Théorie d'estimation

Benchikh Tawfik

Faculté de Médecine, UDL, SBA
1ère année Médecine

21 Février 2024

Plan de cours

1. Estimation par intervalle de confiance
   - Définition d'un intervalle de confiance
   - Construction d'un intervalle de confiance
2. Exemples d'intervalles de confiance
   - Intervalle de confiance pour les paramètres d'une loi normale
   - Intervalle de confiance pour la moyenne d'une loi quelconque
   - Intervalle de confiance pour une proportion
   - Estimation et intervalle de confiance dans le cas d'une population d'effectif fini
3. Exercice

INTRODUCTION

- L'estimation ponctuelle d'un paramètre θ donne une valeur numérique unique à ce paramètre, mais n'apporte aucune information sur la précision des résultats, c'est-à-dire qu'elle ne tient pas compte des erreurs dues aux fluctuations d'échantillonnage, par exemple.
- Pour évaluer la confiance que l'on peut avoir en une estimation, il est nécessaire de lui associer un intervalle qui contient, avec une certaine probabilité, la vraie valeur du paramètre, c'est l'*estimation par intervalle de confiance*.

- L'estimation par intervalle de confiance d'un paramètre θ consiste donc à associer à un échantillon, un intervalle aléatoire I , choisi de telle façon que la probabilité pour qu'il contienne la valeur inconnue du paramètre soit égale à un nombre fixé à l'avance, aussi grand que l'on veut.
- On écrit:

$$\Pr(\theta \in I) = 1 - \alpha$$

  $1 - \alpha$ est la probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre, c'est le *seuil de confiance* ou la *quasi-certitude*.

INTERVALLE DE PROBABILITÉ: RAPPEL

- Soit X une variable aléatoire, f la densité de sa loi de probabilité.
- Étant donnée une probabilité α, on choisit deux nombres α_1 et α_2 ayant pour somme α ($\alpha_1 + \alpha_2 = \alpha$) et on définit deux valeurs x_1 et x_2 de la variable X telles que:

$$\Pr(X < x_1) = \alpha_1 \quad \Pr(X > x_2) = \alpha_2 \,.$$

- L'intervalle $I = [x_1, x_2]$ a une probabilité égale à $(1 - \alpha)$ de contenir une valeur observée de la variable X.
- En négligeant la probabilité α, on résume la distribution de la variable X en ne considérant que les valeurs appartenant à l'intervalle I, on définit un intervalle de probabilité au seuil $(1 - \alpha)$ pour la variable X, la valeur α est le *seuil critique*.

INTERVALLE DE PROBABILITÉ: PRINCIPE

- Pour construire un intervalle de probabilité, deux questions se posent:
  - quel est le seuil de probabilité α susceptible d'être valablement considéré comme négligeable?
  - pour une loi de probabilité et pour un seuil α donnés, il existe une infinité d'intervalles $[x_1, x_2]$ qui dépendent du choix de α_1 et α_2. Comment choisir ces deux valeurs ?
- Les réponses à ces deux questions dépendent des problèmes traités.

Intervalle de probabilité: exemple

- On suppose qu'un dosage sanguin est une variable aléatoire X suivant la loi normale $\mathcal{N}(100, 20)$. On considère comme *normales* les valeurs de X comprises entre deux limites a et b telles que $\Pr(a < X < b) = 0,95$, les autres valeurs étant considérées comme *pathologiques*.
- La donnée du seuil critique $\alpha = 0,05$ sans précision supplémentaire ne permet pas de calculer les limites a et b; une infinité d'intervalles de probabilité répondent à la question. En revanche, la probabilité de mesurer une valeur pathologique est égale à $\alpha = 0,05$, quel que soit l'intervalle.

Intervalle de probabilité: exemple

- On suppose maintenant que les valeurs a et b sont symétriques par rapport à la moyenne $m = 100$.
- En introduisant la variable aléatoire centrée réduite $U = \frac{X-100}{20}$, on sait que:

$$\Pr(-1.96 < U < 1.96) = 0.95$$

- D'où: $a = 100 - 1.96 * 20 = 60.80$ et $b = 100 + 1.96 * 20 = 139.2$ et l'intervalle de probabilité correspondant est $\Pr(60,80 < X < 139,20) = 0,95$.
- Cependant, les faibles valeurs de X ne présentant pas un caractère pathologique, on garde seulement la valeur supérieure $b = 139,20$.
- La probabilité d'observer une valeur pathologique devient égale à 0,025 et un intervalle de probabilité unilatéral au seuil de confiance 0,975 est alors:

$$\Pr(X < 139,20) = 0,975\,.$$

CONSTRUCTION D'UN INTERVALLE DE CONFIANCE

- X est une v.a. dont la densité, $f(x;\theta)$, dépend du paramètre θ et $X = (X_1, \ldots, X_n)$ est un échantillon de taille n de cette variable.
- Soit $T = \varphi(X)$ un estimateur du paramètre θ et $g(t;\theta)$ la loi de probabilité de cet estimateur.
- Étant donnée une probabilité α, on peut, à partir de cette loi et si on suppose le paramètre α connu, construire un intervalle de probabilité pour la variable aléatoire T:

$$\Pr(\theta - h_1 < T < \theta + h_2) = 1 - \alpha \qquad (1.1)$$

Propriétés des intervalles de confiance

- Un intervalle de confiance est un intervalle aléatoire car les bornes de cet intervalle sont des variables aléatoires, fonctions des observations.
- Le seuil α étant donné, il faut définir les nombres α_1 et α_2. Leur choix dépend des problèmes à traiter, des risques encourus à négliger les petites ou les grandes valeurs du paramètre. Si on choisit $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, on construit un intervalle de confiance bilatéral à risques symétriques. On peut construire des intervalles de confiance unilatéraux, soit avec $\alpha_1 = 0$, soit avec $\alpha_2 = 0$.

- **Le seuil** α, les nombres α_1 et α_2 et la taille n de l'échantillon étant fixés, on peut construire un intervalle de confiance associé à chaque échantillon. Cependant, parmi ces intervalles, une proportion égale à $\alpha\%$ ne contiendra pas la valeur exacte du paramètre. Ce seuil α représente donc le risque que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre. La situation la plus favorable correspond à choisir un risque α petit, associé à un intervalle de faible étendue.

PROPRIÉTÉS DES INTERVALLES DE CONFIANCE

- On peut diminuer la valeur du seuil α, et même à la limite, choisir $\alpha = 0$ pour avoir la certitude absolue. Dans ce cas, l'intervalle de confiance s'étend a tout le domaine de definition du paramètre, $]-\infty, +\infty[$ pour l'espérance mathématique ou $[0, +\infty[$ pour l'écart-type, par exemple! Donc: diminuer la valeur de $\alpha \Rightarrow$ augmenter l'étendue de l'intervalle.
- Dans la pratique, on donne à α une valeur acceptable, de l'ordre de 5% puis, quand cela est possible, on augmente la taille de l'échantillon.
- La probabilité $(1-\alpha)$ représente le **niveau de confiance** de l'intervalle; ce niveau de confiance est associé à l'intervalle et non à la valeur inconnue du paramètre.
- Pour définir un intervalle de confiance, il faut connaître un estimateur ponctuel du paramètre ainsi que sa loi de distribution.

ESTIMATION ET INTERVALLE DE CONFIANCE DE LA MOYENNE

- La variable aléatoire X suit une loi normale $\mathcal{N}(m;\sigma)$. Les paramètres à estimer sont la moyenne m et l'écart-type σ.
- L'estimateur sans biais de la moyenne m est la statistique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui suit la loi normale, $\mathcal{N}(m;\frac{\sigma}{\sqrt{n}})$.
- Deux cas sont à distinguer, selon que l'écart-type est connu ou estimé.

Cas 1: l'écart-type σ est connu

- Étant donné un seuil α, on construit, pour la moyenne $\overline{X}$ de l'échantillon, un intervalle de probabilité:

$$\Pr(m - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \overline{X} < m - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

  la valeur $u_{\frac{\alpha}{2}}$ étant lue sur la table de la loi normale réduite: $\Pr(Z \leq u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$; où $Z \rightsquigarrow \mathcal{N}(0;1)$.

- On en déduit l'intervalle de confiance pour la moyenne m:

$$\Pr(\overline{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < m < \overline{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

  où $\overline{x}$ est la moyenne arithmétique de l'échantillon.

- D'où

$$IC_\alpha(m) = [\overline{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\,\overline{x} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}].$$

Intervalle de confiance pour la moyenne, σ connu: exemple

- Après des essais antérieurs, on peut supposer que la résistance à l'éclatement d'un certain type de réservoirs est une variable aléatoire suivant une loi normale de moyenne m inconnue et d'écart-type égal à $4kg/cm^2$. Des essais sur un échantillon de 9 réservoirs donnent une résistance moyenne à l'éclatement égale à $215kg/cm^2$.
- Estimation ponctuelle de la moyenne donnée par l'échantillon: $m \simeq \overline{X}_n = 215kg/cm^2$.
- Loi suivie par la moyenne d'un échantillon de taille $n = 9$ (avec l'hypothèse admise sur la loi suivie par la résistance): $\overline{X}_n$ suit la loi normale $\mathcal{N}(215; 4/3)$.
- Niveau de confiance: $1 - \alpha = 0,95$.

INTERVALLE DE CONFIANCE POUR LA MOYENNE, σ CONNU: EXEMPLE

- Intervalle de confiance:

$$\Pr\left(-1.96 < \frac{\overline{X}-m}{4/3} < 1.96\right) = 0.95 \qquad \bar{x} = 215$$

D'où $\Pr(\bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < m < \bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) =$
$\Pr(215 - 1.96 * 4/3 < m < 215 + 4/3 * 1.96)$
$= \Pr(212.386 < m < 217.613) = 0.95$.
i.e.

$$\begin{array}{rcl} IC_\alpha(m) & = & [\bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\,\bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}] \\ & = & [212,386 - 217,613] \end{array}.$$

- L'intervalle $[212,386 - 217,613]$ a une probabilité égale à $0,95$ de contenir la vraie valeur de la résistance à l'éclatement de ce type de réservoirs.

Intervalle de confiance pour la moyenne, σ connu: exemple

- **Remarque:** Cet exemple montre que si la taille n de l'échantillon augmente, α et σ restant constants, l'étendue de l'intervalle diminue; en revanche, si le seuil a diminue l'étendue de l'intervalle augmente.

CAS 2: L'ÉCART-TYPE σ N'EST PAS CONNU

- L'estimateur sans biais de la variance est la statistique

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \hat{\sigma}^2$$

- La variable aléatoire:

$$T_{(n-1)} = \frac{\overline{X} - m}{S^*/\sqrt{n}} = \frac{\overline{X} - m}{S/\sqrt{n-1}}$$

  suit une loi de Student à $(n-1)$ degrés de liberté

- Le seuil α étant donné, on lit sur la table (loi de Student) le nombre $t_{(n-1;1-\alpha/2)} = t_{(1-\alpha/2)}(n-1)$ tel que:

$$\Pr\left(T(n-1) < t_{(n-1;1-\alpha/2)}\right) = 1 - \frac{\alpha}{2}\,.$$

- D'où l'intervalle de confiance pour la moyenne m:

$$\Pr\left(\bar{x} - \frac{s*}{\sqrt{n}} t_{(n-1;1-\alpha/2)} ; \bar{x} + \frac{s*}{\sqrt{n}} t_{(n-1;1-\alpha/2)}\right) = 1 - \alpha \; .$$

que l'on peut écrire en considérant la statistique S^2:

$$\Pr\left(\bar{x} - \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)} ; \bar{x} + \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)}\right) = 1 - \alpha \, .$$

i.e.

$$IC_\alpha(m) = \left[\bar{x} - \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)} ; \bar{x} + \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)}\right]$$

CAS 2: L'ÉCART-TYPE σ N'EST PAS CONNU: REMARQUE

- Quand n est grand ($n \geq 30$), on peut considérer que la loi de Student est proche de la normale et prendre $t(1-\alpha/2) = U_{\alpha/2}$ dans la table de la loi normale.

Intervalle de confiance pour la moyenne, σ estimé: exemple

- Afin d'étudier le salaire journalier (par 1h), en DA, des ouvriers d'un secteur d'activité, on procède à un tirage non exhaustif, d'un échantillon de taille $n = 16$. On a obtenu les résultats suivants:

$$\begin{array}{cccccccc} 41 & 40 & 45 & 50 & 41 & 41 & 49 & 43 \\ 45 & 52 & 40 & 48 & 50 & 49 & 47 & 46 \end{array}$$

- On suppose que la loi suivie par la variable aléatoire "salaire journalier" est une loi normale de moyenne m et d'écart-type σ inconnus.
- Estimation ponctuelle de la moyenne, la moyenne arithmétique: $\overline{X}_n = 45.4375$.
- Estimation ponctuelle de la variance :
  - $S^2 = 15.2460 = (3.907)^2$ (estimateur biaise).
  - $S^{*2} = 16.2625 = (4.0326)^2$ (estimateur non biaise).

Intervalle de confiance pour la moyenne, σ estimé: exemple

- Intervalle de confiance pour la moyenne, avec seuil de confiance 0.95 (intervalle bilatéral à risques symétriques).
- La variable aléatoire $\frac{\overline{X}-m}{S/\sqrt{n-1}}$ suit une loi de Student à $(n-1)$ degrés de liberté.
- D'où la suite des calculs en tenant compte des résultats donnés par l'échantillon:

$$\Pr(T_{(15)} < 2.131) = 1 - \frac{0.05}{2} = 0.975 \qquad (i.e. T_{15;0.975} = 2.131)$$

$$IC_{0.05}(m) = \left[\overline{x} - \tfrac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)} ; \overline{x} + \tfrac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha/2)}\right] = \left[45.438 - 2.131 \times 3.907/\sqrt{15} ; 45.438 + 2.131 \times 3.907/\sqrt{15})\right] = [43.29; 47.59] \; .$$

- $IC_{0.05}(m) = [43.29; 47.59]$ a une probabilité égale à 0.95 de contenir la vraie valeur du salaire moyen journalier des ouvriers de ce secteur d'activité.

Intervalle de confiance unilatéral pour la moyenne, σ estimée:

- D'une façon analogue, on peut déterminer des intervalles de confiance unilatéraux (à droite ou à gauche):
  - intervalle de confiance unilatéral à gauche:
    - $\Pr\left(\mu < \bar{x} + \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha)}\right) = 1 - \alpha$.
    - $IC_\alpha(\mu) = \left]-\infty; \bar{x} + \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha)}\right]$
  - intervalle de confiance unilatéral à droite:
    - $\Pr\left(\mu > \bar{x} - \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha)}\right) = 1 - \alpha$.
    - $IC_\alpha(\mu) = \left[\bar{x} - \frac{s}{\sqrt{n-1}} t_{(n-1;1-\alpha)} ; +\infty \right[$

CAS 1: LA MOYENNE m EST CONNUE

- Comme précédemment, deux cas sont à distinguer, selon que la moyenne m est connue ou estimée.
- Le meilleur estimateur de la variance est la statistique: $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$.
- La variable aléatoire $\frac{nT}{\sigma^2}$ suit une loi du chi-deux à n degrés de liberté.
- Un intervalle de probabilité pour la variable aléatoire $\chi^2(n)$ est (les bornes de l'intervalle sont lues sur la table de la loi du chi-deux):

$$\Pr\left(\chi^2_{\alpha/2}(n) < \chi^2(n) < \chi^2_{1-\alpha/2}(n)\right) = 1 - \alpha\,.$$

- On en déduit un intervalle de confiance bilatéral pour σ^2 à risques symétriques (t est la valeur de la statistique T donnée par l'échantillon):

$$\Pr\left(\frac{nt}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{nt}{\chi^2_{\alpha/2}(n)}\right) = 1-\alpha\,.$$

e.i.

$$IC_\alpha(\sigma) = \left[\frac{nt}{\chi^2_{1-\alpha/2}(n)}\,;\,\frac{nt}{\chi^2_{\alpha/2}(n)}\right]$$

INTERVALLE DE CONFIANCE POUR LA VARIANCE, m CONNUE: EXEMPLE

- Soit X une variable aléatoire suivant la loi normale $\mathcal{N}(40; \sigma)$. Pour estimer la variance, on prélève un échantillon de taille $n = 25$ et on calcule la valeur de la statistique T (définie précédemment) pour cet échantillon. On obtient $t = 12$.
- Intervalle de confiance bilatéral, à risques symétriques, pour la variance.
- Niveau de confiance: $1 - \alpha = 0.95$.

Intervalle de confiance pour la variance, m connue: exemple

- La statistique $\frac{nT}{\sigma^2}$ suit une loi du chi-deux à $n = 25$ degrés de liberté. On obtient successivement:
  $\Pr\left(\chi^2_{0.025}(25) < \chi^2(25) < \chi^2_{0.975}(25)\right)$
  $\Pr(13.120 < \chi^2(25) < 40.644) = 0.95$.

$$\begin{aligned} IC_{0.05}(\sigma) &= \left[\frac{nt}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{nt}{\chi^2_{\alpha/2}(n)}\right] \\ &= \left[\frac{25\times 12}{40.644}; \frac{25\times 12}{13.120}\right] \\ &= [7,381 < \sigma^2 < 22,866] \end{aligned}$$

- L'intervalle $[7.381; 22.866]$ a une probabilité égale à 0.95 de contenir la vraie valeur de la variance de la loi considérée.
- De même, l'intervalle $[2.716; 4.782]$ a une probabilité égale à 0.95 de contenir la vraie valeur de l'écart-type de la loi considérée.

CAS 2 : LA MOYENNE m N'EST PAS CONNUE

- La statistique $\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est une variable chi-deux à $(n-1)$ degrés de liberté. La démarche est la même que dans le cas 1.

$$\Pr\left(\chi^2_{\alpha/2}(n-1) < \chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) = 1-\alpha\,.$$

  Les bornes de l'intervalle sont lues sur la table du chi-deux.
- D'où l'intervalle de confiance, bilatéral, à risques symétriques :
  - pour la variance (s^2 étant la valeur de la statistique S^2 donnée par l'échantillon):

$$IC_\alpha(\sigma^2) = \left[\frac{ns^2}{\chi^2_{1-\alpha/2}(n-1)};\frac{ns^2}{\chi^2_{\alpha/2}(n-1)}\right]\,.$$

  - pour l'écart-type:

$$IC_\alpha(\sigma) = \left[\sqrt{\frac{ns^2}{\chi^2_{1-\alpha/2}(n-1)}};\sqrt{\frac{ns^2}{\chi^2_{\alpha/2}(n-1)}}\right]\,.$$

Intervalle de confiance pour la variance, m estimée: exemple "salaire journalier"

- La variable $\frac{nS^2}{\sigma^2}$ suit une loi du chi-deux à $16-1=15$ degrés de liberté.
- D'où la suite des calculs:
  $\Pr\left(\chi^2_{0.025}(15) < \chi^2(15) < \chi^2_{0.975}(15)\right) =$
  $\Pr(6.262 < \chi^2(15) < 27.488) = 0.95\,.$

$$
\begin{aligned}
IC_\alpha(\sigma^2) &= \left[\frac{ns^2}{\chi^2_{1-\alpha/2}(n-1)} \,;\, \frac{ns^2}{\chi^2_{\alpha/2}(n-1)}\right]. \\
&= \left[\frac{16\times 15.246}{27.488} ; \frac{16\times 15.246}{6.262}\right] \\
&= [8.874; 39.955]
\end{aligned}
$$

- L'intervalle $[8.874, 38.955]$ a une probabilité égale à 0,95 de contenir la vraie valeur de la variance du salaire moyen horaire des ouvriers de ce secteur d'activité et l'intervalle $[2.98; 6.24]$ a la même propriété pour l'écart-type.

INTERVALLE DE CONFIANCE POUR LA VARIANCE, m ESTIMÉE: EXEMPLE

- D'une façon analogue, on peut déterminer des intervalles de confiance unilatéraux (à droite ou à gauche):
  - A droite:
    - $\Pr\left(\chi^2_\alpha(n-1) < \chi^2(n-1)\right) = 1-\alpha\,.$
    - $\Pr\left(0 < \sigma^2 < \frac{ns^2}{\chi^2_\alpha\,(n-1)}\right) = 1-\alpha\,.$
    - $\Pr\left(0 < \sigma < \sqrt{\frac{ns^2}{\chi^2_\alpha\,(n-1)}}\right) = 1-\alpha\,.$
  - A gauche
    - $\Pr\left(0 < \chi^2(n-1) < \chi^2_{1-\alpha}(n-1)\right) = 1-\alpha\,.$
    - $\Pr\left(\frac{ns^2}{\chi^2_{1-\alpha}\,(n-1)} < \sigma^2\right) = 1-\alpha\,.$
    - $\Pr\left(\sqrt{\frac{ns^2}{\chi^2_{1-\alpha}\,(n-1)}} < \sigma\right) = 1-\alpha\,.$

Intervalle de confiance unilatéral pour la variance: exemple

- Reprenons l'exemple (salaire journalier) déterminons une valeur a telle que (seuil de confiance 0.95):

$$\Pr(0 < \sigma < a) = 1 - \alpha\,.$$

$$\begin{aligned} \Pr(\chi^2_{0.05}(15) < \chi^2(15)) &= \Pr(7.26 < \chi^2(15)) \\ &= \Pr(7.26 < \tfrac{16*15.264}{\sigma^2}) = 0.95 \end{aligned}$$

$$\Pr(\sigma^2 < 33,64) = \Pr(\sigma < 5,80) = 0,95$$

Remarque

- Si la taille n de l'échantillon est supérieure à 30, la variable aléatoire:

$$\sqrt{2\chi^2(n)} - \sqrt{2n-1}$$

suit une loi normale centrée reduite. On utilisera donc la table de la loi normale pour trouver les bornes de l'intervalle.

DÉTERMINATION DE LA TAILLE D'UN ÉCHANTILLON POUR ESTIMER UN PARAMÈTRE AVEC UN NIVEAU DE CONFIANCE DONNÉ

- On suppose que la durée de vie d'ampoules éléctriques suit une loi normale d'écart-type 100 heures. Quelle est la taille minimale de l'échantillon à prélever pour que l'intervalle de confiance, a 95%, de la durée de vie moyenne de ces ampoules ait une longueur inferieure à 20 heures ?
- Les hypothèses faites entrainent que la largeur de l'intervalle de confiance (bilateral, à risques symétriques) pour la moyenne est égale à: $2 \times 1.96 \times \frac{\sigma}{\sqrt{n}}$.
- D'où léquation: $2 \times 1,96 \times \frac{100}{\sqrt{n}} = 20$ et $n = 385$.

INTERVALLE DE CONFIANCE POUR LA MOYENNE: LOI QUELCONQUE

- Quelle que soit la taille de l'échantillon, on prendra pour estimateurs sans biais de la moyenne, la statistique $\overline{X}$, et de la variance, la statistique S^{*2}.
- Si la taille de l'échantillon est grande, en pratique $n > 30$, grace au théorème central limite, on utilise les résultats précédents pour calculer l'intervalle de confiance pour la moyenne.
- Ainsi, dans le cas où l'on souhaite estimer l'espérance lorsque la variance est connue, l'IC est identique à celui déterminé lorsque les v.a.r. $X_1; \ldots ; X_n$ suivent la loi $\mathcal{N}(m; \frac{\sigma^2}{n})$.
- Si la série prélevée est faible, il faut utiliser les lois suivies par les paramètres étudiés.

Intervalle de confiance pour la moyenne, loi quelconque: exemple

- On considère un échantillon de 40 paquets de biscuits provenant d'une production de 2 000 unités. Le poids moyen obtenu pour cet échantillon est égal à 336 g et l'écart-type empirique, c'est-à-dire la quantité s, est égal à 0,86 g.
- Quelle est l'estimation, par intervalle de confiance, du poids moyen de ces paquets de biscuits, pour l'ensemble de la fabrication, avec les seuils de confiance 0,98 ?
- La distribution du poids de ces paquets est inconnue ainsi que la variance de la population. Cependant, pour un grand échantillon $(n = 40 > 30)$, l'intervalle de confiance est de la forme:

$$\bar{x} - t\frac{s}{\sqrt{n-1}} < m < \bar{x} - t\frac{s}{\sqrt{n-1}}$$

avec $\bar{x} = 336$, $s = 0.86$ et $n = 40$.

Intervalle de confiance pour la moyenne, loi quelconque: exemple

- Le fractile h est lu sur la table de Student, degré de liberté $n-1=39$, il dépend du seuil choisi.
- Ainsi, l'intervalle de confiance, qui a 98 chances sur 100 de contenir la vraie valeur du poids moyen de cette fabrication de biscuits, est: $[335.666, 336.334]$.

INTRODUCTION

- Ce problème apparaît dans de nombreuses situations quand on veut estimer par exemple (liste non exhaustive):
  - la proportion de pièces défectueuses dans une fabrication donnée,
  - la proportion d'électeurs qui voteront pour un candidat déterminé,
  - la proportion de ménagères qui achèteront une nouvelle marque de lessive.

- Nous supposons que la proportion p d'individus présentant un certain caractère C au sein d'une population est inconnue.
- Le meilleur estimateur de p est la fréquence empirique F, que l'on peut définir par:

$$F = \overline{X} = \frac{\displaystyle\sum_{i=1}^{n} X_i}{n}$$

où X_i est une v.a. de Bernoulli de paramètre p, définie par:

$$X_i = \begin{cases} 1 & \text{si si l'individu i possède la caractère C} \\ 0 & sinon \end{cases}$$

Intervalle de confiance pour une proportion

- Comme X suit une loi de Bernoulli $\mathcal{B}(p)$, $nF = \sum_{i=1}^{n} X_i$ suit une loi binomiale $\mathcal{B}(n, p)$.
- Selon les valeurs de n et de p, cette loi admet différentes lois limites qui sont utilisées pour déterminer un intervalle de confiance. Dans la pratique, on peut:
  - utiliser les tables statistiques qui donnent les limites inférieures et supérieures d'un intervalle de confiance calculées pour différents seuils et différentes valeurs de n et k, k étant le nombre de succès obtenus au cours de ces n épreuves,
  - utiliser et justifier l'approximation normale.

Intervalle de confiance d'une proportion calculée avec l'approximation normale:

- Si n est faible, on utilisera les tables de la loi binomiale (ou des abaques)..
- Si n est suffisamment grand, de sorte que $np > 5$ et $n(1-p) > 5$, on peut consid´erer (loi des grands nombres) que $\sum_{i=1}^{n} X_i$ suit une loi normale $\mathcal{N}\left(np, \sqrt{np(1-p)}\right)$, d'où F suit une loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, et donc $T = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$ suit une loi $\mathcal{N}(0,1)$.

Intervalle de confiance d'une proportion calculée avec l'approximation normale:

- On obtient alors, en fonction des quantiles $\Pr(-U_{\frac{\alpha}{2}} < T < U_{\frac{\alpha}{2}}) = 1 - \alpha$ l'intervalle de confiance sur p:

$$IC_{1-\alpha}(p) = \left[F - U_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, F + U_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$$

- Cet intervalle recouvre p avec la probabilité $1 - \alpha$, mais il est toutefois inopérant puisque ses bornes dépendent de p.
- En pratique, il existe trois façons d'obtenir l'intervalle de confiance. Nous retiendrons celle qui remplace p par son estimateur F.
- Ainsi, on obtient l'intervalle de confiance sur la proportion p en fonction de la valeur f de F sur notre échantillon:

$$IC_{1-\alpha}(p) = \left[f - U_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}, f + U_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right]$$

Intervalle de confiance d'une proportion calculée avec l'approximation normale: exemple

- Dans un échantillon pris au hasard de 100 automobilistes, on constate que 25 d'entre eux possèdent une voiture de cylindrée supérieure à 1600 cc.
- Quel est l'intervalle de confiance pour la proportion d'automobilistes possédant une voiture de cylindrée supérieure à 1600 cc (intervalle bilatéral, à risques symétriques, seuil de confiance 95%)?

Exemple: solution

- Estimation ponctuelle de p: le nombre d'automobilistes possédant une voiture de cylindrée supérieure à 1600 cc dans un échantillon de taille $n = 100$ suit la loi binomiale $\mathcal{B}(100, p)$.
  Un estimateur non biaisé pour la proportion p est donné par la fréquence $\hat{p} = f_n = \frac{K}{n}$.
  D'où l'estimation ponctuelle de p: $\hat{p} = f_n = \frac{25}{100} = 0.25$.

EXEMPLE: SOLUTION

- Intervalle de confiance pour p, on utilise l'approximation normale:

$$\Pr\left(f_n - u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < f_n + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

$$\Pr\left(f_n - 1.96\sqrt{\frac{p(1-p)}{n}} < p < f_n + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

- on remplace p par son estimation (0,25). Intervalle de confiance:

$$\Pr(0.25-0.085 < p < 0.25+0.085) = \Pr(0.165 < p < 0.335) = 0.9$$

CAS D'UNE POPULATION D'EFFECTIF FINI

- La plupart des cas étudiés (estimation d'une proportion, d'une moyenne ou d'une variance) supposait implicitement que la population était d'effectif infini.
- Certains résultats ne sont plus valables si l'on suppose que la population est d'effectif fini.

- Soit une variable aléatoire X distribuée sur une population d'effectif fini N.
- Cette variable ne prend qu'un nombre fini de valeurs, elle est donc discrète.
- On considère un échantillon de taille n.

ESTIMATION DE LA MOYENNE m ET DE L'ÉCART-TYPE σ

- **Tirage avec remise**
  - Estimateur sans biais de m: la statistique $\overline{X}$

$$\mathbb{E}(\overline{X}) = m \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

  - Estimateur sans biais de σ^2: la statistique $S^{*2} = \frac{nS^2}{n-1}$
- **Tirage sans remise**
  - Estimateur sans biais de m: la statistique $\overline{X}$

$$\mathbb{E}(\overline{X}) = m \qquad Var(\overline{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

  - Estimateur sans biais de σ^2: la statistique $\frac{N-1}{N}\frac{nS^2}{n-1}$

- **Tirage sans remise**
  - La fréquence empirique f est un bon estimateur de p:

$$\mathbb{E}(f) = p \qquad Var(f) = \frac{p(1-p)}{n}$$

- **Tirage sans remise**
  - La loi concernée est la loi hypergéométrique. La fréquence empirique f est un estimateur sans biais de p:

$$\mathbb{E}(f) = p \qquad Var(f) = \frac{N-n}{N-1}\frac{p(1-p)}{n}$$

  - Pour déterminer les intervalles de confiance, on utilise des tables spéciales ou un ordinateur.

EXERCICE-1

Dans le cadre d'une étude sur la santé au travail, on a interrogé au hasard 500 salariés de différents secteurs et de différentes régions de l'Algerie. 145 d'entre eux déclarent avoir déjà subi un harcèlement moral au travail.

1. Identifier la population, la variable, son type et son/ses paramètre(s).
2. Donner une estimation ponctuelle de la proportion de salariés ayant déjà subi un harcèlement moral au travail.
3. Donner une estimation de cette proportion par un intervalle de confiance à 90%.
4. Si avec les mêmes données on calculait un intervalle de confiance à 95%, serait-il plus grand ou plus petit que celui trouvé à la question précédente ? (justifier sans calcul.)

EXERCICE-2

En vue de réaliser un programme de rééducation, des chercheurs ont soumis un questionnaire de neuropsychologie cognitive à 150 enfants dyslexiques tirés au sort. Le questionnaire comporte 20 questions et les chercheurs ont recueilli pour chaque enfant dyslexique le nombre x_i de bonnes réponses. Les résultats ainsi récoltés sont tels que: $\sum_i x_i = 1502$; $\sum_i x_i^2 = 19486.$

1. La population statistique étudiée est:
   (A) les chercheurs (B) le programme de rééducation
   (C) les enfants dyslexiques
   (D) le questionnaire de neuropsychologie (E) le nombre de bonnes réponses.

2. La variable statistique X étudiée est:
   (A) les chercheurs (B) le programme de rééducation
   (C) les enfants dyslexiques
   (D) le questionnaire de neuropsychologie (E) le nombre de bonnes réponses.

3. Les valeurs possibles de X sont: (A) une seule valeur $\{20\}$ (B) $\{0, 1, \ldots ; 20\}$ (C) $\{0, 1, \ldots, 50\}$ (D) {juste,fausse} (E) A, B, C et D sont faux.
4. La taille de l'échantillon est:
   (A) 20 (B) 150 (C) 1502 (D) 3000 (E) 19486.
5. L'estimation ponctuelle sans biais de la moyenne (μ) de la variable dans la population est:
   (A) 1502 (B) 10.8 (C) 10.01 (D) 75.01 (E) 79.05.
6. L'estimation ponctuelle sans biais de la variance de la variable dans la population est:
   (A) 28.3 (B) 29.17 (C) 29.7 (D) 29.9 (E) 30.59.

7. La moyenne de la variable dans la population a 95% de chance de situer dans l'intervalle: (A) $[9.13 - 10.89]$ (B) $[9.27 - 10.75]$ (C) $[10.01 - 12.29]$ (D) $[3.77 - 16.25]$ (E) $[8.87 - 11.15]$.
8. La probabilité que la moyenne de la variable dans la population ce situe dans l'intervalle $[9.27 - 10.75]$ est: (A) 0.05 (B) 0.1 (C) 0.9 (D) 0.95 (E) 0.99.
9. La marge d'erreur dans l'estimation de la moyenne de la variable dans la population au niveau 99% est: (A) 0.99 (B) 0.95 (C) 0.88 (D) 6.24 (E) 1.16.