partie 4:ECHANTILLONNAGE et ESTIMATIONS

Nous allons pouvoir répondre aux questions suivantes :

- que peut-on attendre d'un échantillon aléatoire issu d'une population connue?
- comment la population-mère peut-elle être estimée à partir d'un échantillon?

I - Théorie de l'échantillonnage

Chaque observation individuelle dans un échantillon aléatoire a la même distribution de probabilité que la population.
Echantillonnage aléatoire : soit une population de N unités statistiques sur laquelle on prélève un échantillon de taille n. Supposons que l'on dispose de la liste de toutes les unités qui constituent la population, sans omission ni répétition. Cette liste constitue une **base de sondage**. On peut attribuer à chaque individu un numéro unique puis prélever par tirage au sort n individus pour constituer un **échantillon aléatoire** où chaque unité de la population a une probabilité connue, non nulle d'être choisie.

On peut construire un échantillon aléatoire comme suit :

- **tirage sans remise** : (tirage exhaustif) les unités tirées successivement ou ensemble ne sont pas remises dans la population
- **tirage avec remise** : (tirage indépendant) chaque unité tirée au hasard dans la base de sondage est observée puis remise à la population avant qu'une autre unité soit tirée.

Taux de sondage : $\frac{n}{N}$*100 **(<10% = avec remise)**

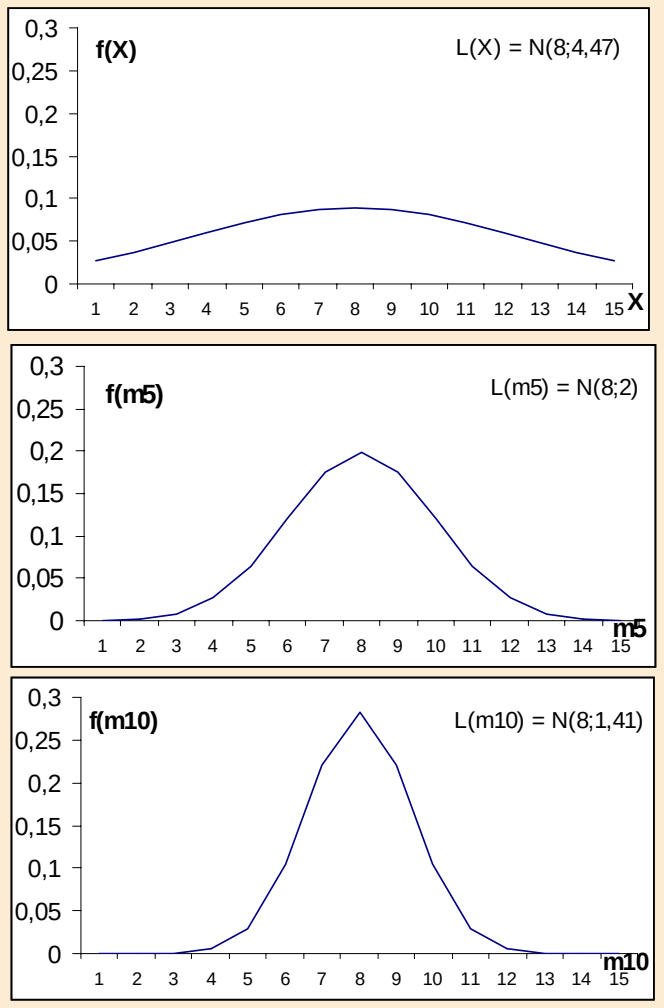
I – 1 - distribution d'échantillonnage de la moyenne

Tous les échantillons de taille n sont issus d'une population dans laquelle la variable aléatoire X est centrée sur **E(X)** = μ et dont la variance est égale à $\sigma^2(x)$.

On considère comme variable aléatoire : **la moyenne mj** de l'échantillon j de taille n.
On considère tous les échantillons de taille n (**k** échantillons possible).
Sur chaque échantillon on mesure la moyenne notée **mj** ou $\bar{x}_j$

n	nombre d'individus par échantillon
E(m)	moyenne des **k moyennes** m_j
$\sigma^2(m)$	variance des **k moyennes** m_j
$\sigma(m)$	écart-type des **k moyennes** m_j (noté aussi σ_m)

Règle de l'approximation normale : " dans les échantillons aléatoires de taille n, la moyenne de l'échantillon m varie autour de la moyenne de la population μ avec un écart-type égal à σ (m) = σ (x) / $\sqrt{n}$. Donc, quand n augmente, la distribution d'échantillonnage de m est de plus en plus concentrée autour de son objectif μ et devient de plus en plus proche de la distribution normale de Gauss. " (consulter les courbes qui suivent).



Pour un échantillonnage aléatoire et indépendant :

Espérance mathématique : **E(m) = μ = E(X)**

écart-type (tirage avec remise) : $\sigma(m) = \frac{\sigma(x)}{\sqrt{n}}$

écart-type (tirage sans remise) : $\sigma(m) = \frac{\sigma(x)}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Si L (X) = N (μ ; σ (x)) ou **Si L (X) ≠ N** **Mais avec n ≥ 30**

alors : **L (mn) = N (E (m) ; σ (m))**

et : $P[\mu - t_{(1-\alpha/2)}\,\sigma_m < m < \mu + t_{(1-\alpha/2)}\,\sigma_m] = (1-\alpha)$

C'est à dire que la moyenne observée m dans un échantillon aléatoire et indépendant issu d'une population centrée sur μ, a α% de risque de ne pas appartenir à l'intervalle de prédiction calculé à partir des données de la population.

I – 2 - distribution d'échantillonnage des fréquences

Tous les échantillons de taille n sont issus d'une population dans laquelle la variable aléatoire X suit une loi Binomiale approximée par une loi Normale centrée sur **np** et dont la variance est égale à **npq**.

On considère comme variable aléatoire : **la fréquence fj** de l'échantillon j de taille n, c'est à dire la fréquence d'apparition de la variable dans l'échantillon.
On considère tous les échantillons de taille n (**k** échantillons possible).
Sur chaque échantillon on mesure la fréquence notée **fj**

n	nombre d'individus par échantillon
E(f)	moyenne des **k fréquences** f_j
σ^2(f)	variance des **k fréquences** f_j
σ(f)	écart-type des **k fréquences** f_j (noté aussi σ_f).

Dans les échantillons aléatoires de taille n, la fréquence f de l'échantillon varie autour de la proportion p de la population avec un écart-type égal à:

$$\sigma\,(f) = \sqrt{pq/n}\ .$$

Donc, quand n augmente, la distribution d'échantillonnage de f est de plus en plus concentrée autour de p et devient plus proche de la loi Normale.

Pour un échantillonnage aléatoire et indépendant :

Espérance mathématique : **E(f) = p**

écart-type (avec remise) : $\sigma(f) = \sqrt{\frac{pq}{n}}$.

écart-type (sans remise) : $\sigma(f) = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$

Si $\dfrac{\left|\sqrt{\frac{p}{q}} - \sqrt{\frac{q}{p}}\right|}{\sqrt{n}} \le 0.34$ et $n > 5$ ou Si np et $n(1-p) \ge 5$

alors : **L (fn) = N (E (f) ; σ (f))**

et : $P\left[\, p - t_{(1-\alpha/2)}\,\sigma_f < f < p + t_{(1-\alpha/2)}\,\sigma_f \,\right] = (1-\alpha)$

C'est à dire que la fréquence observée f dans un échantillon aléatoire et indépendant issu d'une population centrée sur p, a α% de risque de ne pas appartenir à l'intervalle de prédiction calculé à partir des données de la population.

II - Les estimations

II – 1 - définitions et propriétés

On appelle **estimateur** toute fonction statistique des valeurs observées sur un échantillon utilisée pour estimer un paramètre inconnu de la population. Toute valeur prise par l'estimateur est une **estimation** du paramètre de la population Un estimateur est donc une variable aléatoire dont les propriétés sont les suivantes :

1. un estimateur est **sans biais** si son espérance mathématique est égale au paramètre que l'on cherche à estimer quelque soit n.
2. un estimateur est **convergent** (ou correct) si son espérance mathématique tend vers le paramètre que l'on cherche à estimer et si sa variance tend vers 0 lorsque la taille de l'échantillon tend augmente.
3. un estimateur est **absolument correct** s'il est sans biais et si sa variance tend vers 0 lorsque la taille de l'échantillon tend augmente.
4. un estimateur est **efficace** s'il est absolument correct et si sa variance est minimale parmi les estimateurs sans biais possibles.

II – 2 - estimations ponctuelles :

L'estimation ponctuelle est la réalisation d'un estimateur dans un échantillon donné. C'est donc la valeur que l'on attribue au paramètre inconnu que l'on cherche à définir à l'aide d'un échantillon, si l'on doit fournir une valeur unique de ce paramètre.

Estimation d'une **moyenne** : $\hat{\mu}$ = m

Estimation d'une **proportion** : $\hat{p}$ = f = $\frac{k}{n}$

Estimation d'une **variance** : $\hat{\sigma}^2(x) = \frac{ns^2(x)}{n-1} = \frac{SCEx}{n-1}$

Estimation d'un **écart-type** : $\hat{\sigma}$ (x)

II – 3 - estimation par intervalle de confiance

Il s'agit de calculer une fourchette de valeurs estimées d'un paramètre inconnu d'une population-mère, généralement centrée sur l'estimation ponctuelle de ce paramètre, et dont l'amplitude est définie par le choix d'un coefficient de confiance.

Si la loi de probabilité de l'estimateur est connue, il est possible de déterminer, à partir de la valeur calculée sur un échantillon (ou estimation) des limites entre lesquelles se trouve presque certainement comprise la vraie valeur de la caractéristique. Quand n est suffisamment grand, l'estimateur bien choisi d'une caractéristique se distribue généralement suivant une loi voisine de la loi Normale. Celle-ci est alors définie par sa moyenne et son écart-type. Soient θ la vraie valeur inconnue et d l'estimateur. Si l'on peut déterminer l'écart-type σ (d) on peut

affirmer que l'intervalle [d - $t_{1-\alpha/2}$ σ (d) ; d + $t_{1-\alpha/2}$ σ (d)] a une probabilité égale à α de ne pas contenir la vraie valeur θ.

Tout intervalle de confiance à (1 - α)% peut être reformulé sous une forme unilatérale en plaçant toute le risque d'erreur de α% d'un seul côté. Ceci signifie qu'on est beaucoup plus exigeant d'un côté alors que l'on reste très vague de l'autre.

II – 3 – 1 - estimation d'une moyenne ; variance de la population connue.

Pour un échantillonnage aléatoire et indépendant :

Si L(X) = N (μ ; σ(x)) ou Si n > 30

alors $L(m_n) = N(E(m) ; \sigma(m))$

et on obtient : $P[m - t_{1-\alpha/2}\, \sigma(m) < \mu < m + t_{1-\alpha/2}\, \sigma(m)] = 1 - \alpha$

avec $\sigma(m) = \dfrac{\sigma(x)}{\sqrt{n}}$ (avec remise)

ou $\sigma(m) = \dfrac{\sigma(x)}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$ (sans remise)

C'est à dire que l'intervalle calculé, à partir de l'observation de la variable dans un échantillon aléatoire et indépendant, a α% de risque de ne pas contenir la vraie valeur μ de la population.

II – 3 – 2 - estimation d'une moyenne; variance de la population inconnue

Pour un échantillonnage aléatoire et indépendant :

Si L(X) = N (μ ; $\hat{\sigma}$ (x)) et n < 30

alors $L(m_n) = S(\nu)\,(E(m) ; \sigma(m))$

et on obtient : $P[m - t_{1-\alpha/2}\,(\nu)\, \sigma(m) < \mu < m + t_{1-\alpha/2}\,(\nu)\, \sigma(m)] = 1 - \alpha$

ν = ddl = n – 1

avec $\sigma(m) = \dfrac{\hat{\sigma}(x)}{\sqrt{n}} = \dfrac{s(x)}{\sqrt{n-1}}$ (avec remise)

ou $\sigma(m) = \dfrac{\hat{\sigma}(x)}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}} = \dfrac{s(x)}{\sqrt{n-1}} \sqrt{\dfrac{N-n}{N-1}}$ (sans remise)

C'est à dire que l'intervalle calculé, à partir de l'observation de la variable dans un échantillon aléatoire et indépendant, a α% de risque de ne pas contenir la vraie valeur μ de la population.

N.B. Si le ddl est supérieur à 30 on utilisera $t_{1-\alpha/2}$ lu dans la table de la loi normale.

II – 3 – 3 - estimation d'une proportion

On peut poser d'une manière générale :

$$P [p_1 < p < p_2] = (1 - \alpha)$$

C'est à dire que l'intervalle calculé, à partir de l'observation de la variable dans un échantillon aléatoire et indépendant, a α% de risque de ne pas contenir la vraie valeur p de la population.

Il existe 3 cas selon la taille de l'échantillon **n** et la valeur **f = k/n**, **k** représentant le nombre de fois où le caractère est présent dans l'échantillon.

. n < 100

Le prélèvement doit être avec remise. L'estimation met en jeu la **loi F** de Fisher de telle sorte que :

$$p_1 = \frac{k}{k + (n-k+1)F_{1-\alpha/2}(\nu_1, \nu_2)}$$ avec $\nu_1 = 2 (n - k + 1)$ et $\nu_2 = 2k$

$$p_2 = \frac{(k+1)F_{1-\alpha/2}(\nu_1, \nu_2)}{n-k+(k+1)F_{1-\alpha/2}(\nu_1, \nu_2)}$$ avec $\nu_1 = 2 (k + 1)$ et $\nu_2 = 2 (n - k)$

. n ≥ 100 et 0.1 ≤ f ≤ 0.9

L'estimation met en jeu la **loi N** de Gauss de telle sorte que :

$$P [f - t_{1-\alpha/2}\, \sigma (f) < p < f + t_{1-\alpha/2}\, \sigma (f)] = (1 - \alpha)$$

écart-type (avec remise) : $\sigma(f) = \sqrt{\frac{pq}{n}}$.

écart-type (sans remise) : $\sigma(f) = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$

. n ≥100 et f ≤ 0.1

Le prélèvement doit être avec remise. L'estimation met en jeu la **loi du X^2** de telle sorte que :

$$p_1 = \frac{1}{2n} X^2_{\alpha/2}(\nu) \qquad \nu = 2k$$

$$p_2 = \frac{1}{2n} X^2_{1-\alpha/2}(\nu) \qquad \nu = 2k + 2$$

II – 3 – 4 - estimation d'une variance

Si on a n prélèvements indépendants et si la distribution de la variable suit une loi Normale dans la population.

$$P\left[\frac{\sum_{i=1}^{i=n}(xi-\bar{x})^2}{X^2_{1-\alpha/2}(\nu)} < \sigma^2(x) < \frac{\sum_{i=1}^{i=n}(xi-\bar{x})^2}{X^2_{\alpha/2}(\nu)}\right] = 1-\alpha \qquad \nu = n-1$$

C'est à dire que l'intervalle calculé, à partir de l'observation de la variable dans un échantillon aléatoire et indépendant, a α% de risque de ne pas contenir la vraie valeur de la variance $\sigma^2(x)$ de la population.

II – 3 – 5 - estimation d'un écart-type

Si on a n prélèvements aléatoires et indépendants (n > 15) et si la distribution de la variable suit une loi Normale dans la population. Les limites de l'intervalle de confiance d'un écart-type sont les racines carrées des limites correspondantes de l'intervalle de confiance de la variance.

II – 3 – 6 - précision et taille d'échantillon

La précision d'une mesure ou marge d'erreur peut s'apprécier en termes de **marge d'erreur absolue** ou alors **marge d'erreur relative**. Si la précision d'une estimation est fixée on pourra calculer la taille de l'échantillon nécessaire à l'estimation.

La marge d'erreur absolue d'une estimation est utilisée dans le cas d'une estimation par intervalle de confiance d'un paramètre; elle est égale à la demi-différence entre les limites supérieure et inférieure de l'intervalle.

Exemple : marge d'erreur absolue de l'estimation d'une moyenne :

$$MEabs = t_{1-\alpha/2} * \sigma(m) = \frac{t_{1-\alpha/2} * \sigma(x)}{\sqrt{n}}$$

La marge d'erreur relative d'une estimation est égale au rapport entre la marge d'erreur absolue et la valeur sur laquelle est centré l'intervalle (multiplié par 100 pour un pourcentage).

Exemple : marge d'erreur relative de l'estimation d'une moyenne :

$$MErel = 100 * \frac{t_{1-\alpha/2} * \sigma(m)}{m} = 100 * \frac{t_{1-\alpha/2} * \sigma(x)}{m\sqrt{n}}$$