INTERVALLES DE CONFIANCE POUR UNE PROPORTION: POURQUOI L'INTERVALLE "STANDARD" DE WALD EST-IL LE PLUS COURAMMENT ENSEIGNÉ?

[1] *Département STID, IUT de Vannes, 8 rue Montaigne, 56000 Vannes,*
Laboratoire de Mathématiques de Bretagne Atlantique, UMR 6205, UBS
jean-francois.petiot@univ-ubs.fr

[2] *Département STID, IUT de Pau, avenue de l'Université, 64000 Pau*
Laborataire de Mathématiques et de leurs applications UMR 5142, UPPA
jean-christophe.turlot@univ-pau.

Résumé.

La rénovation des programmes dans l'enseignement secondaire et dans les IUT est achevée. Dans les lycées, il a été introduit l'intervalle de confiance pour une proportion. Pour le programme des DUT STID on s'est demandé s'il fallait adapter notre enseignement. Les intervalles de confiance couramment enseignés sont l'intervalle « exact » de Clopper-Pearson et l'intervalle de Wald. Or il se trouve que le premier est bien trop conservatif et que le second a une probabilité de recouvrement généralement très insuffisante même lorsque le paramètre p se trouve dans l'intervalle [0.2, 0.8]. De plus, les oscillations de la probabilité exacte de recouvrement en fonction de la taille de l'échantillon comme en fonction de p ont une forte amplitude. Nous confrontons quelques-uns des intervalles parmi les plus intéressants au sens de leur propriété de recouvrement : des intervalles dits « exacts » au sens où ils ne font pas appel à une distribution approchée ; des intervalles fondés sur une approximation normale dont on peut mesurer le biais et l'amplitude des oscillations au moyen d'un développement d'Edgeworth ; enfin une approche de type bootstrap. Nous étudions la précision de ces intervalles soit par l'espérance de leur longueur, soit par la probabilité de recouvrement de valeurs éloignées de la vraie valeur du paramètre. La plupart de ces intervalles sont relativement complexes à comprendre pour les étudiants ; cependant il en est un d'expression similaire à l'intervalle de Wald et simple d'interprétation, c'est l'intervalle Mid-P ou intervalle d'Agresti-Coull présentant de bonnes qualités de recouvrement.

Mots-clés. Intervalle de confiance, proportion, enseignement, loi binomiale, niveau réel de confiance, biais, oscillations, développement d'Edgeworth, précision

Abstract

The updating of the syllabus in the secondary schools and in the IUT is done. The estimation of a proportion has been introduced in the formers. In the IUT we wondered whether we had to change or not our teaching. The most often teached confidence intervals are the "exact" one of Clopper-Pearson and the Wald one. Well, the first one is too much conservative and the second one has a coverage probability which is generally poor, even when the parameter p is within [0.2 0.8]. Moreover, the variations of the exact coverage probability according to the sample size, as to p, have a large range. We compare some intervals, among the more interesting, using their coverage property: some "exact" intervals which do not use an approached distribution; some approached intervals based on a Gaussian approximation whose the biais and the amplitude of the fluctuations may be measured with an Edgeworth expansion; finally a bootstrap approach. We study the precision of these intervals either through the expectancy of the length or through the coverage probability of values away of the

actual value of p. Most of these intervals are of a quite difficult understanding for the students; however there is one whose the expression looks like the Wald interval and with an easy interpretation, it is the Mid-P or Agresti-Coul interval with good coverage qualities.

Key-words : Confidence interval, proportion, teaching, binomial distribution, actual confidence level, bias, fluctuations, Edgeworth expansion, precision

1 Les deux intervalles les plus enseignés et utilisés : Wald et Clopper-Pearson

Les nouveaux programmes des filières générales dans l'enseignement secondaire introduisent les intervalles de confiance pour une proportion p . Le programme rénové du DUT STatistique et Informatique Décisionnelle (STID) suit cette évolution. Une question naturelle était : ne pourrions-nous pas davantage insister sur la notion d'intervalle de confiance et sur ses pratiques ? Les deux types d'intervalles de confiance les plus couramment enseignés ou utilisés dans la pratique sont l'intervalle de Wald et l'intervalle « exact » de Clopper et Pearson. Or, il se trouve que si l'on calcule la probabilité réelle de recouvrement du paramètre p en fonction de la taille de l'échantillon, ces deux méthodes font apparaitre un écart important entre celle-ci et le taux de recouvrement nominal ($\eta = 0.95$) lorsque la taille de l'échantillon est modérée. L'intervalle de Wald standard est obtenu par inversion de la famille de régions d'acceptation du test de l'hypothèse $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ fondé sur la statistique $T = (\hat{p} - p_0)/\sqrt{\hat{p}(1-\hat{p})/n}$ où n désigne la taille de l'échantillon. Cet intervalle de niveau nominal $\eta = 1-\alpha$ est exprimé de manière simple en utilisant l'approximation normale :

$$S_{Ws}(\hat{p},\eta) = \hat{p} \pm \kappa\sqrt{\hat{p}(1-\hat{p})/n}$$

où κ désigne le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite.

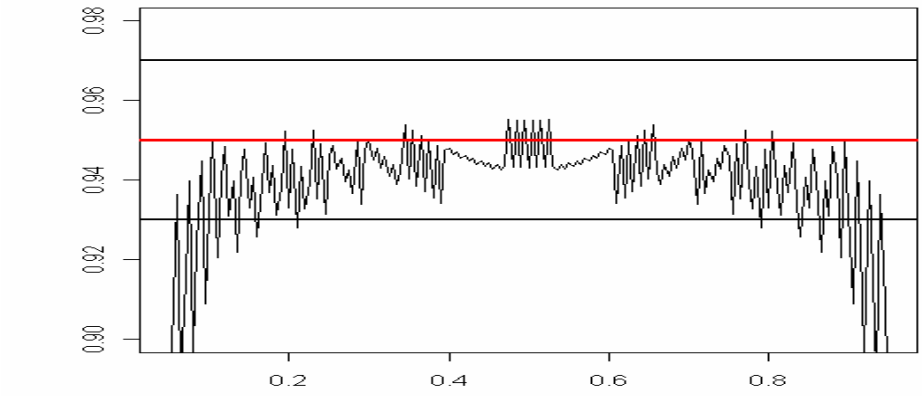

Fig. 1 : Le taux réel de recouvrements de l'Intervalle de Wald standard en fonction de *n* pour *p=0.5*, le taux de recouvrement nominal étant de 0.95

Si cet intervalle de Wald a l'avantage d'être facilement interprétable, il est cependant trop laxiste [Fig.1] en ce sens que la probabilité de recouvrement de p est généralement faible, même lorsque le paramètre p se trouve dans l'intervalle $[0.2-0.8]$. Pour assurer une probabilité de recouvrement de 0.93, la taille de l'échantillon doit dépasser $n = 118$ si $p = 0.2$. Les oscillations observées en fonction de la taille de l'échantillon peuvent dérouter le lycéen ou l'étudiant débutant en statistique. En effet, pour un échantillon de petite taille, $n = 23$, cette probabilité vaut 0.951, très proche de la valeur nominale $\eta = 0.95$, alors qu'en doublant sa taille $(n = 48)$, elle chute à 0.906.

L'intervalle « exact » $S_{CP}(t)$ de Clopper-Pearson ne requiert pas l'approximation normale. La valeur $t = \sum x_i$ étant observée, $S_{CP}(t) = [l_{CP}(t), u_{CP}(t)]$ où $l_{CP}(t)$ et $u_{CP}(t)$ sont les solutions en p_0 des

équations : $\Pr_{p_0}(T \geq t) = \alpha/2$ et $\Pr_{p_0}(T \leq t) = \alpha/2$. Ces équations ont une solution explicite : $l_{CP}(t)$ est le quantile $\alpha/2$ d'une distribution $Beta(t, n-t+1)$ et $u_{CP}(t)$ le quantile $1-\alpha/2$ d'une loi $Beta(t+1, n-t)$ [1].

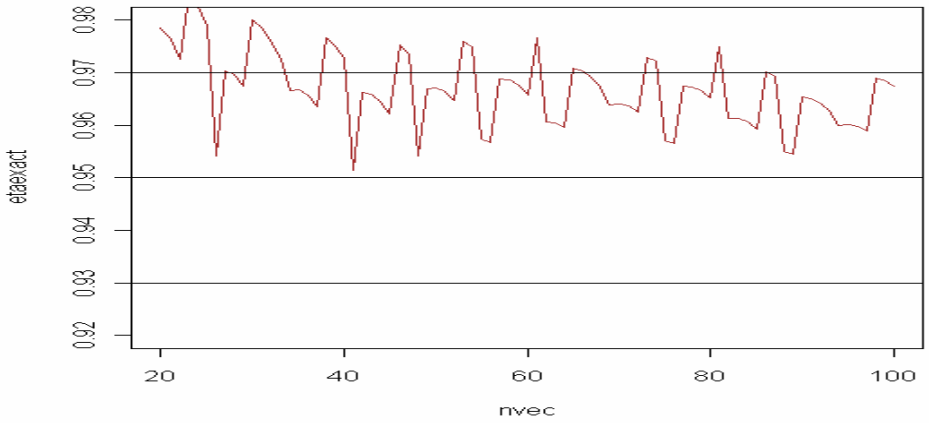

Fig. 2 : Le taux réel de recouvrements de l'Intervalle de Clopper-Pearson en fonction de *n* pour *p=0.5*, le taux de recouvrement nominatif étant de 0.95.

L'intervalle de Clopper et Pearson [Fig.2], très utilisé et que l'on trouve par exemple dans la norme AFNOR [ISO 3534] apparait bien trop conservatif : si $p = 0.5$, ce n'est que pour une taille d'échantillon $n > 85$ que l'on est sûr que la probabilité réelle de recouvrement ne dépasse pas 0.97. Il en résulte un manque de précision.

2. Autres intervalles de confiance

Bien d'autres intervalles de confiance pour une proportion ont été proposés : parmi ceux appelés exacts (i.e., ne faisant pas appel à une approximation normale), l'intervalle mid-P [6] est moins conservatif que le précédent ; l'intervalle Bayésien fondé sur un a priori non informatif de Jeffreys présente un taux de recouvrement bien plus satisfaisant : dès que $n > 31$, la probabilité vraie de recouvrement est dans l'intervalle $[0.93 - 0.97]$. Comme méthode approchée, en remplaçant l'écart type estimé (Wald) par l'écart type sous H_0, on obtient l'intervalle de Wilson appelé aussi l'intervalle du score, bien préférable à l'intervalle de Wald standard : dès que $n > 40$, le taux de recouvrement se trouve dans l'intervalle $[0.93 - 0.97]$, même pour p petit $(p = 0.05, p = 0.10)$. Cet intervalle est recommandé, cependant son expression ne présente pas la simplicité de l'intervalle de Wald ; il s'écrit :

$$S_W(\hat{p}, \eta) = \hat{p} + \frac{\kappa^2}{2n} \pm \frac{\kappa}{\sqrt{n}} \sqrt{\left[\hat{p}(1-\hat{p}) + \kappa^2/4n\right]}$$

On trouve dans certains ouvrages récents [8] l'intervalle d'Agresti-Coull [1] qui est intéressant du point de vue enseignement. Il s'interprète simplement, comme l'intervalle de Wald, mais il est bien plus satisfaisant que celui-ci et que l'intervalle de Clopper-Pearson. L'idée de ces auteurs est de recentrer l'intervalle de Wilson. En écrivant $\tilde{p}$ sous la forme :

$$\tilde{p} = \hat{p}\left(\frac{n}{n+\kappa^2}\right) + \frac{1}{2}\left(\frac{\kappa^2}{n+\kappa^2}\right)$$

On peut interpréter $\tilde{p}$ comme un rétrécissement de l'estimateur $\hat{p}$ vers $1/2$. Soit $\tilde{T} = T + \kappa^2/2$ et $\tilde{n} = n + \kappa^2$; on a : $\tilde{p} = \tilde{T}/\tilde{n}$. L'intervalle d'Agresti et Coull est défini par :

$$S_{AC}(\tilde{p}) = \tilde{p} \pm \kappa \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$$

Si l'on choisit comme niveau de confiance $\eta = 0.95$, alors $\kappa \cong 2$ et cet intervalle s'identifie à l'intervalle de Wald standard avec la règle « plus deux succès, plus deux échecs » ajoutés à l'échantillon $\tilde{X}$ observé. Cet intervalle tend toutefois à être conservatif en regard de l'intervalle de Wilson.

Signalons encore l'intervalle du rapport de vraisemblance S_{rv} dont la performance est comparable à celle de l'intervalle de Wilson et l'intervalle de Stevens. Ce dernier est construit de manière différente [7] : il repose sur la statistique $T = \sum X_i + U$ où U est une variable aléatoire suivant une loi uniforme sur $[0,1]$. La loi de T est continue à rapport de vraisemblance monotone ; il en résulte un intervalle de confiance uniformément le plus précis dans la classe des intervalles sans biais de niveau exactement égal à la valeur nominal η. Il ne présente donc ni biais ni oscillations contrairement à tous les autres, mais du fait de l'introduction de U, sa précision est moindre.

3. Comparaison des intervalles de confiance

3.1 Développement d'Edgeworth

Une manière très convaincante de comparer les performances de ces différents intervalles est de réaliser un développement d'Edgeworth à l'ordre 2 de la probabilité exacte de recouvrement $C_*(p,n) = \Pr_p(p \in S_*(\hat{p},\eta))$. Cette approximation s'écrit [2] :

$$\begin{aligned} C_*(p,n) = (1-\alpha) + & \text{ « oscillations en } O\left(n^{-1/2}\right) \text{ »} \\ & + \text{ « biais en } O\left(n^{-1}\right) \text{ » } + \text{ « oscillation en } O\left(n^{-1}\right) \text{ »} \\ & + O\left(n^{-3/2}\right) \end{aligned}$$

Les calculs sont assez techniques [5] et la qualité de l'approximation est remarquable pour tous les intervalles déduits d'une approximation normale. La composante de biais en $O\left(n^{-1}\right)$ est représentée sur la figure 3. On observe qu'il n'y a quasiment pas de biais pour les intervalles de Wilson, de Jeffreys et du rapport de vraisemblance ; que le bais de l'intervalle d'Agresti-Coull est positif, mais peu important si $0.2 \le p \le 0.8$; que l'intervalle de Wald est trop laxiste. L'intervalle simplifié $\hat{p} \pm \frac{1}{\sqrt{n}}$ présenté dans les programmes de formations technologiques, également représenté, n'est pas un réel intervalle de confiance.

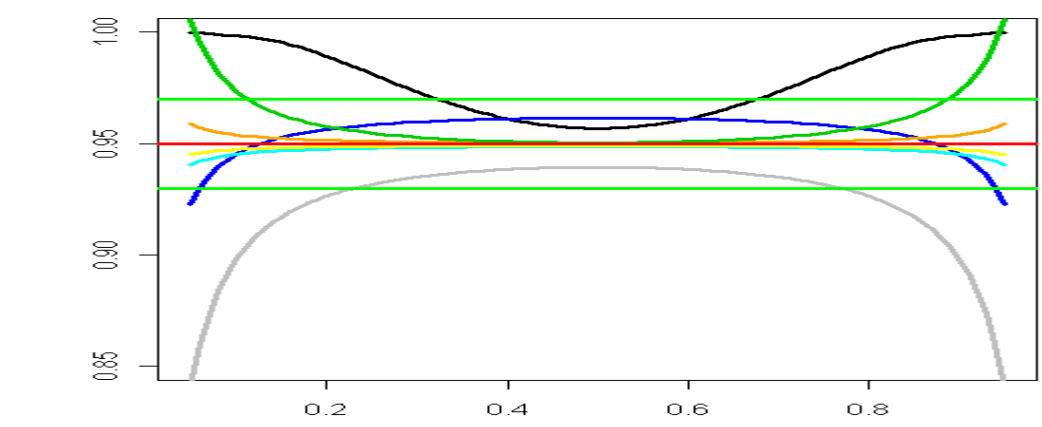

Fig. 3: Approximation d'Edgeworth à l'ordre 2 de la partie non oscillatoire de la probabilité de recouvrement des différents intervalles. De haut en bas : l'intervalle simplifié (noir), Agresti-Coull (vert foncé), Wilson (orange), la ligne 0.95 Stevens (rouge), Jeffreys (jaune), maximum de vraisemblance (bleu ciel), Wald recentré (bleu foncé), Wald standard (gris)

L'approximation oscillatoire $O_*(p,n)$ donnée par le développement d'Edgeworth est une fonction de p de haute fréquence, de sorte que p étant inconnu, une approximation explicite en p ne présente aucun intérêt pratique. On peut cependant apprécier l'amplitude des oscillations dans la probabilité de couverture des intervalles en calculant l'amplitude moyenne des oscillations par lissage selon une densité sur le paramètre p, notée $f(p)$ sur [0-1].

$$O_*(n) = \int \left| C_*(p,n) - (1-\alpha) - Biais\left(O\left(n^{-1}\right)\right) \right| f(p)dp$$

On observe que cette amplitude moyenne est plus faible que le biais et comparable pour les différents intervalles, hormis Wald standard.

3.2 Précision des intervalles de confiance :espérance de la longueur de l'intervalle

Le défaut de l'intervalle de confiance de Wald est de présenter un taux de couverture insuffisant. Celui de Clopper-Pearson est d'engendrer un manque de précision. La précision est généralement mesurée soit par l'espérance de la longueur de l'intervalle (critiquée par Lehmann) soit par la fonction de probabilité de recouvrement de valeurs erronées de p, directement liée à la théorie des tests. On montre comment ces deux notions sont liées, la seconde étant plus exigeante que la première.
L'espérance de la longueur d'un intervalle $S_*(\hat{p},\eta)$ s'écrit :

$$L_*(n,p) = \sum_{x=0}^{n} (u_*(x) - l_*(x)) \binom{n}{x} p^x (1-p)^{n-x}$$

On observe [Fig.4], que l'intervalle de Clopper-Pearson, trop conservatif est bien peu précis ; il en est de même de l'intervalle de Wald qui, lui pourtant est laxiste. Les intervalles de Wilson, du rapport de vraisemblance et de Jeffreys sont plus précis et assez comparables en termes de longueur. L'intervalle d'Agresti-Coull leur est comparable pourvu que la valeur de p ne se rapproche pas trop de 0 ou de 1 ($0{,}25 \le p \le 0.75$).

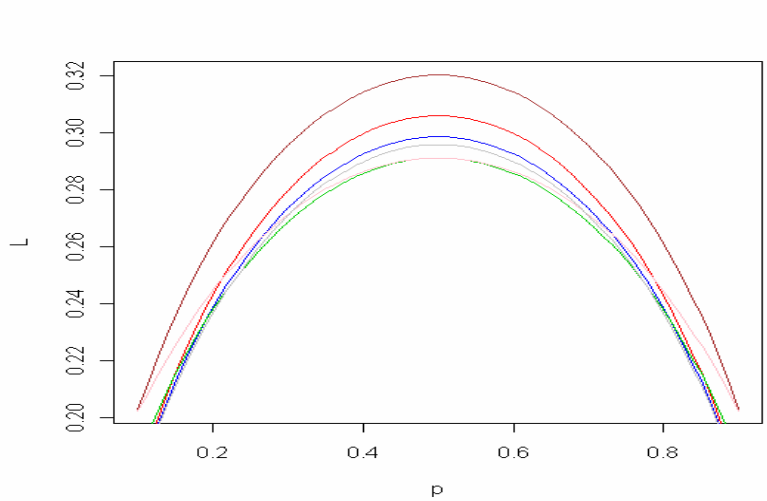

Fig. 4: Les espérances des longueurs des différents IC pour n=40. De haut en bas : en brun, l'IC de Clopper et Pearson ; en rouge, l'IC de Wald ; en bleu, l'IC du rapport de vraisemblance ; en gris , l'IC de Jeffreys ; en rose, l'IC d'Agresti-Coull ; en vert, l'IC de Wilson.

3.3 Recouvrement des valeurs erronnées

Une seconde approche de la précision est la fonction de recouvrement de valeurs erronées, directement liée à la notion de puissance de la famille de tests dont l'inversion de la région d'acceptation conduit aux intervalles de confiance considérés ici. La probabilité de recouvrement d'une valeur p_A

erronées $(p_A \neq p)$ est donnée par :

$$C_*(p_A; p, n) = \Pr_p(p_A \in S_*(\hat{p})) = \sum_{x=0}^{n} I_*(p_A, x) \binom{n}{x} p^x (1-p)^{n-x}$$

où $I_*(p_A, x) = 1$ si l'intervalle contient p_A lorsque $\sum X_i = x$, $I_*(p_A, x) = 0$ sinon.

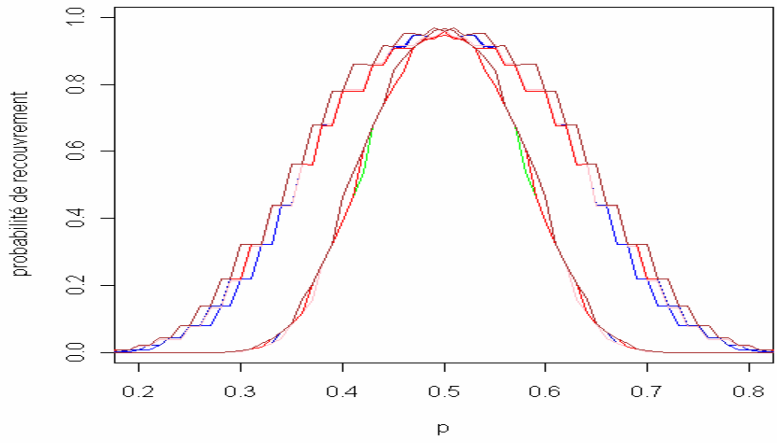

Fig.5 : Probabilités de recouvrement de valeurs erronées de *p*, lorsque la vraie valeur de p vaut 0.5. Les courbes supérieures sur le graphique correspondent à n=40, les courbes inférieures à n=120. En vert, l'*IC* de Wilson ; en bleu, le maximum de vraisemblance ; en rouge, Wald standard; en rose, Agresti-Coull ; en brun, l'IC exact de Clopper-Pearson.

Si l'intervalle de Clopper-Pearson conduit à des probabilités de recouvrement de valeurs erronées supérieures aux autres intervalles considérés, on voit que pour $p_0 = 0.5$ l'intervalle le plus précis dépend de p_A. Pour des valeurs éloignées de $p_0 = 0.5$, l'intervalle du maximum de vraisemblance, l'intervalle de Wilson, d'Agresti-Coull semblent préférables.

4. Conclusion

Pour conclure, plusieurs intervalles satisfont à la condition d'un recouvrement ni insuffisant (défaut de couverture) ni excessif (conservatif) à partir d'une taille d'échantillon modérée. C'est le cas en particulier de l'intervalle d'Agresti-Coull, dont l'expression est de type « intervalle de Wald », le plus simple à intégrer pour un étudiant débutant en statistique. Nous pensons que la notion de précision d'un intervalle doit être prise en compte de la même manière que la puissance d'un test l'est. L'intervalle de Wald simplifié figurant dans le Bulletin Officiel pour les classes technologiques ne prend pas en compte cette exigence. Il reste à comparer ces approches très variées de la construction d' un intervalle de confiance pour une loi binomiale à diverses techniques Bootstrap.

Bibliographie

[1] A.AGRESTI, B.A.COULL (1998), Approximate is Better than "exact" for Interval Estimation of Binomial Proportions, *American Statistician*, Vol. 52 n°2 (1998) pp 119-126

[2] R. BHATTACHARYA, R.R. RAO (1976), Normal approximation and asymptotic expansions, Wiley

[3] G. BERRY, P. ARMITAGE (1995), Mid-P confidence intervals: a brief review, *The Statistician*, 44, N°4, pp 417-423

[4] L.D. BROWN, T. CAI, A. DASGUPTA (2001), Interval Estimation for a Binomial Proportion,

*Statistical Science***,** Vol. 16 (2001), pp 101-117

[5] L.D.BROWN, T.CAI, A. DASGUPTA (2002), Confidence Intervals for a Binomial Proportion and Asymptotic Expansions, *The Annals of Statistics***,** Vol. 30 (2002), pp 160-201

[6] H.O. LANCASTER (1961), Significance tests in discrete distributions, *J. Am. Statist. Ass.,* 56,

pp 223-234

[7] E.L. LEHMANN (1959), Testing Statistical Hypotheses. *J. Wiley*, 1959.

[8] R. G. NEWCOMBE (1998), Two-Sided confidence intervals for the single proportion : comparison of seven methods, Statistics in Medicine, 17, 857-872

[9] R.G. NEWCOMBE (2013), Confidence Intervals for Proportions and Related Measures of Effect Size, Chapman & Hall, CRC Biostatistics Series

[10] M.L.SAMUELS, TAI-FANG C.LU (1992), Sample Size Requirements for the Back-of-the Envelope Binomial Confidence Interval, *The American Statistician,* Vol. 46, n°3 (1992) pp228-231

[11] T.J. SANTNER (1998), A note on teaching binomial confidence intervals. *Teaching Statistics*, **20**, pp 20-23