STATISTIQUES

Statistique Inférentielle

Dans des domaines trés variés, comme l'économie, la gestion, la biologie, la médecine, ... on cherche des modèles mathématiques pour analyser, prévoir et décider. Très souvent, les facteurs intervenant sur le caractère étudié sont plus ou moins aléatoires, et les modèles sont probabilistes. Ainsi, les caractères étudiés sont modélisés par des variables aléatoires. Il s'agit du domaine des statistiques qui est la science des données.

Supposons maintenant que l'on souhaite étudier un caractère donné sur une population Ω. On modélise souvent ce caractère par une variable aléatoire X afin de rendre compte de la variabilité entre les individus. Un recensement complet nous permettrait de conna'tre les valeurs correspondant à tous les individus de cette population. Cela nous permettrait de proposer une modélisation très précise de la distribution de X sur la population toute entière. Cependant, la plupart du temps on n'observe qu'un échantillon de la population totale. En effet, pour des raisons de temps ou de coû t, il est en général impossible de recourir à un recensement.

Le but essentiel de la statistique (inférentielle) est d'obtenir des renseignements (ou une modélisation) d'un certain caractère noté X défini sur une population Ω à l'aide de mesures effectuées sur une sous-population $\omega_1, \cdots, \omega_n$ (appelée échantillon). Ainsi, la statistique inférentielle étudie les relations entre échantillons et population parente . On peut distinguer 3 types de démarche :

1. L'échantillonnage permet de faire le lien entre la population et l'échantillon. Si on connaissait parfaitement la population, que pourrait-on en déduire sur l'échantillon?

2. L'estimation permet de faire le lien entre l'échantillon et la population. Aprés avoir étudié un échantillon, que peut-on dire sur la population?

3. Les tests permettent des prises de décision sur la population à partir des valeurs observées sur un échantillon.

1 L'échantillonnage

Définition 1.1 *Un échantillon est une fraction d'individus de la population. Le prélèvement des éléments d'un échantillon peut-être effectué :*

avec remise : l'individu prélevé est immédiatement remis dans la population avant de prélever le suivant. Un individu pouvant éventuellement être prélevé plusieurs fois, les tirages sont indépendants

et l'échantillon est dit non exhaustif.

sans remise : l'échantillon est exhaustif, mais les tirages ne sont pas indépendants puisque la composition de la population parente (i.e. où le tirage est effectuée) est modifiée à chaque tirage.

On supposera dans ce cours que l'échantillonnage est aléatoire et simple c'est-à-dire que d'une part, tous les individus de la population ont la même probabilité de faire partie de l'échantillon, et d'autre part que les choix successifs des individus composant l'échantillon sont réalisés indépendamment les uns des autres (tirage avec remise, ou sans remise dans une population très grande). Par conséquent, on pourra considérer que les variables $X_1, X_2, \cdots, X_n$ (représentant le caractère propre aux individus numéros $1, 2, \cdots, n$ d'un échantillon de taille n) sont indépendantes et de même loi que X.

Remarque 1.2 *Dans la pratique, voici quelques recommandations pour obtenir un bon échantillon :*
- si le cardinal de Ω est petit (taux de sondage élevé), les individus doivent être choisis au hasard (avec remise).
- si le cardinal de Ω est suffisamment grand (taux de sondage faible) on peut se contenter d'un tirage sans remise. On considère en effet que la population ne change pas beaucoup quand a retiré quelques individus.
- les individus doivent être indépendants, donc on évite les liens qui pourraient influencer les valeurs du caractère (ne pas interroger mari et femme dans un sondage d'opinion).
- les mesures doivent être faites dans les mêmes conditions. Pour que les variables aient toutes la même loi.
- on peut améliorer l'échantillon avec la notion de sondage stratifié qui consiste à découper la population en classes plus homogènes, et à étudier chaque classe séparément.
- on verra aussi que la taille de l'échantillon influe aussi sur sa qualité. On est souvent limité par des problèmes budgétaires (institut de sondages politiques n = 1000) ; organismes d'état n = 10000; la plus grosse enqûete de l'INSEE "emploi " 150000.

1.1 Distribution d'échantillonnage

Nous savons comment décrire les caractéristiques importantes de jeux de données au travers des notions de moyenne, variance, médiane, quartiles, ... Lorsque l'on procède par sondage, ces caractéristiques ne sont plus déterministes. Ce sont des variables aléatoires qui prendront des valeurs aléatoires en fonction de l'échantillon considéré. Une variable calculée sur un échantillon (moyenne, somme, variance, ...) est elle-même une variable aléatoire. En effet, si on prélève plusieurs échantillons de même taille n dans une même population, la variable calculée pour chacun des échantillons prendra une valeur différente d'un échantillon à l'autre en raison du caractère aléatoire des prélèvements. La distribution de probabilité de cette statistique est appelée distribution d'échantillonnage. La démarche d'échantillonnage consiste à déduire (entre autres choses) les distributions des variables aléatoires

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

à partir de ce que l'on connait ou que l'on suppose sur la population entière. On suppose à partir de maintenant que l'échantillonnage est aléatoire et simple donc les variables $X_1, X_2, \cdots, X_n$ (mesurées

sur les individus numéros $1, 2, \cdots, n$ d'un échantillon de taille n) sont indépendantes et de même loi que X d'espérance m et de variance σ^2.

1.2 Distribution d'échantillonnage de la moyenne

Définition 1.3 *On définit la moyenne empirique de l'échantillon de la manière suivante :*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On montre facilement les propriétés suivantes

- $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m$
- Var $(\bar{X})$ = Var $(\frac{1}{n}\sum_{i=1}^{n} E(X_i)) = \frac{1}{n^2} n$ Var $(X_1) = \frac{\text{Var}(X_1)}{n}$

Donc $\bar{X}$ vaut m en moyenne (i.e. son espérance vaut m) et sa variance est inversement proportionnelle à n. Ainsi, plus l'échantillon est grand, plus $\bar{X}$ approche la moyenne m. Plus précisément, si la variable X suit une loi $\mathcal{N}(m, \sigma^2)$, la variable $\bar{X}$ suit la loi $\mathcal{N}(m, \frac{\sigma^2}{n})$).

Si la variable X suit une loi quelconque, on peut quand même appliquer le théorème de la limite centrale et quand n est "grand", faire l'approximation

$$\frac{\sqrt{n}(\bar{X} - m)}{\sigma} \sim \mathcal{N}(0, 1)$$

1.3 Distribution d'échantillonnage de la variance.

On peut montrer que

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

On dit (voir plus loin) que S^2 est un estimateur biaisé de la variance. Si la variable X suit une loi $\mathcal{N}(m, \sigma^2)$

$$\frac{nS^2}{\sigma^2}$$

On utilisera plus souvent l'estimateur corrigé de la variance $S^2_{n-1} = \frac{n}{n-1}S^2$ qui est sans biais et pour lequel on a

$$(n-1)\frac{S^2_{n-1}}{\sigma^2} \sim \chi^2(n-1)$$

1.4 Distribution d'échantillonnage de T

Dans de nombreux cas on ne connait pas la valeur de σ et on peut légitimement vouloir l'estimer en utilisant S^2. On est alors amené à considérer la variable T:

$$T = \sqrt{n}\frac{\bar{X} - m}{S}$$

On peut montrer que si X suit une loi $\mathcal{N}(m, \sigma^2)$, la variable T suit une loi de student de paramètre $n-1$. On peut montrer également, que si X suit une loi quelconque et que n est grand, T suit approximativement une loi $\mathcal{N}(0, 1)$.

2 Estimation

On suppose maintenant que l'on a choisi de modéliser le caractère que l'on souhaite étudier par une variable dont la loi de probabilité n'est pas totalement connue mais est supposée appartenir à une famille paramétrique . On dispose de données $(x_1, ..., x_n)$ qui sont les observations des variables aléatoires $(X_1, ..., X_n)$ modélisant le caractère étudié pour les n individus de l'échantillon. On s'intéresse donc naturellement à la manière dont on peut tirer profit des informations recueillies pour donner une valeur vraisemblable (estimation) des paramètres inconnus de la loi.

Exemple : On étudie par exemple la durée du trajet "domicile fac ". Soit D ce caractère. On dispose d'un échantillon de taille n = 80.

Temps de trajet	[0, 10]	]10, 20]	]20, 30]	]30, 40]	]40, 50]	]50, 60]
Effectif	32	30	10	6	1	1
Pourcentage	0.40	0.375	0.125	0.075	0.0125	0.0125

En regardant l'histogramme, on pense à une loi exponentielle. Le paramètre noté $\lambda > 0$ de cette loi exponentielle est inconnu et on doit en donner une estimation. On cherche pour λ une valeur vraisemblable compte tenu des valeurs observées $d_1 = D_1(\omega), d_2 = D_2(\omega), \cdots d_n = D_n(\omega)$ avec $n = 80$ le nombre d'individus observés et D_i l'observation du caractère D sur l'individu i. On sait par exemple que plus λ est grand, plus D prend des petites valeurs. De façon plus précise, si l'on suppose que D suit une loi exponentielle de paramètre λ , alors $E(D) = 1/\lambda$. Ainsi, d'aprés la loi des grands nombres, on a $\frac{D1+\cdots+Dn}{n} \to \frac{1}{\lambda}$ si n est grand. Par conséquent, l'inverse de la moyenne empirique approche λ (car la fonction inverse est continue en $\lambda > 0$). Et donc dans notre exemple on est tenté d'utiliser comme estimateur $\frac{n}{D_1+\cdots D_n}$ et d'approcher la valeur inconnue de λ par l'inverse de la moyenne empirique obsrvée $\frac{n}{d_1+\cdots d_n} = \frac{1}{14.6}$. (que l'on appelle estimation de λ). Nous donnons dans la section suivante une définition plus générale de la notion d'estimateur ponctuel.

2.1 L'estimateur ponctuel

Définition 2.1 *On appelle statistique une fonction de l'échantillon $(X_1, \cdots, X_n)$. C'est donc une variable aléatoire de la forme $T(X_1, \cdots, X_n)$.*

Définition 2.2 *Soit un modèle paramétrique $\{P_\theta, \theta \in \Theta\}$, on appelle estimateur de $g(\theta)$ toute statistique à valeurs dans $g(\Theta)$.*

Dans cette définition le paramètre θ peut être un vecteur de paramètres réels. C'est par exemple le cas pour les lois normales pour lesquelles $\theta = (m, \sigma)$. D'autre part, cette définition ne requiert pas que l'estimateur (souvent noté $\hat{\theta}$) soit "proche" ni même qu'il soit lié au paramètre θ. Cela est un peu étonnant car notre objectif est de l'utiliser pour donner une valeur vraisemblable de θ. D'autre part, on voit bien que l'on peut définir différents estimateurs pour un même paramètre inconnu θ. On a donc besoin de critères permettant de comparer la qualité de ces différents estimateurs. En pratique

on voudrait que l'estimateur soit "proche" de θ, mais on peut donner plusieurs sens au mot "proche" car l'estimateur est une variable aléatoire. On définit plus précisément ci-dessous différentes qualités des estimateurs qui motivent leur utilisation pour estimer de manière pertinente le paramètre inconnu.

A- La convergence : si on interrogeait toute la population, l'estimateur vaudrait θ. Ainsi, si on a un échantillon suffisamment grand, on voudrait que l'estimateur soit proche de θ

Définition 2.3 . *On dit que l'estimateur est convergent si* $T(X_1, ..., X_n) \to \theta$ *quand* $n \to \infty$

Remarque 2.4 *Il existe différentes notions de convergence (Lp, presque sûre, en probabilité en loi, ...) pour des variables aléatoires mais nous ne les évoquerons pas en détail ici.*

Premiers exemples : La moyenne empirique est d'après la loi des grands nombres un estimateur convergent de la moyenne. En effet, sous certaines hypothèses sur la loi de X, nous savons que

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to E[X] \text{ quand } n \to \infty$$

Il en est de même pour la variance empirique.

B. Pas de biais : on voudrait qu'il n'y ait pas d'erreur systématique, c'est à dire qu'en moyenne (i.e. espérance), l'estimateur soit égal à θ.

Définition 2.5 *On dit que l'estimateur est sans biais si* $E[T(X_1, \cdots, X_n)] = \theta$. *Dans le cas contraire on dit que l'estimateur est biaisé. Le biais est la quantité* $E[T(X_1, \cdots, X_n)] - \theta$.

Exemples:
La moyenne empirique est un estimateur sans biais de l'espérance. En effet, grace à la linéarité de l'espérance on a vu que $E(\bar{X}_n) = \frac{1}{n}(\sum_i E[X_i]) = E[X]$.
Par contre, si on veut estimer $\sigma^2 = E(X - E(X))^2$ la variance d'une variable X, on peut proposer la variance empirique

$$S^2 = \frac{(X_1 - \bar{X}_n)^2 + \cdots + (X_n - \bar{X}_n)^2}{n}$$

. C'est un estimateur convergent d'après la loi des grands nombres, mais on sait que ce n'est pas un estimateur sans biais. En effet, $E[(X_1 - \bar{X}_n)^2 + \cdots + (X_n - \bar{X}_n)^2] = (n-1)\sigma^2$. Pour avoir un estimateur sans biais, , on prendra plutôt

$$S^2_{n-1} = \frac{(X_1 - \bar{X}_n)^2 + \cdots + (X_n - \bar{X}_n)^2}{n-1}$$

comme estimateur de la variance (notamment dans un échantillon Gaussien).

Remarque 2.6 *Quand n est grand, les deux estimateurs sont équivalents !* S^2_{n-1} *est aussi un estimateur convergent. Le biais de* S^2 *tend vers 0 lorsque n tend vers l'infini. On dit que c'est un estimateur asymptotiquement sans biais.*

C- Efficacité d'un estimateur : Il est légitime de demander en plus des qualités précédentes que l'estimateur ait de bonnes qualités en terme de précision. Cette précision se mesure au travers du moment d'ordre 2 suivant $E[(T(X_1, \cdots, X_n) - \bar{X}_n)^2]$. Dans le cas d'un estimateur sans biais l'expression ci-dessus n'est autre que la variance de l'estimateur. Par conséquent, pour comparer la précision de deux estimateurs sans biais, on peut comparer leur variance. Si on dispose de plusieurs estimateurs sans biais et convergents, on choisit celui de variance minimale. La précision d'un estimateur est cependant limitée. On peut montrer, sous certaines conditions, que pour chaque valeur θ du paramètre et pour un biais b donné, le moment d'ordre 2 centré sur θ d'un estimateur de biais b ne peut pas être inférieur à une certaine valeur appelée borne de Cramér–Rao (qui se calcule à partir de la densité de la loi des X_i, de θ et de b). Dans le cas des estimateurs sans biais, cette borne minimale est appelée variance minimale.

Définition 2.7 *Un estimateur sans biais de variance minimale est dit efficace.*

Un tel estimateur n'existe pas toujours. On appelle efficacité d'un estimateur sans biais le rapport de la variance minimale à la variance de l'estimateur.

Premiers exemples :

Pour la moyenne empirique, l'erreur quadratique est $\frac{\text{Var }(X)}{n}$ Lorsque X suit une loi normale, il est aussi pertinent de considérer la médiane $m(X)$ comme estimateur de m. On pourrait montrer que c'est un estimateur sans biais dont la variance (lorsque n est assez grand) est supérieure à celle de la moyenne empirique. C'est donc un estimateur moins précis que la moyenne. De manière plus générale, on peut montrer que la moyenne empirique est un estimateur efficace de m lorsque X est de loi normale. Supposons à nouveau que X suit une loi normale. L'erreur quadratique de S_{n-1} est $\frac{2(\sigma^2)^2}{n-1}$. On peut montrer que dans ce cas la variance minimale est de $\frac{2(\sigma^2)^2}{n}$. Par conséquent S^2_{n-1} n'est pas un estimateur efficace de σ^2. Cependant son efficacité tend vers 1 lorsque n tend vers l'infini. On dit que c'est un estimateur asymptotiquement efficace.

Supposons enfin que l'on cherche à estimer la proportion p d'individus correspondant à un critère donné au sein d'une population. On introduit des varabales de Bernoulli $X_1, \cdots, X_n$ qui valent 1 si l'individu correspondant de l'échantillon correspond au critère et 0 sinon. Ces variables sont indépendantes (echantillonnage aléatoire simple) et de loi de Bernoulli(p) ($X \sim \mathcal{B}(p)$). Il semble assez logique de considérer comme estimateur la proportion d'individus de l'échantillon correspondant au critère c'est à dire

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

La somme des X_i suit une loi Binômiale(n, p) ($S_n \sim \mathcal{B}(n, p)$) d'espérance np donc $\hat{p}$ est un estimateur sans biais. On peut montrer que c'est un estimateur efficace.

Remarque 2.8 *Des résultats généraux (non évoqués ici) donnent des critères précis permettant d'obtenir des estimateurs efficaces pour une grande variété de lois de probabilités (familles exponentielles).*

Question : Les estimateurs proposés dans les cas précédents sont, pour la plupart, assez intuitifs; comment obtenir, de façon générale, des estimateurs possédant les qualités énoncées précédemment ? Nous présentons deux méthodes permettant d'obtenir ces estimateurs

2.2 Méthode des moments :

Une première approche, appelée méthode des moments consiste à tirer profit de la loi des grands nombres qui nous dit (sous certaines hypothèses, voir la leçon sur l'echantillonnage) que les moments empiriques d'ordre k convergent presque sûrement vers les moments théorique $E(X^k)$

$$\bar{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \to E(X^k) \text{ quand } n \to \infty$$

La méthode des moments consiste à exprimer les p premiers moments μ_j de la loi de X en fonction des p paramètres $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ inconnus. Cela amène à un système de p équations à p inconnues :

$$\begin{cases} \mu_1 = h_1(\theta_1, \cdots, \theta_p) \\ \mu_2 = h_2(\theta_1, \cdots, \theta_p) \\ \vdots \\ \mu_p = h_1(\theta_1, \cdots, \theta_p) \end{cases}$$

que l'on cherche à résoudre afin d'obtenir l'expression de chaque paramètre θ_j en fonction des moments $\mu_1, \cdots \mu_p$. On obtient ensuite des estimateurs en remplaçant les moments théoriques μ_l par les moments empiriques $\bar{X^l}$ dans les expressions obtenues pour les θ_j.

Remarque : il peut arriver que l'on ne considère pas (ou pas seulement) les p premiers moments afin d'obtenir un système d'équations permettant d'aboutirà une unique solution. Nous n'en parlerons pas en détails dans ce cours car en général les premiers moments suffisent.

Premiers exemples : La méthode que nous avons utilisée au début de ce TP pour la loi exponentielle est un exemple de l'utilisation de la méthode des moments. On a remarqué que $\mu_1 = \frac{1}{\lambda}$ puis on a déduit que $\hat{\lambda} = \frac{1}{\mu_1}$ et enfin remplacé μ_1 par $\bar{D}_n$ pour aboutir à l'estimateur $\hat{\lambda} = \frac{1}{\bar{D}_n}$. Supposons que l'on étudie un échantillon $(X1, ..., Xn)$ constitué de variables aléatoires indépendantes de même loi $\mathcal{N}(m, \sigma^2)$. Les premiers moments s'expriment en fonction de m et σ de la manière suivante :

$$\begin{cases} \mu_1 = m \\ \mu_2 = \sigma^2 + m^2 \end{cases}$$

On obtient facilement en résolvant ce système les identités suivantes

$$\begin{cases} m = \mu_1 \\ \sigma^2 = \sqrt{\mu_2 - \mu_1^2} \end{cases}$$

ce qui nous amène à considérer les estimateurs $\hat{m} = \bar{X}$ et $\hat{\sigma} = \sqrt{\bar{X^2} - (\bar{X})^2}$. On peut remarquer que dans ce cas on retrouve les estimateurs $\bar{X}$ et $\sqrt{S^2}$que nous avions déjà considéré.

2.3 Maximum de vraisemblance :

La méthode des moments décrite dans la section précédente est assez intuitive et permet de proposer un certain nombre d'estimateurs convergents. Cependant, ces estimateurs n'ont pas tou- jours d'aussi

bonnes propriétés que ceux que l'on obtient en considérant les estimateurs obtenus par la méthode du maximum de vraisemblance que nous allons voir maintenant.

La méthode du maximum de vraisemblance permet d'aboutir dans de nombreux cas à des estimateurs efficaces. Son principe consiste à choisir comme estimation du paramètre θ, la valeur la plus vraisemblable, c'est-à-dire celle qui à la plus forte probabilité de provoquer les valeurs observées dans l'échantillon.

La loi d'une variable X est caractérisée par la probabilité des valeurs possibles (cas discret) ou par sa densité (cas continu). De la même façon, si X suit une loi P_θ de paramètre θ, la loi du n-échantillon (composé de variables indépendantes) est caractériséeée par $\prod_{i=1}^{n} P_\theta(x_i)$. On notera

$$L(x_1, x_2, \cdots x_n, \theta)$$

ce produit. Il s'agit de la vraisemblance.

Définition 2.9 *On appelle estimateur du maximum de vraisemblance de θ une valeur (de t) qui maximise la fonction de vraisemblance $t \to L(X_1, ..., X_n; t)$.*

L'estimation de θ par le maximum de vraisemblance revient donc à chercher le paramètre avec lequel $L(x_1, ..., x_n, t)$ est maximale, c'est à dire "la valeur du paramètre avec laquelle on avait le plus de chance d'obtenir ce qu'on a obtenu".

Remarque : pour des raisons de commodité de calcul, on utilise souvent la fonction de log vraisemblance

$$LL : t \to ln(L(X_1, ..., X_n, t))$$

qui est le logarithme népérien de la fonction de vraisemblance, elles sont maximales en même temps (car le logarithme népérien est strictement croissant sur R_*^+). Reprenons l'exemple de la durée du trajet domicile fac. Si D suit une loi exponentielle de paramètre θ, la "probabilité" infinitésimale (ou densité) d'avoir obtenu les résultats $\vec{x} = (x_1, \cdots, x_n)$ est

$$L(\vec{x}, \theta) = \theta e^{-\theta x1} \theta e^{-\theta x2} \cdots \theta e^{-\theta xn} = \theta^n e^{-\sum x_i}$$

La log-vraisemblance est donc

$$LL(\vec{x}, \theta) = nln(\theta) - \theta \sum_{i=1}^{n} x_i \, .$$

Pour maximiser la log-vraisemblance qui est concave, on dérive par rapport à θ et on annule la dérivée ce qui donne

$$\frac{\partial}{\partial \theta} LL(\vec{x}, \theta) = 0 \iff \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

et donc

$$\hat{\theta} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}} \, .$$

Sous des hypothèses très générales on peut montrer que l'estimateur du maximum de vraisemblance est efficace ou asymptotiquement efficace et que sa distribution d'échantillonnage est asymptotiquement normale.

2.4 L'estimation par intervalle de confiance.

L'estimation ponctuelle d'un paramètre, c'est-à-dire la connaissance de la seule valeur estimée de ce paramètre, n'a d'intérêt que si l'on a une idée de la précision avec laquelle il a été estimé. La plupart du temps, on complète cette estimation en donnant une fourchette

$$[a(x_1, \cdots, x_n); b(x_1, \cdots, x_n)]$$

appelée intervalle de confiance. Elle correspond à la valeur prise sur l'échantillon par l'intervalle aléatoire $[a(X_1, \cdots, X_n); b(X_1, \cdots X_n)]$ que l'on construit de telle manière qu'il y ait une grande probabilité que la vraie valeur du paramètre se trouve à l'intérieur.

Rappels sur la notion de quantile

Les notions de fonction de répartition et de fonction quantile associées à la loi d'une variable aléatoire X sont réciproques l'une de l'autre. Elles permettent de répondre à deux problématiques complémentaires:

- **Quelle est la probabilité que X soit inférieure ou égale à une valeur seuil donnée ?** On appelle fonction de répartition associée à la loi la variable aléatoire X la fonction $F_X : s \to P(X \leq s)$. La valeur de la fonction de répartition au point $s = 5$ correspond donc à la probabilité que la variable X soit inférieure ou égale à 5.

- **Quelle valeur seuil faut-il considérer pour que X lui soit inférieure ou égale avec (au moins) une probabilité donnée ?** Remarquons tout d'abord que la fonction de répartition F_X n'est pas toujours continue et strictement croissante. Par conséquent il arrive que la solution, en s, de $F_X(s) = t$ n'existe pas ou ne soit pas unique. En d'autres termes, il se peut que pour certaines valeurs de t il n'existe pas de valeur seuil s telle que la probabilité que X lui soit inférieure ou égale soit exactement égale à t. Il se peut également que l'on ait plusieurs valeurs seuil possibles. La fonction quantile associée la loi de X correspond à la fonction inverse généralisée de la fonction de répartition. On la définit de manière générale par

  $$Q_X(t) = \inf\{s, F_X(s) \geq t\}$$

  On appelle donc quantile d'ordre t (noté $Q_X(t)$) la plus petite valeur seuil s telle que la probabilité que $X \leq s$ soit au moins t. Remarquons cependant que lorsque la variable X est de loi continue et si la densité est strictement positive, alors F_X est strictement croissante et l'on peut définir $Q_X(t)$ comme la solution en s de $F_X(s) = t$. Le quantile $Q_X(t)$ correspond alors exactement à la seule valeur seuil pour laquelle on a exactement une probabilité t que X lui soit inférieure ou égale. La probabilité que X soit entre deux valeurs a et b (avec $a \leq b$) s'obtient de la manière suivante :

  $$P(a < X \leq b) = F_X(b) - F_X(a)\,.$$

  Cela est toujours vrai pour les variables de loi continues si on considère des inégalités strictes ou larges car $P(X = a) = P(X = b) = 0$. Attention, ce n'est pas le cas lorsque X est une variable discrète car dans ce cas $P(X < s) < F_X(s)$.

On se concentre pour ce paragraphe sur le cas de variables dont la loi est continue. On peut visualiser la valeur de la fonction de répartition au point s comme l'aire comprise sous le graphe de la densité jusqu'à la valeur s (c'est à dire sur l'intervalle $]-\infty, s]$). Chercher le quantile $Q_X(t)$, consiste à trouver la plus petite valeur s telle que l'aire comprise sous la courbe soit égale à t. Enfin, la probabilité que X se trouve dans l'intervalle $[a; b]$ est l'aire comprise sous la courbe de la densité entre les points a et b.

Les fonctions de répartition de certaines lois de probabilité n'ont pas de forme explicite. Leurs valeurs sont données dans des tables. On peut se servir de ces tables pour trouver la valeur approchée des quantiles. Vous trouverez par exemple les tables de la loi normale centrée réduite $\mathcal{N}(0, 1)$ ainsi que des lois du χ^2 et des lois de Student. Elles sont présentées de manières un peu différentes. La table de la loi $\mathcal{N}(0, 1)$ dont vous disposez donne les valeurs de la fonction de répartition $F_X(t)$ pour différentes valeurs de t. La première colonne de la table correspond aux premières décimales de t tandis que la première ligne correspond aux centièmes. On trouve par exemple la valeur de $F_X(0.25)$ dans la case correspondant à la ligne correspondant à 0.2 (dans la première colonne) et dans la colonne correspondant à 0.05 (dans la première ligne), c'est à dire la valeur 0.5987. La table ne donne que les valeurs de la fonction de répartition pour des valeurs de t positives. Comme la densité de la loi $\mathcal{N}(0, 1)$ est symétrique par rapport à 0 on a $F_X(-t) = 1 - F_X(t)$, ce qui nous permet d'obtenir les valeurs $F_x(t)$ pour $t < 0$. Par exemple, $F_X(-0.25) = 1 - F_X(0.25) = 1 - 0.5987 = 0.4013$. Enfin, lorsque la variable Z est de loi $\mathcal{N}(m, \sigma^2)$, la variable $\frac{Z-m}{\sigma}$ est de loi $\mathcal{N}(0, 1)$. Par conséquent,

$$F_Z(t) = P(Z \leq t) = P(\frac{Z-m}{\sigma} \leq \frac{t-m}{\sigma}) = F_X(\frac{t-m}{\sigma})$$

où X est une variable aléatoire de loi normale centrée réduite $\mathcal{N}(0, 1)$. Supposons par exemple que Z suit une loi $\mathcal{N}(1, 2^2)$. La probabilité que Z soit inférieur ou égal à 2 est

$$F_Z(2) = F_{\frac{Z-1}{2}}(\frac{2-1}{2}) = F_{\frac{Z-1}{2}}(0.5) = 0.6915$$

La table de la loi de Student donne seulement les valeurs de t telles que $P(|X| > t) = P$ pour différentes valeurs de P et du degré liberté ν. Nous verrons l'intérêt de donner simplement ce type de valeurs dans ce qui suit. Enfin, la table du χ^2 donne quant à elle les valeurs de t telles que $P(X \leq t) = P$ (c'est à dire de $Q_X(P)$) pour différentes valeurs de P et du degré de liberté (ddl).

2.5 Construction d'intervalles de confiance

Exemple introductif :
Un sondage effectué à quelques jours du second tour d'une élection présidentielle sur un échantillon de 100 personnes donne 47% d'intentions de vote pour Mr Dupont et 53% pour Madame Durand. On supposera que l'échantillonnage est aléatoire et simple. Notons p la probabilité qu'un électeur vote Madame Durand. Les observations recueillies sur notre échantillon nous amènent à estimer p par 53%. Toutefois, afin de pouvoir conclure quelque chose de ce sondage, il est nécessaire d'avoir plus d'informations sur la marge d'erreur de notre estimation. En d'autres termes, quel risque court-on en concluant que p est supérieur à 50% ? Pour répondre à ces questions, on donne en plus un

intervalle de confiance $[a(x_1, \cdots, x_n); b(x_1, \cdots, x_n)]$. Il s'agit de la valeur observée (calculée à partir de nos observations) d'un intervalle aléatoire $[a(X_1, \cdots, X_n); b(X_1, \cdots, X_n)]$ construit à partir de notre échantillon $(X_1, \cdots, X_n)$ de telle sorte que l'on ait une grande probabilité (appelée niveau de confiance) que p y appartienne.

Démarche générale :

On fixe un nombre α compris entre 0 et 1, généralement proche de 0 et appelé risque. Le nombre $1 - \alpha$, généralement proche de 1 est appelé niveau de confiance. Etant donné un échantillon $(X_1, X_2, \cdots, X_n)$ de variables aléatoires, on cherche un intervalle aléatoire

$$I(X_1, \cdots, X_n) = [a(X_1, \cdots, X_n), b(X_1, \cdots, X_n)]$$

tel que α étant fixé, on ait :

$$P(\theta \in I(X_1, \cdots, X_n)) \geq 1 - \alpha\,. \tag{1}$$

Une observation $I(x_1, x_2, \cdots x_n)$ de $I(X_1, \cdots, X_n)$ est appelée intervalle de confiance de θ au niveau $1 - \alpha$. On appelle $1 - \alpha$ le niveau de confiance de l'intervalle de confiance I, et on dit que θ est dans l'intervalle I avec un niveau de confiance $1 - \alpha$.

Remarque 2.10 *Lorsque l'on considère des risques plus faibles, la largeur de l'intervalle de confiance est plus grande. Afin d'avoir moins de risques de se tromper en concluant que $\theta \in I(X_1, \cdots, X_n)$, on considère un intervalle plus large et on perd donc en précision.*

Remarque 2.11 *Il existe en général une infinité d'intervalles de confiance vérifiant* (1) *(même lorsque la probabilité vaut $1 - \alpha$). Afin d'avoir un maximum de précision, on choisit en général les intervalles d'amplitude la plus courte. Par conséquent on privilégira notamment les intervalles tels que $P(\theta \in I(X_1, \cdots, X_n)) = 1 - \alpha$ s'ils existent (voir remarque précédente). D'autre part, parmi les intervalles vérifiant cette dernière égalité, on considèrera plutôt certains types d'intervalles en fonction de la nature de la densité de la loi que l'on considère (voir exemples suivants).*

Remarque 2.12 *La largeur des intervalles de confiance décroît lorsque la taille de l'échantillon augmente.*

2.6 Estimation d'une moyenne par intervalle de confiance :

On construit un intervalle de confiance pour la moyenne en utilisant l'estimateur de la moyenne $\bar{X}$

2.6.1 Cas d'un échantillon de grande taille

On peut utiliser le théorème de la limite centrale qui dit que $\sqrt{n}(\bar{X} - m)/\sigma$ suit approximativement la loi $\mathcal{N}(0, 1)$ quand n est grand ($n > 30$). De plus, quand n est grand on peut approximer σ^2 par la variance empirique S^2_{n-1}. La densité de la loi normale $\mathcal{N}(0, 1)$ est symétrique et concentrée autour de sa moyenne 0. Par conséquent, l'intervalle de confiance de largeur minimale est symétrique par rapport à 0. On recherche donc une constante δ_α positive telle que

$$P(\sqrt{n}\frac{\bar{X} - m}{S_{n-1}} \in [-\delta_\alpha; \delta_\alpha]) \geq 1 - \alpha$$

Cela revient à chercher δ_α telle que

$$P(\sqrt{n}\frac{|\bar{X}-m|}{S_{n-1}} > \delta_\alpha) \leq 1-\alpha$$

Comme la densité de la loi gaussienne est symétrique par rapport à 0, pour toute variable aléatoire Z de loi $\mathcal{N}(0,1)$ et $\delta > 0$ on a $\phi(-\delta) = P(Z \leq -\delta) = P(Z \geq \delta)$. Par conséquent la valeur de δ_α correspond à la valeur t telle que $P(Z \geq t) = 1 - \frac{\alpha}{2}$ et δ_α est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ que l'on note $\phi^{-1}(1-\frac{\alpha}{2})$.. On peut se servir de la table pour obtenir les valeurs du quantile et on en déduit finalement que

$$m \in [\bar{X} - \frac{S_{n-1}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), \bar{X} + \frac{S_{n-1}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]$$

avec une sécurité de $1-\alpha$.. On donnera donc comme intervalle de confiance de niveau de risque $\alpha = 0.05$ et pour m l'intervalle

$$[\bar{x} - \frac{s_{n-1}}{\sqrt{n}} \times 1.96, \bar{x} + \frac{s_{n-1}}{\sqrt{n}} \times 1.96]$$

On voit clairement que la largeur de l'intervalle de confiance décroît lorsque la taille de l'échantillon augmente.

2.6.2 Cas d'un échantillon de petite taille

On ne peut plus utiliser le théorème de la limite centrale et donc on ne peut pas faire grand chose. Le seul cas résolvable est celui où on suppose que le modèle est Gaussien. Dans ce cas, si l'écart-type est supposé connu, on centre et réduit et on utilise la loi $\mathcal{N}(0,1)$ comme dans l'exemple précédent. Par contre si on ne connait pas σ^2, on l'estime par l'estimateur S^2_{n-1} et on centre et on réduit. On peut montrer que dans ce cas la variable centrée réduite suit la loi de student de paramètre $n-1$. La densité de la loi de student est elle aussi symétrique et concentrée autour de 0. En suivant les mêmes étapes que dans le paragraphe précédent, on obtient les intervalles de confiance suivants où les quantiles $t_{1-\frac{\alpha}{2}}$ de la loi de Student(n-1) remplacent ceux de la loi $\mathcal{N}(0,1)$

$$[\bar{x} - \frac{s_{n-1}}{\sqrt{n}}t_{1-\frac{\alpha}{2}}, \bar{x} + \frac{s_{n-1}}{\sqrt{n}}t_{1-\frac{\alpha}{2}}]$$

Remarque : On lit facilement la valeur de $t_{1-\frac{\alpha}{2}}$ sur la table de la loi de student sur la ligne 2 corespondant au degré de liberté $n-1$ et dans la colonne correspondant à $P = \alpha$.

2.6.3 Estimation d'une proportion p

Pour fixer les idées, revenons à notre exemple introductif et considérons la proportion p d'individus favorables à Mme Durand. Le nombre N de personnes favorables dans l'échantillon considéré, suit une loi binomiale $B(n,p)$. L'estimateur ponctuel de p est $F = \frac{N}{n}$. Si n est grand ($n > 30$) et

$nf(1-f) > 12$, on peut approximer la loi binomiale par la loi normale $\mathcal{N}(np, np(1-p))$. Ainsi, $\frac{F-p}{\sqrt{p(1-p)/n}}$ peut-être approximé par une loi $\mathcal{N}(0,1)$ et on a donc :

$$P\left(\frac{|F-p|}{\sqrt{p(1-p)/n}} > \delta\right) \approx 2\phi(\delta) - 1$$

On choisit donc $\delta_\alpha = \phi^{-1}(1-\frac{\alpha}{2})$ et l'on a:

$$P(p \in [F - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), F + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})] \approx 1-\alpha$$

Cependant, les bornes de l'intervalle dépendent de p qui est inconnue.

Une première approche consiste à trouver analytiquement les valeurs de p telles que

$$p \in [F - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), F + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]$$

en étudiant des inégalités faisant intervenir des polynômes d'ordre 2 en p. Cette approche est cependant assez fastidieuse. On utilisera plutôt l'une des deux méthodes "approchées" suivantes qui ont l'avantage d'être plus simples. On peut tout d'abord donner un intervalle un peu plus large en utilisant que $p(1-p) \leq 1/4$:

$$p \in [f - \frac{1/2}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), f + \frac{1/2}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]$$

avec une sécurité de $1-\alpha$. Si on veut être plus précis, on remplace $p(1-p)$ par $f(1-f)$. Lorsque l'on ne peut pas utiliser une approximation de la loi binomiale par une loi normale, il est possible d'envisager dautres approches mais nous n'en parlerons pas ici.

Revenons à notre exemple : Nous avons $n = 100 > 30$ et $nf(1-f) = 100 \times 0.53 \times 0.47 = 24.91 > 12$. On peut donc appliquer l'approximation de la loi binomiale par la loi gaussienne, ce qui nous donne pour $\alpha = 0.05$ comme intervalle de confiance

$$\begin{aligned}
&[f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})] \\
&= [0.53 - \frac{\sqrt{0.53 \times 0.47}}{\sqrt{100}} \times 1.96; 0.53 + \frac{\sqrt{0.53 \times 0.47}}{\sqrt{100}} \times 1.96] \\
&\approx [0.4322; 0.6278]
\end{aligned}$$

On ne peut donc pas vraiment conclure que Madame Durand a plus de 50% d'intentions de vote dans la population toute entière avec un risque de 5%.

2.6.4 Estimation d'une variance par intervalle de confiance pour un échantillon Gaussien

1. Estimation de σ^2 avec m connue. On utilise que

$$T = \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2}$$

est une somme de n variables indépendantes de loi $\mathcal{N}(0,1)$ au carré. Or une telle somme suit une loi $\Gamma(n,1)$, appelée aussi loi de Khi-deux à n degrés de liberté. Puis dans la table de Khi-deux on cherche u et v tels que $P(u \leq \chi^2(n) \leq v) = 1-\alpha$, on choisit par exemple u et v tels que $P(\chi^2(n) \leq u) = P(\chi^2(n) \geq v) = \alpha/2$.

Par exemple, $n = 30$, $1-\alpha = 0.95$ et $u = 16.79$ et $v = 46,97$. D'où l'intervalle de confiance pour σ^2:

$$[\frac{\sum(x_i-m)^2}{v}, \frac{\sum(x_i-m)^2}{u}]$$

Quand le degré de liberté $\nu > 30$, on ne peut plus utiliser la table du khi-deux et on utilise l'approximation suivante : Le quantile $\chi^2_{\nu,\alpha}$ ayant la proba α d'être dépassé peut-être approximé quand $\nu > 30$ par

$$\frac{(t_\alpha + \sqrt{2\nu-1})^2}{2}$$

où t_α est est la valeur ayant la proba α d'être dépassée par la loi $\mathcal{N}(0,1)$. Ainsi

$$u = \frac{(t_{\frac{\alpha}{2}} + \sqrt{2\nu-1})^2}{2} \tag{2}$$

$$v = \frac{(t_{1-\frac{\alpha}{2}} + \sqrt{2\nu-1})^2}{2} \tag{3}$$

2. Estimation de σ^2 quand m est inconnue. On procède exactement de la même façon en remplaçant m par son estimateur $\bar{X}$ et $\sum \frac{(X_i-\bar{X})^2}{\sigma^2}$ suit une loi de Khi-deux à $n-1$ degrés de liberté.

2.6.5 Détermination du nombre d'observations nécessaires pour une certaine précision.

L'amplitude d'un intervalle de confiance est une fonction décroissante de α et de n. Par conséquent, pour augmenter la précision de l'estimation c'est-à-dire réduire l'amplitude de l'intervalle de confiance, il faut :

- soit augmenter le risque α (ou ce qui revient au même, diminuer le niveau de confiance $1-\alpha$)
- soit augmenter la taille de l'échantillon. Comme on souhaite en général ne pas augmenter le risque, on peut souhaiter déterminer le nombre d'observations nécessaires pour atteindre une précision donnée pour un risque α fixé.

Revenons à notre exemple concernant les élections. On a vu qu'avec un échantillon de taille 100, l'intervalle de confiance de risque $\alpha = 0.05$ a une précision de 0.0978 et donc une amplitude de 0.1956 autour de la valeur 0.53. Il contient des valeurs plus petites que 50% donc on ne peut conclure que Mme Durand a plus d'électeurs en sa faveur dans la population totale avec un risque de 5%. On pourrait voir que pour que avoir une précision de 3% de notre intervalle de confiance et qu'il ne contienne pas de valeur plus faible que 0.5 il nous faut prendre un risque d'environ 0.548 ce qui est tout de même assez élevé. On préfère plutôt considérer un échantillon de taille plus grande qui nous permettrait d'avoir une précision de 3% autour de la valeur 0.53 pour le même risque $\alpha = 0.05$. On voit par exemple que pour n = 1064, on a une précision supérieure à 3%. On voit cependant clairement que la précision de notre intervalle de confiance dépend de la valeur de p et donc de celle de f. Pour avoir une précision de 3% indépedamment des valeurs de p et f, on doit prendre n tel que $\frac{\sqrt{1/4}}{\sqrt{n}} \times \Phi^{-1}(1-\frac{\alpha}{2}) \leq 0.03$ c'est c'est à dire tel que $n \geq \frac{1/4}{0.03^2} \times 1.96^2 \approx 1067.111$.