7.❀ Déterminer la probabilité à priori qu'on ait simultanément les faits suivants : l'assassin est HCKJ ; il n'habite pas le village ; et un habitant un village exactement est HCKJ.

8.❀ Déterminer la probabilité à priori qu'on ait simultanément les faits suivants : l'assassin est HCKJ, il n'a pas de mobile, il habite le village, et aucun autre habitant du village n'est HCKJ.

9. En déduire la probabilité à postériori que Gérald soit l'assassin. Conclure.

10. Regarder comment la valeur numérique de la formule ci-dessus aurait été modifiée si on avait changé les valeurs de p, q, ε, h. Cela confirme-t-il votre intuition ? Cela a-t-il tendance à renforcer ou à amoindrir la pertinence que vous attribuez à ce modèle ?

11. De manière générale, nous savons néanmoins que tous les modèles sont faux, ou du moins, seulement approximatifs... De manière générale (pas forcément pour les valeurs spécifiques que p, q, ε, h de l'énoncé !), dans quel sens est-ce que la prise en compte du fait que tous les modèles sont faux devrait modifier votre croyance à postériori que le suspect soit l'assassin ?

Pour finir, nous voulons nous assurer que nous aurions pu arriver aux mêmes conclusions dans un cadre plus classique de statistique inférentielle (bayésienne). Observons qu'ici la question est simplement de savoir « quelle est la probabilité à postériori que Gérald soit l'assassin » ?

(12).★★ Retrouver les éléments de la question précédente à l'aide d'une "authentique" inférence bayésienne sur le paramètre caché d'un modèle statistique (j'entends par là, en reprenant le cadre formel du cours), que vous développerez. Observer les subtilités de la modélisation à bien prendre en compte.

Inférence statistique / Séance 5

Estimation & prédiction fréquentistes

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy *

28 avril 2025

EXERCICE 1 — Trois estimateurs, sinon rien !

Dans cet exercice, on observe les réalisations $x_{0\checkmark}, \dots, x_{(n-1)\checkmark}$ *de* n *v.a.i.i.d.* $\mathrm{Unif}^{\mathrm{me}}(]\alpha_\checkmark, \beta_\checkmark[)$*, avec* $\alpha_\checkmark, \beta_\checkmark$ *inconnus vérifiant* $-\infty < \alpha_\checkmark < \beta_\checkmark < \infty$*. Pour les applications numériques, on prendra* $n = 8$ *et*

$$\vec{x}_{]\!]0,8]\!]\checkmark} = (3{,}63, -0{,}06, 0{,}56, 1{,}74, 3{,}70, -0{,}44, 3{,}64, 3{,}91).$$

Le but de cet exercice est de calculer des estimateurs pour $\boldsymbol{\alpha}$*,* $\boldsymbol{\beta}$*, et* $\boldsymbol{\delta} \coloneqq \boldsymbol{\beta} - \boldsymbol{\alpha}$*. Dans un premier temps, nous allons utiliser l'estimation par maximum de vraisemblance.*

1. Calculer la vraisemblance de $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ au vu des observations ; en déduire les estimateurs du maximum de vraisemblance pour $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\delta}$, ainsi que leurs réalisations (donner les valeurs numériques de ces dernières).

Indication : Salon les conventions que vous aurez prises, il est possible que vous trouviez que le maximum de la vraisemblance n'est jamais atteint. Dans ce cas, on définira *quand même* un estimateur du maximum de vraisemblance en considérant la *limite* des "estimateurs de presque-maximum de vraisemblance". (Il sera clair de comprendre ce que la locution ci-devant signifie, le cas échéant, dans le contexte de cette question).

2. Observer qu'en l'occurrence, le maximum de vraisemblance pour $\boldsymbol{\delta}$ correspond à un estimateur empirique : dire en particulier quelle est la fonctionnelle à utiliser pour le voir comme tel.

Indication : Il est possible que vous ne connaissiez pas le nom de cette fonctionnelle : dans ce cas, expliquez simplement comment cette fonctionnelle est définie.

3. Montrer que l'estimateur du maximum de vraisemblance pour $\boldsymbol{\alpha}$ est biaisé vers le haut.

(4).* Montrer que, dans l'asymptotique $n \to \infty$, l'estimateur du maximum de vraisemblance pour $\boldsymbol{\alpha}$ est convergent.

Nous allons passer maintenant à la méthode des moments.

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

5.❀ Nous introduisons $f_1(\vec{X}) \coloneqq n^{-1}\sum_{i=0}^{n-1} X_i$ et $f_2(\vec{X}) \coloneqq n^{-1}\sum_{i=0}^{n-1} X_i^2$. Comment appelle-t-on ces quantités dans le jargon de la statistique ? (La réponse ne doit contenir aucun symbole mathématique).

6. Pour (α, β) quelconques dans l'espace du paramètre caché, calculer $\mathbb{E}_{\alpha,\beta}(f_1(\vec{X}))$ et $\mathbb{E}_{\alpha,\beta}(f_2(\vec{X}))$.

Indication : On pourra alléger les notations en recourant à l'astuce rédactionnelle consistant à faire le raisonnement pour les vraies valeurs $\alpha_✓$ et $\beta_✓$ (ce qui revient à se placer sous le véritable contexte probabiliste $\mathbb{P}_✓(\bullet)$), puis à revenir au cas de valeurs quelconques pour (α, β) seulement à la fin, en argüant que le calcul a été opéré sans se servir d'aucune information spécifique sur $(\alpha_✓, \beta_✓)$.

7. En déduire des quantités d'intérêt dont $f_1(\vec{X})$ et $f_2(\vec{X})$ sont respectivement des estimateurs par tendance. Observer qu'en l'occurrence, ces estimateurs par tendance peuvent aussi être vus comme des estimateurs empiriques.

8. Déduire de la question précédente des estimateurs par moments de $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\delta}$. Calculer numériquement les réalisations de ces estimateurs.

Indication : Il sera avisé de commencer par chercher un estimateur de $\boldsymbol{\alpha\beta}$, puis un estimateur de $(\boldsymbol{\beta}-\boldsymbol{\alpha})^2$.

9. Justifier que, lorsque $n \to \infty$, les v.a. $f_1(\vec{X})$ et $f_2(\vec{X})$ convergent en probabilité (sous le véritable contexte probabiliste $\mathbb{P}_✓$) vers resp. $(\alpha_✓ + \beta_✓)/2$ et $(\alpha_✓^2 + \alpha_✓\beta_✓ + \beta_✓^2)/3$; et en déduire que l'estimateur trouvé ci-dessus pour $\boldsymbol{\alpha}$ (que nous noterons $\hat{\boldsymbol{\alpha}}^{\mathrm{mom}}$) est convergent.

(10).★★ Montrer que l'estimateur par moments de $\boldsymbol{\alpha}$ trouvé ci-dessus est biaisé vers le haut.

11. En partant de la priore (impropre)

$$\mathbb{P}(\boldsymbol{\alpha} \in \mathrm{d}\alpha \text{ et } \boldsymbol{\beta} \in \mathrm{d}\beta) \propto \mathbf{1}_{\alpha<\beta}(\beta-\alpha)^{-2}\,\mathrm{vol}_1(\mathrm{d}\alpha)\,\mathrm{vol}_1(\mathrm{d}\beta),$$

calculer, selon la méthode bayésienne, la loi à postériori de $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (à constante multiplicative près) au vu des observations.

12.★ En déduire les lois à postériori (toujours à constante multiplicative près) de resp. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\delta}$ au vu des observations.

Indication : Pour alléger les calculs, il pourra être judicieux de poser $Min \coloneqq \inf\{X_i \mid i \in [\![0, n[\![\}$, resp. $Max \coloneqq \inf\{X_i \mid i \in [\![0, n[\![\}$.

Indication : On rappelle va formule de changement de variables multidimensionnel : si $\varphi : U \to V$ est une fonction bijective de régularité C^1 entre deux ouverts de $\mathbb{R}^n$ (avec la même dimension à la source et à la cible), alors pour $g : V \to \mathbb{R}$ une fonction intégrable, on a

$$\int_{y\in V} g(y)\,\mathrm{vol}_n(\mathrm{d}y) = \int_{x\in U} g(\varphi(x))|\det J_x(\varphi)|\,\mathrm{vol}_d(\mathrm{d}x),$$

où $J_x(\varphi)$ est la matrice jacobienne de φ en x, autrement dit la matrice des dérivées partielles des différentes coordonnées de φ, évaluées en x.

[Cela peut aussi d'interpréter en termes d'infinitésimaux : pour $\mathrm{d}x$ un voisinage infinitésimal de $x \in U$, si $\varphi : U \to \mathbb{R}^n$ est une fonction de régularité C^1 au voisinage de x, alors le volume de l'image directe de $\mathrm{d}x$ par φ est donné (au premier ordre) par

$$\mathrm{vol}_d(\varphi(\mathrm{d}x)) = |\det J_x(\varphi)| \, \mathrm{vol}_d(\mathrm{d}x)].$$

13. En déduire les médianes à postériori de $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, ainsi que la formule algébrique caractérisant la médiane à postériori de $\boldsymbol{\delta}$.

14.❀ En quoi les médianes calculées ci-dessus sont-elles bien des estimateurs de resp. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}, (\boldsymbol{\beta} - \boldsymbol{\alpha})$? En vous référant au polycopié, rappeler selon quel critère ces estimateurs peuvent être considérés comme optimaux.

EXERCICE 2 — Gros-Tony contre Docteur Jean

Trois amis se sont réunis devant la télévision pour regarder les championnats du monde de ski alpin, et plus précisément l'épreuve de descente. Ces trois amis sont le Docteur Jean*, éminent scientifique parfois un peu égaré dans l'abstraction de ses modèles, Gros-*Tony*, un* self-made-man *de l'immobilier, homme d'action et d'instinct*[*]*, et la Mère* Poulard*, célèbre spécialiste ès omelettes et biscuits, qui héberge les deux premiers à cette occasion. Les trois amis s'interrogent sur la performance que fera leur champion favori, Alexis* Pinturault*.*

Dans notre histoire, le championnat du monde de descente se déroule en deux manches[†]*, sur le même parcours ; en outre, il s'agit d'un parcours que les skieurs connaissent bien et sur lequel ils ont l'habitude de s'entrainer. L'analyse des compétitions sur la piste du championnat du monde montre que la performance d'un skieur donné se modélise bien de la façon suivante : le skieur a un niveau de performance $\theta_{\swarrow}$, fonction de son niveau, inobservable directement ; et le chrono qu'il fera lors de chaque manche est indépendant, suivant la loi* Normale$(\theta_{\swarrow}, \sigma^2)$*, où σ est connu est vaut* 1 s*. Nous notons respectivement X_0 et X_1 les temps du skieur pour chaque manche.*

Avant le début de la compétition, nos trois amis (que l'ambiance détendue du moment pousse fort logiquement à faire de la statistique ☆), se posent la question suivante : quelle est la meilleure façon de prédire le chrono de Pinturault à la seconde manche sachant son chrono à la première ? Autrement dit, ils se placent dans le cadre d'un modèle de prédiction où l'observation passée est X_0 et l'observation future X_1, et ils cherchent un prédicteur *pour X_1, sous la forme $\hat{X}_1 = \hat{x}_1(X_0)$. Notez que nos amis sont d'accord sur le fait que la fonction de perte qu'ils souhaitent utiliser pour jauger de la qualité d'un prédicteur est la fonction de perte quadratique : $\ell(x_1, \hat{x}_1) \coloneqq (\hat{x}_1 - x_1)^2$.*

Le Docteur Jean, Gros-Tony et la Mère Poulard argüent respectivement en faveur des prédicteurs suivants :

D$^{\mathrm{r}}$ Jean : *Pinturault a fait de nombreuses fois ce parcours à l'entrainement ; son chrono moyen lors des entrainements est de $x_{\mathrm{ref}} \coloneqq 90$ s. Le fait qu'il fasse mieux ou moins bien lors de la première manche est uniquement question de chance, et ne présage en rien de sa performance à la seconde manche : pour moi, le prédicteur le plus pertinent est donc la statistique constante $\hat{X}_1^{\mathrm{J}} \coloneqq x_{\mathrm{ref}}$.*

[*]. Les noms du Docteur Jean et de Gros-Tony sont inspirés de l'ouvrage *Le Cygne Noir* de Nassim Taleb.
[†]. Note sportive : Dans la vraie vie, le format en une seule manche est beaucoup plus fréquent ☠

Gros-Tony : *Pas d'accord ! Ce qui compte, ce n'est pas la moyenne des entrainements passés, mais les circonstances du jour : forme du skieur, état de la piste, effet du stress, … Il faut complètement jeter aux oubliettes tout ce que nous savions sur les habitudes de Pinturault et se focaliser uniquement sur les données du jour ! Dès lors, le meilleur prédicteur pour sa performance de la seconde manche est de reprendre la performance de la premère : je propose donc* $\hat{X}_1^{\mathrm{T}} \coloneqq X_0$.

Mère Poulard : *Eh bien moi, l'expérience des omelettes m'a appris qu'on gagnait toujours à mélanger le blanc et le jaune ! Je me place donc entre vos deux approches, et propose donc* $\hat{X}_1^{\mathrm{P}} \coloneqq (X_0 + x_{\mathrm{ref}})/2$. *Qui veut un biscuit ?…*

1.❀ Rappeler ce qu'est la fonction de risque d'un prédicteur, et donner la formule correspondante dans le contexte de cet énoncé.

2.❀ Rappeler aussi ce qu'est la fonction de biais d'un prédicteur, et donner la formule correspondante dans le contexte de cet énoncé.

Dans la suite, nous noterons $B_{\hat{X}_1}(\theta)$ *la fonction de biais d'un prédicteur* $\hat{X}_1$.

3. Démontrer, pour notre modèle, la formule suivante, appelée *décomposition biais-variance* : pour tout prédicteur $\hat{X}_1$ de X_1, on a l'égalité suivante entre fonctions de θ :

$$R_{\hat{X}_1}(\theta) = B_{\hat{X}_1}(\theta)^2 + \mathrm{Var}_\theta(\hat{X}_1) + \mathrm{Var}_\theta(X_1).$$

On pourra souligner au passage les hypothèses indispensables à la validité de cette formule. Interpréter ce que signifie cette formule quant aux contraintes que devra satisfaire un bon estimateur en termes de biais et de variance.

4. Calculer les fonctions de biais des prédicteurs respectifs de nos trois amis, ainsi que leurs fonctions de variance. Interpréter les forces et les faiblesses respectives des prédicteurs des uns et des autres à cette aune.

5. Calculer les fonctions de risques respectives de nos trois amis ; tracer le résultat pour $\theta \in [87\ \mathrm{s}, 93\ \mathrm{s}]$. Y a-t-il un prédicteur uniformément meilleur, ou pire, qu'un ou deux des autres ?

6.★ Calculer l'estimateur bayésien optimal en s'appuyant sur une priore de la forme $\mathrm{Normale}(x_{\mathrm{ref}}, \tau^2)$ (faire les calculs en laissant la valeur de τ en toutes lettres). En déduire que, dans un sens, le Docteur Jean, Gros-Tony et la Mère Poulard ont "tous les trois raison", au sens où les estimateurs qu'ils proposent sont bel et bien *intelligents*, modulo les hypothèses qu'on accepte...

Indication : Il sera utile d'utiliser la formule de densité de la loi normale (confer théorème 6.3.14 du polycopié) :

$$\mathbb{P}\big(\mathrm{Normale}(\mu, \sigma^2) \in \mathrm{d}x\big) = 1/\sqrt{2\pi}\,\sigma \times \exp\big(-(x-\mu)^2/2\sigma^2\big)\,\mathrm{vol}_1(\mathrm{d}x).$$

EXERCICE 3 — Note de gymnastique

Lors d'un concours de gymnastique, un jury de $5 =: J$ *juges est chargé d'estimer la qualité technique de la performance d'une compétitrice. Chaque juge doit évaluer cette performance par une note entre 0 et 10 (avec une précision de* 0,05 *point). On suppose que, par rapport à la performance* réelle *de la gymnaste, notée* $\mu_{\checkmark}$ *(évidemment inaccesible en tant que telle), les notes des juges, désignées par* $n_{0\checkmark}, \dots, n_{(J-1)\checkmark}$*, sont les réalisations de variables aléatoires i.i.d.* $N_0, \dots, N_{J-1}$ *suivant une loi de Laplace (voir l'annexe en fin de feuille sur* Arche*) de paramètre de position* $\mu_{\checkmark}$ *et de paramètre d'échelle* $\beta_{\checkmark}$ *(*$\beta_{\checkmark}$ *étant inconnu), la note de chaque juge étant indépendante.*

On donne pour les applications numériques :

$$\vec{n}_{\checkmark} = (7{,}15; 6{,}95; 7{,}10; 7{,}05; 6{,}35)$$

(1).❀ Formaliser le problème comme un modèle d'inférence statistique explicative : quel est le paramètre caché (« $\boldsymbol{\theta}$ ») et dans quel espace vit-il (« $\boldsymbol{\Theta}$ ») ; y a-t-il un paramètre de taille (« n ») ; quelle est l'observation (« X ») et dans quel espace vit-elle (« $\mathcal{X}^{(n)}$ ») ; et quelle est la loi de l'observation sachant le paramètre caché (« $\mathrm{Loi}_\theta(X^{(n)})$ ») ? Quel est, en l'occurrence, notre quantité d'intérêt (« $\gamma(\boldsymbol{\theta})$ ») ?

(2). Observer que notre modélisation ne saurait être une modélisation rigoureuse de la réalité, car elle souffre de deux défauts, liés respectivement à l'étendue des notes possibles et à la précision avec laquelle celles-ci sont données. Sous quelles conditions peut-on s'attendre à ce que ces défauts ne soient pas réellement gênants pour notre analyse statistique ?

3. Pour $\mu \in \mathbb{R}$, $\beta \in \mathbb{R}_+^*$, montrer que

$$\mathbb{P}\big(\mathrm{Laplace}(\mu, \beta) \in \mathrm{d}x\big) = \tfrac{1}{2}\beta^{-1}\mathrm{e}^{-|x-\mu|/\beta}\,\mathrm{vol}_1(\mathrm{d}x)\,;$$

et en déduire l'espérance et la variance de la loi $\mathrm{Laplace}(\mu, \beta)$.

4. Déduire de la question précédente des estimateurs empiriques pour $\boldsymbol{\mu}$ et pour $\boldsymbol{\beta}$.

5.★ Calculer une fonction de (log-)vraisemblance pour $(\boldsymbol{\mu}, \boldsymbol{\beta})$. Dessiner, à valeur de $\boldsymbol{\beta}$ fixée, l'allure de la fonction de log-vraisemblance pour $\boldsymbol{\mu}$: en quels points la pente de cette fonction change-t-elle, et quelle est la valeur de cette pente sur les différents intervalles ?

6. En déduire, pour une valeur fixée de $\boldsymbol{\beta}$, quel est le maximum de vraisemblance pour $\boldsymbol{\mu}$.

(7). En déduire quel est le maximum de vraisemblance "tout court" pour $(\boldsymbol{\mu}, \boldsymbol{\beta})$. Faire l'application numérique.

EXERCICE 4 — Sol invictus

*Nous sommes en 45 av. J.-C. Toute la Gaule est occupée par les Romains. Jules César, fort de la gloire dont l'a auréolé cette conquête, souhaite à présent laisser sa marque dans l'histoire en réformant le calendrier ! Une question essentielle est de déterminer la date appropriée pour la fête du soleil invaincu (*Sol invictus*), censée tomber le lendemain du solstice d'hiver... Or, déterminer la date du solstice est bien plus difficile qu'il n'y parait.*

Théoriquement, certes, c'est simple : c'est le jour de l'année auquel le Soleil se couche le plus au sud. Le problème, c'est qu'au voisinage de cet extrêmum, la direction du coucher de soleil est pratiquement constante… Sans compter qu'il n'est pas évident de déterminer précisément la direction de coucher du soleil, notamment à cause des effets de la réfraction atmosphérique, qui ne sont pas toujours les mêmes à cause de la météo !

Fort heureusement, dans la réalité alternative de cet exercice, les Romains ont déjà inventé la statistique ! ◡̈ Ils vont donc utiliser une technique permettant d'estimer précisément la date du solstice en dépit des difficultés de mesure[‡] *◡̈. L'idée centrale est d'observer la direction du coucher de soleil sur l'*ensemble *de la période décembre-janvier, d'en déduire la tendance régulière qui s'en dégage, et d'inférer la date du solstice à partir de cette tendance, sans se laisser perturber par le « bruit » qui parasite les observations.*

Le modèle mis au point par les astronomes romains comporte pas moins de six paramètres cachés. Le premier, $\boldsymbol{\eta}$*, est la date du solstice (repérée à partir du premier janvier), mesurée en jours, à valeurs dans* $\mathbb{R}$ [§]*. Les trois suivants,* $\boldsymbol{\gamma}_0$*,* $\boldsymbol{\gamma}_2$ *et* $\boldsymbol{\gamma}_3$*, donnent un développement limité de l'azimut*[¶] *"théorique" auquel on est censé voir se coucher le soleil au jour* t *: ils sont respectivement à valeurs dans* $\mathbb{R}$*,* $\mathbb{R}_+^*$ *et* $\mathbb{R}$*, et* $\boldsymbol{\gamma}_k$ *est mesuré en degrés d'angle par jour à la puissance* k*. Le deux derniers, enfin, notés* $\boldsymbol{\sigma}_{\mathrm{r}}$ *et* $\boldsymbol{\sigma}_{\mathrm{o}}$ *(tous les deux à valeurs dans* $\mathbb{R}_+^*$*, mesurés en degrés d'angle), sont relatifs aux déviations des mesures réelles par rapport à cet azimut "théorique" :* $\boldsymbol{\sigma}_{\mathrm{r}}$ *se référant aux effets de la réfraction atmosphérique (dépendant des conditions météorologiques), et* $\boldsymbol{\sigma}_{\mathrm{o}}$ *à l'imprécision des observateurs.*

L'observation est constituée, quant à elle, par le 62*-uplet* $(A_{-30}, A_{-29}, \dots, A_{31}) =: \vec{A}$*, où* A_t *dit à quel azimut on a vu se coucher le soleil au jour* t*. Sous le véritable contexte probabiliste* $\mathbb{P}_{\checkmark}$*, l'observation* $\vec{A}$ *est la loi normale* 62*-dimensionnelle suivante, où* $\boldsymbol{P}_{\checkmark}(T)$ *désigne le*

[‡]. En réalité, cet exercice n'est pas si anachronique que cela : si on fait abstraction de l'habillage statistique, en effet, la méthode est plus ou moins celle utilisée par Ptolémée dès le II$^{\mathrm{e}}$ siècle de l'ère commune !

[§]. On notera donc que le solstice n'est pas littéralement un *jour*, mais plutôt un *instant*. De fait, c'est ainsi qu'il est défini rigoureusement : le solstice est l'instant auquel la direction Soleil → Terre forme un angle minimal avec l'axe de rotation nord → sud de notre planète.

[¶]. Çàd. la direction, exprimée comme un angle par rapport au nord.

polynôme $\gamma_{0\checkmark} + \gamma_{2\checkmark} T^2 + \gamma_{3\checkmark} T^3$ *:*

$$\mathrm{Loi}_{\checkmark}(\vec{A}) = \mathrm{Normale}\left(\begin{pmatrix} P_{\checkmark}(-30-\eta_{\checkmark}) \\ P_{\checkmark}(-29-\eta_{\checkmark}) \\ P_{\checkmark}(-28-\eta_{\checkmark}) \\ \cdots \\ P_{\checkmark}(29-\eta_{\checkmark}) \\ P_{\checkmark}(30-\eta_{\checkmark}) \\ P_{\checkmark}(31-\eta_{\checkmark}) \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{r}\checkmark} & 0 & \cdots & 0 & 0 & 0 \\ \sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{r}\checkmark} & 0 & \ddots & 0 & 0 \\ 0 & \sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{r}\checkmark} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & 0 & \sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{r}\checkmark} & 0 \\ 0 & 0 & \ddots & 0 & \sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{r}\checkmark} \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2_{\mathrm{r}\checkmark} & \sigma^2_{\mathrm{o}\checkmark}+2\sigma^2_{\mathrm{r}\checkmark} \end{pmatrix} \right).$$

1. Essayez de vérifier si vous avez bien compris le modèle et son interprétation physique, en répondant aux questions suivantes :

(a) Quel est l'interprétation "physique" du paramètre $\boldsymbol{\gamma}_0$?

(b) Pourquoi n'y a-t-il pas de paramètre $\boldsymbol{\gamma}_1$?

(c) Pourquoi le paramètre caché $\boldsymbol{\gamma}_2$, contrairement à $\boldsymbol{\gamma}_0$ et $\boldsymbol{\gamma}_3$, vit-il dans $\mathbb{R}_+^*$?

(d) Les A_t sont-ils indépendants (sous la loi $\mathbb{P}_{\checkmark}$) ?

(e) Quelle interprétation physique peut-on donner à aux coefficients valant "$\sigma^2_{\mathrm{r}\checkmark}$" autour de la diagonale principale de la matrice des covariances de $\mathrm{Loi}_{\checkmark}(\vec{A})$?

On souhaite estimer $\boldsymbol{\eta}$ *par la méthode des moindres carrés.*

2. Écrire, la valeur de la fonction de constraste qu'il s'agira de minimiser, pour une valeur générique $(\eta, \gamma_0, \gamma_2, \gamma_3, \sigma_{\mathrm{r}}, \sigma_{\mathrm{o}})$ du paramètre caché (et pour l'observation effectivement réalisée).

Indication : La formule est un peu laide ; mais on essaiera néanmoins de se faire une idée suffisamment précise de la façon dont elle se développe pour pouvoir répondre aux questions suivantes.

3.$^{\bigstar}$ Au vu de l'allure de la fonction trouvée, mettre en évidence une contrainte concernant les quantités d'intérêt explicatives qu'on aura une chance, ou pas, de pouvoir estimer par moindre carrés. Constater que $\boldsymbol{\eta}$ satisfait effectivement cette contrainte, mais que d'autres quantités d'intérêt explicatives (lesquelles, par exemple ?) n'auraient pas pu être estimées par cette méthode.

4. Montrer que la fonction de contraste à optimiser (sachant la vraie valeur de l'observation) peut se mettre sous la forme

$$\begin{pmatrix}\gamma_0 & \gamma_2 & \gamma_3 & 1\end{pmatrix}\begin{pmatrix}P_{00}(\eta) & P_{02}(\eta) & P_{03}(\eta) & P_{0+}(\eta)\\ P_{20}(\eta) & P_{22}(\eta) & P_{23}(\eta) & P_{2+}(\eta)\\ P_{30}(\eta) & P_{32}(\eta) & P_{33}(\eta) & P_{3+}(\eta)\\ P_{+0}(\eta) & P_{+2}(\eta) & P_{+3}(\eta) & P_{++}(\eta)\end{pmatrix}\begin{pmatrix}\gamma_0\\ \gamma_2\\ \gamma_3\\ 1\end{pmatrix},$$

où les $P_{ij}(\bullet)$ sont des polynômes, avec $P_{ji} = P_{ij}$ pour tous i, j. Histoire de vérifier que vous avez bien compris, vous répondrez aux questions suivantes :

(a) Quels sont les polynômes dont l'expression fera intervenir, ou pas, la valeur de l'observation ?

(b) Quels sont les degrés des différents polynômes ?

(5). Démontrer le lemme suivant : si $\boldsymbol{Q}$ est une matrice symétrique définie positive de taille $d \times d$, $\vec{\Lambda}$ un vecteur colonne de dimension d, et c une constante, alors le minimum de l'application de $\mathbb{R}^d$ dans $\mathbb{R}$ donnée par

$$\vec{X} \mapsto \begin{pmatrix}\vec{X}^\top & 1\end{pmatrix}\begin{pmatrix}\boldsymbol{Q} & \vec{\Lambda}\\ \vec{\Lambda}^\top & c\end{pmatrix}\begin{pmatrix}\vec{X}\\ 1\end{pmatrix}$$

est égal à $c - \vec{\Lambda}^\top \boldsymbol{Q}^{-1}\vec{\Lambda}$.

6. En déduire que, à valeur fixée de η, le minimum atteint par la fonction de contraste est une certaine fraction rationnelle de η, dont les coefficients peuvent être déterminés à partir de l'observation effective. (On expliquera juste comment on *pourrait* calculer ces coefficients : il n'est pas demandé de déterminer manuellement les formules, qui seraient très fastidieuses...!).

7. Quel est le lien entre la fraction rationnelle évoquée ci-dessus et l'estimateur des moindres carrés pour $\boldsymbol{\eta}$?

En pratique, il y a certains jours pour lesquels, en raison de la couverture nuageuse, il n'est pas possible d'observer la direction du coucher du soleil... On a donc des données incomplètes : seuls certains A_t sont effectivement observés. Une idée pour pallier ce problème est de considérer l'ensemble des jours pour lesquels on pourra observer le coucher du soleil comme un paramètre du modèle — *quand bien même il s'agit en réalité de quelque chose d'aléatoire !*

(8).★ Donner un argument en vertu duquel ce choix de modélisation est pertinent, et un argument en vertu duquel cela pourrait néanmoins manquer de réalisme au niveau physique...

9. Qu'est-ce qui change par rapport aux questions 1 à 7 quand on se place dans ce nouveau cadre ?

(10).★ À l'aide d'un ordinateur, déterminer numériquement l'estimation des moindres carrés pour $\boldsymbol{\eta}$ pour les données du fichier `solstice.tsv`.

11.★ En quoi le problème aurait-il été nettement plus compliqué à résoudre si on avait voulu raisonner par maximum de vraisemblance plutôt que par moindres carrés ?

Annexe : La loi de Laplace

La *loi de Laplace* (également appelée « loi double-exponentielle ») *de paramètre de position* μ *et de paramètre d'échelle* β, que nous noterons ici $\text{Laplace}(\mu, \beta)$ est la distribution de probabilité sur $\mathbb{R}$ telle que :

(i) La loi est symétrique autour de μ : autrement dit, on a l'égalité de distributions $\text{Laplace}(\mu, \beta) - \mu = \mu - \text{Laplace}(\mu, \beta)$;

(ii) La probabilité de s'éloigner de μ de plus d'une certaine distance décroit exponentiellement, et ce avec le taux $1 / \beta$: autrement dit, $\mathbb{P}\big(|\text{Laplace}(\mu, \beta) - \mu| \geqslant x\big) = \mathrm{e}^{-x/\beta}$ pour tout $x \in \mathbb{R}_+$.

Inférence statistique / Séance 6

Tests d'hypothèses nulles

Énoncé

Formation d'Ingénieur Civil des Mines de Nancy *

12 mai 2025

☛ *Dans toute cette feuille de travaux dirigés, on suppose que, si* TelleLoi(α, β, γ) *est une loi de probabilité "classique" à valeurs réelles (spécifiée par certains paramètres* α, β, γ*), alors on dispose d'outils numériques permettant de calculer les « fonctions de tables » de cette loi, à savoir :*

- *La fonction de répartition de la loi, dont la valeur en* x *sera notée* `répartTelleLoi`$(\alpha, \beta, \gamma; x)$ *;*
- *La fonction de quantile de la loi, dont la valeur en* p *sera notée* `qtileTelleLoi`$(\alpha, \beta, \gamma; p)$ *;*
- *Si la loi est discrète, la fonction de masse de la loi, dont la valeur en* x *sera notée* `masseTelleLoi`$(\alpha, \beta, \gamma; x)$ *;*
- *Si la loi est à densité par rapport à la mesure de Lebesgue, la fonction de densité de la loi, dont la valeur en* x *sera notée* `densitéTelleLoi`$(\alpha, \beta, \gamma; x)$.

Par ailleurs, concernant les fonctions de répartition, on s'efforcera de privilégier des notations qui ne dépendent pas du choix de convention utilisé pour ces fonctions au niveau des points de discontinuité. J'entends par là que, si $F(\bullet)$ *est définie comme la fonction de répartition d'une loi de probabilité* $P \in \mathscr{M}_1(\mathbb{R})$*, pour* x *est un atome de* P *[i.e., un point tel que* $\mathbb{P}(P = x) > 0$*], on évitera d'écrire « *$F(x)$ *» (qui serait ambigu) : dans un tel cas, soit on veut se référer à* $\mathbb{P}(P < x)$ *et on écrira «* $F(x-)$ *», soit on veut se référer à* $\mathbb{P}(P \leqslant x)$ *et on écrira alors «* $F(x+)$ *».*

EXERCICE 1 — Test d'une pièce

On souhaite procéder à un test statistique (fréquentiste) pour savoir si une pièce de monnaie est équilibrée (i.e., si elle retombe aussi souvent sur "pile" que sur "face"). Pour ce faire, on lance la pièce $360 =: n$ *fois, et on regarde si le nombre de "pile" obtenus (quantité notée* S*) est compatible avec ce qu'on attend d'une pièce équilibrée.*

1. Modéliser le problème sous forme statistique. On dira qui est le paramètre caché, qui est l'observation et quelle est sa loi pour une valeur donnée du paramètre caché, d'une part ; quelles sont l'hypothèse nulle et la statistique de test utilisée, d'autre part.

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

Indication : Attention, il se peut que certaines de ces quantités n'aient pas été introduites clairement par l'énoncé : à vous de les mettre en évidence dans ce cas !

2. Quel sera le sens de suspicion à utiliser concernant notre statistique de test ?
Indication : On ne demande pas ici de calculs, juste une explication informelle.

3.❀ Soit θ une valeur possible pour le paramètre caché. (On n'impose pas que θ appartienne à l'hypothèse nulle ni à l'hypothèse alternative). Quelle est la loi de la statistique de test sous la probabilité $\mathbb{P}_\theta$?

On souhaite tester l'hypothèse nulle au risque de 8 %. La procédure standard expliquée par le cours nous conduit alors à rejeter l'hypothèse nulle si S *est inférieure (au sens large) à* 162 *ou supérieure à* 198*, et à l'accepter si elle est comprise (au sens large) entre* 163 *et* 197*.*

4. Quelle(s) fonction(s) de table(s), avec quels arguments, avons-nous évaluée(s) pour arriver à ces valeurs ? (On indiquera non seulement le(s) calcul(s) à faire, mais aussi le(s) résultat(s) que, au vu de l'énoncé, ce(s) calcul(s) est censé avoir donné(s)).

5. En pratique, on a observé $s_\checkmark := 197$ "pile", de sorte que l'hypothèse nulle est acceptée. Écrire une phrase de conclusion retranscrivant ce verdict en langage ordinaire, d'une façon qui utilise aussi peu de jargon que possible, mais qui explique néanmoins ce que signifie notre résultat.

6. À l'aide des fonctions de tables appropriées, écrire la formule permettant de calculer le niveau *exact* du test que nous avons défini.

7. *Sans* exécuter numériquement les calculs de la question précédente, le niveau réel du test sera-t-il strictement inférieur, rigoureusement égal, ou strictement supérieur à 8 % ? (Justifier).
Indication : On admettra qu'on n'est pas en présence ici d'un cas "pathologique".

8. Par une approche similaire à celle de la question 6, on pourrait montrer que, si la pièce a une probabilité de 55 % de tomber sur "pile", la probabilité qu'on conclue que la pièce est déséquilibrée vaut 52,2 % environ. Reformuler cela en dans le jargon statistique. Commenter la valeur numérique trouvée en ce qu'elle montre clairement la dissymétrie entre le traitement de hypothèse nulle et cela de l'hypothèse alternative par la théorie des tests.

Les calculs à faire pour délimiter la zone d'acceptation et de rejet requérant des logiciels spécialisés en statistique, on va chercher un test plus simple, basé sur l'idée que l'intervalle de fluctuation pour la loi normale, au niveau de confiance 92 %*, est légèrement plus petit que* $[-1{,}76, +1{,}76]$.

9.❀ Rappeler pourquoi, lorsque n tend vers l'infini, on a

$$\mathrm{Loi}\left(2 \times \frac{\mathrm{Binom}^{\mathrm{le}}(n, 1/2) - n/2}{n^{1/2}}\right) \to \mathrm{Normale}(0, 1).$$

10.* En déduire que, si on prend pour zone d'acceptation

$$[n/2 - 0{,}88\, n^{1/2}, n/2 + 0{,}88\, n^{1/2}],$$

on définira un test asymptotiquement de niveau 8 %.

11.* Démontrer que le test de la question précédente sera consistant.

EXERCICE 2 — Le complot des choses contre les êtres

Dans une enquête des Dingodossiers [*] *intitulée « méchantes choses », Réné Goscinny & Marcel Gotlib font la constatation suivante, qui prouve selon eux la méchanceté des objets inanimés envers nous autres pauvres humains :*

1. Expliquer en quoi le texte de la vignette (modulo la nuance précisée dans l'indication ci-dessous) s'apparente à la philosophie du test statistique d'hypothèse nulle.

Indication : Attention ; dans le texte de la vignette, le mot « toujours » devra être compris comme une hyperbole signifiant en réalité « dans l'écrasante majorité des cas ». D'autre part, il faut considérer ici que les enquêteurs des *Dingodossiers* ne disposent, pour étayer leur affirmation, que de quelques dizaines de témoignages à tout casser : ils ne sont donc pas en mesure de donner des chiffres *précis* sur la probabilité d'être soumis à la malédiction de l'ultime boulon rétif...

Nous allons maintenant faire la modélisation statistique suivante. On suppose qu'on a interrogé $7 \coloneqq n$ *automobilistes à plat, dont les pneus comportaient respectivement* 3, 5, 4, 4, 6, 4 *et* 5 *boulons (valeurs notées resp.* $b_0, \dots, b_{n-1}$*). On a demandé à chacun de ces automobilistes combien de boulons ils avaient eu des difficultés à desserrer : ces quantités sont notées resp.* $X_0, \dots, X_{n-1}$.

Dans notre modèle, l'hypothèse nulle que nous souhaitons tester est le fait que le nombre de boulons qui se desserrent difficilement est, justement, toujours le fruit du hasard. On note Θ_0 *l'espace du paramètre caché correspondant à cette hypothèse nulle. Dans ce cas, le*

[*]. Vous ne connaissez pas les *Dingodossiers* ?! Vous ratez quelque chose... Foncez les lire dès la fin du TD ! ☺

paramètre caché est composé de six valeurs $\pi_0, \dots, \pi_{n-1} \in [0,1]$, π_i *décrivant la probabilité que, sur une roue du modèle de l'automobiliste* i*, un boulon donné se desserre difficilement. Nous ne chercherons pas à modéliser ce qui se passe sous l'hypothèse alternative, notée* Θ_1 *: nous dirons simplement que cette hypothèse alternative correspond à l'ensemble de cas où les objets font exprès d'être méchants envers les humains.*

Nous prenons pour statistique de test $T \coloneqq \sum_{i=0}^{n-1} \mathbf{1}_{X_i = 1}$.

2. Quel sera le sens de suspicion approprié pour cette statistique de test ?

3. À supposer qu'on soit sous l'hypothèse nulle et que le paramètre caché vaille $(\pi_0, \dots, \pi_{n-1})$, exprimer la loi de la statistique de test comme une somme de v.a. de Bernoulli indépendantes (mais pas forcément de même loi), dont on précisera les paramètres.

(4).❀ Pour $k \in \mathbb{N}^*$, démontrer que la fonction définie sur $[0,1]$ par $x \mapsto kx(1-x)^{k-1}$ a pour maximum $(1-1/k)^{k-1}$.

Pour la question suivante, on note $\vec{\pi}_!$ *la valeur, au sein de* Θ_0*, pour laquelle* $\mathrm{Loi}(T \mid \vec{\pi} = \vec{\pi}_!)$ *"prend les plus grandes valeurs". (On admettra qu'on peut en l'occurrence donner un sens précis, et conforme à l'intuition, à cette locution*[†]*).*

5. En vertu de la question 3, nous savons que, sous le contexte $\mathbb{P}_{\vec{\pi}_!}(\bullet)$, la statistique de test T s'écrit comme une somme de variables de Bernoulli indépendantes[‡], dont nous noterons les paramètres respectifs $q_0^!, q_1^!, \dots, q_{n-1}^!$. Exprimer la probabilité $\mathbb{P}_{\vec{\pi}_!}(T > n-2)$ en fonction des $q_i^!$.

6. Montrer que, pour les paramètres du modèle qu'on considère, le niveau du test lorsqu'on décide de rejeter l'hypothèse nulle pour $\{T > 5\}$ vaut numériquement $1{,}6 \times 10^{-2}$ (arrondi supérieurement).

7. Si les enquêteurs ont choisi de fixer le seuil de leur test à 5, à quelle conclusion arriveront-ils si, parmi les sept automobilistes interrogés, 6 ont été confrontés au boulon maudit ? Au vu du niveau du test pour un tel seuil, à quel point leur conclusion pourra-t-elle être considérée comme forte ?

EXERCICE 3 — Homoscédasticité (poil au nez)

En modélisation statistique, on voit généralement les données observées comme la somme d'un terme de tendance et d'un terme de "bruit" : par exemple, dans un jeu de données dont l'objectif serait de comprendre pourquoi certaines femmes sont plus grandes ou plus petites

[†]. Pour ceux que cela intéresse, sur le plan technique, l'idée de « prendre les plus grandes valeurs » peut être formalisée rigoureusement par la notion de « dominance stochastique » : et en l'occurrence, nous sommes dans un cas où, parmi les lois $\mathrm{Loi}_{\vec{\pi}}(T)$ pour $\vec{\pi} \in \Theta_0$, il y en a une qui domine stochastiquement toutes les autres.

[‡]. Qui sont nécessairement au nombre de n, vu les valeurs pouvant être prises par T

que d'autres, on pourrait modéliser la taille d'une femme comme valant 0,50 *fois la taille de sa mère, plus* 0,42 *fois la taille de son père, plus un terme de « bruit » modélisant tous les autres facteurs indépendants de la taille des parents qui interviennent et qu'on ne sait (ou qu'on ne veut) pas préciser.*

*Dans ce genre de contexte, pour permettre une analyse statistique efficace, une hypothèse fréquemment faite est celle d'*homoscédasticité *: cela consiste à dire que le bruit a le même comportement pour toutes les sous-populations ou pour tous les individus. Dans l'exemple sur la taille des femmes, on pourrait par exemple imaginer que le terme de bruit soit plus important lorsque les parents d'une femme sont de tailles très différentes (auquel cas la fille pourrait aussi bien être très petite que très grande) : l'hypothèse d'homoscédasticité consiste alors à supposer que ce n'est pas le cas.*

Les hypothèses d'homoscédasticité sont très pratiques, mais elles ont le mauvais gout d'être facilement mises en défaut… C'est pourquoi il est important d'être capable de tester de telles hypothèses. C'est ce que fait cet exercice, à l'aide du test dit de Fisher-Snedecor.

On considère une étude s'intéressant à la façon dont la température corporelle des gens varie d'un individu à l'autre et d'une population à l'autre : pour ce faire, on sélectionne $20 =: n_{\mathrm{J}}$ *individus japonais pris au hasard et* $30 =: n_{\mathrm{L}}$ *individus libanais pris au hasard ; et, par un protocole précis, on détermine la température moyenne au repos de chacun de ces individus. Ces températures sont vues comme les réalisations de variables aléatoires* $J_0, \dots, J_{n_{\mathrm{J}}-1}$ *(pour les Japonais) et* $L_0, \dots, L_{n_{\mathrm{L}}-1}$ *(pour les Libanais), toutes ces variables aléatoires étant supposées indépendantes, avec*

$$\begin{cases} J_i \hookrightarrow \mathrm{Normale}(\mu_{\mathrm{J}\checkmark}, \sigma^2_{\mathrm{J}\checkmark}) & \text{pour } i \in [\![0, n_{\mathrm{J}}[\![\,; \\ L_i \hookrightarrow \mathrm{Normale}(\mu_{\mathrm{L}\checkmark}, \sigma^2_{\mathrm{L}\checkmark}) & \text{pour } i \in [\![0, n_{\mathrm{L}}[\![. \end{cases}$$

L'hypothèse d'homoscédasticité revient alors à dire que $\sigma_{\mathrm{J}\checkmark}$ *et* $\sigma_{\mathrm{L}\checkmark}$ *sont égaux.*

1. Quel est le paramètre caché, et dans quel espace vit-il ? À quel sous-ensemble de l'espace du paramètre caché l'hypothèse nulle correspond-elle ? L'hypothèse nulle est-elle simple ? Rappeler les difficultés techniques que soulève le fait d'avoir une hypothèse nulle composite.

La théorie des probabilités montre que, pour le modèle considéré, on a :

$$\mathrm{Loi}_{\checkmark}\left(\frac{\mathrm{var}_{\mathrm{B}}(J_0, \dots, J_{n_{\mathrm{J}}-1}) \,/\, \sigma^2_{\mathrm{J}\checkmark}}{\mathrm{var}_{\mathrm{B}}(L_0, \dots, L_{n_{\mathrm{L}}-1}) \,/\, \sigma^2_{\mathrm{L}\checkmark}} \right) = \mathrm{F}_{\mathrm{Sn}}(n_{\mathrm{J}} - 1, n_{\mathrm{L}} - 1), \qquad (*)$$

où $\mathrm{var}_{\mathrm{B}}(x_0, \dots, x_{n-1})$ *désigne ce qu'on appelle la* variance bessélisée *d'un jeu de données réelles, qui s'obtient en divisant sa variance empirique par* $1 - 1/n$ *(*n *étant la taille du jeu de données) ; et* $\mathrm{F}_{\mathrm{Sn}}(n_{\mathrm{J}} - 1, n_{\mathrm{L}} - 1)$ *désigne* la loi F de Snedecor [§] *à* $n_{\mathrm{J}} - 1$ *(numérateur) et* $n_{\mathrm{L}} - 1$ *(dénominateur) degrés de libertés : les lois de Snedecor étant une famille de lois indexée par* $(\mathbb{N}^*)^2$ *(dont nous vous passons la définition), dont les tables sont implémentées dans tous les bons logiciels de statistique.*

[§]. Cette loi a été inventée par George W. Snedecor, qui a appelé cette loi 'F' en hommage à Ronald A. Fisher : pour cette raison, elle est également appelée « loi de Fisher », ou « loi de Fisher-Snedecor ».

2.* Déduire de la formule précédente une statistique pour tester l'hypothèse d'homoscédasticité. Quelle(s) loi(s) cette statistique suit-elle sous l'hypothèse nulle ? (On prendra soin de préciser si la loi suivie est la même quelle que soit la modalité de l'hypothèse nulle, ou s'il y a des variations). Quel est le sens de suspicion approprié ?

3. Donner les formules (à l'aide de « fonctions de tables ») disant quelles valeurs de la statistique de test doivent conduire à l'acceptation, resp. au rejet de l'hypothèse nulle, si on se fixe un niveau de risque à 8 %.

Indication : Voici, pour la suite de l'exercice, les valeurs seuils qu'on trouverait numériquement : 0,459 647 9 et 2,045 852 3.

4. Pour notre jeu de données, on a $\mathrm{var}_{\mathrm{B}}(j_{0\checkmark}, \dots, j_{(n_{\mathrm{J}}-1)\checkmark}) = 0{,}017\,114\,74$, resp. $\mathrm{var}_{\mathrm{B}}(\ell_{0\checkmark}, \dots, \ell_{(n_{\mathrm{L}}-1)\checkmark}) = 0{,}032\,725\,17$. Faire l'application numérique, et écrire la phrase de conclusion appropriée.

5. Le test que nous avons effectué était-il asymptotique ou exact ? Dans une perspective asymptotique, ce test serait-il consistant, à votre avis ? (on ne demande pas une preuve formelle sur ce dernier point).

(6).** Démontrer rigoureusement votre réponse à la question précédente concernant la consistance (ou la non-consistance) du test.

Indication : Le point central consistera à démontrer que, lorsque n_{J} tend vers l'infini, on a

$$\mathrm{Loi}_{\checkmark}\big(\mathrm{var}_{\mathrm{B}}(\boldsymbol{J}_0, \dots, \boldsymbol{J}_{n_{\mathrm{J}}-1})\big) \to \delta_{\sigma^2_{\mathrm{J}\checkmark}}$$

(et similairement concernant les Libanais). Encore faut-il savoir comment le démontrer, et comment en déduire le résultat...

EXERCICE 4 — Des gouts et des couleurs

Une société de marketing se demande s'il existe une corrélation entre les préférences des gens en matière de couleurs et d'orientation politique : si cela s'avérait exact, on pourrait s'en servir pour proposer des publicités ciblées... La société fait donc appel à un institut de sondage qui interroge un panel aléatoire de $500 =: n$ *Français pour leur demander, d'une part, la tendance politique à laquelle ils s'identifient le mieux entre les adjectifs « conservateur », « libéral » et « socialiste » (chaque orientation sera repérée par son initiale), et d'autre part, leur couleur préférée parmi « bleu », « jaune », « rouge » ou « vert » (chaque couleur sera là encore repérée par son initiale). Les résultats sont résumés par le tableau suivant :*

	$j = \mathrm{B}$	$j = \mathrm{J}$	$j = \mathrm{R}$	$j = \mathrm{V}$	total
$i = \mathrm{C}$	$n_{\mathrm{CB}\checkmark} = 67$	$n_{\mathrm{CJ}\checkmark} = 26$	$n_{\mathrm{CR}\checkmark} = 26$	$n_{\mathrm{CV}\checkmark} = 55$	$n_{\mathrm{C}\checkmark} = 174$
$i = \mathrm{L}$	$n_{\mathrm{LB}\checkmark} = 43$	$n_{\mathrm{LJ}\checkmark} = 25$	$n_{\mathrm{LR}\checkmark} = 25$	$n_{\mathrm{LV}\checkmark} = 25$	$n_{\mathrm{L}\checkmark} = 118$
$i = \mathrm{S}$	$n_{\mathrm{SB}\checkmark} = 57$	$n_{\mathrm{SJ}\checkmark} = 33$	$n_{\mathrm{SR}\checkmark} = 44$	$n_{\mathrm{SV}\checkmark} = 74$	$n_{\mathrm{S}\checkmark} = 208$
total	$n_{\mathrm{B}\checkmark} = 167$	$n_{\mathrm{J}\checkmark} = 84$	$n_{\mathrm{R}\checkmark} = 95$	$n_{\mathrm{V}\checkmark} = 154$	$n = 500$

Au niveau de la modélisation du problème, le paramètre caché est constitué par le 12-uplet des valeurs $(\pi_{\mathrm{CB}\checkmark}, \pi_{\mathrm{CJ}\checkmark}, \dots, \pi_{\mathrm{SV}\checkmark}) =: \theta_\checkmark$ *qui représentent la probabilité qu'une personne aléatoire ait, pour couple de préférences* (*politique*, *couleur*), *les valeurs* (conservateur, bleu), …, (socialiste, vert). *Les sondés sont numérotés de* 0 *à* $n-1$, *et pour* $k \in [\![0, n[\![$, *on désigne par* Pol_k *l'orientation politique de l'individus* k *et par* $Coul_k$ *sa couleur favorite.*

On utilise également les notations suivantes :

- $\boldsymbol{\pi}_{\mathrm{C}}, \boldsymbol{\pi}_{\mathrm{L}}, \boldsymbol{\pi}_{\mathrm{S}}$ *désignent respectivement la probabilité (vue comme variable aléatoire) qu'une personne aléatoire s'identifie comme conservatrice, resp. libérale, resp. socialiste. On introduit de même les notations* $\boldsymbol{\pi}_{\mathrm{B}}$, $\boldsymbol{\pi}_{\mathrm{J}}$, *etc. pour les couleurs.*
- *Le nombre d'orientations politiques sera noté* $p := 3$, *et le nombre de couleurs sera noté* $q := 4$. *Ces paramètres du modèle seront comme fixés dans nos analyses asymptotiques.*
- *On notera resp.* $\mathscr{P} := \{\mathrm{C}, \mathrm{L}, \mathrm{S}\}$ *et* $\mathscr{Q} := \{\mathrm{B}, \mathrm{J}, \mathrm{R}, \mathrm{V}\}$ *l'ensemble des orientations politiques et des couleurs. Les éléments de ces ensembles seront génériquement désignés, si besoin, par les indices respectifs* i *et* j. *Ainsi, la définition des* $\boldsymbol{\pi}_{\mathrm{C}}$, $\boldsymbol{\pi}_{\mathrm{L}}$ *et* $\boldsymbol{\pi}_{\mathrm{S}}$ *ci-dessus peut se réécrire en disant que*
$$\forall i \in \mathscr{P} \quad \boldsymbol{\pi}_i := \sum_{j \in \mathscr{Q}} \boldsymbol{\pi}_{ij}.$$
- *On notera* $n_{\mathrm{C}\checkmark}$, $n_{\mathrm{J}\checkmark}$, $n_{\mathrm{LR}\checkmark}$, *etc. le nombre de sondés ayant exprimé une opinion conservatrice, resp. une préférence pour la couleur jaune, resp.* à la fois *une opinion libérale et une préférence pour le rouge, etc. Ces quantités seront vues comme les réalisations des v.a.* N_{C}, N_{J}, N_{LR}, *etc.*

(1). Formellement, $\boldsymbol{\theta}$ vit dans un espace de dimension pq ; cependant, ses composantes sont soumises à certaines contraintes... Quelle est la véritable dimension de l'espace du paramètre caché ; autrement dit, quel est le nombre de paramètres réels indépendants permettant de caractériser θ ?

(2). Pour $(i, j) \in \mathscr{P} \times \mathscr{Q}$, quelle est la loi (sous $\mathbb{P}_\checkmark$) de N_{ij} ? En déduire, sous réserve que $\pi_{ij\checkmark} > 0$, une expression de la loi-limite, lorsque $n \to \infty$, de $(N_{ij} - n\pi_{ij\checkmark})^2 / n\pi_{ij\checkmark}$ (qu'on pourra écrire comme la mesure-image d'une loi classique).

3. L'hypothèse nulle qu'on souhaite tester est le fait que les préférences en matière de couleurs et de politique sont indépendantes. Formaliser cette hypothèse en disant à quel sous-espace $\boldsymbol{\Theta}_0$ de l'espace du paramètre caché (l'espace tout entier du paramètre caché étant quant à lui noté '$\boldsymbol{\Theta}$') elle correspond. L'hypothèse nulle est-elle simple ?

(4).* Quelle est la dimension de $\boldsymbol{\Theta}_0$? (autrement dit, de combien de nombres réels a-t-on besoin pour caractériser un point de $\boldsymbol{\Theta}_0$). En déduire que, du point de vue intuitif, l'hypothèse nulle regroupe "plusieurs conditions à la fois" sur le paramètre caché.

(5).** Supposons que tous les $\pi_{i\checkmark}$ et tous les $\pi_{j\checkmark}$ soient strictement positifs (au passage, justifier la naturalité de cette hypothèse). Montrer que dans ce cas, si $\theta_\checkmark \in \boldsymbol{\Theta}_0$, on a, lorsque $n \to \infty$:
$$\mathrm{Loi}_\checkmark\left(\sum_{(i,j)\in\mathscr{P}\times\mathscr{Q}} \frac{(N_{ij} - n\pi_{i\checkmark}\pi_{j\checkmark})^2}{n\pi_{i\checkmark}\pi_{j\checkmark}}\right) \overset{n\to\infty}{\to} \chi^2(11),$$

où $\chi^2(11)$ est la loi de la somme des carrés de 11 variables normale standard indépendantes : $\chi^2(11) \coloneqq \bigoplus_{d=1}^{11} \text{Normale}(0,1)^2$.

6. À l'inverse, montrer que si $\theta_{\checkmark} \notin \Theta_0$, on s'attend à ce que les valeurs de $\sum_{(i,j)\in\mathscr{P}\times\mathscr{Q}} \frac{(N_{ij}-n\pi_{i\checkmark}\pi_{j\checkmark})^2}{n\pi_{i\checkmark}\pi_{j\checkmark}}$ seront (sous la loi $\mathbb{P}_{\checkmark}$), pour n grand, en général beaucoup plus grandes que les valeurs typiques de la loi $\chi^2(11)$.

7.❀ Pourquoi est-ce que la variable aléatoire

$$\sum_{(i,j)\in\mathscr{P}\times\mathscr{Q}} \frac{(N_{ij} - n\boldsymbol{\pi}_i\boldsymbol{\pi}_j)^2}{n\boldsymbol{\pi}_i\boldsymbol{\pi}_j}$$

ne peut *pas* servir de statistique de test pour notre hypothèse nulle ?

8. Quels sont les estimateurs empiriques pour les $\boldsymbol{\pi}_i$ et les $\boldsymbol{\pi}_j$? En déduire qu'une statistique de test pertinente pourrait être

$$\sum_{(i,j)\in\mathscr{P}\times\mathscr{Q}} \frac{(N_{ij} - N_i N_j / n)^2}{N_i N_j / n} =: T.$$

On peut montrer que, si $\theta_{\checkmark} \in \Theta_0$, la loi de la statistique T ne tend pas vers la loi $\chi^2(pq-1)$, mais vers la loi $\chi^2\big((p-1)(q-1)\big)$, qui tend à prendre des valeurs légèrement plus petites.

(9).★ Pourquoi le point ci-dessus était-il prévisible ?

10. Grâce à la loi-limite pour T établie à la question précédente, on est en mesure de construire un test pour l'hypothèse d'indépendance entre les opinions politiques et les préférences en matière de couleur. Quel sera le sens de suspicion ?

11. Le test sera-t-il asymptotique ou exact ? Faut-il s'attendre à ce que le test soit consistant ?

12. Donner les formules, à partir de fonctions de tables, qui permettraient de faire l'application numérique sur notre jeu de données (les valeurs sont données en indication). On prendra un niveau de risque de 5 %. Écrire la phrase de conclusion appropriée.

Indication : Numériquement, le quantile à calculer vaudra 12,591 59.

13.❀ En général, on considère que l'approximation asymptotique de la loi de T par la loi $\chi^2\big((p-1)(q-1)\big)$ donne des résultats assez fiables (du moins pour des calculs au niveau de risque 5 %) dès lors qu'on a, pour tous $(i,j) \in \mathscr{P}\times\mathscr{Q}$, $n\hat{\pi}^{\text{emp}}_{i\checkmark}\hat{\pi}^{\text{emp}}_{j\checkmark} \geqslant 5$. Cette condition est-elle vérifiée ici ?

Inférence statistique / Séance 7

Familles de tests : p-valeurs ; intervalles fréquentistes

Énoncé

Formation d'Ingénieur Civil des Mines de Nancy *

19 mai 2025

EXERCICE 1 — La gouteuse de thé

Nous sommes en Angleterre, en 1922. M^rs Muriel Bristol*, une éminente biologiste, affirme à son ami Ronald* Fisher*, un non moins éminent mathématicien, qu'elle préfère le goût du thé quand le lait est versé en premier dans la tasse. M^r Fisher est cependant fort sceptique quant à cette affirmation : il lui semble physiquement impossible qu'on puisse distinguer gustativement les deux situations... Mais, en bon scientifique, il décide de procéder à une expérience pour y voir plus clair ! Ce que M^rs Bristol, en tout aussi bonne scientifique, accepte bien volontiers*[*] ⌣

Nos amis décident du protocole suivant. Il demandent M^r Roach *(un ami commun, également présent sur place), de préparer 8 tasses de thé : quatre avec le lait versé en premier, et quatre avec le fait versé en dernier. Seul M^r Roach sait quelle tasse a suivi quel ordre de préparation. Puis M^r Fisher apporte les tasses à M^rs Bristol, qui a pour mission d'identifier correctement les quatre tasses où le lait a été versé en premier. En fonction du nombre d'erreurs commises par M^rs Bristol, M^r Fisher et M^r Roach seront alors en mesure d'être réellement convaincus, ou pas, par l'affirmation de leur amie !*

1. À ce stade de l'énoncé, nous somme d'ores et déjà en mesure de dire quelles sont les hypothèses complémentaires que nous souhaitons confronter (même si nous ne savons pas encore dire qui sera l'hypothèse nulle et qui sera l'hypothèse alternative), quelle sera l'observation du modèle statistique, quelle sera la statistique de test, et quel sera le sens dans lequel nous ferons pencher notre conclusion en faveur de telle ou telle hypothèse. Précisez donc tout cela !

Nous allons considérer le modèle statistique suivant : il existe une valeur $\theta_{\checkmark} \in \mathbb{R}_+^*$, inobservable directement, décrivant la "finesse" du palais de M^rs Bristol. Lorsque M^rs Bristol

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

[∗]. Cette anecdote est authentique ! R. Fisher est, en fait, l'inventeur de la notion de test d'hypothèse nulle ; et dans son ouvrage *The Design of Experiments*, c'est à partir de cette expérience vécue qu'il expose, pour la première fois, la notion en question ! ⌣

goute une tasse avec le lait en premier, ses neurones gustatifs produisent un signal distribué selon une loi Normale$(\theta_{\checkmark}, 1)$ *; tandis que, lorsqu'elle goute une tasse avec le lait en dernier, ses neurones produisent un signal selon une loi* Normale$(0, 1)$. *(L'aléa dans la façon dont ces lois sont réalisées étant indépendant d'une tasse à l'autre). Avec ce modèle, le fait que* M^{rs} *Bristol ait la capacité de distinguer que le lait a été versé en premier se traduit donc par le fait que* $\theta_{\checkmark}$ *soit non nul. Lors de l'épreuve,* M^{rs} *Bristol étiquètera comme « lait en premier » les quatre tasses ayant produit le signal le plus élevé sur ses neurones.*

(2).** Déterminer numériquement la loi du nombre d'erreurs commises par M^{rs} Bristol en fonction de la valeur du paramètre caché (appelant $\boldsymbol{E}$ la variable aléatoire représentant le nombre d'erreurs, on tracera plus précisément les courbes représentant $\theta \mapsto \mathbb{P}_\theta(\boldsymbol{E} \geqslant 2)$, $\theta \mapsto \mathbb{P}_\theta(\boldsymbol{E} \geqslant 4)$ et $\theta \mapsto \mathbb{P}_\theta(\boldsymbol{E} \geqslant 6)$).

La question ci-dessus est indiquée comme très difficile ; et de fait, je n'attends pas de vous que vous la traitiez ! ☺ *Néanmoins, un point essentiel est justement qu'il n'y aura pas besoin de traiter cette question pour mener à bien notre procédure de test d'hypothèse...*

3. Déterminer la loi du nombre d'erreurs $\boldsymbol{E}$ sous l'hypothèse où M^{rs} Bristol n'a en réalité aucune capacité à distinguer l'ordre de versement du lait.

4. Comment allons-nous, grâce au seul résultat de la question précédente (sans utiliser la question 2, donc ! ☺), pouvoir mettre à l'épreuve les capacités de M^{rs} Bristol ? Expliquer en quoi la dissymétrie des rôles que les tests d'hypothèse fréquentistes font jouer à l'hypothèse nulle et à l'hypothèse alternative s'avèrera cruciale dans cette optique.

5. Selon que M^{rs} Bristol commettra 0, 2, 4 ou 6 erreurs, montrer que la p-valeur de notre test vaudra resp. 1,5 %, 25 %, 76 % ou 98,6 %.

6. Que convient-il de conclure dans chacun de ces quatre cas décrits ci-dessus ?

7. Sans faire de cclcul supplémentaire, expliquer, si on avait plutôt procédé par test booléen, comment nous aurions défini notre test pour un niveau de risque de 20 %, resp. de 5 %, resp. de 1 % ?

EXERCICE 2 — La lutte des classes

Dans cet exercice, on se demande si la formation initiale suivie avant Mines Nancy a un impact sur les performances des élèves à l'examen de M. Peyre. Plus précisément, on se demande s'il y a une différence statistiquement significative entre les résultats des "matheux" (élèves issus de CPGE MP ou MPI, ou d'une licence de mathématiques) et les "physiciens" (tous les autres), de quelque nature que ce soit[†]*. Pour les applications numériques éventuelles, on utilisera les données du fichier* `lutte.tsv`.

[†]. On peut imaginer en effet que les matheux soient plus à l'aise avec l'intense formalisme probabiliste du cours, mais on peut aussi imaginer que les physiciens soient moins gênés par l'importance que ce cours accorde à la modélisation et aux raisonnements informels...

Le modèle est le suivant. Il y a n élèves dans la promotion considérée, qu'on peut diviser en n_{M} matheux et n_Φ physiciens. On note $\mathscr{E}_{\mathrm{M}} =: \{\mu_0, \dots, \mu_{n_{\mathrm{M}}-1}\}$ l'ensemble des matheux, resp. $\mathscr{E}_\Phi =: \{\varphi_0, \dots, \varphi_{n_\Phi-1}\}$ l'ensemble des physiciens ; et l'ensemble de tous les élèves est noté $\mathscr{E} := \mathscr{E}_{\mathrm{M}} \sqcup \mathscr{E}_\Phi$; on notera aussi ce dernier ensemble $\{e_0, \dots, e_{n-1}\}$, avec $e_i := \mu_i$ pour $i \in [\![0, n_{\mathrm{M}}[\![$ et $e_{n_{\mathrm{M}}+i} := \varphi_i$ pour $i \in [\![0, n_\Phi[\![$. La note de l'élève e est notée $z_{e\checkmark}$, vue comme la réalisation d'une v.a. Z_e. On note $\vec{Z}_{\mathrm{M}} := (Z_\mu)_{\mu\in\mathscr{E}_{\mathrm{M}}}$, $\vec{Z}_\Phi := (Z_\varphi)_{\varphi\in\mathscr{E}_\Phi}$, $\vec{Z} := (Z_e)_{e\in\mathscr{E}} = (\vec{Z}_{\mathrm{M}}, \vec{Z}_\Phi)$. Toutes les variables Z_e sont indépendantes, sachant que Z_e suit la distribution de probabilité

$$\begin{cases} 20 \times \mathrm{BêtaDéc}(\alpha_{\mathrm{M}\checkmark}, \beta_{\mathrm{M}\checkmark}, \lambda_{\mathrm{M}\checkmark}) =: P_{\mathrm{M}\checkmark} & \text{pour } e \in \mathscr{E}_{\mathrm{M}}, \text{ resp.} \\ 20 \times \mathrm{BêtaDéc}(\alpha_{\Phi\checkmark}, \beta_{\Phi\checkmark}, \lambda_{\Phi\checkmark}) =: P_{\Phi\checkmark} & \text{pour } e \in \mathscr{E}_\Phi, \end{cases}$$

où, pour $(\alpha, \beta, \lambda) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}$, la loi bêta décentrée de paramètres de forme α et β, avec décentrage λ, *notée* $\mathrm{BêtaDéc}(\alpha, \beta, \lambda)$, *est une certaine distribution de probabilité à densité sur $]0, 1[$, dont ni la définition ni les propriétés ne nous intéresseront ici.*

1. Expliciter la façon dont le texte ci-dessus mathématise, dans le cadre de la statistique inférentielle, une situation "industrielle", en répondant aux questions suivantes :

(a) Que représentent, de façon plus parlante, les objets $P_{\mathrm{M}\checkmark}$ et $P_{\Phi\checkmark}$?

(b) En quoi la description ci-dessus fournit-elle bien un modèle statistique (explicatif, fréquentiste) : quels sont les paramètres du modèle, le paramètre caché, l'observation, et la loi de l'observation sachant le paramètre caché ?

(c) Avec nes notations du modèle, comment se formalise l'hypothèse nulle qu'on souhaite tester ? (et qui exprime, en l'occurrence, le fait qu'il n'y ait pas de différence statistique entre les résultats des matheux et ceux des physiciens). S'agit-il d'une hypothèse nulle simple ou composite ?

Notre stratégie de test va reposer sur la statistique de Mann-Whitney, *définie par*

$$U := \sum_{(\mu,\varphi)\in\mathscr{E}_{\mathrm{M}}\times\mathscr{E}_\Phi} \mathbf{1}_{Z_\varphi > Z_\mu}.$$

On supposera dans la suite qu'il ne peut pas y avoir d'ex-æquo (ce qui est le cas en particulier dès lors que les distributions $P_{\mathrm{M}\checkmark}$ et $P_{\Phi\checkmark}$ sont sans atomes), de sorte que $\mathbf{1}_{Z_\varphi > Z_\mu} = 1 - \mathbf{1}_{Z_\mu > Z_\varphi}$.

Le lemme de Wilcoxon *affirme que, si $P_{\mathrm{M}\checkmark} = P_{\Phi\checkmark}$, alors U suit une distribution appelée* distribution U de Wilcoxon [‡] de paramètres n_{M} et n_Φ, *notée* $\mathrm{Wilc}^{\mathrm{xn}}(n_{\mathrm{M}}, n_\Phi)$, *dont la plupart des logiciels équipés d'un module de statistiques savent calculer les tables.*

2. Quel sera notre critère de suspicion concernant cette statistique ? (Motiver votre réponse).

Sur le jeu de donnees qui nous intéresse [§], *on a $n_{\mathrm{M}} = 55, n_\Phi = 109$; et la réalisation de $u_\checkmark$ la statistique de Wilcoxon vaut* 3 077.

[‡]. Ou parfois « Distribution de Mann-Whitney ».

[§]. Pour ceux que cela intéresse de traiter les données par eux-mêmes, j'ai mis le fichier de données en sur *Arche* : `lutte.tsv`.

3. Écrire, en termes de fonction de tables, la formule fournissant l'expression de la p-valeur du test.

4. L'application numérique donne 0,884 983 3. Que convient-il d'en conclure ?

5. Notre test était-il asymptotique ou non asymptotique ?

On considère généralement que, dès lors que $n_1, n_2 \geqslant 20$, la loi $\mathrm{Wilc}^{\mathrm{xn}}(n_1, n_2)$ *est raisonnablement bien approchée par la loi normale d'espérance $n_1 n_2/2$ et de variance $\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)$ (qui sont les espérance et variance de la véritable loi de Wilcoxon). Plus précisément, on peut démontrer le résultat suivant :*

$$\sup_{\substack{n_0, n_1 \geqslant m \\ x \in \mathbb{R}}} \left| \texttt{répartWilc}^{\mathrm{xn}}(n_0, n_1; x) - \texttt{répartNormale}\left(\tfrac{1}{2} n_0 n_1, \tfrac{1}{12} n_0 n_1 (n_0 + n_1 + 1); x\right) \right| \overset{m \to \infty}{\to} 0.$$

Ce résultat permet en particulier, si on ne dispose pas d'un logiciel implémentant les tables des lois de Wilcoxon, de procéder à une analyse statistique approximative en recourant à l'approximation par une loi normale (dont, pour le coup, les tables sont implémentées d'une manière ou d'une autre dans quasiment tous les logiciels).

(6). En utilisant la table de la loi normale standard fournie en annexe (on procèdera, si nécessaire, par interpolation linéaire entre les valeurs tabulées), regarder ce qu'aurait été notre p-valeur avec une telle approximation. Commenter.

7.* Démontrer que le test qui consiste à recourir à l'approximation normale est également valable en tant que test *asymptotique*, dans un régime asymptotique qu'on précisera.

(8).** Argüer qu'il est possible que $(\alpha_{\mathrm{M}\checkmark}, \beta_{\mathrm{M}\checkmark}, \lambda_{\mathrm{M}\checkmark}) \neq (\alpha_{\Phi\checkmark}, \beta_{\Phi\checkmark}, \lambda_{\Phi\checkmark})$, mais que pourtant notre test échoue à mettre cette différence en valeur, même pour des valeurs arbitrairement grandes de n_{M} et n_{Φ}.

9. Exprimer de façon plus concise, à l'aide du jargon de la statistique, la phrase : « Il est possible que $(\alpha_{\mathrm{M}\checkmark}, \beta_{\mathrm{M}\checkmark}, \lambda_{\mathrm{M}\checkmark}) \neq (\alpha_{\Phi\checkmark}, \beta_{\Phi\checkmark}, \lambda_{\Phi\checkmark})$, mais que pourtant notre test échoue à mettre cette différence en valeur, même pour des valeurs arbitrairement grandes de n_{M} et n_{Φ} ». Pourquoi ce phénomène n'est-il pas si grave en pratique ?

EXERCICE 3 — Accidents d'avion

Un avionneur lance une enquête de fiabilité sur le système électronique un modèle d'avion commercial, le Z377*, qu'il a récemment lancé sur le marché. En un total de* 25 920 h $=: T$ *de vol, on a signalé* $6 =: z_{\checkmark}$ *pannes électroniques de l'avion (heureusement sans conséquences tragiques* ☺*). Pour son analyse statistique, la compagnie considère que les pannes se produisent avec un certain taux $\lambda_{\checkmark}$ inconnu, et que le nombre de pannes pendant la période testée était la réalisation d'une variable aléatoire de loi* $\mathrm{Poisson}(\lambda_{\checkmark} T)$*. À partir de cette information, on cherche à déterminer un intervalle de confiance à* 90 % *pour la valeur de $\boldsymbol{\lambda}$.*

(1). Justifier le choix de modélisation.

2.❀ Dans cette question, on suppose pour fixer les idées que $\lambda_{\checkmark}$ est égal à $14\,/\,T$, valeur qui vaut numériquement 54×10^{-5} h^{-1} (i.e. 54 pannes par cent-mille heures de vol) et que nous noterons λ_0 dans la suite. Quel est, dans ce cas, la loi du nombre de pannes sur la durée T? À l'aide du graphe de la loi de Poisson donné en annexe, donner l'intervalle de fluctuation, à 90 % de confiance, de cette loi.

3. En déduire la façon la plus naturelle de tester, au risque 10 %, l'hypothèse $\{\boldsymbol{\lambda} = \lambda_0\}$.

4. L'observation qu'il y a eu 6 pannes est-elle compatible, au risque 10 %, avec l'hypothèse que $\boldsymbol{\lambda}$ vaut λ_0 ? Que peut-on en déduire concernant l'intervalle de confiance sur $\boldsymbol{\lambda}$? (Étant entendu ici qu'on parle de l'intervalle de confiance construit à partir de la statistique « nombre de pannes »).

5. Soit $\lambda \in \mathbb{R}_+^*$ une valeur quelconque. En vous appuyant sur les questions précédentes, donner les conditions que doit satisfaire λ pour être situé dans (la réalisation, au vu de ce qu'on a observé 6 pannes, de) l'intervalle de confiance pour $\boldsymbol{\lambda}$. Ces conditions seront écrites en faisant intervenir la fonction de quantile d'une certaine loi classique (pour laquelle nous n'aurons pas de formule fermée simple, mais que nos logiciels sauraient calculer sans souci ! ☺).

(6).★ Reformuler les conditions obtenues à la question précédente en faisant cette fois-ci intervenir, non pas une certaine fonction de quantile, mais une certaine fonction de *répartition*. (On veillera à écrire le dernier argument de la fonction de répartition d'une façon qui ne laisse aucune ambigüité quant au comportement de ladite fonction de répartition aux points de discontinuité).

7. Justifier informellement que notre intervalle de confiance sera effectivement un intervalle, et calculer ses bornes, en vous appuyant sur les éléments donnés en annexe.

Indication : Si la partie « Justifier que ce sera effectivement un intervalle » vous perturbe, passez directement à la seconde partie de la question.

(8). Pour quelle valeur $\lambda_{\checkmark}$ l'espérance (sous la vraie loi) du nombre de pannes observées aurait-elle précisément été de 6 ? Lier cette valeur avec les concepts vus en cours.

(9). Calculer également l'estimateur du maximum de vraisemblance pour $\boldsymbol{\lambda}$.

Indication : On donne $\mathbb{P}(\text{Poisson}(\theta) = k) = \theta^k \mathrm{e}^{-\theta} \,/\, k!$.

(10). Comparer les résultats des questions 7 et 8-9, et souligner les points qui vous semblent en valoir la peine.

EXERCICE 4 — Bâoum !

☛ *Bien que le contexte de cet exercice soit évidemment inspiré du développement de l'arme atomique par les États-Unis au cours de la seconde guerre mondiale, il n'y a à ma connaissance eu aucune telle interrogation statistique dans la réalité historique !* ☺

Au cours d'une guerre, les physiciens d'un certains pays sont parvenus à mettre au point une nouvelle bombe, extrêmement onéreuse, mais potentiellement extrêmement puissante. Les caractéristiques de la bombe font qu'on ne peut pas prévoir à l'avance quelle sera l'énergie (et donc les dégâts) délivrée par l'explosion de la bombe : tout au plus parait-il raisonnable de supposer que cette énergie suivra une loi log-normale $\mathrm{LogNorm}(\mu_{\checkmark}, \sigma_{\checkmark})$ *de paramètres inconnus. (On "rappelle" que la loi* $\mathrm{LogNorm}(\mu, \sigma)$ *est la mesure-image de la loi* $\mathrm{Normale}(\ln \mu, \sigma^2)$ *par la fonction exponentielle). Les physiciens sont arrivés à mettre au point trois prototypes (de conception identique) de la bombe, qui ont été testés dans le désert, et dont les énergies ont été resp. de* 21, 15 *et* 21 kt *équivalent TNT.*

L'état-major, voyant que les résultats d'une explosion à l'autre sont irréguliers, souhaite savoir dans quel intervalle on peut raisonnablement s'attendre à ce que se situe l'énergie d'une éventuelle nouvelle bombe avant d'en commander la fabrication. C'est à cette question que cet exercice va s'intéresser, à l'aide de la technique de l'intervalle de prédiction. Sauf mention explicite du contraire, nous utiliserons pour les différents intervalles statistiques un niveau de confiance de 94 %.

Questions préliminaires

(1). Mathématiser le problème : quel est le paramètre caché (qu'on notera $\boldsymbol{\theta}$), et dans quel espace vit-il (qu'on notera $\boldsymbol{\Theta}$) ; quelle est l'observation passée (qu'on notera $\boldsymbol{X}$) (on précisera aussi sa réalisation, notée $x_{\checkmark}$), quelle est l'observation future (qu'on notera $\boldsymbol{Y}$), et quelle est la loi du couple de ces observations ? On précisera au passage quelle est la quantité d'intérêt prédictive qui constitue l'objet de l'étude statistique, qu'on notera $g(\boldsymbol{Y})$ dans la suite de l'énoncé.

Pour la suite de l'exercice, nous aurons besoin de divers lemmes :

Lemme 1. *Pour* $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $n \geqslant 2$, *soient* $(X_i)_{1 \leqslant i \leqslant n}$ *des v.a.i.i.d.* $\mathrm{Normale}(\mu, \sigma^2)$.

Alors la variable aléatoire

$$\frac{\mathrm{moy}(X_i)_i - \mu}{\mathrm{var}_{\mathrm{emp}}^{1/2}(X_i)_i}$$

suit la loi

$$\mathrm{T}_{\mathrm{St}}(n-1) \,/\, \sqrt{n-1},$$

où $\mathrm{T}_{\mathrm{St}}(n-1)$ *désigne la « loi t de Student à* $(n-1)$ *degrés de libertés », dont les tables peuvent être calculées par tout logiciel de statistique digne de ce nom.*

Lemme 2. *Pour $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $n \geqslant 2$, soient $(X_i)_{1 \leqslant i \leqslant n}$ des v.a.i.i.d.* Normale(μ, σ^2). *Alors la variable aléatoire*

$$\frac{\mathrm{var}_{\mathrm{emp}}(X_i)_i}{\sigma^2}$$

suit la loi

$$\chi^2(n-1)\,/\,n,$$

où $\chi^2(n-1)$ désigne la « loi du chi-deux[¶] *à $(n-1)$ degrés de liberté », dont les tables peuvent être calculées par tout logiciel de statistique digne de ce nom.*

Lemme 3. *Pour $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $n \geqslant 2$, soient $(X_i)_{1 \leqslant i \leqslant n}$ et X' des v.a.i.i.d.* Normale(μ, σ^2). *Alors la variable aléatoire*

$$\frac{X' - \mathrm{moy}(X_i)_i}{\mathrm{var}_{\mathrm{emp}}^{1/2}(X_i)_i}$$

suit la loi

$$\sqrt{\frac{n+1}{n-1}} \times \mathrm{T}_{\mathrm{St}}(n-1).$$

2. Si $\hat{\boldsymbol{\mu}}$ et $\hat{\boldsymbol{\sigma}}$ sont des estimateurs (par maximum de vraisemblance, mettons) pour resp. $\boldsymbol{\mu}$ et , expliquer pourquoi l'intervalle de fluctutation (à 94 %, s'entend) de la loi LogNorm($\hat{\mu}_{\checkmark}$, $\hat{\sigma}_{\checkmark}$) devrait être *trop étroit* pour constituer (la réalisation de) un intervalle de prédiction (à 94 %) pour $g(Y)$: qualitativement parlant, qu'aurions-nous "oublié" de prendre en compte ?

3. Expliquer pourquoi est-ce qu'un intervalle de confiance (toujours à 94 %) sur $\boldsymbol{\mu}$ devrait, là encore, être centré à peu près convenablement par rapport à $g(Y)$, mais présenter néanmoins une fenêtre trop étroite pour en être un intervalle de prédiction : cette fois-ci, qu'aurion-nous "oublié" de prendre en compte ?

(4).* Expliquer (très informellement) pourquoi, néanmoins, le plus larges des deux intervalles ci-dessus devrait être "pas trop loin" de constituer un intervalle de fluctuation.

(5).* À l'aide des lemmes ci-dessus et des tables numériques fournies en annexe, calculer les réalisations numériques des intervalles évoqués aux questions numéro 3 et 2.

Indication : Concernant la partie « estimation par maximum de vraisemblance », on se réfèrera avec profit à l'exemple (`XM`) du polycopié.

Indication : Les valeurs numériques à trouver sont resp. $[13{,}9, 25{,}4]$ kt et $[12{,}1, 29{,}1]$ kt.

Après avoir vu ci-dessus trois stratégies incorrectes pour déterminer l'intervalle de confiance (même si la dernière ne devrait pas être "trop fausse" en général), voyons à présent deux stratégies réellement correctes. La première de ces deux stratégies, qui est assez intuitive [et qu'on adopte souvent "naïvement" quand on ne fait pas attention], sera malheureusement peu efficace ; la seconde, en revanche [qui est celle décrite par le cours], sera réellement pertinente ! ☺

[¶]. Prononcer « ki-deux ».

Première stratégie : Intervalle de prédiction par combinaison d'intervalles

La troisième stratégie donne un véritable intervalle de prédiction, mais elle passe à côté de plusieurs optimisations importantes.

6. En utilisant le lemme 2, donner (la réalisation de) une majoration à 94 % de confiance pour $\boldsymbol{\sigma}$.

7.* En utilisant l'intervalle de confiance pour $\boldsymbol{\sigma}$ calculé ci-dessus et celui pour $\boldsymbol{\mu}$ calculé à la question 5, et en utilisant les propriétés de fluctuation de la loi normale standard (voir l'annexe), en déduire un intervalle de prédiction pour $g(\boldsymbol{Y})$.

8. Expliquer qu'en réalité, nous ne sommes capables de prouver la fiabilité de l'intervalle de prédiction obtenu qu'avec une confiance de 82 %, et non pas de 94 %. Refaire les calculs pour obtenir, par la même méthode, un intervalle de prédiction à 94 % de confiance. Commenter le résultat obtenu.

Seconde stratégie : Intervalle via une variable de loi fixe

9. À l'aide du lemme 3, trouver une fonction $f(\boldsymbol{X}, g(\boldsymbol{Y}))$ de l'observation passée et de la quantité d'intérêt, dont la loi, qu'on saura tabuler à l'aide de l'annexe, ne dépende pas de la valeur du paramètre caché. Dire qui est la loi en question, que nous noterons $\boldsymbol{P}_f$ dans la suite de l'énoncé.

10. À l'aide de l'annexe, calculer l'intervalle de fluctuation à 94 % pour la distribution de probabilité $\boldsymbol{P}_f$ trouvée à la question précédente.

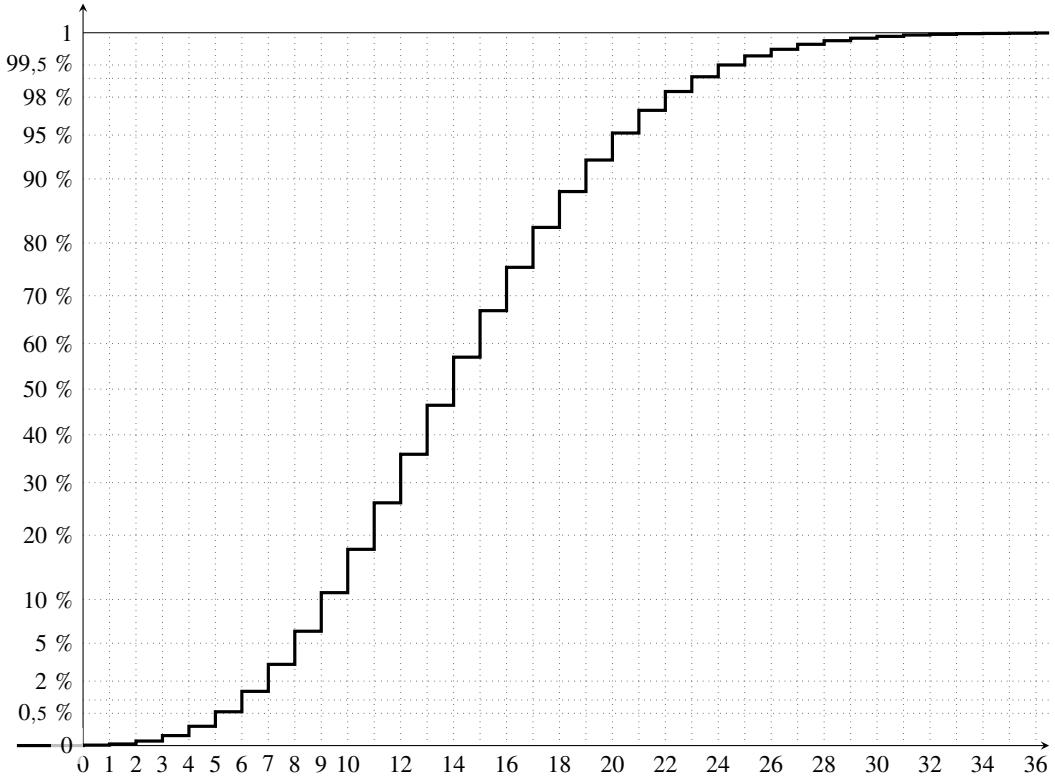
11. En déduire un intervalle de prédiction à 94 % pour $g(\boldsymbol{Y})$. (On donnera la réponse sous forme d'intervalle aléatoire).

12. Donner la réalisation de l'intervalle de prédiction en question pour les observations (passées) effectives indiquées dans l'énoncé. Commenter le résultat obtenu.

Annexes

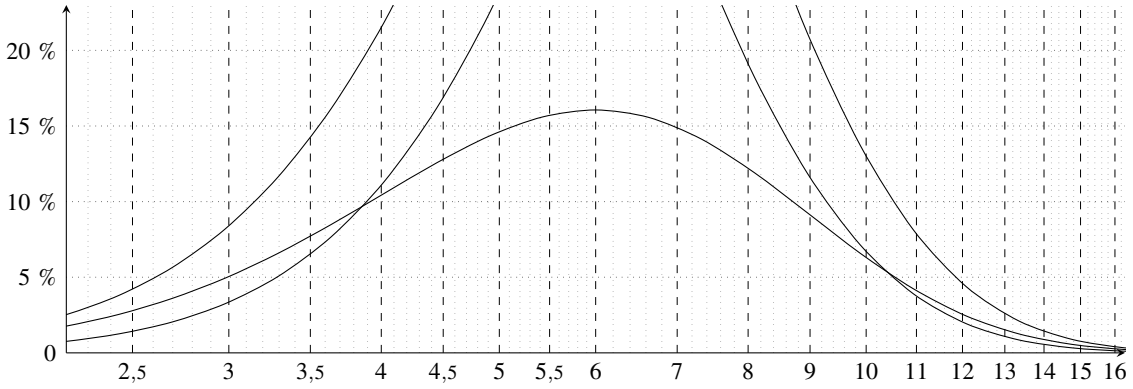
Loi de Poisson de paramètre 14

Le graphique ci-dessous représente la fonction de répartition de la loi Poisson(14). Attention, la graduation de l'axe vertical n'est pas linéaire.



Fonctions de répartition et de masse de la loi de Poisson en fonction de son paramètre

Le diagramme ci-dessous montre comment évoluent les probabilités qu'une loi Poisson(θ) soit resp. strictement inférieure, inférieure ou égale, égale, supérieure ou égale, ou strictement supérieure à 6, selon la valeur de θ (je n'ai pas indiqué quelle courbe correspond à quel cas, car cela peut se retrouver par des considérations élémentaires ☺). Seules les valeurs suffisamment faibles figurent sur les tracés.



Quantiles des lois normales, Student et χ^2

On donne ci-dessous quelques quantiles de la loi normale standard, de la loi t de Student à 2 degrés de liberté, de la loi du chi-deux à deux degrés du liberté, et de l'opposé de cette dernière [|||]. Par ailleurs, les lois Normale$(0,1)$ et $\mathrm{T}_{\mathrm{St}}(2)$ sont symétriques par rapport à zéro : elles sont donc leurs propres opposées, ce qui permet de déduire les quantiles de niveaux $\geqslant 1/2$ des quantiles de niveau $\leqslant 1/2$.

Toutes les valeurs données sont arrondies par défaut

α	$\texttt{qtileNormale}(0,1;\alpha)$	$\texttt{qtileT}_{\texttt{St}}(2;\alpha)$	$\texttt{qtile}\chi^2(2;\alpha)$	$-\texttt{qtile}\chi^2(2;1-\alpha)$
1 %	$-2{,}327$	$-6{,}965$	$0{,}020$	$-9{,}211$
1,5 %	$-2{,}171$	$-5{,}643$	$0{,}030$	$-8{,}400$
2 %	$-2{,}054$	$-4{,}849$	$0{,}040$	$-7{,}825$
3 %	$-1{,}881$	$-3{,}897$	$0{,}060$	$-7{,}014$
4 %	$-1{,}751$	$-3{,}320$	$0{,}081$	$-6{,}438$
6 %	$-1{,}555$	$-2{,}621$	$0{,}123$	$-5{,}627$
9 %	$-1{,}341$	$-2{,}027$	$0{,}188$	$-4{,}816$
12 %	$-1{,}175$	$-1{,}654$	$0{,}255$	$-4{,}241$
18 %	$-0{,}916$	$-1{,}178$	$0{,}396$	$-3{,}430$
24 %	$-0{,}707$	$-0{,}861$	$0{,}548$	$-2{,}855$
36 %	$-0{,}359$	$-0{,}413$	$0{,}892$	$-2{,}044$

Fonction de répartition de la loi normale standard

La table ci-dessous indique, pour toutes les nombres allant de 0 à 3,99 par incréments de 0,01, ce que vaut la fonction de répartition de la loi normale standard en chacun de ces nombres. Par exemple, la valeur $\texttt{répartNormale}(0,1;1{,}23)$ se trouve à l'intersection de la ligne « 1,20 » et de la colonne « 0,03 », et est donc égale à 0,890 65. On rappelle par ailleurs

[|||]. Notez que le quantile le niveau α de la loi $-P$ est la même chose que l'opposé du quantile de niveau $1-\alpha$ de la loi P, ce qui explique l'intitulé de la dernière colonne dans le tableau de données.

que la loi normale standard est diffuse, de sorte que qu'il n'y a pas d'ambigüité concernant la convention de continuité de sa fonction de répartition ; et qu'elle est symétrique par rapport à l'origine, de sorte que $\texttt{répartNormale}(0,1;-x) = 1 - \texttt{répartNormale}(0,1;x)$.

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,10	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,20	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,30	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,40	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,50	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,60	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,70	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,80	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,90	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,00	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,10	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,20	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,30	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,40	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,50	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,60	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,70	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,80	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,90	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,00	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,10	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,20	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,30	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,40	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,50	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,60	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,70	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,80	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,90	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,00	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,10	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,20	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,30	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,40	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,50	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,60	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,70	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,80	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,90	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Inférence statistique / Séance 8

Du bon usage de la statistique fréquentiste

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy*

26 mai 2025

EXERCICE 1 — L'affaire Sally Clark

En 1998, M^me Sally Clark*, citoyenne britannique âgée de 33 ans, fut arrêtée pour meurtre après qu'on se fut aperçu que ses deux enfants étaient tous les deux décédés quelques semaines à peine après leur naissance (le premier en 1996, le second en 1998), soi-disant de la mort subite du nourrisson. (La mort subite du nourrisson est le décès brutal et inexpliqué d'un enfant de moins d'un an dans son sommeil, possiblement lié au fait que le réflexe respiratoire n'est pas encore bien mis en place à cet âge). Aucun autre élément à charge ne témoignait contre elle, sinon cette "coïncidence" suspecte.*

Lors du procès de M^me Clark, le procureur argüa de la façon suivante. Sachant que, à l'époque, environ un bébé sur $8\,500$ *décédait de mort subite de nourrisson, la probabilité qu'une femme perde ses deux enfants par ce mécanisme était d'environ une sur* $7{,}3 \times 10^7$*. Vu cette très faible probabilité, l'affirmation de M^me Clark que ses deux enfants étaient morts naturellement n'était pas crédible, et il y avait donc lieu de la condamner, au-delà de tout doute raisonnable.*

1. Formaliser le raisonnement du procureur en termes statistiques.

(2).★ Comment pourrait-on lier le nombre $7{,}3 \times 10^7$ à une question de vraisemblance ?

3. Mettre en doute la valeur de $7{,}3 \times 10^7$ avancée par le procureur.

4. Identifier la faiblesse globale dans le raisonnement du procureur, et montrer qu'il n'y avait en fait pas lieu de condamner M^me Clark sur la seule foi de ce raisonnement.

☛ *Historiquement, le raisonnement du procureur a convaincu le jury, et M^me Clark fut condamnée pour meurtre, aussi bien en première instance qu'en appel. Ce n'est qu'après que des statisticiens eurent eu vent de l'affaire et eurent critiqué l'argument du procureur dans les médias* [*] *que le procès*

*Équipe pédagogique : Rémi Peyre, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

[*]. ... Et qu'on eut retrouvé le rapport microbiologique de l'autopsie du second enfant, qui prouvait que celui-ci était bien décédé de causes naturelles !

fut révisé et M^me^ Clark finalement acquittée et libérée après trois années de détention. Elle mourut cependant quatre ans plus tard d'une overdose d'alcool, dont beaucoup estiment qu'elle était liée aux dégâts psychologiques causés par l'erreur judiciaire tragique dont elle avait été victime.

EXERCICE 2 — Prescience ?

Un devin affirme être capable de prédire à l'avance sur quel côté une pièce de monnaie lancée par un expérimentateur impartial va retomber — ou du moins, de fournir des prédictions meilleures que le hasard. On réalise une série de $14\,000 =: n$ *lancers à pile ou face, et le devin obtient la bonne réponse dans* $7\,153 =: a_{\checkmark}$ *cas (réalisation de la v.a.* A*). Cet exercice vise à déterminer quel degré de certitude nous pouvons alors avoir sur les capacités du devin : en particulier, nous y comparerons l'approche fréquentiste et l'approche bayésienne.*

1. Modéliser le problème sous forme statistique : donner la signification du paramètre caché et l'espace dans lequel il vit, la signification de l'observation et sa loi sachant le paramètre caché, et l'hypothèse nulle qu'on souhaite tester.

2. Observer que l'hypothèse nulle est simple, et rappeler pourquoi cela est particulièrement pratique d'un point de vue mathématique.

3. Proposer la statistique de test la plus naturelle dans ce contexte, en précisant le critère de suspicion approprié.

4. Calculer la p-valeur de notre test (pour l'observation effective), et interpréter celle-ci. On écrira la formule donnant la p-valeur à partir de fonctions de tables ; pour l'application numérique, voir l'indication.

Indication : L'application numérique donne une p-valeur de $4{,}971\,474 \times 10^{-3}$.

Un souci pratique de la démarche ci-dessus est qu'on a eu besoin de calculer la loi binomiale pour une très grande valeur de son premier paramètre. Le logiciel R *sait le faire ; cependant on peut parfois être confronté à des situations où on ne dispose que d'outils plus rudimentaires, par exemple juste de tables de la loi normale… Les deux questions suivantes nous expliquent comment faire dans un tel cas.*

(5). En utilisant le théorème-limite central, justifier qu'il est raisonnable d'approcher la loi $\mathrm{Binom}^{\mathrm{le}}(n, 1/2)$ par une certaine loi normale dont on précisera les paramètres.

(6). Donner la formule pour la p-valeur lorsqu'on utilise l'approximation normale (voir l'indication pour l'application numérique) ; comparer la valeur trouvée avec celle du calcul exact.

Indication : Application numérique : $4{,}852\,424 \times 10^{-3}$.

Maintenant, nous allons reprendre l'analyse en suivant une approche bayésienne. Nous décidons d'être le plus neutres possible dans notre choix de loi à priori. Plus précisément,

notant $\boldsymbol{\theta}$ la proportion de succès que le devin est capable d'atteindre, nous proposons pour $\boldsymbol{\theta}$ la distribution (à priori) suivante :

$$\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\theta) \coloneqq \begin{cases} 1/2 & \textit{pour } \theta = 1/2 \textit{ ;} \\ \frac{1}{2}\,\mathrm{vol}(\mathrm{d}\theta) & \textit{pour } \theta \in]1/2, 1[\textit{ ;} \\ 1/4 & \textit{pour } \theta = 1. \end{cases}$$

7. Proposer un raisonnement qui pourrait conduire à ce choix de priore, en insistant sur le fait que les choix faits peuvent être considérés comme aussi neutres que possibles.

8. Au fait, dans la question précédente, nous avions dit que nous voulions être aussi neutres que possibles, et faire table rase de nos connaissances antérieures. Personnellement, en votre for intérieur, à quoi aurait ressemblé la priore que vous auriez attribuée sur $\boldsymbol{\theta}$?

9. Calculer la fonction de vraisemblance du paramètre caché (au vu de notre observation effective). À l'aide de la méthode du point critique, trouver pour quelle valeur de $\boldsymbol{\theta}$ la vraisemblance maximale est atteinte, et donner un développement limité de la fonction de log-vraisemblance au second ordre au voisinage de cette valeur.

(10).** Argüer qu'il est raisonnable d'approximer la fonction de vraisemblance par l'exponentielle du développement limité trouvé à la question précédente, sur tout $\boldsymbol{\Theta}$.

11. En déduire, à constante multiplicative près, la loi à postériori de $\boldsymbol{\theta}$.

Indication : On ne s'intéressera pas à ce qui se passe en $\theta = 1$, dans la mesure où la vraisemblance y est tellement faible [†] que la probabilité à postériori est complètement négligeable. De même, afin d'alléger les calculs, on pourra se contenter d'un résultat approximatif donnant une probabilité à postériori non nulle, mais ridiculement faible, à des valeurs de θ supérieures à 1 (!).

(12).* Vu la formule pour la densité des lois normales :

$$\mathbb{P}\big(\mathrm{Normale}(\mu, \sigma^2) \in \mathrm{d}x\big) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{vol}(\mathrm{d}x),$$

exprimer, à l'aide d'une des fonctions de tables de la loi normale, la constante de normalisation pour la loi à postériori de $\boldsymbol{\theta}$; et en déduire, après application numérique, que la probabilité à postériori que $\{\boldsymbol{\theta} = 1/2\}$ vaut environ 77 %.

Indication : Le calcul de la constante de normalisation requiert de déterminer la valeur numérique d'une intégrale qu'on ré-écrira comme une certaine constante (connue) multipliée par la probabilité qu'une certaine loi normale (de paramètres connus) tombe au-delà d'une certaine valeur (connue), ce qui nous permettra donc de calculer tout cela à partir des tables de la loi normale.

Indication : N'hésitez pas à introduire des quantités intermédiaires — explicitement définies, s'entend — pour rendre vos formules moins horribles... ☺

13. Comparer les résultats des approches bayésienne et non bayésienne. Comment interprèteriez-vous vos constatations ?

[†]. En réalité, la vraisemblance en $\theta = 1$ est rigoureusement nulle : quand je parle de « vraisemblance tellement faible », je fais ici référence à la vraisemblance telle qu'approximée à la question précédente.

EXERCICE 3 — Le paradoxe de Simpson

Une étude scientifique cherche à comparer l'efficacité de deux méthodes d'élimination des calculs rénaux[‡] *(que nous appellerons plutôt « cristaux rénaux » dans cette exercice, pour éviter tout risque de confusion avec les calculs mathématiques) : la chirurgie et les ultrasons. En l'occurrence, il s'agit d'une étude* rétrospective[§] *: on ne va pas mener de nouvelles expériences, mais simplement étudier les données passées, en l'occurrence les compte-rendus d'opérations concernant tous les cristaux rénaux ayant été traités par un certain nombre d'hôpitaux sur une certaine plage de temps. (En ne considérant, pour chaque cristal, que le traitement appliqué en première intention, afin d'éviter de créer des données corrélées). Essentiellement, l'idée est simplement de comparer les taux de succès (a-t-on réussi à éliminer le cristal ?), en appliquant une procédure statistique pour contrôler les effets aléatoires liés aux particularités de chaque cristal.*

1.★ Expliciter le modèle sur lequel s'appuie implicitement cette étude, et dire à quelle procédure statistique on voudrait procéder, sur quelle quantité d'intérêt.

2. Quel argument, cité dans le cours, fait que les statisticiens n'aiment pas trop les études rétrospectives en général ? Expliquer pourquoi cet argument n'est pas pertinent ici, et qu'une étude rétrospective semble donc appropriée, du moins à l'aune de cet argument.

Cela étant éclairci, on collecte les données sur les différentes interventions médicales sur des cristaux rénaux ; on les traite ; et on trouve que le taux de succès du traitement par ultrasons est supérieur d'environ 5 points de pourcentage, ce qui constitue en l'occurrence (vu le détail des données) un très haut niveau de significativité quant à l'hypothèse alternative d'une différence entre les deux traitements. Vous racontez cela à une chercheuse en néphrologie[¶] *de vos amis, qui dit « qu'est-ce que c'est que cette étude débile ?! C'est pourtant connu que la chirurgie est la plus efficace : c'est d'ailleurs pour ça que, plus le cristal est compliqué à opérer, plus les chirugiens ont tendance à préférer cette seconde méthode... ! ».*

☛ *Ici je tiens à signaler qu'il arrive plus souvent qu'on ne le croit que des choses « bien connues », et que des milliers de personnes sincères prétendent observer au quotidien, soient en réalité fausses, à cause des biais d'observation de notre esprit qui nous poussent à plus remarquer les données confortant nos croyances que celles les contredisant. (C'est notamment ce qui explique que, parmi les traitements des différentes médecines traditionnelles dont l'usage a semblé validé par les siècles, un certain nombre soient en réalité totalement inefficaces...). La science a par exemple permis de réfuter sans aucune ambiguïté les affirmations suivantes : que les phases de la lune ont un impact sur les naissances des bébés, que le rasage*

[‡]. Un *calcul rénal* est un dépôt minéral obstruant un canal du rein : une affection extrêmement douloureuse, quoique rarement grave en termes de risque de mortalité ou de séquelles. Si cette affection porte le même nom que la procédure consistant à appliquer des opérations mathématiques, c'est parce que dans les deux cas, l'étymologie provient du latin *calculus* signifiant « caillou » : pour le calcul mathématique, cette étymologie à priori étrange provient de ce que, autrefois (avant l'introduction des chiffres arabes en Occident), on procédait aux additions en deplaçant des cailloux ou des jetons sur un abaque (c'est le même principe que pour un boulier, mais à plat, ce qui permet d'utiliser des jetons plutôt que des boules).

[§]. Au passage, l'antonyme de « rétrospective » est *prospective*.

[¶]. La néphrologie est branche de la médecine traitant des reins : du grec ancien *νεφρός* signifiant « rein ».

augmente la repousse des poils, que les cancers de la thyroïde en France ont plus augmenté dans les régions les plus impactées par le nuage radioactif de Tchernobyl en 1986, ou que les crises de rhumatismes permettent anticiper les changements de météo (et on ne parle pas là d'arguments théoriques*, mais bien d'*observations *empiriques !) : pourtant beaucoup de gens « constatent bien que » ce phénomènes sont (croient-ils) vérifiés autour d'eux, et ce, en toute bonne foi... Mais fermons cette paranthèse : cet exercice ne parle pas de psychologie et d'esprit critique, mais bien de statistique !* ◡̈

3.★ De fait, la chercheuse a raison : c'est effectivement la chirurgie qui est la plus efficace ! Mais alors, quelle peut bien être la faille dans notre modélisation qui nous a échappé ?...

Indication : J'ai fait exprès de glisser, quelque part dans l'énoncé, un indice pour vous mettre sur la voie... ◡̈

☛ *Ce piège dans l'analyse des données est appelé* paradoxe de Simpson*. L'exemple des cristaux rénaux ci-dessus est un exemple authentique où une étude très sérieuse a été publiée sans subodorer le piège !*

4.❀ Imaginez d'autre exemples de paradoxes de Simpson qu'un ingénieur inattentif pourrait commettre dans une situation d'entreprise.

5. Comment une étude prospective aurait-elle permis d'éviter le paradoxe de Simpson ? Pourquoi cela aurait-il été une idée douteuse en pratique ?

6.★ Comment pourrait-on essayer de contourner le paradoxe de Simpson tout en restant dans le cadre de notre étude observationnelle ?

EXERCICE 4 — Le biais des familles nombreuses

Dans un village de campagne, une institutrice de CM2 regarde les fiches que ses élèves viennent de remplir, où elle demandait à chaque enfant, notamment, combien de frères et sœurs vivent dans son foyer. Elle remarque qu'il semble y avoir beaucoup de familles nombreuses : en effet, la moyenne du nombre d'enfants par foyer, dans sa classe, vaut 2,6 !

Or, notre institutrice sait bien que, à notre époque, les femmes françaises ont en moyenne 1,9 enfants environ (à l'issue de leur période de fertilité, s'entend). Sachant que, pour les enfants de 10 ans, dans la très grande majorité des cas, les frères et sœurs vivant dans son foyer correspondent exactement aux autres enfants de sa mère, et que leur mère n'aura plus d'autre enfant, notre institutrice se dit qu'elle aurait dû, logiquement, obtenir une moyenne d'environ 1,9...

L'institutrice va consulter son statisticien de frère pour lui demander si les chiffres qu'elle a obtenus sont suffisamment significatifs pour qu'on puisse conclure que, dans son village, les familles sont plus nombreuses qu'ailleurs. Mais, à sa grande surprise, son frère lui dit que la valeur qu'elle a trouvé est parfaitement "normale" et que, à tout prendre, les familles de son village seraient plutôt légèrement moins *nombreuses qu'ailleurs...*

1. Expliquer quel phénomène statistique faisait que le statisticien s'attendait à avoir un nombre d'enfants moyen par fratrie nettement supérieur à 1,9.

Indication : Par « phénomène statistique », j'entends qu'il ne faut pas chercher un phénomène lié aux défauts de la modélisation (en particulier sur les hypothèses expliquant la façon dont les familles sont composées) : en particulier, il ne faut chercher ni du côté des subtilités sociétales (familles recomposées, différences démographiques entre ville et campagne, ...), ni biologiques (possibilité de décès des mères ou des enfants, naissances de jumeaux, ...).

(2). Proposer un modèle (pas forcément ultra-réaliste, mais suffisant pour l'usage qu'on veut en faire) qui permettrait de rendre compte de la différence entre le nombre moyen d'enfants par femme et le nombre moyen d'enfants par fratrie.

3. Si on note π_n la probabilité qu'une femme ait n enfants, exprimer le nombre moyen d'enfants par fratrie en fonction des π_n.

(4). Application : Supposons que les π_n correspondent à une distribution Poisson(λ) [‖] :

$$\forall n \in \mathbb{N} \quad \pi_n = \frac{\lambda^n}{\mathrm{e}^\lambda n!}.$$

Montrer que dans ce cas, les femmes ont en moyenne λ enfants, mais que le nombre moyen d'enfants par fratrie est $\lambda + 1$.

(5).* Retrouver les résultats de la question précédente *sans calcul*, en utilisant uniquement la définition de la loi de Poisson.

Indication : Le point de départ pour résoudre cette question est de considérer un modèle où, les tailles des différentes classes ayant été fixées comme paramètres du modèle, chaque enfant de chaque classe serait attribué aléatoirement à une femme tirée uniformément dans la population adulte, indépendamment pour chaque enfant. (Bien entendu, ce n'est absolument pas un modèle réaliste, puisque les enfants d'une même femme seraient alors dispersés dans toute la France !).

On montrerait alors :

- D'une part, que dans ce modèle, le nombre d'enfants d'une femme est bien distribué (asymptotiquement) selon une loi de Poisson, dont le paramètre peut aisément être rendu égal à λ sous réserve de bien régler les paramètres du modèle ;
- D'autre part, que, un couple (*femme*, *classe*) étant donné, conditionnellement au fait que la femme en question ait bien un enfant dans la classe en question, alors le nombre *de frères et sœurs* de cet enfant sera lui aussi distribué selon la loi Poisson(λ) ; ce dont on déduira que l'espérance de la taille moyenne des fratries du point de vue des instituteurs vaut $\lambda + 1$;
- Enfin, que la relation entre les π_n et la taille moyenne des fratries du point de vue des instituteurs reste bien valable dans le cas de notre modèle.

Le soir, le mari de notre institutrice tombe sur les fiches de ses élèves et se dit, comme elle, qu'il y a décidément beaucoup de familles nombreuses dans le village ! Suite à sa discussion avec son frère, notre institutrice a compris son erreur : mais, son mari n'ayant rigoureusement aucune compétence en statistique, elle ne peut pas lui refaire l'explication dans les mêmes termes que son frère a utilisés...

[‖]. Étonnamment, ce modèle simpliste est particulièrement proche de la réalité observée !

6. Essayez de reformuler l'explication du paradoxe dans des termes aussi "grand public" que possible. (Facultatif : En-dehors des heures de cours, vérifiez si vous arrivez ainsi à faire passer l'explication du phénomène aux personnes non scientifiques de votre entourage : parents, frères et sœurs, compagnons, ...).

EXERCICE 5 — Le biais du joggeur

Votre ami Hercule, comme de nombreux habitants de la ville, va régulièrement faire du jogging au parc. Il se demande comment ses performances personnelles se situent par rapport des autres coureurs : en particulier, il voudrait savoir quelle est la proportion de coureurs qui vont moins vite, resp. plus vite, que lui. L'expérience statistique que propose Hercule est la suivante : au cours du prochain mois où il ira courir, il comptabilisera au cours de sa course, combien de coureurs il dépasse (nombre noté N_-), resp. combien de coureurs l'ont dépassé (nombre noté N_+), et estimera la proportion de coureurs courant moins vite que lui par

$$\frac{N_-}{N_- + N_+}.$$

Hercule vous explique doctement son modèle : « Je fais d'abord l'hypothèse, vous explique-t-il, que le sens dans laquel les coureurs choisissent de faire le tour du parc est indépendant de leur vitesse de course : de la sorte, peu importe si mes données ne prennent pas en compte ceux courant en sens inverse de moi. En outre, je suppose que les coureurs (dont moi-même) courent toujours à vitesse constante, et également constante d'un jour sur l'autre. Mon modèle comporte trois paramètres cachés : υ (« upsilon »), ma vitesse de course personnelle ; ξ, un paramètre multi-dimensionnel servant à décrire la distribution $P_{\xi_\checkmark}$ des vitesses des coureurs du parc[**] *; et λ, le nombre moyen de coureurs avec qui je dois m'attendre à avoir des dépassements pendant le mois. Mon observation est constituée, d'une part, du nombre de coureurs avec qui je vais effectivement avoir des dépassements, nombre noté N, que je suppose suivre (sous $\mathbb{P}_\checkmark$) une loi* Poisson$(\lambda_\checkmark)$*, et d'autre part, d'un N-uplet $(S_1, \dots, S_n)$ à valeurs dans $\{-,+\}$ indiquant, pour chaque coureur, si c'est moi qui l'ai dépassé ou lui qui m'a dépassé. La loi de ce N-uplet est construite de la façon suivante : la vitesse V_i du i-ième coureur avec qui j'ai un dépassement est tirée, de façon i.i.d. (conditionnellement à la connaissance du paramètre caché et de N), selon la loi $P_{\xi_\checkmark}$; et S_i vaut '$-$' si on a $V_i < \upsilon_\checkmark$, resp. '$+$' si on a $V_i > \upsilon_\checkmark$; de sorte qu'au final les S_i sont i.i.d.* Bernoulli$\big(P_{\xi_\checkmark}(]0, \upsilon_\checkmark[)\big)$*. Or ma quantité d'intérêt correspond précisément à $P_{\xi_\checkmark}(]0, \upsilon_\checkmark[)$! Comme je suis dans un modèle d'échantillonnage, j'utilise alors l'estimateur empirique pour déterminer cette valeur, qui*

[**]. Ici je ne détaille pas précisément comment P_ξ est définie en fonction de ξ ; mais il est entendu qu'Hercule, lui, l'a fait. Par exemple, on peut imaginer qu'Hercule a choisi de prendre $\xi := (\mu, \sigma)$ avec μ à valeurs dans $\mathbb{R}_+^*$ (homogène à une vitesse) et σ à valeurs dans $\mathbb{R}_+^*$ (sans dimension), et défini $P_{\mu,\sigma}$ comme l'exponentielle de la loi Normale$(\ln \mu, \sigma^2)$: dans ce cas, la famille $(P_\xi)_{\xi \in \mathbb{R}_+^* \times \mathbb{R}_+^*}$ des distributions de vitesses que le modèle d'Hercule envisage correspond à la famille *log-normale*, qui permet de modéliser très efficacement un grand nombre de situations réalistes. Et Hercule peut même avoir envisagé une modélisation encore plus subtile pour la distribution des vitesses des coureurs du parc, où ξ vivrait dans un espace de dimension 3, 4 ou plus, permettant encore plus de souplesse dans la description de cette distribution ! Néanmoins, dans la mesure où la modélisation précise de la distribution des vitesses des coureurs n'a, en l'occurrence, aucun impact sur l'estimateur construit par Hercule et sur son raisonnement, nous ne nous apesantirons pas sur ces subtilités ☺

correspond à $N_- / (N_- + N_+)$; et comme N sera d'au moins une centaine, la convergence asymptotique de mon estimateur fait que celui-ci sera assez précis !

1.* Porter un regard critique sur le modèle d'Hercule. On fera le distinguo entre les hypothèses complètement raisonnables ; celles qui sont inexactes mais qui ne devraient pas avoir d'incidence catastrophique sur la qualité de l'estimation, et l'"hypothèse" qui ruine complètement l'analyse (il y en a exactement une) !

Indication : Le point le plus important de cette question est d'identifier le point qui ruine complètement l'analyse. Indice : ce point relève de la même idée générale que le paradoxe des familles nombreuses... !

2. De quelle manière la faille du raisonnement d'Hercule l'induit-elle en erreur sur sa performance : se croira-t-il meilleur qu'il n'est ? Moins bon ? Et pourquoi ?...

3. Bien que vous ayez correctement détecté l'erreur de raisonnement d'Hercule, celui-ci n'est pas convaincu par votre explication. Pour achever de le convaincre, essayez de proposer un exemple de situation où le même genre d'erreur de raisonnement conduirait à une conclusion flagramment absurde !

Indication : Cette question est plus importante qu'elle n'en a l'air. Dans le monde de l'entreprise, détecter l'erreur de collègues, c'est bien, mais si vous n'êtes pas capable de les *convaincre* (eux, ou leurs supérieurs) qu'ils ont fait une erreur et donc de la réparer, ça ne sert à rien... ! Or en statistique, on a souvent tendance à s'accrocher à certains raisonnements qui semblent "de bon sens", mais qui sont faux (l'histoire du problème de Monty Hall en est un très bon exemple : de nombreuses personnes très savantes ayant longuement soutenu mordicus que le fait d'ouvrir une porte à chèvre ne changeait rien quant à l'égalité entre les deux autres portes !). Donc, avoir la capacité de démontrer la fausseté de façon convaincante est particulièrement important... !

(4).** En supposant que la distribution des vitesses des coureurs suit approximativement une loi normale, estimer plus raisonnablement quelle est la proportion de joggeurs qui courent plus vite qu'Hercule, sachant que celui-ci double 15 fois plus souvent qu'il n'est doublé. On explicitera les approximations auxquelles on aura eu recours ; certaines de ces approximations pourront être grossières si c'est nécessaire pour résoudre le problème.

Indication : La réponse ne peut pas être calculée sous une forme algébriquement fermée : on se contentera donc de donner l'équation que doit satisfaire la réponse (on pourra y faire intervenir la fonction de distribution Φ de la loi normale standard), avant éventuellement de résoudre celle-ci numériquement (à ce sujet, notez que la fonction Φ s'obtient en écrivant que $\Phi(z) = (1 + \mathrm{erf}(z/\sqrt{2}))/2$, où erf est la fonction d'erreur, laquelle est pré-implémentée dans de nombreux logiciels de calcul numérique).

EXERCICE 6 — Accidentologie

Pour être pleinement efficace, la prévention routière doit avoir une activité de veille : les comportements des automobilistes se sont-ils mis à changer ? Le cas échéant, cela peut justifier un changement dans la façon de tenter de prévenir les accidents...

Dans le cadre de cette activité de veille, on peut comparer le nombre de décès sur les routes de France pour un mois donné avec celui qu'il y avait eu à la même période (pour contrebalancer les fluctuations saisonnières) l'année précédente : si ce nombre indique un changement significatif dans le taux de mortalité, il y aura alors lieu d'en investiguer les causes…

Ici, mettons qu'on compare les décès entre mars 2024 et mars 2025. On note N_4 (la variable aléatoire qui donne) le nombre de morts pour mars 2024, resp. N_5 celui pour mars 2025.

1. Justifier que le modèle suivant semble assez pertinent : le paramètre caché est $(\lambda_4, \lambda_5) \in (\mathbb{R}_+^*)^2$ (que représente-t-il ?), et on a N_4, N_5 indépendants sous $\mathbb{P}_{\checkmark}$, avec $N_4 \sim \text{Poisson}(\lambda_{4\checkmark})$ et $N_5 \sim \text{Poisson}(\lambda_{5\checkmark})$.

(2).★ Montrer le résultat suivant : pour tous n, m,

$$\mathbb{P}_{\checkmark}(N_4 = n \mid N_4 + N_5 = m) = \mathbb{P}\big(\text{Binom}^{\text{le}}\big(m, \tfrac{\lambda_{4\checkmark}}{\lambda_{4\checkmark}+\lambda_{5\checkmark}}\big) = n\big).$$

3.★ En déduire que la statistique

$$\texttt{répartBinom}^{\texttt{le}}\big(N_4 + N_5, \tfrac{1}{2}; N_4+\big)$$

est une p-valeur (confer l'indication) pour l'hypothèse nulle $\{\lambda_5 \leqslant \lambda_4\}$.

Indication : Exceptionnellement, nous construisons *directement* la p-valeur, sans passer par une statistique de test ! Se référer à la définition (`FW'`) du polycopié (version imprimée) pour savoir ce que cela veut dire dans ce cas. (En fait, la construction qu'on fait revient à regarder la statistique de test N_4 sous la loi *conditionnée* par la valeur de $N_4 + N_5$).

4. Application numérique : pour $N_4 = 225$ et $N_5 = 240$, on tombe sur une p-valeur de 25,9 %. Qu'en concluez-vous dans ce cas ?

5. Au vu de ce qui précède, que vous inspire l'entrefilet de presse suivant, sur lequel je suis tombé un jour [††] :

SÉCURITÉ ROUTIÈRE — Le nombre de morts dérape en février

Le nombre de tués sur les routes en France a augmenté de 6,7 % en février [2015] par rapport à 2014. 240 personnes ont perdu la vie contre 225 l'année dernière.

6. Supposons que, pour mars 2025, on ait observé une augmentation incontestablement significative du nombre de victimes d'accidents de la route. Cela suffit-il à en déduire qu'il y a une dégradation de la sécurité routière en France ?

EXERCICE 7 — L'affaire Brian Wansink

En 2016, le P^r Brian Wansink*, chercheur étasunien en psychologie, était considéré comme un expert de niveau mondial en science du comportement alimentaire. Dans une*

[††]. *20 Minutes* n° 2809 du 13 mars 2015, édition « Grand Paris » ; p. 6.

note de blog publiée en novembre de cette année-là, il se livra à quelques considérations sur l'attitude qu'il souhaitait encourager auprès des jeunes chercheurs de son domaine. Voici (modulo adaptation pour les besoins de ce problème) ce qu'il écrivit :

> Une doctorante [‡‡] turque appelée Özge Sığırcı est venue travailler six mois dans mon laboratoire comme chercheuse visiteuse. À son arrivée, je lui ai confié un jeu de données issu d'une étude qui avait donné des résultats négatifs. [Il s'agissait d'une expérience où diverses personnes ont consommé le même repas, payé tantôt à prix plein, tantôt à moitié prix. L'hypothèse était que les personnes ayant payé moitié prix trouveraient leur repas moins bon ; mais les données recueillies avaient échoué à la valider]. J'ai dit à M^lle^ Sığırcı : « Cette étude nous a couté beaucoup de temps et d'argent. Il s'agit d'un jeu de données riche et original : même si l'angle d'analyse initial a échoué, je suis sûr qu'on peut en sauver quelque chose malgré tout ! ». J'avais en effet eu, suite à l'échec de mon "plan A", trois nouvelles idées ("plan B", "plan C" et "plan D") d'angles d'analyse de ces données. J'écrivis ainsi à Özge :
>
> > *« Je ne pense pas avoir jamais mené à bien d'étude intéressante où les résultats se soient révélés de façon flagrante au premier coup d'œil. Ce qui sera intéressant, c'est de voir dans quelles situations le prix réduit a un effet et dans lesquelles il n'en a pas. Il faudrait que vous trouviez certaines situations ou types de personnes pour lesquelles le prix réduit a un effet. Essayez d'imaginer toutes les façons dont on pourrait découper le jeu de données pour en analyser des sous-ensembles et voir quand est-ce que l'effet se manifeste. Par exemple, si cela marche pour les hommes mais pas pour les femmes, nous avons une condition intéressante. Voici quelques autres sous-groupes qu'on pourrait étudier séparément : (…) ».*
>
> Chaque jour, M^lle^ Sığırcı revenait avec de nouveaux résultats intrigants ; nous y réfléchissions ; et en déduisions une nouvelle façon de ré-analyser le jeu de données à l'aune d'un nouveau jeu d'hypothèses plausibles. En fin de compte, nous sommes arrivés à trouver des conclusions statistiquement solides. Nous avons écrit ensemble deux articles de recherche, puis un troisième, ce dernier s'appuyant sur une découverte qu'Özge avait faite de façon complètement autonome en fouillant les données.
>
> À l'issue de ses six mois de visite, M^lle^ Sığırcı avait écrit 5 articles de recherche acceptés pour publication. En comparaison, une post-doctorante [∗] de mon équipe, qui avait décliné ma proposition d'étudier ce même jeu de données en argüant qu'elle n'avait pas le temps de se consacrer à un tel « projet annexe » en sus de sa recherche principale, a fini par quitter le monde de la recherche un an plus tard, en ayant eu un rythme de publication quatre fois moindre que

[‡‡]. Une doctorante est une apprentie-chercheuse, diplômée d'un master, qui travaille pendant quelques années sous la direction de chercheurs confirmés afin de monter pleinement en compétence — cette période d'apprentissage étant appelée « doctorat ».

[∗]. Une post-doctorante est une jeune chercheuse qui a déjà terminé son doctorat, mais qui, faute d'offre d'emploi ferme, travaille provisoirement "en CDD" dans un laboratoire de recherche.

> M^lle^ Siğirci... Moralité : « Ne soyons pas si difficiles : les plus accomodants, ce sont les plus habiles » ! Pour réussir dans la recherche, il faut dire « oui » à toutes les opportunités qui se présentent, même si on ne sait pas bien à première vue comment on va les exploiter !

La première réponse à ce post de blog, écrite le P^r^ Paul KIRSCHNER*, un collègue néerlandais du P^r^ Wansink, disait : « Brian, j'espère que tu n'es pas sérieux et qu'il s'agit d'une satire pince-sans-rire...! ».*

1. Quel problème (en lien avec l'inférence statistique, s'entend ! ◡̈) a détecté Paul Kirschner dans le post de blog de Brian Wansink ? Bien expliquer celui-ci, dans deux styles différents :

— D'une part, en version "pour statisticiens", en employant un vocabulaire technique précis vous permettant de désigner le souci par une formulation concise ;

— D'autre part, en version "pour Béotiens", en utilisant des tournures qui vous permettraient de faire comprendre le souci à des personnes n'ayant pas reçu de formation particulière en sciences [†], et de les convaincre qu'il y a clairement quelque chose qui ne va pas...

EXERCICE 8 — Richard est-il un connard ?

Le personnage central de cet exercice est un certain RICHARD*. C'est un statisticien extrêmement brillant, mais présentant des traits autistiques assez prononcés, ce qui l'amène fréquemment à heurter la sensibilité de ses interlocuteurs en adoptant un discours froidement analytique sur des sujets hautement sensibles, sans pour autant penser à mal... Si le manque de délicatesse de Richard ne mérite certes pas d'être pris en exemple, ses réflexions reflètent néanmoins, quand on prend le soin de bien les comprendre, des démarches fort pertinentes sur le plan épistémique [‡]. C'est pourquoi nous allons nous pencher sur certains exemples de propos surprenants tenus par Richard, pour essayer d'en tirer des idées qu'il peut être utile de garder à l'esprit quand on cherche à analyser un sujet de façon aussi fiable que possible, en particulier dans des circonstances où notre instinct social nous pousserait "viscérablement" à proposer des conclusions qui se seraient forcément pas fondées épistémologiquement...*

☛ *Veuillez noter que le contexte de cet exercice est intégralement fictif ! Le personnage de Richard est imaginaire (et n'est pas censé représenter quelqu'un en particulier) ; les discussions que je lui prête avec ses amis sont inventées de toutes pièces ; et les données des « études » mentionnées dans ces discussions sont tout aussi fictives...*

Histoire de bien charger la barque et de montrer la tension qui peut se produire entre analyse statistique et convenances sociales, je vais imaginer ci-dessous que Richard s'est intéressé à un sujet super-polémique : à savoir, la différence éventuelle d'intelligence entre les femmes et les hommes [§] ! On supposera ici, pour les besoins de l'exercice, que l'« intelli-

[†]. Par exemple, des décideurs politiques à qui vous vous adressez.

[‡]. « Épistémique » signifie « relatif à la connaissance ».

[§]. De fait, il s'agit là d'un sujet d'étude bien réel, mais sur lequel les scientifiques s'aventurent avec la plus grande prudence, et ont beaucoup de peine à se mettre d'accord... Pour vous faire une idée de la complexité de la question, confer par exemple la page *Wikipédia* « Influence du sexe sur l'intelligence ».

gence » est un concept défini de manière univoque et mesurable précisément par des tests de type QI.

En 2022, Richard eut avec son beau-frère Nicolas *la discussion suivante :*

Nicolas — *Dis-moi, Richard, à conditions d'éducation rigoureusement égales, les hommes et les femmes ont-ils rigoureusement la même intelligence en moyenne, ou existe-t-il une petite différence liée au sexe ?*

Richard — *Non ; l'intelligence moyenne des hommes et des femmes, même indépendamment des conditions d'éducation, n'est pas rigoureusement égale !*

Nicolas — *Ah ; une différence a donc été prouvée par des études scientifiques ?*

Richard — *Pas que je sache, non.*

Nicolas — *Alors c'est juste une vague intuition de ta part ?*

Richard — *Ah non ; j'en suis certain à 100 % : je pourrais y mettre ma tête à couper !*

Nicolas — *Euh... Ça me parait complètement non scientifique, comme point de vue, mais admettons... Et qui sont les plus intelligents, selon toi ? Les hommes ou les femmes ?*

Richard — *Ça, par contre, je n'en ai aucune idée ! Je sais juste qu'ils ne sont pas égaux.*

Nicolas — *Mais qu'est-ce que tu racontes ?! Qu'est-ce qui te permet d'affirmer cela ?...*

Richard — *Eh bien, c'est la conséquence logique du fait que je raisonne de façon bayésienne, d'une part ; et que je sois athée, d'autre part !*

1.* Expliquer en quoi est-ce que Richard a, dans un sens, raison : dès lors que, à ses yeux, l'humain est simplement le fruit de processus physiques liés à l'évolution, et pas la création d'une entité supérieure, les probabilités subjectives qu'il mettra sur la valeur de la différence d'intelligence moyenne entre hommes et femmes l'amèneront *forcément* à considérer que celle-ci ne peut pas être nulle, et ce, d'une façon qu'aucune étude statistique ne pourra remettre en cause !

En 2023, nouvelle réunion de famille ; nouvelle discussion entre Richard et Nicolas :

Nicolas — *Eh, Richard, j'ai vu une étude scientifique très sérieuse sur la différence d'intelligence entre hommes et femmes... Les auteurs disent que leurs données ne mettent en évidence absolument aucune différence d'intelligence entre hommes et femmes : tu vois, tu avais tort de m'affirmer qu'il y en avait forcément une !*

Richard — *Ah oui ; j'ai vu cette étude : données de très grande qualité, et analyse statistiquement impeccable ! Cependant, ça ne contredit en rien mes affirmations de l'année dernière...*

2. Quel type de procédure a-t-elle été suivie, manifestement, par les auteurs de l'étude mentionnée par Nicolas ? Pourquoi est-ce que les auteurs de cette étude, malgré le fait qu'ils soient plutôt d'obédience fréquentiste, seraient néanmoins parfaitement d'accord avec Richard sur un point : leurs résultats ne prouvent en rien que la différence moyenne d'intelligence entre hommes et femmes soit inexistante !

Mais venons-en à la conversation la plus intrigante entre Richard et Nicolas, qui eut lieu à la réunion de famille de 2024 :

NICOLAS — *Les auteurs de l'étude dont je t'avais parlé l'an dernier ont poussé leur étude sur un échantillon de taille beaucoup plus grande ; et cette fois-ci, ils t'ont donné raison : il y a bien une différence moyenne d'intelligence entre hommes et femmes ; et en l'occurrence, ce sont les femmes qui sont les plus intelligentes...*

RICHARD — *Oui, en effet ! Étude d'excellente qualité, à nouveau.*

NICOLAS — *Les auteurs de l'étude disent que leurs données prouvent leur conclusion au-delà de tout doute raisonnable : au sens où l'écart qu'ils ont observé n'aurait eu qu'une chance sur deux mille de se produire si les hommes avaient une intelligence supérieure ou égale à celle des femmes !*

RICHARD — *En effet ! Ils ont raison, cela tranche définitivement la question : ce sont les femmes qui sont les plus intelligentes ! Combien de sujets ont-ils interrogés, déjà ?...*

NICOLAS — *Ha ha ; j'étais sûr que tu me poserais la question, alors j'ai appris la réponse par cœur... 14 731 femmes et 16 694 hommes, très exactement !*

RICHARD, après quelques secondes de calcul mental — *Oui ; donc en fait, ils ont prouvé qu'il n'y avait essentiellement aucune différence. À la bonne heure*[¶] *!*

NICOLAS — *Mais tu te fiches de moi ou quoi ?! Puisque je te dis qu'ils ont fait une étude super-rigoureuse, sur une quantité de cobayes super-grande, et que leurs conclusions sont super-décisives !...*

RICHARD — *Ah mais ; je suis bien d'accord avec ces trois points ! N'empêche, que, nonobstant les choux gras que la presse a faits de ces résultats, en fait, ce que ces chiffres montrent, c'est* justement *une quasi-absence de différence entre les deux sexes !*

3. À votre avis (et sans regarder l'énoncé des question suivantes ! ☺), qu'est-ce que Richard a en tête en disant cela ?

Pour mieux comprendre la dernière réflexion de Richard, nous allons avoir besoin d'expliquer le modèle statistique de l'étude évoquée dans sa discussion avec Nicolas... On suppose que l'étude a échantillonné n_X femmes et n_Y hommes, à chaque fois de façon parfaitement uniforme et indépendante, et que les effets liés à l'environnement ont pu être entièrement éliminés par quelque procédé mystérieux[‖]*. Sous ces hypothèses, pour $i \in [\![0, n_X[\![$, on note X_i la mesure d'intelligence de la femme numéro i, resp. Y_j la mesure d'intelligence de l'homme numéro j pour $j \in [\![0, n_Y[\![$. Le modèle (dont nous ne discuterons pas la pertinence ici) postule que, sous le véritable contexte probabiliste, les X_i suivent chacune la loi* Normale$(\mu_{X\checkmark}, \sigma^2_\checkmark)$*, tandis que les Y_j suivent chacune la loi* Normale$(\mu_{Y\checkmark}, \sigma^2_\checkmark)$*, avec $\mu_{X\checkmark}, \mu_{Y\checkmark}, \sigma_\checkmark$ inconnus, les deux premiers paramètres cachés susnommés étant à valeurs dans $\mathbb{R}$, et le troisième à valeurs dans $\mathbb{R}_+^*$*[**]*.*

[¶]. « À la bonne heure ! » est une expression familière marquant une forme d'approbation. En l'occurrence, Richard veut dire qu'il trouve que c'est plutôt une bonne nouvelle d'apprendre qu'il n'y a essentiellement aucune différence entre hommes et femmes : car cette absence de différence enlève un argument à ceux qui auraient prétendu à la supériorité intrinsèque d'un sexe sur un autre : alors qu'à l'inverse, l'existence d'un argument scientifique pouvant être exploité à mauvais escient par des groupes à visées sexistes (dans un sens ou dans l'autre) n'aurait pas manqué de créer des frictions supplémentaires dans la société...

[‖]. Je précise que ce dernier point n'est pas réaliste *du tout* : au contraire, il est *extrêmement* difficile de moduler convenablement les effets de l'environnement social quand on s'intéresse à de telles des données psychométriques : alors, les éliminer complètement, n'en parlons pas...!

[**]. On notera donc que, dans ce modèle, on suppose que la variance des intelligences entre femmes, resp. entre hommes, est exactement identique.

L'outil approprié pour tester l'hypothèse nulle $\{\boldsymbol{\mu}_{\mathrm{X}} = \boldsymbol{\mu}_{\mathrm{Y}}\}$ *est alors le* test de Student, *qui exploite le théorème suivant :*

Théorème. *Pour le modèle décrit ci-dessus, notant* $n_{\mathrm{tot}} \coloneqq n_{\mathrm{X}} + n_{\mathrm{Y}}$, *on a*

$$\frac{(\mathrm{moy}(Y_j)_j - \mathrm{moy}(X_i)_i) - (\mu_{\mathrm{Y}\checkmark} - \mu_{\mathrm{X}\checkmark})}{(n_{\mathrm{X}}\,\mathrm{var}_{\mathrm{emp}}(X_i)_i + n_{\mathrm{Y}}\,\mathrm{var}_{\mathrm{emp}}(Y_j)_j)^{1/2}} \overset{\mathbb{P}_\checkmark}{\sim} \left(\frac{n_{\mathrm{tot}}}{n_{\mathrm{tot}}-2} \times \frac{1}{n_{\mathrm{X}} n_{\mathrm{Y}}}\right)^{1/2} \times \mathrm{T}_{\mathrm{St}}(n_{\mathrm{tot}}-2),$$

où $\mathrm{T}_{\mathrm{St}}(n_{\mathrm{tot}}-2)$ *est ce qu'on appelle la* loi de Student à $(n_{\mathrm{tot}}-2)$ *degrés de liberté*[††], *qui est tabulée dans tous les logiciels destinés à traiter des données statistiques.*

4. D'après ce qu'en explique Nicolas, quelle hypothèse nulle les auteurs de l'étude ont-ils testée, et quelle p-valeur ont-ils obtenue ? D'après les explications ci-dessus concernant le test de Student, quelle statistique de test a-t-elle été utilisée pour mener la procédure à bien ?

5. Au vu des résultats avancés par les auteurs de l'étude, quelle est la réalisation qu'ils ont obtenue pour leur statistique de test ? On écrira celle-ci sous la formule d'une certaine fonction de table, évaluée en les valeurs appropriées. Pour l'application numérique, confer l'indication ci-dessous.

Indication : Si vous avec bien fait votre travail, la valeur de la fonction de table que vous allez vouloir connaitre vaudra numériquement 3,291 (en valeur absolue). Remarque : Selon la façon dont vous aurez formalisé la statistique de test, cela pourra ensuite conduire à des variantes concernant la valeur de la statistique de test ; mais dans tous les cas, cela n'impactera pas la suite du problème.

6. En déduire la réalisation d'un estimateur naturel concernant la valeur de $(\boldsymbol{\mu}_{\mathrm{X}} - \boldsymbol{\mu}_{\mathrm{Y}})/\boldsymbol{\sigma}$. Au vu de sa valeur numérique confer indication ci-dessous), en quoi est-il légitime de considérer qu'il s'agit en effet d'une différence négligeable ? (Expliquer pourquoi ce qui compte, c'est effectivement la valeur de $(\mu_{\mathrm{X}\checkmark} - \mu_{\mathrm{Y}\checkmark})/\sigma_\checkmark$), pas celle de la différence $\mu_{\mathrm{X}\checkmark} - \mu_{\mathrm{Y}\checkmark}$ elle-même).

Indication : L'application nuémrique donne $3{,}96 \times 10^{-2}$.

(7).* Calculer, à l'aide d'une fonction de table appropriée, une borne *supérieure* à 95 % de confiance (autrement dit, un intervalle de confiance dissymétrique de la forme $]-\infty, *]$, qui peut alors être simplement décrit par sa borne de droite) sur la quantité d'intérêt $(\boldsymbol{\mu}_{\mathrm{X}} - \boldsymbol{\mu}_{\mathrm{Y}})/\boldsymbol{\sigma}$. L'application numérique est donnée ci-dessous.

Indication : L'application numérique donné $5{,}93 \times 10^{-2}$.

8. En supposant qu'on ait $(\mu_{\mathrm{X}\checkmark} - \mu_{\mathrm{Y}\checkmark})/\sigma_\checkmark = 0{,}059$, quelle est la probabilité que, si on prend une femme au hasard et une homme au hasard, ce soit la femme qui soit la plus intelligente des deux ? Et, parmi les personnes surdouées (i.e., dont l'intelligence dépasse l'intelligence moyenne de la population mondiale d'au moins deux écarts-type de la distribution des intelligences de la population mondiale), quelle est la proportion de femmes ? Exprimer le résultat à l'aide des fonction de tables de la loi normale standard. Commenter les applications numériques (qui sont fournies en indication).

Indication : Les valeurs numériques sont resp. égales à 51,66 % et 53,50 %.

[††]. Intuitivement, il s'agit d'une espèce de loi normale standard qu'on aurait légèrement déformée.

Inférence statistique / Séance 9

Vers l'analyse de données

Énoncé des exercices

Formation d'Ingénieur Civil des Mines de Nancy *

2 juin 2025

EXERCICE 1 — Puits liquide d'un lingot VAR

La refusion à l'arc sous vide (VAR ou Vacuum Arc Remelting*) est un procédé métallurgique consistant à faire fondre une électrode d'alliage métallique dans une lingotière refroidie par eau, sous vide. Un arc électrique éclate entre le fond de la lingotière et le bas de l'électrode, ce qui permet de fondre cette dernière et de faire tomber du métal liquide qui solidifie au contact de la lingotière. Au fur et à mesure que le processus avance, le lingot croît en même temps que l'électrode décroît, l'arc électrique se maintenant entre le sommet du lingot et le bas de l'électrode. Le fond du lingot et ses bords sont solides, mais au sommet de celui-ci, perdure du fait de la présence de l'arc une zone liquide en forme de U ou de V appelée* puits liquide*. Confer le schéma en figure 1.*

Ce procédé est extrêmement complexe et dépend de nombreux paramètres (masse fondue au cours du temps, puissance électrique injectée, type d'alliage concerné, etc.). En fixant tous les paramètres par ailleurs, un industriel a pu obtenir l'évolution du volume du puits liquide en fonction de l'intensité du courant de fusion pour 26 *essais. Ce volume est un indicateur important sur la "qualité" de l'alliage refondu car en fonction de sa valeur, le refroidissement sera plus ou moins rapide et pourra entraîner des structures de solidification ou des défauts différents*[*]*, et donc des usages différents pour le lingot finalement obtenu.*

L'évolution du puits liquide en fonction de l'intensité est représentée sur la figure 1, où les x_i sont les différentes intensités du courant électrique (supposées parfaitement connues), et les $y_{i\checkmark}$ les volumes de puits liquides mesurés pour chaque intensité. La corrélation linéaire semble bien respectée, mais il faut le vérifier. En tant qu'ingénieur(e) de l'entreprise, vous souhaitez appliquer le modèle de la régression linéaire sur ces données. Vous diposez pour cela d'une section complémentaire de votre polycopié dédiée à ce modèle, fournie en annexe A ; ainsi que de certains résultats numériques pré-calculés à partir de vos données, listés dans la table 2.

*Équipe pédagogique : Rémi Peyre, Virgile Brodu, Bernard Dussoubs, Valentin Féray, Mathilde Gaillard, Abdelkader Metakalard, Anouk Rago, Pierre-Adrien Tahay.

[*]. Confer TD sur les alliages binaires de votre cours de physique statistique ☺

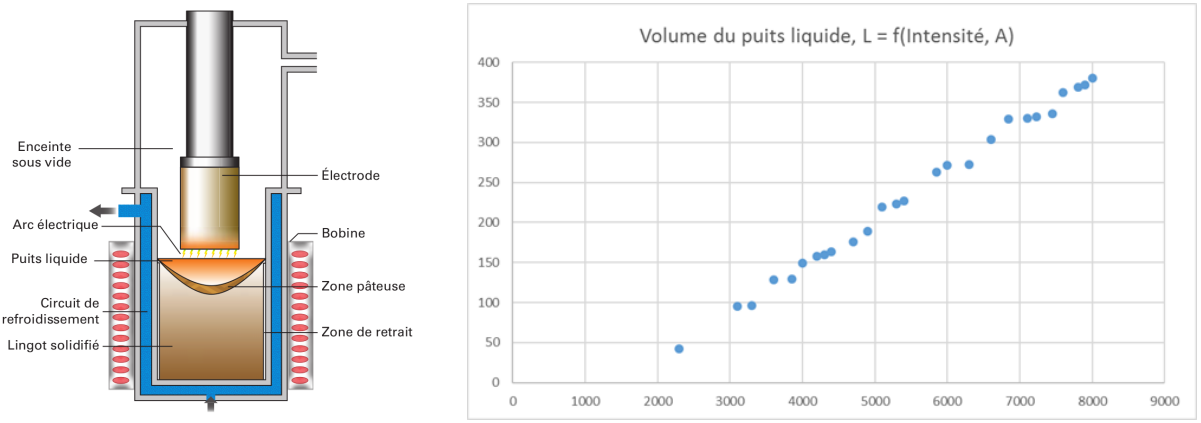


FIGURE 1 – À gauche : Schéma de principe de la refusion à l'arc sous vide, d'après Jardy & al. 2022. À droite : Volume du puits liquide en fonction de l'intensité du courant.

n	26
$\sum_{i=0}^{n-1} x_i$	143 120 A
$\sum_{i=0}^{n-1} y_{i\checkmark}$	6 074 ℓ
$\sum_{i=0}^{n-1} x_i^2$	857 876 000 A^2
$\sum_{i=0}^{n-1} y_{i\checkmark}^2$	1 667 540 ℓ^2
$\sum_{i=0}^{n-1} x_i y_{i\checkmark}$	37 597 720 $\mathrm{A} \cdot \ell$
$\sum_{i=0}^{n-1} (x_i - \bar{x})^2$	70 055 446,15 A^2
$\hat{\sigma}^2_{\mathrm{B}\checkmark}$	50,576 9 ℓ^2

TABLE 2 – Quelques valeurs utiles calculées à partir des paramètres du modèle et des réalisations des observations.

On va considérer que les $y_{i\checkmark}$ sont reliés aux x_i selon le modèle de la régression linéaire : sachant les x_i, les $y_{i\checkmark}$ sont les réalisations de variables aléatoires Y_i indépendantes, avec

$$\text{Loi}(Y_i) = \text{Normale}(\alpha_\checkmark x_i + \beta_\checkmark, \sigma_\checkmark^2),$$

pour certaines quantités $(\alpha_\checkmark, \beta_\checkmark, \sigma_\checkmark)$ inconnues.

1. Déterminez les paramètres $\hat{\alpha}_\checkmark$ et $\hat{\beta}_\checkmark$ estimant par moindre carrés la droite de régression $y = \alpha_\checkmark x + \beta_\checkmark$.

2. Si « $y = \hat{\alpha}_\checkmark x + \hat{\beta}_\checkmark$ », droite des moindres carrés représentant $y(x)$, est la meilleure estimation possible, est-ce que « $x_\checkmark = (y - \hat{\beta}_\checkmark) / \hat{\alpha}_\checkmark$ » est la meilleure droite des moindres carrés représentant $x(y)$ au vu des données fournies ? Pourquoi ?

3. On souhaite tester si la régression est « significative ». À quelle hypothèse nulle cela correspond-il ?

4. Indiquer qui est, d'après l'annexe, la statistique de test appropriée pour tester cette hypothèse nulle, et calculer sa réalisation.

5. Quel est le critère de suspicion approprié pour notre test ? (et pourquoi ?).

On cherche maintenant une valeur seuil que la statistique de test n'ait qu'une probabilité 10^{-4} de dépasser (en valeur absolue) sous l'hypothèse nulle. En l'occurrence, on calcule numériquement que cette valeur vaut 4,66.

6. À quel calcul de fonction de table (et avec quelles valeurs des arguments) a-t-on procédé pour déterminer cette valeur ?

Indication : On admettra la propriété suivante : dans le contexte probabiliste qui nous intéresse, la loi de la statistique de test est ici symétrique par rapport à 0.

7. Concluez, si cela est possible, sur la significativité de la régression.

Imaginons maintenant qu'on s'intéresse à ce qui se passe si on considère une intensité de 5 000 A. *Lorsqu'on consulte l'annexe, on s'aperçoit qu'il y a* deux *situations qui semblent se référer à un tel cas : celle qui parle de l'estimation de $\alpha x_\star + \beta$ (en prenant le cas échéant $x_\star = 5\,000$ A), et celle où on parle de l'estimation de Y_n (en prenant cette fois-ci $x_n = 5\,000$ A). En outre, les théorèmes relatifs à ces deux situations sont extrêmement similaires également : en particulier, à part le fait évident qu'on parle tantôt de $x_\star$ et tantôt de x_n, les formules pour les erreurs standard ne diffèrent que par un terme « $n+$ » à l'intérieur d'une des parenthèses…*

8. Bien expliquer la différence entre les deux quantités d'intérêt auxquelles se réfèrent ces deux situations.

9. Déterminez le volume de puits liquide qu'on peut attendre pour une intensité de courant de 5 000 A. [On demande ici une réponse sous forme de nombre unique, pas sous forme

d'intervalle ni de distribution de probabilité]. Comment s'appelle, dans le jargon statistique, la quantité que vous venez de calculer ?

On s'apprête maintenant procéder à une expérience avec une intensité de courant de 5 000 A *et à déterminer le volume du puits liquide lors de cette expérience.*

10. Déterminer la réalisation de l'intervalle de prédicition, à 5 % de risque, pour le volume du puits qu'on est censé trouver.

Indication : On donne l'intervalle de fluctuation à 95 % de confiance pour la loi $\mathrm{T}_{\mathrm{St}}(24)$: $[\pm 2{,}07]$.

Lors de la mesure, la mesure du volume du puits a donné 223 ℓ. *L'opérateur est ennuyé, car cette valeur n'appartient pas à l'intervalle de fluctuation calculé ci-dessus…*

11. Quelles sont les différentes attitudes qu'il serait possible d'adopter face à une telle mesure ? En l'occurrence, laquelle vous semblerait la plus appropriée ?

Lorsqu'on cherche à déterminer l'ordonnée de la (vraie) droite de régression pour une intensité de 1 000 A*, on trouve une réalisation de l'intervalle de confiance de* $[-36{,}93, -31{,}17]$ ℓ. *Il semble donc que pour une telle intensité, le volume du puits liquide attendu soit clairement* négatif*…*

12. Comment expliquez-vous cela ? Quelle est la bonne attitude à adopter face à un tel cas selon vous ?...

EXERCICE 2 — Les iris [†]

Vous venez de cueillir un magnifique spécimen d'iris, que vous avez collé dans votre herbier, et que vous voudriez légender. Vous savez que, dans votre région, il existe trois espèces d'iris différentes : Iris setosa *(S),* Iris versicolor *(C) et* Iris virginica *(G) ; vous aimeriez donc identifier la bonne espèce. Comme vous manquez de connaissances en botanique, vous voulez vous reposer sur une procédure standardisée, consistant à mesurer quatre caractéristiques de votre fleur, à savoir : la longueur des sépales* [‡] *(*$\mathfrak{X}$*), la largeur des sépales (*$\mathfrak{W}$*), la longueur des pétales (*$\mathfrak{V}$*) et la largeur des pétales (*$\mathfrak{U}$*). En effet, vous disposez d'une base*

[†]. Le contexte de cet article est inspiré d'un célèbre jeu de données utilisé par le statisticien Ronald Fisher dans un article de 1936 où il introduisait le concept d'*analyse discriminante linéaire* — que vous étudierez l'an prochain dans votre cours d'analyse de données. Ici cet exercice s'intéresse à une forme d'analyse discriminante *non* linéaire, où nous ne supposons pas l'homoscédasticité entre espèces.

[‡]. Les *sépales* sont les composants de la couronne externe (le « calice ») d'une fleur. Chez la plupart des fleurs, ils sont peu développés, épais et verts (ressemblant à des petites feuilles "soutenant" les pétales) ; mais chez les iris, ils ont une apparence similaire aux pétales [dans ce cas, on parle plus volontiers de *tépales* — extérieurs pour les sépales, resp. intérieurs pour les pétales]. Les sépales d'un iris sont donc ce qu'on appellerait informellement les "pétales extérieurs" de la fleur (qui sont larges, horizontaux, incurvés vers le bas, avec une tache colorée près de la tige) ; tandis que les pétales *stricto sensu* sont les tépales intérieurs de la fleur (plus fins, plus verticaux, plus rectilignes, de couleur à peu près homogène). [En fait, les iris contiennent même une troisième structure « pétaloïde » — c'est le terme technique utilisé ! —, située juste au-dessus des sépales, correspondant à ce qu'on appelle le *style* de la fleur]. Je remercie Céline Pesteil, grâce à qui j'ai pu vous transmettre ces explications ! ☺

de données comportant $n = 50$ *iris de chaque espèce (espèces dument identifiées par des botanistes compétents) pour lesquelle chacune de ces caractéristiques a été mesurée, et dont vous pouvez par conséquent vous servir pour essayez d'identifier votre fleur.*

Le modèle est le suivant. Nous commençons par appliquer un logarithme à toutes les données de longueur : $\log \mathfrak{X}$ *devient alors* X, $\log \mathfrak{W}$ *devient* W, *etc. Les fleurs de la collection de référence sont numérotées de* 0 *à* $3n-1$, *où les numéros de* 0 *à* $n-1$ *correspondent aux fleurs de l'espèce* S, *les numéros de* n *à* $2n-1$ *aux fleurs* C, *et les numéros* $2n$ *à* $3n-1$ *aux* G ; *quant à la fleur de notre herbier, on la repère par le numéro* $3n$, *et son espèce est notée* $\varepsilon_{\checkmark}$.

Pour toutes les fleurs (aussi bien celles de la collection de référence que celle que nous venons de cueillir), nous supposons que (sous la véritable distribution de probabilité $\mathbb{P}_{\checkmark}$, *s'entend) les quadruplets de mesures (logarithmiques) obtenus sont indépendants d'une fleur à l'autre, et que le quadruplet d'une fleur donnée est distribué selon une loi normale multivariée dont l'espérance et la matrice de covariance ne dépendent que de l'espèce de la fleur : les mesures (logarithmiques) d'un* I. setosa *sont ainsi supposées être la réalisations d'une loi* Normale$(\ln \vec{\mu}_{\mathrm{S}\checkmark}, \boldsymbol{\Gamma}_{\mathrm{S}\checkmark})$, *celles d'un* I. versicolor *sont les réalisations d'une loi* Normale$(\ln \vec{\mu}_{\mathrm{C}\checkmark}, \boldsymbol{\Gamma}_{\mathrm{C}\checkmark})$, *etc.* [§]

(1). Identifier en totalité le paramètre caché du modèle, et donner la dimension de l'espace du paramètre caché (voir l'indication). De même, identifier en totalité l'observation, et donner la dimension de son espace. Comparer les valeurs numériques des dimensions du paramètre caché et de l'observation : diriez-vous que nous sommes en mesure d'obtenir des analyses statistiques précises dans ces conditions ?

Indication : La dimension d'un sous-ensemble "régulier" de $\mathbb{R}^k$ (ce qu'on appelle une *variété*) est le nombre de paramètres *réels* dont on a besoin pour le décrire [¶] (via des applications supposées de régularité $\mathbf{C}^1$), quitte à devoir découper ce sous-ensemble en un nombre discret (fini ou dénombrable) de blocs ayant chacun son propre paramétrage (auquel cas, attention, la dimension globale du sous-ensemble ne sera pas la somme des dimensions de chaque bloc, mais leur *maximum* !).

(2). L'auteur de l'énoncé a choisi de travailler sur les logarithmes des grandeurs mesurées plutôt que sur les données brutes elles-mêmes : quel est l'idée sous-jacente à cette transformation ; et vous semble-t-elle pertinente en l'occurrence ? Citer une raison qui peut rendre conceptuellement *nécessaire* de travailler avec les logarithmes, resp. une raison qui peut rendre conceptuellement nécessaire de ne *pas* appliquer le logarithme aux données brutes.

(3). Le choix de la notation '$\boldsymbol{\Gamma}_{\mathrm{S}\checkmark}$' doit vous apparaitre bizarre : en effet, les conventions utilisées dans ce cours sont d'utiliser le gras *ou* la coche, mais jamais les deux ensemble ; et de

[§]. Pour un quadruplet $\vec{\mu} \coloneqq (\mu_{\mathrm{X}}, \mu_{\mathrm{W}}, \mu_{\mathrm{V}}, \mu_{\mathrm{U}})$, ce que j'entends par « $\ln \vec{\mu}$ » est simplement le quadruplet des logarithmes $(\ln \mu_{\mathrm{X}}, \ln \mu_{\mathrm{W}}, \ln \mu_{\mathrm{V}}, \ln \mu_{\mathrm{U}})$. Si j'ai choisi d'exprimer les espérances des lois normales multivariées comme des logarithmes, c'est afin que $\vec{\mu}_{\mathrm{S}}$, $\vec{\mu}_{\mathrm{C}}$ et $\vec{\mu}_{\mathrm{G}}$ soient physiquement homogènes à des longueurs, et correspondent ainsi à des quantités ayant un vrai sens concret.

[¶]. La dimension des espaces revêt une grande importance en analyse numérique, comme vous l'avez vu au premier semestre, car la pertinence de l'utilisation de telle ou telle méthode est susceptible de dépendre radicalement de la dimension. En analyse de données, les considérations de dimension sont importantes notamment en raison de la question du phéonème de *surapprentissage*, un phénomène que vous étudierez l'année prochaine.

même, dans le cours, les lettres grecques apparaissent toujours en minuscule... [‖] Pouvez-vous donc expliquer pourquoi le compositeur de l'énoncé a fait ce choix ici ?

Pour traiter ce problème à l'aide d'un logiciel, nous allons utiliser la technique appelée analyse discriminante quadratique [**]*. Dans cette technique, on va procéder en deux temps, avec une phase d'*apprentissage *où on ne s'intéressera qu'aux données et aux paramètres relatifs aux fleurs de la collection de référence, puis à une phase d'*application *où on se s'intéressera qu'à la nouvelle fleur*[††]*. Voici par exemple à quoi peut ressembler un code appliquant cette technique avec la bibliothèse* `MASS` *du logiciel* R *:*

```
> # Chargement de la bibliothèque
> library(MASS)
> # Lecture des données
> read.delim2("iris.tsv", row.names = 1) -> df
> # Calcul des logarithmes
> df$Log_long_sep = log(df$Long_sepales_cm)
> df$Log_larg_sep = log(df$Larg_sepales_cm)
> df$Log_long_pet = log(df$Long_petales_cm)
> df$Log_larg_pet = log(df$Larg_petales_cm)
> # Application du modèle de l'analyse discriminante quadratique
> analyse = qda(Espece ~
+                 Log_long_sep + Log_larg_sep + Log_long_pet + Log_larg_pet,
+               data = df)
```

Le logiciel effectue alors un certain nombre de calculs relatifs au modèle de l'analyse discriminante quadratique, à partir desquels il crée un objet `analyse`*, au sein duquel sont synthétisées différentes valeurs et fonctionnalités, que l'utilisateur pourra ensuite consulter ou appliquer pour en tirer des conclusions effectives sur ses données. Par exemple, si l'on souhaite savoir ce que le modèle pense de la logueur typique des sépales de* I. setosa*, on peut lui demander :*

```
> exp(analyse$means)["setosa", "Log_long_sep"]
[1] 4.993841
```

Ce résultat signifie que cette longueur typique vaut 4,99 cm*.*

4. À quoi correspond, dans le jargon technique, la « longueur typique » de sépales de *I. setosa* que le logiciel a calculée comme valant 4,99 cm ? Si on en croit le nom « `means` »

[‖]. Rappelons au passage que les conventions utilisées par le cours ne sont que des choix destinés à faciliter la compréhension, mais que ces choix ne sont en rien *obligatoires* : la plupart des références que vous rencontrerez dans votre vie d'ingénieur(e) ne respecteront pas la totalité de ces conventions, et il se peut très bien que l'énoncé de l'examen choisisse de ne pas les respecter par endroits pour éviter de vous "mâcher le travail"... !

[**]. Au S7, vous verrez un cas particulier de ce modèle, appelé « analyse discriminante linéaire », qui correspond au cas où le modèle suppose l'égalité entre $\boldsymbol{\Gamma}_{\mathrm{S}}$, $\boldsymbol{\Gamma}_{\mathrm{C}}$ et $\boldsymbol{\Gamma}_{\mathrm{G}}$.

[††]. Cette séparation entre apprentissage et application peut aussi se retrouver en intelligence artificielle : cela correspond au cas où le modèle d'IA s'entraine sur un grand nombre de données de qualité bien contrôlée, puis où ses paramètres sont figés une fois pour toutes lorsqu'on le déploie pour les utilisateurs. Le célèbre système *ChatGPT*, par exemple, fonctionnait selon une telle séparation lors de son lancement. L'inconvénient est qu'on se prive de l'apport de ce que font effectivement les utilisateurs avec cette IA ; l'avantage est qu'en l'absence d'une telle séparation, il serait possible à des utilisateurs malintentionnés de faire "désapprendre" ce qu'elle a appris : l'empêcher apporte donc une certaine garantie de qualité...

de la rubrique qu'on a consultée dans l'objet `analyse`, comment cette valeur a-t-elle été calculée ? À quelle technique vue en cours cela correspond-il ?

Essayons maintenant de comprendre ce que notre calcul pense de la matrice $\boldsymbol{\Gamma}_{\mathrm{C}}$. *Là, c'est plus compliqué... On peut lire dans la documentation de la bibliothèque* `MASS` *un passage où il est question de matrice des covariances :*

```
scaling: for each group 'i', 'scaling[,,i]' is an array which
         transforms observations so that within-groups covariance
         matrix is spherical.
```

On peut alors exécuter la lecture suivante :

```
> analyse$scaling[ , , "versicolor"]
                    1          2          3          4
Log_long_sep -11.46667   7.136408  12.065309  -3.311445
Log_larg_sep   0.00000 -10.055573   2.992892   5.436870
Log_long_pet   0.00000   0.000000 -13.995850  11.172778
Log_larg_pet   0.00000   0.000000   0.000000 -11.914679
```

5.$^{\bigstar}$ Appelons $\boldsymbol{R}$ la matrice ci-dessus. Comme subodoré ci-dessus, $\boldsymbol{R}$ est effectivement liée à l'estimation $\hat{\boldsymbol{\Gamma}}_{\mathrm{C}\checkmark}$ — même s'il ne s'agit évidemment pas de $\hat{\boldsymbol{\Gamma}}_{\mathrm{C}\checkmark}$ *elle-même*, puisque ce n'est pas une matrice symétrique ! Quel est, à votre avis, le lien entre $\boldsymbol{R}$ et $\hat{\boldsymbol{\Gamma}}_{\mathrm{C}\checkmark}$?...

Maintenant, on se penche sur la question de la détermination de la fleur que nous avons cueillie. Les dimensions mesurées pour celle-ci sont $(\mathfrak{x}_{3n\checkmark}, \mathfrak{w}_{3n\checkmark}, \mathfrak{v}_{3n\checkmark}, \mathfrak{u}_{3n\checkmark}) = (4{,}9, 3{,}0, 3{,}1, 1{,}7)$ cm. *Pour identifier la fleur à l'aide de l'analyse discriminante quadratique, on procède comme suit :*

```
> # Jeu de données à tester avec les dimensions de la nouvelle fleur
> df.test = data.frame(Long_sepales_cm = 5.0, Larg_sepales_cm = 3.0,
+                      Long_petales_cm = 3.1, Larg_petales_cm = 1.8)
> # Calcul des logarithmes
> df.test$Log_long_sep = log(df.test$Long_sepales_cm)
> df.test$Log_larg_sep = log(df.test$Larg_sepales_cm)
> df.test$Log_long_pet = log(df.test$Long_petales_cm)
> df.test$Log_larg_pet = log(df.test$Larg_petales_cm)
> # Prédiction du modèle...
> predict(analyse, df.test)
$class
[1] versicolor
Levels: setosa versicolor virginica

$posterior
       setosa versicolor virginica
1 0.009795703   0.876036 0.1141683
```

6. L'apparition du mot « posterior » dans la réponse du logiciel nous laisse entendre qu'on a ici procédé à une analyse *bayésienne*... Et c'est effectivement le cas, même si dans le cadre

de cet exercice nous ne nous apesantirons pas sur la priore utilisée. Quel est, à votre avis, le modèle utilisé ? On précisera bien, en particulier, qui est le (ou les) paramètre(s) caché(s).

7.★ Au vu des questions précédentes, il s'avère donc que logiciel a suivi une approche fréquentiste concernant les données de la collection, puis bayésienne pour déterminer l'espèce de la nouvelle fleur. Quand on met les deux étapes bout à bout, au final, le traitement opéré est-il fréquentiste ou bayésien ?...

On peut maintenant se demander en quoi aurait consisté un traitement mathématique où nous aurons directement *considéré le modèle comme un tout,* sans *séparer l'étape relative à la collection de celle relative à la nouvelle fleur. Dans ce cas, un calcul qui aurait pu être fait aurait été de calculer la* vraisemblance *des différentes hypothèses* $\{\boldsymbol{\varepsilon} = \mathrm{S}\}$, $\{\boldsymbol{\varepsilon} = \mathrm{C}\}$ *et* $\{\boldsymbol{\varepsilon} = \mathrm{G}\}$.

8.❀ Pour $\boldsymbol{\gamma} \coloneqq \gamma(\boldsymbol{\theta})$ une quantité d'intérêt explicative (à valeurs dans un certain espace $\mathscr{G}$), rappeler comment est définie la fonction de vraisemblance de la quantité d'intérêt $\boldsymbol{\gamma}$, i.e. la fonction $\gamma_* \mapsto \mathscr{L}(\boldsymbol{\gamma} = \gamma_*)$ (définie sur $\mathscr{G}$, à valeurs dans $\mathbb{R}_+$).

9. Expliquer avec suffisamment de détail, mais sans résoudre les questions d'optimisation mathématique que cela soulève, la procédure qu'il faudrait suivre pour déterminer les vraisemblances respectives des trois hypothèses concernant la nouvelle fleur.

(10).★★ Résoudre les problèmes d'optimisation sous forme théorique, puis, à l'aide de l'ordinateur, procéder aux applications numériques pour mener à bien le calcul de la vraisemblance pour les hypothèses ci-dessus.

Indication : Pour la suite de l'énoncé, mettons qu'on doit ait trouvé des log-vraisemblances (en base $\mathrm{e}^{1/2}$) de respectivement $-19{,}1$ pour l'espèce S, $-9{,}0$ pour l'espèce C et $-13{,}8$ pour l'espèce G.

11. Un élève ayant calculé les log-vraisemblances ci-dessus propose d'aller plus loin dans sa conclusion, en écrivant le raisonnement suivant :

> « Si je suppose à priori que ma fleur avait la même probabilité d'appartenir à chacune des trois espèces, la probabilité à postériori qu'elle appartienne resp. aux espèces S, C ou G sera proportionnelle aux vraisemblance des hypothèses respectives correspondants : ce qui me donne des probabilités à postériori de resp. 0,58 %, 8,3 % et 91,1 % ».

Pourquoi ce raisonnement n'est-il, mathématiquement parlant, pas correct ?

12. De manière générale, qu'est-ce qui rendrait particulièrement délicat de procéder à une analyse bayésienne pour ce modèle, pris "en bloc" ?

Nous voyons donc que l'approche en deux temps suivie, si elle n'est pas très "propre" du point de vue mathématique, est néanmoins considérablement plus facile à comprendre et à implémenter : ce qui explique pourquoi c'est souvent elle qui est proposée par les logiciels...

13.★ Dans quel genre de cas peut-on s'attendre à ce que le caractère approximatif de l'approche en deux temps soit en fait innocent, les approximations fournies étant en fait

quasiment exactes ? À quel type d'approche statistique, mentionnée à plusieurs reprises dans le cours, cela se raccorde-t-il ?

EXERCICE 3 — Sélection d'emprunteurs et sélection de variables

Un établissement bancaire décide de changer sa politique de prêts aux particuliers : alors que, jusque-là, la banque requérait que les prêts avancés à ses clients fussent couverts par une garantie (patrimoine immobilier, engagement d'un tiers, ...), la banque envisage à présent de prêter sans exiger de garantie de la part de ses emprunteurs ! Bien entendu, dans ce cadre, le risque que l'emprunteur ne soit pas en mesure de rembourser sa dette (on parle alors de « défaut de paiement ») devient beaucoup plus problématique ; et la banque souhaite donc ne prêter qu'aux clients ayant les dossiers les plus "solides". Mais comment évaluer la « solidité » d'un dossier ?... C'est ici qu'intervient la statistique ! ⌣

Nous considérons ici que le « dossier » d'un emprunteur potentiel (appelé e) est décrit par un certain nombre (disons, $26 =: p$) de variables quantitatives ([logarithme du] montant du prêt demandé, durée du prêt, [logarithme du] revenu annuel de l'emprunteur, durée depuis laquelle l'mprunteur potentiel est client de la banque, âge de l'emprunteur, nombre d'enfants, [logarithme de la] taille de la commune de résidence, statut salarié ou non[‡‡]*, ...), la valeur de la variable numéro j pour l'emprunteur e étant notée $x_j^{(e)}$.*

Pour savoir à quel point le dossier d'un emprunteur est susceptible de présenter un risque de défaut, la banque s'appuie sur les prêts accordés au cours des années précédentes (pour lesquels il y avait un nombre assez élevé de défauts[∗]*, car à l'époque on pouvait toujours se rabattre sur la garantie le cas échéant). On introduit alors, pour chaque variable e, la variable $Y^{(e)}$ (dont on connait la réalisation au moment de l'étude : c'est une observation* passée*), définie comme valant 1 si l'emprunteur e a fait défaut, et 0 sinon. Les $x_j^{(e)}$, quant à eux, sont vus comme des paramètres du modèle.*

Le modèle utilisé par la banque pour expliquer les défauts en fonction des dossiers est le suivant : on introduit $(p+1)$ paramètres cachés, dont les valeurs effectives sont notées resp. $\alpha_{0\checkmark}, \dots, \alpha_{(p-1)\checkmark}, \beta_\checkmark$; et on considère que, sous la véritable loi $\mathbb{P}_\checkmark$, les différents $Y^{(e)}$ sont indépendants, avec

$$\mathbb{P}_\checkmark(Y^{(e)} = 1) = 1/\left(1 + \exp\left(\sum_{j=0}^{p-1} \alpha_{j\checkmark} x_j^{(e)} + \beta_\checkmark\right)\right).$$

[‡‡]. Le fait d'être salarié ou non est évidemment un critère *qualitatif* ; néanmoins rien n'empêche de le *coder* en associant la valeur 0 au fait d'être non salarié et la valeur 1 au fait d'être salarié. Attention toutefois, ce genre de codage ne fonctionne que s'il n'existe que *deux* modalités à considérer ! Si nous avions affaire à trois modalités qualitatives (p.ex. « vanille », « chocolat » ou « fraise »), coder resp. « vanille », « chocolat » et « fraise » par (mettons) 0, 1 et 2 n'aurait aucun sens, car on ne peut pas dire que « chocolat » soit la moyenne de « vanille » et de « fraise » ! En fait, il est tout de même possible d'encoder quantitativement des variables qualitatives à $k > 2$ modalités, mais à condition de le faire dans un espace de dimension $\geqslant k-1$: ainsi, si on considère les indicateurs « vanillitude », « chocolatitude » et « fraisitude », nos trois parfums seront encodés resp. $(1,0,0)$, $(0,1,0)$ et $(0,0,1)$ ⌣

[∗]. Le fait que les défauts soient assez fréquents sur les données dont on dispose est utile du point de vue de l'analyse statistique, car notre but est de distinguer les dossiers risquant le défaut des dossiers solides : on a donc tout intérêt à avoir des exemples variés des deux situations... ! ⌢

(Il s'agit là d'un modèle extrêmement classique pour ce genre d'applications, qu'on appelle le modèle de la régression logistique [†] *).*

Les quatre premières questions de cet exercice discutent de la façon d'estimer les paramètres $\boldsymbol{\alpha}_j$ et $\boldsymbol{\beta}$ par maximum de vraisemblance.

(1). Écrire la fonction de log-vraisemblance pour ce modèle.

Indication : Maintenant que vous êtes "grands", il y aura éventuellement des notations à introduire par vous-mêmes...! ☿

(2).* Observer que cette fonction de log-vraisemblance est concave sur $\mathbb{R}^{p+1}$.

Indication : On pourra utiliser avec profit les trois propriétés suivantes : (1°) $1-1/(1+\mathrm{e}^{x}) = 1/(1+\mathrm{e}^{-x})$; (2°) la somme de plusieurs fonctions concaves est concave ; (3°) si $f\colon \mathbb{R}^{d} \to \mathbb{R}$ est une fonction concave et $g\colon \mathbb{R}^{d'} \to \mathbb{R}^{d}$ une fonction affine, alors la fonction $f \circ g$ est concave.

(3).$^{\text{❀}}$ En quoi la propriété de concavité de la log-vraisemblance est-elle une bonne nouvelle du point de vue de l'utilisation numérique de ce modèle ?

(4).** Donner un jeu de valeurs pour les $y_{j\checkmark}^{(e)}$ pour lesquelles la fonction de vraisemblance n'atteint nulle part son supremum. Énoncer cependant une condition suffisante relativement "souple" sous laquelle on a l'assurance que le maximum de vraisemblance sera bien atteint. (Nous admettrons qu'en pratique, pour des jeux de données "raisonnables", cette condition est bien satisfaite).

Dans la suite, nous considèrons que nous savons calculer efficacement, pour ce modèle, les estimateurs du maximum de vraisemblance pour l'ensemble des composantes du paramètre caché.

On s'intéresse maintenant à une cliente ⋆ qui vient de solliciter un prêt (sans garantie cette fois-ci), cliente dont la banque connait le dossier (les caractéristiques de ce dossier étant vues comme des paramètres du modèle) : la banque se demande alors, au vu du dossier de la cliente ⋆, s'il est opportun de lui accorder, ou pas, le prêt qu'elle demande. Du point de vue de la modélisation statistique, on considère le fait que la cliente ⋆ fasse défaut ou pas (en supposant qu'on lui accorde le prêt, s'entend) comme une observation future $\boldsymbol{Y}^{(\star)}$, qui est liée aux $x_j^{(\star)}$ exactement de la même façon que les $\boldsymbol{Y}^{(e)}$ étaient liés aux $x_j^{(e)}$ pour les observations passées.

5.$^{\text{❀}}$ Soit $\ell(\bullet,\bullet)$ la fonction de perte définie sur $\{0,1\}^2$ par :

$$\ell(x,\hat{x}) = \begin{cases} 0 & \text{si } \hat{x} = x\,; \\ 1 & \text{si } x = 0 \text{ et } \hat{x} = 1 \text{ (cas d'un « faux positif »)}\,; \\ A & \text{si } x = 1 \text{ et } \hat{x} = 0 \text{ (cas d'un « faux négatif »)}, \end{cases}$$

[†]. De manière intéressante, même pour les modèles plus "modernes" à base de réseaux de neurones, on utilise souvent *quand même* une couche de régression logistique en sortie du modèle, même si ce n'est pas présenté sous ce vocable — dans ce contexte, on dit plutôt qu'on applique la « fonction softmax » à la sortie du réseau de neurones. En fait, le modèle de la régression logistique peut être vu comme un cas particulier de réseau de neurones sans couche intermédiaire (et avec, en l'occurrence, 2 neurones dans la couche de sortie : un décrivant le défaut, et l'autre l'absence de défaut.

pour un certain $A \in \mathbb{R}_+^*$ connu, en général pris largement supérieur à 1 (penser p. ex. à $A = 20$). Dans ce contexte, pour X une v.a. suivant la loi Bernoulli(p), dire à quelle condition sur p (et A) est-ce que la fonction $\hat{x} \mapsto \mathbb{E}\big(\ell(X, \hat{x})\big)$ sera minimale en $\hat{x} = 0$.

6. De la question précédente, déduire un prédicteur pertinent pour $Y^{(\star)}$, associé à la fonction de perte de la question précédente.

Indication : On se remémorera avec profit le passage « de l'estimation à la prédiction » de la présentation en amphithéâtre du 28 avril...

7. Quel est le rapport entre le prédicteur et la décision de la banque de décider, ou pas, d'accorder le prêt à la cliente $\star$? Quel rôle joue la fonction de perte dans ce contexte ; et en particulier, comment faut-il interpréter la valeur choisie pour A ?

8. Pourquoi ne serait-il pas correct de dire que « la probabilité (à postériori, au vu des données passées dont dispose la banque) que la cliente $\star$ fasse défaut vaut $1/\big(1 + \exp\big(\sum_{j=0}^{p-1} \hat{\alpha}_{j\checkmark}^{\mathrm{mv}} x_j^{(\star)} + \hat{\beta}_{\checkmark}^{\mathrm{mv}}\big)\big)$ », pas plus qu'il ne serait correct de dire que « l'"intervalle" de prédiction pour $Y_\star$ consistant à renvoyer $\{0\}$ lorsque $1/\big(\exp\big(\sum_{j=0}^{p-1} \hat{\alpha}_{j\checkmark}^{\mathrm{mv}} x_j^{(\star)} + \hat{\beta}_{\checkmark}^{\mathrm{mv}}\big)\big) \leqslant \alpha$, resp. $\{0, 1\}$ sinon, est fréquentistement valable au niveau de risque α » ?

9. Quelles conséquences pratique devrait entrainer la constatation soulevée à la question ci-dessus du point de vue de la politique de la banque ?

10. Indépendamment de la constatation de la question 8, quelle autre raison (encore plus importante !) doit rendre la banque prudente vis-à-vis des résultats de son analyse statistique ?...

Dans la fin de cet exercice, la banque se demande si, plutôt que d'utiliser un modèle où 26 *facteurs sont pris en compte pour étudier le dossier de chaque emprunteur potentiel, on ne pourrit pas utiliser un modèle plus simple, ou seule une partie de ces facteurs seraient pris en compte... Outre que cela simplifierait la constitution des dossiers, les statisticiens de la banque l'ont en effet alertée sur le fait qu'utiliser trop de facteurs différents pouvait, paradoxalement,* dégrader *la qualité des prédictions effectuées*[‡] *! Bref ; on souhaite donc se débarrasser des facteurs j qui n'apportent rien de probant au modèle.*

Des logiciels de statistique comme R *permettent d'obtenir une analyse du modèle de régression logistique, analyse qui fournit notamment l'estimation du maximum de vraisemblance pour les différents paramètres cachés du modèle, ainsi que des « tests de significativité » pour chacun des facteurs associés à ces paramètres cachés. Par exemple, si nous appelons « Facteur A », ..., « Facteur Z » les éléments du dossier bancaitre associés aux indices resp. $j = 0$, ..., $j = 25$, on pourra obtenir, via le logiciel, un résultat comme celui-ci :*

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.338010   5.716762  -1.109   0.2676
Facteur A    0.002668   0.001674   1.594   0.1110
```

[‡]. C'est ce qu'on appelle le phénomène de *surapprentissage*, que vous étudierez l'année prochaine.

```
Facteur B    0.147992   0.212364   0.697   0.4859
Facteur C    0.021770   0.928327   0.023   0.9813
Facteur D    0.234944   2.670193   0.088   0.9299
Facteur E    0.002260   0.001091   2.071   0.0384 *
Facteur F   -1.340203   0.345317  -3.881   1.0e-4 ***
Facteur G   -1.818404   0.832570  -2.184   0.0289 *
Facteur H    0.239472   0.137345   1.744   0.0812 .
Facteur I    0.042303   0.006570   6.439  1.2e-10 ***
Facteur J   -0.131607   0.350234  -0.376   0.7071
Facteur K    0.079380   0.268305   0.296   0.7673
Facteur L   -0.012916   0.193211  -0.067   0.9467
Facteur M   -0.675441   0.316495  -2.134   0.0328 *
Facteur N    0.001902   0.000704   2.700   0.0069 **
Facteur O    0.045718   0.062184   0.735   0.4620
Facteur P   -2.043545   1.115824  -1.831   0.0670 .
Facteur Q   -4.713474   2.864827  -1.645   0.0999 .
Facteur R   -1.551464   0.417839  -3.713   1.9e-4 ***
Facteur S    1.813444   1.310591   1.384   0.1664
Facteur T   -0.002815   0.005317  -0.529   0.5965
Facteur U    2.028505   1.883201   1.077   0.2814
Facteur V    2.490210   1.641601   1.517   0.1293
Facteur W    0.015926   0.008324   1.913   0.0557 .
Facteur X    0.804042   0.331823   2.423   0.0153 *
Facteur Y   -4.548428   2.071250  -2.196   0.0280 *
Facteur Z    0.013706   0.003870   3.541   3.4e-4 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11.★ Dans le résultat d'analyse ci-dessus, identifier :

(i) Les informations relatives à l'analyse du paramètre caché $\boldsymbol{\beta}$;

(ii) L'estimation du paramètre caché dans $\mathbb{R}^{p+1}$ (en l'occurrence, il s'agit de l'estimation du maximum de vraisemblance, confer question 3) ;

(iii) Les quantités calculées par le logiciel qui décrivent des p-valeurs ;

(iv) Les hypothèses nulles respectives des tests auxquels correspondent ces p-valeurs.

Au vu des résultats affichés par R*, un élève affirme que « s'il y a bien* un *élément dont on peut se passer dans l'étude du dossier bancaire des clients, c'est le facteur C ». Sa voisine de table, pour sa part, estime au contraire l'élément le plus superflu est le facteur I !*

12. Qui a raison ? (Et pourquoi ?).

13.★ En fait, l'élève que nous avons mentionné ci-dessus (resp. sa voisine) va même jusqu'à considérer que l'analyse du logiciel prouve qu'on pourrait contenter des facteurs E, F, G, H, I, M, N, P, Q, R, W, X, Y et Z (et sa voisine pense, pour sa part, qu'on pourrait de contenter des facteurs A, B, C, D, H, J, K, L, P, Q, S, T, U, V et W). Mais pour le coup, ils ont tous les deux tort !... Pourquoi ?