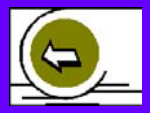
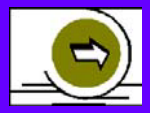
STATISTIQUES INFERENTIELLES

Version 2 - 02/05

■ Outils de navigation


Affiche la page précédente


Affiche la boîte de dialogue "Recherches"


Affiche la page suivante


Affiche la page d'accueil (permet de changer de cours)

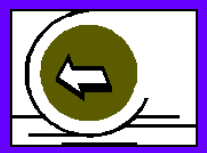
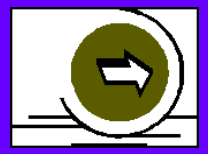

Affiche le plan du cours

■ Restrictions

Ce cours interactif est prévu pour une utilisation sur machine (en ligne ou sur Cd-rom). Il n'est donc pas possible de lancer une impression à partir du cours. Le cours imprimable permet de conserver une trace papier, il doit être complété à l'aide du cours interactif. Le support papier permet également de compléter, a posteriori, les connaissances acquises à l'aide du cours interactif.

Il n'est pas possible de modifier une partie du cours ou les commentaires, seuls les champs des exercices peuvent être remplis mais ils ne seront pas conservés lors d'une utilisation future.

Il n'est pas possible de copier le cours ou toute partie (graphiques, tableaux, vidéos …).

Pré-requis indispensables pour ce cours :

- Les statistiques descriptives
- Les lois de probabilités

iutenligne.net


Commencer le cours ...

Cours imprimable

Exercices corrigés

Tables statistiques

Q.C.M.

STATISTIQUES INFERENTIELLES

1. Echantillonnage d'une population

Voici quelques raisons d'échantillonner une population :

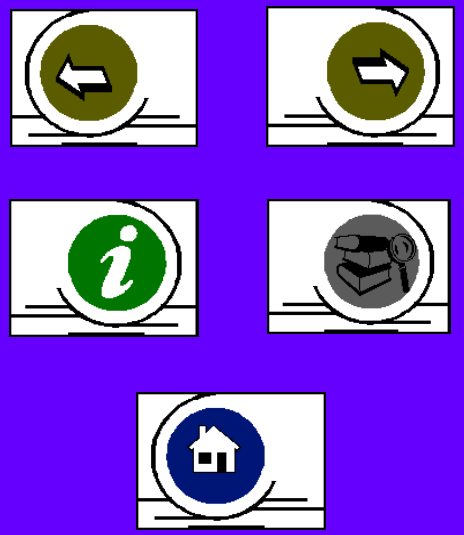
- Le budget est limité et le coût de collecte important.
- Il faut user ou détruire les éléments pour en mesurer la qualité (par exemple résistance d'un article…).
- Les résultats sont plus fiables sur un nombre peu élevé d'observations.

Pour prélever un échantillon, il existe plusieurs méthodes :

- **Echantillon au hasard**
  Chacun des N éléments de la population a les mêmes chances d'être tiré.

- **Méthode des quotas**
  L'échantillon représente en miniature la population vis-à-vis des caractéristiques qui influent sur le phénomène étudié.

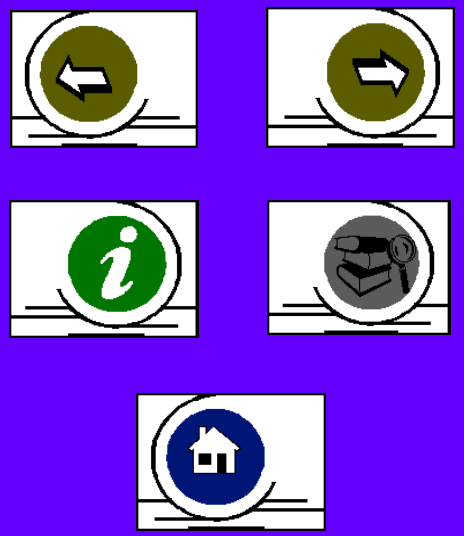
- **Méthode en cascade**
  On tire au hasard un échantillon de quelques villes, puis dans chaque ville un échantillon de quelques quartiers, puis dans chaque quartier un échantillon de quelques individus. Tous les individus de ces groupes sont sondés.



Lorsque le tirage de l'échantillon se fait sans remise, on dit que l'échantillon est exhaustif, au contraire, lorsque le tirage de l'échantillon se fait avec remise, on dit que l'échantillon est non exhaustif.

STATISTIQUES INFERENTIELLES

1. Echantillonnage d'une population

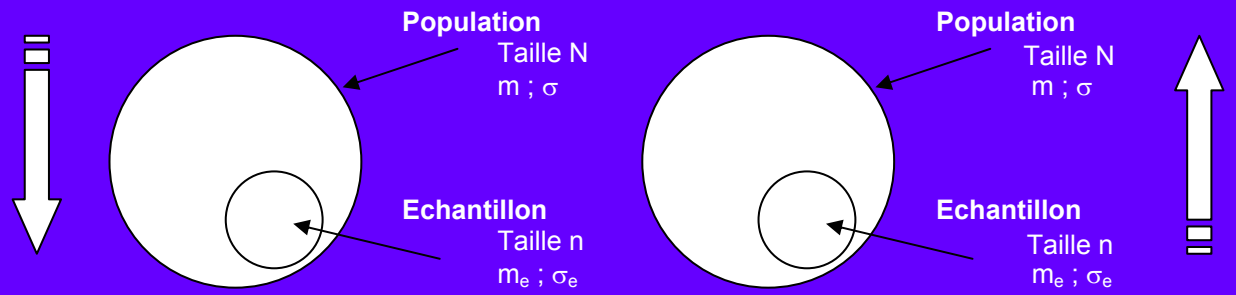


Bien sûr les résultats obtenus sur l'échantillon ne sont pas identiques à ceux qui auraient pu l'être sur la population. Le but du cours est de fournir des outils permettant de déduire les paramètres de la population, connaissant ceux de l'échantillon.

Il existe cependant deux sources de distorsions importantes :

- **Le biais sur les mesures**
  Les résultats sont mesurés avec des erreurs (l'appareil qui prend les mesures est défectueux, la personne est incompétente…)
- **Le biais de recrutement**
  L'échantillon prélevé n'est pas représentatif de la population vis-à-vis du caractère étudié.

Etant donnée une population de taille N et un échantillon de taille n, deux configurations sont possibles :

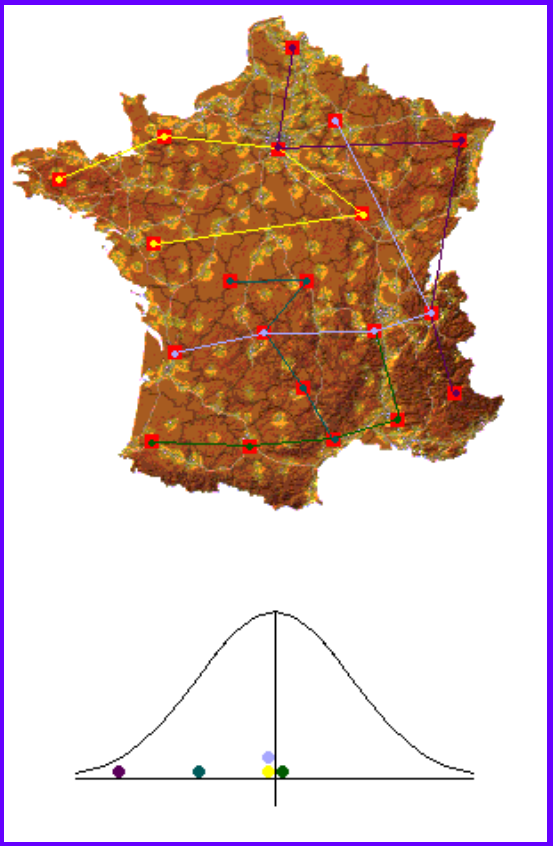


Induction : on déduit les paramètres de l'échantillon de ceux de la population.

Estimation : on estime les paramètres de la population de ceux de l'échantillon.

STATISTIQUES INFERENTIELLES

2. Induction



Dans ce paragraphe, on considère que les paramètres de la population sont connus (moyenne, écart-type ...).

On considère tous les échantillons de taille n que l'on peut extraire d'une population de taille N, de moyenne m et d'écart type σ.

Pour chaque échantillon, on peut calculer la moyenne d'une certaine propriété qui varie d'un échantillon à l'autre.

On obtient alors, une variable aléatoire $\overline{X}$, égale à la moyenne des éléments sur tout échantillon de taille n, dont la loi de probabilité est appelée distribution d'échantillonnage des moyennes.

Remarque :

On peut aussi définir F la variable aléatoire égale à la proportion de réussite sur tout échantillon de taille n, dont la loi de probabilité est appelée distribution d'échantillonnage des fréquences.

Exemple :

Soient les notes de mathématiques d'un étudiant : 4 ; 5 ; 8 ; 10 ; 12 ; 13.

1. Calculer la moyenne et l'écart type de la population des notes.
2. Former tous les échantillons exhaustifs possibles de taille 2.
3. Calculer l'espérance et l'écart type de la distribution d'échantillonnage des moyennes.
4. Recommencer dans le cas d'un échantillon non exhaustif.

STATISTIQUES INFERENTIELLES

2. Induction

Solution :

1. On a m = ☐ et σ = ☐ . Solution

2.

4 ; 5	m_e = 4,5	5 ; 8	m_e = 6,5	8 ; 12	m_e = 10
4 ; 8	m_e = 6	5 ; 10	m_e = 7,5	8 ; 13	m_e = 10,5
4 ; 10	m_e = 7	5 ; 12	m_e = 8,5	10 ; 12	m_e = 11
4 ; 12	m_e = 8	5 ; 13	m_e = 9	10 ; 13	m_e = 11,5
4 ; 13	m_e = 8,5	8 ; 10	m_e = 9	12 ; 13	m_e = 12,5

3. La variable aléatoire $\overline{X}$, égale à la moyenne des éléments sur tout échantillon exhaustif de taille 2 suit la loi de probabilité :

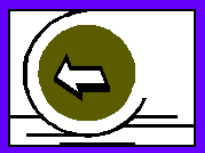
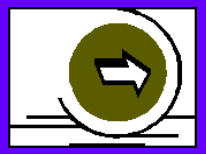
4,5	6	6,5	7	7,5	8	8,5	9	10	10,5	11	11,5	12,5
2/30	2/30	2/30	2/30	2/30	2/30	4/30	4/30	2/30	2/30	2/30	2/30	2/30

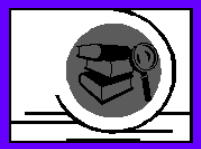
Donc $E(\overline{X})$ = ☐ et $\sigma(\overline{X})$ = ☐ . Solution

4. De manière identique, la variable aléatoire $\overline{X}$, égale à la moyenne des éléments sur tout échantillon non exhaustif de taille 2 suit la loi de probabilité :

4	4,5	5	6	6,5	7	7,5	8	8,5	9	10	10,5	11	11,5	12	12,5	13
1/36	2/36	1/36	2/36	2/36	2/36	2/36	3/36	4/36	4/36	3/36	2/36	2/36	2/36	1/36	2/36	1/36

Donc $E(\overline{X})$ = ☐ et $\sigma(\overline{X})$ = ☐ . Solution




STATISTIQUES INFERENTIELLES

2. Induction

Exercice

⊙ Cas d'une moyenne

On considère tous les échantillons non exhaustifs de taille n que l'on peut extraire d'une population de taille N, de moyenne m et d'écart-type σ. On appelle X la variable aléatoire décrivant le caractère étudié sur la population.

On considère la variable aléatoire $\overline{X}$, égale à la moyenne des éléments sur tout échantillon de taille n ; on s'intéresse donc à la distribution d'échantillonnage des moyennes.

Un n-échantillon est la réalisation d'un n-uplet $(X_1 ; \dots ; X_n)$ les X_i étant des variables aléatoires indépendantes de même loi que X. On a alors $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.

$$E(\overline{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} \cdot E\left(\sum X_i\right) = \frac{1}{n} \cdot \sum E(X_i) = \frac{n \cdot E(X)}{n} = m.$$

et

$$V(\overline{X}) = \left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \cdot V\left(\sum X_i\right) = \frac{1}{n^2} \cdot \sum V(X_i) = \frac{n \cdot V(X)}{n^2} = \frac{V(X)}{n}.$$

- Pour un échantillon non exhaustif de taille n, on a : $E(\overline{X}) = m$ et $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$.

- Pour un échantillon exhaustif de taille n, on a : $E(\overline{X}) = m$ et $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}} \times \sqrt{\dfrac{N - n}{N - 1}}$.

Le théorème de la limite centrale nous permet d'affirmer que pour de grandes valeurs de n ($n \geq 30$) la distribution d'échantillonnage des moyennes est approximativement la loi normale $\mathcal{N}\left(m ; \dfrac{\sigma}{\sqrt{n}}\right)$ [respectivement $\mathcal{N}\left(m ; \dfrac{\sigma}{\sqrt{n}} \times \sqrt{\dfrac{N - n}{N - 1}}\right)$].

STATISTIQUES INFERENTIELLES

2. Induction

⊙ Cas d'une proportion

Soit une population de taille N avec une probabilité de réussite p et une probabilité d'échec (1 – p).

Considérons tous les échantillons non exhaustifs de taille n extraits de cette population et pour chaque échantillon, on détermine la proportion de succès.

On considère Y la VA égale au nombre de succès sur un échantillon de taille n donné. Y suit la loi binomiale $\mathcal{B}(n ; p)$.

Si $n \geq 30$ on peut considérer que Y suit approximativement la loi normale $\mathcal{N}(n \times p ; \sqrt{n \times p \times (1 - p)})$

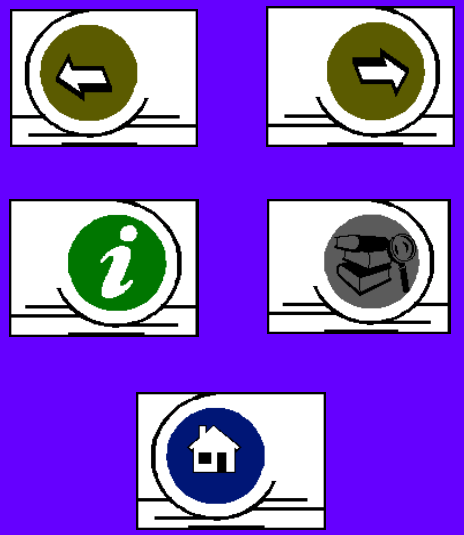
On a alors $F = \frac{Y}{n}$ qui représente la variable aléatoire égale à la proportion de réussites sur l'échantillon.

On obtient alors une variable aléatoire F telle que :

Lorsque n est assez grand ($n \geq 30$) la loi de F est approximativement $\mathcal{N}\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$ [respectivement $\mathcal{N}\left(p ; \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}\right)$].

Donc :

- Pour un échantillon non exhaustif : $E(F) = p$ et $\sigma(F) = \sqrt{\frac{p(1-p)}{n}}$
- Pour un échantillon exhaustif : $E(F) = p$ et $\sigma(F) = \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$



STATISTIQUES INFERENTIELLES

2. Induction

Exercice :

Une enquête écrite est réalisée sur un échantillon de 250 clients ciblés. Une étude statistique a montré que le taux de défection moyen, dans ce genre d'enquête, est de 15 %.

Les questions peuvent alors être :

- Quelle est la gamme plausible des défections sur l'échantillon ?
- L'échantillon est-il bien ciblé ?

Solution :

On considère Y la VA égale au nombre de défections sur l'échantillon. Y suit la loi binomiale $\mathcal{B}(250 ; 0,15)$. Comme $n \geq 30$ on peut considérer que Y suit la loi normale $\mathcal{N}(250 \times 0,15 ; \sqrt{250 \times 0,15 \times (1 - 0,15)})$

On a alors $F = \frac{Y}{n}$ qui représente la variable aléatoire égale à la proportion de défections sur l'échantillon.

On a F qui suit la loi $\mathcal{N}(0,15 ; 0,02)$.

Si on veut avoir une plage de valeurs où doit se trouver la proportion f_e de l'échantillon, il faut choisir un risque, par exemple 4,5 %, et alors :

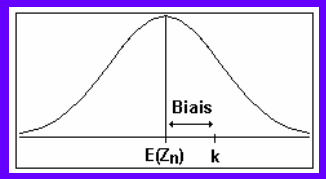
$P(0,15 - a \leq F \leq 0,15 + a) = 0,955$ si et seulement si $a = 0,04$.

Donc, au risque de 4,5 %, f_e doit se trouver dans l'intervalle [11% ; 19%]. Cet intervalle est appelé intervalle de confiance à 95,5 % (ou au seuil de risque 4,5 %).

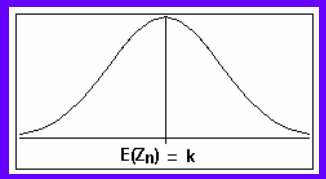
Après l'enquête, il est alors possible de valider, ou non, l'échantillon.

STATISTIQUES INFERENTIELLES

3. Estimation ponctuelle



Estimateur biaisé.



Estimateur sans biais.

L'objet de l'estimation est d'obtenir les paramètres d'une population à partir d'observations établies sur un échantillon de cette population.

On appelle estimateur d'un paramètre k, une variable aléatoire dont le but est d'estimer au mieux la valeur du paramètre k.

On dit que l'estimateur Z_n est un estimateur sans biais du paramètre k si $E(Z_n) = k$.

Si de plus, $V(Z_n) \to 0$ quand $n \to \infty$, l'estimateur Z_n converge (en probabilité) vers k. L'estimateur Z_n est alors un estimateur absolument correct du paramètre k.

⊙ Estimation ponctuelle d'une moyenne

On note m la moyenne de la population mère (paramètre inconnu) et X la variable aléatoire décrivant le caractère étudié.

On prélève au hasard un échantillon non exhaustif de taille n.

Un estimateur naturel de m est la variable aléatoire $\overline{X}$ telle que :

$\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ où $(X_1 ; X_2 ; \ldots ; X_n)$ forme un n-échantillon c'est à dire que les variables aléatoires X_i sont indépendantes et de même loi que X.

L'estimateur $\overline{X}$ est sans biais : $E(\overline{X}) = E\left(\dfrac{\sum X_i}{n}\right) = \dfrac{1}{n} \cdot E\left(\sum X_i\right) = \dfrac{1}{n} \cdot \sum E(X_i) = \dfrac{n \cdot E(X)}{n} = m.$

La réalisation de $\overline{X}$ sur un échantillon donné est m_e la moyenne sur l'échantillon.

Une estimation ponctuelle de m est alors m_e

STATISTIQUES INFERENTIELLES

3. Estimation ponctuelle

Remarque : $V(\bar{X}) = \left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \cdot V\left(\sum X_i\right) = \frac{1}{n^2} \cdot \sum V(X_i) = \frac{n \cdot V(X)}{n^2} = \frac{V(X)}{n} \to 0$ *quand* $n \to +\infty$.

Exemple :

On s'intéresse à la durée d'attente à un centre de renseignements téléphoniques, avant que la communication ne soit amorcée. On a prélevé, au hasard, la durée d'attente (en s) de 100 contacts :

Durée(s)	[7,5 ; 11,5[	[11,5 ; 15,5[	[15,5 ; 19,5[	[19,5 ; 23,5[	[23,5 ; 27,5[
Effectif	12	25	36	18	9

La moyenne sur l'échantillon est m_e = ☐ donc on peut dire que l'estimation ponctuelle de la durée d'attente moyenne sur la population des appels est de ☐ s.

Solution

⊙ Estimation ponctuelle d'une proportion

On sait que la population mère contient une proportion p (inconnue) d'individus ayant une propriété donnée.

On prélève, au hasard, un échantillon non exhaustif de taille n et on note Y la variable aléatoire égale au nombre d'individus ayant la propriété dans l'échantillon.

Y suit la loi binomiale $\mathcal{B}$ (n ; p) qui peut être approchée pour $n \geq 30$ par $\mathcal{N}$ $(np ; \sqrt{n \times p \times (1 - p)})$.

La variable aléatoire $F = \frac{Y}{n}$ suit la loi $\mathcal{N}$ $\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$ donc F est un estimateur sans biais de p (en effet E(F) = p).

La réalisation de F sur un échantillon donné est f_e la proportion sur l'échantillon.

Une estimation ponctuelle de p est f_e

STATISTIQUES INFERENTIELLES

3. Estimation ponctuelle

Exemple :

Dans un centre de renseignements téléphoniques, une étude statistique a été réalisée pour déterminer le pourcentage de communications n'aboutissant pas pour des raisons techniques.

Sur un échantillon de 2 000 communications, 60 n'ont pas abouti. Estimer le pourcentage de communications qui n'ont pas abouti.

On a $f_e =$ ☐ . Donc l'estimation ponctuelle sur la population est de ☐ %.

Solution

⊙ Estimation ponctuelle d'une variance – d'un écart-type

La population admet une moyenne m et un écart-type σ inconnus. On considère un échantillonnage non exhaustif.

Un estimateur logique de la variance σ^2, est : $S_n^2 = \frac{1}{n}\cdot\sum(X_i - \overline{X})^2$.

Or $E(S_n^2) = \frac{n-1}{n}\sigma^2$, il y a donc un biais.

Pour obtenir un estimateur sans biais de σ^2, on prend $\frac{n-1}{n} S_n^2$; on le note S'^2_n.

Une estimation ponctuelle de σ est $s = \sqrt{\frac{n}{n-1}\cdot\sigma_e}$ (σ_e = écart-type de l'échantillon).

Définition : s est appelée déviation standard et est souvent notée σ_{n-1} sur les calculatrices.

Exemple : reprenons les durées d'attente aux standards téléphoniques, $\sigma_e = 4{,}48$ donc une estimation ponctuelle sur la population de l'écart-type des durées d'attente est s = 4,5.

STATISTIQUES INFERENTIELLES

4. Estimation par intervalle de confiance

L'estimation ponctuelle dépend, bien sûr, de l'échantillon choisi, ce qui sera toujours le cas mais ne donne aucune information sur la pertinence du résultat.

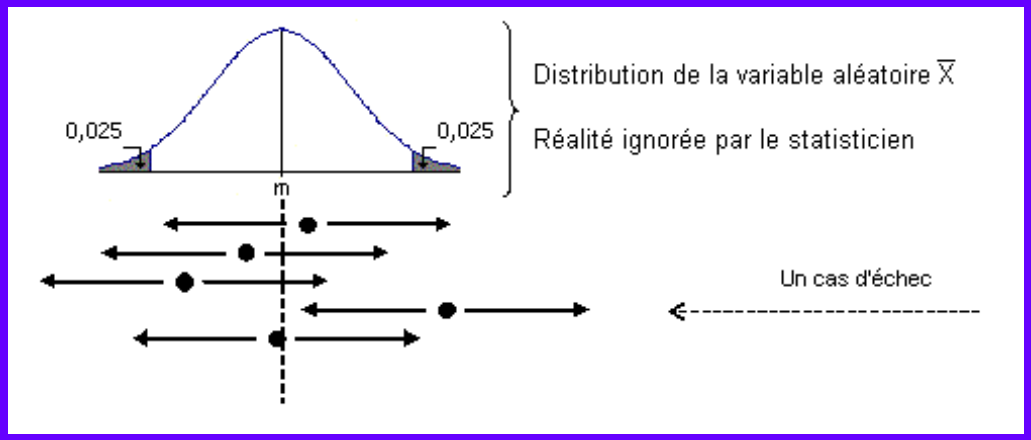
L'estimation par intervalle de confiance va nous permettre de donner une gamme de valeurs susceptibles d'être prises par le paramètre. On lui affecte de plus un coefficient de crédibilité appelé coefficient de confiance.

Exemple :

Calculer un intervalle de confiance à 95 % de la moyenne (on dit aussi au seuil de risque de 5 %).

Ceci veut dire que l'intervalle cherché doit vérifier la propriété suivante :

- Si on prélève un grand nombre d'échantillons de même taille, dans la même population, on trouve pour chacun d'eux un intervalle de confiance différent mais 95 % d'entre eux contiennent la vraie valeur du paramètre à estimer, ici la moyenne.
- Pour un échantillon donné, le risque que le paramètre à estimer soit en dehors de l'intervalle est de 5 %.



Le paramètre m appartient à 95 % des intervalles.

STATISTIQUES INFERENTIELLES

4. Estimation par intervalle de confiance

⊙ Intervalle de confiance d'une moyenne

On note X la variable aléatoire mesurant le caractère étudié sur la population mère et on considère un échantillonnage non exhaustif.

Cas 1 : X suit la loi $\mathcal{N}(m ; \sigma)$ et σ est connu

On sait qu'un estimateur de m est $\overline{X}$ qui suit la loi $\mathcal{N}\left(m ; \frac{\sigma}{\sqrt{n}}\right)$.

Il faut trouver l_α tel que :

$P(\overline{X} - l_\alpha \leq m \leq \overline{X} + l_\alpha) = 1 - \alpha$ si et seulement si $P(-l_\alpha \leq \overline{X} - m \leq l_\alpha) = 1 - \alpha$

si et seulement si $P\left(\frac{-l_\alpha}{\sigma/\sqrt{n}} \leq \frac{\overline{X} - m}{\sigma/\sqrt{n}} \leq \frac{l_\alpha}{\sigma/\sqrt{n}}\right) = 1 - \alpha$, on pose $T = \frac{\overline{X} - m}{\sigma/\sqrt{n}}$

si et seulement si $P\left(T \leq \frac{l_\alpha}{\sigma/\sqrt{n}}\right) = \frac{1 + (1 - \alpha)}{2}$ où T suit la loi normale centrée réduite.

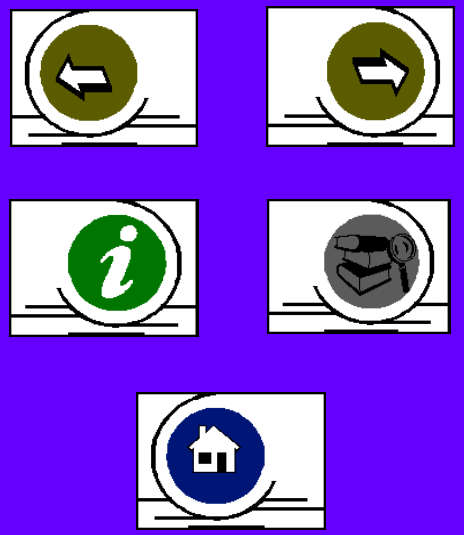
Par lecture sur la table de la loi normale centrée réduite on obtient une valeur t_α :

$l_\alpha = \frac{\sigma}{\sqrt{n}} \cdot t_\alpha$.

En remplaçant $\overline{X}$ par sa réalisation sur l'échantillon de taille n :

$P(m_e - \frac{\sigma}{\sqrt{n}} \cdot t_\alpha \leq m \leq m_e + \frac{\sigma}{\sqrt{n}} \cdot t_\alpha) = 1 - \alpha$

On a donc l'intervalle de confiance : $\left[m_e - \frac{\sigma}{\sqrt{n}} t_\alpha ; m_e + \frac{\sigma}{\sqrt{n}} t_\alpha\right]$.



STATISTIQUES INFERENTIELLES

4. Estimation par intervalle de confiance

Exemple :

Reprenons l'exemple des durées d'attente aux standards téléphoniques. Donnons un intervalle de confiance bilatéral à 95 % de la moyenne des temps d'attente sur la population des appels.

On suppose que le temps d'attente sur la population mère suit la loi normale $\mathcal{N}(m ; \sigma)$ où σ est connu et vaut 4.

On sait alors que la variable aléatoire $\overline{X}$ suit la loi normale $\mathcal{N}(m ; \frac{4}{\sqrt{100}} = 0{,}4)$.

On pose $l_\alpha = a$ et on obtient : $P(-\frac{a}{0{,}4} \leq \frac{\overline{X} - m}{0{,}4} \leq \frac{a}{0{,}4}) = 0{,}95$ donc $P(T \leq \frac{a}{0{,}4}) = 0{,}975$.

Par lecture sur la table de la loi normale centrée réduite : on trouve $\frac{a}{0{,}4} = 1{,}96$ donc :

$P(16{,}98 - 0{,}78 \leq m \leq 16{,}98 + 0{,}78) = 0{,}95$.

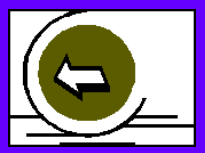
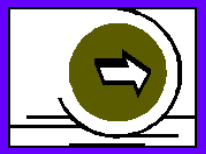
Donc l'intervalle de confiance à 95% sur la population est l'intervalle [16,2 ; 17,76].

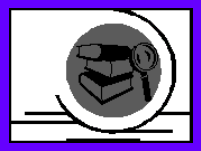
Cas 2 : X suit $\mathcal{N}(m ; \sigma)$ et σ est inconnu

Dans ce cas, on doit utiliser une estimation de σ.

L'estimation naturelle est la déviation standard s, mais l'utilisation de s introduit une source supplémentaire de non-fiabilité. Afin de palier cet inconvénient on doit "élargir" l'intervalle de confiance bilatéral de la moyenne m.

On obtient alors : $\left[m_e - \frac{s}{\sqrt{n}} t_\alpha^* ; m_e + \frac{s}{\sqrt{n}} t_\alpha^* \right]$ où la valeur t_α^* est lue dans la table de la loi de Student à n – 1 degrés de liberté (ddl).

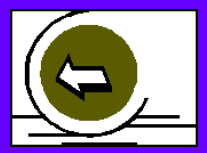
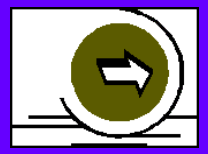


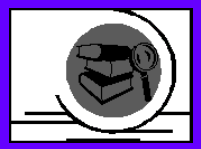

STATISTIQUES INFERENTIELLES

4. Estimation par intervalle de confiance

Exercices interactifs

Exercices interactifs




Cas 3 : X suit une loi quelconque

Il faut : $n \geq 30$ et alors :

- Si σ est connu $\overline{X}$ suit approximativement la loi $\mathcal{N}\left(m ; \frac{\sigma}{\sqrt{n}}\right)$.
- Si σ est inconnu $\overline{X}$ suit approximativement la loi $\mathcal{N}\left(m ; \frac{s}{\sqrt{n}}\right)$.

⊙ Intervalle de confiance d'une proportion

La proportion p d'une propriété donnée est inconnue, l'échantillonnage non exhaustif.

Cas 1 : $n \geq 30$

Comme $n \geq 30$, on sait que F suit (approximativement) la loi $\mathcal{N}\left(p ; \sqrt{\frac{f_e(1-f_e)}{n-1}}\right)$.

Pour un échantillon donné de taille n, l'intervalle de confiance au seuil de risque α est donné par $\left[f_e - t_\alpha \sqrt{\frac{f_e(1-f_e)}{n-1}} ; f_e + t_\alpha \sqrt{\frac{f_e(1-f_e)}{n-1}}\right]$.

Exemple :

On reprend l'exemple des appels n'aboutissant pas pour des raisons techniques.

Donner un intervalle de confiance bilatéral de la proportion p d'appels n'aboutissant pas pour des raisons techniques, au seuil de risque 2 %.

F suit la loi normale $\mathcal{N}\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$ et $\sqrt{\frac{f_e(1-f_e)}{n-1}} = \sqrt{\frac{0{,}03 \times 0{,}97}{1999}} = 0{,}0038$.

On a donc : $[0{,}03 - 2{,}33 \times 0{,}0038 ; 0{,}03 + 2{,}33 \times 0{,}0038] = [0{,}021 ; 0{,}039]$.

STATISTIQUES INFERENTIELLES

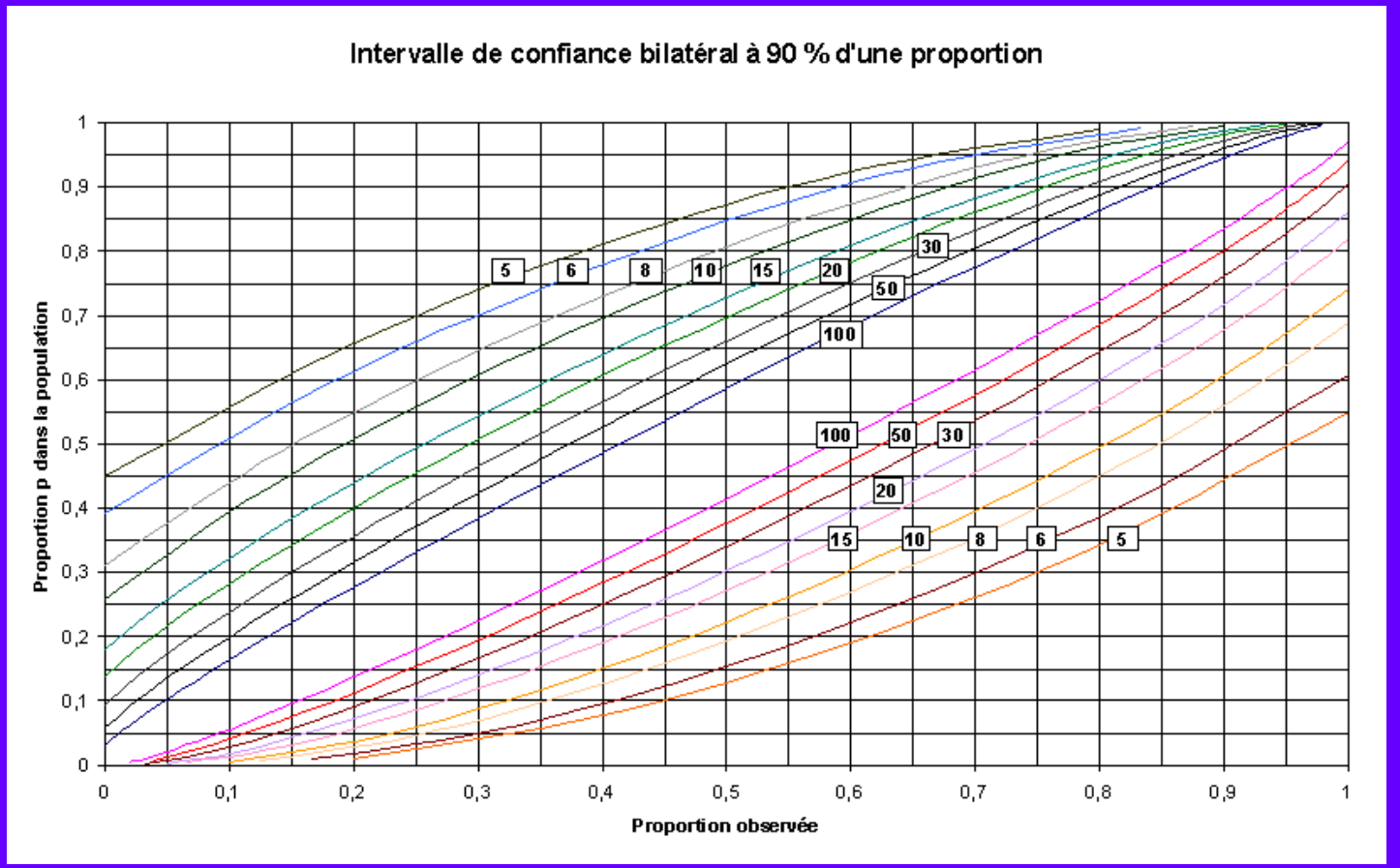
4. Estimation par intervalle de confiance

Cas 2 : n < 30

Dans ce cas là, la seule approche est graphique. On utilise des abaques pour conclure.

Exemple :

Donner l'intervalle de confiance à 90 % de la proportion de l'absentéisme sur l'année dans l'entreprise Alpha (comprenant 500 employés) sachant que sur un échantillon représentatif de 20 employés la proportion est de 15 %.



L'intervalle de confiance de p est donc [____ % ; ____ %].

Solution