UE Statistiques pour la biologie : application à des problèmes de Biologie
Résumé du cours - exercices en biologie

N. Castelle, C. Dillmann, M. Gallopin, J. Legrand, E. Marchadier, Y. Pautrat, C. Vassiliadis

October 1, 2024

Contents

1 Probabilités, variables aléatoires **2**

2 Echantillon, n-échantillon **4**

3 Tests d'hypothèses **8**

4 Test Gaussien **12**

5 Test du chideux d'indépendance **13**

6 Séances de TD de bio-statistiques **16**

7 Exercices complémentaires **31**

8 Annexe : Tables de nombres au hasard **36**

La nature même de l'approche expérimentale en Biologie implique que toute mesure est le résultat d'une expérience aléatoire.

- Du fait de la reproduction, chaque être vivant est unique, mais ses caractéristiques ressemblent à celles de ses parents, selon des lois de probabilité définies par le mode de reproduction et la distribution des effets des mutations.
- Tout au long de sa vie, chaque être vivant acquiert des caractéristiques qui lui sont propres en fonction des conditions environnementales qu'il rencontre. Plusieurs individus peuvent partager un environnement commun, mais la succession exacte des conditions qui finiront par déterminer finement les caractéristiques d'un individu est le résultat du hasard.
- Les mesures sont réalisées sur un échantillon d'individus, choisis au hasard dans la population que l'on cherche à caractériser et ne représentent pas l'ensemble de la population.
- Les instruments de mesures ne sont jamais complètement fiables.

Le biologiste doit donc s'appuyer sur des méthodes prenant en compte ces aléas pour analyser les résultats de ses expériences.

Statistiques : *Ensemble des méthodes permettant d'obtenir, de décrire et d'analyser des observations (ou données). Ces observations consistent généralement en la mesure d'une ou plusieurs caractéristiques sur un ensemble d'individus (unités statistiques).*

La *théorie des probabilités* est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude ; la *statistique* est l'activité qui consiste à recueillir, traiter et interpréter un ensemble de données. Avec la théorie des probabilités, les statistiques forment les sciences de l'aléatoire.
Par ailleurs, à l'ère du *Big-Data*, la description du vivant nécessite la mise en place d'approches pluri-disciplinaires et fait largement appel à la modélisation mathématique. Les objectifs de l'UE "Statistiques pour la biologie" sont

1. Apporter des connaissances de base en statistique pour traiter en autonomie des problèmes simples d'analyse de données.
2. Maîtriser l'utilisation d'une écriture formelle. Être capable de traduire une question de biologie en utilisant un formalisme mathématique.

Les cours et Travaux-Dirigés sont dispensés à la fois par des enseignants de Mathématique, et des enseignants de Biologie. L'enseignement est en grande partie basé sur la pratique, à travers l'étude de données réelles publiées dans des articles scientifiques sur des questions de biologie correspondant aux cours de niveau L2. Ce poly vous présente un résumé du cours, insistant sur les notions que l'on vous demande d'être capable de maîtriser. Il est accompagné d'un poly de mathématiques, disponible sur eCampus, qui aborde aussi des questions allant au delà du contenu de cette ue, et qui vous servira dans la suite de vos études.

1 Probabilités, variables aléatoires

Variable aléatoire : *On appelle variable aléatoire toute variable dont la valeur dépend du résultat d'une expérience probabiliste. Une variable aléatoire est caractérisée par*

1. *les valeurs qu'elle peut prendre, que l'on appelle le* ***support*** *de la variable aléatoire.*
2. *la probabilité d'observer chaque valeur dans la population ou* ***loi de probabilité****.*

On peut distinguer plusieurs catégories de variables aléatoires

Variable qualitative : Caractéristiques non numériques. Peuvent être ***nominales*** (couleur des yeux,..) ou ***ordinales*** lorsque l'ensemble des catégories est muni d'un ordre total (peu confortable, assez confortable, très confortable). Les différents niveaux d'une variable qualitative s'appellent des modalités. Attention, les modalités peuvent être codées sous forme de valeurs numériques mais dans le cas d'une variable nominale, cela n'a pas de sens de faire une opération sur ces valeurs.

Variable quantitative : Caractéristiques numériques (taille, âge,...) qui résulte d'une mesure sur des individus. Elle s'exprime par des nombres sur lesquels les opérations arithmétiques de base (somme, moyenne,...) ont un sens. Une variable aléatoire quantitative peut être ***discrète*** (nombre de descendants d'un individu, nombre de soies thoraciques chez la drosophile) ou ***continue*** (poids, taille).

1.1 Variable aléatoire qualitative

Pour une variable aléatoire qualitative, il est possible d'énumérer toutes les valeurs possibles. Ces valeurs sont appelées des ***modalités*** de la variable aléatoire.
Soit X une variable aléatoire pouvant prendre les modalités $\{a_1, a_2, ..., a_J\}$. On peut calculer la probabilité pour que X prenne une valeur a_j particulière ($j = 1..J$). La loi de probabilité de X décrit $P(X = a_j)$ pour chaque modalité a_j.

1.2 Variable aléatoire quantitative discrète

Le support de ces variables aléatoires sont des valeurs discrètes. C'est le cas des variables de dénombrement, dans ce cas les valeurs sont entières. On peut énumérer ces valeurs et les appeller comme précédemment a_1, a_2, ..., a_J. On peut associer une probabilité $P(X = a_j)$ à chaque valeur a_j. Une réalisation x de X ne pourra prendre qu'une seule valeur parmis a_1, ..., a_J. Le support peut être infini, par exemple $\mathbb{N}$. Lorsque le nombre de modalités devient très grand, une variable de comptage peut être assimilée à une variable continue.

1.3 Variable aléatoire quantitative continue

Ils s'agit de variables à valeurs dans $\mathbb{R}$, ou, le plus souvent en biologie, un intervalle inclus dans $\mathbb{R}^+$ (mesures biométriques, concentrations, ...). La probabilité pour qu'une variable aléatoire X (exemple, le taux de glucose du sang) prenne très exactement la valeur x (0.3846mg/l) est nulle. Par contre, on peut calculer la probabilité $F(x)$ pour que X soit plus petit qu'une certaine valeur x,

$$F(x) = P(X \leq x)$$

que l'on appelle ***fonction de répartition*** de X. La fonction de répartition permet de calculer la probabilité que X se trouve dans un intervalle compris entre x et $x + dx$: $P(X \in]x, x + dx]) = F(x + dx) - F(x)$.
La ***fonction de densité*** notée $f(x)$ est la dérivée de F :

$$f(x) = lim_{dx \to 0} \frac{(F(x + dx) - F(x))}{dx}$$

La connaissance de la loi de probabilité d'une variable aléatoire permet de calculer des grandeurs telles que l'espérance ou la variance.

1.4 Couple de variables aléatoires

Lorsque l'on réalise des mesures différentes (pex taille et poids à la naissance) sur le même individu statistique, celles-ci peuvent dépendre l'une de l'autre. Par exemple, il est clair que pour des raisons biologiques (allométrie), on s'attend à une relation positive entre la taille et le poids à la naissance chez l'Homme. Ainsi, deux variables aléatoires mesurées chez les mêmes individus ne sont pas forcément indépendantes entre elles.

Couple de variables aléatoires : (X, Y) *est un couple de variables aléatoires lorsque deux mesures différentes X et Y sont réalisées sur le même individu.*

On dira qu'il y a indépendance entre les deux variables aléatoires X et Y lorsque, quelle que soit la valeur prise par X, la loi de Y ne change pas, et réciproquement. De cette définition découle, pour les variables aléatoires discrètes : $P(X = a \quad \& \quad Y = b) = P(X = a) \cdot P(Y = b)$ et pour les variables aléatoires continues $f(x, y) = f(x) \cdot f(y)$.
On peut aussi dire pour les variables aléatoires discrètes que la probabilité conditionnelle que X soit égale à a sachant que $Y = b$ est égale à la probabilité que X soit égale à a, ce qui donne la formule $P(X = a/Y = b) = P(X = a)$. Pour les variables aléatoires à densité la définition de l'indépendance est $f_{(X,Y)}(x, y) = f_X(x) \cdot f_Y(y)$ ce qui donne l'autre définition, plus intuitive : la densité conditionnelle de X au point x sachant que $Y = y$ est égale à la densité de X au point x ce qui donne la formule $f_{(X/Y=y)}(x) = f_X(x)$.

1.5 A retenir

Cette partie correspond au chapitre 1 du poly de mathématiques, et aux séances de travaux-dirigés 1 et 2. On vous demande de retenir les choses suivantes

- Connaître les lois de probabilités usuelles pour des variables aléatoires discrètes (loi Binomiale, loi de Poisson) ou continues (loi Normale, Student, Chi-deux) et leurs caractéristiques : les paramètres qui caractérisent ces lois, et, dans chaque cas, l'espérance et la variance de la variable aléatoire.
- Être capable de calculer la probabilité pour qu'une valeur se situe dans un certain intervalle en utilisant les tables de valeurs numériques et les axiomes des probabilités.
- Savoir reconnaître la nature d'une variable aléatoire (discrète, continue) et proposer une loi de distribution pour cette variable.

Vous devez maîtriser le calcul suivant :

$$P(a < X <= b) = F(b) - F(a)$$

Pour les variables quantitatives, peut aussi écrire

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$

qui n'est pas valable pour les variables qualitatives.

Le tableau ci-dessous résume les lois de probabilités usuelles selon la nature des variables aléatoires :

Mesure	support	effectif	loi	paramètres
qualitative	binaire (O/N)	$n = 1$	Bernoulli	$\mathcal{B}(p)$
	A/B/C/D		définie par	(p_A, p_B, p_C, p_D)
quantitative	comptage binaire	n petit	Binomiale	$B(n, p)$
		n grand, p petit	Poisson	$\mathcal{P}(np)$
		n grand, p ni trop petit ni trop grand	Normale	$\mathcal{N}(np, np(1-p))$
quantitative	nombre réel		Normale	$\mathcal{N}(m, \sigma^2)$

La démarche en statistique consiste à utiliser des mesures d'une variable aléatoire pour **estimer les paramètres** de la loi de distribution ou **tester des hypothèses** sur les ordres de grandeur ou les valeurs de ces paramètres.

2 Echantillon, n-échantillon

2.1 Individu, population, échantillon et estimation

Individu statistique : *unité de base sur laquelle la mesure est réalisée.*

Les individus peuvent être des personnes, des lunettes (exemple, on veut tester la qualité d'une monture), des bactéries. L'unité statistique peut aussi être un groupe d'individus (exemple, une classe de 25 élèves dont on étudie le comportement, ou toutes les plantes d'*Arabidopsis thaliana* contenues dans une terrine) ...

Population : *Ensemble des individus sur lesquels on souhaite dégager des informations*

Cette population peut être très grande. Par exemple, la population française des hommes adultes, ou l'ensemble des lunettes qui sortent d'une chaîne de montage sur une période donnée. Une population peut aussi être un ensemble plus restreint d'individus. **La loi de probabilité d'une variable aléatoire est toujours définie pour une population**, qu'il convient de décrire au début de chaque expérience. Quand la population est de très grande taille, il est impossible de faire des mesures sur l'ensemble des individus de la population. Par contre, on peut faire des mesures sur un petit nombre d'individus pris ou tirés au hasard dans la population, qui seront considérés comme représentatifs de la population.

n-échantillon : *sous-ensemble de n individus tirés au hasard et indépendamment dans la population de référence. On appelle X_i la variable aléatoire associée au tirage de l'individu i ($i = 1..n$), et x_i la valeur observée chez l'individu i. La façon de constituer l'échantillon (tirages aléatoires indépendants) permet de faire l'hypothèse que les variables aléatoires X_i sont indépendantes et de même loi.*

Les observations x_i peuvent servir à faire des hypothèses concernant la loi commune des X_i, c'est à dire la loi de X, ou bien à **estimer** les paramètres de la loi de X.

2.2 Description d'un n-échantillon

Lorsque la taille de l'échantillon devient grande, il est impossible de représenter chaque individu. Par contre, on peut décrire l'échantillon par des mesures qui le résument, ou en réaliser une représentation graphique.

2.2.1 Variables aléatoires discrètes (quantitative ou qualitative)

On peut résumer l'échantillon par un **tableau de contingence**, en comptant les effectifs observés pour chaque modalité de la variable aléatoire.

Exemple Le frêne présente un régime de reproduction mixte avec des arbres mâles, femelles et aussi des arbres hermaphrodites. Les modèles d'évolution du sexe supposent souvent l'existence d'un coût à la fonction mâle et prédisent un succès reproducteur plus important pour les femelles que pour les hermaphrodites. Dans l'étude suivante, on s'interesse aux arbres femelles et hermaphrodites. On a mesuré pour chaque arbre le sexe (femelle ou hermaprhodite) et la densité de fruits, répartie en quatre classes, de zéro (absence de fruits) à 3 (très nombreux fruits). La densité de fruit est traitée ici comme une variable qualitative ordinale. Les données peuvent être représentées de deux façons différentes, comme illustré dans le tableau 1.

Variables aléatoires associées au n-échantillon Le tableau précédent décrit ce qui a été obtenu avec un échantillon. En échantillonnant la même population une seconde fois, on aurait obtenu des valeurs différentes. On observe donc bien le résultat d'un tirage aléatoire. On peut définir plusieurs variables aléatoires qui décrivent des n-échantillons de la population de frênes.

- Population : les frênes femelles et hermaphrodites
- Echantillon : 102 frênes échantillonnés aléatoirement dans la population
- Individu statistique : un frêne de l'échantillon
- Le sexe de l'arbre i, X_i, est une variable de Bernoulli de loi $\mathcal{B}(p_F)$ qui prend la valeur F avec la probabilité p_F.

A.

arbre	sexe	densité de fruits
1	F	1
2	F	3
3	H	3
...	...	...
102	H	2

B.

	densité de fruits				
sexe	0	1	2	3	effectifs marginaux
F	7	8	7	8	30
H	10	24	20	18	72
eff. marginaux	17	32	27	26	102

Table 1: **Variables discrètes : données brutes et tableau de contingence. A.** Représentation du fichier de données récoltées sur le frêne. Chaque ligne correspond à un arbre. La première ligne donne les intitulés des colonnes. La première colonne est l'identifiant de l'arbre. Les colonnes suivantes contiennent les valeurs des variables mesurées. **B.** Tableau de contingence résumant le fichier de données sans perte d'information.

- la densité en fruits Y_i de l'arbre i est une variable qualitative dont les modalités sont $(0, 1, 2, 3)$. On peut associer à Y_i la loi de probabilité définie par les probabilités de chaque modalité : (q_0, q_1, q_2, q_3).
- (X_i, Y_i) $(i \in \{1, ..., 102\})$ est le couple de variables aléatoires qui définit le sexe (X_i) et la densité de fruit (Y_i) d'un arbre i tiré au hasard dans la population. La loi conjointe de X_i et Y_i est inconnue dans le cas général, mais peut-être calculée sous l'hypothèse d'indépendance entre X_i et Y_i :

$$P(X_i = k \cap Y_i = l) =_{indep} p_k \cdot q_l$$

- Z est le nombre d'arbres femelles dans un n-échantillon de $n = 102$ frênes. Z suit une loi binomiale $B(102, p_F)$.
- W_{kl} est le nombre d'individus de sexe k (H ou F) produisant une densité de fruits l (0, 1, 2 ou 3) dans un n-échantillon de $n = 102$ frênes. La loi de probabilité des W_{kl} peut être calculée à partir de la loi conjointe de X_i et Y_i.

Fréquences Dans le cas des frênes, si chaque arbre de l'espèce était connu, on pourrait calculer la probabilité p_F de trouver un arbre femelle comme le nombre total d'arbres femelles sur l'effectif total de l'espèce. La fréquence empirique d'arbres femelles dans l'échantillon donne une estimation de la probabilité p_F.

2.2.2 Variables aléatoires continues

ID	sexe	taille
1	F	164
2	F	170
3	F	165
...	...	...
172	F	167

Table 2: **Variables continues : données brutes.** Représentation du fichier de données récoltées sur les étudiantes de L2. Chaque ligne correspond à une étudiante. La première ligne donne les intitulés des colonnes. La première colonne est l'identifiant de l'étudiante. Les colonnes suivantes contiennent le sexe (F pour fille) et la taille en cm. La moyenne de l'échantillon vaut $\bar{x} = 165.49$ et la variance $s_n^2 = 37.52$.

Un n-échantillon d'une variable aléatoire continue peut se résumer de deux façons différentes :

- Il est possible de *résumer* les valeurs de la variable aléatoire dans l'échantillon par une moyenne et une variance.
- Les données peuvent être regroupées en classes discrètes, en comptant le nombre d'individus se trouvant dans un intervalle donné de valeurs, et représentées sous la forme d'un histogramme.

Les deux méthodes ne sont pas équivalentes. Lorsque l'on regroupe des données en classe, on perd une partie de l'information.

Exemple Une enquête biométrique relève la taille (en cm) d'un échantillon de 172 étudiantes de L2 Biologie à l'université Paris-Sud dans le but de définir s'il convient de changer le mobilier des salles de classe.

Variables aléatoires associées au n-échantillon Le tableau précédent décrit ce qui a été obtenu avec un échantillon. En échantillonnant la même population une seconde fois, on aurait obtenu des valeurs différentes. On observe donc bien le résultat d'un tirage aléatoire. On peut définir plusieurs variables aléatoires qui décrivent des n-échantillons de la population d'étudiantes.
Voici une liste des variables aléatoires usuellement définies dans un n-échantillon pour une mesure continue :

- La taille de la i-ème fille de l'échantillon, X_i est une variable aléatoire gaussienne de loi $\mathcal{N}(m, \sigma^2)$.
- La moyenne de la taille dans l'échantillon :

  $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  est aussi une variable aléatoire de loi $\mathcal{N}(m, \frac{\sigma^2}{n})$. On peut noter que la loi de $\bar{X}$ est différente de celle de X, la variance est divisée par n car on a beaucoup moins de chances d'obtenir des valeurs extrêmes en calculant une moyenne sur n individus qu'en réalisant une observation sur un individu.
- La variance empirique à moyenne inconnue

  $$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

  est une variable aléatoire d'espérance σ^2. On ne connaît pas sa loi, mais la loi de $\frac{(n-1)S_n^2}{\sigma^2}$, qui suit $\mathcal{X}^2_{n-1}$.

intervalle	$[150-155[$	$[155-160[$	$[160-165[$	$[165-170[$	$[170-175[$	$[175-180[$	$[180-190[$
effectif	13	27	50	53	27	7	1

Table 3: **Variables continues : données regroupées en classes.** Tableau des effectifs observés dans chaque classe, après avoirs regroupé les tailles en classes constituées d'intervalles de 5cm. Noter la dernière classe qui comprend un intervalle de 10cm. La moyenne de l'échantillon vaut $\bar{x} = 164.74$ et la variance $s_n^2 = 41.01$.

Moyenne et variance de l'échantillon La moyenne $\bar{x}$ de l'échantillon donne une estimation de la moyenne μ de la population.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La mesure qui permet d'estimer la variance de la population est la variance corrigée de l'échantillon

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A noter le terme correctif en $n-1$. Il permet de s'assurer que l'espérance de la variable aléatoire associée S_n^2 est égale à la variance de la population. En d'autres termes, s'il était possible de refaire l'échantillon un nombre de fois infini, la moyenne des valeurs de s_n^2 serait égale à la variance σ^2 de la variable aléatoire dans la population.
Lorsque les données sont représentées en classes de valeurs (Table 3) sous la forme (a_k, n_k), où a_k est la valeur du centre de la classe k et n_k l'effectif observé dans la classe, on utilise les formules suivantes pour calculer la moyenne et la variance de l'échantillon :

$$\bar{x} = \frac{1}{n}\sum_{k} n_k \cdot a_k$$

La moyenne est la somme des valeurs des centres de classe pondérée par les effectifs de la classe.

$$s_n^2 = \frac{1}{n-1}\sum_{k=1}^{K} n_k \cdot (a_k - \bar{x})^2$$

La variance est la somme du carré des écarts pondérés de centres de classe à la moyenne.
A noter que l'on perd toujours un peu d'information en regroupant les données en classes de valeurs, ce qui explique la différence entre les moyennes et variances calculées à partir des données brutes (Table 2), et les moyennes et variances calculées à partir des données regroupées en classes (Table 3).

2.3 A retenir

Cette partie correspond au chapitre 2 du poly de math. Il faut retenir les éléments suivants :

- La définition d'un n-échantillon.
- Savoir décrire la (les) variable(s) aléatoire(s) et leur loi de probabilité à partir de la description de l'échantillon. Faire la différence entre une variable discrète et une variable continue.
- L'échantillon sert à tester des hypothèses sur les paramètres des lois des variables aléatoires. Identifier dans un énnoncé le paramètre de la loi sur lequel porte la question biologique.
- Savoir représenter ou calculer des statistiques qui résument un échantillon.
- Faire la différence entre une variable aléatoire décrivant une mesure réalisable sur un échantillon (par exemple, effectif d'une classe ou moyenne de l'échantillon), et la valeur observée de cette variable aléatoire dans un échantillon particulier.

3 Tests d'hypothèses

Le principe des tests d'hypothèse est détaillé dans le chapitre 3 du poly de math. Nous résumons ici la démarche générale à suivre pour la réalisation de n'importe quel test statistique. Elle peut se décomposer en 7 étapes :

1. Poser un modèle : quelle(s) variables sont étudiées ? Quelles sont leur loi ? Sur quel(s) paramètres de la loi porte la question biologique ? Comment est défini l'individu statistique ?
2. Formulation des hypothèses H0/H1.
3. Choix d'une statistique de test et détermination de sa loi sous H0.
4. Choix du risque de première espèce α (appelé aussi niveau) et définition de la zone de rejet

5. Calcul du seuil
6. Calcul de la valeur observée de la statistique de test, comparaison avec le seuil et calcul de la pvalue
7. Conclusion statistique (rejet ou non rejet de H0) et biologique (réponse à la question posée - retour aux données)

Les étapes les plus difficiles sont les étapes (1) et (3).

3.1 Modèle

L'étape de modélisation consiste à déterminer les paramètres qui influencent la loi de distribution des observations. Nous allons voir dans ce cours un certain nombre de cas standard qui pourront vous servir dans la plupart des cas que vous rencontrerez. De façon plus générale, l'étape de modélisation consiste à savoir décrire comment les données expérimentales sont générées. Si vous êtes capable de simuler une expérience, vous serez capable de la modéliser.
Une bonne façon de procéder pour décrire les modèle est d'utiliser la démarche suivante :

- Définir la nature de la variable aléatoire considérée (qualitative/quantitative, discrète/continue)
- Proposer une loi de probabilité pour cette variable.
- Déterminer quels paramètres de cette loi de probabilité sont inconnus.

3.2 Hypothèses H0/H1

L'hypothèse H0 doit toujours être formulée de façon à pouvoir donner une valeur numérique au paramètre (inconnu) de la loi de probabilité de la variable aléatoire considérée. L'hypothèse H1 dépend de la question posée.

- Dans un test bilatéral, on teste l'hypothèse H0 que la valeur du paramètre est égale à une valeur connue, contre l'hypothèse H1 que la valeur du paramètre est différente de cette valeur.
- Dans un test unilatéral, on teste l'hypothèse H0 que la valeur du paramètre est inférieure ou égale (ou supérieure ou égale) à une valeur connue, contre l'hypothèse H1 que la valeur du paramètre est supérieure (inférieure) à cette valeur. Pour calculer la statistique de test sous H0, on se placera alors dans la situation la plus défavorable en supposant que le paramètre est égal à la valeur de la borne supérieure (inférieure) que l'on cherche à tester.

3.3 Choix d'une statistique de test

L'étape de choix d'une statistique de test est un domaine de recherche en soi. Il faut trouver **une variable aléatoire qui résume les données expérimentales et qui peut se calculer uniquement à partir de ces données, dont la loi sous H0 est connue**. Le choix d'une statistique de test dépend intimement du modèle. Ci dessous, les différents modèles qui sont abordés dans cette ue, avec leur statistique de test :

- **Tests de conformité :** La question est de savoir si l'un des paramètres de la variable aléatoire étudié est égal (conforme) à une valeur connue.
  - *v.a. quantitative discrète : test binomial.* La variable aléatoire est une variable de comptage X, qui suit une loi binomiale $\mathcal{B}(n, p)$. La taille de l'échantillon n est connue. Le test porte sur la probabilité de succès p. La statistique de test est le nombre de succès X observés dans l'échantillon. Sous l'hypothèse H0 ($p = p_{theor}$), la loi de X est $\mathcal{B}(n, p_{theor})$. Ce test sera vu au TD4.

- *v.a. qualitative discrète : test du chideux de conformité.* La variable aléatoire N_j, décrit le nombre d'individus d'un n-échantillon ayant la modalité a_j avec la probabilité q_j. Sous l'hypothèse H0 (l'échantillon est un échantillon conforme d'une certaine population, $q_j = q_{j_{theor}}$ pour toutes les modalités), l'effectif théorique de la classe j est $M_j = n \cdot q_{j_{theor}}$. La statistique

$$Q = \sum_j \frac{(N_j - M_j)^2}{M_j}$$

suit une loi du chideux $\mathcal{X}^2_{J-1}$ à $J-1$ degrés de libertés, avec J le nombre de modalités. Il faut vérifier que tous les effectifs théoriques sont supérieurs à cinq. Ce test est décrit au chapitre 4 du poly de math.

- *v.a. quantitative continue, variance connue : test gaussien.* La variable aléatoire X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, et la variance de la **population**, σ^2 est connue. Le test porte sur la moyenne de la population. La statistique de test utilise la moyenne de l'échantillon $\bar{X}$. Sous l'hypothèse H0 ($\mu = \mu_{theor}$), la loi de

$$Z = \sqrt{n}\left(\frac{\bar{X} - \mu_{theor}}{\sigma}\right)$$

est une loi normale $\mathcal{N}(0, 1)$. Ce test est décrit au chapitre 5 du poly de math.

- *v.a. quantitative continue, variance inconnue : test de Student.* La variable aléatoire X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, et la variance de la population σ^2 est inconnue. Le test porte sur la moyenne de la population. La statistique de test utilise la moyenne de l'échantillon $\bar{X}$ et la variance **estimée de la population à partir de l'échantillon**, S_n^2 (cette variance estimée peut vous être donnée dans l'énoncé ou vous pouvez avoir à la calculer à partir des données observées). Sous l'hypothèse H0 ($\mu = \mu_{theor}$), la loi de

$$T = \sqrt{n}\left(\frac{\bar{X} - \mu_{theor}}{S_n}\right)$$

est une loi de Student $\mathcal{T}(n-1)$ à $n-1$ degrés de liberté. Ce test est décrit au chapitre 5 du poly de math.

- **Test d'homogénéité :** Deux variables aléatoires indépendantes décrivent la même mesure dans deux échantillons différents. La question est de savoir si les deux variables aléatoires ont la même loi de probabilité, c'est à dire si les deux échantillons proviennent de la même population.

  - *v.a. qualitative discrète : test du chideux d'homogénéité.* Les deux variables aléatoires ont les mêmes modalités. On veut savoir si les probabilités de chaque modalité sont égales pour les deux variables aléatoires. Les effectifs théoriques se calculent, pour chaque modalité, en utilisant les proportions observées de chaque modalité après avoir réuni les deux échantillons. Si W_{ij} est la variable aléatoire qui décrit le nombre d'observations de la modalité j dans l'échantillon i, et M_j la variable aléatoire qui décrit l'effectif théorique de la modalité j sous l'hypothèse H0 (les probabilités sont identiques pour les deux variables aléatoires), la statistique

$$Q = \sum_i \sum_j \frac{(X_{ij} - M_j)^2}{M_j}$$

suit approximativement une loi du chideux $\mathcal{X}^2_{(J-1).(I-1)}$ à $(J-1).(I-1)$ degrés de libertés, avec J le nombre de modalités et I le nombre de populations étudiées. Il faut vérifier que tous les effectifs théoriques sont supérieurs à cinq. Ce test est décrit au chapitre 7 du poly de math.

- **Test d'indépendance :** Le couple de variables aléatoires (X, Y) décrit deux mesures différentes sur un même individu. La question est de savoir si X et Y sont indépendantes.

  - *v.a. qualitative discrète : test du chideux d'indépendance.* Sous l'hypothèse d'indépendance entre X et Y, la probabilité d'observer chez un même individu la modalité a_j de X (probabilité p_j) et la modalité b_k de Y (probabilité q_k) est égal au produit des probabilités $p_j \cdot q_k$. Les effectifs théoriques se calculent en utilisant, pour chaque couple de modalité (a_j, b_k) du couple (X, Y), le produit des proportions observées des modalités a_j pour la v.a. X et b_k pour la variable Y. Si W_{jk} est la variable aléatoire qui décrit le nombre d'observations de la modalité j de la variable X et la modalité k de la variable Y, et M_{ij} la variable aléatoire qui décrit l'effectif théorique associé sous l'hypothèse H0 (indépendance), la statistique

$$Q = \sum_j \sum_k \frac{(W_{jk} - M_{jk})^2}{M_{jk}}$$

    suit approximativement une loi du chideux $\mathcal{X}^2_{(J-1)\cdot(K-1)}$ à $(J-1)\cdot(K-1)$ degrés de libertés, avec J le nombre de modalités de X, et K le nombre de modalités de Y. Il faut vérifier que tous les effectifs théoriques sont supérieurs à cinq. Ce test est décrit au chapitre 7 du poly de math.

3.4 Choix du risque et définition de la zone de rejet

On décide de rejeter H0 lorsque la valeur observée de la statistique de test est peu probable sous l'hypothèse H0. La zone de rejet est définie comme un intervalle. Si la valeur observée de la statistique de test appartient à cet intervalle, on rejette l'hypothèse H0.
Le risque α est défini comme la probabilité que la statistique de test appartienne à la zone de rejet si H0 est vraie. Une fois le risque choisi, les bornes de l'intervalle se calculent en utilisant les quantiles de la loi sous H0. Dans le cas d'un test bilatéral, la zone de rejet est un intervalle disjoint.

3.5 Calcul du seuil

Après avoir déterminé la forme de la zone de rejet, on calcule le seuil grâce au niveau du test fixé précédemment et de la loi suivie par la statistique de test sous H0.

3.6 Calcul de la valeur observée de la statistique de test et de la pvalue

Les valeurs observées de la variable aléatoire dans le n-échantillons sont utilisées, ainsi que les paramètres de la loi de probabilité sous H0, pour calculer la valeur observée de la statistique de test.
Si la valeur observée de la statistique de test est comprise dans l'intervalle de rejet, on rejette H0.
La pvalue est définie comme la valeur minimale du risque qui permettrait de rejeter l'hypothèse H0 sachant la valeur observée de la statistique de test. Avec cette définition, lorsque la valeur observée de la statistique de test est dans la zone de rejet, la pvalue est inférieure au risque α. Inversement, lorsque la valeur observée de la statistique de test n'est pas dans la zone de rejet, la pvalue est supérieure au risque α. La règle de décision du test peut donc être énnoncée de la façon suivante :
Si la pvalue est inférieure au risque, on rejette H0.
A noter que la pvalue permet aussi de quantifier le risque pris en rejetant H0 à partir de l'échantillon observé. Cette notion sera abordée au TD4.

3.7 Décision et conclusion

Après avoir énoncé la décision prise, à savoir rejet ou acceptation de l'hypothèse H0, on peut revenir à la question biologique et y répondre en fonction du résultat du test.

3.8 A retenir

Les TD seront consacrés à la pratique de tests statistiques.

- Connaître les 7 étapes d'un test statistique
- Savoir formuler correctement des hypothèses H0 et H1 et dessiner la zone de rejet.
- Connaître les cinq tests statistiques vu en cours.
- Savoir prendre une décision en se basant sur l'intervalle de rejet, ou bien sur la pvalue.

4 Test Gaussien

Le déroulement complet d'un test gaussien est présenté ici à partir d'un exemple concernant le déterminisme du sexe chez les plantes.

4.1 Problème

Chez les plantes à fleurs, le régime de reproduction, qui décrit la façon dont se rencontrent les gamètes au moment de la reproduction sexué, est extrêmement variable, et dépend du déterminisme du sexe. Voici quelques exemples:
hermaphrodisme : à l'intérieur de chaque fleur on observe des organes reproducteurs mâles et femelles. L'autofécondation est possible. C'est le cas le plus répandu.
monoécie : fleurs unisexuées mâles (à étamines) et femelles (à pistil) portées par le même pied (ex. le maïs).
dioécie : les fleurs unisexuées mâles (à étamines) et femelles (à pistil) portées par des pieds différents (ex. le palmier-dattier, l'asperge).
gynodioecie : existence d'individus hermaphrodites et d'individus femelles, avec un déterministe génétique du sexe (ex. le thym).
Une hypothèse pour expliquer ces variations est le coût de fabrication des organes reproducteurs mâles et femelles. En général, le plantes hermaphrodites produisent moins de pollen que les plantes mâles, et moins de graines que les plantes femelles. Le gypsophylle est une plante de rocaille gynodioique. Les plantes femelles sont incapables de produire du pollen du fait d'une mutation dans un gène mitochondrial. Il s'agit donc d'un cas de déterminisme maternel du sexe. La descendance d'une plante hermaphrodite est constituée de plantes hermaphrodites. La descendance d'une plante femelle est constituée de plantes femelles uniquement. Le données présentées ici sont tirées de l'article de Lopez-Villavicencio et al (2005, Am. J. Bot 92(12):1995-2002).
Les plantes hermaphrodites produisent en moyenne 14 graines par fruit, avec une écart-type de 11 *graines*. On a mesuré la production de graines d'un échantillon de 20 fruits récoltées sur des plantes femelles. Les données sont fournies dans le tableau ci-dessous.

ech	nb graines	ech	nb graines	ech	nb graines	ech	nb graines
1	18	6	8	11	26	16	24
2	6	7	8	12	7	17	42
3	15	8	5	13	13	18	11
4	1	9	18	14	0	19	21
5	0	10	17	15	26	20	12

On aimerait savoir si les plantes femelles produisent plus de graines par fruit que les plantes hermaphrodites.

4.2 Réalisation du test

Question biologique : Le cadre général de l'étude est la variation du déterminisme du sexe chez les plantes à fleur. Il s'agit de tester l'hypothèse que les plantes femelles ont un avantage par rapport aux plantes hermaphrodites car elles produisent plus de graines. A noter que pour produire plus de graines au total, une plante peut produire plus de graines par fruit, ou plus de fleurs donnant des fruits. Le caractère observé ici est le nombre de graines par fruit.

1. **Modèle :** On considère que le gypsophylle est un mélange de deux populations, la population des plantes hermaphrodites, et la population des plantes femelles. On dispose d'un échantillon de $n = 20$ fruits récoltés sur des plantes femelles de gypsophylle. Soit X la variable aléatoire représentant le nombre de graines d'une fruit. On considère l'échantillon $(X_1, ..., X_i, ..., X_{20})$. On suppose que X suit une loi normale $\mathcal{N}(m, \sigma^2)$, où m est la moyenne du nombre de graines produites dans la population des gypsophylles femelles. On connaît par ailleurs la moyenne (14) et la variance ($11^2 = 121$) du nombre de graines dans la population des gypsophylles hermaphrodites. On fait l'hypothèse supplémentaire que $\sigma^2 = 121$ (les variances du nombre de graines sont les mêmes dans les deux populations).

2. **Hypothèses H_0 et H_1 :** On cherche à comparer les moyennes des deux populations de gypsophylle, avec un a priori qui serait $m \geq 14$. On pose $H_0 : m = 14$ et $H_1 : m \geq 14$. Sous l'hypothèse H_0, il n'y a pas de différence entre les plantes femelles et hermaphrodites et l'on dispose d'une valeur numérique pour les deux paramètres de la loi de X, $m = 14$ et $\sigma^2 = 121$.

3. **Statistique de test :** La moyenne de l'échantillon $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ suit une loi normale $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$. Sous l'hypothèse H_0, cette loi est connue. Il s'agit de la loi $\mathcal{N}\left(14, \frac{121}{20}\right)$. Il est souvent plus facile de travailler avec des variables centrées réduites.

   On propose donc $Z = \frac{\bar{X}-14}{\sqrt{\frac{121}{20}}}$ comme statistique de test. Sous H_0, Z suit la loi $\mathcal{N}(0, 1)$.

4. **Choix du risque de première espèce et région de rejet de H_0 :** Si H_0 est vraie, on s'attend à ce que Z prenne des valeurs proches de 0. Si H_1 est vraie, Z devrait prendre des valeurs positives. On rejettera H_0 si Z est trop grand (test unilatéral). L'intervalle de rejet est donc de la forme

   $$I = [a, +\infty[$$

   On choisit un seuil $\alpha = 0.05$.

5. **Calcul d'un seuil :** On cherche a tel que $P_{H_0}(Z \geq a) \leq 0.05$, ce qui revient à chercher a tel que $P_{H_0}(Z \leq a) \geq 0.95$. Dans les tables de la loi normale centrée réduite, on cherche la valeur de a telle que $F(a) \geq 0.95$. On trouve $F(1.65) = 0.9505$. On choisit donc $a = 1.65$.

6. **Valeur observée de la statistique de test et pvalue:** On utilise le tableau des données réparties en classes pour calculer la moyenne observée de l'échantillon $\bar{x}_{obs}$.

   $$\bar{x}_{obs} = \frac{1}{20}(18 + 8 + 26 + 24 + 6 + 8 + 7 + 42 + 15 + 5 + ... + 0 + 21 + 0 + 17 + 26 + 12) = 13.90$$

   On en déduit $z_{obs} = -0.04$. z_{obs} est plus petit que 1.65. On peut utiliser la table de la loi normale centrée réduite pour calculer $P_{H_0}(Z \geq z_{obs}) = 1 - F(-0.04) = 0.5159$. Il était donc assez probable d'observer cette moyenne pour l'échantillon sous l'hypothèse H_0.

7. **Conclusions :** On ne peut donc pas rejeter H_0. Dans cet échantillon, les plantes femelles ne produisent pas plus de graines par fruit que les plantes hermaphrodites. Pour mettre en évidence un avantage femelle sur la production de graines, il faudrait regarder aussi le nombre de fruits par plante.

5 Test du chideux d'indépendance

Le déroulement complet d'un test de chideux d'indépendance est présenté ici à partir d'un exemple tiré d'une étude sur l'efficacité d'un vaccin contre la grippe saisonnière chez l'homme.

5.1 Problème

Les résultats suivants ont été obtenus lors d'une étude visant à évaluer l'efficacité respective de deux vaccins contre la grippe saisonnière (Monto et al, 2009, The New England Journal of Medecine). Le vaccin **TIV** contient le virus atténué. La vaccin **LIV** contient le virus vivant atténué et peut être

administré par brumisation. Les données ont été recueillies sur des participants volontaires âgés en moyenne de 23 ans et fréquentant les campus universitaires de l'état du Michigan (USA). Il s'agit d'une expérience en double aveugle. Les participants ont été divisés en trois lots selon le vaccin administré : **TIV**, **LIV** ou **Placebo**. Le traitement a été administré en novembre 2008, et les patients ont été suivis jusqu'en avril 2009. En cas de syndrome grippal, un prélèvement était effectué pour confirmer qu'il s'agissait bien de la grippe.

Lot	**TIV**	**LIV**	**Placebo**	Total
grippe avérée	28	56	35	119
absence symptomes	785	758	290	1833
Total	813	814	325	1952

Peut-on dire que le nouveau vaccin est efficace ?

5.2 Réalisation du test

Question biologique : Il s'agit de tester l'efficacité d'un nouveau conditionnement du vaccin contre la grippe permettant une administration par brumisation. On se demande si, toutes choses égales par ailleurs, la probabilité d'attrapper la grippe dépend du type de traitement.

1. **Modèle :** On suppose que le choix du traitement pour un patient donné est fait au hasard. Soit $X \in \{presence, absence\}$ la variable aléatoire décrivant la présence ou non de symptomes de la grippe, et $Y \in \{TIV, LIV, Placebo\}$, la variable aléatoire décrivant le traitement reçu. Un patient donné est caractérisé par une valeur de X et une valeur de Y. Si au moins un des vaccins est efficace, alors la loi de probabilité de Y va dépendre du traitement, c'est à dire de la valeur de X. On dispose d'un échantillon de $n = 1952$ patients. Les valeurs de X et Y pour chaque patient sont synthétisées dans le tableau disjonctif. On appelle n_{ij} le nombre de patients pour la ligne i ($i \in \{1, 2\}$) et la colonne j ($j \in \{1, 2, 3\}$) du tableau.

2. **Hypothèses H_0 et H_1 :** On cherche à savoir si la loi de probabilité de Y dépend de X (au moins l'un des traitements est efficace) ou non (les deux variables aléatoires sont indépendantes). On pose H_0 : X et Y sont indépendantes et H_1 : X et Y ne sont pas indépendantes.

3. **Statistique de test :** Sous l'hypothèse H_0, on peut définir $p_i = P(Y = i)$ et $q_j = P(Y = j)$. La probabilité d'une observation du tableau disjonctif s'écrit alors $P(X = i \cap Y = j) = p_i \cdot q_j$. On peut estimer ces probabilités dans un échantillon par

$$\hat{p}_i = \frac{\sum_{j=1}^{3} n_{ij}}{n} \qquad ; \qquad \hat{q}_j = \frac{\sum_{i=1}^{2} n_{ij}}{n}$$

   De même, on peut calculer

$$m_{ij} = n \cdot \hat{p}_i \cdot \hat{q}_j = \frac{\left(\sum_{j=1}^{3} n_{ij}\right) \cdot \left(\sum_{i=1}^{2} n_{ij}\right)}{n}$$

   Sous l'hypothèse H_0, m_{ij} est l'effectif théorique pour la ligne i et la colonne j du tableau. La statistique de test proposée est

$$Z = \sum_{i=1}^{2} \sum_{j=1}^{3} \left(\frac{(n_{ij} - m_{ij})^2}{m_{ij}} \right)$$

   Lorsque l'hypothèse H_0 est vérifiée, Z suit approximativement une loi de chi-deux à $(2-1)(3-1) = 2$ degrés de liberté $\mathcal{X}^2(2)$.

   Remarque : Pour calculer le nombre de degrés de libertés, on peut aussi utiliser

   ddl = nombre de termes de la somme pour le calcul de Z - nombre de paramètres estimés pour le calcul des effectifs théoriques m_{ij}

Ici, pour calculer m_{ij}, il faut estimer $\hat{p}_1$, $\hat{q}_1$ et $\hat{q}_2$ et connaître n, soit 4 paramètres. Il y a 6 termes dans la somme (6 cases dans le tableau disjonctif).

4. **Région de rejet de H_0 et choix du risque de première espèce :** Si H_0 est vraie, on s'attend à ce que la différence entre n_{ij} et m_{ij} soit petite, c'est à dire que Z prenne des valeurs proches de 0. Si H_1 est vraie, Z devrait prendre des valeurs positives. On rejettera H_0 si Z est trop grand. L'intervalle de rejet est donc de la forme

$$I = [a, +\infty[$$

On choisit un seuil $\alpha = 0.05$.

5. **Calcul d'un seuil :** On cherche a tel que $P_{H_0}(Z \geq a) \leq 0.05$, ce qui revient à chercher a tel que $P_{H_0}(Z \leq a) \geq 0.95$. Dans les tables de la loi du $\mathcal{X}^2(2)$, on cherche la valeur de a telle que $F(a) \geq 0.95$. On trouve $F(5.9915) = 0.95$. On choisit donc $a = 5.9915$.

6. **Valeur observée de la statistique de test, pvalue et décision** On utilise le tableau disjonctif pour calculer les m_{ij}.

Lot	**TIV**	**LIV**	**Placebo**	Total
grippe avérée	49	50	20	119
absence symptomes	764	764	305	1833
Total	813	814	325	1952

On trouve $z_{obs} = 21.91$. z_{obs} est plus grand que 5.9915. $P_{H_0}(Z \geq z_{obs}) = 1 - F(21.91)$.
Cette valeur n'est pas tabulée dans la table du $\mathcal{X}^2(2)$. On peut dire que la probabilité critique est plus petite que 0.001. On peut donc rejeter H_0.

7. **Conclusion biologique :** Les patients n'ont pas la même probabilité d'attraper la grippe selon le traitement administré. Si l'on compare les effectifs observés et théoriques (sous l'hypothèse H0), on constate qu'il y a moins de patients atteints de la grippe qu'attendu sous H0 dans le cas du traitement TIV (28/49), alors qu'il y a plus de patients atteints qu'attendu sous H0 pour les deux autres traitements, en particulier pour le placébo (35/20). On peut donc conclure que le nouveau vaccin est efficace.

6 Séances de TD de bio-statistiques

Les exercices qui seront abordés lors de ces séances de TD sont des exemples tirés d'études dans différents domaines de la biologie. L'objectif est de vous familiariser avec l'étape de modélisation qui consiste exprimer une question de biologie en utilisant un formalisme mathématique permettant de déboucher sur la réalisation d'un test statistique.

Les TD2, TD3 et TD4 seront consacrés à des exercices de biologie nécessitant l'application de tests vus en TD de mathématiques. A la suite du TD3, vous trouverez des exercices d'application à réaliser chez vous pour vous entrainer.
Le TD5 sera composé de deux parties en demi-groupes : une partie avec encadrement et une partie en autonomie (avec le passage ponctuel d'un enseignant). Cette séance vous permettra de réviser l'ensemble des tests traités dans la partie "biostat" de cette UE et de vous entraîner à identifier le test à appliquer pour répondre à une question biologique. A l'issue du TD5, vous aurez un DST de biomaths de 30min dont le but sera de vous évaluer sur votre capacité à identifier et à réaliser le test à appliquer pour un problème biologique donné.

Ne négligez pas les exercices abordés dans ces TD, à l'examen final sur les 3h d'examen, il y aura des exercice de maths et de biomaths. Le contrôle continu WIMS portera sur la partie maths ET la partie biomaths.

Pour rappel, voici la liste des OAV correspondant à cette UE :

- Savoir identifier la nature d'une variable aléatoire (quantitative, qualitative discrète ou continue) et sa loi de distribution.
- Savoir identifier la population étudiée et l'échantillon qui y est associé à partir d'un énoncé biologique
- Savoir estimer la moyenne et la variance d'une population à partir d'un n-échantillon
- Savoir identifier et écrire sous forme mathématique les variables et les paramètres sur lesquels porte une question biologique à partir d'un énoncé.
- Connaître les 7 étapes d'un test statistique (modèle, hypothèses, choix d'une statistique de test, détermination de la zone de rejet, calcul de la statistique de test, décision et pvalue, conclusion)
- Connaître les tests statistiques vu en cours (tests de conformité gaussien, binomial, de Student et du Chi2, test du chi2 d'indépendance)

TD2:BIO1. Lois de probabilités : Binomiale, Poisson
La dérive génétique et l'évolution des espèces vivantes
Durée : 1h15

Nous allons étudier au cours de cette séance des populations de bactéries cultivées en laboratoire, pour comprendre un processus majeur de l'évolution, l'évolution des espèces. On considèrera une situation très simplifiée et les impacts de deux forces évolutives : la dérive génétique et la mutation.

Exercice 1. La dérive génétique
On s'intéresse à un gène particulier noté A. Ce gène A est responsable de la transformation de l'indole en tryptophane. Dans les populations bactériennes étudiées, on observe trois génotypes $a1$, $a2$ et $a3$ (les bactéries ne possèdent qu'un seul allèle au locus étudié). On considère qu'il n'y a pas de mutation chez ces bactéries et qu'elles se divisent à la même vitesse. Une population correspond à une culture bactérienne en phase stationnaire contenue dans un Erlen de 100 ml de milieu minimum liquide. Dans ce milieu, la croissance s'arrête lorsque les bactéries de la culture ont atteint une densité de 10^8 cellules/ml (phase stationnaire).

1. Dans cet Erlen, on constate que la fréquence des bactéries $a1$ est 0,40 et celle des $a2$ 0,10. Quelle est la fréquence des bactéries $a3$ dans cette population ?

2. On repique un échantillon de $N = 10$ bactéries de cette population dans un nouvel Erlen de 100 ml de milieu minimum liquide. La croissance de la nouvelle population s'effectue jusqu'à la phase stationnaire. On s'intéresse à la fréquence des bactéries $a1$ dans la population issue du repiquage.

   Pour simuler cet échantillonnage, vous allez tirer au hasard des bactéries dans une *population infinie* (celle des bactéries présentes dans l'Erlen en phase stationnaire avant repiquage) contenant 40 % de bactéries $a1$. On utilisera une boite de grains de maïs dans laquelle les graines rouges sont en fréquence 0.4. Chaque étudiant effectue 3 tirages de taille $N = 10$. On peut également utiliser la table des nombres au hasard fournie à la fin du poly.

   Soit X la variable aléatoire décrivant le nombre de bactérie a1 dans un échantillon de taille 10.

   En réunissant les valeurs trouvées par l'ensemble des étudiants, construisez la distribution expérimentale de X.

   Donnez la probabilité estimée, d'après votre simulation expérimentale, pour $X = 4$.

   Quelle loi de probabilité théorique attendez-vous pour la variable X ? Quelles sont les moyennes et les variances de la distribution théorique de cette loi ? Donnez $P(X = 4)$ d'après la loi théorique.

3. On s'intéresse maintenant à la variable aléatoire P qui représente la proportion de bactéries $a1$ dans un échantillon de taille 10. Ecrivez P en fonction de la variable X. Définir la loi de probabilité de P. Quelles sont la moyenne et la variance de la distribution de cette loi ? Quelle est la probabilité d'obtenir une proportion de 40% de bactéries $a1$ dans l'échantillon repiqué ?

4. A votre avis, pour quelle raison s'est-on intéressé au calcul de la probabilité d'avoir une fréquence de 0,40 bactéries $a1$ dans l'échantillon repiqué ? Pensez-vous qu'une population issue d'un seul repiquage va être exactement de même composition génétique que la population dont elle est issue ?

5. On va simuler l'évolution d'une population de bactéries au cours du temps en utilisant la table des nombres aléatoires. On part d'une population où $f(a1) = 0.4$. Chaque étudiant effectue des simulations au cours du temps (10 générations, soit 10 repiquages successifs) de l'évolution de 2 populations de taille $N = 10$ et $N = 2$.

   Que va devenir l'allèle $a1$ dans cette population ? Est-ce que la taille de la population a une influence sur son devenir ?

6. Le phénomène étudié dans cet exercice s'appelle la dérive génétique. Il est impliqué dans l'évolution des espèces. Expliquez pourquoi.

Exercice 2. La mutation

On sait actuellement qu'il existe en moyenne 4000 gènes dans un génome bactérien. Considérons une culture contenant au départ 10^3 bactéries. Pour simplifier on va supposer que toutes les bactéries se divisent en même temps et que le milieu n'est pas limitant. On s'intéresse à la variable aléatoire X=nombre de gènes mutés dans cette culture pendant l'étape de division.

1. En supposant que le taux de mutation par gène et par division est de 10^{-6}, donnez la loi de X à la l'issue de la première étape de division. Donnez sa moyenne et sa variance.

2. Dites par quelle loi connue la loi que vous avez trouvée peut être approchée. Lire dans la table correspondante la probabilité qu'il y ait eu au moins une mutation dans la culture lors de cette division.

TD3:BIO2. Principe d'un test
Etudes sur les écureuils de la forêt de Sénart
Durée : 1h15

Exercice 3. Présence de tiques chez les écureuils
Un chercheur soupçonne que les tiques soient davantage présentes chez les écureuils mâles que chez les femelles car ceux-ci sont connus pour être davantage en mouvement. Il aimerait réaliser l'expérience suivante : capturer 20 écureuils porteurs de tiques et compter le nombre de mâles et de femelles. Pour essayer de répondre à sa question, il doit mettre en place un test statistique. On admettra que le sex-ratio des écureuils est de 0.5.

1. Quelle est la variable qui doit être étudiée ?
2. Quelles hypothèses H_0 et H_1 doit-on tester et pourquoi ?
3. Pour tester l'hypothèse H_0, on s'intéresse à la loi de probabilité de la variable dans des échantillons de 20 écureuils porteurs d'une population comportant autant de mâles que de femelles. Utilisez les tables pour construire le diagramme en bâtons donnant la proportion d'échantillons en fonction du nombre de mâles dans l'échantillon sous l'hypothèse H_0.
4. Construisez un test statistique qui permettra de répondre à la question posée.
5. Il réalise ensuite son expérience, il a capturé 20 écureuils porteurs de tiques dans la forêt : 6 sont des femelles et 14 des mâles.
6. Quel est le degré de signification α_0 de l'événement étudié (la probabilité de l'évènement sous l'hypothèse $H0$ testée)? Qu'en concluez-vous ?
7. Comment répondre à la question : y a-t-il autant de femelles que de mâles chez les écureuils porteurs de tiques en France ?

Après la séance TD3:BIO2 : pour vous entraîner

Exercice 4. Sex ratio des étudiants de Biologie
On voudrait savoir si le sex-ratio (proportion de femmes) de la population des étudiants de Biologie est égal au sex-ratio à la naissance dans la population humaine (50% de femmes). On utilise pour cela l'échantillon constitué par votre groupe de TD. On notera X la variable aléatoire qui décrit le nombre d'hommes dans l'échantillon.

1. Sur quel intervalle est définie la variable X ?
2. Quelle loi suit la variable X ? Quels sont les paramètres de cette loi ? Sont-ils connus ici ?
3. Pour chacune des valeurs k que peut prendre X, calculez $P(X = k)$ dans le cas où le sex ratio des étudiants de biologie est égal à celui de la population humaine.
4. Calculez l'espérance et la variance estimée de X dans le cas où le sex ratio des étudiants de biologie est de 0.5.

Exercice 5. Biométrie : Variations de la taille chez l'homme
La taille d'un individu (hauteur) est une donnée anthropométrique ayant une variance relativement faible (comparée au poids par exemple). Elle est déterminée par le sexe, les gènes et l'environnement des individus.
Au XXème siècle, on attribue l'augmentation de taille des Hommes à des facteurs environnementaux tels que l'hygiène, la alimentation, les progrès de la médecine. Cependant, la taille d l'Homme n'a pas toujours augmenté depuis son apparition sur Terre, au Néolithique (début de l'agriculture jusqu'à l'apparition de l'écriture) la taille des Homme a diminué en raison d'une diminution de l'apport protéique (moins d'animaux issus de la chasse, davantage d'animaux d'élevage, plus gras et moins riches en protéines).
Les femmes atteignent leur taille adulte vers 15 ans et les hommes vers 20 ans. Même si les hommes sont en moyenne plus grands que les femmes, à la puberté, les femmes sont donc "temporairement" plus grandes.

On s'intéresse à la taille (en cm) des étudiantes de l'université Paris-Sud que l'on modélisera par la variable aléatoire X. On considère pour cela le n-échantillon constitué des filles du groupe de TD.

1. Réalisez un sondage au sein de votre groupe et complétez le tableau ci-dessous à partir des effectifs observés dans votre groupe de TD.

intervalle	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$
effectif				

2. Quelle est la population étudiée ici ? Décrivez le n-échantillon tiré dans cette population.
3. Donnez la loi de la variable aléatoire associée à la mesure de la taille d'une étudiante.
4. A partir de votre échantillon (effectifs renseignés dans le tableau), estimez la moyenne et la variance de la population d'étudiantes.
5. Des études antérieures ont montré que l'espérance de la taille des étudiantes était $\mu = 165.49$ cm, et la variance $\sigma^2 = 37.42$. En quelle unité est exprimée cette variance ? En utilisant les classes définies plus haut et les tables de valeurs numériques, calculez la probabilité de chaque classe, sous l'hypothèse que les chiffres donnés par l'étude antérieure décrivent bien la population actuelle d'étudiante.

6. Comparer graphiquement les probabilités théoriques aux fréquences observées pour chaque classe dans l'échantillon. Discutez ensemble des différences observées.

Exercice 6. Génétique : rôle des mutations dans la diversité du vivant [A faire à la maison]
Pendant longtemps, on a supposé que les mutations, du fait de leur faible fréquence, ne pouvaient jouer un rôle dans l'évolution d'une population que sur des échelles de temps longues. Cet exercice a pour but de quantifier l'importance de la mutation comme pression évolutive.
On sait actuellement qu'il existe en moyenne 4000 gènes dans un génome bactérien. Considérons une culture contenant au départ 10^3 bactéries. Pour simplifier on va supposer que toutes les bactéries se divisent en même temps et que le milieu n'est pas limitant. On s'intéresse à la variable aléatoire X=nombre de gènes mutés dans cette culture pendant l'étape de division.

1. En supposant que le taux de mutation par gène et par division est de 10^{-6}, donnez la loi de X à la l'issue de la première étape de division. Donnez sa moyenne et sa variance.

2. Dites par quelle loi connue la loi que vous avez trouvée peut être approchée. Lire dans la table correspondante la probabilité qu'il y ait eu au moins une mutation dans la culture lors de cette division.

TD4:BIO3 - Plan d'expérience et tests statistiques sur des variables quantitatives continues et des variables qualitatives
Durée : 1h30

Exercice 7. Choix du partenaire chez les guppy

A

B

Figure 1: **Dimorphisme sexuel chez les guppy A femelles. B mâles**.

Un chercheur s'intéresse aux caractères sélectionnés par les femelles de guppy (*Poecilia reticulata*, petits poissons d'eau douce) pour choisir leur partenaire mâle. La longueur de la nageoire caudale, la taille totale et la couleur sont des caractères qui pourrait influencer ce choix.

1. Pour tester l'effet de la longueur de la nageoire caudale du mâle dans le choix des femelles, le chercheur réalise l'expérience suivante. Dans un aquarium, deux mâles différents sont proposés à une femelle, et le choix de la femelle est enregistré. L'expérience est répétée 60 fois. Il s'intéresse ensuite à l'échantillon des 60 mâles choisis et mesure leurs nageoires caudales. Le chercheur sait par ailleurs que la longueur des nageoires caudales chez les guppy mâles suit une loi Normale de moyenne 12,57 mm et de variance inconnue.

   Quelles hypothèses cherche-t-il à tester ?
   Quel test statistique faut-il conduire pour tester ces hypothèses ?
   Quelle loi suit la statistique de test sous H0 ?

2. La longueur moyenne des nageoires caudales dans l'échantillon est de 12,64 mm et la variance estimée est de 0,098 mm^2. Que pouvez vous conclure de cette expérience ?

3. L'étudiante en stage de L2 dans le laboratoire propose une modification du protocole expérimental. Plutôt que de mesurer la longueur de la nageoire des mâles choisis, il s'agirait de mesurer, pour chaque paire, la différence de longueur entre la nageoire du mâle choisi et celle du du mâle rejeté par la femelle. Proposez un modèle, des hypothèses H_0 et H_1, et une statistique de test correspondant à ce protocole.

4. A votre avis, lequel de ces deux protocoles faut-il suivre ?

Exercice 8. Choix du partenaire chez les guppy : taille et coloration du mâle
Les chercheurs s'intéressent à deux caractères phénotypiques : la taille et la couleur des mâles. Ils aimeraient tester si ces deux traits sont liés. Ils prévoient de pêcher plusieurs (n) individus mâles dans une population sauvage du Vénézuela. Les mâles seront mesurés et leur catégorie de taille sera notée : $<$ 2 cm, entre 2 et 4 cm et $>$ 4 cm. Leur intensité de coloration sera également notée : peu ou très colorés. Ils seront ensuite relâchés.

1. Décrivez la population qu'ils cherchent à étudier, l'échantillon qui sera observé et l'individu statistique.

2. Décrivez les variables qui seront observées et leurs types.

3. Les observations sont les suivantes :

Taille	Très colorés	Peu colorés
0 à 2 cm	325	62
2 à 4 cm	212	24
$>$ 4 cm	42	4

   Comment pourront-il analyser s'il existe une relation entre la taille et la couleur des nageoires ?

4. Proposez une représentation graphique de ces données.

TD5:BIO4 - Choix d'un test statistique
Partie en autonomie
Exercice de génétique

Exercice 9. Etude des relations entre caractères des grains de maïs
On observe des plants de maïs présentant des compositions en types de grains différentes, on échantillonne des plantes dans le champ et on observe les caractéristiques des graines récoltées. On compte 156 grains rouges lisses, 55 grains rouges ridés, 58 grains blancs lisses et 15 grains blancs ridés. On se demande si les deux phénotypes observés sur les grains (couleur et aspect) sont indépendants. C'est à dire, on se demande si un grain à l'aspect ridé à la même probabilité d'être rouge ou blanc qu'un grain à l'aspect lisse, ou bien si un grain rouge à la même probabilité d'être lisse ou ridé qu'un grain à blanc.

1. Construisez les données sous la forme d'un tableau de contingence résumant les effectifs observés.
2. Identifiez la population, l'échantillon et les deux variables aléatoires mesurées par l'observateur. Quelle est la nature de ces variables ?
3. Formulez les hypothèses biologiques H0 et H1 permettant de répondre à la question posée.
4. Dans un premier temps, on s'intéresse à chaque caractère séparément. Calculez la proportion de grains rouges dans l'échantillon et déduisez-en la proportion de grains blancs. Faites de même pour le caractère "aspect du grain".
5. Sous l'hypothèse d'indépendance entre les deux caractères, proposez une estimation de la proportion attendue de chaque type de grain (rouge-lisse, blanc-lisse, rouge-ridé, blanc-ridé) ?
6. Identifiez et réalisez le test statistique permettant de répondre à la question.

Exercice 10. Génétique des grains de maïs
En discutant avec les personnes en charge de l'expérience, on apprend que ces grains constituent une population F2 : des croisements entre des plants de maïs appartenant à deux lignées pures différentes (nommées lignée I et lignée II) avaient permis de produire une F1 l'année précédente, et un croisement F2 (F1xF1) a été réalisée cette année. Ils se souviennent que les lignées I et II produisaient l'une des grains rouges lisses et l'autre des grains blancs ridés. Par contre, en F1, tous les épis étaient à grains rouges lisses.

1. Dans un premier temps, on s'intéresse à la couleur des grains. A partir de ces nouvelles informations, proposez une hypothèse génétique simple mettant en jeu deux allèles permettant d'expliquer les phénotypes observés en F1 et en F2.
2. Selon votre hypothèse, on peut maintenant calculer les effectifs théoriques attendus en F1 et en F2. Pour le caractère "couleur", faites un tableau montrant les genotypes attendus en F1 et un tableau montrant les genotypes attendus en F2.
3. Déduisez en les proportions théoriques de grains blancs et de grains rouges en F1 et en F2 sous cette hypothèse.
4. Comparez les proportions théoriques aux proportions observées calculées dans l'exercice précédent.
5. **Indépendance génétique des 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. Si ils sont effectivement indépendants, calculez les proportions théoriques pour chacun des quatre types de grains en F2 en mobilisant vous cours de génétique. Vous regrouperez les proportions théoriques dans le tableau suivant.

Phénotypes des grains	Proportions attendues
rouge-lisse	
rouge-ridé	
blanc-lisse	
blanc-ridé	

6. Sachant que les proportions théoriques de chaque type de grain sont connues sous l'hypothèse d'indépendance entre les deux loci, quel test statistique proposez vous pour tester l'indépendance génétique entre les deux loci ?

7. Réalisez le test permettant de déterminer si les deux loci en jeu présentent une liaison génétique.

TD5:BIO4 - Choix du test statistique
Etudes des interactions entre espèces associées aux arbres *Macaranga*

A partir des articles
- Fiala et al. (1989) Studies of a South East Asian Ant-Plant Association: Protection of Macaranga Trees by Crematogaster borneensis. Oecologia 79(4):463-470. doi: 10.1007/BF0037866
- Heil et al. (1997) Food Body Production in Macaranga Triloba (Euphorbiaceae): A Plant Investment in Anti-Herbivore Defence via Symbiotic Ant Partners. Journal of Ecology 85:847-861. doi: 10.2307/2960606

L'objectif de cette séance est de vous faire travailler, par petits groupes à l'identification d'une question biologique et du test à mettre en œuvre pour y répondre (Tests 1 à 4). Vous vous attacherez donc plus particulièrement à énoncer la question biologique, le modèle, H0/H1. Chaque groupe travaillera sur une question différente et un bilan permettra de résumer les résultats de l'ensemble du groupe de TD. Les tests 5 à 8 peuvent être traités en TD également si vous avez le temps mais leur correction vous sera fournie pour vous permettre de réviser l'examen.
La compréhension de la dynamique des écosystèmes nécessite une bonne connaissance des interactions entre espèces, ou symbioses. Elles peuvent être de différents ordres allant des interactions bénéfiques pour les deux espèces à des interactions bénéfiques pour l'une d'elles et négatives pour l'autre.

- Mutualisme obligatoire : avantage pour les deux espèces et obligatoire (*E. coli et l'Homme : digestion, défense)*
- Mutualisme : avantage pour les deux espèces (*plantes / insectes lors de la pollinisation*)
- Commensalisme : avantage pour une espèce et sans effet sur la seconde(*les plantes épiphytes comme les lianes qui se développement sur d'autres plantes sans que ce soit au détriment des plantes support*)
- Neutralisme : deux espèces cohabitent sur un même territoire sans exercer d'influence entre elles (*l'escargot et le cheval dans un pré*)
- Amensalisme : neutre pour une espèce et délétère pour l'autre (*piétinement de végétaux par un mammifère*)
- Compétition pour une ressource limitée (nourriture, espace...) (*deux espèces bactériennes dans un milieu de culture*)
- Parasitisme : avantage pour une espèce et délétère pour l'autre (*Plasmodium falciparum et l'Homme (paludisme)*)
- Prédation : une espèce se nourrit de l'autre (*lynx et lapins*)

Les arbres du genre *Macaranga tribola* sont originaires des régions tropicales humides d'Asie du Sud Est. En Malaisie occidentale, ils vivent en étroite association avec les fourmis de l'espèce *Crematogaster borneensis*. Les bénéfices de cette association pour les communautés de fourmis sont relativement bien connus : celles-ci se nourrissent de protéines et lipides produits par l'arbre sous forme de corps nourriciers (Heil et al. 1997, *Journal of Ecology*)et se nichent dans les entrenœuds creux des tiges. Les effets de cette association sur les arbres *Macaranga tribola* sont plus difficiles à étudier. On sait que la production de corps nourriciers est coûteuse pour l'arbre, ils constituent en moyenne 5.4% de la biomasse aérienne sèche de la plante et que leur production représente 8.6% des dépenses énergétiques de la plante (Heil et al. 1997, *Journal of Ecology*). Cependant, de précédentes études tendent aussi à démontrer un effet positif également pour l'arbre hôte. Différentes hypothèses pourraient l'expliquer :

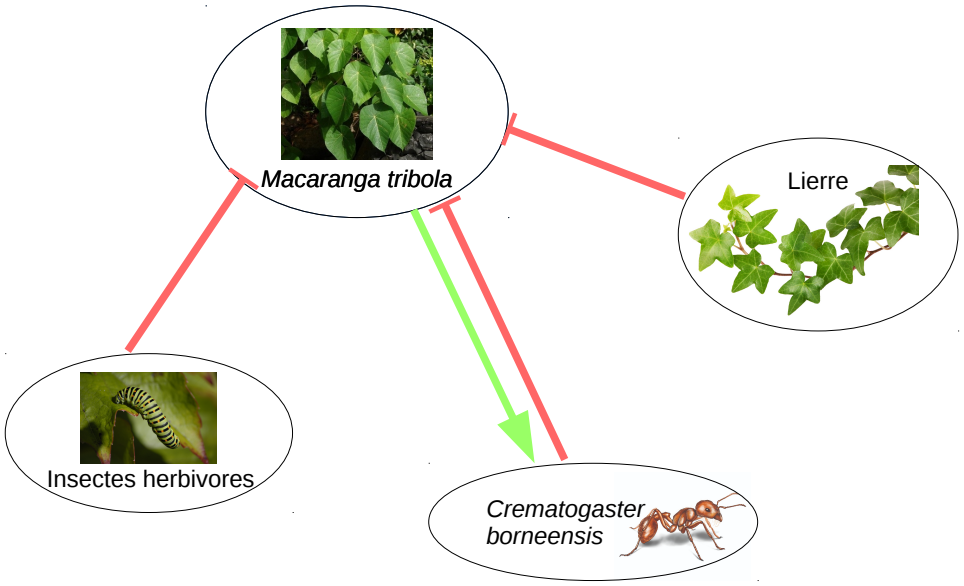


Figure 2: **Schéma représentant les interactions connues entre les 4 espèces étudiées.** Les flèches vertes représentent un effet positif de l'espèce de départ vers l'espèce cible et les flèches rouges représentent un effet négatif de l'espèce source vers l'espèce cible.

- effet bénéfique direct des fourmis sur la croissance
- effet protecteur contre une ou plusieurs espèces parasites de l'arbre

Pour mieux comprendre cet effet, les chercheurs ont cherché à mieux caractériser les interactions en les organismes associés à l'espèce *Macaranga tribola* : les fourmis, le lierre et les insectes herbivores.

Au début de leur étude, ils ont déjà connaissance des bénéfices de l'association avec *Macaranga tribola* pour les fourmis et également de l'effet négatif du lierre et des insectes herbivores sur la croissance de *Macarange tribola* (compétition dominante du lierre et dégâts foliaires par les insectes) - cf schéma ci-dessous.
Les chercheurs ont mesuré différentes variables sur les différents organismes. Les résultats de leurs mesures sont détaillés ci-dessous.

Test 1 : Effet de la présence de fourmis sur *Macaranga tribola*

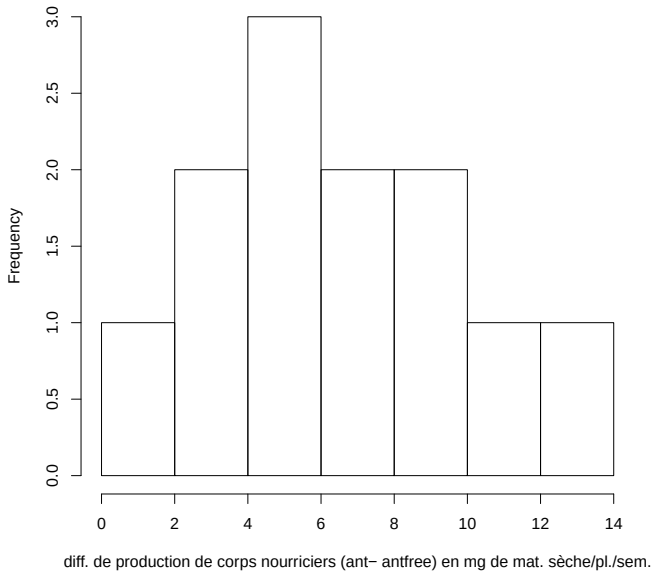
Les chercheurs ont réalisé des mesures sur 31 sites sur lesquels ils ont suivi deux arbres *Macaranga tribola*: un arbre non-colonisé et un arbre colonisé par les fourmis. Ils ont ensuite calculé, pour chaque arbre, la variation du nombre de dégats causés (en pourcentage de surface foliaire consommée) en 6 semaines.
Les statistiques résumées de ces mesures sont présentées dans le tableau ci-dessous.
Une nouvelle variable a été calculée : la différence de pourcentages foliaires consommés en 6 semaines entre l'arbre colonisé et l'arbre non colonisé d'un même site.
Ces résultats sont présentés sur la dernière ligne du tableau ci-dessous.

fourmis	nombre d'observations	% surface foliaire consommée
avec	31	+5.1 % ($s_{(n-1)} = 5.2$)
sans	31	+16.4% ($s_{(n-1)} = 11.8$)
couples "avec-sans"	31	-11.3 % ($s_{(n-1)} = 10$))

Déterminer et appliquer le test permettant de savoir si la consommation de surface foliaire est liée à la présence de fourmis sur l'arbre ?

Test 2 : Effet de la présence de fourmis sur *Macaranga tribola*

Dans une étude indépendante, des chercheurs ont étudié l'effet de la présence fourmis sur une plante sur sa production de corps nourriciers. Pour ce faire, ils ont constitué 12 couples de plantes *Macaranga tribola* de même taille et en ont placé une en présence de fourmis et l'autre en absence de fourmis pendant une semaine. A l'issue de cette semaine, ils ont mesuré différence de masse sèche de corps nourriciers entre la plante occupée par les fourmis et celle dépourvue de fourmis. La distribution de cette variable (moyenne observée = 6.775 mg et variance observée = 14.88 mg^2) est représentée ci-dessous. On peut considérer que cette variable a une distribution gaussienne.



On peut tenter de répondre à la question suivante : la production de corps nourriciers et la présence de fourmis sur l'arbre sont-elles liées ?

Test 3 : Attaque d'insectes herbivores par les fourmis

Des chenilles ont été placées sur des plantes hôtes en présence et en absence de fourmis. La présence de la chenille 1h plus tard a été observée.

fourmis	placées	présentes après 1h	disparues
oui	36	7	29
non	20	10	10

Test 4 : Lierre et fourmis

208 *Macaranga tribola* colonisés par des fourmis ont été observés et la présence ou l'absence de lierre chaque plante a été notée. Les résultats sont présentés dans le tableau ci-dessous.

fourmis	avec lierre	sans lierre
oui	10	198

On sait, de précédentes études qu'en absence de fourmis, 20% des arbres sont colonisés par du lierre.

Les tests 5 à 8 suivants peuvent être traités après la séance, pour vous entrainer.
Leur correction sera déposée sur ecampus.

Test 5 : Effet de la présence de fourmis sur *Macaranga tribola*

Les chercheurs ont réalisé des mesures sur 29 arbres *Macaranga tribola* colonisés par les fourmis. Ils ont calculé, pour chaque arbre, la variation de hauteur de l'arbre observée en 6 semaines, elle est estimée à 11 cm. Ils savent par ailleurs qu'un arbre non colonisé par les fourmis grandit en hauteur en moyenne de 9.7 cm en 6 semaines, et que cette variable suit une loi normale de variance et égale à $\sigma^2 = 5.1\ cm^2$ Identifiez et appliquez le test permettant de déterminer si la pousse de l'arbre diffère en présence de fourmis sur l'arbre.

Test 6 : Attaque d'insectes herbivores par les fourmis

Des œufs d'insectes ont été placés sur différents organes des plantes hôtes et en présence de fourmis. Les fourmis sont capables de les attraper entre leurs pattes avant et leurs mandibules pour les retirer de la plante. Les fourmis fréquentent préférentiellement la tige et la stipule (petites feuilles situées sur la tige, à la base des pétioles) de la plante et relativement moins les feuilles de *Macaranga tribola* Après 1h de suivi, les oeufs restants sur chacun des organes ont été comptabilisés et les résultats sont présentés dans le tableau ci-dessous.

organe	initialement	restants après 1h
face supérieure de la feuille	61	18
face inférieure de la feuille	9	9
tige	16	2
stipules	5	0

Test 7 : Lierre et fourmis

20 *Macaranga tribola* colonisés par des fourmis ont été observés et la présence ou l'absence de lierre chaque plante a été notée. Les résultats sont présentés dans le tableau ci-dessous.

fourmis	avec lierre	sans lierre
oui	2	18

On sait, de précédentes études qu'en absence de fourmis, 20% des arbres sont colonisés par du lierre.

Test 8 : Lierre et fourmis

42 sites ont été sélectionnés et à chaque site, deux *Macaranga tribola* ont été suivis : l'un sur lequel des fourmis étaient présentes et l'autre non. Sur chacun des 84 *Macaranga tribola* on a artificiellement enroulé du lierre dont la longueur avait été préalablement notée. Après 2 semaines, le lierre a de nouveau été mesuré et son accroissement a été calculé. Pour chaque site, on a ensuite calculé la différence entre l'accroissement du lierre sur l'arbre associé aux fourmis et l'accroissement du lierre sur l'arbre non-associé aux fourmis.
A partir des ces 42 différences calculés à partir des mesures, on obtient une valeur moyenne de $-0.74cm$ et une variance de $0.37cm^2$.

7 Exercices complémentaires

Vous pouvez vous entraîner à résoudre ces exercices complémentaires, qui sont tirés des annales de l'examen de Biologie des années antérieures. Si vous avez des questions, contactez votre chargé de TD biologiste par mail.

Exercice 11. Régulation de la température chez les crabes
Des chercheurs analysent la température corporelle de 100 crabes prélevés dans la zone intertidale (zones des marées) et se demandent si leur température corporelle moyenne de 20 degrés Celsius s'ajuste sur la température ambiante de 23 degrés Celsius. Décrire la variable étudiée. On supposera que cette variable suit une loi Normale $\mathcal{N}(\mu, \sigma^2)$. Posez les hypothèses H0 et H1 permettant de répondre à la question des chercheurs. Quelle hypothèse supplémentaire faut-il faire concernant l'échantillonnage des crabes ? Quel test statistique feriez-vous pour répondre à la question ?

Exercice 12. Fiabilité des appareils de mesure : cytométrie de flux
La cytométrie de flux est utilisée pour caractériser des échantillons biologiques. Un cytomètre comprends trois parties : un réseau fluidique constitué d'une veine liquide s'écoulant à vitesse constante qui entraîne et focalise un deuxième flux liquide contenant l'échantillon, un banc optique permettant la détection, un microprocesseur qui convertit les signaux électriques en signaux numériques. Pour comparer des échantillons biologiques, il est important que le débit du fluide reste constant. On peut mesurer ce débit en utilisant un échantillon d'étalonnage contenant un nombre connu de micro-billes de même calibre.
Le débit (en centaines de billes par seconde ou **CBS**) du cytomètre de flux du laboratoire A est modélisé par une loi normale $\mathcal{N}(m_A; \sigma^2)$, et celui du cytomètre de flux du labo B par une loi normale $\mathcal{N}(m_B; \sigma^2)$.

1. Interpréter les paramètres m_A, m_B, σ.
2. D'après les constructeurs, $\sigma = 1$ CBS et le débit annoncé est $m = 5$ CBS. En utilisant l'étalon, on relève les débits de 12 mesures faites avec le cytomètre A : $X_1, X_2, ..., X_{12}$ et de 13 mesures faites avec le cytomètre B : $Y_1, Y_2, ..., Y_{13}$ et l'on note :

$$\bar{X} = \frac{X_1+...+X_{12}}{12} \quad ; \quad \bar{Y} = \frac{Y_1+...+Y_{13}}{13}$$

   On observe $\bar{X}_{obs} = 5.83$ CBS et $\bar{Y}_{obs} = 5.03$ CBS. Les cytomètres de flux de chacun des laboratoires sont-ils conformes à l'annonce des constructeurs ?

Exercice 13. Efficacité des méthodes de prélèvement sanguin
Le paludisme est l'endémie la plus étendue au monde. Pour mieux comprendre la biologie du parasite *Plasmodium falciparum*, qui réalise une partie de son cycle chez un hôte humain, il est nécessaire de disposer de moyens de prélèvement sanguins simples et efficaces pour les analyses. Une étude a été menée dans deux villages du Burkina-Faso pour déterminer le mode de prélèvement sanguin (capillaire ou veineux) permettant d'obtenir la plus forte concentration pour le dosage de la charge parasitaire. Des écoliers agés de 6 à 9 ans ont été examinés et ont subi des prélèvements sanguins au pli du coude pour le sang veineux et au bout du doigt pour le sang capillaire. Deux frottis minces ont été confectionnés par élève. Les frottis identifiés par numéro sont immédiatement ramenés au laboratoire pour la coloration et la lecture. La numération des formes asexuées de *Plasmodium falciparum* est exprimée en nombre de globules rouges parasités (GRP) par microlitre (μl) de sang sur la base de 200 hématies par champ microscopique et de 4 000 000 hématies par microlitre de sang. Sur la base d'une lecture de 200 champs par échantillon, le seuil de détection des parasites est de 100 globules rouges parasités par microlitre de sang. Tous les élèves fébriles ou ayant une infection mixte ont été exclus de l'étude. Au total, 489 écoliers âgés de 6 à 9 ans ont été examinés et 108 ont été trouvés porteurs de plasmodium et retenus pour l'étude. On dispose, pour chaque enfant, d'une mesure de la densité parasitaire (GRP/μl) pour les prélèvements veineux et capillaires.

1. En utilisant le tableau ci-dessous, tracez l'allure des distributions empiriques de la densité parasitaire pour les prélèvements veineux et capillaires.

Prélèvement	minimum	médiane	quantile 75%	maximum
capillaire	0	300	900	50000
veineux	0	200	900	36100

Quantiles de la distribution de la densité parasitaire dans l'échantillon, selon le type de prélèvement.

2. A votre avis, à quoi ressemble la distribution empirique de la différence de densité parasitaire entre les prélèvements veineux et capillaires ?

3. On cherche à savoir quelle méthode fournit la plus forte concentration parasitaire. Construisez un test statistique pour répondre à cette question. Vous vous arrêterez à la proposition d'une statistique de test et sa loi sous H0.

tiré de : Médecine d'Afrique Noire : 1991, 38 (8/9)

Exercice 14. Biologie de la conservation - Effet de l'urbanisation sur la diversité des espèces

L'urbanisation est une des causes majeures d'extinction, car elle est généralement synonyme de destruction d'habitat. Une étude du ministère de l'agriculture réalisée entre 1992 et 2002 montre que l'urbanisation en France s'est beaucoup développée. Des chercheurs ont mesuré le nombre d'espèces d'oiseaux spécialistes d'un habitat, et le nombre d'espèces d'oiseaux généralistes, c'est-à-dire adaptés à plusieurs habitats. Ils ont réalisé ces mesures en 1992 dans une zone non urbaine et en 2002 dans cette même zone urbanisée. Ils veulent savoir si les effectifs des espèces généralistes et des espèces spécialistes sont affectés de la même façon par l'urbanisation. Décrire les variables étudiées, donner le test statistique à faire pour répondre à cette question et poser H0 et H1. Donnez les paramètres de la loi de probabilité que vous utiliseriez pour réaliser ce test avec ces données.

Exercice 15. Répartition des naissances au long de l'année chez l'homme

Le tableau ci-dessous est un extrait de données recueillies par l'INSEE et donne la répartition des naissances au cours de l'année 2002 dans le département de la Creuse, en France métropolitaine, ainsi que en Martinique. La dernière ligne du tableau donne les fréquences observées sur l'ensemble de la population de France Métropolitaine.

Mois	jan	fev	mar	avr	mai	juin	juil	aout	sept	oct	nov	dec
Creuse	71	63	81	82	85	95	82	94	84	73	75	72
Martinique	467	415	408	421	391	353	362	405	523	511	487	512
Métropole	0.083	0.077	0.079	0.079	0.080	0.082	0.089	0.086	0.085	0.087	0.084	0.086

1. Les statisticiens de l'INSEE voudraient savoir si l'échantillon de la Creuse est représentatif de la population française métropolitaine pour ce qui concerne la répartition des naissances au cours de l'année.

   (a) Donnez les premières étapes du test statistique, jusqu'au choix de la statistique de test et sa loi sour H_0.

   (b) Les chercheurs réalisent un test de Chideux de conformité. Ils trouvent une valeur observée de la statistique de 12.52 et une probabilité critique de 0.32. Quelle conclusion faites-vous au seuil de 5% ? au seuil de 1% ? Quelle est la probabilité de vous tromper sur votre conclusion ?

2. Les statisticiens cherchent ensuite à savoir si la distribution des naissances est influencée par la lattitude de la région dans laquelle vivent les futurs parents. Pour cela, ils comparent la répartition des naissances en 2002 dans le département de la Creuse et en Martinique.

   (a) Quelles sont les variables aléatoires étudiées ? Donnez les valeurs possibles et faites une hypothèse sur les lois de probabilité associées.

(b) Décrivez les échantillons étudiés (de quelle population est issue chaque échantillon, taille de l'échantillon)

(c) Ennoncez les hypothèses H_0 et H_1 correspondant à la question que se posent les statisticiens

(d) Proposez une statistique de test et donnez sa loi sous H_0.

(e) Calculez les effectifs attendus de naissances au mois de février 2002 dans le département de la Creuse sous l'hypothèse H_0

(f) Les données permettent-elles de répondre à la question posée ?

Exercice 16. Génétique de la longueur des ailes chez la drosophile
Les drosophiles *D. melanogaster* de la souche sauvage ont les ailes longues. On dispose d'une souche pure mutante appelée *miniature* où les individus possèdent des ailes de petite taille (phénotype noté [*mini*]). Le croisement entre des mouches de la souche *miniature* et des mouches de la souche sauvage donne des descendants F1 de phénotype sauvage, quelque soit le sens du croisement. On procède au croisement entre deux individus F1 et parallèlement au croisement entre une femelle F1 et un mâle de souche pure [mini]. On obtient les résultats suivants pour chaque population F2:

Croisement	Phénotype	
	Ailes longues	Ailes miniature
F1 x F1	460	149
F1 x [mini]	256	242

1. Pourquoi ces deux types de résultats conduisent-ils à la même interprétation génétique ? Décrivez dans chaque cas la variable aléatoire et le modèle utilisé. Proposez des hypothèses H0/H1.

2. Réalisez les test statistiques correspondants.

Exercice 17. Régime de reproduction chez *Lloydia serotina*
Le régime de reproduction décrit la façon dont se rencontrent les gamètes au moment de la reproduction sexuée. Il est extrêmement variable chez les êtres vivants. Chez les mammifères, les sexes sont séparés (dioecie) et déterminés génétiquement par des chromosomes sexuels hétéromorphes (femelle XX et mâle XY). Chez d'autres espèces animales (nématodes, crustacées) ou végétales (60 espèces répertoriées), on observe dans les populations des mâles (Mm) et des hermaphrodites (mm). Les individus portant l'allèle M sont des mâles. Les hermaphrodites peuvent s'autoféconder. Ce régime de reproduction est appellé androdioecie. Enfin, il existe des espèce entièrement hermaphrodites (monoecie). Une hypothèse pour expliquer cette grande variabilité est que le régime de reproduction d'une espèce est le résultat d'un compromis entre l'avantage à la reproduction sexuée (variabilité génétique) et la difficulté à trouver un partenaire sexuel. L'androdioecie serait un régime de reproduction transitoire qui pourrait évoluer soit vers la dioecie, soit vers l'hermaphroditisme.

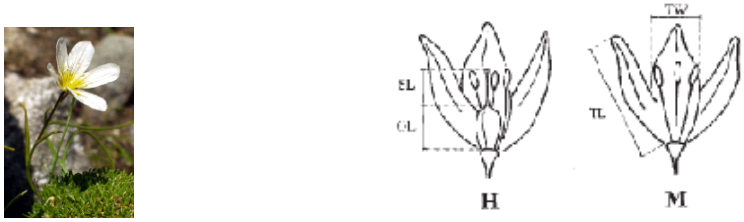


Figure 3: **Morphologie des fleurs de *Lloydia serotina*** H = Fleur hermaphrodite, M= Fleur mâle.

1. Chez une espèce dioïque, il est possible de montrer que le sex-ratio attendu est 50/50 (pour vous en convaincre, calculez le sex ratio attendu dans un troupeau de jeunes bovins issus de la reproduction entre 1 mâle et 50 femelles). Que pensez-vous du sex-ratio attendu chez une espèce androdioïque ?

2. Lloydia serotina est une espèce végétale artique-alpine pérenne, qui ne produit qu'une seule fleur par plante et par année. On observe dans les populations de Lloydia serotina des plantes hermaphrodites et des plantes mâles (Figure 3). Le régime de reproduction pourrait être l'androdioécie. Une hypothèse alternative est que les hermaphrodites de cette espèce sont en réalité des femelles avec des organes mâles non fonctionnels (dioecie). Des chercheurs ont compté dans une population alpine le nombre de plantes mâles et le nombre de plantes hermaphrodites. A quelle question biologique peuvent-ils répondre à partir de ces données ? Décrire les variables étudiées, poser H0 et H1, et donner le test statistique à faire pour répondre à cette question.

d'après Manicacci et Després, Can. J. Bot. Vol. 79, 2001

Exercice 18. L'efficacité des vaccins contre la grippe saisonnière chez l'homme
Les résultats suivants ont été obtenus lors d'une étude visant à évaluer l'efficacité respective de deux vaccins contre la grippe saisonnière (Monto et al, 2009, The New England Journal of Medecine). Le vaccin **TIV** contient le virus atténué. La vaccin **LIV** contient le virus vivant atténué et peut être administré par brumisation. Les données ont été recueillies sur des participants volontaires âgés en moyenne de 23 ans et fréquentant les campus universitaires de l'état du Michigan (USA). Il s'agit d'une expérience en double aveugle. Les participants ont été divisés en trois lots selon le vaccin administré : **TIV**, **LIV** ou **Placebo**. Le traitement a été administré en novembre 2008, et les patients ont été suivis jusqu'en avril 2009. En cas de syndrome grippal, un prélèvement était effectué pour confirmer qu'il s'agissait bien de la grippe.

Lot	**TIV**	**LIV**	**Placebo**	Total
grippe avérée	28	56	35	119
absence symptomes	785	758	290	1833
Total	813	814	325	1952

1. (a) Décrivez le modèle permettant de tester s'il y a une différence entre les trois traitements, notamment les variables aléatoires considérées.
   (b) Quelle est l'hypothèse H0 ?
   (c) Calculez les effectifs théoriques attendus dans chaque cas sous l'hypothèse H0
   (d) La statistique de test calculée est $Z_{obs} = 21.91$ et la probabilité critique associée est 0.000001. Quelle conclusion en tirez-vous au seuil 5% ?

2. On peut utiliser ces données pour tenter de prédire la proportion de personnes qui seront atteintes par la grippe saisonnière entre novembre et avril, en fonction du nombre de personnes qui acceptent de se faire vacciner par le vaccin **TIV**. On considère ici uniquement les personnes "jeunes" (moins de 65 ans) qui ont fait l'objet de l'étude.
   (a) Dans l'échantillon étudié, estimez la probabilité P_V d'avoir la grippe sachant que l'on est vacciné avec **TIV**
   (b) Dans l'échantillon étudié, estimez la probabilité P_{NV} d'avoir la grippe sachant que l'on n'est pas vacciné
   (c) Quelle serait la proportion de cas de grippe saisonnière dans le pays si 26% de la population était vaccinée avec **TIV** ? Pour simplifier, on supposera que la probabilité d'attraper la grippe est indépendante du nombre de personnes vaccinées.
   (d) En France, environ 26% des personnes de moins de 65 ans se font vacciner tous les ans contre la grippe saisonnière par un vaccin de type **TIV**. Les données recueillies par le réseau Sentinelle montrent que la proportion de personnes atteintes par la grippe entre novembre et avril varie selon les années mais se situe aux environs de 5% de la population. Quelles hypothèses peut-on faire pour expliquer la différence entre l'incidence de la grippe en France et l'incidence prédite à partie des données de l'étude américaine ?

Exercice 19. Les paramécies et leurs parasites

Holospora undulata est un parasite du micronoyau des paramécies. Après l'infection, il se multiplie à l'intérieur de son hôte sous une forme encapsulée **FR** et se propage par transmission verticale lors de la division cellulaire des paramécies. Ce n'est qu'au bout d'un certain temps que les formes **FR** se transforment en formes filamenteuses **FI** infectieuses. Petit à petit, les formes **FI** remplissent le micronoyau. Puisque les paramécies sont apparement capables de se reproduire même en étant infectées, des chercheurs se demandent si le parasite affecte réellement la survie de l'hôte et dans quelle mesure il modifie la dynamique des populations. Ils mesurent les densités de populations dans des cultures de paramécies infectées par le parasite. Le caractère mesuré est le nombre de paramécies (en milliers d'individus) dans une culture après 16h d'incubation. Au départ, toutes les cultures contiennent le même nombre de paramécies. On sait par ailleurs que, dans les cultures non infectées et réalisées avec le même protocole, la moyenne est $\mu = 23$ milliers d'individus. Le tableau ci-dessous donne la distribution du nombre d'individus après 16h de culture pour 55 échantillons de cultures infectées par *Holospora undulata*. Pour simplifier les calculs, les données ont été regroupées en classes.

nb. paramécies à $T16$ (en milliers)	[7-11[	[11-15[	[15-19[	[19-23[	[23-27[	[27-31[
nb échantillons	1	6	20	18	9	1

1. Calculez la moyenne et la variance empirique de la population.
2. On aimerait savoir si les cultures infectées contiennent moins d'individus que les cultures non infectées. Donnez le modèle statistique et les hypothèses H_0 et H_1 permettant de répondre à cette question.
3. Ecrivez la statistique de test et la loi de cette statistique sous l'hypothèse H_0. (Aucun calcul n'est attendu)

8 Annexe : Tables de nombres au hasard

Table 5_1 : nombres au hasard

53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 09
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60	75 91 12 81 19
35 30 58 21 46	06 72 17 10 94	25 21 31 75 96	49 28 24 00 49	55 65 79 78 07
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98	00 83 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 27
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80	13 29 54 19 28
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57	65 65 80 39 07
03 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 95 30 27 59	37 75 41 66 48	86 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	98 77 27 85 42	28 88 61 08 84	69 62 03 42 73
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39	48 03 45 15 22
50 44 66 44 21	66 06 58 05 62	68 15 54 35 02	42 35 48 96 32	14 52 41 52 48
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51	03 37 18 39 11
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	78 35 34 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
55 19 68 97 65	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 98
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 00
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 82	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 68 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 02 60

Table 5_2 : nombres au hasard

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62
36 67 40 67 14	64 05 71 95 86	11 05 65 09 68	76 83 20 37 90	57 16 00 11 66
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08	07 52 74 95 80
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90	49 37 38 44 59
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36	47 95 93 13 30
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62	02 67 74 17 33
05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 56 78 06 83	52 91 05 70 74
07 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93	58 05 77 09 51
68 71 86 85 85	54 87 66 47 54	73 32 08 11 12	44 95 92 63 16	29 56 24 29 48
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74	94 44 67 16 94
14 65 52 68 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50	15 29 39 39 43
17 53 77 58 71	71 41 61 50 72	12 41 94 96 26	44 95 27 36 99	02 96 74 30 83
90 26 59 21 19	23 52 23 33 12	96 93 02 18 39	07 02 18 36 07	25 99 32 70 23
41 23 52 55 99	31 04 49 69 96	10 47 48 45 88	13 41 43 89 20	97 17 14 49 17
60 20 50 81 69	31 99 73 68 68	35 81 33 03 76	24 30 12 48 60	18 99 10 72 34
91 25 38 05 90	94 58 28 41 36	45 37 59 03 09	90 35 57 29 12	82 62 54 65 60
34 50 57 74 37	98 80 33 00 91	09 77 93 19 82	74 94 80 04 04	45 07 31 66 49
85 22 04 39 43	73 81 53 94 79	33 62 46 86 28	08 31 54 46 31	53 94 13 38 47
09 79 13 77 48	73 82 97 22 21	05 03 27 24 83	72 89 44 05 60	35 80 39 94 88
88 75 80 18 14	22 95 75 42 49	39 32 82 22 49	02 48 07 70 37	16 04 61 67 87
90 96 23 70 00	39 00 03 06 90	55 85 78 38 36	94 37 30 69 32	90 89 00 76 33