

Feuille 1 — Normales, Poisson et binomiales

Groupe inversé : les exercices 1, 2, 3 sont à traiter en autonomie. Les exercices 4, 5 et éventuellement début du 6 sont à traiter en séance encadrée.

Groupe classique : les exercices 4, 5 sont à traiter en séance encadrée. Les exercices 1, 2, 6 sont à traiter en autonomie.

Pour tous les groupes : un exercice du type de l'exercice 6 devra être rédigé et rendu lors du TD sur campus suivant, un énoncé individualisé sera envoyé à chaque étudiant.

Exercice 1 — Espérance et variance.

Soient X et Y deux variables aléatoires indépendantes, d'espérance 2 et de variance 1. En utilisant les propriétés de l'espérance et de la variance vues en cours, calculer $\mathbb{E}(5X - 4Y)$ et $\mathbb{V}(5X - 4Y)$.

Exercice 2 — Lecture tables.

- Soit X une variable aléatoire de loi $\mathcal{P}(8)$.
 - Que vaut $\mathbb{P}(X \leq 8)$? $\mathbb{P}(X = 8)$? $\mathbb{P}(X > 8)$? $\mathbb{P}(X > 9)$?
 - Trouver le plus petit entier u tel que $\mathbb{P}(X \leq u) \geq 85\%$.
 - Trouver le plus petit entier v tel que $\mathbb{P}(X > v) \leq 10\%$.
 - Trouver le plus grand entier w tel que $\mathbb{P}(X \geq w) \geq 70\%$.
- Soit X une variable aléatoire de loi $\mathcal{B}(20; 0,5)$.
 - Que vaut $\mathbb{P}(2 < X \leq 6)$?
 - Que vaut $\mathbb{P}(5 \leq X \leq 15)$?
 - Que vaut $\mathbb{P}(X = 13)$?
 - Quel est le plus grand entier v tel que $\mathbb{P}(X \geq v) \geq 90\%$?
- Soit X une variable aléatoire de loi $\mathcal{N}(0; 1)$.
 - Que vaut $\mathbb{P}(X \leq 1,63)$?
 - Que vaut $\mathbb{P}(X \geq 0,53)$?
 - Que vaut $\mathbb{P}(X \leq -1,14)$?
 - Que vaut $\mathbb{P}(|X| \geq 1,27)$?
 - Trouver le réel u tel que $\mathbb{P}(-u \leq X \leq u) = 87\%$.

Exercice 3 — Calcul des probabilités pour une loi normale quelconque.

Soit X une variable aléatoire de loi normale $\mathcal{N}(178; 3,1)$.

- En centrant et réduisant X (cours), calculer $\mathbb{P}(X \geq 180)$.
- Soit maintenant X_1, X_2, X_3 des variables aléatoires indépendantes et de même loi que X .

Cours : donner la loi de $\bar{X} := \frac{\sum_{i=1}^3 X_i}{3}$.

En déduire $\mathbb{P}(\bar{X} \geq 180)$.

Exercice 4 — Normale(s).

La mesure de la concentration d’ozone dans l’air (en $\mu\text{g}/\text{m}^3$) est modélisée par une variable aléatoire X de loi $\mathcal{N}(m ; \sigma^2)$ avec $\sigma^2 = 3,1$. Un niveau d’alerte pollution est franchi si la concentration moyenne d’ozone dans l’air dépasse $180 \mu\text{g}/\text{m}^3$.

1. Quelle est l’unité de m ? de σ ? Que représentent m et σ ?
2. On effectue des mesures un jour donné, et l’on suppose que ce jour là la concentration moyenne d’ozone dans l’air est de $178 \mu\text{g}/\text{m}^3$ (mais l’expérimentateur ne le sait pas sinon il n’aurait pas besoin de faire des mesures).
 - (a) Quelle est la probabilité qu’une mesure unique soit supérieure à 180 ?
 - (b) Quelle est la probabilité que la moyenne de trois mesures soit supérieure à 180 ?
 - (c) (**question facultative, selon le temps restant**) Combien de mesures faut-il réaliser pour que la probabilité que la moyenne de ces mesures dépasse 180 soit inférieure à 1% ?

Exercice 5 — Binomiale, Normale, Poisson.

Dans une forêt, les champignons non comestibles sont quatre fois plus nombreux que les champignons comestibles.

1. On cueille au hasard 20 champignons. Quelle est la probabilité de trouver parmi ces 20 champignons un nombre de champignons non comestibles inférieur ou égal à 13 ?
2. On cueille au hasard 100 champignons. Quelle est la probabilité que sur les 100 champignons cueillis, le nombre de champignons comestibles soit compris entre 15 et 20 ?
3. On cueille au hasard 150 champignons. Parmi les champignons comestibles, un dixième sont des cèpes. Quelle est la probabilité que sur les 150 champignons cueillis, au moins cinq soient des cèpes ?
4. (**question facultative, selon le temps restant**) Combien faut-il cueillir de champignons pour que la probabilité d’avoir dans son panier 40 champignons comestibles soit supérieure ou égale à 90% ?

Exercice 6 — Binomiale et Normale.

Chez l’Alligator d’Amérique, la proportion de mâles et de femelles varie en fonction de la température durant une période de l’incubation (entre le 7^{ème} et le 21^{ème} jour). La température moyenne relevée dans un marécage durant cette période étant de 33°C , la proportion théorique de femelles est 25% . On observe dans ce marécage, les animaux issus de 100 oeufs. On note S le nombre de mâles issus de ces 100 oeufs.

Quelle est la loi de S ? Son espérance ? Sa variance ? Peut-on approximer la loi de S ?

Dans les questions suivantes, on pourra ou non utiliser la correction de continuité. Son utilisation correcte est comptée comme un point bonus.

Quelle est la probabilité d’observer au moins 80 mâles ?

Quelle est la probabilité que le nombre de mâles soit compris (au sens large) entre 68 et 82 ?

Feuille 2 — Tests gaussiens

Groupe inversé : l'exercice 1 est à traiter en autonomie. L'exercice 2 question 2 et l'exercice 3 sont à traiter en séance encadrée.

Groupe classique : l'exercice 2 en entier et éventuellement le début de l'exercice 3 sont à traiter en séance encadrée. L'exercice 3 est à finir en autonomie.

Pour tous les groupes : un exercice du type de l'exercice 4 devra être rédigé et rendu lors du TD sur campus suivant, un énoncé individualisé sera envoyé à chaque étudiant.

Exercice 1 — Gaussien en autonomie.

Lors d'une réunion sportive, C.L. et B.J. subissent un contrôle sanguin inopiné : on mesure leur taux d'hématocrite (en %). Les médecins estiment que la technique de mesure est précise à environ 2% près. Ils savent par ailleurs qu'un individu normal a un taux d'hématocrite de 45%, mais que certains produits permettent de l'augmenter et d'améliorer ainsi ses performances. On note τ_1 (resp. τ_2) le vrai taux de C.L. (resp. B.J.). Ci-dessous est construit un test statistique en 7 points.

1. Expliquer pourquoi on modélise les mesures de C.L. et de B.J. par des variables aléatoires X_1 et X_2 de loi gaussienne centrée respectivement en τ_1 et τ_2 , et de variance $\sigma^2 = 4$.

Par la suite, on note X l'une quelconque des deux mesures, et τ le vrai taux d'hématocrite de la personne associée à la mesure.

2. Parmi les trois jeux d'hypothèses ci-dessous, choisir le jeu d'hypothèses pertinent.

$$\begin{array}{lll} H_0 : \tau = 45 & \text{contre} & H_0 : \tau > 45 \\ H_0 : \tau = 45 & \text{contre} & H_0 : \tau < 45 \\ H_0 : \tau = 45 & \text{contre} & H_0 : \tau \neq 45 \end{array}$$

3. La statistique de test est au choix X ou sa version centrée réduite sous H_0 que l'on notera Y :

$$Y := \frac{X - 45}{2}.$$

Quelle est la loi de X sous H_0 ? Quelle est la loi de Y sous H_0 ?

4. * Sur le même repère, représenter l'allure de la densité de X sous H_0 et l'allure de la densité de X sous H_1 .

* Sur le même repère, représenter l'allure de la densité de Y sous H_0 et l'allure de la densité de Y sous H_1 .

* Choisir (en fonction de X ou de Y) la forme de la région de rejet de H_0 .

Rappel : les trois formes possibles de région de rejet vues en cours sont rejet à gauche, rejet à droite ou rejet bilatéral.

5. Les médecins ont choisi pour ce test statistique un niveau 1% (pourquoi pas 5% ?). Calculer le(s) seuil(s) de la région de rejet de H_0 pour le niveau 1%.

6. Revenons au cas de C.L. et de B.J. pour qui le résultat des mesures est le suivant : $X_1^{obs} = 49$ et $X_2^{obs} = 50$. Calculer Y_1^{obs} et Y_2^{obs} . Expliquer quelle décision doivent prendre les médecins en fonction de ces observations. On répondra par deux méthodes :

- en comparant X^{obs} (ou Y^{obs} selon la statistique de test choisie) au seuil de la région de rejet de H_0 associée au niveau 1% (méthode 1) ;

- en calculant la p-valeur associée à ces observations (méthode 2).

7. Rédiger une conclusion.

Exercice 2 — Gaussien unilatéral.

1. Lors d'une réunion sportive, C.L. et B.J. subissent un contrôle sanguin inopiné : on mesure leur taux d'hématocrite (en %). Les médecins estiment que la technique de mesure est précise à environ 2% près. Ils savent par ailleurs qu'un individu normal a un taux d'hématocrite de 45%, mais que certains produits permettent de l'augmenter et d'améliorer ainsi ses performances. On note τ_1 (resp. τ_2) le vrai taux de C.L. (resp. B.J.).

Expliquer pourquoi on modélise les mesures de C.L. et de B.J. par des variables aléatoires X_1 et X_2 de loi gaussienne centrée respectivement en τ_1 et τ_2 , et de variance $\sigma^2 = 4$.

Expliquer comment les médecins doivent prendre leur décision en fonction des observations, et appliquer la procédure proposée aux cas de C.L. et B.J. pour les valeurs observées suivantes : $X_1^{obs} = 49$, $X_2^{obs} = 50$.

2. Après un contrôle sanguin ayant déclaré leur taux d'hématocrite trop élevé, C.L. et B.J. demandent une contre-expertise. Neuf mesures du taux d'hématocrite de C.L. vont être réalisées ainsi que neuf mesures du taux d'hématocrite de B.J.. La technique de mesure est précise à environ 2% près et les mesures sont indépendantes.

Quelle variable utiliser pour savoir si les taux d'hématocrite de C.L. et de B.J. sont normaux ?

Décrire le test statistique réalisant la contre-expertise.

Que conclut ce test pour les observations suivantes :

C.L.	47	46	45	50	44	46	49	45	46
B.J.	48	50	49	46	45	47	46	48	44

Exercice 3 — Student bilatéral.

On considère qu'en moyenne, la tension artérielle systolique au repos des français est 12. On la mesure chez 28 sujets montpelliérains et on obtient les résultats suivants :

Somme des 28 mesures : 363,67 et Somme des carrés des 28 mesures : 4880,1818

On veut tester le fait que la tension artérielle systolique moyenne au repos des montpelliérains est identique à celle des français. D'après les 28 observations dont nous disposons, quelle est la conclusion du test de niveau 5% ?

Indication pour les étudiants en autonomie : la variance de la tension artérielle systolique au repos est ici inconnue, il faut utiliser la statistique de Student. Pour calculer la valeur observée de cette statistique, il faut calculer la valeur observée de la variance empirique S^2 . On utilisera la seconde écriture de la formule suivante :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} (\bar{X})^2.$$

Exercice 4 — Gaussien bilatéral.

On considère qu'en moyenne, la tension artérielle systolique au repos des français est 12 et que l'écart-type associé est 2. On la mesure chez 28 sujets montpelliérains et on obtient les résultats suivants :

Somme des 28 mesures : 363,67

On veut tester le fait que la tension artérielle systolique moyenne au repos des montpelliérains est identique à celle des français. On suppose que l'écart-type associé à la tension artérielle systolique au repos est la même pour les montpelliérains que pour les français pris dans leur ensemble. D'après les 28 observations dont nous disposons, quelle est la conclusion du test de niveau 5% ?

Feuille 3 — Tests du χ^2

Groupe inversé : L'exercice 1 et l'exercice 2 sont à traiter en autonomie. L'exercice 3 et l'exercice 4 sont à traiter en séance encadrée.

Groupe classique : L'exercice 3 et l'exercice 4 sont à traiter en séance encadrée. L'exercice 1 et l'exercice 2 sont à traiter en autonomie.

Pour tous les groupes : un exercice avec un des trois tests du chi-deux vus en cours (chi-deux d'adéquation, chi-deux d'indépendance avec effectifs espérés sous H_0 inconnus, chi-deux d'indépendance avec effectifs espérés sous H_0 connus) devra être rédigé et rendu à son enseignant, un énoncé individualisé sera envoyé à chaque étudiant.

Exercice 1 — à traiter en autonomie (chi-deux d'adéquation)

La structure d'âge de la population française est donnée en pourcentage dans le tableau ci-dessous :

Classe d'âge	0-19	20-39	40-59	60-74	> 74
Pourcentage	24,6	28,1	26	13,6	7,7

Données issues de *Statistique et Epidémiologie*, T.Ancelle, ed. Maloine.

On veut vérifier qu'un sous-groupe particulier de la population française noté \mathcal{G} n'a pas une structure d'âge significativement différente de la structure d'âge de la population française. Pour cela, on tire au hasard $n = 284$ individus du sous-groupe \mathcal{G} et on relève à quelle classe d'âge ils appartiennent.

- * On appellera 1, 2, 3, 4, 5 les 5 classes d'âge et on notera j l'indice de classe ($j = 1, \dots, 5$). On définit N_j les variables aléatoires égales au nombre de sujets tombant dans la classe d'âge j parmi les $n = 284$ sujets de \mathcal{G} considérés. Les variables N_j suivent des lois binomiales $\mathcal{B}(n = 284 ; p_j)$, où p_j est le pourcentage (inconnu) d'individu de \mathcal{G} dans la classe d'âge j . Ces variables N_j sont appelées les effectifs observables.
- * On va tester H_0 : la structure d'âge de \mathcal{G} n'est pas significativement différente de celle de la population française contre H_1 : la structure d'âge de \mathcal{G} est significativement différente de celle de la population française.

Exprimer mathématiquement ces deux hypothèses à l'aide d'égalité ou d'inégalité sur les paramètres p_j et des opérateurs logiques ET et OU.

- * Quel est le nom du test statistique à réaliser ici ? On notera Z la statistique de test.

Si la structure de \mathcal{G} n'est pas significativement différente de celle de la population française, quel est l'effectif « espéré » de chaque classe d'âge dans un échantillon de $n = 284$ personnes issues de \mathcal{G} ? Reporter ces effectifs, notés n_j dans le tableau suivant :

Classe d'âge	0-19	20-39	40-59	60-74	> 74
Effectifs espérés sous H_0					

Pourquoi peut-on approximer la loi de Z sous H_0 par une loi du chi-deux ? Préciser les degrés de liberté de cette loi.

- * Dessiner la densité de la loi de Z sous H_0 ainsi que la densité de la loi de Z sous H_1 (décalée à droite et plus aplatie). On rappelle que pour un test du chi-deux, la région de rejet de H_0 est toujours de la forme $\mathcal{R} = \{Z \geq a\}$. Poser a sur l'axe des abscisses et hachurer l'aire entre la courbe de la densité sous H_0 et l'axe des abscisses sur l'intervalle $[a ; +\infty[$. Que représente cette aire ?

- * Pour un risque de première espèce (ou niveau) égal à 0,05, calculer la valeur de a .
- * Conclusion du test. Les effectifs observés N_j^{obs} sur les $n = 284$ individus issus de \mathcal{G} sont présentés dans le tableau suivant :

Classe d'âge	0-19	20-39	40-59	60-74	> 74
Effectifs observés	73	82	75	36	18

Calculer Z_{obs} .

- Comparer Z_{obs} au seuil a précédemment trouvé. Que conclut le test statistique de niveau 5% ?
 - Calculer la p-valeur $\mathbb{P}_{H_0}(Z \geq Z_{obs})$. Que conclut le test statistique de niveau 5% ?
 - Représenter une fois encore la densité de la loi de Z sous H_0 . Poser a et Z_{obs} sur l'axe des abscisses. Hachurer avec des couleurs différentes l'aire entre la courbe de la densité sous H_0 et l'axe des abscisses sur l'intervalle $[a ; +\infty[$ et l'aire entre la courbe de la densité sous H_0 et l'axe des abscisses sur l'intervalle $[Z_{obs} ; +\infty[$.
- Retrouver les deux conclusions précédentes à partir de ce dessin.
- * Conclure en français. Par exemple : d'après cette étude, il n'existe aucun argument permettant d'affirmer que la sous population \mathcal{G} présente une structure d'âge différente de celle connue dans la population française. Ou encore : les 284 individus tirés au hasard dans \mathcal{G} peuvent être considérés comme représentatifs de la population française du point de vue de la structure d'âge.

Exercice 2 — à traiter en autonomie (chi-deux d'indépendance)

La société Mac Quick lance son nouveau produit FolBurger aux USA et en Europe. Elle effectue un sondage dans quatre villes (2 aux USA et 2 en Europe) en demandant une appréciation (Mauvais, Correct, Bon) à chaque consommateur sondé :

	Mauvais	Correct	Bon
San Francisco	29	124	87
New York	74	278	208
Rome	114	277	87
Paris	182	417	123

1. En faisant des tests de niveau 1% :
 - (a) L'appréciation du Folburger par le consommateur dépend-elle de la ville où habite le consommateur ?
 - (b) Les habitants de New-York et de San Francisco ont-ils une appréciation différente du Folburger ?
Indication : reformuler la question en utilisant le mot indépendance.
Indication : ne considérer que les deux premières lignes du tableau.
 - (c) Les habitants de Paris et de Rome ont-ils une appréciation différente du Folburger ?
2.
 - (a) Représenter l'histogramme de l'appréciation des consommateurs de San Francisco, des consommateurs de New York, puis des consommateurs américains.
 - (b) Représenter l'histogramme de l'appréciation des consommateurs de Rome, des consommateurs de Paris, puis des consommateurs européens.
 - (c) Commenter ces six histogrammes.

Exercice 3 — Seconde session juin 2014**Seconde session juin 2014, partie I :**

On a repéré la séquence d'ADN suivante

```

C A G C G G A A T C T T C T G A C T C G T G T T C
G C T T T T A C T G A G T T G C C G G G T C A C C
A G G A A G T A C C G C T A A G G T C G C A G G G
G G T T T T G A T A G G T A G C T C C A A T G G G

```

mais on ne connaît pas l'organisme de provenance. On suppose qu'elle peut venir d'un organisme qu'on appellera \mathcal{O} . On sait que pour \mathcal{O} les probabilités théoriques des différentes bases sont :

	A	T	C	G
prob	0,2	0,3	0,24	0,26

Décrire et appliquer un test de niveau 5% pour déterminer si la séquence provient de \mathcal{O} .

Seconde session juin 2014, partie II :

On a repéré les deux séquences d'ADN suivantes :

```

C A G C G G A A T C T T C T G A C T C G T G T T C
G C T T T T A C T G A G T T G C C G G G T C A C C
A G G A A G T A C C G C T A A G G T C G C A G G G
G G T T T T G A T A G G T A G C T C C A A T G G G

```

Echantillon 1

```

G A C G G G G G A A G G G T G A G C G G
G C A A T T C A C C G C T A G G C A A A
G T G C G T C C C T T A G A A A G A A A

```

Echantillon 2

1. Compléter le tableau suivant :

Bases	A	T	C	G
Nombre de bases observées dans le premier échantillon				
Nombre de bases observées dans le second échantillon				

2. Proposer et appliquer un test de niveau 5% pour savoir si les deux séquences ont pour origine le même organisme. *Indication : reformuler la question en utilisant le mot indépendance.*

Exercice 4 — Génétique des grains de maïs (à la demande de l'unité GBMB)

Des croisements entre des plants de maïs appartenant à deux lignées pures différentes (nommées lignée I et lignée II), l'une à grains rouges lisses et l'autre à grains blancs ridés, produisant en première génération des épis à grains rouges lisses, ont donné en deuxième génération ($F_1 \times F_1$) des épis présentant les compositions suivantes :

Population F2 issue du croisement de lignée I par lignée II : 156 grains rouges lisses, 57 grains rouges ridés, 56 grains blancs lisses, 15 grains blancs ridés.

1. **Analyse par caractère.** Quelles données utiliseriez-vous pour étudier le déterminisme génétique de la couleur des graines chez le maïs? Faites un tableau montrant, pour chaque classe phénotypique, les proportions attendues sous l'hypothèse H_0 et les effectifs espérés sous l'hypothèse H_0 .

Il est en principe obligatoire d'établir le protocole d'un test statistique (points 1–5) avant d'appliquer le protocole aux données recueillies. Néanmoins, pour une fois, nous allons déroger à cette règle. Que valent les effectifs observés de chaque classe phénotypique? Quelle est la p-valeur du test?

2. **Relations entre caractères.** On suppose que chaque caractère (couleur et forme du grain) est déterminé par un locus, avec deux allèles. On veut savoir si les deux loci sont indépendants. Donnez le modèle et l'hypothèse H_0 . Faites un tableau montrant, pour chaque classe phénotypique, les effectifs observés et les proportions attendues sous l'hypothèse H_0 , puis réalisez le test.