Calcul de la taille d'un échantillon

Pr. A. ILIADIS
Laboratoire de Pharmacocinétique
U.F.R. de Pharmacie, Université de la Méditerranée
iliadis@pharmacie.univ-mrs.fr http://pharmapk.pharmacie.univ-mrs.fr/

Résumé du cours dispensé dans le cadre du Diplôme d'Université intitulé Expérimentation Animale. Principale référence: Saporta, G. (1990). Probabilités, Analyse des Données et Statistique. Paris, Technip.

1 Introduction

En calcul des probabilités, deux questions se posent:

- $\Pr[x_1 \leq X < x_2] = ?$ dans le sens de la *prévision* et
- $\Pr[x_1 \leq X < ?] = 1 - \alpha$ dans le sens de l'*estimation*,

x_1 et x_2 étant deux niveaux ($x_1 < x_2$) de la variable aléatoire appelée X et α une probabilité ($0 < \alpha < 1$). On peut répondre à ces deux questions par le même outil mathématique qui prend deux formes différentes selon que la variable aléatoire est discrète ou continue:

- Variable aléatoire discrète (p.ex. nombre d'animaux restant malades après traitement):

$$\Pr[x_1 \leq X < x_2] = \sum_{x_1 \leq x_i < x_2} f(x_i) = F(x_2) - F(x_1)$$

  $f(x)$ est la *fonction de distribution des probabilités* et $F(x)$ est la *fonction cumulative des probabilités* (CDF).

- Variable aléatoire continue (p.ex. taille tumorale après traitement):

$$\Pr[x_1 \leq X < x_2] = \int_{x_1}^{x_2} f(x)\,dx = F(x_2) - F(x_1)$$

  $f(x)$ est la *fonction de densité des probabilités* (PDF) et $F(x)$ est la *fonction cumulative des probabilités* (CDF). Pour une variable aléatoire continue, ces deux fonctions sont liées par la relation:

$$\frac{dF(x)}{dx} = f(x)$$

Parmi toutes les fonctions mathématiques possibles, celles qui peuvent être utilisées comme modèles de distribution et de densité des probabilités doivent être non-négatives et vérifier les conditions:

$$\sum_i f(x_i) = 1 \qquad \text{et} \qquad \int_x f(x)\,dx = 1$$

Ces modèles sont caractérisés par:

- l'*espérance* $E[X]$ qui exprime la tendance centrale,
- la *variance* $V[X]$ qui exprime la dispersion,
- l'*asymétrie* $\Gamma_1[X]$ et l'*aplatissement* $\Gamma_2[X]$ qui expriment la forme du modèle.

2 Statistique exploratoire - échantillon

Comment obtenir ces outils-modèles mathématiques $f(x)$? Sans doute, après avoir analysé la population (ensemble des résultats possibles issus d'une expérience aléatoire) de la variable aléatoire X. Deux cas de figure se présentent:

- Population dénombrable: analyser l'expérience aléatoire et appliquer l'*approche théorique*.
- Population de taille infinie: utiliser l'*approche expérimentale* qui consiste à:
  - Echantillonner (procédure aléatoire).
  - Analyser les propriétés de l'échantillon et les résumer par des:
    - graphiques, les histogrammes,
    - indices numériques qui caractérisent:
      - la tendance centrale, t.q. $\overline{x}$, la *moyenne arithmétique*,
      - la dispersion, t.q. s^2, la *variance de l'échantillon*,
      - la forme, t.q. γ_1 et γ_2, *coefficients d'asymétrie et d'aplatissement*, respectivement.
  - Extrapoler ces propriétés pour définir une fonction mathématique $f(x)$ qui les vérifie.

On se place dans le cadre d'une population de taille infinie, sinon le problème se réduit à un problème d'analyse combinatoire ou de recensement.

3 Statistique inférentielle - population

Parmi les modèles disponibles $f(x)$, nous devons choisir celui qui est compatible avec les caractéristiques de l'échantillon. Ce choix se fait en deux étapes:

- Choix structurel: Fixer la forme mathématique en fonction du contexte expérimental, de la forme de l'histogramme et surtout en fonction des indices de forme γ_1 et γ_2. Par exemple, si $\gamma_1 \approx 0$ et $\gamma_2 \approx 3$ on choisit la PDF normale:

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right] \qquad \text{ou} \qquad X \sim N\left(m, \sigma^2\right)$$

  car sous l'hypothèse de normalité, nous avons $\Gamma_1[X] = 0$ et $\Gamma_2[X] = 3$. m et σ^2 sont les deux paramètres de la PDF dont les valeurs numériques doivent être calculées à partir des observations sur l'échantillon. Ces paramètres expriment les caractéristiques de la PDF, car $E[X] = m$ et $V[X] = \sigma^2$.

- Choix paramétrique: Assigner des valeurs numériques aux paramètres de la structure fixée. Par exemple, pour la PDF normale, trouver les valeurs de ses paramètres m et σ^2. Dans ce cas, la théorie de l'estimation nous permet d'affecter $\overline{x}$ à m (tous les deux exprimant la tendance centrale) et s^2 à σ^2 (tous les deux exprimant la dispersion).

A la fin, on dispose d'un modèle complet $f_N(x)$ qui peut être utilisé comme il a été indiqué en introduction.

4 Intervalles de confiance

Soit α la probabilité pour que X se réalise à l'extérieur de l'intervalle $[L_g, L_d]$:

$$\Pr[X < L_g \text{ ou } X \geq L_d] = \alpha$$

ou

$$\Pr[L_g \leq X < L_d] = 1 - \alpha$$

L_g et L_d sont appelés les limites de confiance et $[L_g, L_d]$, l'intervalle de confiance associé au niveau α. En général, on partitionne α en $\alpha/2$ à gauche de L_g et en $\alpha/2$ à droite de L_d. La relation précédente peut être alors écrite:

$$\Pr[L_g \leq X < L_d] = F(L_d) - F(L_g) = 1 - \alpha$$

et

$$F(L_g) = \alpha/2 \qquad \text{et} \qquad F(L_d) = 1 - \alpha/2$$

et par inversion de la fonction CDF:

$$L_g = F^{-1}(\alpha/2) \qquad \text{et} \qquad L_d = F^{-1}(1 - \alpha/2) \tag{1}$$

5 Echantillonnage

Cette théorie revient sur la constitution d'un échantillon représentatif de la population à partir de laquelle il a été tiré. En effet, on pourrait imaginer plusieurs échantillons $E_1, E_2, ...$ issus de la même population. L'analyse de chacun donnera des résultats différents: $\overline{x}_1, \overline{x}_2, ...$ puis $s_1^2, s_2^2, ...$ etc. C'est ainsi que $\overline{x}$, s^2 et les autres indices numériques deviennent des véritables variables aléatoires. Les questions que nous allons nous poser sont:

- Quelles sont les caractéristiques distributionnelles des indices numériques de l'échantillon?
- Si $E_1, E_2, ...$ peuvent être envisagées de taille différente, $n_1, n_2, ...$, quelle est la taille critique qui garantit une certaine performance?

La seconde question constitue l'objet de ce cours. La performance sera exprimée par la largeur de l'intervalle de confiance $L_d - L_g$. Pour une taille donnée d'un échantillon, l'étude distributionnelle de ces indices permet le calcul des intervalles de confiance soit pour des variables aléatoires continues, soit discrètes.

5.1 Variables aléatoires continues

Dans le cas d'un modèle général $f(x)$, un certain nombre de résultats existent, mais nous ne ferons pas mention dans ce cours. Ici, nous allons présenter les résultats obtenus sous *l'hypothèse d'un modèle PDF normal*. En fait, selon le théorème central-limite des statistiques, le cas d'un modèle normal est très attendu quand l'expérience aléatoire est composé de plusieurs processus aléatoires mélangés, c'est le cas en biologie. Nous allons donc *fixer la structure normale* et établir des intervalles de confiance sur ses *paramètres* m et σ^2.

1. Dispersion exprimée par σ^2: Ayant à notre disposition $\overline{x}$, s^2 etc., l'objectif est d'obtenir un intervalle de confiance pour σ^2. On établit que:

$$\left[\frac{ns^2}{\sigma^2}\right] \sim \chi^2(n-1) \tag{2}$$

   En d'autres termes, l'intervalle de confiance pour σ^2 peut être calculé à l'aide de χ^2 à $n-1$ degrés de liberté et en utilisant s^2 calculé à partir de l'échantillon de taille n. Soit $F_{\chi^2}(n-1)$ la CDF de $\chi^2(n-1)$.

2. Tendance centrale exprimée par m: Ayant à notre disposition $\overline{x}$, s^2 etc., l'objectif est d'obtenir un intervalle de confiance pour m. On établit que:

$$\left[\frac{\overline{x}-m}{\sigma}\sqrt{n}\right] \sim N(0,1)$$

   On se rend compte dans cette relation que la réponse dépend de σ. Si σ est connu, il suffit tout simplement d'utiliser sa valeur; le cas le plus intéressant est quand σ est inconnu. Il doit alors être remplacé par son estimation s, mais la relation précédente doit être modifiée en conséquence. Ceci peut se faire à l'aide de la distribution de Student t. On obtient:

$$\left[\frac{\overline{x}-m}{s}\sqrt{n-1}\right] \sim t(n-1) \tag{3}$$

   C'est ainsi que l'intervalle de confiance pour m peut être calculé à l'aide de la distribution t de Student à $n-1$ degrés de liberté et en utilisant $\overline{x}$ et s^2 calculés à partir de l'échantillon de taille n. Soit $F_t(n-1)$ la CDF de $t(n-1)$.

5.2 Variables aléatoires discrètes

Soit E la variable aléatoire qui exprime le nombre d'événements (associés a priori à une probabilité p) réalisés au cours de n répétitions. C'est le contexte d'une distribution binomiale, $E \sim B(p,n)$. Soit maintenant la fréquence $w = E/n$ d'apparition de E au cours de n répétitions. w est maintenant une variable aléatoire continue. Quand $n \gg 1$, w devient pratiquement une variable aléatoire continue avec une distribution $w \sim N[p, p(1-p)/n]$. Le calcul de l'intervalle de confiance de p pose un problème du fait que p influence à la fois l'espérance et la variance du modèle: il faut *stabiliser* la variance. Soit la transformation $\arcsin\sqrt{w}$. La PDF de $\arcsin\sqrt{w}$ est maintenant:

$$\arcsin\sqrt{w} \sim N\left(\arcsin\sqrt{p}, \frac{1}{4n}\right)$$

A l'aide de cette transformation, p n'influence maintenant que l'espérance.

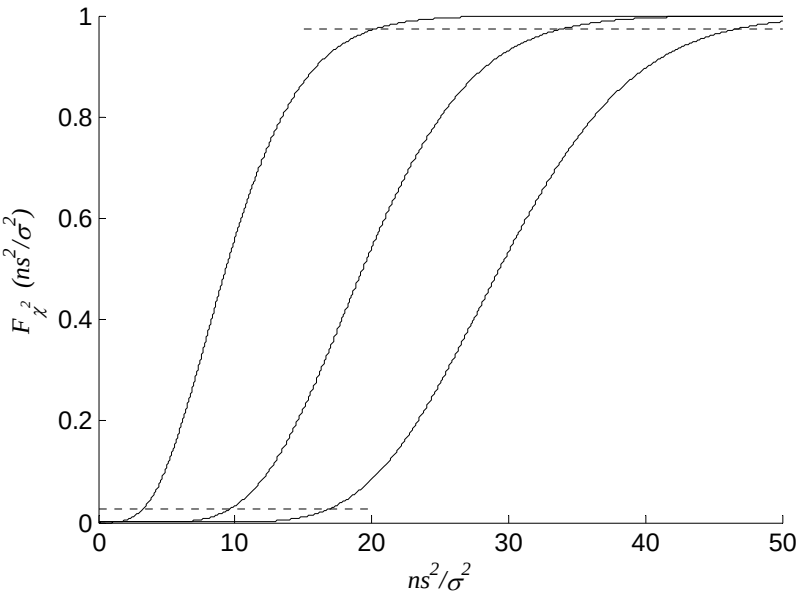


Figure 1: CDF de χ^2 pour $n = 10, 20, 30$.

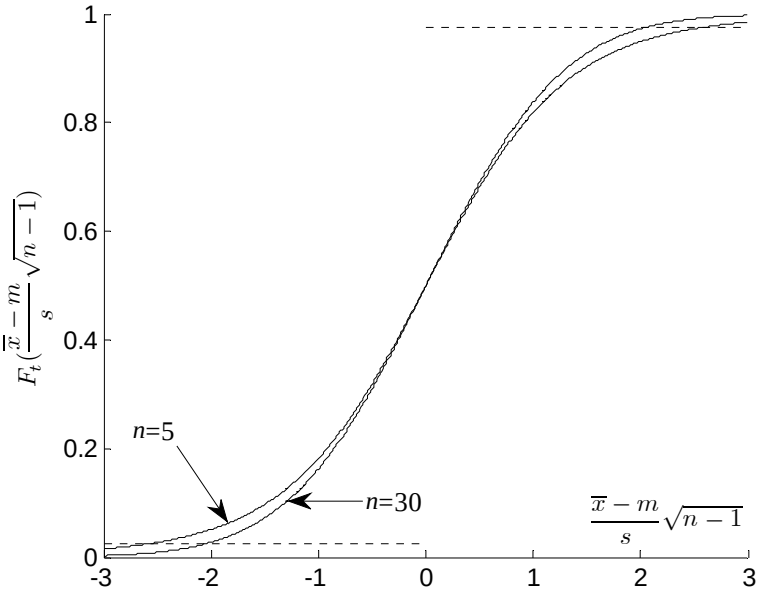


Figure 2: CDF de Student pour $n = 5, 30$.

Ayant à notre disposition w, la fréquence observé sur l'échantillon, l'objectif est d'obtenir un intervalle de confiance sur p. D'après la relation précédente, on établit que:

$$2\sqrt{n}\left(\arcsin\sqrt{w} - \arcsin\sqrt{p}\right) \sim N(0,1) \tag{4}$$

L'intervalle de confiance sur p peut être calculé en utilisant w et n, et faisant valoir la propriété 4. Soit F_U la CDF de $N(0,1)$.

6 Applications

6.1 Intervalles de confiance sur σ

D'après la distribution 2, pour un α fixé et selon la définition 1, les limites de confiance L_g et L_d de la variable aléatoire s^2/σ^2 sont:

$$L_d = \frac{1}{n} F^{-1}_{\chi^2(n-1)}(1-\alpha/2) \qquad \text{et} \qquad L_g = \frac{1}{n} F^{-1}_{\chi^2(n-1)}(\alpha/2) \tag{5}$$

L'influence de la taille de l'échantillon sur l'intervalle de confiance de s^2/σ^2 peut être étudiée en traçant $L_d - L_g$ en fonction de n. La Figure 4 présente cette fonction pour $\alpha = 0.10, 0.05, 0.01$.

Exercise 1 *Intervalles sur la dispersion*

Analyse Pour $\alpha = 0.05$, $s^2 = 4$ et $n = 30$, calculer les limites de confiance de σ. D'après les tables de distribution χ^2 et compte tenu de 5, les limites de confiance de s^2/σ^2 sont:

$$\begin{aligned} L_d &= \frac{1}{30} F^{-1}_{\chi^2(29)}(0.975) = \frac{45.722}{30} = 1.5241 \\ L_g &= \frac{1}{30} F^{-1}_{\chi^2(29)}(0.025) = \frac{16.047}{30} = 0.5349 \end{aligned}$$

ce qui est bien vérifié sur la Figure 4. Ceci se traduit par:

$$\Pr\left[0.5349 \leq \frac{s^2}{\sigma^2} < 1.5241\right] = \Pr\left[\frac{4}{1.5241} < \sigma^2 \leq \frac{4}{0.5349}\right] = \Pr\left[1.6200 < \sigma \leq 2.7346\right] = 0.95$$

Synthèse On demande la taille nécessaire de l'échantillon pour "contrôler le rapport s/σ dans l'intervalle $0.75 \leq s/\sigma \leq 1.12$". Ceci amène à concevoir un intervalle de confiance:

$$L_d - L_g = \left(\frac{s}{\sigma_g}\right)^2 - \left(\frac{s}{\sigma_d}\right)^2 = (1.12)^2 - (0.75)^2 \approx 0.7$$

qui, pour des risques $a = 0.10, 0.05, 0.01$, conduit respectivement à $n = 40, 60, 100$.

6.2 Intervalles de confiance sur m

D'après la distribution 3, pour un α fixé et selon la définition 1, les limites de confiance L_g et L_d de la variable aléatoire $\frac{\overline{x}-m}{s}$ sont:

$$L_d = \frac{1}{\sqrt{n-1}} F^{-1}_{t(n-1)}(1-\alpha/2) \qquad \text{et} \qquad L_g = -L_d \tag{6}$$

Il est ainsi évident que la largeur de l'intervalle dépend de n. L'influence de la taille de l'échantillon sur l'intervalle de confiance peut être étudiée en traçant $L_d - L_g$ en fonction de n. La Figure 5 présente cette fonction pour $\alpha = 0.10, 0.05, 0.01$.

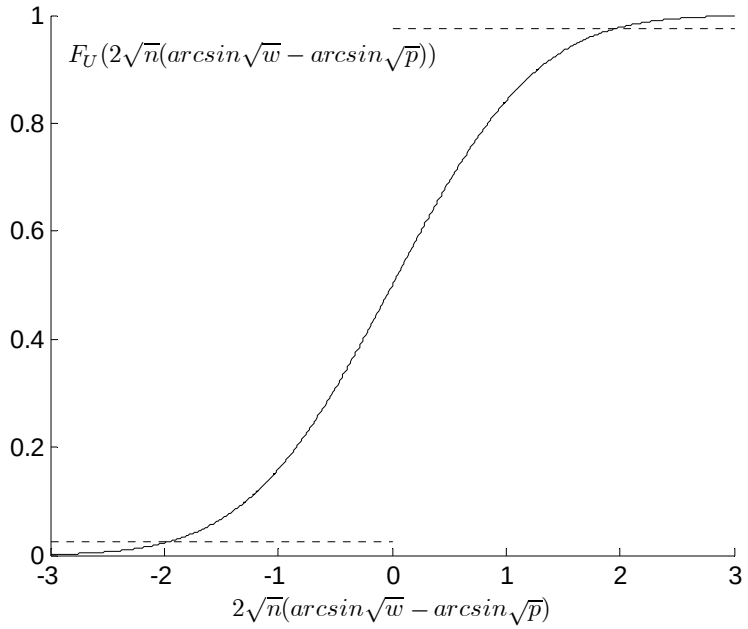
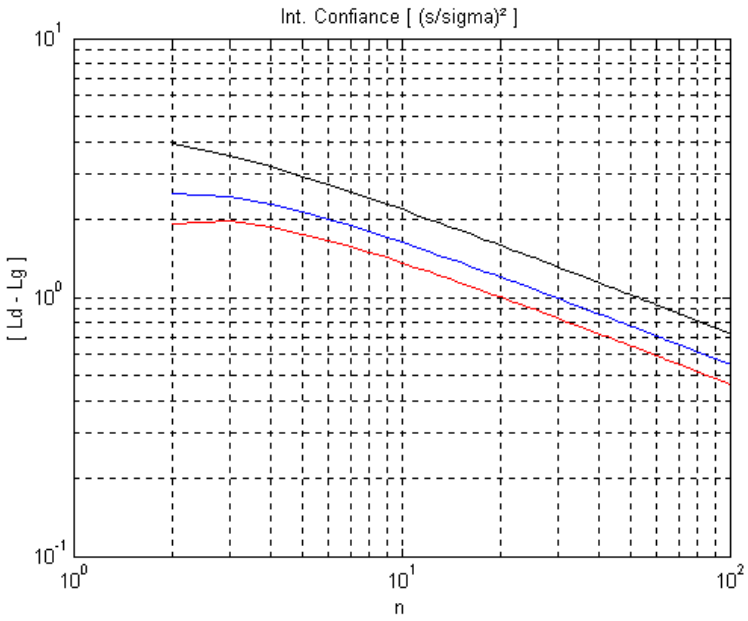


Figure 3: CDF de la distribution normale centrée, réduite.



1

Figure 4: Intervalles de confiance de s^2/σ^2. Pour $\alpha = 0.10, 0.05, 0.01$, $L_d - L_g$ en fonction de n.

Exercise 2 *Intervalles sur la tendance centrale*

Analyse Pour $\alpha = 0.05$, $\overline{x} = 25$, $s^2 = 4$ et $n = 30$, calculer les limites de confiance de m. D'après les tables de distribution t et compte tenu de 6, les limites de confiance de $\frac{\overline{x}-m}{s}$ sont:

$$L_d = \frac{1}{\sqrt{29}} F^{-1}_{t(29)}\left(0.975\right) = \frac{2.0452}{5.3852} = 0.3798 \qquad \text{et} \qquad L_g = -0.3798$$

ce qui est bien vérifié sur la Figure 5. Ceci se traduit par:

$$\Pr\left[-0.3798 \leq \frac{\overline{x}-m}{s} < 0.3798\right] = \Pr\left[24.2404 < m \leq 25.7596\right] = 0.95$$

Synthèse On demande la taille nécessaire de l'échantillon pour "assurer un intervalle de confiance sur m du même ordre de grandeur que s". En d'autres termes:

$$L_d - L_g = \frac{\overline{x}-m_d}{s} - \frac{\overline{x}-m_g}{s} = \frac{m_g - m_d}{s} = 1$$

Pour des risques $a = 0.10, 0.05, 0.01$, ceci conduit respectivement à $n = 12, 20, 30$.

6.3 Intervalles de confiance sur p

D'après la distribution 4, pour un α fixé et selon la définition 1, les limites de confiance L_g et L_d pour la variable aléatoire $\left(\arcsin\sqrt{w} - \arcsin\sqrt{p}\right)$ sont:

$$L_d = \frac{1}{2\sqrt{n}} F^{-1}_{U}\left(1-\alpha/2\right) \qquad \text{et} \qquad L_g = -L_d \tag{7}$$

Il est ainsi évident que la largeur de l'intervalle dépend de n. L'influence de la taille de l'échantillon sur l'intervalle de confiance de $\left(\arcsin\sqrt{w} - \arcsin\sqrt{p}\right)$ peut être étudiée en traçant $L_d - L_g$ en fonction de n. La Figure 6 présente cette fonction pour $\alpha = 0.10, 0.05, 0.01$.

Exercise 3 *Intervalles sur la fréquence*

Analyse Pour $\alpha = 0.05$, $w = 0.9, 0.5, 0.1$ et $n = 30$, calculer les limites de confiance de p. D'après les tables de distribution U et compte tenu de 7, les limites de confiance de $\left(\arcsin\sqrt{w} - \arcsin\sqrt{p}\right)$ sont:

$$L_d = \frac{1}{2\sqrt{30}} F^{-1}_{U}\left(0.975\right) = \frac{1.96}{10.9545} = 0.1789 \qquad \text{et} \qquad L_g = -0.1789$$

ce qui est bien vérifié sur la Figure 6. Ceci se traduit par:

$$\begin{aligned} & \Pr\left[-0.1789 \leq \arcsin\sqrt{w} - \arcsin\sqrt{p} < 0.1789\right] \\ = \; & \Pr\left[\sin^2\left(\arcsin\sqrt{w} - 0.1789\right) < p \leq \sin^2\left(\arcsin\sqrt{w} + 0.1789\right)\right] \end{aligned}$$

et

A $\Pr\left[0.7696 \leq p < 0.9797\right] = 0.95$ pour $w = 0.9$,

B $\Pr\left[0.3249 \leq p < 0.6751\right] = 0.95$ pour $w = 0.5$, et

C $\Pr\left[0.0203 \leq p < 0.2304\right] = 0.95$ pour $w = 0.1$.

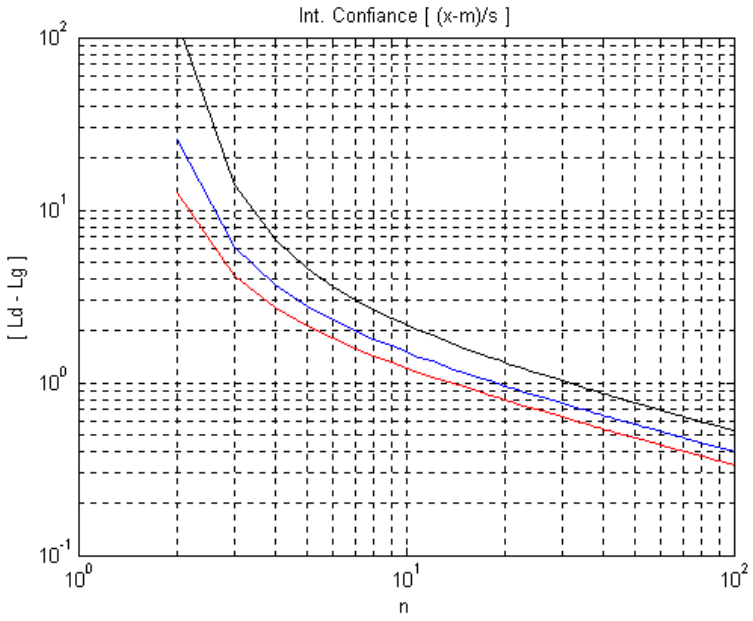


Figure 5: Intervalles de confiance de $\frac{\overline{x}-m}{s}$. Pour $\alpha = 0.10, 0.05, 0.01$, $L_d - L_g$ en fonction de n.

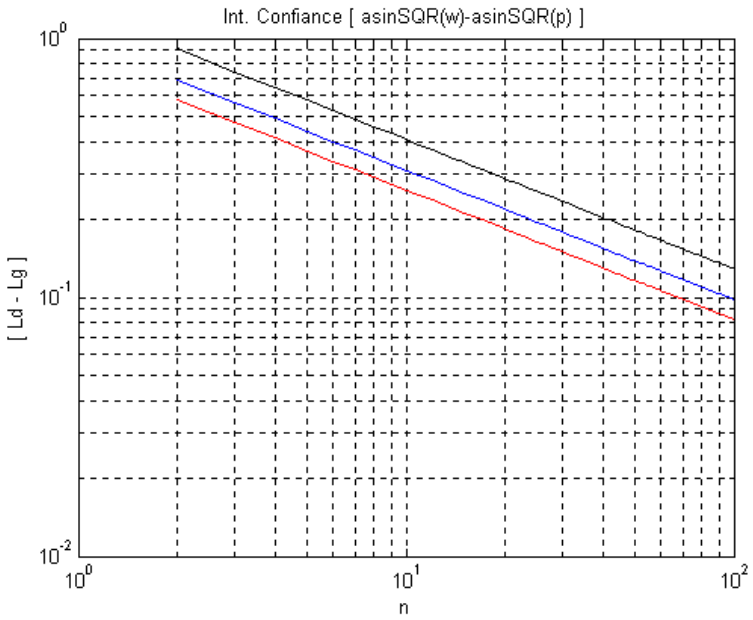


Figure 6: Intervalles de confiance de $\left(\arcsin\sqrt{w} - \arcsin\sqrt{p}\right)$. Pour $\alpha = 0.10, 0.05, 0.01$, $L_d - L_g$ en fonction de n.

Synthèse On demande la taille nécessaire de l'échantillon pour "assurer une probabilité a priori dans un intervalle de 0.1 avec un risque $\alpha = 0.05$". On aura:

$$\begin{aligned} L_d - L_g &= \left(\arcsin\sqrt{w} - \arcsin\sqrt{p_g}\right) - \left(\arcsin\sqrt{w} - \arcsin\sqrt{p_d}\right) \\ &= \arcsin\sqrt{p_d} - \arcsin\sqrt{p_g} \end{aligned}$$

La réponse dépend du niveau de probabilité a priori p. Ainsi si:

A $p \approx 0.9$, $L_d - L_g = \arcsin\sqrt{0.95} - \arcsin\sqrt{0.85} \approx 0.24$ et $n = 12$,

B $p \approx 0.5$, $L_d - L_g = \arcsin\sqrt{0.55} - \arcsin\sqrt{0.45} \approx 0.11$ et $n = 75$, et

C $p \approx 0.1$, $L_d - L_g = \arcsin\sqrt{0.15} - \arcsin\sqrt{0.05} \approx 0.10$ et $n = 100$.

7 Conclusion

Nous avons présenté que la partie "analyse" du problème posé. La partie "synthèse" pose toujours la question comment choisir n. D'une manière pratique, il faut fixer la largeur d'un intervalle sur l'axe des ordonnées des Figures 4, 5 et 6, puis définir approximativement sur l'axe des abscisses une plage de valeurs possibles pour n. Ayant fixé n, une première expérience peut être planifiée suite à laquelle un premier échantillon être obtenu. Les observations récoltées sur cet échantillon peuvent être utilisées:

- pour vérifier les intervalles de confiance pressentis, mais aussi,
- pour affiner davantage le choix de n au cours d'une nouvelle expérience.