Intervalles de confiance d'une proportion et lois binomiales

Que ce soit en Seconde, avec les fourchettes de sondage, ou en BTSA, la notion d'estimation d'une proportion par intervalle de confiance est présente et fait toujours référence (explicitement ou non) à l'utilisation de lois normales dont l'usage n'a de sens que dans le cas de grands échantillons.
Alors que faire si on considère un petit échantillon (le fameux "$n < 30$") ?

Comme nous allons le voir dans cet article, on peut tout de même déterminer un intervalle de confiance mais en utilisant une loi "exacte", en l'occurrence, une loi binomiale.
Dans ce qui suit, nous allons nous intéresser à un échantillon aléatoire et simple de taille n=34 afin, non seulement, de présenter la méthode mais également de comparer le résultat obtenu avec celui obtenu par la méthode "approchée".
Considérons une population quelconque sur laquelle on étudie un caractère prenant deux modalités dont une est notée A. On note p la proportion (inconnue) d'éléments de la population présentant cette modalité.
Supposons que 20 éléments de l'échantillon présentent cette modalité.
La proportion observée est alors égale à $f = \frac{20}{34}$ et la fourchette de sondage, obtenue en utilisant une approximation par une loi normale, $\left[f - 1{,}96 \times \sqrt{\frac{f(1-f)}{n}} ; f + 1{,}96 \times \sqrt{\frac{f(1-f)}{n}} \right]$ nous permet de conclure qu'une estimation par intervalle de confiance à 95 % de la proportion d'éléments de la population présentant cette modalité est $[0{,}422 ; 0{,}754]$.

Intéressons-nous désormais à la loi "exacte" sous-jacente.
On note X la variable aléatoire prenant pour valeurs le nombre d'éléments d'un échantillon de taille 34 présentant la modalité A. X est distribuée suivant la loi binomiale $B(34 ; p)$ qui peut être approchée par une loi normale conduisant à la formule ci-dessus.

Cherchons un intervalle de confiance à 95 % (symétrique en probabilité).
Pour différentes valeurs de p, nous allons donc déterminer deux entiers N_{1p} et N_{2p} tels que N_{1p} est le plus grand des entiers k tels que $P(X < k) \leq 0{,}025$ et N_{2p} est le plus petit des entiers k tels que $P(X > k) \leq 0{,}025$.

On a donc : $P(X < N_{1p}) \leq 0{,}025$ et $P(X > N_{2p}) \leq 0{,}025$ ce qui permet d'en déduire que $P(N_{1p} \leq X \leq N_{2p}) \geq 0{,}95$.

Supposons, par exemple, que $p = 0{,}4$. X est donc distribuée suivant la loi binomiale $B(34\ ;\ 0{,}4)$.

A l'aide d'un tableur, on obtient les résultats suivants :

k	$P(X<k)$
1	$2{,}9 \times 10^{-8}$
2	$6{,}8 \times 10^{-7}$
3	$7{,}8 \times 10^{-6}$
4	$5{,}9 \times 10^{-5}$
5	0,0003
6	0,0014
7	0,0048
8	**0,0138**
9	0,0341
⋮	⋮

k	$P(X>k)$
34	0
33	3×10^{-14}
32	2×10^{-12}
31	4×10^{-11}
30	6×10^{-10}
29	8×10^{-9}
28	7×10^{-8}
27	5×10^{-7}
26	3×10^{-6}
25	2×10^{-5}
24	8×10^{-5}
23	0,0003
22	0,0010
21	0,0031
20	0,0085
19	**0,0205**
18	0,0444
⋮	⋮

$N_{1\,;\,0,4}$ est le plus grand des entiers k tels que $P(X < k) \leq 0{,}025$, donc $N_{1\,;\,0,4}=8$.

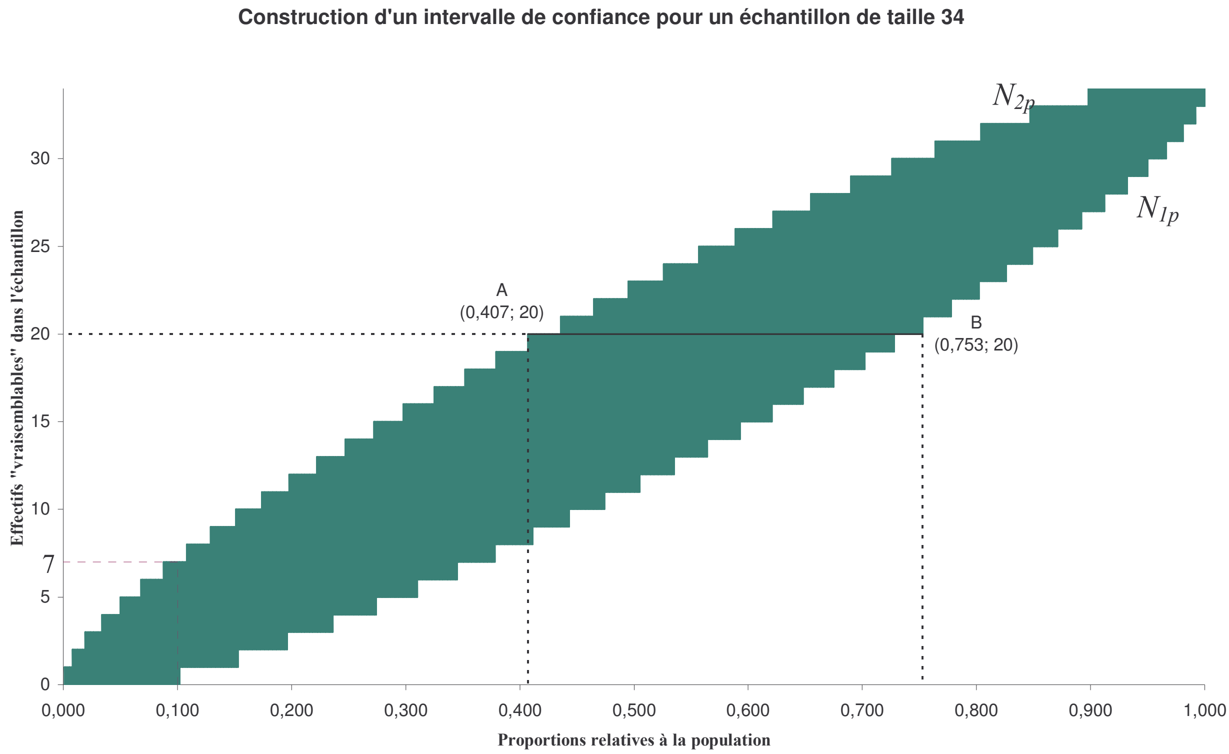
$N_{2\,;\,0,4}$ est le plus petit des entiers k tels que $P(X > k) \leq 0{,}025$, donc $N_{2\,;\,0,4}=19$.

Ainsi, si la proportion p (dans la population) est égale à 0,4 alors la probabilité que le nombre d'observations de la modalité A (dans l'échantillon) soit compris entre 8 et 19 est supérieure ou égale à 0,95.

Dans cet exemple, on a : $P(8 \leq X \leq 19) = P(X \leq 19) - P(X \leq 7) = 0{,}966$ à 10^{-3} près.

Il en résulte qu'il est "peu vraisemblable" qu'il y ait 20 observations de la modalité A dans l'échantillon dans le cas où p serait égale à 0,4.

En réitérant cette méthode pour différentes valeurs de p (balayage de l'intervalle [0 ; 1] avec un pas égal à 0,001), on a obtenu la représentation graphique suivante où on a matérialisé pour chaque valeur de p l'intervalle $[N_{1p}\,;N_{2p}]$ correspondant :



D'après le graphique précédent, pour un échantillon de taille 34, si la proportion dans la population est égale à 0,1 (lue sur l'axe des abscisses) alors le nombre d'observations est "vraisemblablement" compris entre 0 et 7 (valeurs lues sur l'axe des ordonnées).
Il en résulte qu'il est "peu vraisemblable" qu'il y ait 20 observations de la modalité A sur l'échantillon dans le cas où p serait égal à 0,1.

En utilisant le graphique, on constate que la plus petite valeur de p rendant 20 observations "vraisemblables" est 0,407 (cf. le point A de coordonnées (0,407 ; 20) sur le graphique).
De même, la plus grande valeur de p rendant 20 observations "vraisemblables" est 0,753 (cf. le point B de coordonnées (0,753 ; 20)).
Ainsi, sans utiliser d'approximations par des lois normales, on constate qu'une estimation par intervalle de confiance à 95 % de la proportion p d'individus présentant la modalité A dans la population est [0,407 ; 0,753].
Il est à noter que la borne inférieure de l'intervalle correspond à la valeur de p telle que $P(X \geq 20) = 0{,}025$ et que la borne supérieure de l'intervalle correspond à la valeur de p telle que $P(X \leq 20) = 0{,}025$, sachant que la loi de X est la loi binomiale $B(34\,;p)$.

On peut constater que cet intervalle a des bornes voisines de l'intervalle $[0{,}422\ ;\ 0{,}754]$ obtenu avec l'approximation normale pour lequel on a appliqué des formules faciles à mettre en œuvre, ce qui n'est pas le cas dans la méthode exposée dans cet article.

Alors, que dire ? Tout ça pour conclure que notre bonne vieille méthode est plus pratique et propose des résultats satisfaisants ? Bien évidemment non puisque, dans le cas d'échantillons de faible taille, seule la méthode présentée ici est applicable comme en atteste ce qui suit.
En effet, supposons que 2 éléments d'un échantillon de taille 9 présentent une modalité donnée.
Avec la méthode exacte, on obtient [0,028 ; 0,600] comme estimation par intervalle de confiance à 95 % de la proportion contre [0 ; 0,494] avec la "fameuse formule"…