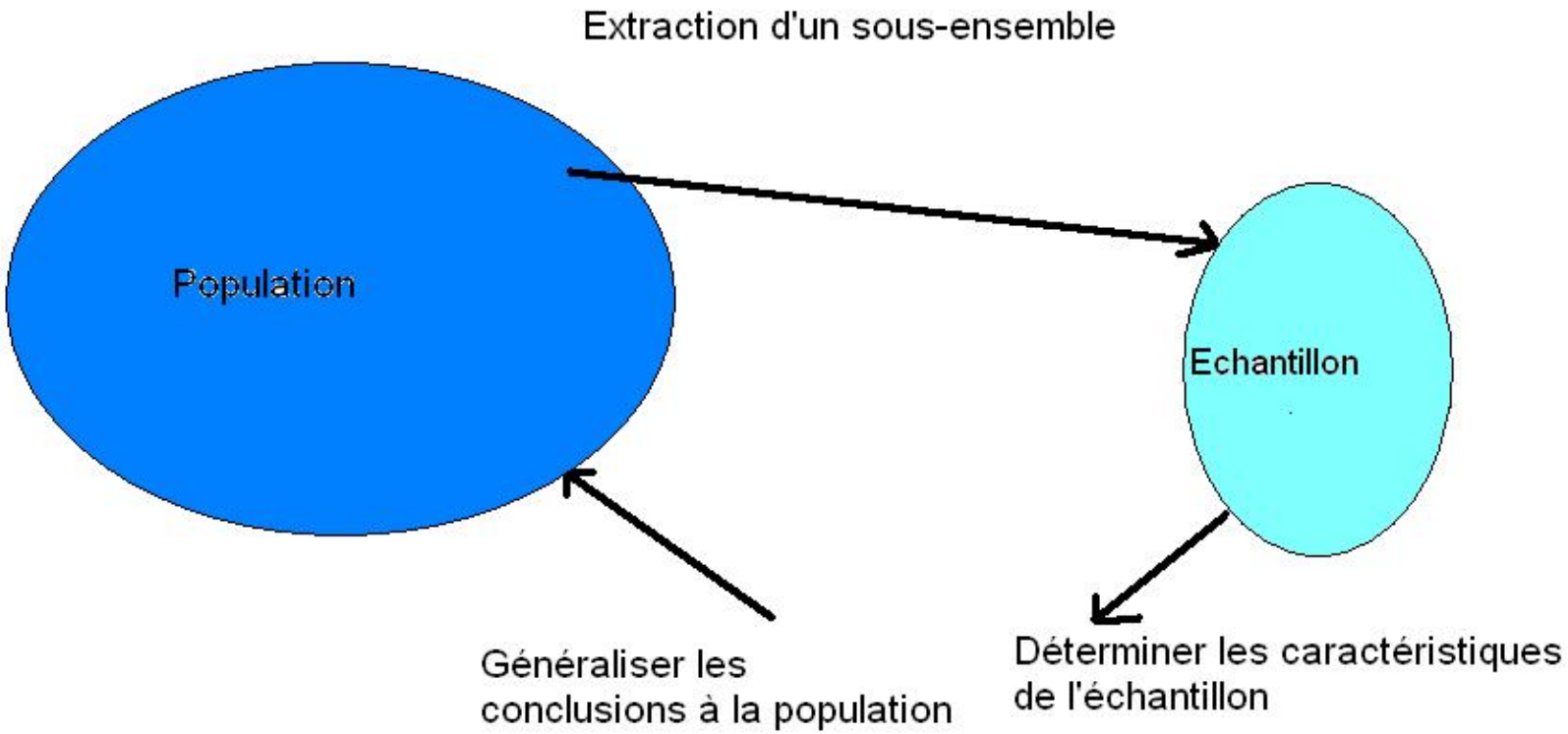
Introduction à la théorie des probabilités et à la statistique

Ségolen Geffray

Ecole doctorale Biologie/Chimie

Année 2008-2009

La démarche


Extraction d'un sous-ensemble
Population
Echantillon
Généraliser les
conclusions à la population
Déterminer les caractéristiques
de l'échantillon

Un problème concret

- Un laboratoire pharmaceutique souhaite vérifier l'efficacité d'un nouveau traitement avant d'envisager une mise sur le marché.
- Pour cela, il propose donc d'évaluer la proportion de patients guéris par le nouveau traitement.
- Comment faire ? on ne peut pas tester le traitement sur tous les malades du monde !
- On tire un échantillon de malades auxquels on alloue au hasard le nouveau traitement ou un placebo.
- On réalise l'expérience, on suit les patients, on calcule la proportion échantillonnale de patients guéris par le nouveau traitement et on la compare à la proportion échantillonnale de patients guéris sous placebo.

Un problème concret (suite)

- On approxime la proportion de patients guéris par le nouveau traitement de la **population** entière par la proportion de patients guéris par le nouveau traitement de l'**échantillon** et on approxime la proportion de patients guéris sous placebo de la **population** entière par la proportion de patients guéris sous placebo de l'**échantillon**.
- Cela permet de généraliser les conclusions de l'étude effectuée sur l'**échantillon** à la **population** entière des malades.
- Comment savoir si ce qu'on vient de faire est licite ? Quelle est la qualité de l'approximation ? Il faut étudier la théorie des probabilités et la statistique !
- Si on tire un autre échantillon, il y a de fortes chances que l'on n'obtienne pas exactement les mêmes résultats. Ces fluctuations (ou erreurs d'échantillonnage) sont dues à la variabilité. Cela signifie que des sujets semblables en apparence peuvent présenter des différences lorsqu'on effectue des mesures.

Distinction Proba/Stat

- **La Théorie des probabilités** :
  - permet de modéliser des phénomènes aléatoires et d'y effectuer des calculs théoriques
  - concerne les **populations** : on ne peut donc pas faire de mesures.
- **La Statistique** :
  - concerne les **échantillons**, le monde réel, la pratique,
  - on fait des mesures (observations) sur des individus,
  - repose sur la modélisation probabiliste des observations.

Les différents aspects de la Statistique

Observer ne suffit pas, il faut interpréter !

- **Statistique descriptive** :
  - Résumer les mesures sur un échantillon (moyenne, variance,...)
  - Représenter les mesures (histogramme, distribution)
- **Statistique inférentielle** :
  - Généraliser les propriétés d'un échantillon à une population en prenant en compte les fluctuations d'échantillonnage
  - il faut modéliser les observations (variables aléatoires)
- **Tests d'hypothèses** :
  - Contrôler la validité d'un modèle
  - Comparer plusieurs échantillons
- **Statistique décisionnelle** :
  - Savoir prendre une décision pour une personne donnée alors que les résultats sont exprimés en termes de probabilités (i.e. de pourcentage de chances, de risques)

Pourquoi la biostatistique ?

A cause de la variabilité omniprésente en biologie et en médecine : cela signifie que les variables d'individus semblables en apparence peuvent prendre en réalité des valeurs différentes

- variabilité = métrologique + biologique
- variabilité biologique = inter-individuelle + intra-individuelle
- variabilité au niveau :
  - populationnel : conséquences de la très grande prématurité en termes de handicap neurologique et comportemental de l'enfant (c'est la base du darwinisme : pas de sélection sans variabilité)
  - fonctionnel : contrôle de la glycémie
  - cellulaire : durée du cycle cellulaire
  - génomique : séquences codant pour une protéine, ex : de nombreux gènes prédisposent aux cancers
  - moléculaire : structure des protéines

Conséquence de la variabilité en médecine : des prognostics variables, efficacité des traitements difficile à évaluer.

$1^{\text{ère}}$ partie

Notions de théorie des probabilités

Expérience aléatoire, évènements

- Une expérience est **aléatoire** si on ne peut pas prévoir à l'avance son résultat, et si, répétée dans des conditions identiques, elle peut donner lieu à des résultats différents. Une expérience permet l'obtention de valeurs appelées **réalisations**.
- Ensemble fondamental (noté Ω) : ensemble de tous les résultats possibles. Il peut être :
  - fini, par ex $\{x_1, ..., x_k\}$
  - infini dénombrable : on peut indicer, numéroter ses éléments jusqu'à l'infini, par ex $\{x_1, , x_2, ..., x_n, ...\}$
  - infini non-dénombrable : ceci signifie qu'il n'est pas possible de décrire l'ensemble sous la forme d'une liste numérotée $\{x_1, x_2, ..., x_k, ...\}$, par ex $[0, 1]$.
- **Evènement** : un des résultats possibles de l'expérience.

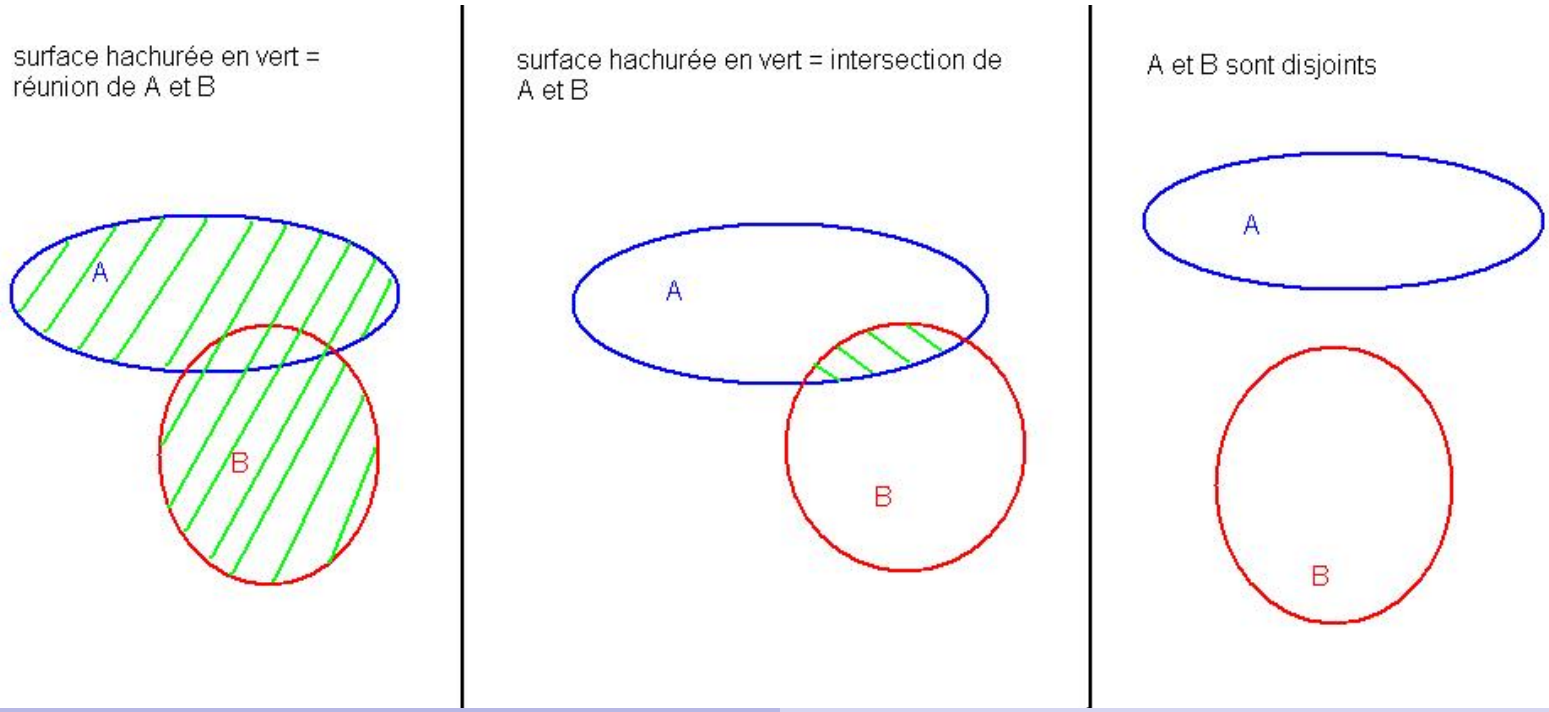
Exemple d'univers

- Soit l'expérience "*un homme se présente au laboratoire d'analyse, effectue un test de groupe sanguin*". L'univers associé à cette expérience est $\{A, B, AB, O\}$.
- Soit l'expérience "*un médecin épidémiologiste compte le nombre de personnes atteintes par la grippe en une journée dans un département donné*". L'univers associé à cette expérience est $\{0, 1, 2, 3, ...\}$.
- Soit l'expérience "*le technicien de laboratoire pèse une souris adulte et note son poids*". Comme une souris adulte pèse entre 10g et 30g, l'univers associé à cette expérience est $[10, 30]$.

Rappel : opérations sur les ensembles

Soient A et B deux évènements d'un ensemble fondamental Ω

- $\{A \text{ ou } B\} = A \cup B =$ réunion de A et B
- $\{A \text{ et } B\} = A \cap B =$ intersection de A et B
- complémentaire de A dans $\Omega = \bar{A} = \Omega - A$
- $\emptyset =$ évènement impossible
- $\Omega =$ évènement certain
- A et B sont incompatibles ou disjoints lorsque $A \cap B = \emptyset$



Règle de calcul des probabilités

- Une probabilité est une **fonction** notée $\mathbb{P}$ qui attribue à tout évènement A une valeur $\mathbb{P}(A)$ désignant la probabilité que A se réalise.
- Une probabilité possède les propriétés suivantes :
  - $0 \leq \mathbb{P}(A) \leq 1$ pour tout évènement A
  - $\mathbb{P}(\Omega) = 1$
  - $\mathbb{P}(\emptyset) = 0$
  - $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$
  - $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
  - en général, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - mais si A et B sont disjoints, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
  - en général, on a $\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$
  - mais si les A_i sont 2 à 2 disjoints, $\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$

Probabilités conditionnelles : introduction

- Considérons une expérience réalisée sur une certaine population et un évènement A qui a une probabilité $\mathbb{P}[A]$ de se réaliser,
  par ex : A = présence d'une maladie M.
- Que devient $\mathbb{P}[A]$ si on se restreint à une sous-population ?
  par ex : sous-population = les individus présentant un signe S.
- On introduit un évènement B conditionnant, qui définit la sous-population,
  par ex : B = présenter le signe S.
- $\mathbb{P}[B]$ ne doit pas être nul.

Probabilités conditionnelles : définition

- La probabilité que l'évènement A se réalise sachant que l'évènement B a eu lieu (=probabilité de A parmi la sous-population caractérisée par B) est définie par :

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

- De même, la probabilité que l'évènement B se réalise sachant que l'évènement A a eu lieu est définie par :

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]}.$$

- Ne PAS confondre $\mathbb{P}[A|B]$= probabilité que A se réalise sachant qu'on a observé B avec $\mathbb{P}[A \cap B]$=probabilité que A et B se réalisent simultanément ! !

Exemple de probabilités conditionnelles

Durant l'hiver, la probabilité pour qu'une personne ait la grippe est estimée à 30%. Le diagnostique clinique est posé lorsque la personne présente les symptômes suivants : courbatures, fièvre subite, signes respiratoires. Une personne ayant la grippe a 80 chances sur 100 d'avoir ces symptômes.

- $\mathbb{P}[\text{grippe}] = 0.3$
- $\mathbb{P}[\text{symptômes}|\text{grippe}] = 0.8$
- On peut alors obtenir :

$$\begin{aligned}
&\mathbb{P}[\text{symptômes et grippe}] \\
&= \mathbb{P}[\text{grippe}] \times \mathbb{P}[\text{symptômes}|\text{grippe}] \\
&= 0.3 \times 0.8 = 0.24
\end{aligned}$$

Autre exemple de probabilités conditionnelles

Considérons les chiffres suivants valables pour la France :

- $\mathbb{P}[\text{être VHC+}] = 600000/60000000 = 1\%$
- $\mathbb{P}[\text{être VHC+ sachant que âge} = 15 \text{ ans}] = \text{faible, certainement} < 10^{-4}$
- $\mathbb{P}[\text{être VHC+ sachant qu'il y a toxicomanie IV depuis} > 5 \text{ ans}] = \text{forte, certainement} > 50\%$
- $\mathbb{P}[\text{être VHC+ sachant que la personne est asthmatique}] = 1\%$

Probabilités conditionnelles : règles de calcul

- Soit B un évènement **fixé**. La fonction $A \longrightarrow \mathbb{P}[A|B]$ est une vraie probabilité i.e. les règles de calcul avec les probabilités conditionnelles sont les mêmes qu'avec les probabilités classiques.
- $0 \leq \mathbb{P}(A|B) \leq 1$ pour tout évènement A
- $\mathbb{P}(\Omega|B) = 1$
- $\mathbb{P}(\emptyset|B) = 0$
- $\mathbb{P}(\overline{A}|B) = 1 - \mathbb{P}(A|B)$
- $A_1 \subset A_2 \Rightarrow \mathbb{P}(A_1|B) \leq \mathbb{P}(A_2|B)$
- en général,
  $\mathbb{P}(A_1 \cup A_2|B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B) - \mathbb{P}(A_1 \cap A_2|B)$
- mais si A_1 et A_2 sont disjoints,
  $\mathbb{P}(A_1 \cup A_2|B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B)$
- en général, on a $\mathbb{P}\left(\cup_{i=1}^{n} A_i|\, B\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i|B)$
- mais si les A_i sont 2 à 2 disjoints,
  $\mathbb{P}\left(\cup_{i=1}^{n} A_i|B\right) = \sum_{i=1}^{n} \mathbb{P}(A_i|B)$

Indépendance de 2 évènements : introduction

- définition sans formule : soient A et B deux évènements. Si, lorsqu'on reçoit l'information que B s'est produit, cela ne modifie pas la probabilité de A, on dit que A et B sont indépendants. Autrement dit, des événements indépendants n'apportent pas d'information l'un sur l'autre.
- ex : l'asthme et le VHC (diapo précédente) sont indépendants : l'un n'aide pas au diagnostic de l'autre. Notons que $\mathbb{P}[\text{être VHC+}] = 1\%$ et que $\mathbb{P}[\text{être VHC+ sachant que la personne est asthmatique}] = 1\%$.

Indépendance de 2 évènements : définition formelle

- 1ère définition avec formule : A et B sont indépendants si

$$\mathbb{P}[A|B] = \mathbb{P}[A] \qquad \text{ou/et} \qquad \mathbb{P}[B|A] = \mathbb{P}[B].$$

- ex : La fréquence du VHC est 1% : $\mathbb{P}[\text{VHC+}] = 0.01$. La fréquence du VHC chez les asthmatiques est également de 1% : $\mathbb{P}[\text{VHC+}|\text{asthmatique}] = 0.01$.
- 2ème définition avec formule : A et B sont indépendants si

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B].$$

Cette formule est symétrique en A et B, on en déduit que si $\mathbb{P}[A|B] = \mathbb{P}[A]$, alors on a aussi $\mathbb{P}[B|A] = \mathbb{P}[B]$.
ex : $\mathbb{P}[\text{asthmatique}|\text{VHC+}] = \mathbb{P}[\text{asthmatique}]$ puisque asthme et VHC sont indépendants.

Indépendance et incompatibilité

- Ne pas confondre des événements incompatibles et des évènements indépendants...
- ex : considérons A=”l'enfant à naître est un garçon” et B=”l'enfant à naître est une fille”. Les évènements A et B sont incompatibles. Mais ils ne sont pas indépendants ! ! ! En effet,

$$\mathbb{P}[A \cap B] = 0 \neq \mathbb{P}[A] \times \mathbb{P}[B] = 0.5 \times 0.5 = 0.25$$

Exercice : test diagnostic

Un laboratoire commercialisant un test diagnostic pour le dépistage d'une maladie M décide d'en vérifier la fiabilité. Les chiffres sont les suivants :

- la prévalence de la maladie dans la population est de 25%
- 95 fois sur 100, le test diagnostic s'est révélé positif alors que la personne était réellement atteinte par M
- 1 fois sur 100, le test diagnostic s'est révélé positif alors que la personne n'était pas atteinte par M

1. Quelle est la probabilité que le test diagnostic donne une indication correcte ?
2. Quelle est la probabilité qu'une personne soit réellement atteinte par M lorsque le test diagnostic est positif ?

Définition d'une variable aléatoire

- Une **variable aléatoire** est le procédé aléatoire qui mène à un nombre ou à une modalité.
- Pour une variable aléatoire X, on note D_X l'ensemble des valeurs possibles pour X.
- La **loi de probabilité** P d'une variable aléatoire décrit quelles sont les valeurs ou modalités possibles prises par la variable et avec quelle probabilité ces différentes valeurs ou modalités sont prises.
- La **théorie des probabilités** vise à évaluer le comportement des variable aléatoires (espérance, moments, probabilités de dépassement, comportement de sommes,...) étant donné la loi de probabilité P.
- La **statistique** fournit des méthodes pour résoudre le problème inverse dit d'inférence statistique : caractériser P au vu des observations des variables.

Exemples de variables aléatoires et d'évènements associés

- Au laboratoire, on dispose d'un lot de 30 prélèvements sanguins sur lesquels on procède au dépistage d'une maladie M. Soit X la variable aléatoire qui compte le nombre de prélèvements positifs (M+).
  - L'ensemble des valeurs acceptables pour X est $D_X = \{0, 1, ..., 30\}$.
  - $\{X = 21\}$ code pour l'évènement "21 prélèvements sont positifs".
  - $\{X = 50\} = \emptyset$
  - $\{8.5 \leq X \leq 10.5\}$ code pour l'évènement "9 ou 10 prélèvements sont positifs".
- On s'intéresse au poids des souris d'une certaine race. Soit X la variable aléatoire représentant le poids (en g) d'une souris adulte.
  - L'ensemble des valeurs acceptables pour X est $D_X = [10, 30]$.
  - $\{X = 21\}$ code pour l'évènement "le poids d'une souris est de 21 g".
  - $\{X = 50\} = \emptyset$
  - $\{8.5 \leq X \leq 10.5\}$ code pour l'évènement "le poids d'une souris est compris entre 8.5g et 10.5g".

Les différents types de variables

Au sens strict, une variable aléatoire est un codage des résultats d'expérience donc produit toujours un résultat numérique. Mais un résultat d'expérience est généralement considéré comme une variable aléatoire dans les cas suivants.

- variables **quantitatives** : elles sont numériques. On distingue :
  - les variables **continues** : toute valeur d'un intervalle de $\mathbb{R}$ est acceptable, ex : taille, poids, volume, temps écoulé...
  - les variables **discrètes** : elles prennent un nombre dénombrable (fini ou infini) de valeurs, ex : nb de bactéries dans 1 ml de solution, nb de colonies dans une boîte de Pétri
- variables **qualitatives** : elles expriment l'appartenance à une catégorie ou une modalité. On distingue :
  - les variables **ordinales** : l'ens des catégories est ordonné, ex : cancer stade $\{I, II, III, IV\}$.
  - les variables **nominales** : pas d'ordre, ex : groupe sanguin $\{A, B, AB, O\}$, sexe $\{H, F\}$,...

Loi de probabilité, fonction de répartition et densité

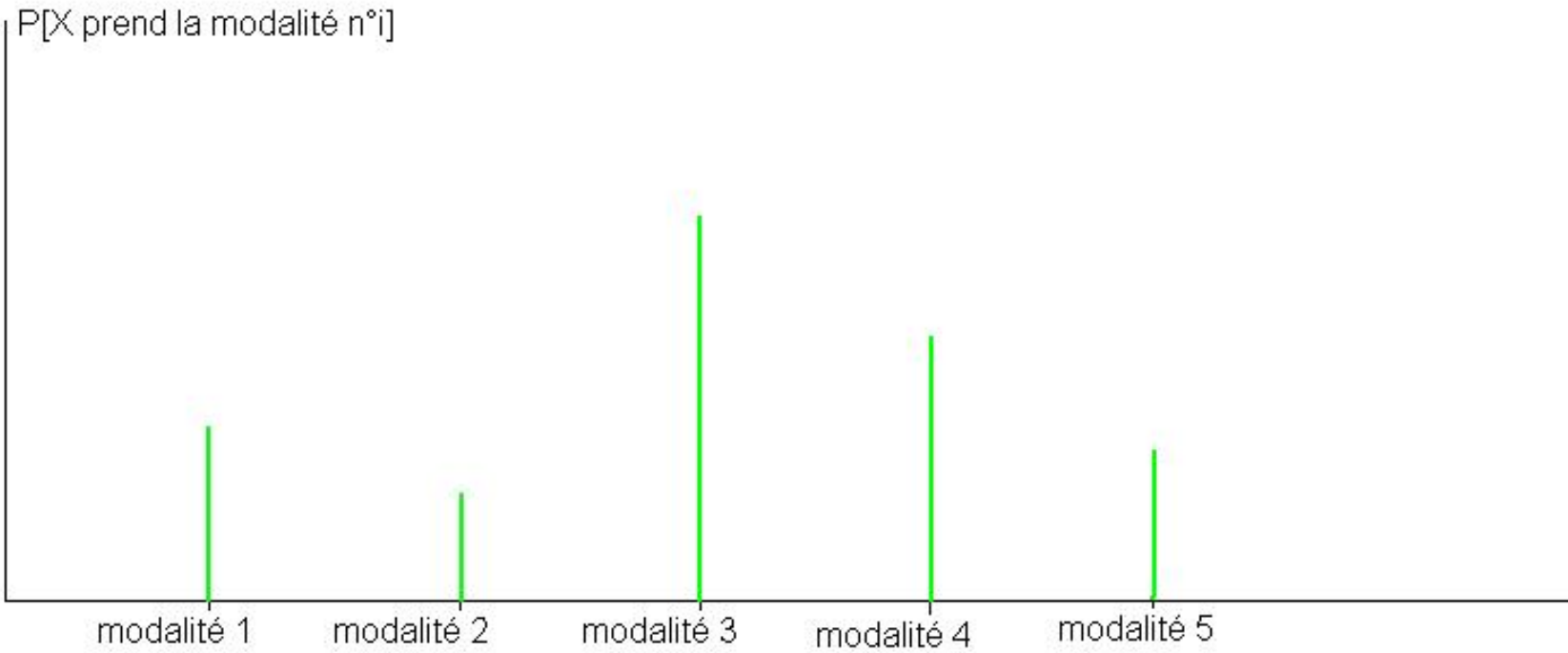
- Pour les variables **qualitatives**, on peut définir la **loi de probabilité** par la donnée des modalités prises et des probabilités associées, ex : X=groupe sanguin d'un sujet, X peut prendre les modalités suivantes : A,B,AB ou O. La loi de probabilité de X est la donnée des nombres suivants : $\mathbb{P}[X = \mathrm{A}]$, $\mathbb{P}[X = \mathrm{B}]$, $\mathbb{P}[X = \mathrm{AB}]$ et $\mathbb{P}[X = \mathrm{O}]$.
- Pour les variables aléatoires **quantitatives discrètes**, on peut définir la **loi de probabilité** par la donnée des valeurs prises et des probabilités associées ou de manière équivalente la **fonction de répartition**.
- Pour les variables aléatoires **quantitatives continues**, on peut définir la **fonction de répartition** ou de manière équivalente la **densité**.

Loi de probabilité d'une variable qualitative

La loi de probabilité d'une variable qualitative est la donnée des nombres

$$p_i = \mathbb{P}[\text{la v.a. } X \text{ prend la modalité n°i}]$$

pour chacune des modalités possibles. Ces nombres sont compris entre 0 et 1 et leur somme vaut 1.



Fonction de répartition d'une variable quantitative

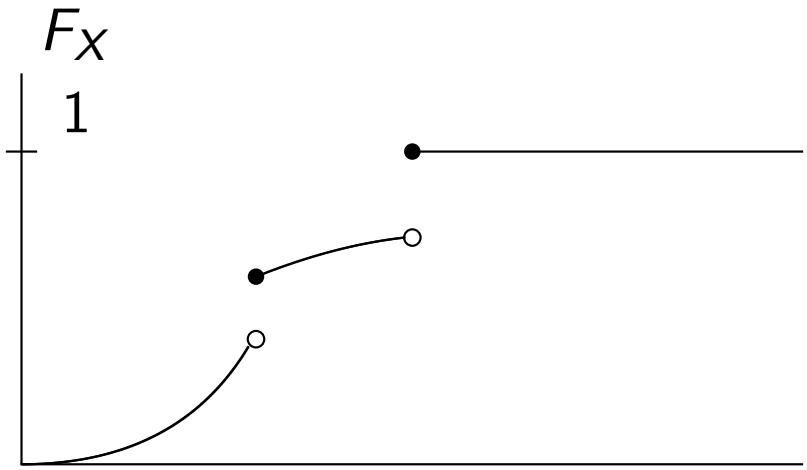
- Soit X une variable aléatoire et soit x un nombre.
- Considérons l'évènement $\{X \leq x\}$ = ensemble des résultats d'expérience dont le codage est inférieur ou égal à x.
- $\mathbb{P}[X \leq x]$ est un nombre qui dépend de la valeur de x
- On définit F_X la fonction de répartition de X par

$$F_X(x) = \mathbb{P}[X \leq x].$$

- Pour tout x, on a $0 \leq F_X(x) \leq 1$ avec $\lim_{x \to +\infty} F_X(x) = 1$ et $\lim_{x \to -\infty} F_X(x) = 0$.
- On note $F_X(x^-) = \mathbb{P}[X < x]$.
- $\mathbb{P}[X > x] = 1 - \mathbb{P}[X \leq x]$ i.e. $\mathbb{P}[X > x] = 1 - F_X(x)$
- $\mathbb{P}[X \leq x] - \mathbb{P}[X < x] = \mathbb{P}[X = x]$ i.e. $F_X(x) - F_X(x^-) = \mathbb{P}[X = x]$
- Pour $a < b$, on a :
  - $F_X(b) - F_X(a) = \mathbb{P}[a < X \leq b]$
  - $F_X(b) - F_X(a^-) = \mathbb{P}[a \leq X \leq b]$
  - $F_X(b^-) - F_X(a) = \mathbb{P}[a < X < b]$
  - $F_X(b^-) - F_X(a^-) = \mathbb{P}[a \leq X < b]$

Quelques propriétés de la fonction de répartition

- F_X est croissante ($x \leq y \Rightarrow F_X(x) \leq F_X(y)$).
- F_X est continue sauf éventuellement en un nombre dénombrable de points isolés $(a_i)_{i=1,..}$ en lesquels $F_X(a_i) \neq F_X(a_i^-)$.
- Si F_X est discontinue en un point b alors on a $\mathbb{P}[X = b] = F_X(b) - F_X(b^-) > 0$.
- Si F_X est continue en un point a alors on a $F_X(a) = F_X(a^-)$ et $\mathbb{P}[X = a] = 0$.



Fonction de répartition d'une variable discrète

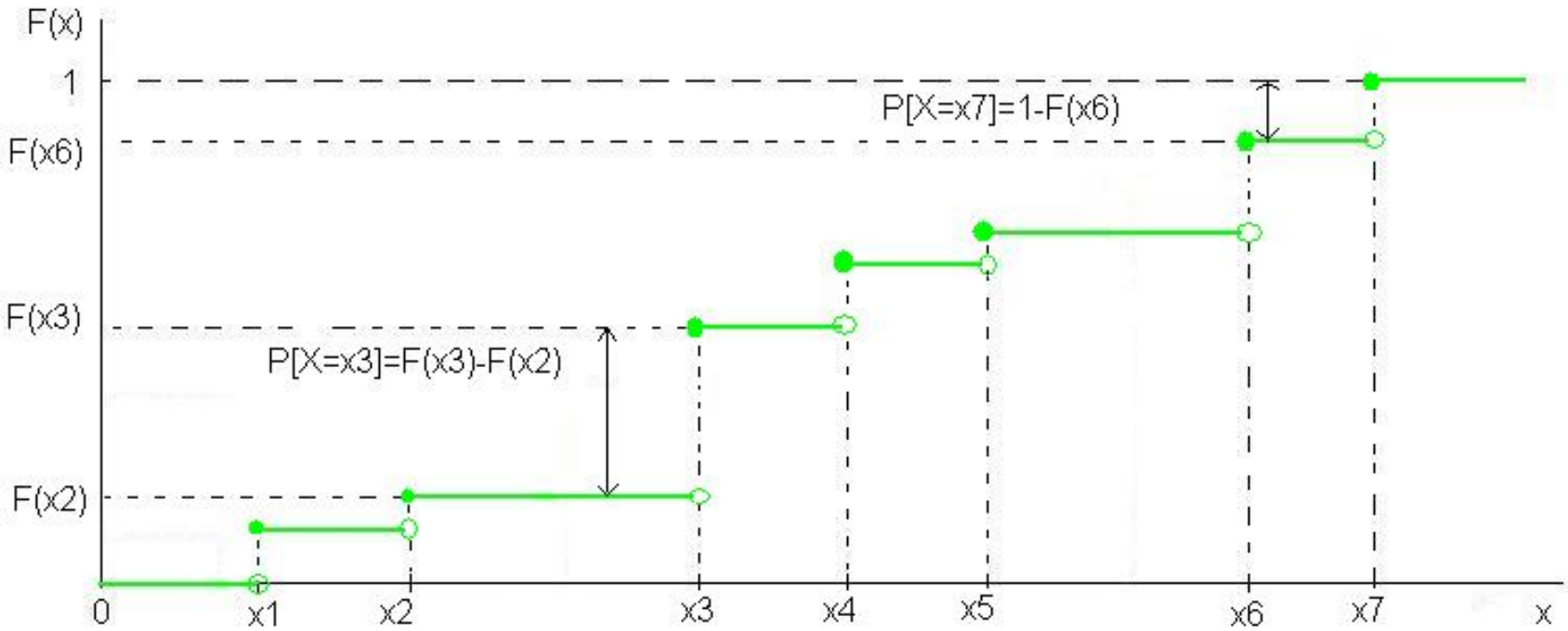
- La fonction de répartition d'une variable discrète au point x correspond à l'accumulation des probabilités des valeurs inférieures ou égales à x :

$$F_X(x) = \sum_{x_i \leq x, x_i \in D_X} \mathbb{P}[X = x_i]$$

- Ainsi, F_X est une fonction en escalier, continue à droite.
- Pour $a < b$, on a :

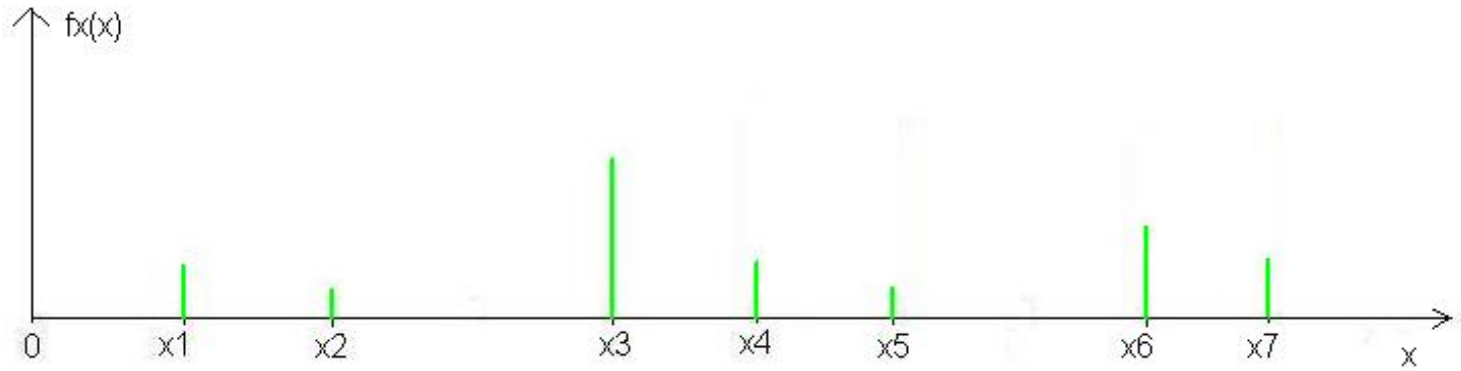
$$\begin{aligned}
F_X(b) - F_X(a) &= \sum_{a < x_i \leq b, x_i \in D_X} \mathbb{P}[X = x_i] \\
F_X(b) - F_X(a^-) &= \sum_{a \leq x_i \leq b, x_i \in D_X} \mathbb{P}[X = x_i] \\
F_X(b^-) - F_X(a) &= \sum_{a < x_i < b, x_i \in D_X} \mathbb{P}[X = x_i] \\
F_X(b^-) - F_X(a^-) &= \sum_{a \leq x_i < b, x_i \in D_X} \mathbb{P}[X = x_i]
\end{aligned}$$

Allure de la fonction de répartition d'une variable discrète



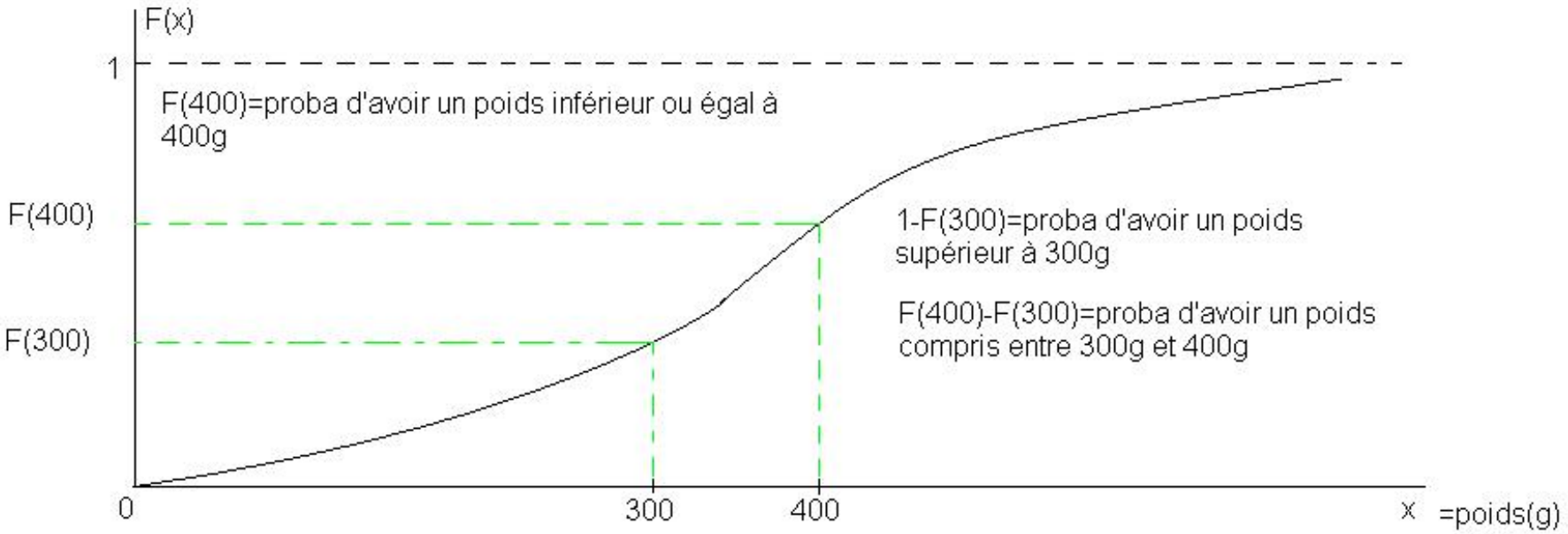
Loi de probabilité d'une variable discrète

- La loi de probabilité d'une variable aléatoire discrète est la donnée des nombres $p_i = \mathbb{P}[X = x_i]$ pour chacune des valeurs possibles x_i pour X. Ces nombres sont compris entre 0 et 1 et leur somme vaut 1.



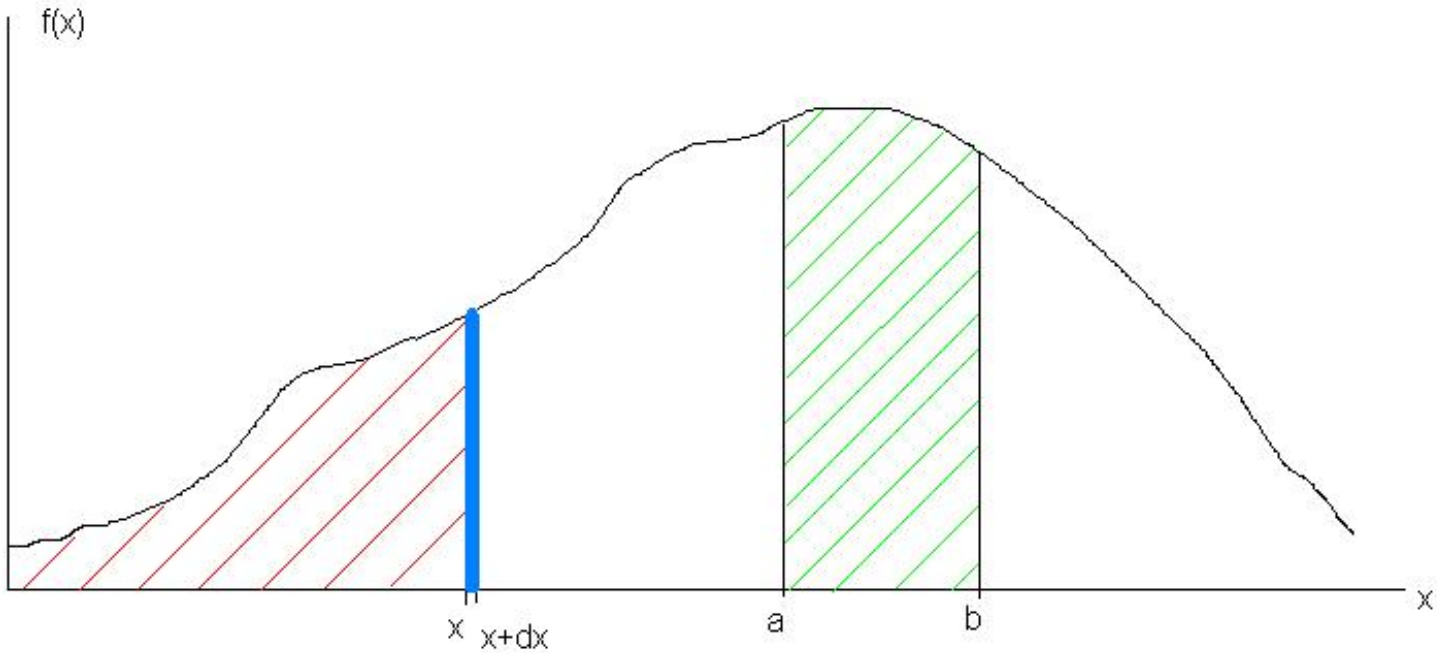
Fonction de répartition d'une variable continue

- La fonction de répartition d'une variable aléatoire continue est une fonction continue, dérivable presque-partout.
- En tout point x, on a $\mathbb{P}[X = x] = 0$ et $\mathbb{P}[X < x] = \mathbb{P}[X \leq x]$.
- Pour tout $a < b$, on a :
  $\mathbb{P}[a \leq X \leq b] = \mathbb{P}[a \leq X < b] = \mathbb{P}[a < X \leq b] = \mathbb{P}[a < X < b]$



Rappels sur le calcul des intégrales

- $f(x)dx$=surface bleue
- $\int_{-\infty}^{x} f(t)dt = F(x)$=surface en rouge : F=primitive de f
- $\int_a^b f(x)dx = F(b) - F(a)$=surface verte $(= \int_a^b f(t)dt)$
- $f(x) = \frac{dF(x)}{dx}$=dérivée de F en x= pente de F en x



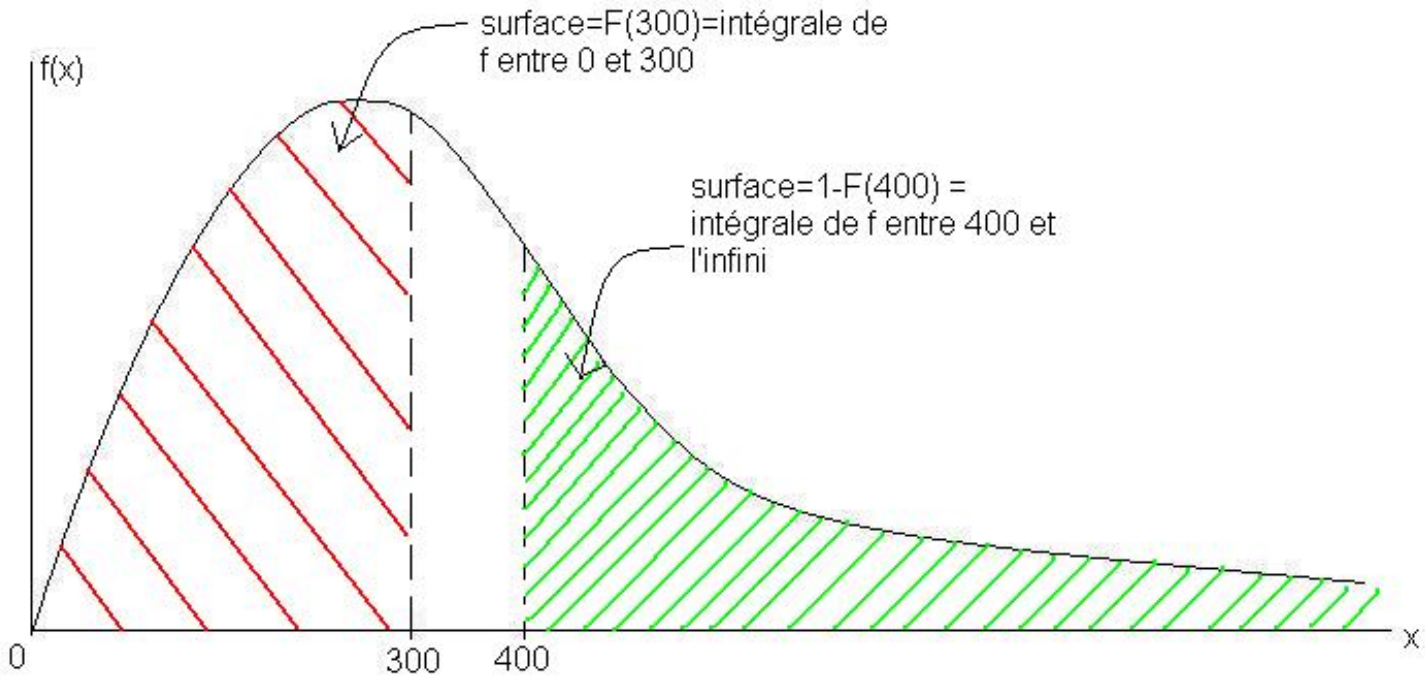
Densité d'une variable aléatoire continue

- La densité d'une variable aléatoire continue est donnée par :

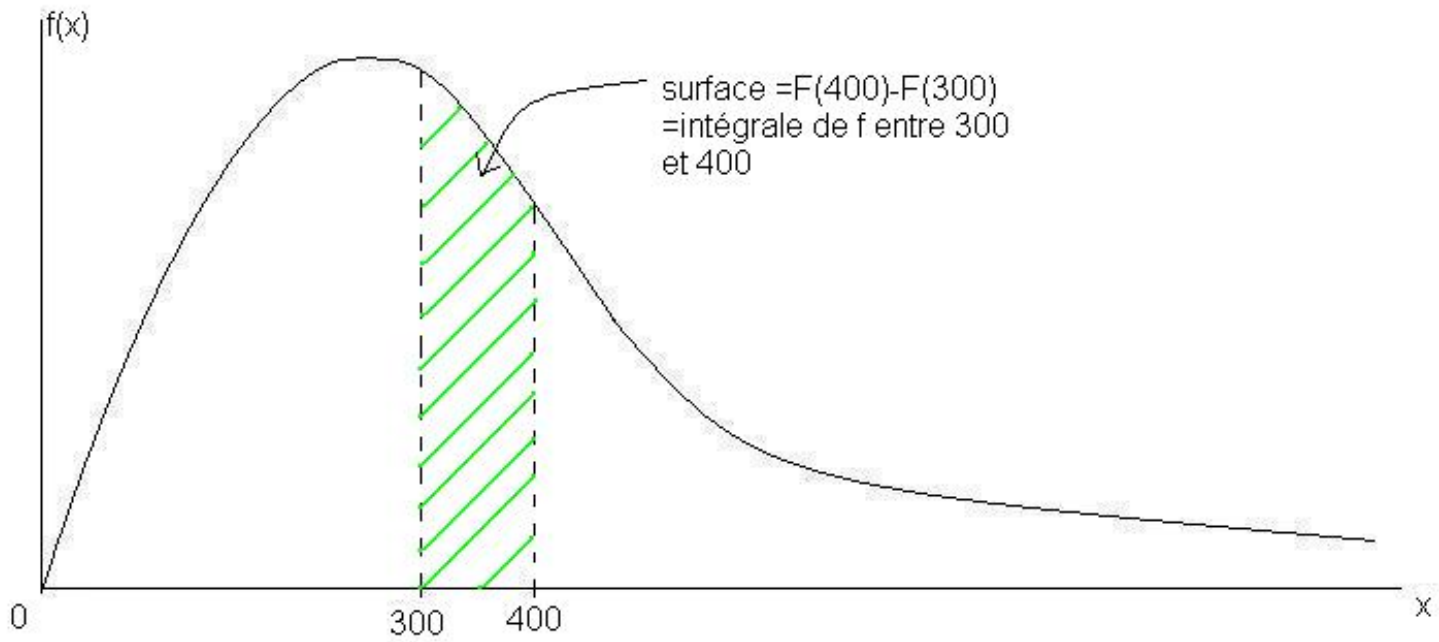
$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx} & \text{pour } x \in D_X \text{ tel que } F_X \text{ est dérivable} \\ 0 & \text{pour } x \notin D_X \text{ et lorsque } F_X \text{ n'est pas dérivable} \end{cases}$$

- $f_X(x)dx = \mathbb{P}[x \leq X \leq x + dx] \approx \mathbb{P}[X = x]$
- $F_X(x) = \int_{-\infty}^{x} f_X(u)du$
- f_X est positive (car $F_X \nearrow$) et $\int_{-\infty}^{\infty} f_X(u)du = 1$
- $\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a) = \int_a^b f_X(u)du$

Densité d'une variable aléatoire continue (suite)



Densité d'une variable aléatoire continue (suite 2)



Variables aléatoires indépendantes

- Deux variable X_1 et X_2 sont **indépendantes** lorsque le fait de connaître la valeur obtenue par X_1 n'apporte aucune information sur la valeur qui sera prise par X_2 et réciproquement.
- ex : X_1="poids d'une souris" et X_2="couleur du pelage" sont indépendantes alors que X_1 et Y_1="taille d'une souris" ne le sont vraisemblablement pas.
- Caractérisation de l'indépendance pour un couple de variables **discrètes** : X et Y sont indépendants lorsque pour tout couple de valeurs (x_i, y_j) pris par (X, Y), on a

$$\mathbb{P}[X = x_i, Y = y_j] = \mathbb{P}[X = x_i]\mathbb{P}[Y = y_j]$$

- Caractérisation de l'indépendance pour un couple de variables **continues** : X et Y sont indépendants lorsqu'on a

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{pour tout } (x, y)$$

où $f_{X,Y}$ représente la densité jointe du couple (X, Y), f_X représente la densité de X et f_Y représente la densité de Y.

Espérance mathématique

- Soit X une variable aléatoire. On note $\mathbb{E}[X]$ l'**espérance** de X. C'est un nombre qui représente la valeur moyenne prise par X.
- Si X est discrète, on calcule $\mathbb{E}[X]$ par la formule :

$$\mathbb{E}[X] = \sum_{x_i \in D_X} x_i \mathbb{P}[X = x_i].$$

- Si X est continue, on calcule $\mathbb{E}[X]$ par la formule :

$$\mathbb{E}[X] = \int x f_X(x) dx.$$

- On dit qu'une variable est **centrée** lorsque $\mathbb{E}[X] = 0$.
- On a toujours $\mathbb{E}[aX] = a\mathbb{E}[X]$, $\mathbb{E}[a + X] = a + \mathbb{E}[X]$ et $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.
- On calcule $\mathbb{E}[X^2]$ par la formule :
  - $\mathbb{E}[X^2] = \sum_{x_i \in D_X} x_i^2 \mathbb{P}[X = x_i]$ si X est discrète,
  - $\mathbb{E}[X^2] = \int x^2 f_X(x) dx$ si X est continue.

Variance mathématique, écart-type mathématique

- La **variance** d'une v.a. X est le nombre **positif** défini par :

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

  La variance d'une v.a. exprime à quel point les valeurs prises sont dispersées autour de la moyenne. Une grande variance indique une grande variabilité. A l'inverse, une variance nulle révèle que la variable aléatoire est en fait certaine.

- Si X est discrète, on calcule $\mathrm{Var}(X)$ par l'une des 2 formules :

$$\mathrm{Var}(X) = \sum (x_i - \mathbb{E}[X])^2 \mathbb{P}[X = x_i] = \sum x_i^2 \mathbb{P}[X = x_i] - \mathbb{E}[X]^2.$$

- Si X est continue, on calcule $\mathrm{Var}(X)$ par l'une des 2 formules :

$$\mathrm{Var}(X) = \int (x - \mathbb{E}[X])^2 f(x) dx = \int x^2 f(x) dx - \mathbb{E}[X]^2.$$

- On a toujours $\mathrm{Var}(X + a) = \mathrm{Var}(X)$ et $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$ mais la relation $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$ n'est vraie que lorsque X_1 et X_2 sont indépendantes ! ! !

- On dit qu'une v.a. est **réduite** lorsque $\mathrm{Var}(X) = 1$.

Covariance

- La **covariance** de X et de Y est le nombre défini par :

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- Le nombre $\mathrm{Cov}(X, Y)$ sert à mesurer le degré de **dépendance linéaire** entre X et Y.
- Lorsque $\mathrm{Cov}(X, Y) > 0$, cela signifie que lorsqu'une des v.a. a tendance à augmenter, l'autre aussi.
- Lorsque $\mathrm{Cov}(X, Y) < 0$, cela signifie que lorsqu'une des v.a. a tendance à augmenter, l'autre a tendance à diminuer.
- Lorsque $\mathrm{Cov}(X, Y) = 0$, on dit que X et Y sont non corrélés.
- Lorsque X et Y sont indépendantes, $\mathrm{Cov}(X, Y) = 0$ mais la réciproque est fausse ! ! ! Si $\mathrm{Cov}(X, Y) = 0$, X et Y ne sont pas forcément indépendantes, par ex : soit X de loi normale et soit $Y = X^2$, alors $\mathrm{Cov}(X, Y) = 0$ mais X et Y ne sont pas indépendantes.

Corrélation, skewness, kurtosis

- On appelle **coefficient de corrélation linéaire** de X et Y le réel (forcément compris entre -1 et 1) :

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

- $\rho_{XY} \pm 1 \Leftrightarrow Y$ est une fonction affine de X avec proba 1 i.e. $Y = aX + b$.
- Le **coefficient d'assymétrie** (skewness) ν_1 est nul lorsque la loi de X est symétrique :

$$\nu_1 = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\mathrm{Var}(X)^{3/2}}$$

- $\nu_1 = 0$ pour les distributions symétriques, $\nu_1 > 0$ lorsque les valeurs prises par X sont très étalées sur la droite, $\nu_1 < 0$ lorsque les valeurs prises par X sont très étalées sur la gauche.
- Le **coefficient d'aplatissement** (kurtosis) ν_2 est défini par :

$$\nu_2 = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathrm{Var}(X)^2}$$

Distributions usuelles discrètes

- Les familles de lois de proba usuelles discrètes sont
  - lois de Bernoulli
  - lois binômiales
  - lois de Poisson
  - lois uniformes discrètes
- Ces lois sont paramétrées. Cela signifie que la famille de lois donne la forme générale mais qu'à l'intérieur de ces familles chaque loi dépend de un ou plusieurs nombres appelés **paramètres**.

Lois de Bernoulli

- On dit qu'une variable X suit une loi de Bernoulli de paramètre p, notée $\mathcal{B}(p)$, lorsqu'elle prend les valeurs 0 avec probabilité $(1-p)$ et 1 avec probabilité p : $\mathbb{P}[X=1]=p$ et $\mathbb{P}[X=0]=1-p$.
- $\mathbb{E}[X]=p$ et $\text{var}(X)=p(1-p)$.
- Définir une variable de Bernoulli revient à coder par 1 la réalisation d'un évènement et par 0 sa non-réalisation, autrement dit, $X=1$ si l'évènement est réalisé et $X=0$ sinon.
  - La loi de Bernoulli peut être utilisée pour coder des caractéristiques qualitatives à 2 modalités, par ex : vivant/ décédé, masculin/féminin, traité/ non traité, par ex : $X=1$ si l'individu est vivant et $X=0$ sinon.
  - La loi de Bernoulli peut être utilisée pour coder des caractéristiques quantitatives à 2 modalités, par ex : $X=1$ si le poids d'un sujet est inférieur à un seuil et $X=0$ sinon.

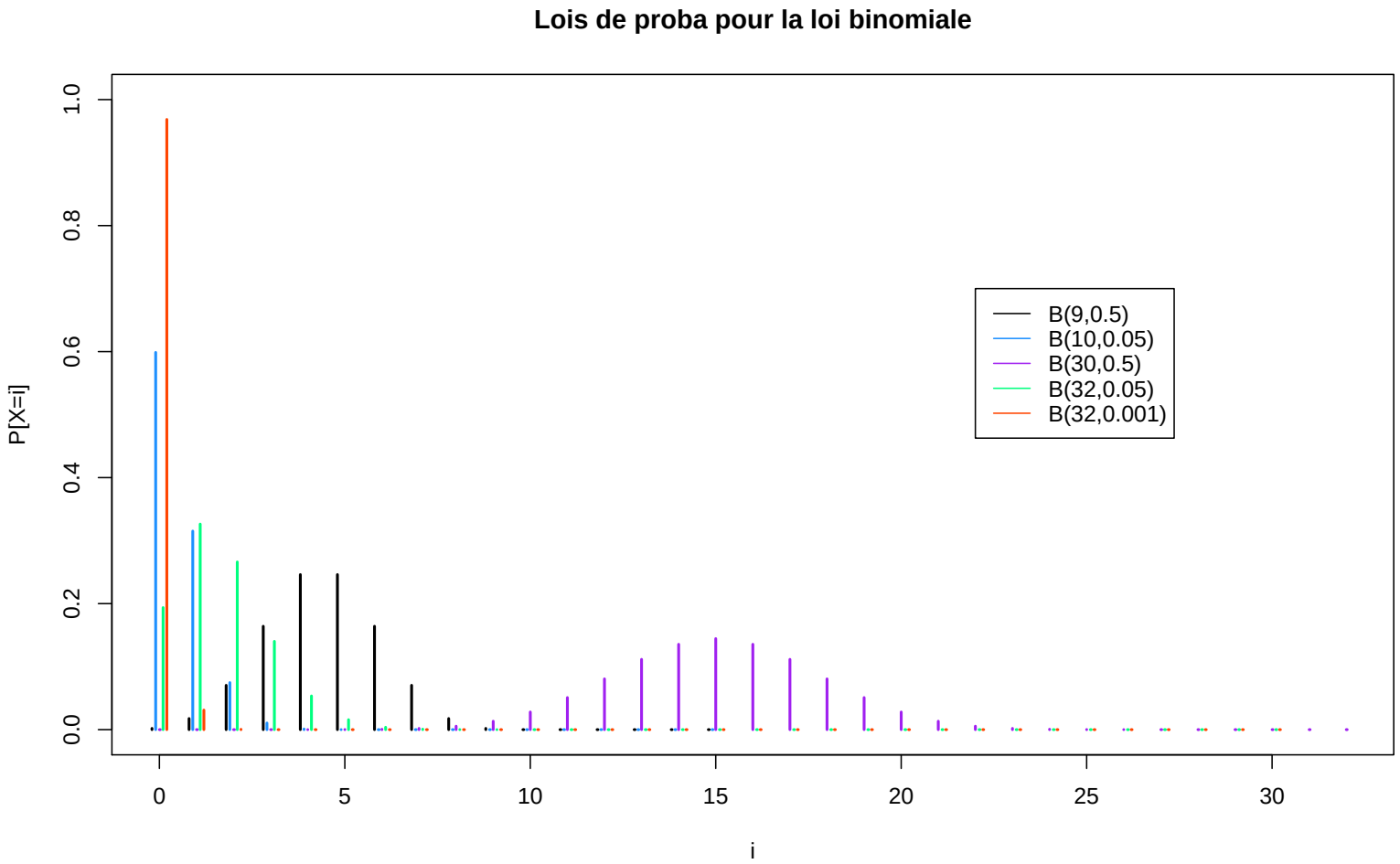
Lois binômiales

- On dit qu'une variable X suit une loi binômiale de paramètres n et p, notée $\mathcal{B}(n,p)$, lorsqu'elle prend ses valeurs parmi $\{0,...,n\}$ avec les probabilités suivantes pour $k \in \{0,...,n\}$:

$$\mathbb{P}[X=k] = C_n^k p^k (1-p)^{n-k}.$$

- $\mathbb{E}[X] = np$ et $\text{var}(X) = np(1-p)$.
- X représente le nombre de fois où un évènement se réalise en n répétitions indépendantes de l'expérience, par ex : X représente le nombre de prélèvements positifs à une infection parmi un lot de n prélèvements.
- Si $Y_1,...,Y_n$ est une suite de v.a. indépendantes de loi de Bernoulli de même paramètre p, $\mathcal{B}(p)$, alors $\sum_{i=1}^n Y_i$ suit une loi $\mathcal{B}(n,p)$.

Lois binômiales



Lois de Poisson

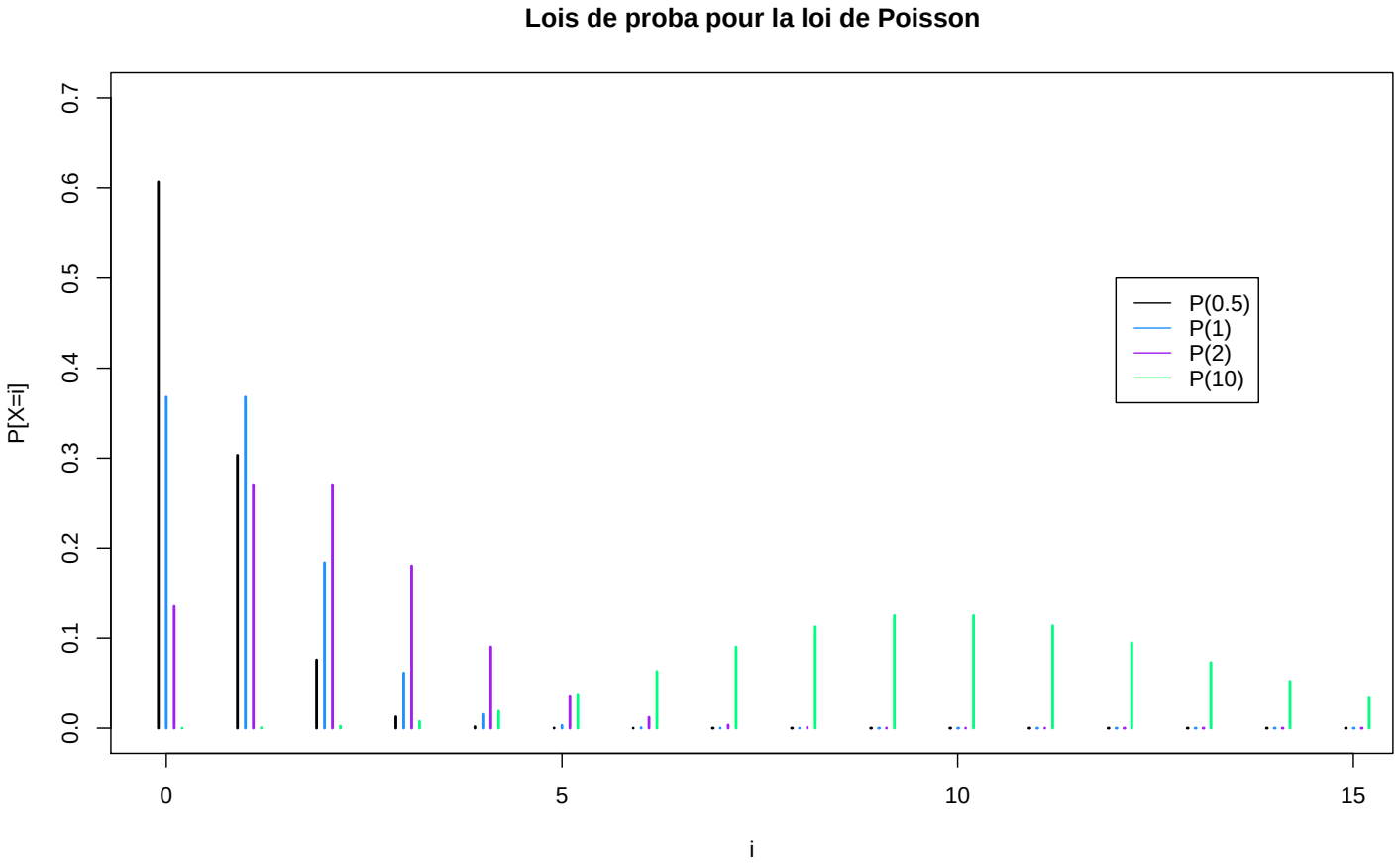
- On dit qu'une variable X suit une loi de Poisson de paramètre λ, $\mathcal{P}(\lambda)$, lorsqu'elle prend ses valeurs dans $\mathbb{N} = \{0, 1, ..., n, ...\}$ (infinité de valeurs possibles) avec les probabilités suivantes pour $k \in \mathbb{N}$:

$$\mathbb{P}[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}.$$

- $\mathbb{E}[X] = \lambda$ et $\text{var}(X) = \lambda$.
- X représente le nombre de fois où un évènement se produit au cours d'un e certaine durée d'observation. Par ex, X représente le nombre de fois où un évènement survient au fil du temps, par ex : X compte le nombre de colonies dans une boite de Pétri au bout d'un temps donné de culture (1 semaine), X compte le nombre d'évènements indésirables graves associés à une prise de médicaments au terme d'une durée d'observation de 1 an.

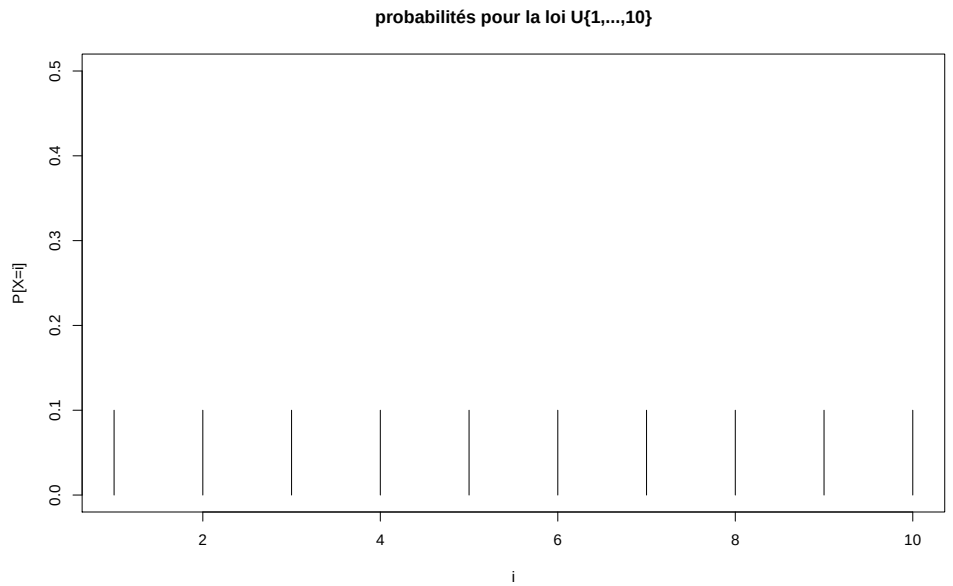
Loi de Poisson (suite)

- Si $Y_1, ..., Y_n$ est une suite de v.a. indépendantes de loi de Poisson où Y_i a pour paramètre λ_i, $\mathcal{P}(\lambda_i)$, alors $\sum_{i=1}^{n} Y_i$ suit une loi $\mathcal{P}(\sum_{i=1}^{n} \lambda_i)$.



Loi uniforme discrète

- Une variable X suit une loi uniforme discrète, $\mathcal{U}(\{1, ..., n\})$, lorsque X prend ses valeurs dans $\{1, ..., n\}$ avec les proba suivantes pour $k \in \{1, ..., n\}$: $\mathbb{P}[X = k] = \frac{1}{n}$.
- Lorsqu'on randomise le traitement des patients rentrant dans un essai clinique, on tire au sort le traitement que chacun reçoit. X représente le numéro de traitement parmi n traitements différents. Tous les traitements ont la même probabilité d'être attribué.



Distributions usuelles continues

- Les familles de lois de proba usuelles continues sont
  - lois normales
  - lois de Student
  - lois du chi-deux
  - lois uniformes continues
  - lois exponentielles
  - lois de Fisher
- Ces lois sont paramétrées. Cela signifie que la famille de lois donne la forme générale mais qu'à l'intérieur de ces familles chaque loi dépend de un ou plusieurs nombres appelés **paramètres**.

Loi normale

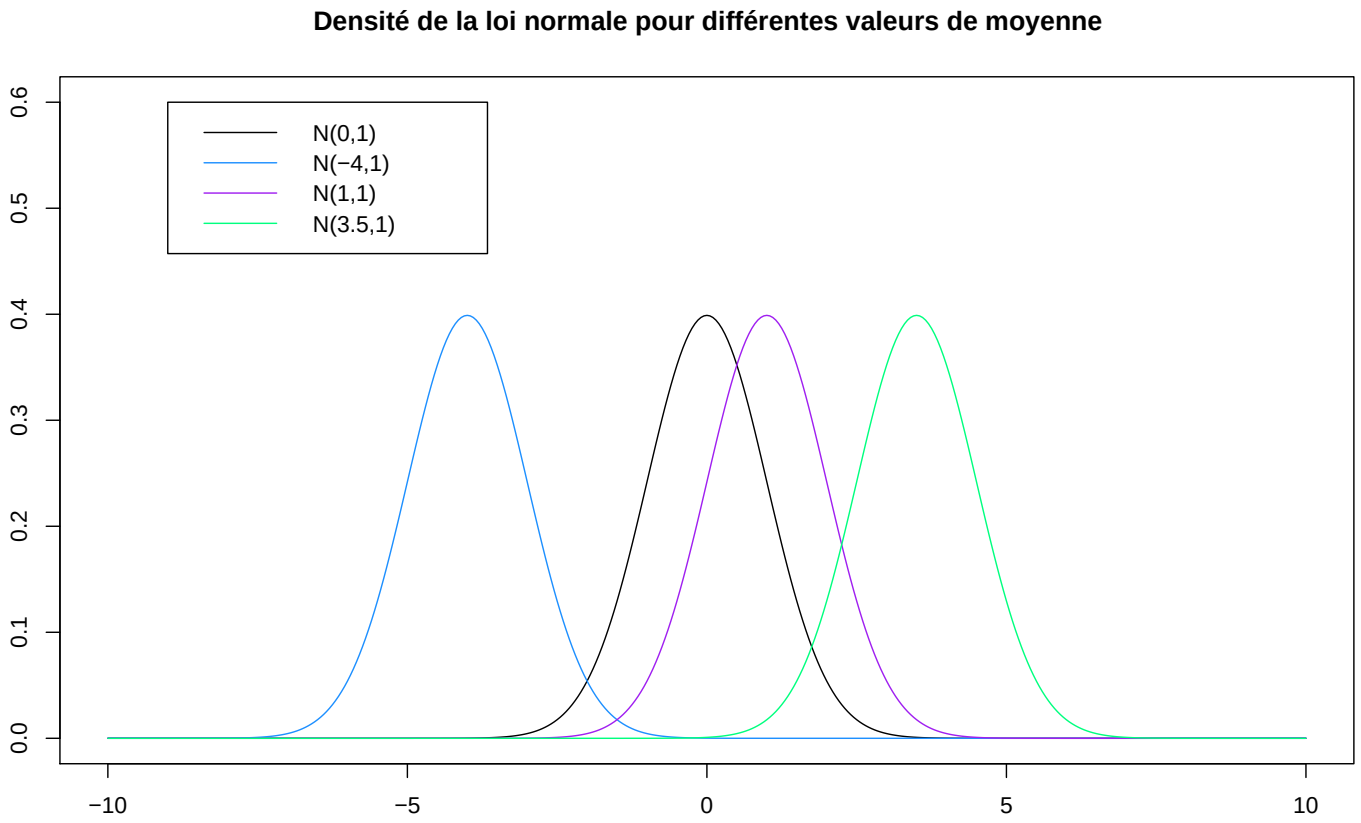
- On dit qu'une variable X suit une loi normale de paramètres m et σ^2, notée $\mathcal{N}(m, \sigma^2)$, lorsqu'elle prend ses valeurs dans $\mathbb{R}$ avec la densité suivante pour $x \in \mathbb{R}$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

- La densité est symétrique par rapport à la droite verticale d'absisse $x = m$.
- Une variable X de loi normale $\mathcal{N}(m, \sigma^2)$ représente une variable qui oscille de façon symétrique autour de sa moyenne.
- $\mathbb{E}[X] = m =$ médiane $=$ mode et $\text{var}(X) = \sigma^2$.
- Si $X_1, ..., X_n$ est une suite de variables indépendantes de loi normales telle que $X_i \simeq \mathcal{N}(m_i, \sigma_i^2)$, alors $\sum_{i=1}^{n} X_i$ suit une loi $\mathcal{N}(m_1 + ...m_n, \sigma_1^2 + ... + \sigma_n^2)$.

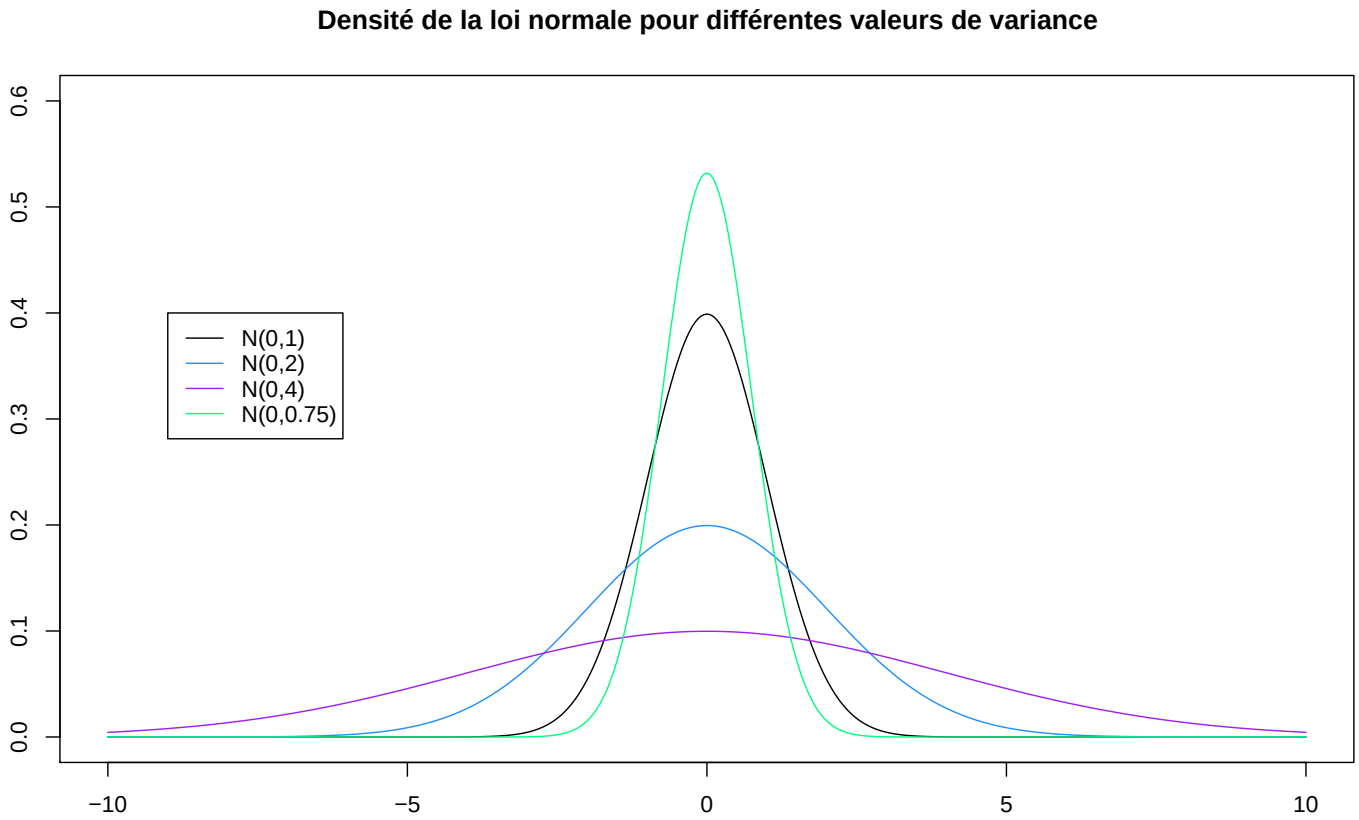
Loi normale : influence de la moyenne

L'allure de la courbe se conserve si on change de moyenne. Il s'agit d'un simple décalage.



Loi normale : influence de la variance

La coube s'aplatit lorsque la variance augmente, elle se resserre si la variance diminue, le maximum s'ajuste pour que la surface vaille 1, le maximum peut dépasser 1.



Loi normale et transformation

- D'une manière générale, si X suit une loi $\mathcal{N}(m, \sigma^2)$, alors $aX + b$ suit une loi $\mathcal{N}(am + b, a^2\sigma^2)$.
- On ne peut déterminer la fonction de répartition de X de loi $\mathcal{N}(m, \sigma^2)$ que par approximations numériques (par ordinateur).
- Un cas important est la loi $\mathcal{N}(0, 1)$ appellée **loi gaussienne standard** ou **loi normale centrée réduite** qui est tabulée.
- Pour se ramener à la loi $\mathcal{N}(0, 1)$ à partir d'une variable X de loi $\mathcal{N}(m, \sigma^2)$, on utilise la variable $Y = \frac{X-m}{\sigma}$: Y suit une loi $\mathcal{N}(0, 1)$ et s'appelle la variable centrée réduite associée à X.

Loi normale et tabulation

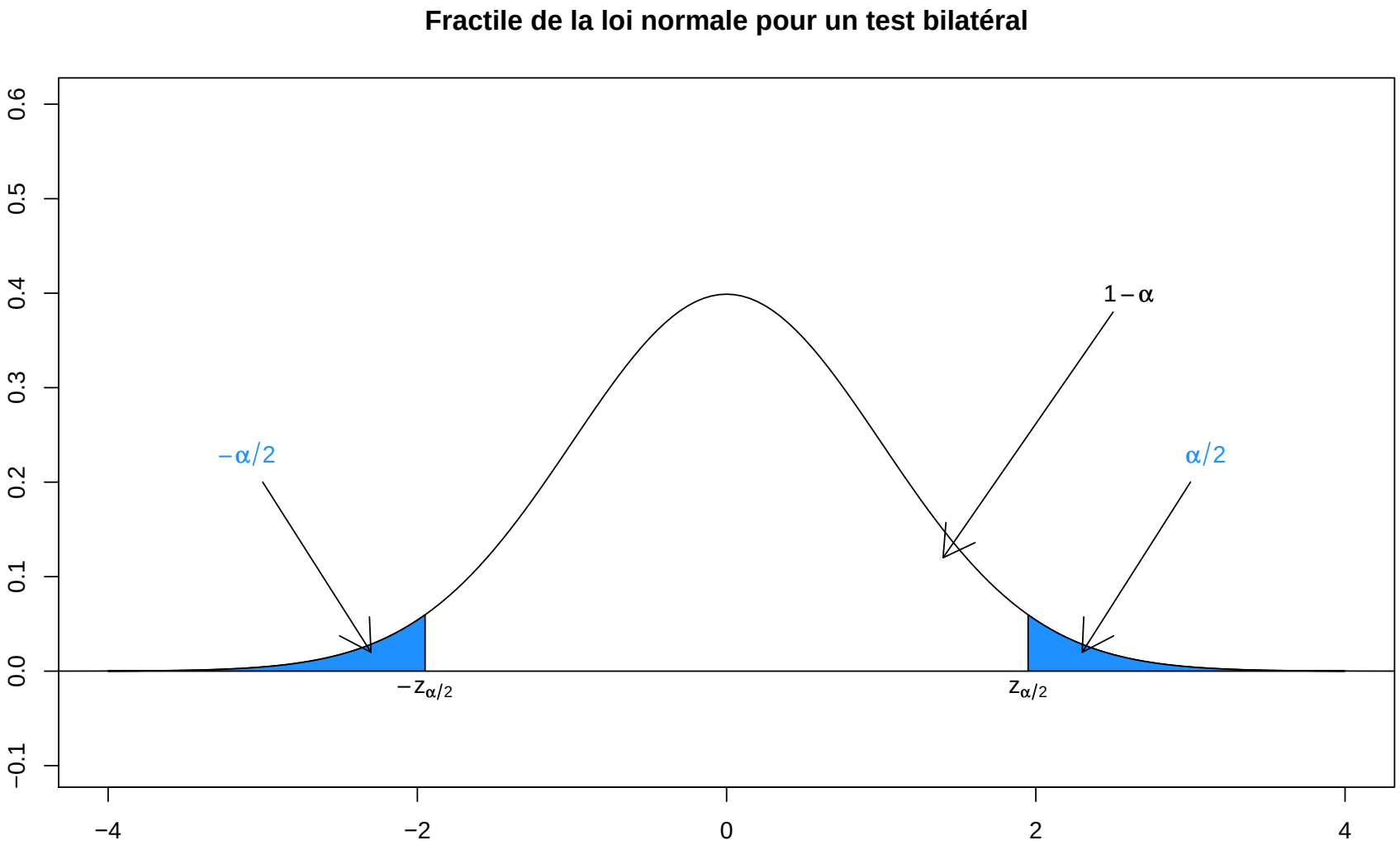
- On cherche la valeur de $\mathbb{P}[a \leq X \leq b]$ pour X de loi $\mathcal{N}(m, \sigma^2)$.
- Notons Φ la fonction de répartition associée à la variable Y de loi $\mathcal{N}(0, 1)$.
- Or, seule Φ est tabulée et en plus seulement pour $x > 0$.
- Pour $x < 0$, on utilise la formule $\Phi(x) + \Phi(-x) = 1$.
- Pour $x > 0$, on dispose aussi de la relation $\mathbb{P}[-x \leq Y \leq x] = 2\Phi(x) - 1$.
- Pour la variable X de loi $\mathcal{N}(m, \sigma^2)$, on utilise $Y = (X - m)/\sigma$,la variable centrée réduite associée à X et on obtient :

$$\begin{aligned} \mathbb{P}[a \leq X \leq b] &= \mathbb{P}\left[\frac{a-m}{\sigma} \leq \frac{X-m}{\sigma} \leq \frac{b-m}{\sigma}\right] \\ &= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right). \end{aligned}$$

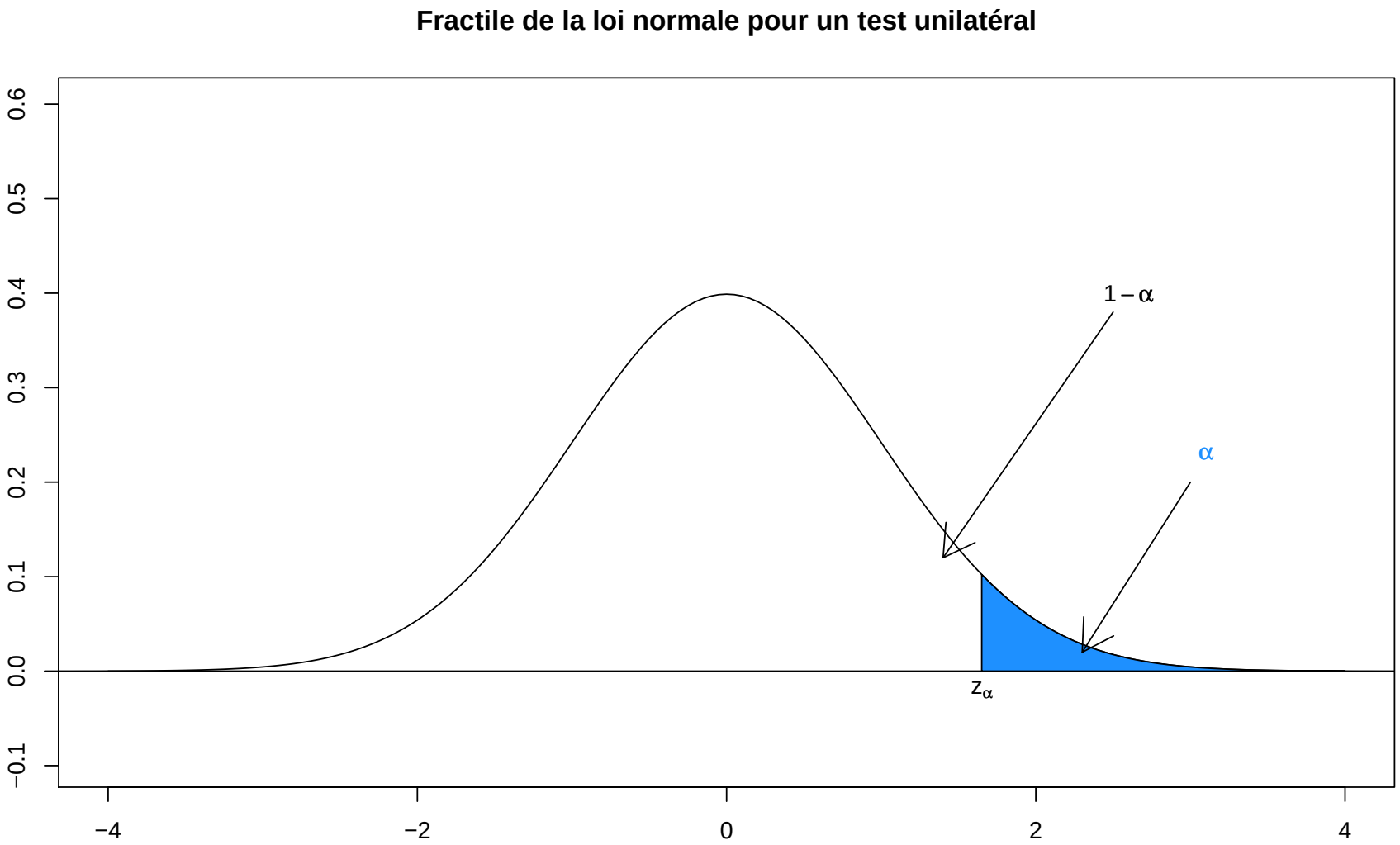
Lecture de la table $\mathcal{N}(0,1)$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

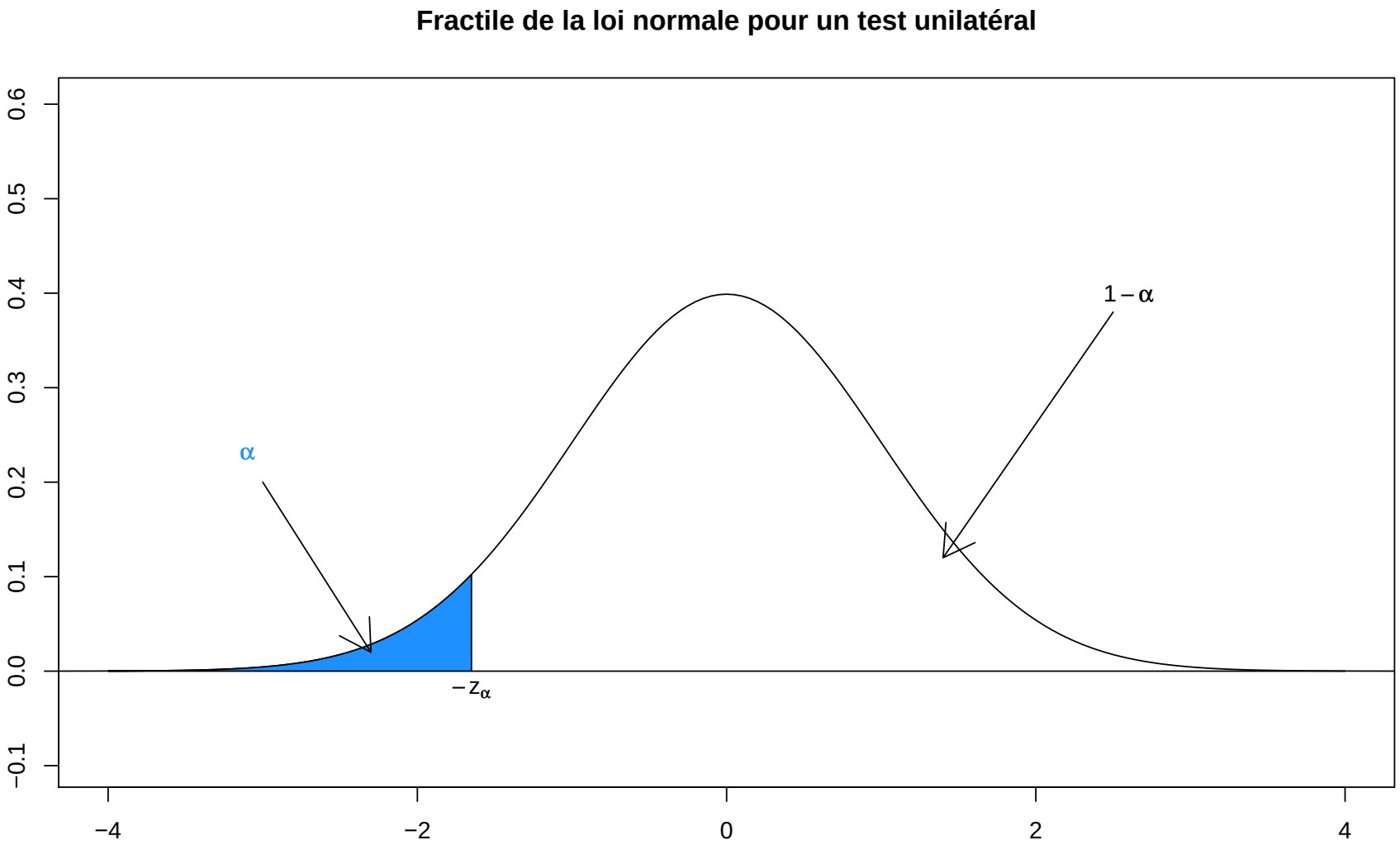
Loi normale : fractile bilatéral



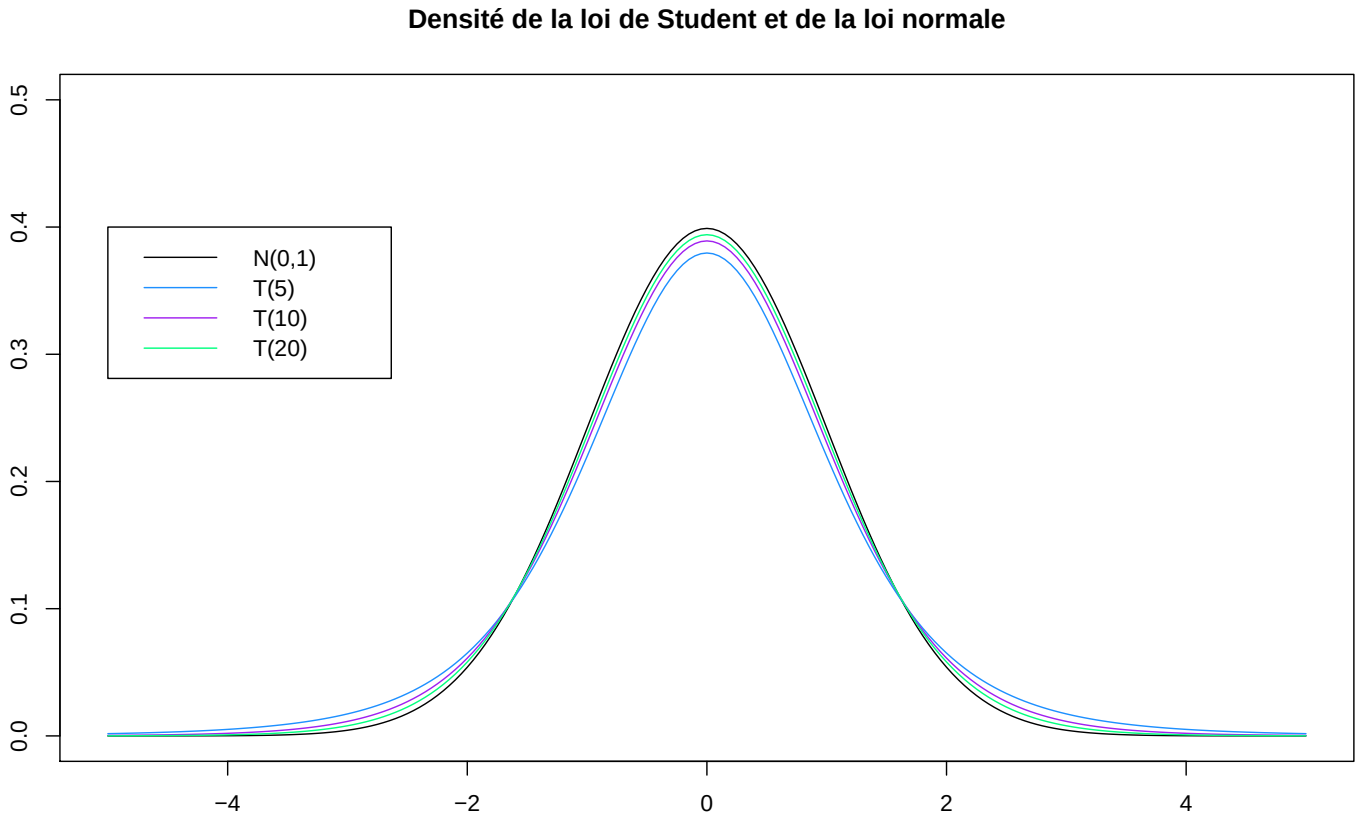
Loi normale : fractile unilatéral supérieur



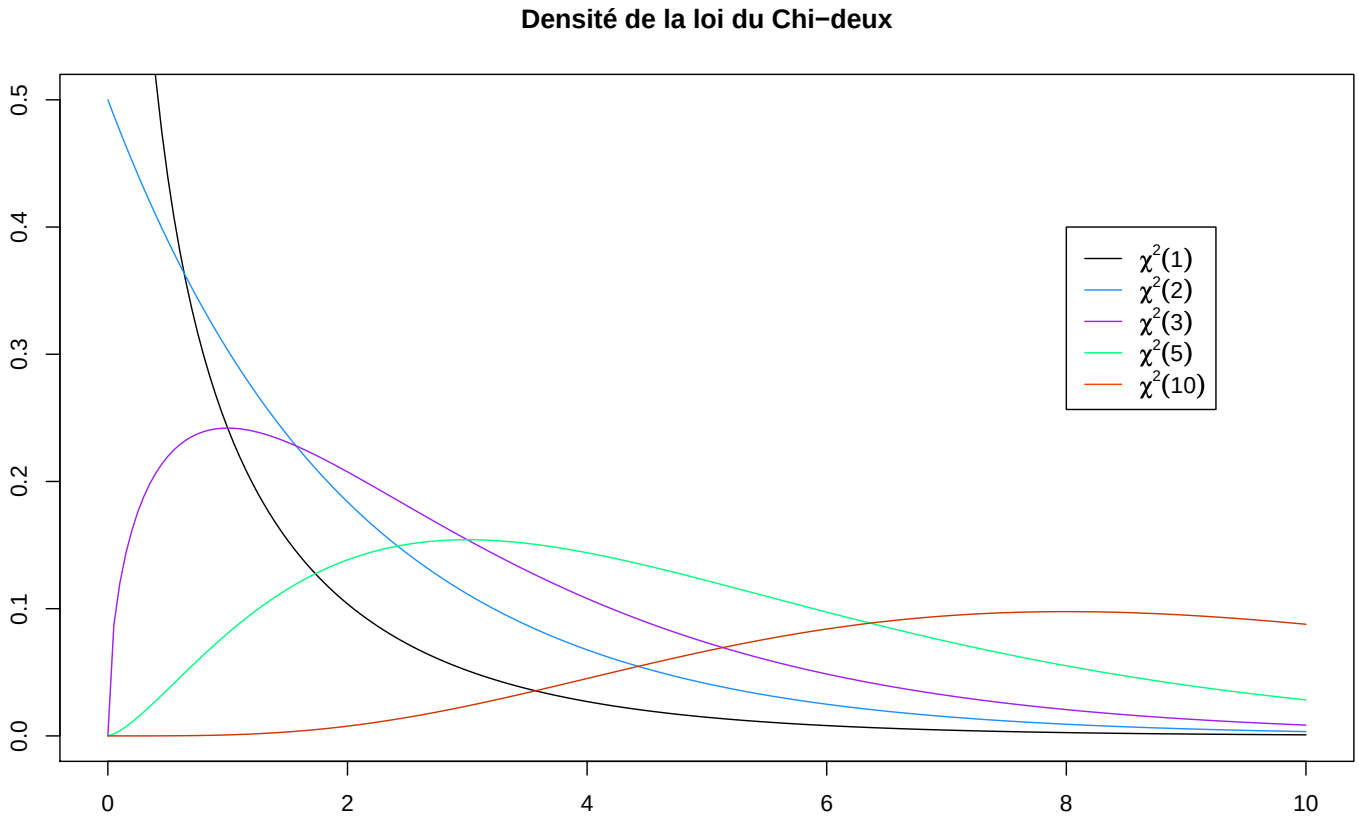
Loi normale : fractile unilatéral inférieur



Loi de Student



Loi du chi-deux



Loi uniforme

- On dit qu'une variable X suit une loi uniforme sur un intervalle $[a, b]$, notée $\mathcal{U}(a, b)$, lorsqu'elle prend ses valeurs dans $[a, b]$ avec la densité et la f.r. suivantes pour $x \in \mathbb{R}$:

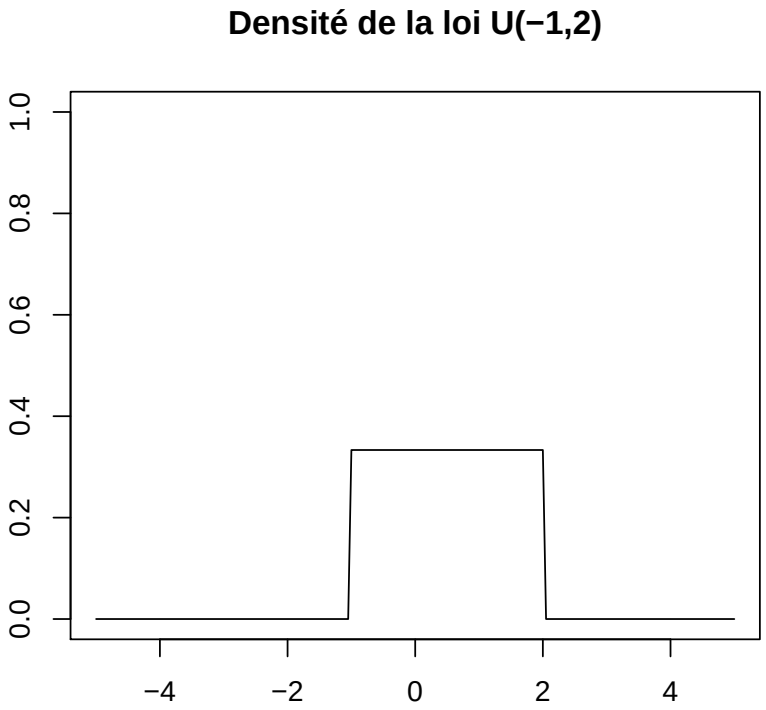
$$f(x) = \begin{cases} 1/(b-a) & \text{si } x \in [a, b] \\ 0 & \text{sinon} \end{cases}$$

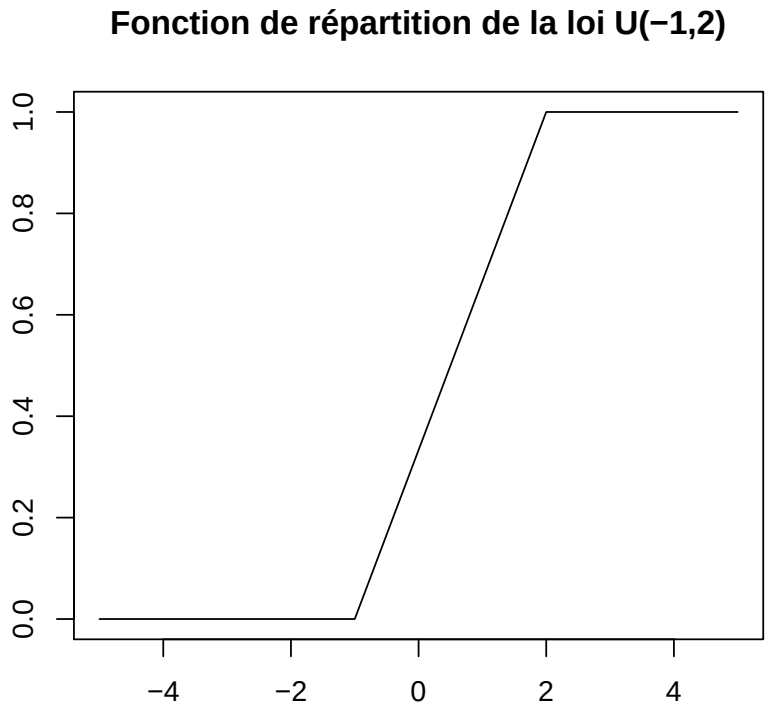
$$F(x) = \begin{cases} 0 & \text{si } x < a \\ (x-a)/(b-a) & \text{si } x \in [a, b] \\ 1 & \text{si } x > b \end{cases}$$

- $\mathbb{E}[X] = \frac{a+b}{2}$ et $\text{var}(X) = \frac{(b-a)^2}{12}$.

Loi uniforme

- Les valeurs de l'intervalle $[a, b]$ ont toutes la même proba d'être prises.





Lois exponentielles

- On dit qu'une variable X suit une loi exponentielle de paramètre λ, notée $\mathcal{E}(\lambda)$, lorsqu'elle prend ses valeurs dans $\mathbb{R}^+ = [0, +\infty[$ avec la densité et la f.r. suivantes pour $x \in \mathbb{R}$:

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

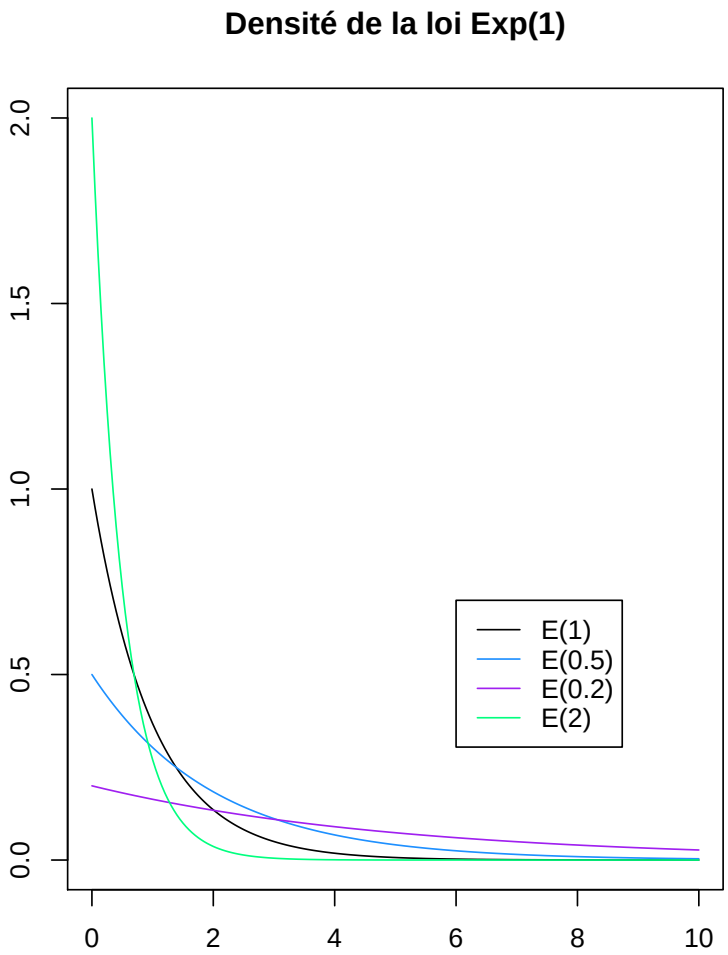
$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

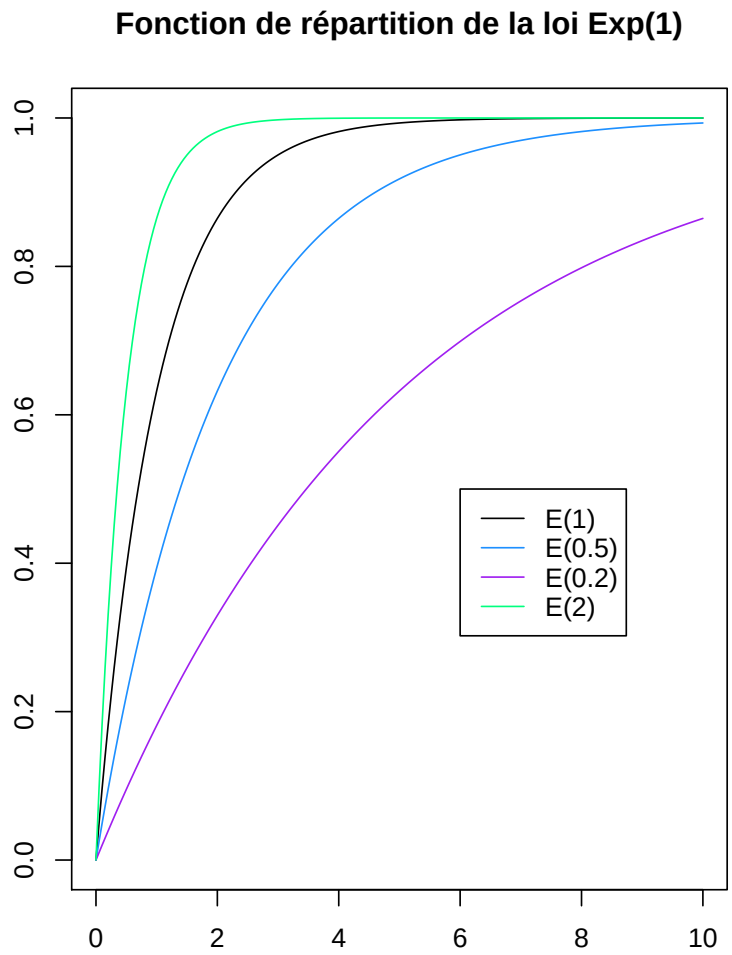
- $\mathbb{E}[X] = \frac{1}{\lambda}$ et $\text{var}(X) = \frac{1}{\lambda^2}$.
- La loi exponentielle est une loi sans mémoire, sans vieillissement :

$$\mathbb{P}[T > t + s | T > s] = \mathbb{P}[T > t].$$

Elle ne convient donc pas pour modéliser des durées de vie humaines considérées sur de grands intervalles de temps.

Lois exponentielles (suite)





Approximations de lois

- Pour $n = 5, 10, 15, 20, 30, 40, 50$ et $p = 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5$, la loi binômiale est tabulée.
- Pour n et p satisfaisant $np > 5$ et $n(1-p) > 5$, on approxime la loi binômiale $\mathcal{B}(n, p)$ par la loi normale $\mathcal{N}(np, np(1-p))$.
- Pour n grand ($n \geq 30$) et p petit ($p < 0.1$), on approxime la loi binômiale $\mathcal{B}(n, p)$ par la loi de Poisson $\mathcal{P}(np)$.
- Pour n grand ($n \geq 30$) et p petit ($p < 0.1$) et $np \geq 10$, on approxime la loi binômiale $\mathcal{B}(n, p)$ par la loi normale $\mathcal{N}(np, np)$.
- Pour λ grand ($\lambda \geq 10$), on approxime la loi de Poisson $\mathcal{P}(\lambda)$ par la loi normale $\mathcal{N}(\lambda, \lambda)$.

Variables aléatoires indépendantes

- Deux v.a. X_1 et X_2 sont **indépendantes** lorsque le fait de connaître la valeur obtenue par X_1 n'apporte aucune information sur la valeur qui sera prise par X_2 et réciproquement.
- ex : X_1="poids d'une souris" et X_2="couleur du pelage" sont indépendantes alors que X_1 et Y_1="taille d'une souris" ne le sont vraisemblablement pas.
- Caractérisation de l'indépendance pour un couple de variables **discrètes** : X et Y sont indépendants $\Leftrightarrow$ pour tout couple de valeurs (x_i, y_j) pris par (X, Y), on a

$$\mathbb{P}[X = x_i, Y = y_j] = \mathbb{P}[X = x_i]\mathbb{P}[Y = y_j]$$

- Caractérisation de l'indépendance pour un couple de variables **continues** : X et Y sont indépendants $\Leftrightarrow$ on a

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{pour tout } (x, y)$$

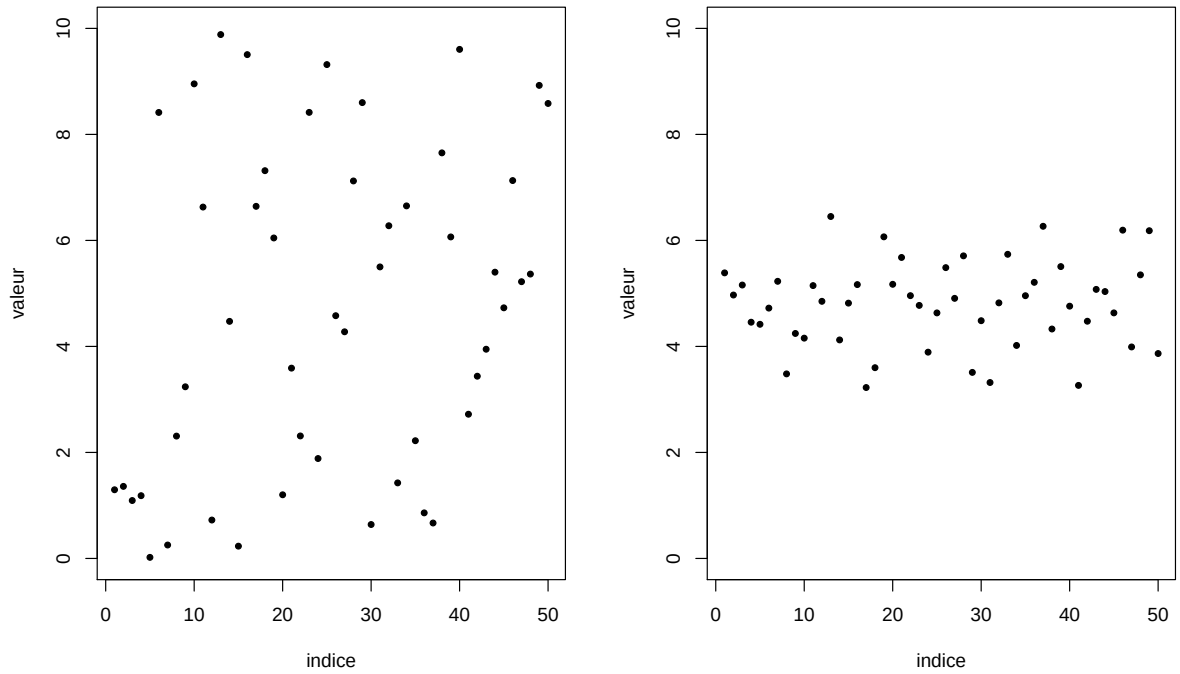
où $f_{X,Y}$ représente la densité jointe du couple (X, Y), f_X représente la densité de X et f_Y représente la densité de Y.

$2^{\text{ème}}$ partie

Notions de statistique descriptive

Regarder ses données ! ! !

- Considérons un jeu de données $(x_1, ..., x_n)$. Première chose à faire : tracer le nuage de points. Y a-t-il des points d'accumulation ou non ? Y-a-t-il des tendances ou non ? Y-a-t-il des valeurs extrêmes ?
- ex : les deux nuages de points ci-dessous comportent $n = 50$ observations.



Caractéristiques de position

- **moyenne empirique** (=moyenne de l'échantillon) :

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

- Cas des données groupées : lorsque les données sont présentées sous la forme (x_i, n_i) pour $i = 1, ..., k$ où x_i représente la valeur obtenue avec un effectif n_i, on calcule $\overline{X}$ au moyen de la formule :

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}$$

- Si les données sont fournies en classes $[a_i, a_{i+1}[$ pour $i = 1, ...k$, on approxime $\overline{x}$ par $\frac{\sum_{i=1}^{k} n_i c_i}{\sum_{i=1}^{k} n_i}$ où n_i=effectif de la classe n°i et c_i=centre de la classe n°i : $c_i = (a_i + a_{i+1})/2$.
- La moyenne empirique est très sensible aux valeurs extrêmes (peu robuste).

Caractéristiques de position (suite)

- **médiane empirique** (=médiane de l'échantillon) : m=valeur qui partage l'effectif total rangé par ordre croissant en deux classes de même taille. Notons $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ l'échantillon ordonné. La médiane empirique est donnée par

$$q_2 = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ est impair} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{si } n \text{ est pair} \end{cases}$$

  La médiane empirique est un indicateur robuste.
- Pour une distribution parfaitement symétrique, on a : moyenne=médiane. C'est utile en particulier pour vérifier rapidement la plausibilité d'une hypothèse de normalité des données : pour la loi $\mathcal{N}(0,1)$, moyenne=médiane.

Caractéristiques de dispersion

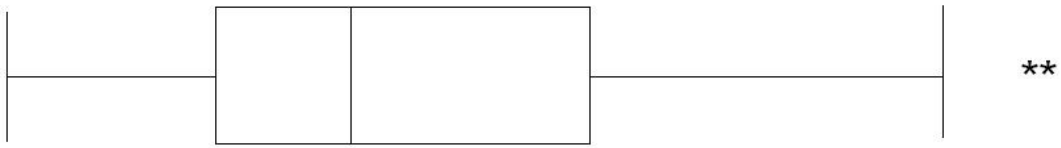
- **variance empirique** :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

  La variance d'un jeu de données exprime à quel point les valeurs sont dispersées autour de la valeur moyenne. Plus la variance est grande, plus les données sont dispersées.
- **écart-type empirique** : $s = \sqrt{s^2}$.
- **étendue** : $r = x_{(n)} - x_{(1)}$=value maximale-valeur minimale. C'est un indicateur instable car il ne dépend que des valeurs extrêmes.
- **intervalle interquartile** : les quartiles q_1, q_2, q_3 sont les 3 valeurs partageant l'effectif total ordonné en 4 parties égales (q_2=médiane). L'intervalle interquartile est plus robuste que l'étendue.

Caractéristiques de dispersion (suite)

- **boxplot (boîte à moustaches)** : ce diagramme représente schématiquement les principales caractéristiques d'un jeu de données en utilisant les quartiles. La partie centrale de la distribution est représentée par une boîte dont la longueur correspond à l'intervalle interquartile. On trace à l'intérieur la position de la médiane. On complète par les moustaches correspondant aux valeurs adjacentes :
  - adjacente supérieure : plus grand x_i inférieur à $q_3 + 1.5(q_3 - q_1)$
  - adjacente inférieure : plus petit x_i supérieur à $q_1 - 1.5(q_3 - q_1)$
- Les valeurs extérieures représentées par des étoiles sont celles qui sortent des moustaches.



Caractéristiques de forme

- Le **coefficient de symétrie empirique** (skewness) est défini par :

$$\widehat{\nu}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(s^2)^{3/2}}.$$

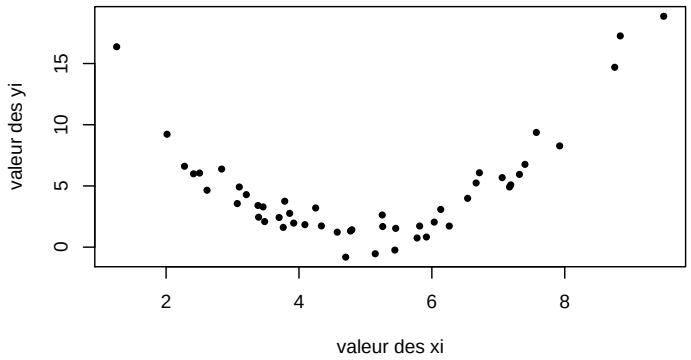
- Le **coefficient d'aplatissement empirique** (kurtosis) est défini par :

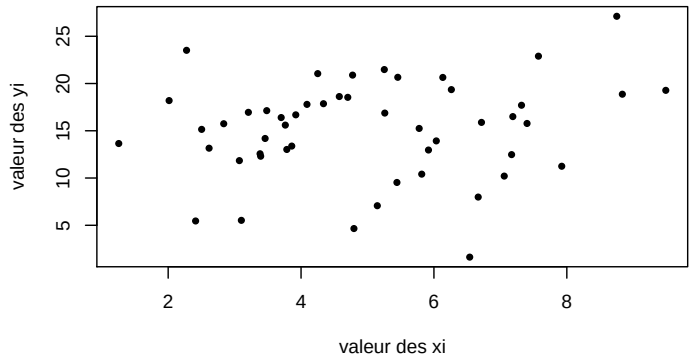
$$\widehat{\nu}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{(s^2)^{2}}.$$

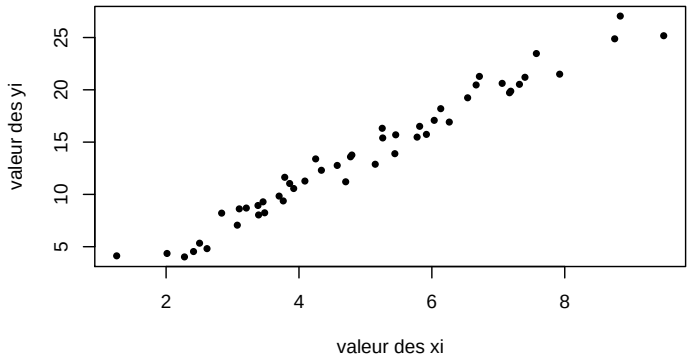
- Utile en particulier pour vérifier rapidement la plausibilité d'une hypothèse de normalité des données : pour la loi $\mathcal{N}(0,1)$, $\nu_1 = 0$ et $\nu_2 = 3$.

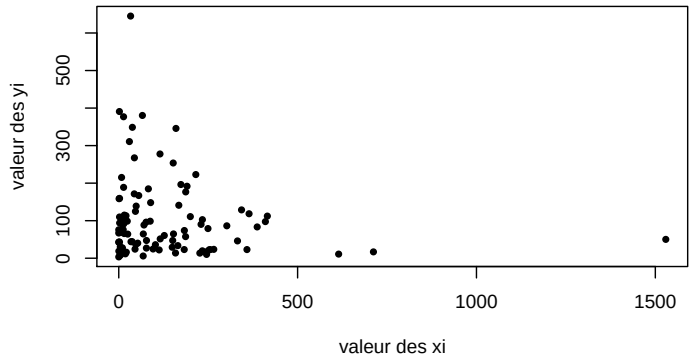
Couple de variables

Regarder ses données ! ! ! Tracer le nuage de points : y a-t-il liaison linéaire, non-linéaire, pas de liaison ? ? ? ?









$3^{\text{ème}}$ partie

Notions de statistique inférentielle

Notion d'échantillon

- On appelle **échantillon** de taille n d'une variable X une succession de n variables aléatoires $(X_1, ..., X_n)$ indépendantes et toutes de même loi.
- Cela correspond aux conditions suivantes :
  1. tous les individus sont sélectionnés dans la même population et sont donc identiques à quelques variations près
  2. les individus sont sélectionnés de manière indépendante
- Si on note P la loi de probabilité commune des X_i, parfois on dit que P est la loi parente de l'échantillon. Parfois on introduit une variable X de même loi que les X_i et on dit que X est la **variable parente** de l'échantillon.
- Après expérience, on recueille un jeu de données constitué des observations $(x_1, ..., x_n)$. C'est une réalisation de l'échantillon aléatoire $(X_1, ..., X_n)$: x_1 est la **réalisation** de la variable aléatoire X_1=1$^{\text{ère}}$ valeur obtenue en tirant au sort n sujets, etc...

Protocole d'une expérience

- Quel protocole expérimental permet obtenir une réalisation d'un échantillon $(X_1, ..., X_n)$? On tire au sort parmi la population n individus indépendants et représentatifs : c'est le mode de constitution de l'échantillon qui importe.
- Protocole d'une expérience : description complète de l'ensemble des matériels et méthodes employés dans l'expérience.
- Cette description doit permettre à un expérimentateur indépendant de reproduire cette expérience : pas de démarche scientifique sans cette **reproductibilité** !
- C'est bien les conditions de l'expérience (et pas forcément les résultats) qui doivent être reproduites.
- Protocole d'un traitement : description complète de l'ensemble des procédures utilisées dans le traitement d'un patient pour une pathologie donnée.

Notion de statistique, d'estimateur

Soit $(X_1, ..., X_n)$ un échantillon aléatoire.

- Une **statistique** est une variable aléatoire qui est une fonction de l'échantillon $(X_1, ..., X_n)$ soit $\widehat{\theta} = \varphi(X_1, ..., X_n)$, par exemple : $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Soit θ un paramètre inconnu à estimer. Pour estimer un paramètre d'une distribution, on utilise forcément l'échantillon ! L'**estimateur** du paramètre est une variable aléatoire qui est une fonction $\widehat{\theta}$ des variables de l'échantillon, soit $\widehat{\theta} = \varphi(X_1, ..., X_n)$, et qui doit "approcher" $\theta \Rightarrow$ est-ce que $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approche $\theta = \mathbb{E}[X]$?
- La réalisation $\widehat{\theta} = \varphi(x_1, ..., x_n)$ s'appelle une **estimation** de θ et fournit une approximation de la valeur de θ.
- A chaque fois qu'on réalise une expérience, on obtient une nouvelle réalisation de l'échantillon et donc vraisemblablement une nouvelle valeur de $\widehat{\theta}$. Ainsi $\widehat{\theta}$ est une variable aléatoire et donc suit une loi de probabilité, a une moyenne, une variance...

Qualités d'un estimateur

- La $1^{\text{ère}}$ qualité d'un estimateur est d'être **convergent** : quand la taille de l'échantillon n augmente, $\widehat{\theta}$ doit avoir tendance à se rapprocher de θ puisque la quantité d'information augmente.
- La $2^{\text{ème}}$ qualité d'un estimateur est d'être précis. La précision d'un estimateur peut se mesurer au moyen du biais et de la variance.
  - Le **biais** est l'écart moyen entre $\widehat{\theta}$ et θ i.e. biais $= \mathbb{E}[\widehat{\theta}] - \theta$. Un estimateur est **sans biais** lorsque $\mathbb{E}[\widehat{\theta}] = \theta$ i.e. si en utilisant $\widehat{\theta}$ un grand nombre de fois, il donne en moyenne la valeur du paramètre recherché. A l'inverse, un estimateur est biaisé si en l'utilisant un grand nombre de fois, il ne donne pas en moyenne la valeur du paramètre recherché. On souhaite qu'un estimateur soit non biasé !
  - On préfère qu'un estimateur sans biais ait une **variance minimale** de sorte que ses réalisations oscillent autour de θ sans jamais s'en éloigner trop.

L'estimateur "moyenne empirique"

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X. On veut estimer le paramètre $\theta = \mathbb{E}[X]$.
- Un estimateur de $\mathbb{E}[X]$ est $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. C'est une variable aléatoire donc $\overline{X}$ une moyenne, une variance, une loi de probabilité...
- Cet estimateur est convergent : $\overline{X} \longrightarrow \theta$ quand $n \to \infty$.
- Cet estimateur est sans biais : $\mathbb{E}[\overline{X}] = \theta$.
- Sa variance satisfait : $\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n}$ donc la variance de $\overline{X}$ décroît vers 0 quand n tend vers $+\infty$.
- Quelque soit la loi de X, la loi de $\overline{X}$ converge toujours vers la loi normale. En pratique, pour n assez grand ($n \geq 30$), la loi de $\overline{X}$ peut être approximée par une loi normale de moyenne θ et de variance $\frac{\mathrm{Var}(X)}{n}$ soit $\mathcal{N}(\theta, \mathrm{Var}(X)/n)$.
- La loi d'échantillonnage révèle la façon dont les réalisations de $\overline{X}$ oscillent autour de θ. Cette loi sert à
  - contrôler la marge d'erreur
  - construire une estimation par intervalle

Cas particulier de la moyenne : l'estimateur d'une proportion

- Soit $(X_1, ..., X_n)$ un échantillon de variables **binaires** de variable parente X codée par $X = 1$ si l'individu a une caractéristique souhaitée (ex : sexe masculin, guéri, vivant,...) et $X = 0$ sinon (ex : sexe féminin, non-guéri, décédé, ..). On veut estimer $\mathbb{E}[X] = p$.
- Un estimateur de p est $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i =$ proportion empirique d'individus ayant la caractéristique.
- Cet estimateur est convergent $\widehat{p} \overset{n\to\infty}{\to} p$.
- Cet estimateur est sans biais : $\mathbb{E}[\widehat{p}] = p$. Sa variance satisfait : $\mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n} \overset{n\to\infty}{\to} 0$.
- Pour n petit ($n < 30$), on utilise la loi exacte de $\sum_{i=1}^{n} X_i$: $\sum_{i=1}^{n} X_i$ suit une loi binômiale $\mathcal{B}(n, p)$.
- Pour n assez grand ($n \geq 30$, $np > 5$ et $n(1-p) > 5$), on utilise la loi approchée de $\widehat{p}$: $\widehat{p}$ suit approximativement une loi normale $\mathcal{N}(p, p(1-p)/n)$.

L'estimateur "variance empirique" et "écart-type empirique"

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X. On veut estimer $\mathrm{Var}(X)$.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$ est un estimateur de $\mathrm{Var}(X)$.
- Cet estimateur est convergent : $S^2 \to \mathrm{Var}(X)$ quand $n \to \infty$.
- S^2 est un estimateur sans biais : $\mathbb{E}[S^2] = \mathrm{Var}(X)$.
- La variance de S^2 satisfait $\mathrm{Var}(S^2) \to 0$ quand $n \to \infty$.
- Pour n assez grand ($n \geq 30$), S^2 suit approximativement une loi normale de moyenne $\mathbb{E}[S^2] = \mathrm{Var}(X)$ et de variance un peu compliquée dans le cas général.
- $S = \sqrt{S^2}$ est un estimateur de $\sigma(X)$.
- Lorsque $\mathbb{E}[X]$ est connu, $T^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X])^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\mathbb{E}[X]^2\right)$ est un estimateur de $\mathrm{Var}(X)$.

L'estimateur "distribution empirique"

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X de loi F.
- On veut estimer F. Un estimateur de F est la fonction de répartition empirique définie par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \quad \text{où} \quad I(X_i \leq x) = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{sinon} \end{cases}$$

- Cet estimateur est convergent : $F_n(x) \to F(x)$ quand $n \to \infty$
- On a : $\mathbb{E}[F_n(x)] = F(x)$ et $\text{Var}(F_n(x)) = \frac{1}{n}F(x)(1 - F(x)) \to 0$ quand $n \to \infty$.
- Les variables $I(X_i \leq x)$ sont des variables de Bernoulli indépendantes donc la loi exacte de $\sum_{i=1}^{n} I(X_i \leq x)$ est une loi binômiale $\mathcal{B}(n, F(x))$. Pour n assez grand ($n \geq 30$), on utilise la loi approchée (TCL) de $F_n(x)$ qui est une loi normale $\mathcal{N}(F(x), F(x)(1 - F(x))/n)$.

Cas des échantillons gaussiens

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X de loi $\mathcal{N}(m, \sigma^2)$. On veut estimer m et σ^2.
- Un estimateur de m est $\overline{X}$ et un estimateur de σ^2 est S^2.
- Dans le cas particulier où la loi de l'échantillon est gaussienne, on connaît la loi exacte de $\overline{X}$ et de S^2 : $\overline{X}$ suit une loi gaussienne $\mathcal{N}(m, \sigma^2/n)$ et $(n-1)S^2/\sigma^2$ suit une loi $\chi^2(n-1)$.
- On retrouve $\mathbb{E}[\overline{X}] = m$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$ et $\mathbb{E}[S^2] = \sigma^2$.
- Dans le cas particulier où la loi de l'échantillon est gaussienne, on a, en plus, $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$.

Théorèmes généraux de convergence

- La LFGN s'énonce comme suit. Soit $(X_1, X_2, ..., X_n, ...)$ des variables aléatoires indépendantes et toutes distribuées comme une même variable parente X. La convergence suivante a lieu quand $n \longrightarrow \infty$:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}[X].$$

- La loi forte des grands nombres (LFGN) permet d'approcher les espérances mathématiques comme $\mathbb{E}[X]$ par les moyennes empiriques correspondantes $\overline{X}$, d'approcher les probabilités théoriques par les proportions empiriques correspondantes...
- Le théorème central limite (TCL) permet d'approximer la loi des sommes de variables aléatoires indépendantes et équidistribuées. Il s'énonce comme suit. Soit $(X_1, X_2, ..., X_n, ...)$ des variables aléatoires indépendantes et toutes distribuées comme une même variable parente X. *(suite à la diapo suivante)*

Théorèmes généraux de convergence (suite)

- Pour n assez grand ($n \geq 30$), la loi de
$$\sqrt{n}\frac{\overline{X} - \mathbb{E}[X]}{\sigma(X)}$$
peut s'approximer par la loi $\mathcal{N}(0,1)$. Autrement écrit, pour n assez grand ($n \geq 30$), la loi de $\overline{X}$ peut s'approximer par la loi $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X)/n)$.
- Pourquoi y a-t-il dans la nature beaucoup de lois normales ?
  - La glycémie à jeun, G, est distribuée normalement. Pourquoi ? Un "modèle" raisonnable est de supposer que G est la somme d'un grand nombre de variables "explicatives" indépendantes : génétiques, environnementales, nutritionnelles,... Le TCL implique que G est alors distribuée normalement.
  - La loi de l'erreur de mesure commise au cours d'une expérience est souvent ditribuée normalement. Un "modèle" raisonnable est de supposer que l'erreur de mesure est la somme de différentes erreurs indépendantes : erreurs dues à chacun des appareils de mesure, erreur due au manipulateur...

Estimation par intervalle de confiance (IC) : fluctuations prévisibles de l'estimation

- Une estimation ponctuelle ne nous renseigne pas ni sur le niveau de confiance que l'on peut avoir en l'estimation, ni sur la marge d'erreur :
  - le niveau de confiance nous dit dans quelle mesure la méthode est fiable en usage répétée,
  - a marge d'erreur nous dit dans quelle mesure la méthode est sensible, i.e. avec quelle précision l'intervalle localise le paramètre en train d'être estimé.
- Lorsqu'on est intéressé non seulement par l'estimation en elle-même mais aussi par le niveau de confiance et la marge d'erreur, on effectue une estimation par intervalle.
- A taille d'échantillon fixé à n, lorsqu'on augmente le niveau de confiance $1 - \alpha$, la largeur de l'IC augmente.
- A niveau de confiance fixé à $(1 - \alpha)$, lorsqu'on augmente la taille de l'échantillon n, la largeur de l'IC diminue.

Estimation par intervalle de $\mathbb{E}[X]$ avec σ **connu**

- Estimation de $m = \mathbb{E}[X]$ à partir d'un échantillon de taille n de variable parente $X \sim \mathcal{N}(m, \sigma^2)$ avec σ **connu**.
  - $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ estime $\mathbb{E}[X]$ et $\overline{X} \sim \mathcal{N}(m, \sigma^2/n)$
  - L'intervalle de proba de $\overline{X}$ au niveau de confiance $(1-\alpha)$ est :

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{\overline{X} - m}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

    où $z_{\alpha/2}$ est le fractile bilatéral de la loi normale $\mathcal{N}(0,1)$ défini par :

$$\mathbb{P}[\mathcal{N}(0,1) \leq z_{\alpha/2}] = 1 - \alpha/2.$$

  - D'où l'IC pour m au niveau de confiance $(1-\alpha)$:

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  - **La probabilité que l'IC contienne m est $(1-\alpha)$.**
- Le TCL implique que cet IC est valable pour estimer $\mathbb{E}[X]$ à partir d'un échantillon de taille n de variable parente X de loi quelconque lorsque n est assez grand ($n \geq 30$).

Estimation par intervalle de $\mathbb{E}[X]$ avec σ **inconnu**

- Estimation de $m = \mathbb{E}[X]$ à partir d'un échantillon de taille n de variable parente $X \sim \mathcal{N}(m, \sigma^2)$ avec σ **inconnu**.
  - L'intervalle de proba de $\overline{X}$ au niveau de confiance $(1-\alpha)$ devient :

$$\mathbb{P}\left[-t_{\alpha/2} \leq \frac{\overline{X} - m}{\sqrt{S^2}/\sqrt{n-1}} \leq t_{\alpha/2}\right] = 1 - \alpha$$

    où $t_{\alpha/2}$ est le fractile bilatéral de la loi de Student $T(n-1)$ défini par :

$$\mathbb{P}[T(n-1) \leq t_{\alpha/2}] = 1 - \alpha/2.$$

  - D'où l'IC pour m au niveau de confiance $(1-\alpha)$:

$$\overline{X} - t_{\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n-1}} \leq m \leq \overline{X} + t_{\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n-1}}$$

  - **La probabilité que l'IC contienne m est $(1-\alpha)$.**
- Le TCL implique que cet IC est valable pour estimer $\mathbb{E}[X]$ à partir d'un échantillon de taille n de variable parente X de loi quelconque lorsque n est assez grand ($n \geq 30$).

Estimation par intervalle d'une proportion pour $n \geq 30$

- Estimation d'une proportion p à partir d'un échantillon de taille n.
- Un estimateur ponctuel de p est $\widehat{p}$.
- L'IC pour p au niveau de confiance $(1-\alpha)$ est approché lorsque n est assez **grand** $(n \geq 30)$ par :

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

où $z_{\alpha/2}$ est le fractile bilatéral de la loi normale $\mathcal{N}(0,1)$ défini par :

$$\mathbb{P}[\mathcal{N}(0,1) \leq z_{\alpha/2}] = 1 - \alpha/2.$$

- **La probabilité que l'IC contienne p est d'environ $(1-\alpha)$.**

Estimation par intervalle de Var(X) avec m **connu**

- Estimation de $\sigma^2 = \text{Var}(X)$ à partir d'un échantillon de taille n de variable parente $X \sim \mathcal{N}(m, \sigma^2)$ avec m **connu**.
  - On utilise $T^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$, estimateur de Var($X$).
  - L'intervalle de proba de T^2 au niveau de confiance $(1-\alpha)$ est :

$$\mathbb{P}\left[k_1 \leq \frac{nT^2}{\sigma^2} \leq k_2\right] = 1 - \alpha$$

    où k_1 et k_2 sont les fractiles bilatéraux de la loi $\chi^2(n)$ définis par :

$$\mathbb{P}[\chi^2(n) \leq k_1] = \alpha/2 \quad \text{et} \quad \mathbb{P}[\chi^2(n) \leq k_2] = 1 - \alpha/2.$$

  - D'où l'IC pour σ^2 au niveau de confiance $(1-\alpha)$:

$$\frac{nT^2}{k_2} \leq \sigma^2 \leq \frac{nT^2}{k_1}$$

  - **La probabilité que l'IC contienne σ^2 est $(1-\alpha)$.**
- Cette formule d'IC pour σ^2 est valable **exclusivement** pour $X \sim \mathcal{N}(m, \sigma^2)$.

Estimation par intervalle de Var(X) avec m **inconnu**

- Estimation de $\sigma^2 = \text{Var}(X)$ à partir d'un échantillon de taille n de variable parente $X \sim \mathcal{N}(m, \sigma^2)$ avec m **inconnu**.
  - on utilise $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, estimateur de Var($X$).
  - L'intervalle de proba de S^2 au niveau de confiance $(1-\alpha)$ est :

$$\mathbb{P}\left[k_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq k_2\right] = 1-\alpha$$

    où k_1 et k_2 sont les fractiles bilatéraux de la loi $\chi^2(n-1)$ définis par :

$$\mathbb{P}[\chi^2(n-1) \leq k_1] = \alpha/2 \quad \text{et} \quad \mathbb{P}[\chi^2(n-1) \leq k_2] = 1-\alpha/2.$$

  - D'où l'IC pour σ^2 au niveau de confiance $(1-\alpha)$:

$$\frac{(n-1)S^2}{k_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{k_1}$$

  - **La probabilité que l'IC contienne σ^2 est $(1-\alpha)$.**
- Cette formule d'IC pour σ^2 est valable **exclusivement** pour $X \sim \mathcal{N}(m, \sigma^2)$.

Estimation par IC de Var(X) pour $n \geq 30$

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X.
- Pour n assez grand, la loi de $\frac{\sqrt{n-1}}{\sqrt{2}} \left(\frac{S^2}{\text{Var}(X)} - 1 \right)$ peut être approchée par une loi $\mathcal{N}(0,1)$.
- L'intervalle de proba au niveau de confiance $(1-\alpha)$ s'approxime par :

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{\sqrt{n-1}}{2} \left(\frac{S^2}{\text{Var}(X)} - 1 \right) \leq z_{\alpha/2} \right] = 1 - \alpha$$

où $z_{\alpha/2}$ est le fractile bilatéral de la loi normale $\mathcal{N}(0,1)$ défini par :

$$\mathbb{P}[\mathcal{N}(0,1) \leq z_{\alpha/2}] = 1 - \alpha/2.$$

- L'IC pour Var(X) au niveau de confiance $(1-\alpha)$ peut être approché par : (valable pour $n \geq 30$! !)

$$\frac{S^2}{1 + z_{\alpha/2} \frac{\sqrt{2}}{\sqrt{n-1}}} \leq \text{Var}(X) \leq \frac{S^2}{1 - z_{\alpha/2} \frac{\sqrt{2}}{\sqrt{n-1}}}$$

- **La probabilité que l'IC contienne la variance de la population est d'environ $(1-\alpha)$.**

Exemple : taux de cholestérol nominal

On considère que la valeur moyenne "normale" du taux de cholestérol est 2g/L. Dans une population de patients bien définie cliniquement, on a prélevé au hasard un échantillon de 150 sujets sur lesquels on a mesuré le taux de cholestérol X en g/L. Les résultats ont été les suivants : $\sum_{i=1}^{n} x_i = 273\text{g/L}$, $\sum_{i=1}^{n} x_i^2 = 512.12(\text{g/L})^2$. Comment modéliser cette situation ? Quelle quantité permettrait d'apprécier si la population de patients bien définie cliniquement présente un taux de cholestérol nominal ?

Exemple : impact de la présence de lipides dans l'alimentation sur le poids

On tire au sort un échantillon de personnes agées entre 30 et 50 ans. On relève le poids X en kg de chaque sujet ainsi que son régime alimentaire classé en deux catégories : régime moyen (R1) et régime riche en lipides (R2). Les données suivantes sont recueillies :
R1 : $n_1 = 80$, $\sum_{i=1}^{n_1} x_{1,i} = 5200\text{kg}$ et $\sum_{i=1}^{n_1} x_{1,i}^2 = 338789\text{kg}^2$
R2 : $n_2 = 90$, $\sum_{i=1}^{n_2} x_{1,i} = 6120\text{kg}$ et $\sum_{i=1}^{n_2} x_{2,i}^2 = 417225\text{kg}^2$
Comment modéliser cette situation ? Quelle(s) quantité(s) permettrai(en)t d'apprécier si la présence de lipides dans l'alimentation influe sur le poids ?

Exemple : comprimés défectueux

Lors du contrôle d'une chaine de médicaments, on s'intéresse au nombre de comprimés défectueux dans un lot. L'étude de 200 lots a donné les résultats suivants :

nb de comprimés défectueux	0	1	2	3	4	5
effectif	75	53	39	23	9	1

Quelle(s) quantité(s) pourrait-on modéliser et estimer pour se faire une idée de la qualité des médicaments produits ?

Exemple : effets secondaires

Sachant que 20% des personnes vaccinées par la formule F d'un vaccin présentent des allergies ou troubles secondaires, un laboratoire pharmaceutique propose une formule améliorée de ce vaccin Fa et espère diminuer le taux d'allergies et de troubles secondaires. Sur un échantillon de 400 personnes prises au hasard et ayant opté pour la formule Fa, on observe 60 cas d'allergies ou troubles secondaires. Comment modéliser cette situation ? Quelle quantité pourrait permettre d'apprécier si la formule améliorée apporte réellement un bénéfice ?

Exemple : impact sur le taux de cholestérol

On réalise une étude sur l'impact des habitudes alimentaires et de la pratique d'une activité physique régulière sur le taux de cholestérol sanguin. Un échantillon de 25 personnes est constitué. Pour chacune de ces personnes, on recueille les informations suivantes :

taux de cholestérol sanguin (en g/L)	1.88 2.09 1.69 1.95 2.23 2.19 1.75 2.20 2.02 1.83 2.03 2.04 2.08 2.07 2.13 1.81 1.97 2.13 2.00 1.97 2.11 2.15 1.84 2.19 2.18
sexe	F F F F F F F F F F H H H H H H H H H H H H H H H
consommation de lipides (en g/jour)	65 70 70 68 80 75 70 75 72 70 95 95 97 96 97 92 95 100 94 94 97 98 93 98 97
activité physique régulière	non non non oui non non oui non oui oui oui oui oui oui non oui oui non oui oui non non oui oui non

Comment modéliser cette situation au moyen de variables aléatoires ? Comment estimer les quantités suivantes : taux de cholestérol moyen pour la population, consommation journalière moyenne de lipides, proportion de personnes pratiquant régulièrement une activité physique ?

Exemple : efficacité d'un régime hypocholestérolémiant

On étudie l'efficacité d'un régime hypocholestérolémiant sur des animaux comparativement à des animaux témoins (non traités). On mesure le taux de cholestérol X dans les deux groupes et on obtient les résultats suivants :
Groupe traité : $n_1 = 100$, $\sum_{i=1}^{n_1} x_{1,i} = 250\text{g/L}$ et $\sum_{i=1}^{n_1} x_{1,i}^2 = 1025(\text{g/L})^2$
Groupe témoin : $n_2 = 100$, $\sum_{i=1}^{n_2} x_{2,i} = 400\text{g/L}$ et $\sum_{i=1}^{n_2} x_{2,i}^2 = 2140(\text{g/L})^2$
Comment modéliser cette situation au moyen de variables aléatoires ? Que pourrait-on faire pour apprécier l'efficacité du régime hypocholestérolémiant ?

Exemple : efficacité d'un traitement

Un essai thérapeutique visant à évaluer l'efficacité de deux traitements a été réalisé en cross-over (i.e. chaque patient est pris comme son propre témoin et reçoit les deux médicaments dont l'ordre d'administration est tiré au sort). Les valeurs suivantes de taux de cholestérol X en g/L ont été obtenues après administration de chaque traitement :
Après traitement A : $n_1 = n = 100$, $\sum_{i=1}^{n_1} x_{1,i} = 269$,
$\sum_{i=1}^{n_1} x_{1,i}^2 = 1370$,
Après traitement B : $n_2 = n = 100$, $\sum_{i=1}^{n_2} x_{2,i} = 302$,
$\sum_{i=1}^{n_1} x_{1,i}^2 = 2899$,
Différence : $\sum_{i=1}^{n}(x_{1,i} - x_{2,i}) = -33$, $\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2 = 194$.
Comment modéliser cette situation au moyen de variables aléatoires ? Que pourrait-on faire pour apprécier l'efficacité du traitement ?