Lois limites Echantillonnage

1 Lois Limites

1.1 Exemple

On considère $X_1, \ldots, X_n$,n variables aléatoires indépendantes et de même loi, de moyenne m et d'écart type σ.

Soit $Y_n = \dfrac{X_1 + \ldots + X_n}{n}$ alors $E(Y_n) = m$ et $\sigma(Y_n) = \dfrac{\sigma}{\sqrt{n}}$

1.2 Deux théorèmes fondamentaux

Th 1 : *Loi faible des grands nombres*

Soient X_1; X_2 ; . . . X_n n variables aléatoires indépendantes et de même loi de moyenne m alors la variable aléatoire $Y_n = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ vérifie pour tout réel positif ε :

$$\lim_{n \to +\infty} p(|Y_n - m| < \varepsilon) = 1$$

Signification: Lorsque n devient très grand, le nombre de valeurs de la variable aléatoire Y_n qui sont éloignées de m devient négligeable devant celui des valeurs proches de m

Th 2 : *Théorème de la limite centrée ou théorème limite central*

Si X_1; X_2 ; . . . X_n sont n variables aléatoires indépendantes et de même loi de moyenne m et d'écart - type σ alors la variable aléatoire $Y_n = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ suit approximativement une loi normale $\mathcal{N}(m, \dfrac{\sigma}{\sqrt{n}})$

2 Loi d'échantillonnage

Problème : Etude du lien entre les caractéristiques d'une population et celle d'un échantillon de cette population. Dans cette partie on suppose la population connue et l'on étudie les échantillons.

2.1 Distribution d'échantillonnage des moyennes

L'effectif de la population étudiée est N, sa moyenne est m et son écart - type σ.
On prélève dans cette population un échantillon aléatoire de taille n. (tirage non exhaustif)
On va alors chercher à estimer la moyenne de cet échantillon ainsi que son écart - type.

On appelle distribution d'échantillonnage des moyennes la variable aléatoire $\overline{X}$ qui, à tout échantillon de taille n prélevé avec remise, associe la moyenne de cet échantillon.

Pour n suffisamment grand $\overline{X}$ suit approximativement la loi $\mathcal{N}(m, \dfrac{\sigma}{\sqrt{n}})$

Si les tirages sont exhaustifs $\overline{X}$ suit approximativement la loi $\mathcal{N}(m, \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}})$

2.2 Distribution d'échantillonnage des fréquences ou des pourcentages

L'effectif de la population est N. une certaine partie de pourcentage p possède une certaine propriété. Soit F la variable aléatoire qui, a tout échantillon aléatoire prélevé avec remise et d'effectif n fixé, associe le pourcentage des éléments de cet échantillon possédant cette propriété.

Pour n suffisamment grand F suit approximativement la loi $\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ où $q = 1 - p$.

Si le tirage est exhaustif alors pour n suffisamment grand F suit la loi $\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$

2.3 Distribution d'échantillonnage des différences et des sommes

On considère deux échantillons A et B issus de deux populations. L'échantillon A est d'effectif n_A, l'échantillon B est d'effectif n_B.
D'après le **1.** les distributions d'échantillonnage sont:

Pour A: $\overline{X_A}$ de moyenne $m'_A = m_A$ d'écart type $\sigma'_A = \frac{\sigma_A}{\sqrt{n_A}}$

Pour B: $\overline{X_B}$ de moyenne $m'_B = m_B$ d'écart type $\sigma'_B = \frac{\sigma_B}{\sqrt{n_B}}$

En prenant toutes les combinaisons possibles de 2 éléments de ces échantillons, on définit les distributions d'échantillonnages des sommes: $S = \overline{X_A} + \overline{X_B}$ ou des différences $D = \overline{X_A} - \overline{X_B}$.

Comme $E(X + Y) = E(X) + E(Y)$ et $E(X - Y) = E(X) - E(Y)$
la moyenne de la somme est : $E(S) = m'_A + m'_B$
la moyenne de la différence est : $E(D) = m'_A - m'_B$

Comme : Si X et Y sont indépendants $V(X + Y) = V(X) + V(Y)$
et $V(X - Y) = V(X) + V(Y)$!

Si les échantillons sont indépendants, l'écart type de la somme et l'écart type de la différence sont égaux : $\sigma_S = \sigma_D = \sqrt{\sigma'^2_A + \sigma'^2_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

2.4 Distribution d'échantillonnage des variances (hors programme?)

On appelle distribution d'échantillonnage des variances la variable aléatoire $\overline{S}$ qui, à tout échantillon de taille n prélevé avec remise, associe la variance de cet échantillon.

La moyenne de la distribution d'échantillonnage des variances est $E(S^2) = \frac{n-1}{n}\sigma^2$

3 Estimation

Dans cette partie, nous allons chercher des informations sur la population connaissant les caractéristiques d'un échantillon. Il s'agit en fait de faire l'inverse de ce qui a été fait dans la partie précédente.

3.1 Estimation ponctuelle

Moyenne: On choisit la moyenne $\bar{x}$ d'un échantillon prélevé au hasard comme estimation de la moyenne de la population.

Fréquence (ou proportion): On choisit la proportion f des éléments possédant une certaine propriété dans un échantillon prélevé au hasard comme estimation de la proportion inconnue de la population.

Variance, écart - type: On choisit le nombre $\frac{n}{n-1}\sigma'^2$ comme estimation de la variance inconnue σ^2 de la population. (n est l'effectif et σ'^2 la variance de l'échantillon prélevé au hasard dans la population)

L'écart type σ de la population est estimé par $\sigma = \frac{n}{n-1}\sigma'$.

rem : Dans le cas où n n'est pas petit par rapport à N et où le tirage des éléments de l'échantillon est sans remise alors on prend pour estimation ponctuelle de l'écart - type de la population le nombre

$$\sigma = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

3.2 Estimation par intervalle de confiance d'une moyenne

On commence par se fixer un coefficient de confiance (95% ou 99% ou . . .) et on va chercher un intervalle de confiance de la moyenne pour ce coefficient , c'est à dire trouver un intervalle [a, b] tel que l'on soit sûr à 95% (ou 99% ou . . .) que la moyenne est dans cet intervalle.

On a vu (**2.**) que pour n assez grand ($n \geq 30$) la distribution d'échantillonnage des moyennes suivait à peu près la loi $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$

donc la v.a.r. $T = \frac{\sqrt{n}}{\sigma}(X - m)$ suit la loi normale $\mathcal{N}(0;1)$ alors si $t \geq 0$ $\text{Prob}(-t \leq T \leq t) = 2\pi(t) - 1$

Trouver un intervalle de confiance à 95% c'est d'abord trouver t tel que

$$\text{Prob}(-t \leq T \leq t) = 0{,}95$$

On sait que ceci est vrai pour $t = 1{,}96$ (voir feuille loi normale)

Il ne reste plus alors qu'à en déduire l'intervalle de confiance sur X

$$-1{,}96 \leq \frac{\sqrt{n}}{\sigma}(\overline{X} - m) \leq 1{,}96 \quad ñ \quad -1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - m \leq 1{,}96\frac{\sigma}{\sqrt{n}} \quad ñ \quad m - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq m + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Donc avant de prélever un échantillon de taille n, il y a 95% de chances que sa moyenne soit dans l'intervalle $\left[m - 1{,}96\frac{\sigma}{\sqrt{n}} ; m + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$

Mais on peut aussi écrire à la place de la dernière inégalité

$- \overline{X} - 1{,}96 \frac{\sigma}{\sqrt{n}} \leq -m \leq -\overline{X} + 1{,}96 \frac{\sigma}{\sqrt{n}}$ d'où $\overline{X} + 1{,}96 \frac{\sigma}{\sqrt{n}} \geq -m \geq \overline{X} - 1{,}96 \frac{\sigma}{\sqrt{n}}$

Ce qui peut s'écrire $\overline{X} - 1{,}96 \frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + 1{,}96 \frac{\sigma}{\sqrt{n}}$

Donc après avoir prélevé un échantillon de taille *n*, on peut estimer qu'il y a 95% de chances que la moyenne de la population soit dans l'intervalle $[\overline{x} - 1{,}96 \frac{\sigma}{\sqrt{n}} ; \overline{x} + 1{,}96 \frac{\sigma}{\sqrt{n}}]$ où $\overline{x}$ représente la moyenne de l'échantillon.

Rem 1 : Ne connaissant pas σ on détermine cette valeur à l'aide d'une estimation ponctuelle.
Rem 2 : Pour les intervalles de confiance aux autres taux il faut utiliser les valeurs de *t* associées aux autres valeurs.

3.3 Estimation par intervalle de confiance d'une fréquence (ou pourcentage)

On a vu que pour *n* assez grand, la distribution d'échantillonnage des fréquences suivait approximativement une loi $N(p; \sqrt{\frac{pq}{n}})$

Il suffit alors de reprendre des calculs identiques à ceux du **3.2** pour déterminer un intervalle de confiance.

Exemple : L'intervalle de confiance au taux de 99% pour une fréquence est l'intervalle :

$$[f - 2.58\sqrt{\frac{f(1-f)}{n-1}} ; f + 2.58\sqrt{\frac{f(1-f)}{n-1}}]$$