Chapitre 5 : Population, échantillon, statistiques, estimation

---

5.1 Population 1

5.2 Échantillon et statistique 1

5.3 Sources d'incertitudes 2

5.4 Estimateurs 3
5.4.1 Propriétés des estimateurs 3
5.4.2 Estimation ponctuelle 3

5.5 Estimation par intervalles de confiance 5
5.5.1 Divers intervalles de confiance pour des cas courants 6
5.5.2 Détermination de la taille d'un échantillon 7

Les chapitres 3 et 4 ont montré différents modèles probabilistes (loi de distribution) définies par quelques paramètres. Ces modèles sont sensés représenter le phénomène étudié (la population dans le jargon statistique). Souvent, les paramètres du modèle sont inconnus et doivent être estimés à l'aide de données. Les données recueillies constituent l'échantillon. L'échantillon sert à deux buts principaux : estimer les paramètres du modèle et valider le modèle lui-même. En effet, les probabilités prédites par le modèle ne devraient pas normalement trop s'écarter des fréquences correspondantes observées dans l'échantillon.

5.1 Population

La population est l'ensemble de tous les éléments identiques dont le modèle probabiliste est sensé décrire le comportement. La population est un concept plus ou moins abstrait, ayant une existence réelle ou virtuelle, de nature finie ou non. Par exemple, la population pourrait être l'ensemble des étudiants de Polytechnique, ou l'ensemble des blocs de $1m^3$ dans une digue, ou l'ensemble des éprouvettes de béton que l'on pourrait former en suivant la recette A. La population résulte souvent d'un choix délibéré de l'ingénieur (par conséquent non réfutable). Par exemple si l'on étudie la distribution d'un contaminant dans un sol donné, il faudra fixer selon un certain critère (ex. droit de propriété, limites naturelles, nature des sols rencontrés, etc.) les limites de la région d'étude. La population sera constituée de l'ensemble des unités de sol que l'on retrouve à l'intérieur de ces limites.

5.2 Échantillon et statistique

Définition : Un **échantillon aléatoire** de taille « n » d'une v.a. X est un ensemble $X_1,...X_n$ de v.a. *indépendantes* dont la distribution est *identique à celle de X*. X est la population et chaque Xi est une observation de X.

L'échantillon est un sous-ensemble de la population. On s'en sert normalement pour estimer les paramètres du modèle probabiliste s'appliquant à la population et pour effectuer la vérification du modèle.

Comme on l'a vu précédemment, plusieurs lois de distribution d'usage courant sont entièrement spécifiées par un ou deux paramètres. Aux paramètres du modèle devant s'appliquer à la population, on peut faire correspondre des quantités analogues calculées sur l'échantillon que l'on nomme statistiques. Concernant les statistiques, il faut distinguer la définition théorique qui est celle d'une fonction de l'échantillon, de la valeur particulière réalisée par un échantillon aléatoire donné.

Définition : Une **statistique** est une *fonction des observations d'un échantillon aléatoire* ne dépendant *d'aucun paramètre inconnu*. Par définition, une statistique est donc aussi une v.a. La distribution d'une statistique est appelée la distribution échantillonnale (ou d'échantillonnage).

Note : Les lois χ^2, Student, et Fisher sont des exemples courants de distribution de statistiques.

Paramètre ou statistique	Population (cas continu)	Échantillon aléatoire	Échantillon particulier
Espérance mathématique ou moyenne	$\mu = \int xf(x)dx$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \int (x-\mu)^2 f(x)dx$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Écart-type	$\sigma = \sqrt{\sigma^2}$	$S = \sqrt{S^2}$	$s = \sqrt{s^2}$
Médiane	$Méd = F^{-1}(0.5)$	Valeur « milieu » de l'échantillon	Valeur « milieu » de l'échantillon
Covariance	σ_{XY}	$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
Corrélation	$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$	$R_{xy} = \frac{S_{xy}}{S_x S_y}$	$r_{xy} = \frac{s_{xy}}{s_x s_y}$
Description de la distribution	Fonction de densité ou fonction de répartition		Histogramme ou courbe des fréquences cumulées

5.3 Sources d'incertitudes

On reconnaît trois grandes sources d'incertitude :

1- L'incertitude naturelle, intrinsèque, irréductible ou fondamentale. C'est l'incertitude liée au caractère probabiliste du modèle. C'est l'incertitude qui subsiste quand on a le bon modèle et les bons paramètres du modèle.
2- L'incertitude de nature statistique. Les paramètres du modèle étant estimés, il y a certainement une différence entre ceux-ci et les « vrais » paramètres du modèle.
3- L'incertitude sur le modèle lui-même. Par exemple, on suppose dans un certain contexte que la population suit une loi normale alors qu'en réalité une loi gamma pourrait être la « bonne » loi à appliquer.

Note : En génie civil, géologique et des mines, la plupart des données s'inscrivent dans le temps ou dans l'espace. Le modèle probabiliste doit pouvoir s'appliquer à la période observée ou au domaine observé. Il faut donc que le phénomène présente une certaine « homogénéité » statistique pour que le modèle décrive bien l'ensemble du phénomène (on parle alors de stationnarité du phénomène). Si les paramètres du modèle semblent bouger de façon notable dans le temps ou l'espace, il y a fort à parier que le modèle probabiliste sera relativement peu efficace partout. Dans ces cas là, il faut soit découper le domaine ou le temps en tranches plus homogènes (c'est rarement simple à réaliser) soit recourir à des modèles moins exigeants en terme de stationnarité (nous en verrons quelques exemples dans le chapitre sur les statistiques spatiales).

Exemple : On s'intéresse à prédire la fréquence de précipitations d'une certaine amplitude. On dispose des données du dernier siècle pour estimer les paramètres de la population. Si l'hypothèse du réchauffement climatique et des bouleversements sur le climat est réaliste, il y a danger à prédire

la fréquence des précipitations très fortes en utilisant l'ensemble des données historiques. Le modèle devrait pouvoir refléter l'évolution récente touchant cette fréquence.

5.4 Estimateurs

Certaines statistiques peuvent servir à estimer des paramètres d'une population. On dit alors que ce sont des estimateurs. Les estimateurs présentent un certain nombre de propriétés. Un même paramètre peut être estimé par deux estimateurs (statistiques) présentant des propriétés différentes. Le choix de l'un ou l'autre dépend des propriétés que l'on veut favoriser.

5.4.1 Propriétés des estimateurs

Biais : Un estimateur T d'un paramètre θ est sans biais si E[T]= θ. Le **biais** est E[T]- θ.

Un estimateur est asymptotiquement sans biais si $\lim(n \to \infty) \quad E[T] = \theta$.

Ex. $\overline{X}$ est sans biais pour μ.

Définition : On définit l'**erreur quadratique moyenne** de l'estimateur T d'un paramètre inconnu θ**:**

$$\text{E.Q.M.(T)} = E[(T-\theta)^2]$$

Note : E.Q.M.(T)= E[({T –E[T]}+{E[T]- θ })^2] = Var(T)+Biais(T)^2

Efficacité : Soit T_1 et T_2, deux estimateurs de θ. T_1 est plus efficace (ou meilleur) que T_2 si :

$$\text{E.Q.M.}(T_1) < \text{E.Q.M.}(T_2)$$

Convergence : Un estimateur est convergent si $\lim(n \to \infty) \quad P[|T-\theta| < \varepsilon] = 1 \quad \forall \varepsilon > 0$.

5.4.2 Estimation ponctuelle

Dans cette section, on s'intéresse à l'estimation des paramètres d'un modèle probabiliste (loi de distribution). Deux méthodes sont examinées: la méthode des moments et la méthode du maximum de vraisemblance.

5.4.2.1 Méthode des moments

Plusieurs des lois vues aux chapitres 3 et 4 sont entièrement définies par un ou deux paramètres. Ces paramètres peuvent correspondre aux deux premiers moments de la v.a. ou du moins y être reliés directement. Ainsi :

Loi	Paramètres
Bernouilli	1er moment
normale	deux premiers moments
lognormale	deux premiers moments du logarithme
exponentielle	inverse du 1er moment
beta	2 paramètres fonction des 2 premiers moments
extrême Type 1	2 paramètres fonction des 2 premiers moments

Dans ce contexte, il semble logique que les paramètres de la population puissent être estimés en utilisant les statistiques analogues, i.e. les moments calculés sur l'échantillon. C'est la méthode des moments.
La méthode des moments consiste donc à poser $E[X^k]=1/n \sum X_i^k$. Ainsi, on estimera μ par $\overline{X}$, $E[X^2]$ par $\frac{1}{n}\sum X_i^2$ et donc $\sigma^2=E[X^2]-\mu^2$ par $(\frac{1}{n}\sum X_i^2) - \overline{X}^2=\frac{1}{n}\sum (X_i-\overline{X})^2 =(n-1)/n\ S^2$.

5.4.2.2 Méthode du maximum de vraisemblance

Plutôt que de considérer l'échantillon comme « n » observations d'une même v.a., on peut aussi le considérer comme une seule réalisation d'une loi conjointe comprenant « n » v.a.. La vraisemblance n'est rien d'autre que la valeur de la fonction de densité conjointe évaluée aux valeurs prises par les « n » observations. Comme les observations sont considérées comme indépendantes, la loi conjointe s'obtient simplement du produit de « n » lois marginales. On a donc :

$$\text{Vraisemblance} = L(\boldsymbol{\theta} \mid x_1,...x_n) = f(x_1, x_2,...x_n \mid \theta_1...\theta_p) = \prod_{i=1}^{n} f(x_i \mid \theta_1...\theta_p)$$

où $\boldsymbol{\theta} = \theta_1...\theta_p$ rappelle la dépendance de la valeur de la fonction de densité sur les paramètres inconnus de la loi.

On note la vraisemblance : $L(\boldsymbol{\theta} \mid x_1,...x_n)$. À une constante près, la vraisemblance peut être interprétée comme une fonction de densité pour les paramètres $\boldsymbol{\theta}$ étant données les observations dont nous disposons.

Il semble naturel de prendre comme estimé pour $\theta_1...\theta_p$ les valeurs $\hat{\theta}_1...\hat{\theta}_p$ qui maximisent la valeur de la fonction de vraisemblance. Dans la plupart des cas, plutôt que maximiser la vraisemblance, on simplifie le problème et les manipulations en maximisant, de façon totalement équivalente, $\ln(L) = \sum_{i=1}^{n} \ln\left(f(x_i \mid \theta_1...\theta_p)\right)$.

Exemple : Considérons un échantillon de « n » observations tirées d'une loi normale.

$$L(\boldsymbol{\theta} \mid x_1,...,x_n) = \prod_{i=1}^{n} f(x_i \mid \theta_1...\theta_p) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

On trouve le maximum de cette fonction en dérivant par rapport à μ et σ. Cependant, il est plus simple ici de maximiser ln(*L*) plutôt que *L* (comme la transformation « ln » est monotone, le maximum demeure le même).

$$\ln(L) = \frac{-n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Dérivant et posant les dérivées partielles égales à 0, on trouve :

$$\frac{\partial \ln(L)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\frac{\partial \ln(L)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

d'où

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

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

Note : Pour la loi normale, la méthode des moments et la vraisemblance maximale fournissent le même estimateur pour μ et σ^2, soit $\bar{X}$ et $\hat{\sigma}^2 = \frac{n-1}{n} S^2$. On peut montrer que S^2 est sans biais (et donc $\hat{\sigma}^2$ est biaisé) mais S^2 est toujours moins efficace que $\hat{\sigma}^2$ dans le cas normal. En effet, on pour une loi normale :

$$E.Q.M.(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2} \quad < \quad \frac{2\sigma^4}{n-1} = E.Q.M.(S^2)$$

Exemple : Considérons l'échantillon suivant supposé provenir d'une loi normale :

Observation	Valeur
1	60
2	70
3	65
4	80
5	74

On calcule : $\bar{x} = 69.8,\ s^2 = 60.2$ et $\hat{\sigma}^2 = 48.2$

5.5 Estimation par intervalles de confiance

Un estimateur est une statistique. Si l'on connaît la distribution de la statistique, on peut fournir une estimation du paramètre inconnu sous la forme d'un intervalle plutôt que d'une valeur unique.

Exemple : Supposons que X_i soit $N(\mu, \sigma^2)$. On a alors $\bar{X} \sim N(\mu, \sigma^2/n)$ (une somme de v.a. normales indépendantes est aussi normale. Il ne reste qu'à calculer la moyenne et la variance de $\bar{X}$ et le résultat suit). Conséquemment :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow P\left[-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right] = 1 - \alpha$$

$$\Rightarrow P\left[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Interprétation : Cette équation signifie que si l'on construisait un grand nombre d'intervalles à partir d'échantillons différents de la manière indiquée, 95% de ces intervalles devraient contenir la vraie valeur inconnue μ.

Note : 1-α est appelé le niveau de confiance de l'intervalle (ou coefficient de confiance).

Exemple : Pour fournir l'intervalle précédent, il faut connaître σ. Qu'arrive-t-il si σ^2 n'est pas connu et qu'on l'estime par s^2 ?

On a vu au chapitre 4 qu'une loi de Student à « n » degrés de liberté est le ratio entre une N(0,1) et la racine d'une $\frac{\chi_n^2}{n}$ indépendantes. On a $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ et $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. On peut démontrer que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ est indépendant de $\frac{(n-1)s^2}{\sigma^2}$. Donc, $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

L'intervalle de confiance est donc tel que :

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \Rightarrow P\left[\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu > \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

Note : Si l'on avait estimé, σ^2 par $\hat{\sigma}^2$, alors on peut déduire du résultat précédent que : $\frac{n}{(n-1)}\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$ (et non pas $\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t_n$ comme on pourrait être porté à le croire). La raison est que $\overline{X}$ et $\hat{\sigma}^2$ ne sont pas indépendants.

5.5.1 Divers intervalles de confiance pour des cas courants

On peut construire des intervalles de confiance pour plusieurs statistiques courantes lorsque la distribution de X est normale ou lorsque « n » est assez grand pour que le théorème central limite entre en jeu.

Intervalle pour	Expression utilisée	Distribution utilisée pour construire l'intervalle
$\mu;\sigma^2$ connu	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	N(0,1)
$\mu;\sigma^2$ inconnu	$\frac{\overline{X}-\mu}{S/\sqrt{n}}$	t_{n-1}
$\mu_X-\mu_Y$; σ_X^2,σ_Y^2 connus	$\overline{X}-\overline{Y}$	$N(\mu_X-\mu_Y, \frac{\sigma_x^2}{n_X}+\frac{\sigma_Y^2}{n_Y})$
$\mu_X-\mu_Y$; σ_X^2,σ_Y^2 inconnus mais $\sigma_X^2=\sigma_Y^2$	$\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{S_p\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)^{1/2}}$	$t_{nx+ny-2}$, n_X : nombre d'observations de X $S_p^2 = \frac{(n_X-1)S_x^2+(n_Y-1)*S_Y^2}{n_X+n_Y-2}$
$\mu_X-\mu_Y$; σ_X^2,σ_Y^2 inconnus mais n_X et n_Y sont grands	$\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^{1/2}}$	t_v avec $v = \frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X-1}+\frac{\left(\frac{S_y^2}{n_Y}\right)^2}{n_Y-1}}$
σ^2 , μ connu	$\frac{n\hat{\sigma}^2}{\sigma^2}$	χ_n^2 avec $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$
σ^2 , μ inconnu	$\frac{(n-1)S^2}{\sigma^2}$	χ_{n-1}^2
σ_X^2/σ_Y^2	$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2}$	$F_{n_Y-1,\,n_X-1}$

p (loi binomiale)	$\hat{p}$	$N(p, p(1-p)/n)$ Note : on utilise $\hat{p}$ au lieu de p dans le calcul de la variance de $\hat{p}$.
p_X-p_Y (2 lois binomiales)	$\hat{p}_X - \hat{p}_Y$	$N(p_X - p_Y, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y})$ Note : on remplace les paramètres inconnus par leurs valeurs estimées pour le calcul de la variance de $\hat{p}_X - \hat{p}_Y$.
Paramètre estimé par vraisemblance maximale		Intervalle approximatif lu directement sur la fonction de vraisemblance $L(\boldsymbol{\theta} \mid x_1, ...x_n)$[1].

5.5.2 Détermination de la taille d'un échantillon

Les résultats précédents peuvent être utilisés pour évaluer à l'avance la taille de l'échantillon qui fournira un intervalle d'une certaine largeur.

Ex. La demi-largeur de l'intervalle de confiance pour $\mu; \sigma^2$ connu, est $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Si l'on désire un intervalle de ± 1 pour μ, on choisira $n = Z^2_{\alpha/2}\sigma^2$.

Ex. La demi-largeur de l'intervalle de confiance pour $\mu; \sigma^2$ inconnu, est $t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$. Si l'on désire un intervalle de ± 1 pour μ, on choisira $n = t^2_{n-1,\alpha/2}S^2$. Ici, il faudrait avoir une valeur préliminaire de S^2 (par exemple avec un pré-échantillon ou par expérience) pour pouvoir estimer « n ».

Des résultats similaires s'obtiennent pour les autres intervalles de confiance.

[1] Basé sur le test du rapport des vraisemblances. [θ_1, θ_2] tel que 2 [ln (L(θ^*)-ln(L(θ_1)] = 2 [ln (L(θ^*) - ln(L(θ_2)] = $\chi^2_{1,\alpha}$