GC – Probabilités et Statistique

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

Cours 6

- VAs simultanées
- VAs indépendantes
- Somme des VAs indépendantes
- Distribution d'échantillonnage d'une statistique T
- Théorème Central Limite (TCL)
- Estimation par intervalle de confiance (IC)

Révision : VAs

- **VA discrète** :
  1. loi de probabilité : $p(x) = P(X = x)$
  2. fonction de répartition : $F(x) = P(X \leq x) = \sum_{i \leq x} p(i)$
- **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 :**
  - $F(x) = P(X \leq x) = \int_0^x f(u)\, du$
  - $F(-\infty) = 0$
  - $F(\infty) = 1$

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*

Fonction de répartition marginale

- La **fonction de répartion marginale** pour une VA est la fonction de répartion de cette VA seule, *sans égard* aux autres VAs

- La fonction de répartition de X est obtenue de la fonction de répartition conjointe de X et Y :

$$\begin{aligned}
F_X(a) &= P(X \leq a) && \text{[définition]} \\
&= P(X \leq a, Y < \infty) && \text{[cdf conjointe]} \\
&= P(\lim_{b->\infty} X \leq a, Y \leq b) && \text{[subst. en limite]} \\
&= \lim_{b->\infty} P(X \leq a, Y \leq b) && \text{[change l'ordre lim / P]} \\
&= \lim_{b->\infty} F(a,b) = F(a,\infty) && \text{[définition]}
\end{aligned}$$

- De manière similaire, on trouve la fonction de répartition de Y, $F_Y(b) = F(\infty, b)$

Loi discrète conjointe

- Au cas où X et Y sont *deux VAs discrètes*, on définit la **loi de probabilité simultanée** (ou **conjointe**) :

$$p(x, y) = P(X = x, Y = y)$$

- La **loi marginale** de X est obtenue de la loi conjointe $p(x, y)$:

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_{\underline{y} : p(x,y)>0} p(x, y), \end{aligned}$$

- *C.-à-d.*, la loi marginale de X est obtenue en additionnant la loi conjointe sur *toutes les valeurs possibles de* Y
- La **loi marginale** de Y est trouvée d'une manière similaire : $p_Y(y) = \sum_{\underline{x} : p(x,y)>0} p(x, y)$

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 :

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

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 :

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

Exemple

Exemple 6.1 La densité conjointe de X et Y est donnée par $f(x,y) = 2e^{-x}e^{-2y}$, $0 < x < \infty$, $0 < y < \infty$ ($f(x,y) = 0$ sinon).
[Astuce : trouver les bornes / limites d'intégration]

(a) $P(X > 1, Y < 1) =$

(b) $P(X < Y) =$

(c) $P(X < a) =$

Variables aléatoires indépendantes

- On a déjà étudié la notion de l'indépendance des *événements*
- Maintenant, on considère l'indépendance des *variables aléatoires*
- Les VAs X et Y sont dites **indépendantes** si, pour tout choix d'une paire d'ensembles A et B de nombres réels, on a

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

- En d'autres termes, X et Y sont indépendantes si, quels que soient A et B, les événements $X \in A$ et $Y \in B$ sont indépendants

Variables aléatoires indépendantes

- **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 $A_1, A_2, \ldots, A_n$,

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

Exemple

Exemple 6.2

(a) Soit la densité conjointe de X et Y
$f(x,y) = 6e^{-2x}e^{-3y},\ 0 < x < \infty,\ 0 < y < \infty$ ($f(x,y) = 0$ sinon). Est-ce que X et Y sont indépendantes *? ?*

(b) Soit la densité conjointe de X et Y
$f(x,y) = 24xy,\ 0 < x < 1,\ 0 < y < 1,\ 0 < x + y < 1$ ($f(x,y) = 0$ sinon). Est-ce que X et Y sont indépendantes *? ?*

Exemple

Exemple 6.3 Si X et Y sont VAs de Poisson indépendantes, $X \sim Pois(\lambda_1), Y \sim Pois(\lambda_2)$, trouver la distribution de $X + Y$.

Solution L'événement $X + Y = n$ est l'union disjointe des événements $(X = k, Y = n - k)$ pour $k = 0, 1, \ldots, n$; donc

$$\begin{aligned} P(X + Y = n) &= \sum_{k=0}^{n} P(X = k, Y = n - k) \\ &= \sum_{k=0}^{n} P(X = k) P(Y = n - k) \\ &= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \end{aligned}$$

Solution, cont.

$$
\begin{aligned}
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n
\end{aligned}
$$

Il s'agit de quelle distribution ? ?

PAUSE

Quel est votre QI ? ?

Test de QI

1. Le Père Noel, existe-t-il ?
2. Qui est le meilleur footballeur du monde ?
3. Évaluer :

$$\int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Échantillonage

- Le but d'une étude statistique est d'obtenir des connaissances sur l'ensemble de la *population*, c.-à-d. **l'estimation d'un paramètre**
- Puisque un dénombrement complet de la population est très souvent pratiquement impossible, il faut d'autres moyens plus pratique
- $\Rightarrow$ Un **échantillonnage** consiste à choisir parmi les éléments de la population un certain nombre d'unités pour lesquelles nous obtiendrons des observations (données)
- Nos données sont considérées comme la suite d'un processus aléatoire : si la collecte de données ont été répétées, le résultat serait probablement différent, qui peuvent influer sur les conclusions tirées sur la base de données
- C.-à-d., nos conclusions sont sujettes à la *variation aléatoire*

Distribution d'échantillonnage

- Une **statistique** est une *fonctionne des données*
- La distribution (exacte) d'une statistique T est appelée la **distribution d'échantillonnage**
- La distribution d'échantillonnage d'une statistique est déterminé par le *programme d'échantillonnage* – c'est ce qui définit la probabilité associée à chaque échantillon possible

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.

Exemple

Exemple 6.4 Un ascenseur (imaginaire) a une capacité de charge maximale de 3,6 tonnes métriques (3600 kg). Une certaine population a un poids moyen de 70 kg, avec un écart-type de 16 kg.

(a) Quelle est la probabilité qu'un échantillon aléatoire de 49 (! !) personnes de cette population dépasse la capacité de l'ascenseur *? ?*

(b) Trouver le nombre maximum de personnes que l'ascenseur devrait accueillir afin que la chance d'être surchargée est au plus 1% **...**

Estimation – Intervalle de confiance

- Il n'est pas très instructif de donner une valeur unique pour estimer la valeur du paramètre (*estimation ponctuelle*)
- Il est également intéressant d'avoir une idée de **l'ampleur probable de l'erreur**
  - **l'erreur standard** (ES) : l'estimation de l'écart-type de la distribution d'échantillonnage de la statistique
  - *p. ex.* $SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, estimé par $ES = \frac{s}{\sqrt{n}}$
- Une autre façon de présenter l'estimation est sous la forme d'un **intervalle de confiance** (IC)

Développement d'un IC pour la moyenne de la population

les détails ne figurent pas sur l'examen

- **TCL** : la *distribution d'échantillonage de la moyenne de l'échantillon* est approximativement normale, de moyenne μ et d'écart-type $\sigma/\sqrt{n}$
- Cela signifie qu'il y a une probabilité de 95% que la (**VA**) $\overline{X}$ se trouve dans un rayon de 1.96 de la vraie moyenne de la population μ :

$$P(\mu - 1.96\sigma/\sqrt{n} \leq \overline{X} \leq \mu + 1.96\sigma/\sqrt{n}) = 0.95.$$

- Si la VA $\overline{X}$ est dans le rayon de $1.96\sigma/\sqrt{n}$ de μ, alors μ est dans le rayon de $1.96\sigma/\sqrt{n}$ de $\overline{X}$
- Donc les probabilités de ces événements sont les mêmes :

$$P(\overline{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

IC pour la moyenne de la population, cont

- L'intervalle $(\overline{x} - 1.96\,\sigma/\sqrt{n},\ \overline{x} + 1.96\,\sigma/\sqrt{n})$ basé sur *la moyenne (observée) de l'échantillon* $\overline{x}$ est appelé un **intervalle de confiance (95%)** pour μ
- La probabilité 0.95 (95%) est appelée le **niveau (ou indice ou taille) de confiance**
- Lorsque l'écart-type de la population σ est inconnu (le cas le plus fréquent), on l'estime en utilisant l'écart-type de l'échantillon s
- $\Rightarrow$ Puisque 1.96 $\approx$ 2, alors on pourrait exprimer l'IC de 95% comme : $\overline{x} \pm 2\,\dfrac{s}{\sqrt{n}}$

Exemple – IC (mécanique)

Exemple 6.5 Supposons que nous voulons estimer le revenu moyen d'une population particulière. Un échantillon aléatoire de taille $n = 16$ soit prise ; $\overline{x} = \$23,412$, $s = \$2000$.

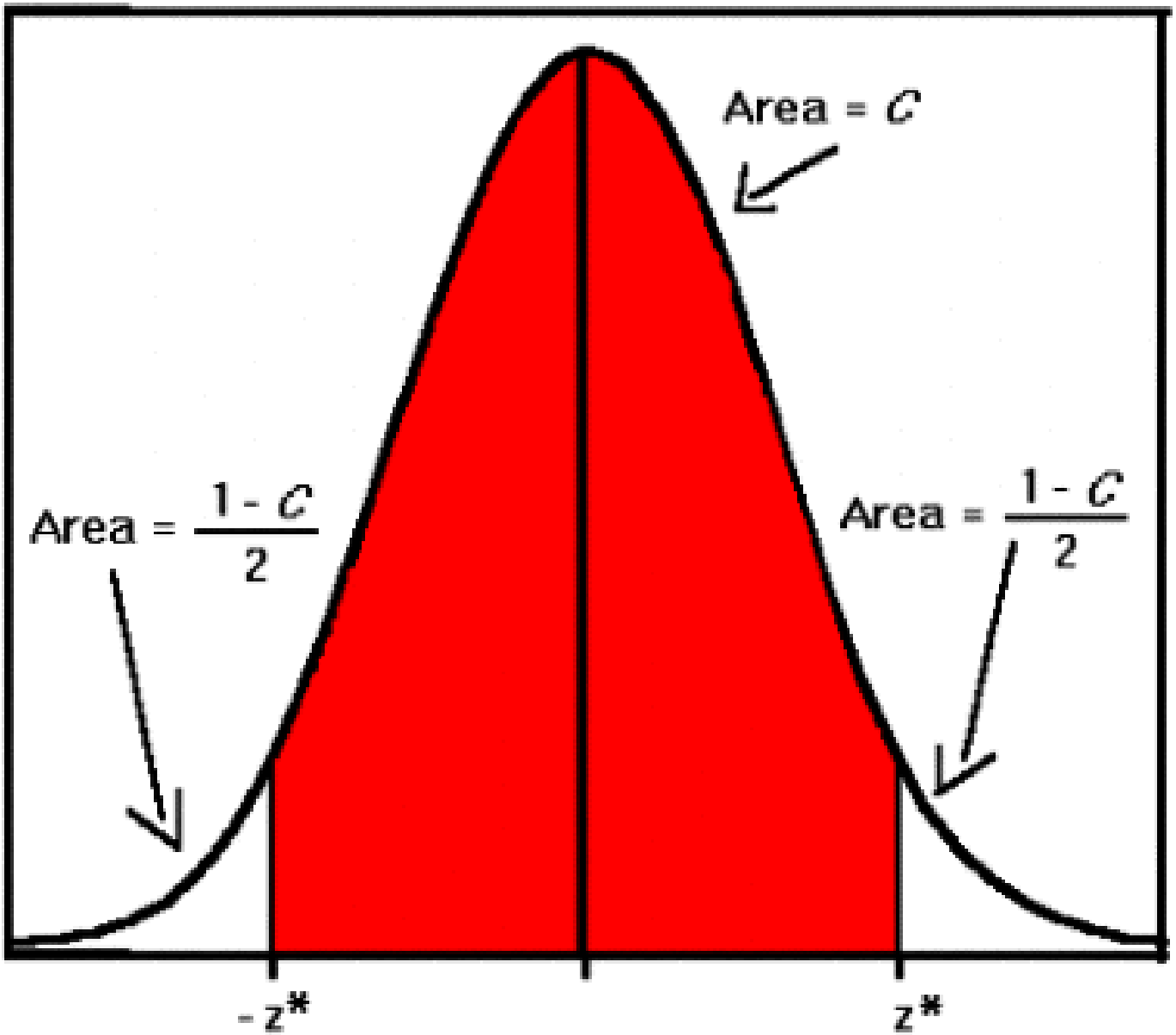
(a) Estimer la moyenne de la population μ

(b) Donner un IC approximative de niveau 95% pour μ **...**

Niveau de confiance ≠ 95% ? ?

- Le niveau de confiance le plus couramment utilisé est de 95% ou 90%, mais il n'y a pas de règle disant qu'il faut utiliser ces niveaux
- Le niveau pourrait être une valeur inférieure à 100%, en fonction de la 'confiance' que vous voulez avoir que la vraie valeur du paramètre sera contenue dans un intervalle de fait selon la procédure décrite ci-dessus
- S'il y a un changement du niveau de confiance, *la valeur z (1.96 pour un IC de 95%) change aussi*

Illustration



Autre exemple – IC mécanique

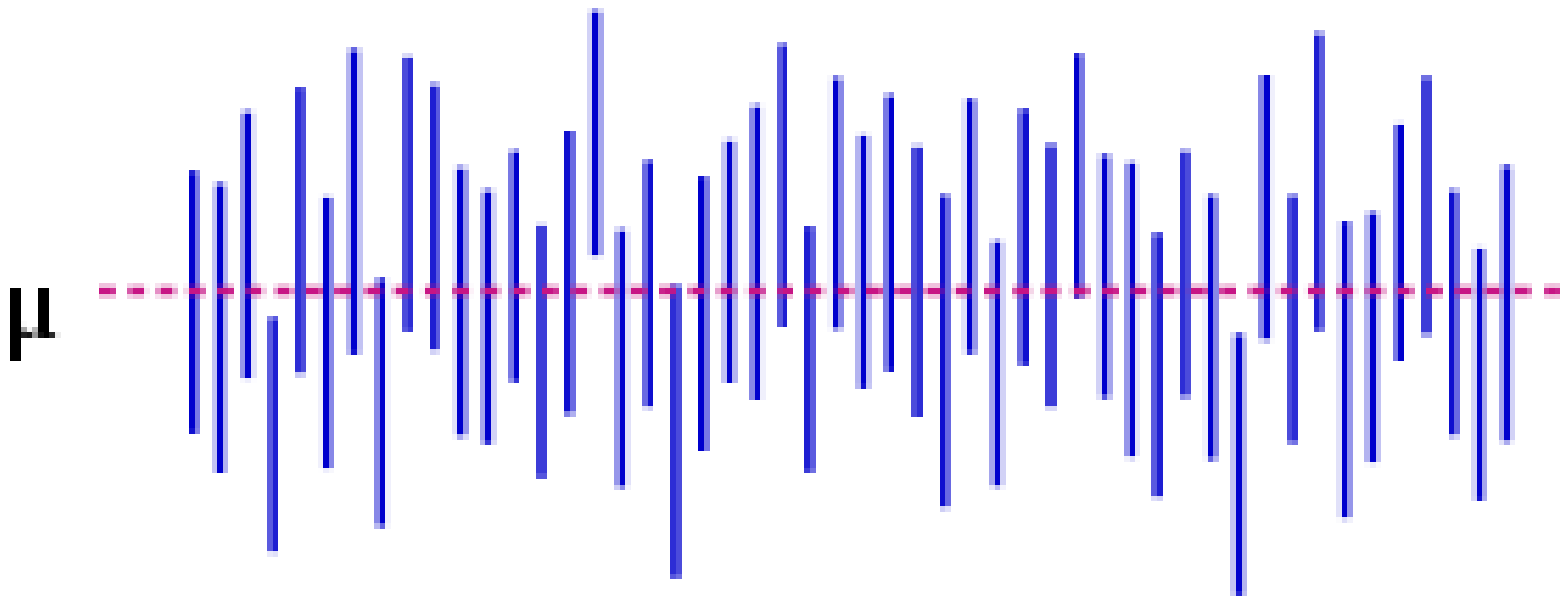
Exemple 6.6 Supposons que nous voulons estimer $\mu =$ la note d'examen moyenne dans une large population. Un échantillon aléatoire de taille 25 est obtenu ; $\overline{x} = 72$, $s = 15$.

Donner un IC approximative de 90% pour μ.

Interprétation d'un IC

- Il est tentant **MAIS FAUX ! ! ! ! !** à croire que pour un IC 95% particulier, la probabilité est de 95% que le vraie valeur du paramètre soit dans l'IC – *théorie fréquenciste prob.*
- Toutefois, dans cette interprétation, le paramétre de la population n'est **PAS** une VA, mais plutôt qu'il s'agit **d'une constante** dont la valeur nous est inconnue
- *Avant* l'échantillonnage, la moyenne de l'échantillon $\overline{X}$: VA
- *Après* l'échantillonnage, **il n'y a plus de VA**
- Le paramètre est soit *dans*, soit *hors de* cet intervalle particulier
- Le 95% se réfère à la **procédure d'échantillonnage** :
  Si l'on ferait la procédure entière plusieurs fois (Obtention d'un échantillon aléatoire et en faisant un IC de 95%), alors **environ 95% des intervalles fait de cette façon contiendrait la vraie valeur du paramètre**

Illustration



Autre exemple

Exemple 6.7 Dans une année donné il y a 100.000 recrues de l'armée. Le poids moyen est de 75 kilos, avec un écart-type de 15 kilos.

(a) Si possible et approprié, donnez un IC de niveau (indice) 95% pour le poids moyen des recrues de l'armée cette année-là. *Expliquer.*

(b) Supposons maintenant que la moyenne de poids dans la population est inconnue, mais un échantillon aléatoire de taille 400 est pris, et le poids moyen de l'échantillon est de 75 kilos avec l'écart-type de 15 kilos. Peut-on faire l'IC maintenant ?

(c) Est-il nécessaire de supposer que la distribution des poids de recrues de l'armée est normale ? *Expliquer.*

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