Chapitre 5 :

L'estimation statistique

Un aspect important de l'inférence statistique est celui d'obtenir à partir des résultats trouvés sur l'échantillon d'une population des estimations fiables de certain paramètre de cette population tels que la moyenne, la variance et la proportion.

Ces estimations peuvent s'exprimer soit par une seule valeur, on parlera alors d'estimation ponctuelle, soit par un intervalle, on parlera dans ce cas d'estimation par intervalle de confiance.

I- L'estimation ponctuelle :

Cette technique consiste à estimer un paramètre θ de la population à l'aide d'un seul nombre déduit des résultats de l'échantillon, ce nombre est appelé estimateur ponctuel de θ et sera noté $\hat{\theta}$.

Un estimateur dois posséder certaines qualités ou propriétés afin de fournire une bonne estimation. Cet estimateur doit être sans biais efficace et convergent.

-Estimateur sans biais :$E(\hat{\theta}) = \theta$

-Estimateur efficace : si sa variance est la plus faible parmi les variances des autres estimateurs sans biais

- estimateur convergent : $\lim_{n \to +\infty} V(\hat{\theta}) = 0$

Exercice :

- Montrer que $\overline{X}$ est un estimateur sans biais et convergent de m
- Montrer que f est un estimateur sans biais et convergent de p

$$E(\overline{X}) = \frac{\sum_1^n E(xi)}{n} = \frac{\sum_1^n m}{n} = \frac{n.m}{n} = m$$

$$V(\overline{X}) = \frac{\sum_1^n V(xi)}{n^2} = \frac{\sum_1^n \sigma^2}{n^2} = \frac{n.\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to+\infty]{} \text{limite V } (\overline{X}) = 0$$

Puisque $f \to N(p, \sqrt{\frac{pq}{n}})$, E (f) = p

Et V(f) = $\frac{pq}{n} \xrightarrow[n\to+\infty]{}$ limite V (f)= 0

Remarque : les estimations ponctuelles ne fournissent aucune information concernant la précision des estimations, c'est à dire qu'elles ne tiennent pas fluctuations d'échantillonnage.

II- L'estimation par intervalle de confiance :

L'estimation par intervalle d'un paramètre inconnu θ consiste à calculer à partir d'un estimateur choisi $\hat{\theta}$, un intervalle dans lequel on a un pourcentage de chance d'y trouver correspondante du paramètre θ .

L'intervalle de confiance est défini par deux limites auxquelles est associée une certaine probabilité de contenir la vraie valeur du paramètre θ .

$$P(LI \le \theta \le LS) = 1 - \alpha$$

LI : Limite inférieure de l'intervalle de confiance

LS : Limite supérieure de l'intervalle de confiance

1- Estimation d'une moyenne :

Trois cas peuvent se présenter selon la loi de probabilité de X :

1er cas : σ connu alors la loi de probabilité de X sera :

$$\overline{X} \to N(m, \frac{\sigma}{\sqrt{n}}) d'où \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

$$P\left(\left| \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour m.

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Application :

Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Des tests effectués il y 2 ans permettent de considérer que le poids d'une boite est distribué normalement avec une variance de 9 grammes.

Un test sur un échantillon de 16 boites a donné un poids moyen de 219 gr.

Estimer par un intervalle de confiance le poids moyen des boites avec un niveau de confiance de 95%.

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est connu, ($\sigma = \sqrt{9}$ = 3) d'où $\overline{X}$ → N (m, σ / $\sqrt{n}$)

$$\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

$$P\left(\left| \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

$1 - \alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha_{/2} = 0.025 \rightarrow F(t_{\alpha/2}) = 0.975 \rightarrow t_{\alpha/2} = 1.96$

La limite inférieure de l'intervalle de confiance est donc :

$LI = \overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 219 - 1.96 \times 3 / \sqrt{16} = 217.53$

Et la limite supérieure de l'intervalle de confiance est donc :

$LS = \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 219 + 1.96 \times 3 / \sqrt{16} = 220.47$

2ème cas : σ inconnu et $n \geq 30$ alors la loi de probabilité de X sera :

$$\overline{X} \rightarrow N(m, \frac{S}{\sqrt{n}})$$

$$P\left(\left|\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \left|\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour m.

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

D'ou

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \quad ; \quad \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

Application : Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Un test sur un échantillon de 36 boites a donné un poids moyen de 219 grammes avec un écart type de 1.5 grammes Estimer par intervalle de confiance le poids moyen des boites avec un niveau de confiance de 99%.

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est inconnu, et n = 36 >30 d'où $\overline{X}$ → N (m, S / $\sqrt{n}$)

$$\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}} \rightarrow N(0,1)$$

$$P\left(\left|\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right]$$

$1 - \alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha_{/2} = 0.005 \rightarrow F(t_{\alpha/2}) = 0.995 \rightarrow t_{\alpha/2} = 2.58$

La limite inférieure de l'intervalle de confiance est donc :

LI = $\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ = 219 – 2.58 x 1.5 / $\sqrt{36}$ = 218.355

Et la limite supérieure de l'intervalle de confiance est donc :

LS = $\overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ = 219 + 2.58 x 1.5 / $\sqrt{36}$ = 219.645

3ème cas : σ inconnu et n<30 alors la loi de probabilité de X sera :

$$\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \rightarrow T(n-1)$$

$$P\left(\left| \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi de Student. Par la suite on pourra dégager un intervalle de confiance pour m.

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right]$$

Application : Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Un test sur un échantillon de 9 boites a donné un poids moyen de 219 grammes avec un écart type de 1.5 grammes

Estimer par intervalle de confiance le poids moyen des boites avec un niveau de confiance de 99%

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est inconnu, et n = 9 < 30 d'où $\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$ → T(n-1)

S = 1.5

$$P\left(\left|\frac{\overline{X}-m}{\frac{s}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

$1 - \alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha_{/2} = 0.005$ et ddl = 8 $\rightarrow t_{\alpha/2} = 3.355$

La limite inférieure de l'intervalle de confiance est donc :

LI = $\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ = 219 – 3.355 x 1.5 / $\sqrt{9}$ = 217.322

Et la limite supérieure de l'intervalle de confiance est donc :

LS = $\overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ = 219 + 3.355 x 1.5 / $\sqrt{9}$ = 220.677

Remarque : L'expression des bornes de l'intervalle de confiance pourrait être modifiée si on se trouve dans l'obligation d'apporter un ajustement à la variance dans le cas ou le tirage de l'échantillon se fait sans remise avec un taux de sondage supérieur à10%.

2-Estimation de la différence de deux moyennes :

La différence de deux moyennes d'un même caractère X mesuré sur deux populations différentes ayant un écart type connu au niveau de chaque population peut être estimée comme suit :

Puisque

$$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

$$P\left(\left|\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \left|\frac{(\overline{X}_1-\overline{X}_2)-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour $m_1 - m_2$.

$$\overline{X}_1-\overline{X}_2-t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \prec m_1-m_2 \prec \overline{X}_1-\overline{X}_2+t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

D'ou

$$m_1-m_2 \in \left[\overline{X}_1-\overline{X}_2-t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad ; \quad \overline{X}_1-\overline{X}_2+t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$ **3-**

Estimation d'une proportion :

Pour déterminer un intervalle de confiance pour une proportion on utilise la distribution suivante :

$$f \rightarrow N(p,\sqrt{\frac{pq}{n}}) \qquad d'ou \qquad \frac{f-p}{\sqrt{\frac{pq}{n}}} \rightarrow N(0,1) \qquad avec\ n \geq 30$$

$$P\left(\left|\frac{f-p}{\sqrt{\frac{pq}{n}}}\right| \leq t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale

$$-t_{\frac{\alpha}{2}} \leq \frac{f-p}{\sqrt{\frac{pq}{n}}} \leq t_{\frac{\alpha}{2}}$$

$$f-t_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} \leq p \leq f+t_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}$$

On remarque que les bornes de l'intervalle de confiance dépendent de p, on va alors estimer p par f uniquement au niveau des deux bornes de l'intervalle.

$$f-t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f+t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

Exercice :

Dans une banque, on a effectué un sondage pour connaître l'opinion des clients sur un nouveau service aux agences. D'une liste de 6000 clients de la banque on extrait 150, sur ces 150 clients interrogés 45 étaient satisfaits de ce service. Déterminer un intervalle de confiance pour la vraie proportion des clients qui sont satisfaits de ce nouveau service avec un niveau de confiance de 99%.

Corrigé :

Soit p la vraie proportion des clients satisfaits et f la proportion empirique de ces clients observé sur l'échantillon

$$f \to N(p, \sqrt{\frac{pq}{n}}) \ \ d'où \ \ \frac{f-p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

f= cas favorables/taille de l'échantillon= k/n = 45/150 = 0.3 → q= 1-p = 0.7

$$P\left(\left|\frac{f-p}{\sqrt{\frac{pq}{n}}}\right| \le t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$f - t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} \le p \le f + t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

$1 - \alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha/2 = 0.005 \rightarrow F(t_{\alpha/2}) = 0.995 \rightarrow t_{\alpha/2} = 2.58$

La limite inférieure de l'intervalle de confiance est donc :

$$LI = f - t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} = 0.3 - 2.58 \times \sqrt{\frac{0.3x0.7}{150}} = 0.2035$$

Et la limite supérieure de l'intervalle de confiance est donc :

$$LS = f + t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} = 0.3 + 2.58 \times \sqrt{\frac{0.3x0.7}{150}} = 0.396$$

3-Estimation de la différence de deux proportions :

La différence de deux proportions ($f_1 - f_2$) relatives à un même caractère mesuré sur deux échantillons n_1 et n_2 ayant chacun une taille supérieure à 30 et extraits de deux populations différentes peut être estimée comme suit :

Puisque

$$\frac{(f_1-f_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \to N(0,1)$$

$$P\left(\left|\frac{(f_1-f_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right| \le t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale

$$-t_{\frac{\alpha}{2}} \le \frac{(f_1-f_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \le t_{\frac{\alpha}{2}}$$

$$f-t_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \le p_1-p_2 \le f+t_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

On remarque que les bornes de l'intervalle de confiance dépendent de p_1 et p_2, on va alors estimer p_1 par f_1 et p_2 par f_2 uniquement au niveau des deux bornes de l'intervalle.

$$(f_1-f_2)-t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}} \le (p_1-p_2) \le (f_1-f_2)+t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}$$

III- La détermination de la taille de l'échantillon :

La taille de l'échantillon est liée à la marge d'erreur E qu'on va tolérer c'est à dire la différence en valeur absolue entre le paramètre à estimer et son estimateur.

$$E=\left|\theta-\hat{\theta}\right|$$

1- Cas d'une moyenne :

On sait que : $\overline{X}-t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \prec m \prec \overline{X}+t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

Donc : $\left|m-\overline{X}\right| \le t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

Ainsi pour un niveau de confiance 1-α la marge d'erreur E est au plus égale à :

$$t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Cette valeur quantifie l'erreur attribuable aux fluctuations d'échantillonnage. Ainsi on peut fixer la marge d'erreur E qu'on ne veut pas excéder et déterminer la taille minimale requise de l'échantillon.

$$E=t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\Rightarrow n=\left(\frac{t_{\frac{\alpha}{2}}\sigma}{E}\right)^2$$

Ainsi plus la marge d'erreur est faible plus la taille de l'échantillon est élevée.

Exemple : On veut estimer la durée de vie moyenne d'un dispositif électronique. D'après le bureau de RD l'écart type de la durée de vie de ce dispositif serait 100 heures. Déterminer le nombre d'essais requis pour estimer avec un niveau de confiance de 95%, la durée de vie moyenne d'une grande production de ce dispositif de sorte que la marge d'erreur dans l'estimation n'excède pas 50 heures.

E= 50, σ = 100, 1-α = 0.95 → $t_{\alpha/2}$ = 1.96 → n = (1.96 x 100/50)² = 15.366≈16

2- Cas d'une proportion :

Dans le cas de l'estimation d'une proportion on sait que :

$$|f-p|\leq t_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

$$E=t_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

La taille de l'échantillon serait donc :

$$n=\frac{t_{\alpha/2}^2\ p\ (1-p)}{E^2}$$

Si on a une information sur la valeur approximative de p on va la remplacer dans la formule, sinon on remplace p par 0.5.

Application :

On veut effectuer un sondage auprès des 8000 étudiants d'une faculté pour estimer le pourcentage des fumeurs. Déterminer la taille de l'échantillon requise pour estimer avec un niveau de confiance de 95% cette proportion avec une marge d'erreur n'excédant pas 5% dans les deux cas suivants :

1- Une enquête similaire effectuée il y a 3 ans indiqua que 32% des étudiants fumaient

2- En supposant qu'on n'a aucune information préalable sur p.

Corrigé :

1- P= 0.32 → 1-p = 0.68, E =0.05

$1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha_{/2} = 0.025 \rightarrow F(t_{\alpha/2}) = 0.975 \rightarrow t_{\alpha/2} = 1.96$

$$\rightarrow n = \frac{t_{\alpha/2}^2 \; p\,(1-p)}{E^2} = 1.96^2 \text{x } 0.32\text{x}0.68 / 0.05^2 = 334.37 \approx 335$$

2- Si p est inconnue, on prend p =0.5 d'où $n = 1.96^2\text{x}0.5\text{x}0.5 / 0.05^2 = 384.16 \approx 385$