GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 'Minime'

- Resumé du semestre

Principe fondamental de dénombrement généralisé (PFDG)

- Si r expériences doivent être réalisées et sont telles que :
  - la première peut produire l'un quelconque de n_1 résultats,
  - si pour *chacun d'entre eux* il y a n_2 résultats possibles pour la deuxième expérience,
  - si pour *chaque résultat des deux premières expériences* il y a n_3 résultats possibles pour la troisième expérience,
  - et ainsi de suite . . .
- Il y aura alors *un total de* $\boxed{n_1 \times n_2 \times \cdots \times n_r}$ résultats possibles pour les r expériences prises ensemble.

Permutations : arrangements ordonnés

- Un *arrangement ordonné* d'objets est appelé une **permutation**
- **Exemple :** Combien existe-t-il d'arrangements ordonnés des lettres a, b et c ?

1. *Énumération directe :* faire la liste de toutes les possibilités, puis compter
2. *Principe fondamental :*
   - la première lettre de la permutation peut être n'importe laquelle des 3,
   - la deuxième peut ensuite être choisie parmi les 2 restantes,
   - tandis que la troisième ne peut plus faire l'objet d'aucun choix (c.-à-d. 1 'choix')

- En utilisant le *principe fondamental généralisé*, on a que : le nombre de permutations de n *objets discernables* est $\boxed{n!}$ (n **factorielle** $= n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$)

Permutations : objets partiellement indiscernables

Exemple 2.4 Combien *d'arrangements différents* peut-on former avec les letters $E\ R\ R\ E\ U\ R$?

Solution

- D'abord, comptons le nombre de permutations quand les 3 R et les 2 E *sont distincts* $(E_1R_1R_2E_2UR_3) =$ _____
- Cependant, considérons l'une quelconque des permutations, p. ex. $E_1R_1E_2R_2UR_3$. Si nous permutions les R entre eux et les E entre eux, l'arrangement résultant serait toujours de la même forme, soit : $ERERUR$
  - Il existe combien de permutations des R et des E : _____
  - Donc, combien de permutations (le total) : $=$ _____
- Le nombre de permutations de n objets, *parmi lesquels* n_1 *sont indiscernables* entre eux, n_2 *sont indiscernables* entre eux, ..., n_r *sont indiscernables* entre eux : $\boxed{\frac{n!}{n_1!n_2!\cdots n_r!}}$

Combinaisons : séléction sans ordre

- Ensuite, déterminons le nombre de groups de r objets qu'il est possible de former sans répétition à partir d'un total de n objets (quand *l'ordre des objets n'est pas significatif*)

Exemple 2.5 Combien de groupes de 3 batons peut-on construire en tirant parmi 5 souris (A, B, C, D, E) ?

Solution On utilise le raisonnement suivant :

Puisqu'il y a ____ façons de choisir le premier batons, puis ____ de choisir ensuite le deuxième et finalement ____ de choisir le dernier, il y a donc _____ *en tenant compte de l'ordre dans lequel ces batons sont choisis*.

Cependent, un triplé donné, p. ex. le triplé constitué des batons A, B, D, apparaîtra _____ fois.

Donc, le nombre total de groupes pouvant être formés est ____

Coefficients binomiaux et multinomiaux

- L'expression $\binom{n}{r}$ (r *parmi* n) pour $r \leq n$, le **coefficient binomial**, est définie par : $\boxed{\binom{n}{r} = \frac{n!}{r!(n-r)!}}$
- Tout sous-ensemble de r objets choisis sans répétition dans un ensemble en contenant n est appelé **combinaison de r objets pris parmi n**
- Le nombre $\binom{n}{r}$ est le nombre de combinaisons de r objets pris parmi n si **l'ordre des objets est sans importance**
- En utilisant *le principe fondamental généralisé*, on déduit que le nombre de répartitions possibles de n objets en r groupes distincts de tailles respectives $n_1, n_2, \ldots, n_r$ est le **coefficient multinomial** : $\boxed{\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\ n_2! \cdots n_r!}}$

Ensemble fondamental

- Considérons une 'expérience' dont l'issue n'est pas prévisible – p. ex., je lance un dé équilibré
- Bien que *l'issue de l'expérience* ne soit pas connue d'avance, admettons que *l'ensemble des issues possibles* est connu
- Cet ensemble des issues possibles de l'expérience est appelé **l'ensemble fondamental** de l'expérience, noté S (ou Ω dans quelques livres)
- L'ensemble fondamental pourrait être *discret* ou *continu*

Événements

- Tout sous-ensemble $E \subset S$ est appelé un **événement**
- Pour toute paire d'événements E et F, le nouvel événement $E \cup F$ (**l'union** de E et F) contient chaque élément se trouvant *dans E, dans F* **OU** *dans les deux à la fois*
- De même pour toute paire d'événements E et F, le nouvel événement $E \cap F$ (**l'intersection** de E et F) est défini comme l'ensemble des réalisations qui sont *à la fois dans E* **ET** *dans F*
  - Si $E \cap F = \varnothing$ (**l'événement vide**), alors E et F sont dits **mutuellement exclusifs (ME)**
- Le nouvel événement E^c, le **complément** de E, contient tous les éléments de S *qui* **ne sont pas** *dans E*

La probabilité

- Théorie fréquenciste des probabilités : $P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$

Axiomes de probabilité

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. Pour chaque séquence d'événements mutuellement exclusifs $E_1, E_2, \ldots$, $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

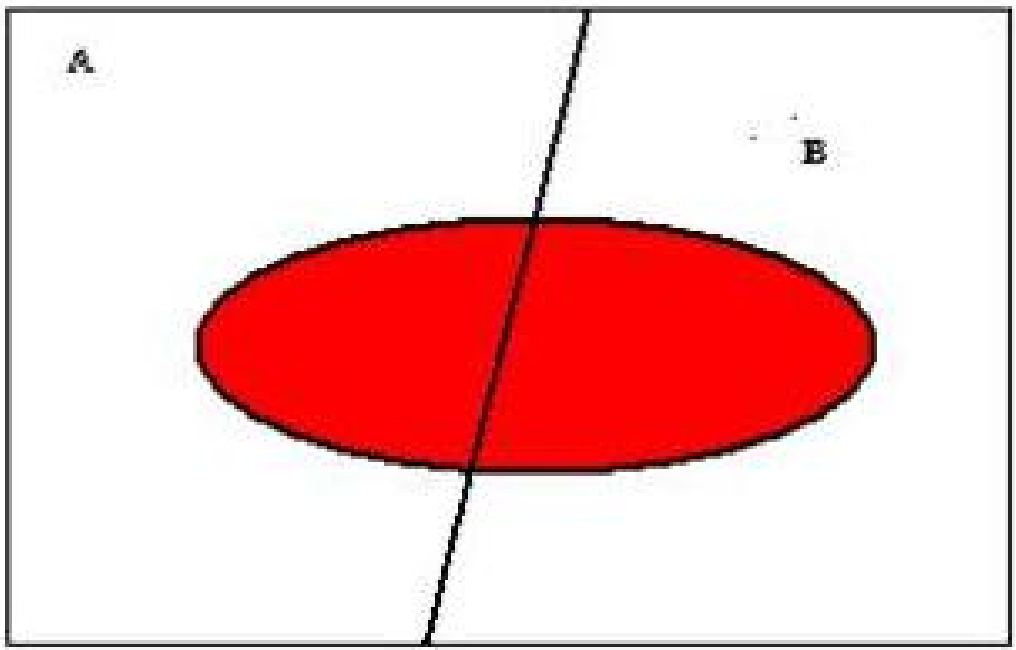
- Événements élémentaires équiprobables : il y a un nombre fini d'éléments élémentaires de l'ensemble fondamental à la *même probabilité d'apparaître* : $P(E) = \dfrac{\text{nombre de points dans } E}{\text{nombre de points dans } S}$

Probabilité conditionnelle et $P(E \text{ et } F)$

- On considère la *définition de la probabilité conditionnelle* :
  1. $P(E \mid F) = \dfrac{P(E \text{ et } F)}{P(F)}$
- On a également :
  2. $P(F \mid E) = \dfrac{P(E \text{ et } F)}{P(E)}$
- Donc on a **deux façons** d'exprimer $P(E \text{ et } F)$:
  1. $P(E \text{ et } F) = P(F)P(E \mid F)$
  2. $P(E \text{ et } F) = P(E)P(F \mid E)$
- Ceci pourrait être généralisé :
  - $P(E_1 \text{ et } E_2 \text{ et } \ldots \text{ et } E_n)$
    $= P(E_1) \times P(E_2 \mid E_1) \times P(E_3 \mid E_1, E_2)$
    $\times \cdots \times P(E_n \mid E_1, E_2, \ldots, E_{n-1})$

Partition

- Une **partition** (cloison) divise l'ensemble fondamental en des *sous-ensembles disjoints* :
  - sans trou (en anglais 'no gaps')
  - sans superflue (en anglais 'no overlaps')

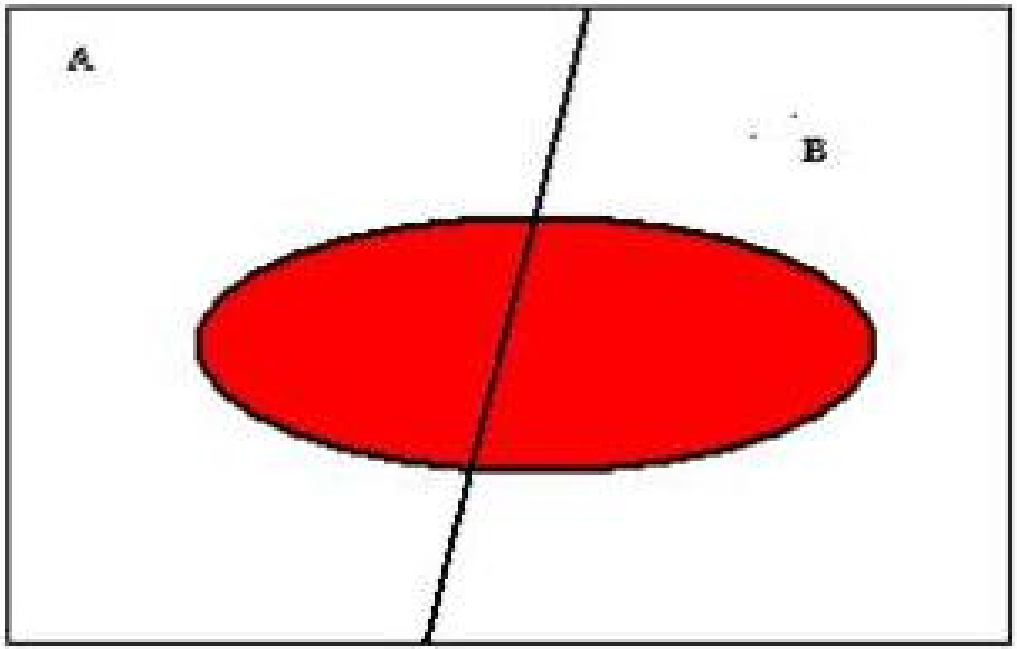


Formule des probabilités totales

- Soit A, B une partition de l'ensemble fondamental et soit R un événement ; alors,

$$\begin{aligned} P(R) &= P(R \text{ et } A) + P(R \text{ et } B) \\ &= P(R \mid A)P(A) + P(R \mid B)P(B) \end{aligned}$$

- La partition pourrait être composée de plus que deux événements / sous-ensembles



Formule de Bayes

- Pour une *partition* $F_1, F_2, \ldots, F_n$ de l'espace fondamental S :

$$\underline{P(F_j \mid E)} = \frac{P(E \text{ et } F_j)}{P(E)}$$

$$= \frac{P(E \mid F_j)P(F_j)}{\sum_{i=1}^{n} P(E \mid F_i)P(F_i)}$$

$$= \boxed{\frac{P(E \mid F_j)P(F_j)}{P(E \mid F_1)P(F_1) + P(E \mid F_2)P(F_2) + \cdots + P(E \mid F_n)P(F_n)}}$$

- La formule de Bayes utilise les deux expressions pour $P(E \text{ et } F)$

Indépendance

- En générale, $P(E \mid F) \neq P(E)$; c.-à-d., le fait de savoir que F est survenu *influe sur* **la probabilité de E**
- Dans le **cas particulier** où $P(E \mid F) = P(E)$, on dit que les événements E et F sont **indépendants**
- Les événements sont indépendants lorsque le fait de savoir que l'un est survenu *ne modifie pas la probabilité* que l'autre se produit
- Pour les événements *indépendants* (mais **PAS** en général !),

$$P(E \text{ et } F) = P(E) \times P(F)$$

- Deux événements sont *dépendants* s'ils ne sont pas indépendants

Epreuves (aléatoires)

- Il arrive parfois que l'expérience étudiée consiste à effectuer une suite d'*expériences partielles* – p. ex. plusieurs lancements d'une pièce
- Il est peut-être raisonnable d'admettre que l'issue de tout groupe d'expériences partielles sont *totalement indépendantes* – encore, lancements d'une pièce
- Si toutes ces expériences partielles sont *identiques* (le même ensemble fondamental, la même fonction de probabilité), elles sont appelées **épreuves**

VAs discrètes

- **Variable Aléatoire (VA)** : une fonction réelle définie sur l'ensemble fondamental
  - VA : MAJUSCULES ; valeur spécifique : miniscules
- **VA discrète** :
  1. loi de probabilité : $p(x) = P(X = x)$
  2. fonction de répartition : $F(x) = P(X \leq x) = \sum_{i \leq x} p(i)$
- **Resolution des problèmes avec VAs**
  1. *Identifier la VA*
  2. Déterminer la *distribution* (loi) de la VA
  3. *Traduire* la question
  4. *Répondre* à la question

Espérance (VA discrète)

- Pour une VA discrète X de loi $p(x)$, on définit l'**espérance** (ou la **moyenne**) par :

$$E[X] = \sum_{\substack{\text{toutes} \\ \text{valeurs } x}} xp(x)$$

- Donc c'est la *moyenne pondérée* des valeurs possibles de X, où les poids sont $P(X = x)$
- C'est également possible à calculer l'espérance d'une *fonction* de la VA (discrète) X (disons $g(X)$) dans la même manière
- $g(X)$ elle aussi est une VA discrète, donc pour calculer $E[g(X)]$ il suffira de trouver sa loi (distribution) $p(g(x))$
- On devrait pouvoir déduire la distribution de celle de X

$E[g(X)]$

- **Théorème :** Si X est une VA discrète pouvant prendre ses valeurs parmi les valeurs x_i, $i \geq 1$, avec des probabilités respectives $p(x_i)$, alors pour toute fonction réelle g on a

$$E[g(X)] = \sum_i g(x_i)p(x_i)$$

- Pour toute paire (a, b) de constantes, $E[aX + b] = aE[X] + b$

Variance et écart-type

- Comme on s'attend à voir toute variable X prendre ses valeurs autour de son espérance $E[X]$, il paraît raisonnable de mesurer les variations en considérant l'écart moyen entre X et $E[X]$, $E[|\, X - \mu \,|]$, où $\mu = E[X]$
- Il est plus facile techniquement (en maths) de considerer le moyen du *carré* de l'écart entre X et sa espérance $E[X]$
- Pour la VA X avec espérance μ, on définit la **variance** de X :

$$Var(X) = E[(X - \mu)^2]$$

- On peut établir une *formule alternative* pour le calcul de $Var(X)$ (plus commode dans la pratique) :

$$Var(X) = E[X^2] - (E[X])^2$$

- L'**écart-type** de X (σ) est la racine carré de $Var(X)$:

$$\sigma = \sqrt{Var(X)}$$

Variance d'une fonction linéaire d'une VA X

- Pour toute paire (a, b) de constantes,

$$Var(aX + b) = a^2 Var(X)$$

- Facile à démontrer :

$$\begin{aligned} Var(aX + b) &= E[(aX + b) - E(aX + b)]^2 \\ &= E[aX + b - (aE[X] + b)]^2 \\ &= E[aX - aE[X]]^2 \\ &= E[a^2(X - E[X])^2]] \\ &= a^2 E[(X - \mu)^2] \\ &= a^2 Var(X) \end{aligned}$$

- Donc pour une fonction linéaire de X, on a :

$$SD(aX + b) = \mid a \mid \ SD(X)$$

- ('SD' = 'écart-type' = 'standard deviation' en anglais)

VA de Bernoulli

- Une VA de *Bernoulli* prend les valeurs 0 et 1
- Sa loi de probabilité est :

x	0	1
$p(x)$	$(1-p)$	p

- Utilisée dans la modelisation des problèmes ayant 2 résultats possibles : pile/face ; oui/non ; succès/échec ; *etc.*
- Pour une VA de Bernoulli X :
  - $E(X) = 0 \times (1-p) + 1 \times p =$ **p**
  - $Var(X) = E[X^2] - (E[X])^2$

    $= [0^2 \times (1-p) + 1^2 \times p] - p^2 = p - p^2 =$ **p(1-p)**

VA binomiale et VA de Poisson

binomiale : $X \sim Bin(n, p)$

- échantillonage *AVEC* remise
  1. nombre fixe n
  2. épreuves de Bernoulli
  3. épreuves indépendantes
  4. chacune a la même probabilité p de 'succès'
- loi : $\boldsymbol{P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}}$; $\boldsymbol{x = 0, 1, \ldots, n}$
- $E[X] =$ **np**, $Var(X) =$ **np(1-p)**

Poisson : $X \sim Pois(\lambda)$

- approximation de la loi binomiale pour n grand, p petit, np $(= \lambda)$ moyen
- loi : $\boldsymbol{p(i) = e^{-\lambda}\lambda^i / i!}$; $\boldsymbol{i = 0, 1, 2, \ldots}$
- $E[X] = \boldsymbol{\lambda}$, $Var(X) = \boldsymbol{\lambda}$

VAs continues

- **VA continue : densité de probabilité :**
  - $P(X \in B) = \int_B f(x)dx$
  - $f(x) \geq 0$ pour chaque x
  - $\int_x f(x)\,dx = 1$
- **VA continue : fonction de répartition :**
  - $F(x) = P(X \leq x) = \int_0^x f(u)\,du$
  - $F(-\infty) = 0$
  - $F(\infty) = 1$

Relation entre densité et répartition

- La relation entre la fonction de répartition F et la densité f d'une VA continue X est donnée par

$$F(a) = P(X \in (-\infty, a]) = \int_{-\infty}^{a} f(x)dx$$

- *La dérivation* des deux membres dans l'équation ci-dessus donne

$$\frac{d}{da}F(a) = f(a)$$

- c.-à-d., la densité d'une VA continue est *la dérivée* de la fonction de répartition

Espérance, variance d'une VA continue

- Si X est une VA *continue* ayant pour densité $f(x)$, la définition analogue de **l'espérance** de X est simplement :

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

- Pour toute fonction réelle g, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
- Facile de montrer que pour toute paire a et b de constantes,

$$E[aX + b] = aE[X] + b$$

- La **variance** d'une VA continue est définie exactement comme celle d'un VA discrète : $Var(X) = E[(X - E[X])^2]$
- On a aussi l'autre formule : $Var(X) = E[X^2] - (E[X])^2$
- Pour les constantes a et b, on a $Var[aX + b] = a^2 Var(X)$

VA uniforme et VA normale

Uniforme :

- densité : $\frac{1}{\beta - \alpha}, \quad \alpha < x < \beta \quad (= 0 \text{ sinon})$
- $E[X] = \frac{\beta + \alpha}{2}$, $Var(X) = \frac{(\beta - \alpha)^2}{12}$

Normale :

- densité : $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \qquad -\infty < x < \infty$
- $E[X] = \mu$, $Var(X) = \sigma^2$

Distribution de Y = aX + b

- Une propriété importante de la famille des variables normales (ce qui *NE TIENT PAS* pour toutes VAs) est que :

$$\textit{si } X \sim N(\mu, \sigma^2), \textit{ alors } Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- Pour démontrer ce résultat, on peut trouver la fonction de répartition F de la VA $Y = aX + b$; $X \sim N(\mu, \sigma^2)$, et a, b constantes
- La dérivation de F donne la densité de Y, qui est de la forme d'une densité normale

VA normale centrée réduite

- L'application la plus utile du résultat précédent consiste à déterminer les probabilités des VAs normalement distribuées
- Si $X \sim N(\mu, \sigma^2)$, la VA $Z = (X - \mu)/\sigma \sim N(0, 1)$
- La distribution de Z est **normale centrée réduite** (ou **standard**)
- On note la fonction de densité $f(z)$ d'une variable normale centrée réduite par le symbole ϕ, et la fonction de répartion par Φ :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz$$

- Cette intégrale n'a pas de forme simple, donc on utilise un *table des valeurs calculées* (ou un logiciel) pour trouver *l'aire sous la courbe*

Résoudre les problèmes en utilisant la table normale

- Étapes à suivre pour résoudre les problèmes impliquant *VAs normalement distribuées* (c.-à-d. étape 4 : "Répondre à la question" pour $X \sim N(\mu, \sigma^2)$) :

1. *Centrer et Réduire* ('*standardiser*') la VA
2. **** DESSINER L'IMAGE ****
3. Utiliser la *table normale* afin de trouver la probabilité
4. (Faire le calcul)

L'approximation normale d'une VA binomiale

- Il s'avère que si n est *suffisament grand**, $S_n \sim Bin(n, p)$ est *approximativement normalement distribuée*, ayant la même moyenne et la même variance que la VA binomiale
- **Théorème limite de DeMoivre-Laplace**
  Soit S_n le nombre de 'succès' lors de la réalisation de n épreuves indépendantes, la probabilité de réussite pour chaque épreuve étant p. Alors pour tout $a < b$,

$$P\left\{a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right\} \to \Phi(b) - \Phi(a) \quad \text{lorsque } n \to \infty$$

- C.-à-d. la distribution d'une VA binomiale standardisée *converge vers la distribution normale standard* lorsque le nombre d'épreuves $n \to \infty$

$^*np \geq 10 \quad \underline{\textbf{et}} \quad n(1-p) \geq 10$

Fonction de répartition conjointe

- On n'a traité jusqu'ici que des distributions de VAs *isolées*
- Dans la pratique, il est souvent nécessaire de considérer des événements relatifs à deux (ou même plus) variables *simultanément*
- Pour traiter de tels problèmes on définit une **fonction F de répartition simultanée** (ou **conjointe**) pour toute paire de VAs X et Y :

$$F(a,b) = P(X \leq a, Y \leq b) \qquad -\infty < a, b < \infty$$

- Tout comme avant, en sachant la fonction de répartion des ensembles de VAs (également la loi ou la densité), *on pourrait répondre aux questions concernant les probabilités*

Densité conjointe

- Les VAs X et Y sont dites **conjointement continues** s'il existe une fonction $f(x, y)$ pour toute paire x et y réels ayant pour tout sous-ensemble C du plan

$$P((X, Y) \in C) = \int \int_{(x,y)\in C} f(x, y)\, dx\, dy$$

- La fonction $f(x, y)$ est appelée **densité conjointe** ou **simultanée** de X et Y (également **pdf**)
- Notons par A et B deux ensembles de nombres réels, $C = \{(x, y) : x \in A,\ y \in B)\}$; on a :

$$P(X \in A, Y \in B) = \int_B \int_A f(x, y)\, dx\, dy$$

- La fonction de densité conjointe peut être obtenue à partir de la fonction de répartition conjointe :

$$f(a, b) = \frac{\partial^2}{\partial a\, \partial b} F(a, b)$$

(pour autant que les dérivées partielles soient définies)

Densité marginale

- Si X et Y sont des VAs conjointement continues, elles sont également *individuellement continues*
- On obtient la **densité marginale** de chaque VA ainsi :

$$\begin{aligned} P(X \in A) &= P(X \in A,\ Y \in (-\infty, \infty)) \\ &= \int_A \left[\int_{-\infty}^{\infty} f(x,y)\, \underline{\boldsymbol{dy}} \right] dx = \int_A f_X(x)\, dx, \end{aligned}$$

où $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, \underline{\boldsymbol{dy}}$ est la densité (marginale) de X

- On obtient de même l'expression de la densité (marginale) de Y :

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, \underline{\boldsymbol{dx}}$$

Variables aléatoires indépendantes – Pratique

- **Théorème** : Les VAs X et Y sont indépendantes *si et seulement si* la loi conjointe (VAs discrètes) ou la densité conjointe (VAs continues) se factorise :

$$p_{X,Y}(x,y) = g(x)\,h(y) \qquad \text{pour tout } x \text{ et tout } y;$$

$$f_{X,Y}(x,y) = g(x)\,h(y), \qquad -\infty < x < \infty,\ -\infty < y < \infty$$

- En général, les VAs $X_1, X_2, \ldots, X_n$ sont dites **indépendantes** si pour tout choix de n ensembles de nombres réels $A_1, A_2, \ldots, A_n$,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

Distribution de la somme des VAs normales indépendantes

- Pour VAs $X_1, \ldots, X_n$:

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

- Pour VAs $X_1, \ldots, X_n$ **indépendantes** :

$$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n]$$

- **Théorème** : Soient $X_1, \ldots, X_n$ les variables aléatoires **indépendantes normales** de paramètres $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$
- Alors,

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Théorème Central Limite (TCL)

Le **Théorème Central Limite (TCL)** est l'un des *résultats les plus importants* de la probabilité et la statistique, et est largement utilisé comme un *outil* pour la résolution de problèmes.

Théorème (TCL) : Soient $X_1, X_2, \ldots$ des variables aléatoires *indépendantes et identiquement distribués (iid)*, et telles que $E[X_i] = \mu$ et $Var(X_i) = \sigma^2 < \infty$ existent. Alors, la distribution de

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

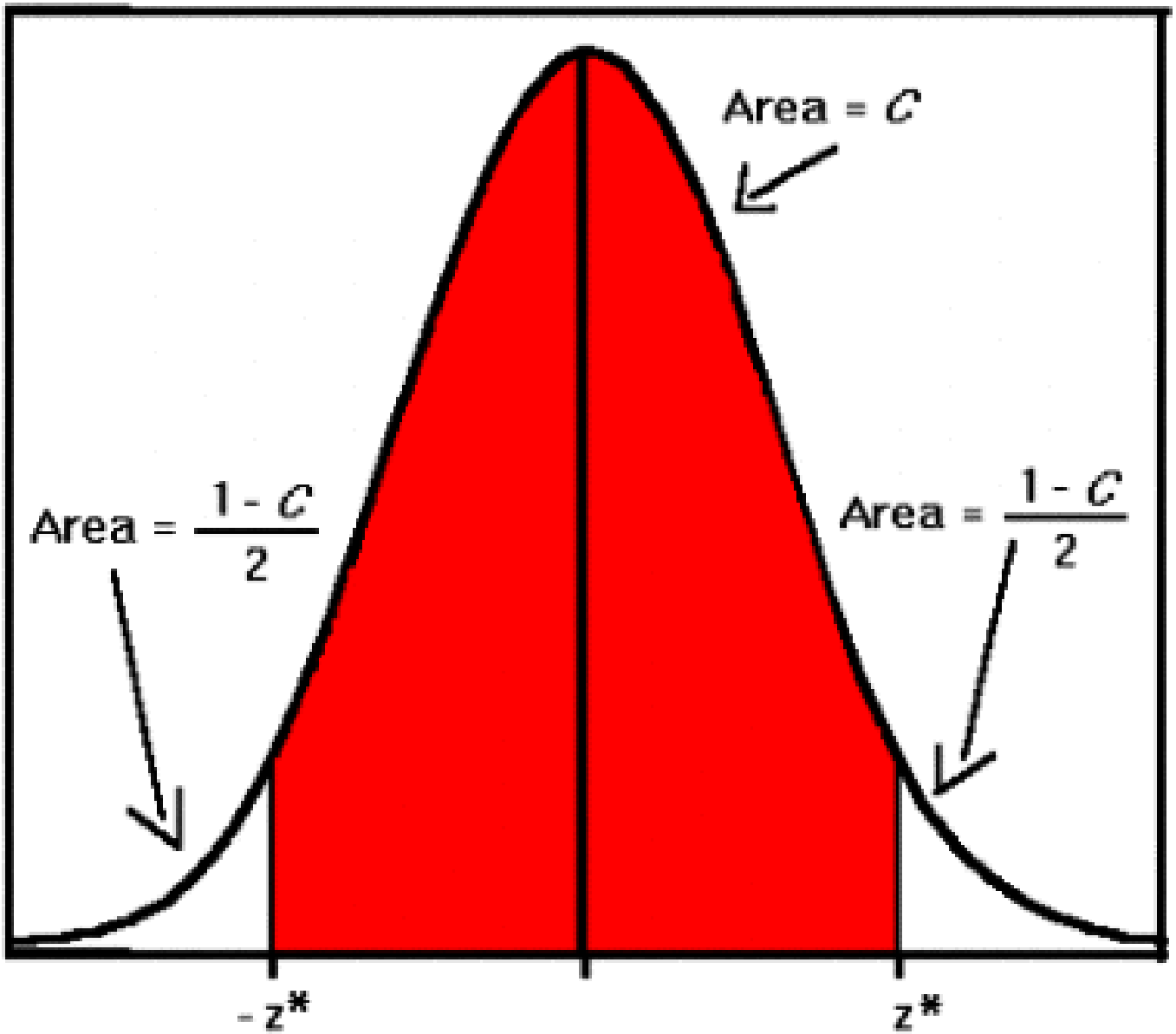
se rapproche d'une distribution normale lorsque $n \to \infty$.

C.-à-d. : Plus n est grand ('suffisament grand'), plus *la loi de la somme (ou la moyenne)* se rapproche d'une distribution normale.

IC – Suppositions

1. Il y a un paramètre de la population dont la valeur est *inconnue*
2. Il y a un *échantillon aléatoire* (observations independantes ou EAS d'une population nombreuse, où la taille de l'échantillon est petite par rapport à celle de la population)
3. TCL s'appliquent

Illustration



Qualité d'un estimateur : Biais / Variance

- Le **biais** d'un estimateur T d'un paramètre θ est définit par :

$$b(T) = E[T] - \theta,$$

(c.-à-d. la différence entre *l'espérance* de la distribution d'échantillonnage de l'estimateur T et *la vraie valeur* du paramètre θ)

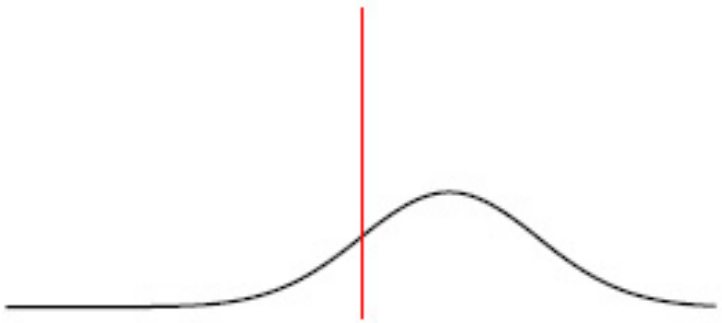
- Un estimateur est **sans biais** (ou **non biaisé**) si le biais égale à 0
- Une autre qualité on peut considérer est le *variance* de l'estimateur :

$$Var(T) = E[(T - E[T])^2]$$

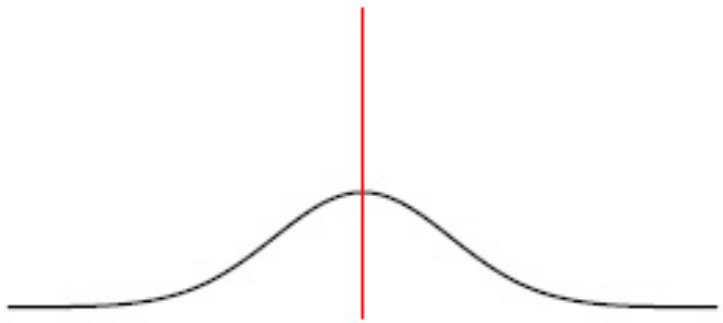
- Parmi deux estimateurs sans biais de θ, l'un sera *plus efficace* que l'autre si *sa variance est plus petite*

Biais et variance d'un estimateur T

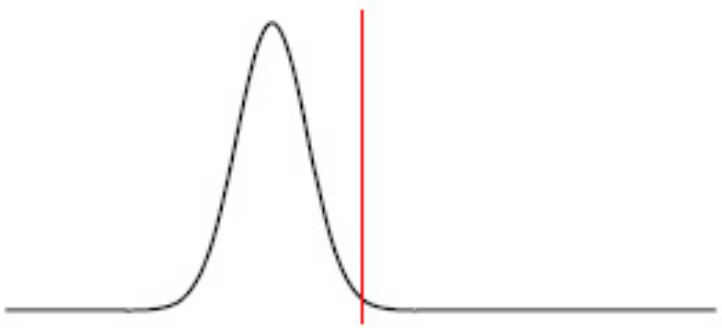
big bias, big variance



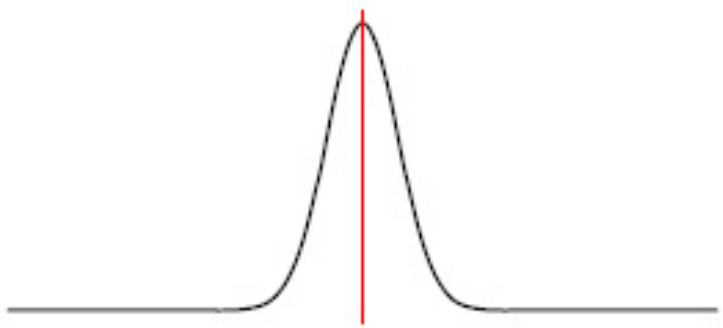
no bias, big variance



big bias, low variance



no bias, low variance



Qualité d'un estimateur : Erreur Quadratique Moyenne (EQM)

- Une autre qualité que nous pouvons considérer est le **erreur quadratique moyenne (EQM)** d'un estimateur

$$EQM(T) = E[(T - \theta)^2]$$

- Ceci est différent de la variance lorsque l'estimateur T est biaisé
- Parfois, nous pourrions utiliser un estimateur qui a un peu de biais s'il a une variance beaucoup plus petite que la meilleure estimateur sans biais (*compromis biais-variance*)
- Il est simple à démontrer que l'EQM peut être exprimée comme une combinaison de biais et la variance :

$$EQM(T) = Var(T) + [b(T)^2]$$

Vraisemblance

- Pour une valeur p connue, on peut exprimer la probabilité de n'importe quelles données possibles
- En revanche, on peut *considérer les observations comme connues* et considérer la probabilité en fonction du paramètre inconnu p
- La fonction de probabilité vue de cette façon est appelée la **vraisemblance**

Définition de la vraisemblance

- **Définition** : Soit $x \sim f(x;\theta)$. La **vraisemblance** et **log vraisemblance** sont :

$$L(\theta) = f(x;\theta), \quad \ell(\theta) = \log L(\theta),$$

considérés comme des fonctions du paramètre θ.

- Soient $x = (x_1, \ldots, x_n)$ une réalisation des VAs $X_1, \ldots, X_n$. Alors

$$L(\theta) = f(x;\theta) = \prod_{j=1}^{n} f(x_j;\theta), \quad \ell(\theta) = \sum_{j=1}^{n} \log f(x_j;\theta),$$

où $f(x_j;\theta)$ est la loi de x_j.

- **À NOTER :** log = log base e = log naturel

Estimation par maximum de vraisemblance

- Une méthode d'estimation intuitive est **l'estimation par maximum de vraisemblance**
- Par exemple, l'estimateur le plus 'évident' p est $\hat{p} = X/n$ se révèle être **l'estimateur du maximum de vraisemblance (EMV / MLE)**
- En général, l'EMV est la valeur qui rend la probabilité aussi grande que possible – c'est la valeur qui *rend les données observées le plus probable*
- La manière habituelle de trouver l'EMV : le calcul – trouver la dérivée de la fonction de (log) vraisemblance, annuler et résoudre :

$$\frac{d \log L(\hat{\theta})}{d\theta} = 0, \quad \frac{d^2 \log L(\hat{\theta})}{d\theta^2} < 0$$

- (Cette méthode ne fonctionne pas dans tous les cas)
- Nous supposons que la première équation a une solution *unique* (ce n'est pas toujours vrai dans la réalité)

EMV, cont

- L'EMV $\hat{\theta}$ remplit la condition

$$L(\hat{\theta}) \geq L(\theta) \quad \text{pour toute } \theta,$$

  ce qui équivaut à $\log L(\hat{\theta}) \geq \log L(\theta)$, car les valeurs maximales de $L(\theta)$ et $\log L(\theta)$ sont obtenues *à la même valeur* θ
- L'EMV peut :
  - exister et être unique,
  - ne pas être unique, ou
  - ne pas exister
- Dans la pratique, il est normalement nécessaire d'utiliser des algorithmes numériques pour obtenir $\hat{\theta}$ et $d^2 \log L(\hat{\theta})/d\theta^2$

Avantages/désavantages de la méthode

- Pour un échantillon 'suffisamment grand', l'EMV est :
  - non-biaisé
  - consistent
  - efficace (EQM minimal ; donc au moins puissant que l'estimateur EMM)
  - *normalement distribué*
  - donc, pratique pour l'inférence statistique
- En revanche, l'EMV :
  - pourrait être très biaisé si la taille de l'échantillon est petite
  - pourrait être très compliqué d'évaluer (il faut le faire numériquement)

Information (statistique)

- L'*information observée* $J(\theta)$ et l'*information espérée* (aussi appelée *Fisher information*) $I(\theta)$ sont :
  - $J(\theta) = \dfrac{-d^2\ell(\theta)}{d\theta^2}$
  - $I(\theta) = E\{J(\theta)\} = E\left\{\dfrac{-d^2\ell(\theta)}{d\theta^2}\right\}$
- Elles sont des mesures de la *courbature* de $-\ell(\theta)$:

  plus les valeurs de $J(\theta)$ et $I(\theta)$ sont *grandes*, plus $\ell(\theta)$ et $L(\theta)$ sont *concentratés*

Propriétés de l'EMV

- **Convergent** : $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1, \forall \epsilon > 0$
- **Invariance** : si $\hat{\theta}$ est l'EMV pour le paramètre θ, alors $h(\hat{\theta})$ est l'EMV pour le paramètre $h(\theta)$
- **Asymptotiquement sans biais** : $b(\theta) \to 0$ lorsque $n \to \infty$ (pour les échantillons 'petits' l'EMV pourrait être biaisé)
- **Efficacité asymptotique optimale** : aucun estimateur asymptotiquement sans biais peut avoir une variance plus petite que celle de l'EMV
- **Normalité asymptotique** : la distribution de $\hat{\theta}_n$ lorsque $n \to \infty$ est *la distribution normale* ; cela nous donne une base pour la statistique inferentielle à partir de l'EMV (p. ex. IC)
- **IC approximatif** (niveau $1-\alpha$) pour θ : $\hat{\theta} \pm z_{1-\alpha/2} \Big/ \sqrt{J(\hat{\theta})}$