Chapitre 2

Echantillonnage

Introduction

Le mot statistique désigne à la fois un ensemble de données et l'activité qui consiste à les recueillir, les traiter et les interpréter.
Faire de la statistique c'est étudier un ensemble d'objets équivalents appelés individus ou unités statistique, sur lesquels on observe des caractéristiques appelées variables.
Recensement : observer toutes les unités statistiques d'une population finie.
Echantillon : Partie de la population.
Sondage : Etudier les unités de l'échantillon.

La démarche statistique

Statistique descriptive : Son but est de synthétiser, résumer l'information contenue dans les données. Elle utilise les tableaux statistiques et les représentations graphiques.

Statistique inférentielle : Son but est d'étendre les propriétés constatées sur l'échantillon à toute la population et valider des hypothèses a priori. Le calcul des probabilités joue un rôle important.

2.1 Position du problème

On veut à partir d'un échantillon, déduire des informations sur la population. Le problème qui se pose alors est comment choisir une partie de la population qui repro-

duit le plus fidèlement possible ses caractéristiques . C'est le problème d'échantillonnage (sondage).

2.1.1 Avantages de l'échantillonnage

- Impossibilité d'étudier toute la population lorsqu'elle est infinie.
- Le coût : Le choix d'un échantillon est de moindre coût qu'un recensement.
- Le temps : la rapidité nécessaire de certaines prises de décisions empêche le recours à un recensement.

2.1.2 Choix de l'échantillon

Il ya 3 grands modes de tirage d'un échantillon :

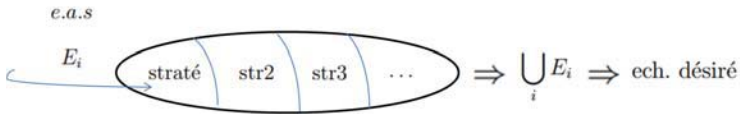
- Tirage empirique.
- Tirage aléatoire.
- Tirage artificiel (simulation).

a) Echantillonnage empirique : Utilise des connaissances préalables qu'on a sur la population. La méthode la plus utilisée est la méthode des quotas. Cette méthode se base sur la construction d'un échantillon de taille n dans lequel les proportions des individus sont égales à celles de la population. Une fois ces quotas déterminés, il faut les respecter dans le choix de l'échantillon.

b) Echantillonnage aléatoire : L'échantillon est dit aléatoire si chaque individu de la population a une probabilité connue d'appartenir à l'échantillon.

* Echantillon aléatoire simple : Chaque individu a la même probabilité d'être choisi (la loi uniforme discrète).
* Echantillon aléatoire non simple : Tient compte de certain facteurs de pondération. On distingue deux types de tirage :
  1. Tirage exhaustif : Sans remise (les observations ne sont pas indépendantes).
  2. Tirage non exhaustif : Avec remise bernoullien (les observations sont toutes indépendantes).

* Echantillonnage stratifié : On subdivise la population en sous ensembles (strates) relativement homogènes. On extrait de chaque strate un échantillon aléatoire simple (e.a.s). La réunion de tous ces échantillons constitue l'échantillon désiré.



Cette méthode a le but d'améliorer la précision de l'estimation.

* L'échantillonnage par grappes : On subdivise la population en sous grappes, on tire un échantillon aléatoire de grappes. L'échantillon désiré est constitué de tous les individus de chaque grappe.

2.2 Notions fondamentales

2.2.1 Population de référence

C'est la totalité des éléments pris en considération, et sur lesquels on désire obtenir des informations.

2.2.2 Echantillon

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ une population de taille N. Soit X le caractère que l'on voudrait étudier sur cette population. Avec l'échantillon aléatoire simple : soit X_k le résultat aléatoire du $k^{\text{ième}}$ tirage, c'est une variable aléatoire qui suit la même loi que X. On note x_k le résultat du $k^{\text{ième}}$ tirage et on note $(X_1, X_2, ..., X_n)$ le résultat aléatoire de ces n tirages.

Définition 2.1. $(X_1, X_2, ..., X_n)$ est un n-uplet de variables aléatoires indépendantes de même loi (celle de X). Il est appelé n-échantillon aléatoire simple.

Définition 2.2. L'unique réalisation $(x_1, x_2, ..., x_n)$ est appelée ensemble des valeurs observées.

Définition 2.3. Une statistique Y sur un échantillon $(X_1, X_2, ..., X_n)$ est une variable aléatoire fonction mesurable des X_i :

$$Y = f(X_1, X_2, ..., X_n) \text{ prend la valeur } f(x_1, x_2, ..., x_n)$$

2.2.3 Distribution d'un échantillon

Soit $(X_1, X_2, ..., X_n)$ un échantillon aléatoire simple issu d'une variable aléatoire X de densité de probabilité f. La distribution (loi) de probabilité du n-échantillon aléatoire simple est :

$$\begin{aligned} g(x_1, x_2, ..., x_n) = g(\underline{x}) &= f_{X_1}(x_1) f_{X_2}(x_2) ... f_{X_n}(x_n) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i) = (f_X(x))^n \end{aligned}$$

g est appelée *vraisemblance du n-échantillon.*

Exemple 2.1. *1. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une variable aléatoire $X \rightsquigarrow \mathcal{N}(0,1)$*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \qquad x \in \mathbb{R}.$$

$$g(\underline{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \frac{1}{(\sqrt{2\pi})^n} e^{\frac{-1}{2}\sum_{i=1}^{n} x_i^2}, \qquad x_i \in \mathbb{R} \quad \forall i = \overline{1,n}.$$

2. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une variable aléatoire $X \rightsquigarrow \mathcal{B}(\theta)$, $f(x) = \theta^x (1-\theta)^{1-x}$ avec $x \in \{0,1\}$.

La vraisemblance de cet échantillon est :

$$g(x) = \prod_{i=1}^{n} f_{X_i}(x_i) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}.$$

2.2.4 Fonction de répartition empirique d'un échantillon

Définition 2.4. On appelle fonction de répartition empirique d'un échantillon (e.a.s avec remise) $F_n(x)$ la proportion des n variables aléatoires $X_1, X_2, ..., X_n$ qui sont inférieures à x. C'est donc une fonction aléatoire (v.a pour tout x) dont les réalisations sont des fonctions

en escalier de sauts égaux à $1/n$.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(X_i<x)} = \begin{cases} 0, & \text{si } x < x_1\,; \\ \frac{1}{n}, & \text{si } x_1 \leq x < x_2\,; \\ \vdots & \\ \frac{(i-1)}{n}, & \text{si } x_{i-1} \leq x < x_i\,; \\ \vdots & \\ 1, & \text{si } x \geq x_n. \end{cases}$$

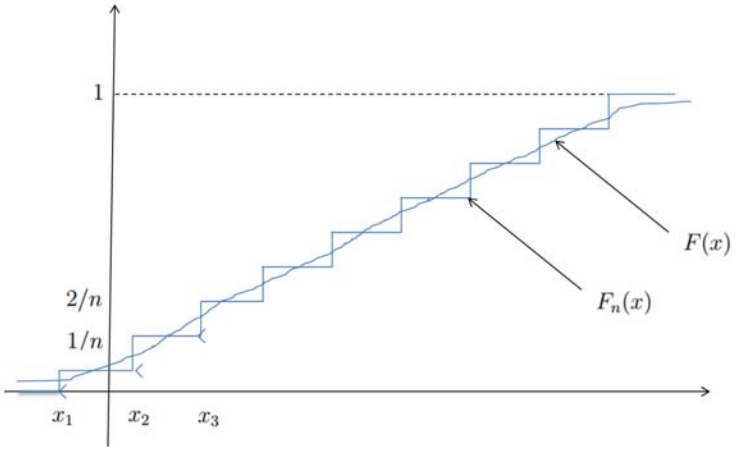


FIG. 2.1 – Graphe de la fonction de répartition empirique

Convergence de $F_n(x)$ vers $F(x)$

Théorème 2.1. *Soit $(X_1, X_2, ..., X_n)$ un n-échantillon de fonction de répartition empirique $F_n(x)$ et $F(x)$ fonction de répartition de X (variable aléatoire parente).*
Alors : $F_n(x) \xrightarrow{p.s} F(x)$. i.e : $\mathbb{P}\left(x : \lim_{x\to\infty} F_n(x) = F(x)\right) = 1$

Démonstration. $F_n(x)$ étant une moyenne empirique de variable aléatoire réelle indépendante (puisque les X_i le sont), d'après la loi forte des grands nombres : $F_n(x) \xrightarrow{p.s} \mathbb{E}[\mathbb{1}_{(X_i<x)}] = \mathbb{E}[\mathbb{1}_{(X<x)}] = F(x)$.

□

Convergence en loi de $F_n(x)$

On a : $\mathbb{1}_{(X_i<x)}$ est une variable aléatoire de Bernoulli $\mathcal{B}(F(x))$,

$$\mathbb{P}(F_n(x) = k/n) = \mathbb{P}(nF_n(x) = k) = C_n^k(F(x))^k(1 - F(x))^{n-k},$$

donc $nF_n(x) \rightsquigarrow \mathcal{B}(n, F(x))$. Alors d'après le théorème central limite :

$$\frac{nF_n(x) - nF(x)}{\sqrt{nF(x)(1 - F(x))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

i.e

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Théorème 2.2. *(Glivenko-Contelli)*
La convergence de F_n vers F est presque sûrement uniforme, i.e :

$$D_n = \sup_x |F_n(x) - F(x)| \xrightarrow{p.s} 0.$$

Théorème 2.3. *(Kolmogorov)*

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n < y) = K(y) = \sum_{-\infty}^{+\infty} (-1)^k e^{-2k^2y^2}.$$

Ce théorème signifie que la distribution asymptotique de la variable aléatoire D_n est connue et ne dépend pas de la variable de départ X, et permet de calculer des limites pour les valeur de D_n. La loi exacte de la variable D_n a été tabulée.

2.2.5 Statistique d'ordre d'un échantillon

Soit $(X_1, X_2, ..., X_n)$ un n-échantillon aléatoire simple issu d'une variable aléatoire X, les réalisation $x_1, x_2, ..., x_n$ peuvent être réordonnées en $y_1, y_2, ..., y_n$ où $y_1 < y_2 < ... < y_n$. Les y_i constituent une permutation particulière des x_i.
Les y_i sont des réalisations du n-uplet de variables aléatoires $(Y_1, Y_2, ..., Y_n)$ qui constitue l'échantillon ordonné de X.
Soit F (resp. f) la fonction de répartition (resp. la densité) de la variable aléatoire X. Soit

H_k (resp. h_k) les fonctions de répartition (resp. les densités) de Y_k.

On cherche la loi de $Y_1 = \inf\limits_{1\leq i\leq n} X_i$ et $Y_n = \sup\limits_{1\leq i\leq n} X_i$.

a) Loi de $Y_1 = \inf\limits_{1\leq i\leq n} X_i$:

On a :

$$\begin{aligned} H_1(y) = \mathbb{P}(Y_1 < y) &= \mathbb{P}(\inf X_i < y) \\ &= 1 - \mathbb{P}(\inf X_i \geq y) \\ &= 1 - \mathbb{P}(\prod_{i=1}^{n}(X_i < y)) \\ &= 1 - \prod_{i=1}^{n}\mathbb{P}(X_i < y) \\ &= 1 - \prod_{i=1}^{n}[1 - F(y)] \end{aligned}$$

$$H_1(y) = 1 - [1 - F(y)]^n.$$

$$\text{D'où : } h_1(y) = n(1 - F(y))^{n-1} f(y).$$

b) Loi de $Y_n = \sup\limits_{1\leq i\leq n} X_i$:

$$H_n(y) = \mathbb{P}(\sup X_i < y) = \prod_{i=1}^{n}\mathbb{P}(X_i \leq y),$$

$$H_n(y) = (F(y))^n,$$

$$\text{et } h_n(y) = n(F(y))^{n-1} f(y).$$

Remarque 2.1. Ces deux lois servent en particulier à détecter les valeurs aberrantes de l'échantillon : valeurs trop grandes ou trop petites.

2.2.6 Moments d'un échantillon

Définition 2.5. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon aléatoire issu d'une variable aléatoire X et r un nombre entier ($r \in \mathbb{N}$). On appelle *moment d'ordre r* de l'échantillon et on note m'_r, la quantité :

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

Théorème 2.4. *Soit $(X_1, X_2, ..., X_n)$ un n-échantillon aléatoire simple issu d'une variable aléatoire X, alors $\mathbb{E}(m'_r) = m_r$, où m_r est le moment d'ordre r de la variable aléatoire X :* $\mathbb{E}(m'_r) = m_r = \mathbb{E}(X_i^r)$. *En effet :*

$$\mathbb{E}(m'_r) = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^r] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^r] = m_r.$$

Moyenne empirique d'un échantillon

Définition 2.6. Soit $m'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$ le moment d'ordre r de l'échantillon. Pour $r = 1$: $m'_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ est la moyenne du n-échantillon aléatoire simple $(X_1, X_2, ..., X_n)$ on la note $\overline{X}$,

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

propriétés :

1. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une variable aléatoire X de moyenne μ et de variance σ^2. Alors $\mathbb{E}(\overline{X}) = \mu$ (on dit que $\overline{X}$ est une statistique sans biais).
   En effet : $\mathbb{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$.
2. $Var(\overline{X}) = \frac{\sigma^2}{n}$.
   En effet : $Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$.
3. $\overline{X} \xrightarrow{p.s} \mu$. (Loi forte des grands nombres).

Variance empirique d'un échantillon

Soit $(X_1, X_2, ..., X_n)$ un n-échantillon aléatoire simple issu d'une variable aléatoire X de moyene μ et de variance σ^2.
On appelle variance empirique de l'échantillon la quantité :

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

Propriétés :

1. $\mathbb{E}(S^2) = \frac{\sigma^2(n-1)}{n}$,
2. $S^2 \longrightarrow^{p.s} \sigma^2$,
3. $Var(S^2) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4]$, où $\mu_4 = \mathbb{E}(X - \overline{X})^4$.

Démonstration de (1) : $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Décomposons S^2 :

$$\begin{aligned} X_i - \mu &= X_i - \overline{X} + \overline{X} - \mu \\ \sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 + 2\sum_{i=1}^{n}(X_i - \overline{X})(\overline{X} - \mu) + \sum_{i=1}^{n}(\overline{X} - \mu)^2 \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 + (\overline{X} - \mu)^2. \end{aligned}$$

D'où :

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

$$\begin{aligned} \mathbb{E}(S^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \overline{X})^2 - \mathbb{E}(\overline{X} - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} Var(X_i) - Var(\overline{X}) \\ &= \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{\sigma^2}{n}(n-1). \end{aligned}$$

D'où :

$$\mathbb{E}(S^2) = \frac{\sigma^2}{n}(n-1).$$

Remarque 2.2. $\mathbb{E}(S^2) \neq \sigma^2$, on dit que S^2 est une statistique biaisée, son biais vaut : $\frac{\sigma^2}{n}$.

Autre démonstration :

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2)-\mathbb{E}(\overline{X}^2) \\
&= \frac{1}{n}n\mathbb{E}(X_i^2)-\mathbb{E}(\overline{X}^2) \\
&= \mathbb{E}(X_i^2)-\mathbb{E}(\overline{X}^2)=\sigma^2\frac{n-1}{n}.
\end{aligned}$$

Cas de grands échantillons

Théorème 2.5. *Pour n assez grand :*

$$\bar{X} \xrightarrow{\mathcal{L}} \mathcal{N}(\mu,\frac{\sigma^2}{n}) \quad i.e \quad \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

(Ceci résulte du théorème central limite)

Cas de petits échantillons

Le théorème central limite ne s'applique pas, alors on fait l'hypothèse : $(X_1,...,X_n)$ est issu d'une variable aléatoire $X \rightsquigarrow \mathcal{N}(\mu,\sigma^2)$.

Alors : $\overline{X} \longrightarrow \mathcal{N}(\mu,\sigma^2/2)$.

D'aprés la décomposition de S^2 :

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2$$

En divisant par σ^2 :

$$\begin{aligned}
\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 &= \sum_{i=1}^{n}(\frac{X_i-\overline{X}}{\sigma})^2+\frac{n}{\sigma^2}(\overline{X}-\mu)^2 \\
&= \frac{nS^2}{\sigma^2}+\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2.
\end{aligned}$$

On a : $\frac{X^i-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1) \Rightarrow \left(\frac{X^i-\mu}{\sigma}\right)^2 \rightsquigarrow \chi_1^2$.
D'où : $\sum_{i=1}^{n} \left(\frac{X^i-\mu}{\sigma}\right)^2 \rightsquigarrow \chi_n^2$, (Comme somme de n carrés de variables aléatoires indépendantes normales centrées réduites).

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1) \Rightarrow \left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \rightsquigarrow \chi_1^2$$

D'où, on en déduit :

$$\frac{nS^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

$$\text{i.e} \quad S^2 \rightsquigarrow \frac{\sigma^2}{n}\chi_{n-1}^2$$

$$\mathbb{E}(S^2) = \frac{\sigma^2}{n}(n-1) \quad \text{et} \quad Var(S^2) = \frac{\sigma^4}{n^2}2(n-1).$$

2.3 Modèle statistique

Définition 2.7. On appelle *modèle statistique paramétrique* tout couple $(\chi, (\mathbb{P}_\theta)_{\theta\in\Theta})$ où χ est un ensemble dit espace des observations et $(\mathbb{P}_\theta)_{\theta\in\Theta}$ est une famille de lois de probabilités définies sur une tribu $\mathcal{A}$ fixée de χ et Θ espace des paramètres.

Exemple 2.2. *Dans une production de* 10000 *pièces mécaniques, une certaine proportion* θ *de ces pièces est défectueuses. On prélève au hasard* 20 *pièces.*
Un modèle courant de cette situation est le modèle binômiale $(\chi, (\mathbb{P}_\theta)_{\theta\in\Theta})$*, où* $\Theta = [0,1]$*,* $\chi = \{0,1,...,20\}$ *et* $\mathbb{P}_\theta$ *est la loi* $\mathcal{B}(20,\theta)$*,* θ *est inconnu, donc on a affaire à un problème de statistique inférentielle.*

Remarque 2.3. Si χ est fini ou dénombrable : $\mathcal{A} = \mathcal{P}(\chi)$, si $\chi = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$.

Définition 2.8. Soit $(\chi, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique. On appelle statistique sur ce modèle toute application φ indépendante de θ, mesurable relativement à la tribu $\mathcal{A}$ considérée sur χ à valeurs dans un certain espace Υ muni d'une tribu $\mathcal{B}$.
Le modèle $(\Upsilon, (q_\theta)_{\theta\in\Theta}) = (\Upsilon, (\varphi(\mathbb{P}_\theta))_{\theta\in\Theta})$, où $\forall\theta \in \Theta : \quad q_\theta = \varphi(\mathbb{P}_\theta)$ est la loi image par φ de la loi $\mathbb{P}_\theta$ est dit modèle image par φ du modèle $(\chi, (\mathbb{P}_\theta)_{\theta\in\Theta})$.

Reprenons l'exemple précédent :

On a chacun des prélèvement est régi par une loi de Bernoulli π_θ de paramètre θ sur l'ensemble $\{0,1\}$ (1 : pièce défectueuse, 0 : pièce non défectueuse). En supposant l'indépendance entre les prélèvements (justifié par le grand effectif de la production).

Le résultat de l'expérience est alors un 20-uplet, soit $(x_1, x_2, ..., x_n)$ régi par la loi puissance $\pi_\theta^{\otimes 20}$ de la loi π_θ, le modèle sous-jacent est ainsi $\left(0, 1^{\otimes 20}, (\pi_\theta)_{\theta\in[0,1]}^{\otimes 20}\right)$ est définit un modèle de Bernoulli.

Le modèle binômial considéré s'en déduit en prenant $\forall\theta \in [0,1]$ par
$\varphi : (x_1, ..., x_{20}) \longmapsto \sum_{i=1}^{20} x_i$ l'image $\mathbb{P}_\theta$ de $\pi_\theta^{\otimes 20}$.

Remarque 2.4. φ joue un rôle de passage d'un modèle statistique vers un autre.

2.4 Exercices

Exercice 2.1. *Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon issu d'une v.a. X de loi* $\exp(1/\lambda)$ *représentant la durée de vie d'un appareil. $\lambda > 0$ inconnu.*

1. *A l'instant 0, on met en fonctionnement un appareil, on le remplace immédiatement par un autre, dès qu'il tombe en panne, et ainsi de suite on poursuit le processus. On observe la v.a. Z représentant le nombre de défaillances pendant un intervalle de temps $[0, a]$, ($a > 0$ fixé). Donner le modèle statistique associé.*
   *(**Indication** : $\mathbb{P}(Z = n) = \mathbb{P}(\sum_{i=1}^{n} X_i \leq a,\ \sum_{i=1}^{n+1} X_i > a)$.*
2. *A l'instant 0 on met en fonctionnement k appareils. Dès qu'un appareil tombe en panne, on le remplace. On observe toujours le nombre de défaillance dans l'intervalle $[0, a]$. Donner le modèle statistique associé.*