UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESTIMATION ET PRÉVISION AMÉLIORÉES DU PARAMÈTRE D'UNE LOI BINOMIALE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

AHMED NEMIRI

MARS 2012

REMERCIEMENTS

En tout premier lieu, je tiens à remercier mon directeur de recherche, Monsieur Glenn Shorrock, et ma codirectrice de recherche, Madame Brenda MacGibbon, qui ont su être disponibles pour répondre à mes nombreuses questions et me conseiller pour prendre la bonne direction. Leur rigueur m'a aidé à éviter de pencher vers le faux.

Mes remerciements s'adressent également à Madame Geneviève Lefebvre et à Monsieur François Watier, Professeurs de mathématiques à l'UQAM, qui ont eu la gentillesse de lire et corriger ce travail.

Bien sûr, j'exprime toute ma reconnaissance à Nabyla, mon épouse, pour son soutien constant, sa compréhension et surtout son encouragement qui m'ont permis de réaliser des études à la maîtrise à temps plein.

Enfin, je dédie ce mémoire à mes enfants : Maryam, Leila, Abdelghani et Taher et à mes défunts parents : ma mère et Khali.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Dans ce mémoire, on présente une étude sur l'estimation et la prévision du paramètre binomial.

Le Chapitre 1 traite de l'estimation ponctuelle et de la prévision du paramètre binomial. En suivant l'approche de Brown (2008*a*), on commence ce chapitre par la description de six estimateurs : trivial, moyenne générale, Bayes empirique paramétrique avec la méthode des moments, Bayes empirique paramétrique avec la méthode du maximum de vraisemblance, Bayes empirique non paramétrique et James-Stein. Ensuite, on évalue ces estimateurs en se servant de la base de données de baseball 2005 de Brown (2008*b*) et on finit par la comparaison des performances de ces estimateurs entre elles, selon leurs écarts quadratiques totaux normalisés.

Le Chapitre 2 traite de l'estimation par intervalle de confiance et de la prévision du paramètre binomial. Dans ce chapitre, on étudie cinq intervalles de confiance en suivant l'approche de Brown, Cai et DasGupta (1999) et (2001) : standard ICs, Wilson ICw, Agresti-Coull $ICac$, maximum de vraisemblance $ICrv$ et Jeffreys bilatéral ICj. En premier, vu l'importance particulière de l'intervalle standard, on calcule théoriquement, avec un n modéré, la déviation du biais, de la variance et des coefficients d'asymétrie et d'applatissement de la variable aléatoire $W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{\text{loi}} N(0,1)$ par rapport à leurs valeurs asymptotiques correspondantes 0, 1, 0 et 3. Ensuite, on approxime la probabilité de couverture et la longueur moyenne de chacun des cinq intervalles de confiance mentionnés plus haut par un développement d'Edgeworth d'ordres 1 et 2. Enfin, en se servant de la même base de données de baseball 2005, on détermine ces intervalles ainsi que leurs probabilités de couverture et leurs longueurs moyennes et on compare leurs performances entre elles, selon leurs probabilités de couverture et leurs longueurs moyennes.

Mots clés : estimateur de Bayes empirique paramétrique, méthode des moments, méthode du maximum de vraisemblance, estimateur de Bayes empirique non paramétrique, estimateur de James-Stein, développement d'Edgeworth d'ordres 1 et 2, intervalle de Wald (standard), intervalle de Wilson, intervalle d'Agresti-Coull, intervalle du rapport de vraisemblance, intervalle de Jeffreys bilatéral, programmes en R.

INTRODUCTION

Le travail mené dans le cadre de ce mémoire est une tentative d'améliorer l'estimation et la prévision du paramètre d'une loi binomiale selon l'approche de Brown, Cai et DasGupta (1999) et (2001) et Brown (2008*a*).

L'idée de base développée dans les deux chapitres composant ce mémoire est l'étude théorique de l'estimation ponctuelle et par intervalle de confiance et son application à la prévision sur la base de données de baseball de Brown (2008*b*). Cette base est relative à la saison régulière de 2005 en ligue majeure de baseball. La saison est divisée en six mois dont le premier est le mois d'avril et le dernier est formé par les rencontres jouées au mois de septembre plus celles jouées au début du mois d'octobre. Les rencontres jouées hors saison et celles des séries mondiales ne sont pas incluses dans le dernier mois. La base inclut 929 joueurs dont 301 lanceurs et 628 non-lanceurs. Parmi toutes les données de cette base, on ne s'intéresse qu'aux colonnes donnant le nombre de présences au bâton et le nombre de coups sûrs de chaque joueur. En se servant des données de ces colonnes, on calcule une moyenne que l'on nomme moyenne au bâton en divisant le nombre de coups sûrs par le nombre de présences au bâton. Cette moyenne est utilisée pour évaluer la performance d'un joueur de baseball. En symbole, la moyenne au bâton du i^e joueur est : $R_i = H_i/N_i$, où H_i est le nombre de coups sûrs et N_i est le nombre de présences au bâton. Il est naturel de modéliser H_i par une variable aléatoire suivant une loi binomiale $Bin(N_i\,,\,p_i)$, où le paramètre p_i est inconnu et représente l'habileté latente du joueur i. Dans toute notre étude, on se sert des données d'une partie du début de la saison pour estimer ce paramètre p_i et prévoir celui du restant de la même saison.

Dans le Chapitre 1, on commence par la stabilisation de la variance de R_i en utilisant la transformation $X_i^{(c)} = \arcsin\sqrt{\frac{H_i+c}{N_i+2c}}$ avec $c \in \mathbb{R}$. Le cas $c = 1/4$ donne

le meilleur compromis biais-variance pour $X_i^{(c)}$ (Figure 1 et Figure 2, Brown (2008a)). Dans ce cas, le biais est presque nul à partir de $N_i = 11$ (Figure 1, Brown (2008a)). C'est pour toutes ces raisons que, dans notre étude, on utilise $X_i^{(1/4)}$ au lieu de R_i et on ne considère que les frappeurs ayant au moins 11 présences au bâton. Aussi, dans notre étude, on sépare le cas des lanceurs de celui des non-lanceurs, car la différence entre leurs moyennes au bâton est très significative. Ensuite, on applique sur la base de données mentionnée plus haut des méthodes d'estimation ponctuelle découlant directement d'une approche de Bayes empirique ainsi que les méthodes de James-Stein, moyenne générale et triviale. Il est à noter que la méthode triviale est tout simplement la méthode qui utilise la moyenne au bâton d'une première partie de la saison comme étant la prévision de celle du restant de la même saison. En se servant de la somme quadratique de prévision comme critère de comparaison entre les performances de tous les estimateurs étudiés, l'estimateur de Bayes empirique non paramétrique sort avec la meilleure performance et l'estimateur trivial avec la pire.

Dans le Chapitre 2, on aborde l'estimation par intervalle de confiance du paramètre d'une loi binomiale selon l'approche de Brown, Cai et DasGupta (1999) et (2001). La méthodologie que l'on adopte pour étudier ce grand classique en statistique est la comparaison de la performance de l'intervalle de confiance standard ICs connu sous le nom de l'intervalle de Wald aux intervalles : de Wilson ICw, d'Agresti-Coull $ICac$, du maximum de vraisemblance $ICrv$ et de Jeffreys bilatéral ICj. Cette comparaison se fait sur la base des probabilités de couverture et des longueurs moyennes de ces intervalles. On rappelle que l'intervalle standard est basé sur l'approximation de la loi binomiale $B(n, p)$ par la loi normale $N\big(np, np(1-p)\big)$. Il est défini par : $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, où $\hat{p} = X/n$ est la proportion échantillonnale des succès et $z_{\alpha/2}$ est le $100\,(1-\alpha/2)^{\text{e}}$ percentile de la loi normale centrée réduite $N(0, 1)$. Il est connu que cet intervalle a une faible probabilité de couverture lorsque p est proche des frontières 0 et 1. C'est pourquoi l'utilisation de cet intervalle est accompagnée de conditions telles que $\min\{np(1-p)\} \geq 5$ ou 10. Mais il est démontré

par de récents articles que la probabilité de couverture de cet intervalle, qui dépend de n et p, a un comportement chaotique et imprévisible lorsque ces deux paramètres varient. Ceci nous amène à suggérer d'autres intervalles de confiance alternatifs, tels que mentionnés plus haut, pour étudier leurs performances.

Vu l'importance de l'intervalle standard, on commence la Section 2 par l'analyse de la déviation du biais, de la variance et des coefficients d'asymétrie et d'applatissement de la variable aléatoire $W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{\text{loi}} N(0,1)$, avec un n modéré, par rapport à leurs valeurs asymptotiques correspondantes 0, 1, 0 et 3. Le reste de la section est consacré au développement d'Edgeworth de la probabilité de couverture et de la longueur moyenne de chacun des cinq intervalles cités plus haut. On montre par des exemples bien choisis que l'approximation de la probabilité de couverture par un développement d'Edgeworth d'ordre 1 n'est pas assez exacte. Quant au développement d'Edgeworth d'ordre 2, il approxime la probabilité de couverture et la longueur moyenne avec une exactitude adéquate même pour un échantillon de taille modeste. On utilise ce développement d'Edgeworth comme moyen d'analyse pour comparer et ordonner les performances des cinq intervalles cités plus haut. Du point de vue de la probabilité de couverture, ce développement d'Edgeworth montre que l'intervalle d'Agresti-Coull a la meilleure performance et celles de Wilson, du rapport de vraisemblance et de Jeffreys bilatéral sont comparables.

Un autre critère pour évaluer la performance d'un intervalle de confiance est la longueur moyenne. Du développement d'Edgeworth d'ordre 2 de la longueur moyenne de chacun de ces cinq intervalles, on voit que l'intervalle d'Agresti-Coull est toujours le plus long. Les intervalles de Wilson et de Wald ont des longueurs moyennes plus ou moins comparables. L'intervalle de Jeffreys bilatéral est toujours le plus court. L'intervalle du maximum de vraisemblance est un peu plus long que celui de Jeffreys bilatéral.

De ce calcul théorique relatif au développement d'Edgeworth de la probabilité de couverture et de la longueur moyenne, on voit que les intervalles alternatifs suggérés performent mieux que l'intervalle standard.

La Section 3 est consacrée à l'application de la théorie développée pour les méthodes d'estimation par intervalle de confiance de la section précédente. Comme dans le Chapitre 1, cette application se fait sur la base de données 2005 de Brown (2008*b*).

Pour les différentes parties du début de saison utilisées comme période de prévision, les cinq intervalles de confiance introduits dans la Section 2 sont calculés pour les moyennes au bâton de chaque frappeur sur la base de la statistique $R_i = H_i/N_i$, où $H_i \sim B(N_i, p_i)$. Les frappeurs pris en compte sont ceux qui ont au moins 11 présences au bâton. Pour chaque période de prévision, comme par exemple celle qui s'étale sur les 2 premiers mois de la saison, les longueurs moyennes et les probabilités de couverture moyennes sont calculées pour les cinq intervalles pour les deux cas : non-lanceurs et tous (lanceurs et non-lanceurs).

Indépendamment de la période de prévision et du cas considéré (non-lanceurs ou tous), les probabilités de couverture moyennes de ces cinq intervalles sont classées dans l'ordre décroissant suivant : $ICac$, ICw, $ICrv$, ICj et ICs.

Lorsque la longueur moyenne est prise pour critère d'évaluation des performances de ces intervalles, on n'a pas un classement unique des longueurs pour toutes les situations. Cependant, on a ces deux résultats : l'intervalle d'Agresti-Coull $ICac$ est toujours le plus long et la longueur moyenne de n'importe quel intervalle est inversement proportionnelle à la largeur de la période de prévision utilisée et ceci pour n'importe quel groupe de frappeurs étudié.

CHAPITRE I

ESTIMATION PONCTUELLE ET PRÉVISION DU PARAMÈTRE BINOMIAL

1.1 Introduction

La moyenne au bâton est une statistique pour évaluer la performance d'un joueur de baseball et plus précisément d'un frappeur. Cette moyenne est calculée en divisant le nombre de coups sûrs par le nombre de présences au bâton, c'est-à-dire le nombre de fois qu'un frappeur atteint l'une des quatre bases avec un coup sûr. À titre d'exemple, une moyenne au bâton de 0.3 (3 coups sûrs sur chaque 10 présences au bâton) est considérée comme excellente. Le frappeur, Nolan Gallagher Reimold, évoluant en ligue majeure de baseball avec les Orides de Baltimore, a réalisé une moyenne au bâton de 0.36 en 2005 et il a été désigné le joueur de l'année en 2005.

Notre étude est basée, comme celle de Brown (2008*a*), sur les données de la saison régulière de 2005 en ligue majeure de baseball qui commence le 3 avril et finit le 2 octobre. Cette saison est divisée en six mois dont le premier est le mois d'avril et le dernier est formé par les rencontres jouées au mois de septembre plus celles jouées au début du mois d'octobre. Les rencontres jouées hors saison et celles des séries mondiales ne sont pas incluses dans le dernier mois. La base de données de Brown (2008*b*) inclut 929 joueurs dont 301 lanceurs et 628 non-lanceurs. Étant donné que les lanceurs et les non-lanceurs ont des moyennes au bâton très différentes, ils seront séparés dans l'étude pour que les prévisions soient plus exactes. Les résultats présentés dans le tableau 1.1.1 et le tableau 1.1.2 ci-dessous illustrent cette différence.

Tableau 1.1.1 Moyenne générale au bâton pour les lanceurs, non-lanceurs et tous (lanceurs et non-lanceurs) lorsque la prévision est basée sur la première moitié de la saison

	1^{re} moitié de la saison	2^{e} moitié de la saison
Non-lanceurs	0.255	0.252
Lanceurs	0.153	0.145
Tous	0.240	0.237

Tableau 1.1.2 Moyenne générale au bâton pour les lanceurs, non-lanceurs et tous (lanceurs et non-lanceurs) lorsque la prévision est basée sur les 2 premiers mois et les 4 premiers mois de la saison

	2 premiers mois	4 derniers mois	4 premiers mois	2 derniers mois
Non-lanceurs	0.256	0.254	0.254	0.253
Lanceurs	0.148	0.146	0.149	0.147
Tous	0.240	0.237	0.238	0.239

Le but de l'étude est d'estimer la moyenne au bâton d'un joueur sur une partie du début de la saison 2005 pour prévoir celle du restant de la même saison. De plus, comme cette saison est déjà terminée, on peut faire la validation en comparant la valeur de notre prévision de la moyenne au bâton à la vraie valeur du restant de la saison. Brown (2008*a*) a étudié trois cas :

- Prévision basée sur la 1^{re} moitié de la saison ;
- Prévision basée sur le 1^{er} mois de la saison ;
- Prévision basée sur les 5 premiers mois de la saison.

Quant à nous, la prévision est basée une fois sur les 2 premiers mois de la saison et une autre fois sur les 4 premiers mois. Ceci nous permet de comparer les résultats obtenus avec ceux de Brown (2008*a*).

Dans ce chapitre, on commence par la description de cinq estimateurs ponctuels et le modèle appliqué pour estimer le paramètre binomial. Ces estimateurs sont : moyenne générale, Bayes empirique paramétrique avec la méthode des moments, Bayes empirique paramétrique avec la méthode du maximum de vraisemblance, Bayes empirique non paramétrique et James-Stein. Ensuite, on évalue ces estimateurs en se servant de la base de données 2005 de Brown (2008*b*) et on finit par l'analyse des résultats obtenus.

1.2 Modèle

Dans tout ce qui suit, on garde la même notation que celle de Brown (2008*a*). Soient H_{ji} et N_{ji} respectivement le nombre de coups sûrs et le nombre de présences au bâton du joueur i à l'intérieur de la période j. Il est naturel que $H_{ji} \sim Bin(N_{ji}\ ,\ p_i)$, où p_i est inconnu, $i = 1, 2, ...n_j$ et $j = 1, 2$ (dans le cas où la saison est divisée en deux périodes quelconques). Le rapport $R_{ji} = \frac{H_{ji}}{N_{ji}}$ est la moyenne au bâton du joueur i à l'intérieur de la période j.

Prenons le cas générique $R = \frac{H}{N}$. La variable aléatoire R suit approximativement la loi normale $N\Big(p, p(1-p)/N\Big)$. Puisque la variance de R dépend de p, inconnu, on doit chercher une transformation qui stabilise cette variance. On cherche cette transformation dans la famille des transformations suivante : $\left\{X^{(c)} = \arcsin\sqrt{\frac{H+c}{N+2c}} \quad \text{avec} \quad c \in \mathbb{R}\right\}$. Nous montrons en premier que $X^{(c)}$ suit approximativement une loi normale dont on déterminera la moyenne et la variance. D'après le théorème central limite, $\sqrt{N}(\frac{H}{N} - p) \overset{Loi}{\to} \sqrt{p(1-p)}\, Z$ où Z est la densité de la loi normale centrée réduite $N(0,1)$. Soit, maintenant, la fonction $g_1(t) = \frac{Nt+c}{N+2c}$. En appliquant la méthode Delta on obtient :

$$\sqrt{N}\Big(g_1(\frac{H}{N}) - g_1(p)\Big) \overset{Approx}{\to} \sqrt{p(1-p)}\,\frac{dg_1(t)}{dt}|_{t=p}\, Z \Longleftrightarrow$$
$$\sqrt{N}\Big(\frac{H+c}{N+2c} - \frac{Np+c}{N+2c}\Big) \overset{Approx}{\to} \sqrt{p(1-p)}\,\frac{N}{N+2c}\, Z.$$

En considérant la fonction $g_2(u) = \arcsin\sqrt{u}$ et en appliquant encore une fois la méthode Delta, on obtient :

$$\sqrt{N}\Big(g_2(\frac{H+c}{N+2c}) - g_2(\frac{Np+c}{N+2c})\Big) \overset{Approx}{\to} \sqrt{p(1-p)}\,\frac{N}{N+2c}\,\frac{dg_2(u)}{du}|_{u=\frac{Np+c}{N+2c}}\, Z \Longleftrightarrow$$
$$\sqrt{N}\left\{\arcsin\sqrt{\frac{H+c}{N+2c}} - \arcsin\sqrt{\frac{Np+c}{N+2c}}\right\} \overset{Approx}{\to} \frac{1}{2}\left\{\frac{p(1-p)}{(p+c/N)\big((1-p)+c/N\big)}\right\}^{1/2} Z.$$

Donc, $X^{(c)} = \arcsin\sqrt{\frac{H+c}{N+2c}} \overset{Approx}{\to} N\left\{\arcsin\sqrt{\frac{Np+c}{N+2c}}\,;\,\frac{1}{4N}\left[\frac{p(1-p)}{(p+c/N)\big((1-p)+c/N\big)}\right]\right\}$.

Remarque : si $c = 0$, on obtient : $X^{(0)} = \arcsin\sqrt{\frac{H}{N}} \overset{Approx}{\rightarrow} N\big(\arcsin\sqrt{p}\,; \frac{1}{4N}\big)$. Dans ce cas, la variance ne dépend plus de p.

Les développements en séries de puissance de la moyenne et de la variance de $X^{(c)}$ sont :

$$\begin{aligned} E(X^{(c)}) &= \arcsin\sqrt{p} + \frac{1-2p}{2N\sqrt{p(1-p)}}(c - 1/4) + O(N^{-2}); \\ Var(X^{(c)}) &= \frac{1}{4N} + O(N^{-2}). \end{aligned}$$

Pour le cas particulier $c = 1/4$, on a donc :

$$X^{(1/4)} = \arcsin\sqrt{\frac{H+1/4}{N+1/2}} \overset{Approx}{\rightarrow} N\left\{\arcsin\sqrt{p}\,; \frac{1}{4N}\right\}.$$

Parmi toutes les transformations $X^{(c)}$, la transformation $X^{(1/4)} = \arcsin\sqrt{\frac{H+1/4}{N+1/2}}$ offre, asymptotiquement, le meilleur compromis biais-variance (Figure 1 et Figure 2, Brown (2008*a*)). Donc, le choix de $c = 1/4$ est le plus adéquat pour ce genre de transformation. Le biais $\sin^2(E(X^{(c)}) - p$, pour le cas $c = 1/4$, est presque nul à partir de $N = 11$ (Figure 1, Brown (2008*a*)). Pour toutes ces raisons, dans la suite de l'étude, la prévision de la moyenne au bâton sera calculée sur la base de la transformation $X_i = \arcsin\sqrt{\frac{H_i+1/4}{N_i+1/2}}$ et ce calcul sera restreint sur les joueurs qui ont au moins 11 présences au bâton. Comme vu précédemment, les variables aléatoires X_i sont indépendantes et identiquement distribuées (i.i.d.) et $X_i \sim N(\theta_i\,, \sigma_i^2)$ approximativement, où $\sigma_i^2 = \frac{1}{4N_i}$ et $\theta_i = \arcsin\sqrt{p_i}$. Dans ce qui suit, on suppose que θ_i ne dépend pas de la période j ; autrement dit $\theta_{ji} = \theta_i$ pour $j = 1, 2$.

Soient, maintenant, S_1 et S_2 deux ensembles tels que S_1 est l'ensemble représentant la partie du début de la saison utilisée comme base pour la prévision et S_2 est l'ensemble représentant la partie du restant de la saison. Les joueurs appartenant à ces deux ensembles ont au moins 11 présences au bâton. Ceci est résumé comme suit :

$S_1 = \{i : N_{1i} \geq 11\}$ et $S_2 = \{i : N_{2i} \geq 11\}$. On définit aussi l'ensemble des joueurs en commun à S_1 et S_2 servant pour l'étape de la validation par $S_1 \cap S_2$. L'estimation des θ_i se fait sur l'ensemble $\{X_{1i}, N_{1i} : i \in S_1\}$. Quant à la validation des estimateurs qui consiste à comparer les estimations à leurs valeurs observées correspondantes X_{2i}, elle se fait sur l'ensemble des indices $i \in S_1 \cap S_2$. On se sert de la somme de l'erreur quadratique de prévision $SEQP$ comme estimateur de l'erreur de prévision et on choisit la méthode d'estimation ayant la plus petite $SEQP$. Notons par $\left\{\hat{\theta}_i : i \in S_1\right\}$ l'ensemble des estimateurs de $\{\theta_i : i \in S_1\}$. Cette estimation est basée sur l'ensemble $\{X_{1i} : i \in S_1\}$. Soit $\hat{\theta} \in \left\{\hat{\theta}_i : i \in S_1\right\}$ un estimateur quelconque. Alors, $SEQP(\hat{\theta}) = \sum_{i \in S_1 \cap S_2}(X_{2i} - \hat{\theta}_i)^2$. On se sert d'une façon indirecte de cette somme en passant par :

$$\begin{aligned} SEQP(\hat{\theta}) &= \sum_{i \in S_1 \cap S_2} \left\{(\hat{\theta}_i - X_{2i}) - (\theta_i - X_{2i})\right\}^2 \\ &= \sum_{i \in S_1 \cap S_2} (\hat{\theta}_i - X_{2i})^2 + \sum_{i \in S_1 \cap S_2} (\theta_i - X_{2i})^2 - 2 \sum_{i \in S_1 \cap S_2} (\hat{\theta}_i - X_{2i})(\theta_i - X_{2i}). \end{aligned}$$

Calculons maintenant l'espérance conditionnelle $E\Big(SEQP(\hat{\theta})|X_1\Big)$.

$E\Big(\sum_{i \in S_1 \cap S_2}(\hat{\theta}_i - X_{2i})(\theta_i - X_{2i})|X_1\Big) = 0$;

$E\Big(\sum_{i \in S_1 \cap S_2}(X_{2i} - \theta_i)^2|X_1\Big) = \sum_{i \in S_1 \cap S_2} \frac{1}{4N_{2i}}$;

$E\Big(\sum_{i \in S_1 \cap S_2}(\hat{\theta}_i - X_{2i})^2|X_1\Big) = \widehat{EEQT}(\hat{\theta})$, où $\widehat{EEQT}(\hat{\theta})$ est l'estimation de l'erreur quadratique totale. Donc, on a :

$$\widehat{EEQT}(\hat{\theta}) = SEQP(\hat{\theta}) - \sum_{i \in S_1 \cap S_2} \frac{1}{4N_{2i}}.$$

Soit, maintenant, $\hat{\theta}_t$ l'estimateur trivial défini par $\hat{\theta}_t(X) = X$. On utilise cet estimateur pour normaliser $\widehat{EEQT}(\hat{\theta})$. Cette normalisation s'obtient en divisant $\widehat{EEQT}(\hat{\theta})$ par $\widehat{EEQT}(\hat{\theta}_t)$ et sert à comparer les différents estimateurs entre eux. Notons cette

estimation de l'erreur quadratique totale normalisée par $\widehat{EEQT}^{(n)}(\hat{\theta})$.

$$\widehat{EEQT}^{(n)}(\hat{\theta}) = \frac{\widehat{EEQT}(\hat{\theta})}{\widehat{EEQT}(\hat{\theta}_t)}.$$

Précisons enfin que $\widehat{EEQT}(\hat{\theta}_t)$ se calcule aussi sur l'ensemble $S_1 \cap S_2$ et sa valeur normalisée est égale à 1.

1.3 Description des estimateurs

En premier, on décrit trois estimateurs : moyenne générale et Bayes empiriques dont les hyper-paramètres sont estimés une fois par la méthode des moments et une autre fois par la méthode du maximum de vraisemblance. Ensuite, on les applique pour le calcul des prévisions basées sur les 2 premiers mois et les 4 premiers mois de la saison. En dernier, on étudie les deux autres estimateurs, Bayes empirique non paramétrique et James-Stein, et on fait une comparaison entre tous ces estimateurs en utilisant l'estimation de l'erreur quadratique totale normalisée telle que mentionnée plus haut.

1.3.1 Moyenne générale

L'utilisation de cet estimateur ne prend pas en compte l'habileté au bâton intrinsèque de chaque joueur. Cet estimateur que l'on note par $\hat{\theta}_{mg}$ est défini comme suit : $\hat{\theta}_{mg} = \frac{\sum_{S_1} X_{1i}}{n_1}$, où n_1 est le cardinal de S_1.

1.3.2 Bayes empirique

On suppose que $\theta_i \sim N(\mu, \tau^2)$, où μ et τ^2 sont des hyper-paramètres. Si ces derniers étaient connus, on calculerait l'estimateur de Bayes que l'on note par θ_i^{bayes} en procédant comme suit : La distribution *a posteriori* de θ_i sachant l'observation x_{1i} est

$$\pi(\theta_i \backslash x_{1i}) \propto N(\theta_i\,,\,\sigma_{1i}^2) * N(\mu\,,\,\tau^2)$$

$$\propto \exp\left\{\frac{\tau^2+\sigma_{1i}^2}{2\tau^2\sigma_{1i}^2}\Big(\theta_i - \frac{x_{1i}\tau^2+\mu\sigma_{1i}^2}{\tau^2+\sigma_{1i}^2}\Big)^2\right\}.$$

Donc, $\theta_i \backslash x_{1i} \sim N\Big(\frac{x_{1i}\tau^2+\mu\sigma_{1i}^2}{\tau^2+\sigma_{1i}^2}\,,\,\frac{\tau^2\sigma_{1i}^2}{\tau^2+\sigma_{1i}^2}\Big)$, où $\sigma_{1i}^2 = \frac{1}{4N_{1i}}$.

L'estimateur de Bayes est $\theta_i^{bayes} = E(\theta_i \backslash x_i) = \frac{x_{1i}\tau^2+\mu\sigma_{1i}^2}{\tau^2+\sigma_{1i}^2}$. On écrit l'expression de cet estimateur de la façon suivante :

$$\theta_i^{bayes} = \mu + \frac{\tau^2}{\tau^2+\sigma_{1i}^2}(X_{1i} - \mu).$$

Si, maintenant, μ et τ^2 sont inconnus, alors on doit les estimer pour calculer un estimateur de Bayes empirique. Tout d'abord, on détermine la distribution marginale de X_{1i} que l'on note par $m(x_{1i})$ pour s'en servir dans les calculs ci-après.

$$\begin{aligned} m(x_{1i}) &= \int_{-\infty}^{+\infty} N(\theta_i\,,\,\sigma_{1i}^2) * N(\mu\,,\,\tau^2) d\theta_i \\ &= \frac{1}{\sqrt{2\pi(\tau^2+\sigma_{1i}^2)}} \exp\left\{\frac{(x_{1i}-\mu)^2}{2(\tau^2+\sigma_{1i}^2)}\right\}. \end{aligned}$$

Donc, $X_{1i} \sim N(\mu\,,\,\tau^2+\sigma_{1i}^2)$, où $\sigma_{1i}^2 = \frac{1}{4N_{1i}}$.
Le principe de détermination de l'estimateur de Bayes empirique est d'estimer μ et τ^2 en utilisant $\{X_{1i}\}$ et de substituer ces estimations dans l'expression de θ_i^{bayes}. On utilise deux méthodes pour estimer μ et τ^2, la méthode des moments et la méthode du maximum de vraisemblance.

1.3.2.1 Méthode des moments

Cette méthode nécessite une résolution itérative du système de deux équations en μ

et τ^2 suivant :

$$\begin{cases} \mu &= \frac{\sum_{i\in S_1} \frac{X_{1i}}{\tau^2+\sigma^2_{1i}}}{\sum_{i\in S_1} \frac{1}{\tau^2+\sigma^2_{1i}}} \\ \\ \tau^2 &= \left[\frac{\sum_{i\in S_1}(X_{1i}-\mu)^2}{n_1-1} - \frac{\sum_{i\in S_1}\sigma^2_{1i}}{n_1}\right]_+ \end{cases}$$

où $[.]_+$ est la partie positive.

En pratique, avec les données que nous avons, une seule itération suffit pour donner une solution aussi exacte que celle donnée par la convergence. Cette itération consiste à remplacer μ par $\hat{\theta}_{mg} = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}$ dans l'expression de τ^2 pour en tirer l'estimation $\tilde{\tau}^2$. Pour en déduire l'estimation $\tilde{\mu}$, il suffit de remplacer τ^2 par $\tilde{\tau}^2$ dans l'expression de μ. Enfin, en remplaçant μ et τ^2 par $\tilde{\mu}$ et $\tilde{\tau}^2$ respectivement dans l'expression de θ_i^{bayes}, on obtient l'estimateur de Bayes empirique avec la méthode des moments que l'on note par $\hat{\theta}_{bemm}$.

1.3.2.2 Méthode du maximum de vraisemblance

Soit $L(\mu, \tau^2 \backslash X_1)$ la fonction de vraisemblance.

$$\begin{aligned} L(\mu,\tau^2\backslash X_1) &= \frac{(2\pi)^{-n_1/2}}{\prod_{i\in S_1}(\tau^2+\sigma^2_{1i})^{1/2}} \exp\left\{-\frac{1}{2}\sum_{i\in S_1}\frac{(x_{1i}-\mu)^2}{\tau^2+\sigma^2_{1i}}\right\}. \\ \log\Big(L(\mu,\tau^2\backslash X_1)\Big) &= -\frac{n_1}{2}log(2\pi) - \frac{1}{2}\sum_{i\in S_1}\log(\tau^2+\sigma^2_{1i}) - \frac{1}{2}\sum_{i\in S_1}\frac{(x_{1i}-\mu)^2}{\tau^2+\sigma^2_{1i}}. \\ \frac{\partial \log\Big(L(\mu,\tau^2\backslash X_1)\Big)}{\partial\mu} &= \sum_{i\in S_1}\frac{(x_{1i}-\mu)}{\tau^2+\sigma^2_{1i}}. \\ \frac{\partial \log\Big(L(\mu,\tau^2\backslash X_1)\Big)}{\partial(\tau^2+\sigma^2_{1i})} &= -\frac{1}{2}\sum_{i\in S_1}\frac{1}{\tau^2+\sigma^2_{1i}} + \frac{1}{2}\sum_{i\in S_1}\frac{(x_{1i}-\mu)^2}{(\tau^2+\sigma_{1i})^2}. \end{aligned}$$

Les estimateurs du maximum de vraisemblance $\hat{\mu}$ et $\hat{\tau}^2$ sont la solution du système de deux équations suivant :

$$\left\{ \begin{array}{rcl} \sum_{i\in S_1} \frac{(x_{1i}-\mu)}{\tau^2+\sigma_{1i}^2} & = & 0 \\ -\frac{1}{2}\sum_{i\in S_1} \frac{1}{\tau^2+\sigma_{1i}^2} + \frac{1}{2}\sum_{i\in S_1} \frac{(x_{1i}-\mu)^2}{(\tau^2+\sigma_{1i}^2)^2} & = & 0 \end{array} \right\} \Longleftrightarrow$$

$$\left\{ \begin{array}{rcl} \mu & = & \frac{\sum_{i\in S_1} X_{1i}/(\tau^2+\sigma_{1i}^2)}{\sum_{i\in S_1} 1/(\tau^2+\sigma_{1i}^2)} \\ \sum_{i\in S_1} \frac{(x_{1i}-\mu)^2}{(\tau^2+\sigma_{1i}^2)^2} & = & \sum_{i\in S_1} \frac{1}{\tau^2+\sigma_{1i}^2} \end{array} \right\}.$$

En remplaçant μ et τ^2 par $\hat{\mu}$ et $\hat{\tau}^2$ respectivement dans l'expression de θ_i^{bayes}, on obtient l'estimateur de Bayes empirique avec la méthode du maximum de vraisemblance que l'on note par $\hat{\theta}_{bemv}$.

1.3.2.3 Bayes empirique non paramétrique

Dans le cas des estimateurs de Bayes empirique avec la méthode des moments et la méthode du maximum de vraisemblance, on avait l'hypothèse $\theta_i \sim N(\mu\,,\,\tau)$. Par contre, dans le cas du non paramétrique, on suppose que $\theta_i \sim G$, où G est une fonction de répartition inconnue et les θ_i sont indépendantes. Tout d'abord, on détermine l'estimateur de Bayes de θ_i que l'on note par $(\theta^G)_i$ pour garder la même notation que Brown (2008*a*). On sait que $X_{1i}\backslash\theta_i \sim N(\theta_i\,,\,\sigma_i^2)$, où $\sigma_i^2 = 1/4N_{1i}$. Donc, la distribution marginale de X_{1i} est :

$$\begin{aligned} g_i(x_{1i}) &= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}\left(\frac{x_{1i}-\theta_i}{\sigma_i^2}\right)^2} dG(\theta_i) \\ &= \int \phi\Big(\frac{x_{1i}-\theta_i}{\sigma_i}\Big) dG(\theta_i)\,, \end{aligned}$$

où ϕ est la densité de la loi normale centrée réduite $N(0,1)$.

Soient, maintenant, δ une règle de décision et $L\big(\theta_i\,,\,\delta(X_{1i})\big) = \big(\theta_i - \delta(X_{1i})\big)^2$, où L est la fonction de perte quadratique associée à la règle de décision δ lorsque le paramètre est θ_i. L'estimateur de Bayes $(\theta^G)_i$ de θ_i est la règle de décision qui minimise $E\big\{L\big(\theta_i\,,\,\delta(X_{1i})\big)\big\}$, où E est l'espérance inconditionnelle.

On note l'espérance conditionnelle sachant X_{1i}, qui représente l'espérance *a posteriori*, par $E^{X_{1i}}$. Cette espérance est définie comme suit :

$$\begin{aligned} E^{X_{1i}} h(X_{1i}, \theta_i) &= \frac{\int h(X_{1i}, \theta_i) e^{-\frac{1}{2\sigma_i^2}(X_{1i}-\theta_i)^2} dG(\theta_i)}{\int e^{-\frac{1}{2\sigma_i^2}(X_{1i}-\theta_i)^2} dG(\theta_i)} \\ &= \frac{\int h(X_{1i}, \theta_i) \phi\left(\frac{X_{1i}-\theta_i}{\sigma_i}\right) dG(\theta_i)}{\int \phi\left(\frac{X_{1i}-\theta_i}{\sigma_i}\right) dG(\theta_i)}, \end{aligned}$$

où h est une fonction et ϕ est la densité de la loi normale centrée réduite $N(0,1)$.

Déterminons l'estimateur de Bayes $(\theta^G)_i$ de θ_i. Cet estimateur est la règle qui minimise

$$E\left(\theta_i - \delta(X_{1i})\right)^2 = E E^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2 = E\left\{\frac{\int h\left(X_{1i}, \theta_i\right) \phi\left(\frac{X_{1i}-\theta_i}{\sigma_i}\right) dG(\theta_i)}{\int \phi\left(\frac{X_{1i}-\theta_i}{\sigma_i}\right) dG(\theta_i)}\right\}.$$

Mais, minimiser $E\left(\theta_i - \delta(X_{1i})\right)^2$ revient à minimiser $E^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2$.

Minimisons donc $E^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2$.

$E^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2 = E^{X_{1i}}(\theta_i^2) - 2\delta(X_{1i}) E^{X_{1i}}(\theta_i) + \delta^2(X_{1i})$

$\frac{dE^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2}{d\delta(X_{1i})} = -2E^{X_{1i}}(\theta_i) + 2\delta(X_{1i})$

$\frac{dE^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2}{d\delta(X_{1i})} = 0 \Longleftrightarrow \delta^*(X_{1i}) = E^{X_{1i}}(\theta_i)$ (la solution).

$E^{X_{1i}}\left(\theta_i - E^{X_{1i}}(\theta_i)\right)^2$ est bien un minimum, car $\frac{d^2 E^{X_{1i}}\left(\theta_i - \delta(X_{1i})\right)^2}{d\left(\delta(X_{1i})\right)^2} = 2 > 0.$

Donc, l'estimateur de Bayes de θ_i est :

$$\begin{aligned} (\theta^G)_i &= E^{X_{1i}}(\theta_i) \\ &= E^{X_{1i}}\left(\theta_i - X_{1i} + X_{1i}\right) \\ &= X_{1i} + E^{X_{1i}}\left(\theta_i - X_{1i}\right) \\ &= X_{1i} + \frac{\int (\theta_i - X_{1i}) e^{-\frac{1}{2\sigma_i^2}(X_{1i}-\theta_i)^2} dG(\theta_i)}{\int e^{-\frac{1}{2\sigma_i^2}(X_{1i}-\theta_i)^2} dG(\theta_i)} \\ &= X_{1i} + \sigma_{1i}^2 \frac{\partial \log\left(g_i(X_{1i})\right)}{\partial X_{1i}}. \end{aligned}$$

Voir Stein (1981) pour plus de détails.

Pour la détermination de l'estimateur de Bayes empirique, Brown (2008*a*) propose la méthode suivante : il part de la formule de l'estimateur de Bayes $\left(\theta^G(X_1)\right)_i = X_{1i} + \sigma_{1i}^2 \frac{\partial \log\left(g_i(X_1)\right)}{\partial X_{1i}}$. Ensuite, il estime g_i par l'estimation à noyau et finit par la substitution de cette estimation, notée $\tilde{g}_i$, dans l'expression de $\left(\theta^G(X_1)\right)_i$ pour en déduire l'estimateur de Bayes empirique non paramétrique, noté $\hat{\theta}_{benp}$. Avec un paramètre de lissage constant h de l'estimateur à noyau utilisé, on a :

$$\begin{aligned}\tilde{g}_i(X_1) &= \left\{\sum_k \frac{I_{\{k:(1+h)\sigma_{1i}^2-\sigma_{1k}^2>0\}}(k)}{\sqrt{(1+h)*\max(\sigma_{1i}^2;\sigma_{1k}^2)-\sigma_{1k}^2}} * \phi\left(\frac{(X_{1i}-X_{1k})}{\sqrt{(1+h)*\max(\sigma_{1i}^2;\sigma_{1k}^2)-\sigma_{1k}^2}}\right)\right\} \\ &* \left\{\sum_k I_{\{k:(1+h)\sigma_{1i}^2-\sigma_{1k}^2>0\}}(k)\right\}^{-1},\end{aligned}$$

où ϕ et I sont respectivement la densité de la loi normale centrée réduite $N(0,1)$ et la fonction indicatrice.

Finalement, l'estimateur de Bayes empirique non paramétrique est :

$$(\hat{\theta}_{benp})_i = X_{1i} + \sigma_{1i}^2 \frac{\partial \tilde{g}_i(X_1)/\partial X_{1i}}{\tilde{g}_i(X_1)}.$$

Pour les applications, Brown (2008*a*) propose l'utilisation de $h \approx 1/log(n_1)$, où $n_1 =$ cardinal (S_1). Mais, dans ses applications, il utilise $h = 0.25$ pour cardinal $(S_1) > 200$ et $h = 0.30$ pour cardinal $(S_1) = 81$. Pour pouvoir faire une comparaison avec les résultats de Brown (2008*a*), on utilise le même h que lui et $h \approx 1/log(n_1)$.

1.3.3 James-Stein

L'estimateur de James-Stein que l'on utilise pour nos applications ci-dessous et que l'on note par $\hat{\theta}_{ejs}$ est :

$$(\hat{\theta}_{ejs})_i = \hat{\mu}_1 + \left\{1 - \frac{n_1 - 3}{\sum_{S_1}(X_{1i}-\hat{\mu}_1)^2/\sigma_{1i}^2}\right\}_+ (X_{1i}-\hat{\mu}_1),$$

où $\hat{\mu}_1 = \frac{\sum_{S_1} X_{1i}/\sigma_{1i}^2}{\sum_{S_1} 1/\sigma_{1i}^2}$ et $\{.\}_+$ est la partie positive.

Tableau 1.4.3 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs. La prévision est basée sur les 2 premiers mois de la saison et h est le paramètre de lissage ($n_1 = 518$ si Tous et $n_1 = 444$ si Non-lanceurs)

	$\widehat{EEQT}^{(n)}$; Tous	$\widehat{EEQT}^{(n)}$; Non-lanceurs
Cardinal(S_1) : pour estimation	518	444
Cardinal($S_1 \cap S_2$) : pour validation	474	410
$\hat{\theta}_t$: Trivial	1	1
$\hat{\theta}_{mg}$: Moy.générale	0.544	0.130
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.403	0.119
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	0.546	0.113
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h = 0.25$	0.320	0.090
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h \approx 1/log(n_1)$	0.379	0.087
$\hat{\theta}_{ejs}$: James-Stein	0.330	0.099

1.4 Application

Dans cette section, on calcule l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ des six estimateurs décrits dans la section précédente (Section 1.3) pour faire une comparaison entre leurs performances.

1.4.1 Prévision basée sur les 2 premiers mois de la saison

Dans cette situation, S_1 et $S_1 \cap S_2$ contiennent peu de lanceurs (74 lanceurs pour S_1 et 64 pour $S_1 \cap S_2$). Alors, il est plus approprié d'exclure les lanceurs de l'analyse. Le tableau 1.4.3 ci-dessus présente l'estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ des six estimateurs. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs. (Prog. A.1, Annexe A)

1.4.2 Prévision basée sur les 4 premiers mois de la saison

Dans cette situation, S_1 contient 93 lanceurs et 509 non-lanceurs mais $S_1 \cap S_2$ ne contient que peu de lanceurs (63 lanceurs et 427 non-lanceurs). Alors, pour la même raison que dans le cas où la prévision est basée sur les 2 premiers mois de la saison, on exclut les lanceurs de l'analyse. Le tableau 1.4.4 ci-dessous présente l'estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ des six estimateurs. Aussi, dans ce cas, l'analyse est faite sur tous les frappeurs et sur les non-lanceurs.

Tableau 1.4.4 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs. La prévision est basée sur les 4 premiers mois de la saison et h est le paramètre de lissage ($n_1 = 602$ si Tous et $n_1 = 509$ si Non-lanceurs)

	$\widehat{EEQT}^{(n)}$; Tous	$\widehat{EEQT}^{(n)}$; Non-lanceurs
Cardinal(S_1) : pour estimation	602	509
Cardinal($S_1 \cap S_2$) : pour validation	490	427
$\hat{\theta}_t$: Trivial	1	1
$\hat{\theta}_{mg}$: Moy.générale	1.067	0.499
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.757	0.494
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	1.062	0.535
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h = 0.25$	0.660	0.498
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h \approx 1/log(n_1)$	0.889	0.550
$\hat{\theta}_{ejs}$: James-Stein	0.735	0.486

Maintenant, on choisit 12 frappeurs parmi les non-lanceurs de la façon suivante :

- Les 4 frappeurs ayant les moyennes au bâton les plus élevées ;
- Les 4 frappeurs ayant les moyennes au bâton les plus basses ;
- 4 frappeurs ayant une performance moyenne au bâton.

Voici ces 12 frappeurs :

```
Moyennes les plus élevées          Moyennes les plus basses          Performance moyenne
First.Name  Last.Name  MB          First.Name  Last.Name  MB         First.Name  Last.Name  MB
-----------------------------      -----------------------------    -----------------------------
Rene        Rivera     0.364       Scott       Spiezio    0.065      Joe         Crede      0.256
Rainer      Olmedo     0.409       Brayan      Pena       0.130      Daryle      Ward       0.262
Jeff        Francoeur  0.413       Ross        Gload      0.130      Rob         Mackowiak  0.269
Matthew     Murton     0.441       Miguel      Ojeda      0.137      Scott       Hatteberg  0.280
```

Dans ce cas, on a $S_1 = S_1 \cap S_2$ et chacun des deux ensembles contient les 12 frappeurs Le tableau 1.4.5 ci-dessous présente l'estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ des six estimateurs. L'analyse est faite sur les 12 frappeurs.

1.4.3 Résultats de Brown

À titre de comparaison, le tableau 1.4.6, le tableau 1.4.7 et le tableau 1.4.8 ci-dessous présentent les résultats de Brown (2008a).

Tableau 1.4.5 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est faite sur 12 frappeurs. La prévision est basée sur les 4 premiers mois de la saison et h est le paramètre de lissage $(n_1 = 12)$

	$\widehat{EEQT}^{(n)}$; Non-lanceurs
Cardinal(S_1) : pour estimation	12
Cardinal($S_1 \cap S_2$) : pour validation	12
$\hat{\theta}_t$: Trivial	1
$\hat{\theta}_{mg}$: Moy.générale	0.568
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.318
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	0.196
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h \approx 1/log(n_1)$	0.592
$\hat{\theta}_{ejs}$: James-Stein	0.462

Tableau 1.4.6 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs. La prévision est basée sur la première moitié de la saison et h est le paramètre de lissage

	$\widehat{EEQT}^{(n)}$; Tous	$\widehat{EEQT}^{(n)}$; Non-lanceurs
Cardinal(S_1) : pour estimation	567	486
Cardinal($S_1 \cap S_2$) : pour validation	499	435
$\hat{\theta}_t$: Trivial	1	1
$\hat{\theta}_{mg}$: Moy.générale	0.853	0.378
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.588	0.359
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	0.887	0.398
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h = 0.25$	0.485	0.358
$\hat{\theta}_{ejs}$: James-Stein	0.535	0.348

Tableau 1.4.7 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est faite sur tous les frappeurs. La prévision est basée sur le premier mois de la saison et h est le paramètre de lissage

	$\widehat{EEQT}^{(n)}$; Tous
Cardinal(S_1) : pour estimation	421
Cardinal($S_1 \cap S_2$) : pour validation	409
$\hat{\theta}_t$: Trivial	1
$\hat{\theta}_{mg}$: Moy.générale	0.250
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.240
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	-
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h = 0.25$	0.169
$\hat{\theta}_{ejs}$: James-Stein	0.218

Tableau 1.4.8 Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$. L'analyse est restreinte sur les non-lanceurs. La prévision est basée sur les 5 premiers mois de la saison et h est le paramètre de lissage

	$\widehat{EEQT}^{(n)}$; Non-lanceurs
Cardinal(S_1) : pour estimation	634
Cardinal($S_1 \cap S_2$) : pour validation	448
$\hat{\theta}_t$: Trivial	1
$\hat{\theta}_{mg}$: Moy.générale	0.955
$\hat{\theta}_{bemm}$: Bayes empirique (méth.moments)	0.904
$\hat{\theta}_{bemv}$: Bayes empirique (méth.max.vrais)	-
$\hat{\theta}_{benp}$: Bayes empirique non paramétrique : $h = 0.25$	0.944
$\hat{\theta}_{ejs}$: James-Stein	0.808

1.5 Analyse des résultats

Que la période de prévision soit réservée à tous les frappeurs ou seulement aux non-lanceurs ayant au moins 11 présences au bâton, l'estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ de chacun des estimateurs est plus petite, à part celle de l'estimateur trivial qui est constante $(= 1)$, lorsque cette période de prévision est moins large. Maintenant, si on compare les estimations des erreurs quadratiques totales normalisées de ces estimateurs entre elles à l'intérieur de chaque période de prévision, les performances de ces estimateurs sont classées dans l'ordre suivant : Bayes empirique non paramétrique $\hat{\theta}_{benp}$, James-Stein $\hat{\theta}_{ejs}$ et Bayes empirique avec la méthode des moments $\hat{\theta}_{bemm}$. Quant à la performance de Bayes empirique avec la méthode du maximum de vraisemblance $\hat{\theta}_{bemv}$, elle reste médiocre par rapport à celles des autres estimateurs, et ceci pour n'importe quelle période de prévision utilisée dans l'analyse. La faible performance de l'estimateur de Bayes empirique avec la méthode du maximum de vraisemblance $\hat{\theta}_{bemv}$ s'explique par deux raisons :

- À chaque période de prévision, en prenant pour analyse tous les frappeurs ou seulement les non-lanceurs ayant au moins 11 présences au bâton, la figure 1.1, la figure 1.2 et la figure 1.3 ci-dessous montrent clairement que les histogrammes des valeurs de X_{1i} qui s'y trouvent ne sont pas bien ajustés à des distributions normales. Bien évidemment, ce manque d'ajustement des valeurs de X_{1i} à une distribution normale concerne aussi les estimateurs de Bayes empirique avec la

méthode des moments $\hat{\theta}_{bemv}$ et James-Stein $\hat{\theta}_{ejs}$, mais leurs performances sont moins affectées que celle de l'estimateur de Bayes empirique avec la méthode du maximum de vraisemblance $\hat{\theta}_{bemv}$. Le seul estimateur qui n'est pas affecté par ce manque d'ajustement est celui de Bayes empirique non paramétrique $\hat{\theta}_{benp}$, et ceci justifie sa meilleure performance par rapport aux autres estimateurs.

- La corrélation est assez élevée entre les valeurs de N_{1i} et les valeurs de X_{1i}, et ceci pour n'importe quelle période de prévision et pour n'importe quel groupe de frappeurs pris pour l'analyse. Bien que cette corrélation affecte tous les estimateurs de Bayes empiriques, il semble que celui de Bayes empirique avec la méthode du maximum de vraisemblance $\hat{\theta}_{bemv}$ est le plus affecté. La figure 1.1, la figure 1.2 et la figure 1.3 ci-dessous illustrent cette corrélation.

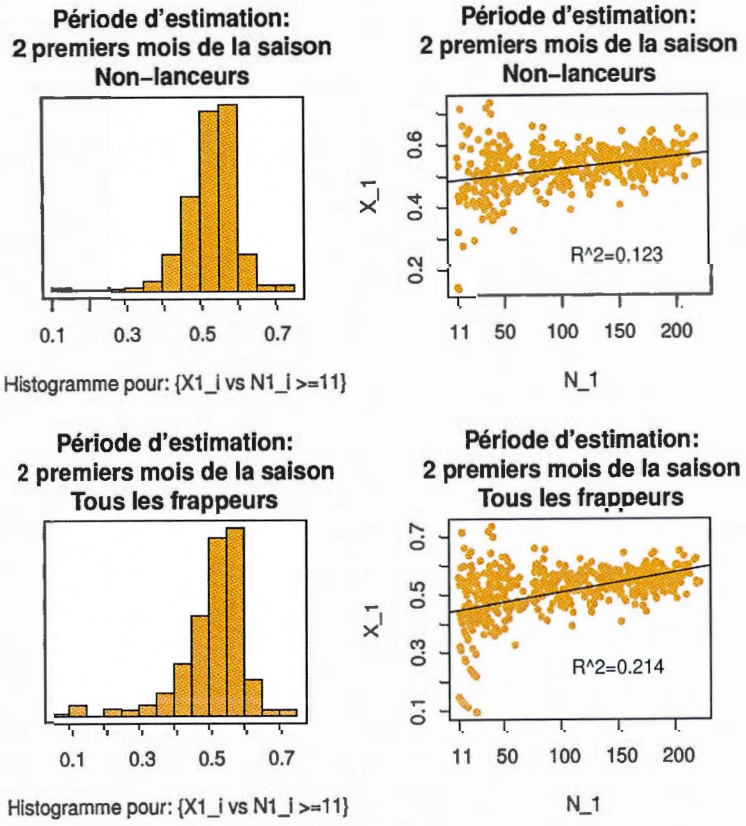

Figure 1.1 Histogrammes pour $\{X_{1i} : N_{1i} \geq 11\}$ et nuage de points pour $X_1 vs\, N_1$ pour tous les frappeurs et les non-lanceurs ayant au moins 11 présences au bâton. La période de prévision est basée sur les 2 premiers mois de la saison et R^2 est égal à 0.123 pour les non-lanceurs et à 0.214 pour tous les frappeurs

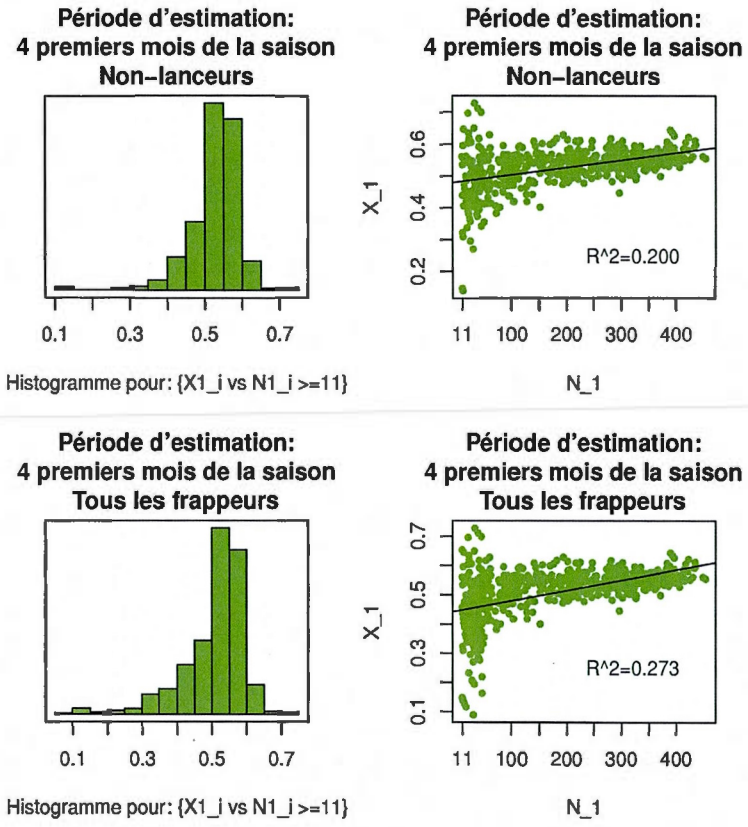

Figure 1.2 Histogrammes pour $\{X_{1i} : N_{1i} \geq 11\}$ et nuage de points pour X_1 *vs* N_1 pour tous les frappeurs et les non-lanceurs ayant au moins 11 présences au bâton. La période de prévision est basée sur les 4 premiers mois de la saison et R^2 est égal à 0.200 pour les non-lanceurs et à 0.273 pour tous les frappeurs

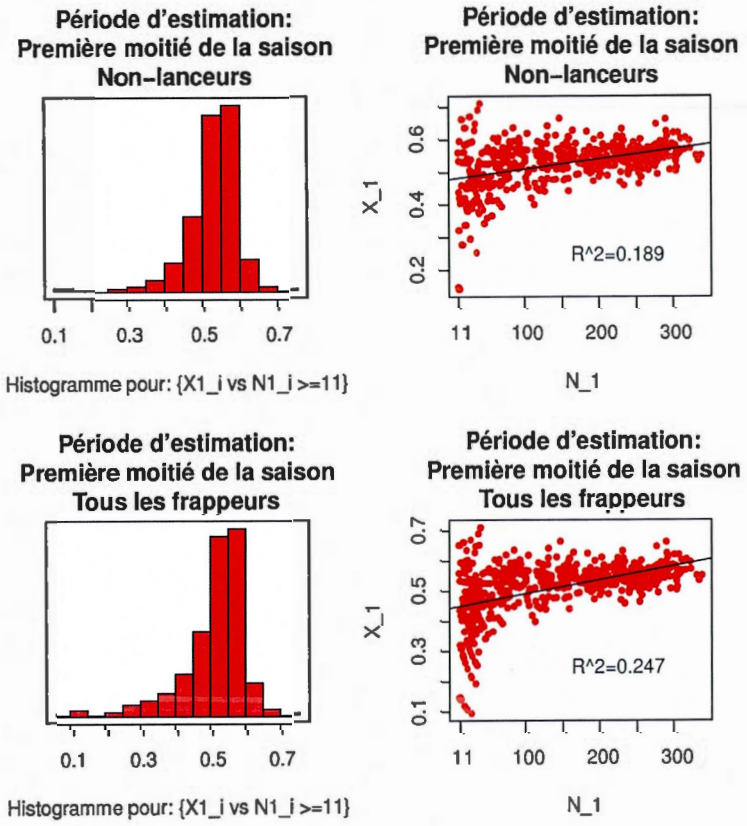

Figure 1.3 Histogrammes pour $\{X_{1i} : N_{1i} \geq 11\}$ et nuage de points pour X_1 *vs* N_1 pour tous les frappeurs et les non-lanceurs ayant au moins 11 présences au bâton. La période de prévision est basée sur la première moitié de la saison et R^2 est égal à 0.189 pour les non-lanceurs et à 0.247 pour tous les frappeurs

CHAPITRE II

ESTIMATION PAR INTERVALLE ET PRÉVISION DU PARAMÈTRE BINOMIAL

2.1 Introduction

L'intervalle de confiance du paramètre binomial p basé sur l'approximation de la loi binomiale $B(n\,,p)$ par la loi normale $N(np\,,\,np(1-p))$ est :

$$\hat{p} - z_{\alpha/2}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}\ ,$$

où $\hat{p} = X/n$ est la proportion échantillonnale des succès et $z_{\alpha/2}$ est le $100\,(1-\alpha/2)^{\text{e}}$ percentile de la loi normale centrée réduite $N(0\,,\,1)$. Cet intervalle, connu sous le nom de l'intervalle standard ou de Wald, est reconnu pour avoir une faible probabilité de couverture lorsque p est proche des frontières 0 et 1. Pour remédier à ce problème, l'application de cet intervalle est accompagnée de conditions telles que $\min\{np(1-p)\} \geq 5$ ou 10. Mais il est démontré, par de récents articles, que la probabilité de couverture de cet intervalle, qui dépend de n et p, a un comportement chaotique et imprévisible lorsque n et p varient. Cette probabilité de couverture peut être faible, même pour n assez grand et p loin des frontières 0 et 1. Pour avoir une meilleure performance que celle de cet intervalle, d'autres intervalles sont suggérés. Il s'agit des intervalles : de Wilson ICw, d'Agresti-Coull $ICac$, du maximum de vraisemblance $ICrv$ et de Jeffreys bilatéral ICj.

En suivant l'approche de Brown, Cai et DasGupta (1999), on commence ce chapitre par un développement d'Edgeworth d'ordres 1 et 2 de la probabilité de couverture et un développement d'Edgeworth d'ordre 2 de la longueur moyenne de l'intervalle

standard ICs et des intervalles suggérés afin de les approximer. Ensuite, on fait une comparaison entre les probabilités de couverture et entre les longueurs moyennes approximées obtenues. Enfin, on applique la théorie développée sur la base de données 2005 de Brown (2008*b*) et on analyse les résultats obtenus.

2.2 Théorie et méthode

Dans cette section, on commence par le calcul approximatif, avec un n modéré, de la déviation du biais, de la variance et des coefficients d'asymétrie et d'applatissement de $W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{\text{loi}} N(0,1)$ et la comparaison des valeurs obtenues aux valeurs asymptotiques correspondantes : 0, 1, 0 et 3. Ensuite, on aborde le côté théorique du développement d'Edgeworth de la probabilité de couverture et de la longueur moyenne des intervalles de confiance : ICs, ICw, $ICac$, $ICrv$ et ICj.

2.2.1 Déviation du biais, de la variance et des coefficients d'asymétrie et d'applatissement de $W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{\text{loi}} N(0,1)$

L'intervalle de confiance standard ICs est basé sur le fait que

$$W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}} \xrightarrow{\text{loi}} N(0,1).$$

Cependant, la distribution de W_n n'est pas normale, même pour n assez grand. Asymptotiquement, W_n a un biais 0, une variance 1, un coefficient d'asymétrie 0 et un coefficient d'applatissement 3. Par contre, avec un n modéré, les déviations de ces derniers paramètres par rapport à leurs valeurs asymptotiques correspondantes sont souvent significatives. En particulier, ces déviations causent un biais négatif de la probabilité de couverture de l'intervalle de confiance standard ICs.

Calculons maintenant le biais de W_n en utilisant les développements limités. Soit $Z_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{pq}}$. En remplaçant $\hat{p}$ par $p + Z_n\sqrt{pq/n}$ dans l'expression de W_n, on

obtient $W_n = \frac{Z_n}{\sqrt{1+(1-2p)Z_n/\sqrt{npq}-Z_n^2/n}}$. Afin d'approximer le biais de W_n, utilisons un développement limité de Taylor d'ordre 3 de $\sqrt{1+(1-2p)Z_n/\sqrt{npq}-Z_n^2/n}$. Posons $\gamma = \frac{1-2p}{\sqrt{pq}}$ et $\kappa = \frac{1-6pq}{\sqrt{pq}}$, respectivement le coefficient d'asymétrie et le coefficient d'applatissement non normalisé d'une variable aléatoire de Bernoulli de paramètre p. On a :

$$\begin{aligned} W_n &= Z_n\left\{1+n^{-1/2}\gamma Z_n - n^{-1}\,Z_n^2\right\}^{-1/2} \\ &= Z_n\Big\{1-n^{-1/2}\left(\tfrac{\gamma}{2}\right)Z_n + n^{-1}\left(\tfrac{1}{2}+\tfrac{3}{8}\gamma^2\right)Z_n^2 - n^{-3/2}\left(\tfrac{3}{4}+\tfrac{5}{16}\gamma^3\right)Z_n^3 + O(n^{-3/2})\Big\}. \end{aligned}$$

Le biais est $E(W_n)$.

Calculons en premier les espérances de : Z_n, Z_n^2, Z_n^3, et Z_n^4. On obtient :

$E(Z_n) = 0$;

$E(Z_n^2) = 1$;

$\mathrm{E}(\mathrm{Z}_n^3) = n^{-1/2}\gamma$;

$E(Z_n^4) = 3 + n^{-1}\kappa$.

En remplaçant ces espérances dans l'expression de $E(W_n)$ on obtient :

$$\begin{aligned} E(W_n) &= -n^{-1/2}\left(\frac{\gamma}{2}\right) - n^{-3/2}\left(\frac{7}{4}\gamma + \frac{9}{16}\gamma^2\right) + O(n^{-3/2}) \\ &= \frac{p-\frac{1}{2}}{\sqrt{npq}}\left(1+\frac{7}{2n}+\frac{9(p-\frac{1}{2})^2}{2npq}\right) + O(n^{-3/2}). \end{aligned}$$

On voit d'après l'expression de $E(W_n)$ que le biais de W_n est négatif pour $p < 0.5$ et positif pour $p > 0.5$.

Calculons maintenant les coefficients d'asymétrie et d'applatissement de W_n. Tout d'abord, on calcule les espérances de : W_n, W_n^2, W_n^3 et W_n^4.

$$\begin{aligned} E(W_n) &= E\left(Z_n\left[1+n^{-1/2}\left(\tfrac{\gamma}{2}\right)Z_n\right]\right) + O(n^{-1}) \\ &= -\tfrac{1}{2}\,n^{-1/2}\gamma + O(n^{-1}). \end{aligned}$$

$$\begin{aligned} E(W_n^2) &= E\left(Z_n^2\left[1 - n^{-1/2}\gamma Z_n + n^{-1}(1+\gamma^2)Z_n^2\right]\right) + O(n^{-3/2}) \\ &= 1 - n^{-1}\gamma^2 + 3n^{-1}(1+\gamma^2) + O(n^{-3/2}) \\ &= 1 + n^{-1}(2\gamma^2 + 3) + O(n^{-3/2}). \end{aligned}$$

$$\begin{aligned} E(W_n^3) &= E\left(Z_n^3\left[1 - n^{-1/2}(\tfrac{3\gamma}{2})Z_n\right]\right) + O(n^{-1}) \\ &= n^{-1/2}\gamma - 3n^{-1/2}(\tfrac{3\gamma}{2}) + O(n^{-1}) \\ &= -\tfrac{7}{2}n^{-1/2}\gamma + O(n^{-1}). \end{aligned}$$

Le calcul de $E(W_n^4)$ dépend de $E(Z_n^5)$ et de $E(Z_n^6)$. D'une manière générale $E(Z_n^j) = \frac{\mu_j}{(npq)^{j/2}}$ avec $j \geq 2$ et μ_j est le moment centré d'ordre j d'une variable aléatoire suivant une loi binomiale de paramètre p. Pour $j \geq 2$, μ_j est calculé à partir de la formule de récurrence de Romanovski ; voir Johnson, Kotz et Balakrishnan (1995). Cette formule est :

$$\mu_{j+1} \;=\; pq\big(nj\mu_{j-1} + \frac{d\mu_j}{dp}\big).$$

Le calcul donne :

$$E(Z_n^5) = n^{-1/2}(10\gamma) + n^{-3/2}\gamma(\kappa - 6).$$
$$E(Z_n^6) = 15 + n^{-1}(20\gamma^2 + 5\kappa - 20) + n^{-2}\left(\gamma^2(\kappa - 18) - 2\kappa + 12\right).$$

Calculons maintenant $E(W_n^4)$.

$$\begin{aligned} E(W_n^4) &= E\left(Z_n^4\left[1 - n^{-1/2}(2\gamma)Z_n + n^{-1}(2+3\gamma^2)Z_n^2\right]\right) + O(n^{-3/2}) \\ &= 3 + n^{-1}\kappa - (2n^{-1/2}\gamma)(10n^{-1/2}\gamma) + 15n^{-1}(2+3\gamma^2) + O(n^{-3/2}) \\ &= 3 + n^{-1}(25\gamma^2 + \kappa + 30) + O(n^{-3/2}). \end{aligned}$$

Enfin, on calcule : $var(W_n)$, $E(W_n - EW_n)^3$ et $E(W_n - EW_n)^4$. On obtient :
$var(W_n) = E(W_n^2) - (EW_n)^2$

$$= 1 + n^{-1}(2\gamma^2 + 3) - \tfrac{1}{4}(-n^{-1/2}\gamma)^2 + O(n^{-2})$$
$$= 1 + n^{-1}(\tfrac{7}{4}\gamma^2 + 3) + O(n^{-2}).$$

$$\begin{aligned} E(W_n - EW_n)^3 &= E(W_n^3) - 3E(W_n)E(W_n^2) + 2(EW_n)^3 \\ &= \tfrac{-7}{2}n^{-1/2}\gamma - 3(\tfrac{-1}{2}n^{-1/2}\gamma)(1+n^{-1}(\tfrac{7}{4}\gamma^2+3)) + 2(\tfrac{-1}{2}n^{-1/2}\gamma)^3 + O(n^{-3/2}) \\ &= -\tfrac{7}{2}n^{-1/2}\gamma + \tfrac{3}{2}n^{-1/2}\gamma + O(n^{-3/2}) \\ &= -2\,n^{-1/2}\,\gamma + O(n^{-3/2}). \end{aligned}$$

$$\begin{aligned} E(W_n - EW_n)^4 &= E(W_n^4) - 4E(W_n)E(W_n^3) + 6(EW_n)^2E(W_n^2) - 3(EW_n)^4 \\ &= 3 + n^{-1}(25\gamma^2 + \kappa + 30) + 6(\tfrac{-1}{2}n^{-1/2}\gamma)^2 + O(n^{-2}) \\ &= 3 + n^{-1}\left(\tfrac{39}{2}\gamma^2 + \kappa + 30\right) + O(n^{-2}). \end{aligned}$$

Désignons par S et K respectivement le coefficient d'asymétrie et le coefficient d'applatissement de W_n. On a : $S = \frac{E(W_n - EW_n)^3}{(var(W_n))^{3/2}}$ et $K = \frac{E(W_n - EW_n)^4}{(var(W_n))^2}$.
Le calcul donne :

$$\begin{aligned} S &= \frac{-2\gamma n^{-1/2} + O(n^{-3/2})}{\left[1 + \left(\frac{7}{4}\gamma^2 + 3\right)n^{-1} + O(n^{-2})\right]^{3/2}} \\ &= \left\{-2\gamma\, n^{-1/2} + O(n^{-3/2})\right\}\left\{1 - \tfrac{3}{2}\left(\tfrac{7}{4}\gamma^2 + 3\right)n^{-1} + O(n^{-2})\right\} \\ &= -2\gamma\, n^{-1/2} + O(n^{-3/2}). \end{aligned}$$

$$\begin{aligned} K &= \frac{3 + \left(\frac{39}{2}\gamma^2 + \kappa + 30\right)n^{-1} + O(n^{-2})}{\left[1 + \left(\frac{7}{4}\gamma^2 + 3\right)n^{-1} + O(n^{-2})\right]^2} \\ &= \left\{3 + \left(\tfrac{39}{2}\gamma^2 + \kappa + 30\right)n^{-1} + O(n^{-2})\right\}\left\{1 - 2\left(\tfrac{7}{4}\gamma^2 + 3\right)n^{-1} + O(n^{-2})\right\} \\ &= 3 + n^{-1}\left(9\gamma^2 + \kappa + 12\right) + O(n^{-2}). \end{aligned}$$

Le calcul numérique et les graphiques de ces paramètres sont présentés dans le tableau 2.2.1 et la figure 2.1 ci-dessous. (Prog. A.2, Annexe A)

Tableau 2.2.1 Variance, coefficient d'asymétrie et d'applatissement de W_n avec $p = 0.1$, $p = 0.25$ et $p = 0.5$ lorsque $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

	p \ n	20	30	40	50	60	70	80	90	100	150	200
var	0.10	1.77	1.51	1.39	1.31	1.26	1.22	1.19	1.17	1.15	1.08	1.08
	0.25	1.27	1.18	1.13	1.11	1.09	1.08	1.07	1.06	1.05	1.03	1.03
	0.50	1.15	1.10	1.08	1.06	1.05	1.04	1.04	1.03	1.03	1.02	1.02
S	0.10	-1.19	-0.97	-0.84	-0.75	-0.69	-0.64	-0.60	-0.56	-0.53	-0.39	-0.38
	0.25	-0.52	-0.42	-0.37	-0.33	-0.30	-0.28	-0.26	-0.24	-0.23	-0.17	-0.16
	0.50	0	0	0	0	0	0	0	0	0	0	0
K	0.10	7.06	5.70	5.03	4.62	4.35	4.16	4.01	3.90	3.81	3.43	3.41
	0.25	4.17	3.78	3.58	3.47	3.39	3.33	3.29	3.26	3.23	3.12	3.12
	0.50	3.50	3.33	3.25	3.20	3.17	3.14	3.13	3.11	3.10	3.05	3.05

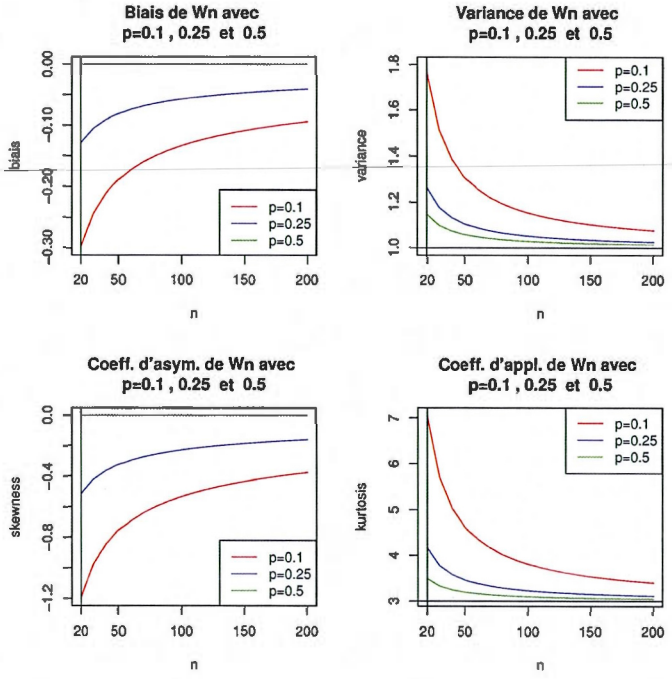

Figure 2.1 Biais, variance, coefficient d'asymétrie et d'applatissement de W_n avec $p = 0.1$, $p = 0.25$ et $p = 0.5$ lorsque n varie de 20 à 200

2.2.2 Développement d'Edgeworth d'ordre 1 de la probabilité de couverture

Nous allons utiliser un développement d'Edgeworth d'ordre 1 pour approximer la probabilité de couverture et en tirer la source de son oscillation. L'étude porte sur la probabilité de couverture de l'intervalle standard ICs et sur celles des intervalles alternatifs suggérés : de Wilson ICw, d'Agresti-Coull $ICac$, du rapport de vraisemblance $ICrv$ et de Jeffreys bilatéral ICj.

Soient $X \sim Bin(n, p)$, $\hat{p} = \frac{X}{n}$ et $Z_n = n^{1/2}\frac{\hat{p}-p}{(pq)^{1/2}}$. Définissons la fonction $g(p,z) = h(np + z(npq)^{1/2})$ avec $h(x) = x - [x]$ et $[x]$ est la partie entière de x ; autrement dit $g(p,z)$ est la partie fractionnelle de x. D'après le théorème 23.1 de Battacharya et Rao (1976), Z_n est développée comme suit :

$$P(Z_n \leq z) \;=\; \Phi(z) + \left[\left(\frac{1}{2} - g(p,z)\right) + \frac{1}{6}(1-2p)(1-z^2)\right]\phi(z)(npq)^{-1/2} + O(n^{-1}),$$

où $\Phi(z)$ et $\phi(z)$ sont respectivement la fonction de répartition et la fonction de densité de la loi normale centrée réduite $N(0,1)$. Les termes $\frac{1}{6}(1-2p)(1-z^2)$ et $\frac{1}{2} - g(p,z)$ représentent respectivement l'erreur d'asymétrie et l'erreur d'arrondi.

Remarque : $|\frac{1}{2} - g(p,z)| \leq \frac{1}{2}$ pour $\forall\ (p,z) \in [0,1]\times\mathbb{R}$.

2.2.2.1 Intervalle standard

La probabilité de couverture $P_p(p \in ICs)$ est $P_p(-z \leq n^{1/2}\frac{\hat{p}-p}{(\hat{p}\hat{q})^{1/2}} \leq z)$,
où z est le $100\,(1-\alpha)^e$ percentile de la loi normale centrée réduite $N(0,1)$.

$|n^{1/2}\frac{\hat{p}-p}{(\hat{p}\hat{q})^{1/2}}| \leq z \Longleftrightarrow n\frac{(\hat{p}-p)^2}{\hat{p}\hat{q}} \leq z^2 \Longleftrightarrow (n+z^2)\hat{p}^2 - (z^2+2np)\hat{p} + np^2 \leq 0$

$\Longleftrightarrow \hat{p}_1 \leq \hat{p} \leq \hat{p}_2$ avec $\hat{p}_{1,2} = \frac{z^2+2np \pm z\sqrt{z^2+4npq}}{2(z^2+n)}$.

Tableau 2.2.2 Comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle ICs à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

n	20	30	40	50	60	70	80	90	100	150	200
$C(p,n)$	0.921	0.946	0.905	0.938	0.922	0.940	0.932	0.947	0.933	0.940	0.941
$e(p,n)$	0.960	0.968	0.934	0.952	0.951	0.951	0.952	0.954	0.942	0.945	0.949
$C(p,n) - e(p,n)$	-0.039	-0.021	-0.029	-0.015	-0.028	-0.010	-0.020	-0.007	-0.009	-0.004	-0.008

Donc,

$$\begin{aligned} P_p(p \in ICs) &= P_p\left(n^{1/2}\frac{\hat{p}_1 - p}{(pq)^{1/2}} \leq Z_n \leq n^{1/2}\frac{\hat{p}_2 - p}{(pq)^{1/2}}\right) \\ &= P_p(ls \leq Z_n \leq us), \end{aligned}$$

où $(ls, us) = \frac{(1/2-p)z^2 n^{1/2} \pm zn\sqrt{pq+z^2/(4n)}}{(pq)^{1/2}(n+z^2)}$.

En développant ls et us, on obtient :

$$(ls, us) = \frac{(1/2-p)z^2}{\sqrt{npq}} \pm (z + \frac{(1/8-pq)z^2}{npq}) + O(n^{-3/2}),$$

où le signe $(-)$ va avec ls et $(+)$ avec us.

Donc, la probabilité de couverture de ICs satisfait :

$$\begin{aligned} P_p(p \in ICs) &= P_p(np + ls(npq)^{1/2} \leq X \leq np + us(npq)^{1/2}) \\ &= \underbrace{(1-\alpha)}_{terme\,1} + \underbrace{[g(p,ls) - g(p,us)]\,\phi(z)(npq)^{-1/2}}_{terme\,2} + O(n^{-1}). \end{aligned}$$

Le terme 2 représente l'erreur d'arrondi et il est la cause principale du phénomène de l'oscillation de la probabilité de couverture. Ce terme est de l'ordre de $n^{-1/2}$ et il est majoré par $\phi(z)(npq)^{-1/2}$, car $|g(p,ls) - g(p,us)| \leq 1$.

Une comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle ICs à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1 est présentée dans le tableau 2.2.2 ci-dessus. (Prog. A.3, Annexe A)

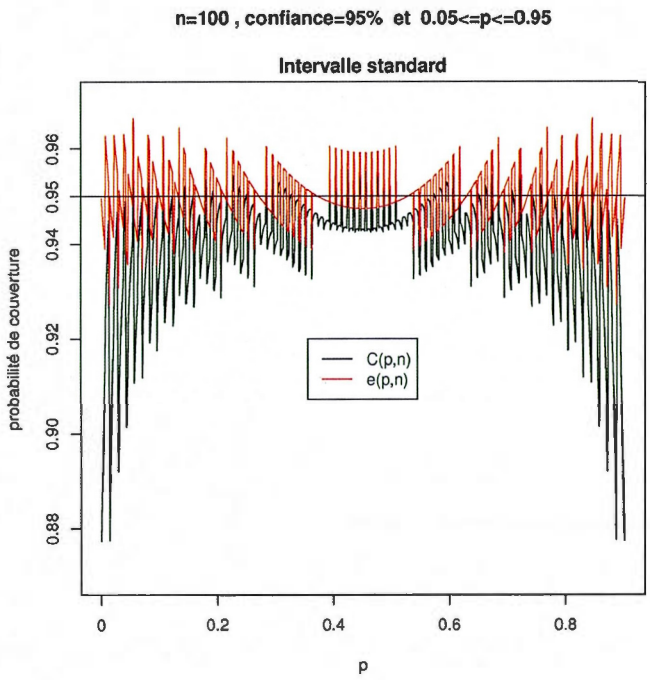

Figure 2.2 Comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle ICs à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 100$

La figure 2.2 ci-dessus nous permet de comparer graphiquement la probabilité de couverture $C(p,n)$ de l'intervalle ICs à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1. (Prog. A.4, Annexe A)

2.2.2.2 Intervalle de Wilson

L'intervalle de confiance de Wilson est basé sur l'inversion du test caractérisé par la région d'acceptation $\left\{p| -z \leq n^{1/2}\frac{\hat{p}-p}{(\hat{p}\hat{q})^{1/2}} \leq z\right\}$, mais en utilisant l'écart type sous l'hypothèse nulle $n^{-1/2}(pq)^{1/2}$ au lieu de son estimation $n^{-1/2}(\hat{p}\hat{q})^{1/2}$. Ceci revient à résoudre l'inégalité en p suivante : $|n^{1/2}\frac{\hat{p}-p}{(pq)^{1/2}}| \leq z$, où z est le $100\,(1-\alpha)^{\text{e}}$ percentile de la loi normale centrée réduite $N(0,1)$, $\hat{p} = X/n$ et $q = 1-p$. Cet intervalle est

Tableau 2.2.3 Comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle ICw à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

n	20	30	40	50	60	70	80	90	100	150	200
$C(p,n)$	0.956	0.964	0.928	0.951	0.966	0.950	0.965	0.953	0.941	0.944	0.958
$e(p,n)$	0.950	0.961	0.929	0.948	0.966	0.948	0.966	0.952	0.940	0.944	0.959
$C(p,n) - e(p,n)$	0.007	0.003	-0.001	0.003	0.000	0.002	-0.001	0.001	0.000	0.000	0.000

$ICw = \tilde{p} \pm \frac{zn^{1/2}}{n+z^2}(\hat{p}\hat{q} + \frac{z^2}{4n})^{1/2}$, où $\tilde{p} = \tilde{X}/\tilde{n}$, $\tilde{X} = X + z^2/2$, $\tilde{n} = n + z^2$ et $\tilde{q} = 1 - \tilde{p}$.
La probabilité de couverture $P_p(p \in ICw)$ est tout simplement égale à $P_p(-z \leq Z_n \leq z)$, où $Z_n = n^{1/2}\frac{\hat{p}-p}{(pq)^{1/2}}$. Donc,

$$\begin{aligned} P_p(p \in ICw) &= P_p(-z \leq Z_n \leq z) \\ &= P_p(np - z(npq)^{1/2} \leq X \leq np + z(npq)^{1/2}) \\ &= (1-\alpha) + [g(p,-z) - g(p,z)]\,\phi(z)(npq)^{-1/2} + O(n^{-1}). \end{aligned}$$

Une comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle ICw à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1 est présentée dans le tableau 2.2.3 ci-dessus.
La figure 2.3 ci-dessous montre une comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle ICw à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1

2.2.2.3 Intervalle d'Agresti-Coull

L'intervalle de confiance $ICac$ est $\tilde{p} \pm z(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}$, où $\tilde{X} = X + z^2/2$, $\tilde{n} = n + z^2$, $\tilde{p} = \tilde{X}/\tilde{n}$ et $\tilde{q} = 1 - \tilde{p}$.
La probabilité de couverture $P_p(p \in ICac)$ est $P_p(-z \leq \tilde{n}^{1/2}\frac{\tilde{p}-p}{(\tilde{p}\tilde{q})^{1/2}} \leq z)$.

$|\tilde{n}^{1/2}\frac{\tilde{p}-p}{(\tilde{p}\tilde{q})^{1/2}}| \leq z \Longleftrightarrow \tilde{n}\frac{(\tilde{p}-p)^2}{\tilde{p}\tilde{q}} \leq z^2 \Longleftrightarrow (\tilde{n}+z^2)\tilde{p}^2 - (z^2 + 2\tilde{n}p)\tilde{p} + \tilde{n}p^2 \leq 0.$

La solution de cette inégalité du deuxième degré à une inconnue p est l'ensemble des $\tilde{p}$ telle que $\tilde{p}_1 \leq \tilde{p} \leq \tilde{p}_2$, où $\tilde{p}_{1,2} = \frac{z^2+2\tilde{n}p \pm z\sqrt{z^2+4\tilde{n}pq}}{2(z^2+\tilde{n})}$. En remplaçant $\tilde{n}$ par $n + z^2$ dans

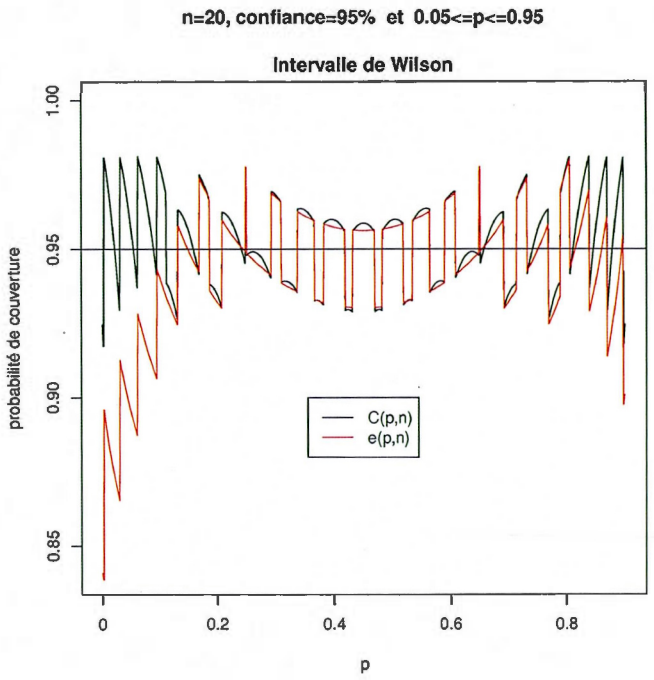

Figure 2.3 Comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle ICw à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 20$

l'expression de $\tilde{p}_{1,2}$, on obtient :

$$\tilde{p}_{1,2} = \frac{(1+2p)z^2 + 2np \pm z\sqrt{(1+4pq)z^2 + 4npq}}{2(2z^2+n)},$$

où le signe $(-)$ va avec $\tilde{p}_1$ et $(+)$ avec $\tilde{p}_2$.

Maintenant, on a : $\tilde{p}_1 \leq \tilde{p} \leq \tilde{p}_2 \Longleftrightarrow \tilde{p}_1 \leq \frac{\tilde{X}}{\tilde{n}} \leq \tilde{p}_2 \Longleftrightarrow \tilde{p}_1 \leq \frac{X+z^2/2}{n+z^2} \leq \tilde{p}_2$

$\Longleftrightarrow (n+z^2)\tilde{p}_1 - z^2/2 \leq X \leq (n+z^2)\tilde{p}_2 - z^2/2$

$\Longleftrightarrow \frac{(n+z^2)\tilde{p}_1 - z^2/2 - np}{(npq)^{1/2}} \leq Z_n \leq \frac{(n+z^2)\tilde{p}_2 - z^2/2 - np}{(npq)^{1/2}}$

$\Longleftrightarrow lac \leq Z_n \leq uac.$

Donc, $(lac, uac) = \frac{z/2}{2z^2+n}\left\{(2p-1)z^3 \pm (z^2+n)\sqrt{(1+4pq)z^2+4npq}\right\}(npq)^{-1/2}$.

Tableau 2.2.4 Comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICac$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

n	20	30	40	50	60	70	80	90	100	150	200
$C(p,n)$	0.956	0.964	0.949	0.951	0.966	0.950	0.965	0.953	0.941	0.944	0.958
$e(p,n)$	0.938	0.953	0.946	0.943	0.962	0.944	0.963	0.949	0.938	0.942	0.957
$C(p,n) - e(p,n)$	0.019	0.011	0.003	0.007	0.004	0.005	0.002	0.004	0.003	0.002	0.001

En développant lac et uac, on obtient :

$$(lac, uac) = \pm z \left\{ 1 + \left(\frac{1}{8pq} - \frac{1}{2} \right) z^2 n^{-1} \right\} + O(n^{-3/2}),$$

où le signe $(-)$ va avec lac et $(+)$ avec uac. Donc,

$$\begin{aligned} P_p(p \in ICac) &= P_p(lac \leq Z_n \leq uac) \\ &= P_p(np + lac(npq)^{1/2} \leq X \leq np + uac(npq)^{1/2}) \\ &= (1-\alpha) + [g(p, lac) - g(p, uac)]\, \phi(z)(npq)^{-1/2} + O(n^{-1}). \end{aligned}$$

Une comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICac$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1 est présentée dans le tableau 2.2.4 ci-dessus.

La figure 2.4 ci-dessous montre une comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICac$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1.

2.2.2.4 Intervalle du rapport de vraisemblance

Cet intervalle est construit par l'inversion de la région d'acceptation du test du rapport de vraisemblance. L'hypothèse nulle $H_0 = p_0$ n'est pas rejetée, si $-2\log(\Lambda_n) \leq z^2$, où Λ_n est le rapport de vraisemblance :

$$\Lambda_n = \left(\frac{p}{\hat{p}} \right)^{n\hat{p}} \left(\frac{q}{\hat{q}} \right)^{n\hat{q}}.$$

Soit $Z_n = n^{1/2} \frac{\hat{p}-p}{(pq)^{1/2}}$. Donc, $\hat{p} = n^{-1/2}\sqrt{pq}\, Z_n + p$ et $\hat{q} = q - n^{-1/2}\sqrt{pq}\, Z_n$. En

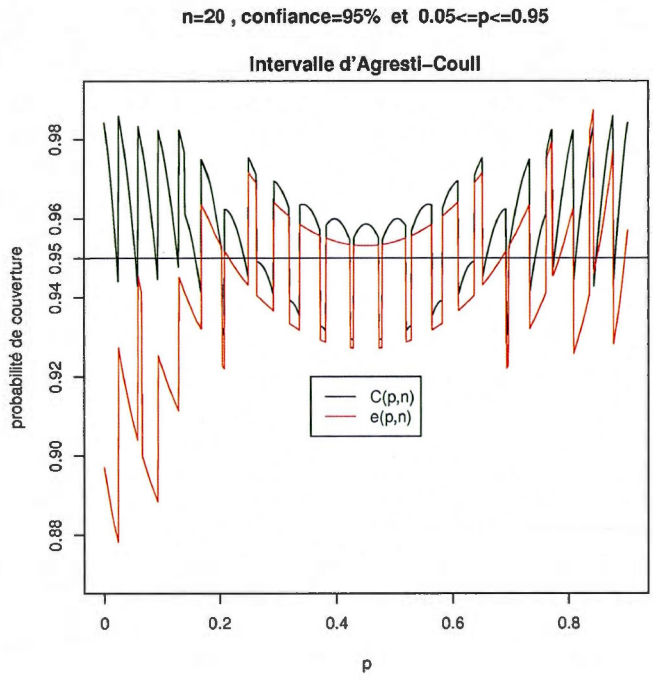

Figure 2.4 Comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICac$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 20$

introduisant ces valeurs dans l'expression de $-2\log(\Lambda_n) - z^2 \leq 0$, on obtient l'inégalité à une inconnue Z_n suivante : $f(Z_n) = p(1 + (\frac{q}{np})^{1/2} Z_n) \log(1 + (\frac{q}{np})^{1/2} Z_n)$ $+ q(1 - (\frac{p}{nq})^{1/2} Z_n) \log(1 - (\frac{p}{nq})^{1/2} Z_n) - \frac{z^2}{2n} \leq 0.$

Le support de $f(Z_n)$ est $I =]-(\frac{np}{q})^{1/2}, (\frac{nq}{p})^{1/2}[$ et sa dérivée seconde est

$\frac{d^2}{dZ_n^2} f(Z_n) = \frac{1/n}{(1+(\frac{q}{np})^{1/2} Z_n)(1-(\frac{p}{nq})^{1/2} Z_n)} \geq 0$, pour $\forall Z_n \in I$. Donc, $f(Z_n)$ est une fonction convexe sur I, et $f(Z_n) = O(n^{-3/2})$ a au plus deux racines, notées par lrv et urv. En développant en série de Taylor les deux fonctions $\log(1 + (\frac{q}{np})^{1/2} Z_n)$ et $\log(1 + (\frac{q}{np})^{1/2} Z_n)$, l'équation $f(Z_n) = O(n^{-3/2})$ devient :

$$Z_n^2 - \frac{1}{3}(1-2p)(pq)^{-1/2} n^{-1/2} Z_n^3 + \frac{1}{6}(1-3pq)(pq)^{-1} Z_n^4 - z^2 = O(n^{-3/2}).$$

Posons $r_1(p) = -\frac{1}{3}(1-2p)(pq)^{-1/2}n^{-1/2}$ et $r_2(p) = \frac{1}{6}(1-3pq)(pq)^{-1}$ et soit $Z_n = a + bn^{-1/2} + cn^{-1}$, où a, b, et c sont des constantes ne dépendant pas de n. Résoudre l'équation $f(Z_n) = O(n^{-3/2})$ revient à déterminer a, b, et c à partir du système d'équations suivant :

$$\begin{cases} a^2 &= z^2 \\ r_1(p)a^2 + 2ab &= 0 \\ 2r_1(p)ac + r_1(p)b^2 + 4r_2(p)a^2b(a+3) &= 0 \end{cases}$$

La solution est :

$$\begin{cases} a &= \pm z \\ b &= \frac{1}{6}(1-2p)(pq)^{-1/2}z^2 \\ c &= \pm\frac{1}{72}(\frac{1}{pq}+2)z^3 \end{cases}$$

Les racines de l'équation $f(Z_n) = O(n^{-3/2})$ sont :

$$(lrv, urv) = \pm\, z\left\{1 + \frac{1}{6}(1-2p)(pq)^{-1/2}\, z\, n^{-1/2} - \frac{1}{72}\left(\frac{1}{pq}+2\right) z^2\, n^{-1}\right\} + O(n^{-3/2}),$$

où le signe $(-)$ va avec lrv et $(+)$ avec (urv).

Donc, la probabilité de couverture se développe comme suit :

$$\begin{aligned} P_p(p \in ICrv) &= P_p(lrv \le Z_n \le urv) \\ &= P_p(np + lrv(npq)^{1/2} \le X \le np + urv(npq)^{1/2}) \\ &= (1-\alpha) + [g(p, lrv) - g(p, urv)]\, \phi(z)(npq)^{-1/2} + O(n^{-1}). \end{aligned}$$

Voir ci-dessous le tableau 2.2.5 et la figure 2.5 qui présentent respectivement une comparaison numérique et une comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICrv$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1.

Tableau 2.2.5 Comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICrv$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

n	20	30	40	50	60	70	80	90	100	150	200
$C(p,n)$	0.956	0.964	0.928	0.951	0.966	0.950	0.932	0.953	0.941	0.944	0.958
$e(p,n)$	0.953	0.963	0.930	0.949	0.967	0.949	0.934	0.952	0.940	0.944	0.959
$C(p,n) - e(p,n)$	0.003	0.001	-0.002	0.002	-0.001	0.001	-0.001	0.001	0.000	0.000	0.000

2.2.2.5 Intervalle de Jeffreys bilatéral

L'intervalle de Jeffreys bilatéral pour p est $ICj = \left[J_{\alpha/2}\,,\, J_{1-\alpha/2}\right]$, où $J_{\alpha/2}$ et $J_{1-\alpha/2}$ sont respectivement les quantiles $\alpha/2$ et $1-\alpha/2$ de la distribution *a posteriori* de p basée sur n observations. Dans le cas binomial, la loi *a priori* de Jeffreys est $B(1/2\,,\,1/2)$ et la loi *a posteriori* est $B(X+1/2\,,\,n-X+1/2)$, où $B(.)$ est la distribution bêta. Donc, l'intervalle ICj est $\left[B_{\alpha/2,X+1/2,n-X+1/2}\,,\, B_{1-\alpha/2,X+1/2,n-X+1/2}\right]$.

Soit $F(.)$ la fonction de répartition de la distribution bêta. La probabilité de couverture de ICj est :

$$P_p(p \in ICj) \;=\; P\{\alpha/2 \leq F\,[p;\, X+1/2,\, n-X+1/2] \leq 1-\alpha/2\}\,.$$

La fonction $F\,[p;\, X+1/2,\, n-X+1/2]$ est strictement décroissante en X; voir Johnson, Kots et Balakrishnan (1995). Par conséquent, il existe un unique $(X_l\,,\,X_u)$ satisfaisant :

$$\begin{cases} F\,[p;\, X_l+1/2,\, n-X_l+1/2] \;\leq\; 1-\alpha/2 \\ F\,[p;\, X_l-1/2,\, n-X_l+3/2] \;>\; 1-\alpha/2 \end{cases}$$

$$\begin{cases} F\,[p;\, X_u+1/2,\, n-X_u+1/2] \;\geq\; \alpha/2 \\ F\,[p;\, X_u+3/2,\, n-X_u-1/2] \;<\; 1-\alpha/2 \end{cases}$$

Donc, la probabilité de couverture de ICj est $P_p(p \in ICj) = P(lj \leq Z_n \leq uj)$, avec $lj = \frac{X_l - np}{npq}$ et $uj = \frac{X_u - np}{npq}$. Un développement de (lj, uj) est donné par :

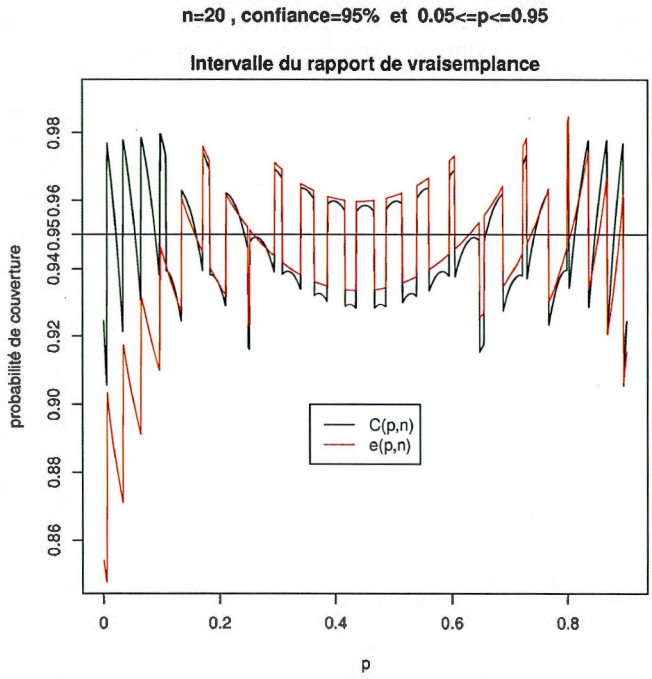

Figure 2.5 Comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle $ICrv$ à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 20$

$(lj, uj) = \pm z + \frac{1}{6}(z^2-1)\frac{(1-2p)}{\sqrt{npq}} \pm \left\{ \left(\frac{8}{pq} - \frac{1}{3} \right) z^3 + z \left(\frac{1}{3} - \frac{1-2p}{2pq} t_1(p) \right) + \frac{t_2(p)}{\sqrt{pq}} \right\} n^{-1} + O(n^{-3/2})$,

avec $t_1(p) = \frac{1}{6}(2z^2+1)(1-2p)$ et $t_2(p) = \frac{1}{36}\left[\frac{(z^2+2)}{pq} - (13z^2+17) \right] \frac{z}{\sqrt{pq}}$, où le signe $(-)$ va avec lj et $(+)$ avec (uj).

Donc, la probabilité de couverture se développe comme suit :

$$\begin{aligned} P_p(p \in ICj) &= P_p(lj \leq Z_n \leq uj) \\ &= P_p(np + lj(npq)^{1/2} \leq X \leq np + uj(npq)^{1/2}) \\ &= (1-\alpha) + [g(p,lj) - g(p,uj)]\, \phi(z)(npq)^{-1/2} + O(n^{-1}). \end{aligned}$$

Une comparaison numérique et une autre graphique de la probabilité de couverture $C(p,n)$ de l'intervalle ICj à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1 sont présentées dans le tableau 2.2.6 et la figure 2.6 ci-dessous.

Tableau 2.2.6 Comparaison numérique de la probabilité de couverture $C(p,n)$ de l'intervalle ICj à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$

n	20	30	40	50	60	70	80	90	100	150	200
$C(p,n)$	0.956	0.930	0.952	0.951	0.947	0.950	0.950	0.953	0.955	0.954	0.948
$e(p,n)$	0.930	0.915	0.936	0.931	0.942	0.939	0.928	0.930	0.927	0.944	0.954
$C(p,n) - e(p,n)$	0.026	0.015	0.016	0.020	0.005	0.011	0.022	0.023	0.028	0.010	-0.006

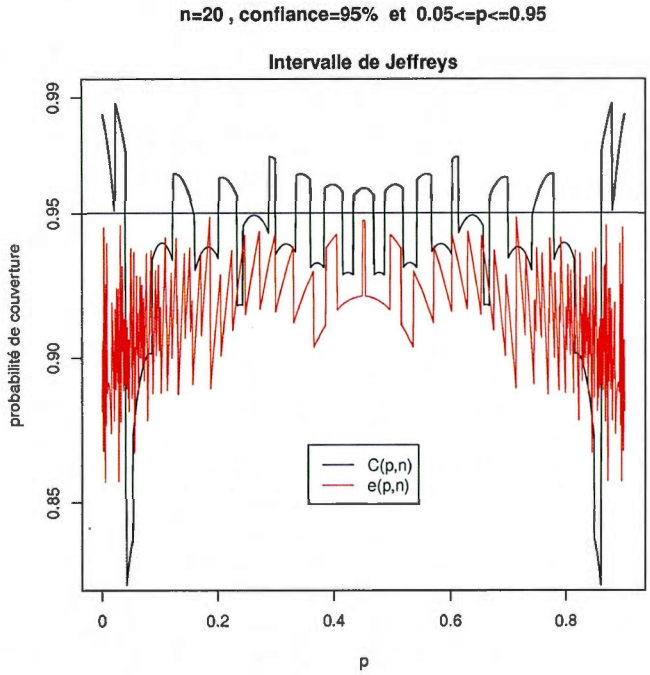

Figure 2.6 Comparaison graphique de la probabilité de couverture $C(p,n)$ de l'intervalle ICj à son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 20$

2.2.3 Développement d'Edgeworth d'ordre 2 de la probabilité de couverture

Le terme $O(n^{-1})$ dans le développement d'Edgeworth d'ordre 1 n'est pas négligeable, surtout lorsque n n'est pas grand. Donc, un développement d'Edgeworth d'ordre 2 est nécessaire pour pouvoir comparer les probabilités de couverture des intervalles de confiance vus dans la section précédente. Notons que nous avons omis les détails du calcul de ce développement. Pour les détails, voir Brown, Cai et DasGupta (1999). D'une manière générale, les approximations par un développement d'Edgeworth d'ordre 2 de la probabilité de couverture $P(p \in IC)$ se présentent comme suit :

$$\begin{aligned} P(p \in IC) &= (1-\alpha) + \ll O(n^{-1/2}) \text{ déviation} \gg + \ll O(n^{-1/2}) \text{ oscillation} \gg \\ &+ \ll O(n^{-1}) \text{ déviation} \gg + \ll O(n^{-1}) \text{ oscillation} \gg + O(n^{-3/2}). \end{aligned}$$

Les termes « $O(n^{-1/2})$ déviation » et « $O(n^{-1})$ déviation » décrivent les mouvements non-oscillatoires liés au biais. En plus des fonctions $h(x) = x - [x]$ et $g(p,z) = h(np + z(npq)^{1/2})$ définies lors du développement d'Edgeworth d'ordre 1, on définit deux autres fonctions, Q_{21} et Q_{22} :

$$\begin{aligned} Q_{21}(l,u) &= 1 - [g(p,l) + g(p,u)]; \\ Q_{22}(l,u) &= -\frac{1}{2}\left[g^2(p,l) + g^2(p,u)\right] + \frac{1}{2}\left[g(p,l) + g(p,u)\right] - \frac{1}{6}. \end{aligned}$$

Dans ce qui suit, on exprime l'approximation de la probabilité de couverture des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj par un développement d'Edgeworth d'ordre 2.

2.2.3.1 Intervalle standard

$$\begin{aligned} P_p(p \in ICs) &= (1-\alpha) + \left[g(p,ls) - g(p,us)\right]\phi(z)(npq)^{-1/2} \\ &+ \left\{\left(\frac{4}{9} - \frac{1}{9pq}\right)z^5 - \left(\frac{11}{18} + \frac{1}{18pq}\right)z^3 + \left(\frac{1}{6} - \frac{1}{6pq}\right)z\right\}\phi(z)n^{-1} \end{aligned}$$

$$+ \left\{-(1-2p)\left(\frac{1}{2}+\frac{z^2}{3}\right)Q_{21}(ls,us)+Q_{22}(ls,us)\right\}\frac{z}{pq}\phi(z)n^{-1}$$
$$+ O(n^{-3/2}).$$

Voir les expressions de ls et us dans le développement d'Edgeworth d'ordre 1.

2.2.3.2 Intervalle de Wilson

$$\begin{aligned}P_p(p \in ICw) &= (1-\alpha)+[g(p,-z)-g(p,z)]\,\phi(z)(npq)^{-1/2}\\ &+ \left\{\left(\frac{1}{9}-\frac{1}{36pq}\right)z^5+\left(\frac{7}{36pq}-\frac{11}{18}\right)z^3+\left(\frac{1}{6}-\frac{1}{6pq}\right)z\right\}\phi(z)n^{-1}\\ &+ \left\{(1-2p)\left(\frac{z^2}{6}-\frac{1}{2}\right)Q_{21}(-z,z)+Q_{22}(-z,z)\right\}\frac{z}{pq}\phi(z)n^{-1}\\ &+ O(n^{-3/2}).\end{aligned}$$

2.2.3.3 Intervalle d'Agresti-Coull

$$\begin{aligned}P_p(p \in ICac) &= (1-\alpha)+[g(p,lac)-g(p,uac)]\,\phi(z)(npq)^{-1/2}\\ &+ \left\{\left(\frac{1}{9}-\frac{1}{36pq}\right)z^5+\left(\frac{4}{9pq}-\frac{29}{18}\right)z^3+\left(\frac{1}{6}-\frac{1}{6pq}\right)z\right\}\phi(z)n^{-1}\\ &+ \left\{(1-2p)\left(\frac{z^2}{6}-\frac{1}{2}\right)Q_{21}(lac,uac)+Q_{22}(lac,uac)\right\}\frac{z}{pq}\phi(z)n^{-1}\\ &+ O(n^{-3/2}).\end{aligned}$$

Voir les expressions de lac et uac dans le développement d'Edgeworth d'ordre 1.

2.2.3.4 Intervalle du rapport de vraisemblance

$$\begin{aligned}P_p(p \in ICrv) &= (1-\alpha)+[g(p,lvr)-g(p,uvr)]\,\phi(z)(npq)^{-1/2}+(\frac{1}{6}-\frac{1}{6pq})z\phi(z)n^{-1}\\ &+ \left\{(p-\frac{1}{2})Q_{21}(lvr,uvr)+Q_{22}(lvr,uvr)\right\}\frac{z}{pq}\phi(z)n^{-1}+O(n^{-3/2}).\end{aligned}$$

Voir les expressions de lvr et uvr dans le développement d'Edgeworth d'ordre 1.

2.2.3.5 Intervalle de Jeffreys bilatéral

$$\begin{aligned} P_p(p \in ICj) &= (1-\alpha) + [g(p,lj) - g(p,uj)]\,\phi(z)(npq)^{-1/2} - \frac{z}{12pq}\phi(z)n^{-1} \\ &+ \left\{\frac{2p-1}{3}Q_{21}(lj,uj) + Q_{22}(lj,uj)\right\}\frac{z}{pq}\phi(z)n^{-1} + O(n^{-3/2}). \end{aligned}$$

Voir les expressions de lj et uj dans le développement d'Edgeworth d'ordre 1.

2.2.4 Exactitude du développement d'Edgeworth d'ordre 2 de la probabilité de couverture

Le développement d'Edgeworth d'ordre 2 approxime la vraie probabilité de couverture avec une erreur de $O(n^{-3/2})$. Lorsque p n'est pas proche des frontières 0 et 1, cette approximation est assez exacte même pour un n petit à modéré. Par contre, lorsque p est proche de ces frontières, le n doit être grand pour que l'exactitude de l'approximation soit bonne. Le tableau 2.2.7 ci-dessous présente l'erreur maximale commise entre la probabilité de couverture avec et sans développement d'Edgeworth d'ordre 2 pour les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj. Le p varie de 0.20 à 0.80, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$.

2.2.5 Comparaison des probabilités de couverture

On utilise le terme non-oscillatoire du développement d'Edgeworth d'ordre 2 de la probabilité de couverture des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj pour montrer la faible performance de l'intervalle standard ICs devant celles des autres. En rejetant l'erreur $O(n^{-3/2})$ et en n'utilisant que l'erreur non-oscillatoire d'ordre n^{-1}, on obtient :

$$\begin{aligned} P_p(p \in ICs) &= \left\{(\frac{4}{9} - \frac{1}{9pq})z^5 - (\frac{11}{18} + \frac{1}{18pq})z^3 + (\frac{1}{6} - \frac{1}{6pq})z\right\}\phi(z)n^{-1} + osci; \\ P_p(p \in ICw) &= \left\{(\frac{1}{9} - \frac{1}{36pq})z^5 + (\frac{7}{36pq} - \frac{11}{18})z^3 + (\frac{1}{6} - \frac{1}{6pq})z\right\}\phi(z)n^{-1} + osci; \end{aligned}$$

Tableau 2.2.7 Erreur maximale commise entre la probabilité de couverture avec et sans développement d'Edgeworth d'ordre 2 pour les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj, où $0.20 \leq p \leq 0.80$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$

n	20	30	40	100	150	200
ICs	0.0112	0.0100	0.0078	0.0017	0.0011	0.0008
ICw	0.0008	0.0004	0.0003	0.0001	0.0000	0.0000
ICac	0.0041	0.0022	0.0014	0.0005	0.0002	0.0001
ICrv	0.0323	0.0247	0.0225	0.0130	0.0105	0.0088
ICj	0.0004	0.0004	0.0003	0.0001	0.0000	0.0000

$$
\begin{aligned}
P_p(p \in ICac) &= \left\{ (\frac{1}{9} - \frac{1}{36pq})z^5 + (\frac{4}{9pq} - \frac{29}{18})z^3 + (\frac{1}{6} - \frac{1}{6pq})z \right\} \phi(z)n^{-1} + osci; \\
P_p(p \in ICrv) &= (\frac{1}{6} - \frac{1}{6pq})z\phi(z)n^{-1} + osci; \\
P_p(p \in ICj) &= -\frac{z}{12pq}\phi(z)n^{-1} + osci.
\end{aligned}
$$

La figure 2.7 ci-dessous représente les courbes des termes non-oscillatoires d'ordre n^{-1} de la probabilité de couverture des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj, où p varie de 0.05 à 0.95 et $\alpha = 0.05$. (Prog. A.5, Annexe A)

D'après cette figure, on voit clairement que l'intervalle d'Agresti-Coull $ICac$ a la plus grande probabilité de couverture. Ceci était prévisible parce que cet intervalle couvre p avec une probabilité au moins égale au niveau de confiance nominal $1 - \alpha$. C'est un intervalle conservateur. L'intervalle standard ICs a un biais négatif très significatif. Quant aux intervalles ICw, ICj, et $ICrv$, ils sont comparables.

2.2.6 Développement d'Edgeworth d'ordre 2 de la longueur moyenne

La longueur moyenne d'un intervalle de confiance est un critère fondamental qui, jumelée à la probabilité de couverture, permet de comparer les performances des différents intervalles. Dans ce qui suit, on donne les expressions du développement d'Edgeworth d'ordre 2 des longueurs espérées des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj. Ce développement renferme deux termes : le terme d'ordre $n^{-1/2}$ et le terme d'ordre $n^{-3/2}$.

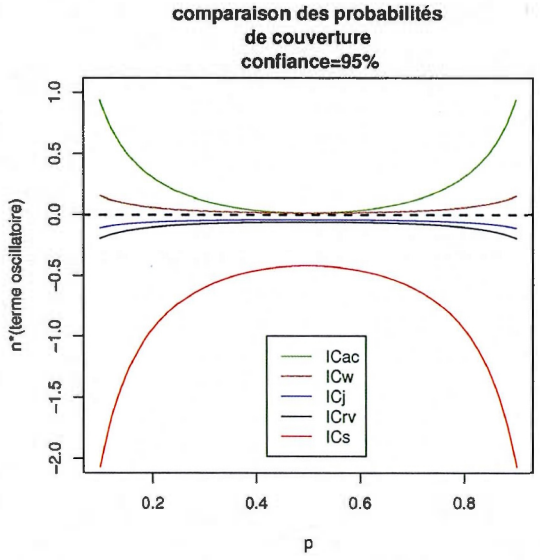

Figure 2.7 Courbes des termes non-oscillatoires d'ordre n^{-1} de la probabilité de couverture des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj, où $0.05 \leq p \leq 0.95$ et $\alpha = 0.05$

Voici ces expressions sans le détail des calculs. Pour le détail des calculs, voir Brown, Cai et DasGupta (1999).

2.2.6.1 Intervalle standard

La longueur est $Ls = 2zn^{-1/2}\sqrt{\frac{X}{n}(1-\frac{X}{n})}$ et le développement de la longueur espérée est :

$$E(Ls) = 2z(pq)^{1/2}n^{-1/2}(1-\frac{1}{8npq}) + O(n^{-2}).$$

2.2.6.2 Intervalle de Wilson

La longueur est $Lw = 2zn^{-1/2}\frac{n}{n+z^2}\sqrt{\frac{X}{n}(1-\frac{X}{n})+\frac{z^2}{4n}}$ et le développement de la longueur espérée est :

$$E(Lw) = 2z(pq)^{1/2}n^{-1/2}\left\{1-\frac{1+z^2(8pq-1)}{8npq}\right\} + O(n^{-2}).$$

2.2.6.3 Intervalle d'Agresti-Coull

La longueur est $Lac = 2z(n+z^2)^{-1/2}\left[\frac{X+z^2/2}{n+z^2}(1-\frac{X+z^2/2}{n+z^2})\right]^{1/2}$ et le développement de la longueur espérée est :

$$E(Lac) = 2z(pq)^{1/2}n^{-1/2}\left\{1-\frac{1+z^2(12pq-2)}{8npq}\right\}+O(n^{-2}).$$

2.2.6.4 Intervalle du rapport de vraisemblance

La longueur est $Lrv = 2z(\frac{\hat{q}}{\hat{p}})^{1/2}n^{-1/2} + \frac{z^3\hat{p}^{-3/2}\hat{q}^{-1/2}}{18}(1-13\hat{p}\hat{q})\,n^{-3/2}+O(n^{-2})$ et le développement de la longueur espérée est :

$$E(Lrv) = 2z(pq)^{1/2}n^{-1/2}\left\{1-\frac{1+z^2(\frac{26}{9}pq-\frac{2}{9})}{8npq}\right\}+O(n^{-2}).$$

2.2.6.5 Intervalle de Jeffreys bilatéral

La longueur est $Lj = 2zn^{-1/2}\sqrt{\frac{X}{n}(1-\frac{X}{n})}+2t_2(\hat{p})\,n^{-3/2}+O(n^{-2})$. La fonction $t_2(.)$ a été définie lors du développement d'Edgeworth d'ordre 2 de la probabilité de couverture de l'intervalle ICj. On remarque que $Lj = Ls + 2t_2(\hat{p})\,n^{-3/2}+O(n^{-2})$.
Le développement de la longueur espérée est :

$$E(Lj) = 2z(pq)^{1/2}n^{-1/2}\left\{1-\frac{1+z^2(\frac{26}{9}pq-\frac{2}{9})+(\frac{34}{9}pq-\frac{4}{9})}{8npq}\right\}+O(n^{-2}).$$

2.2.7 Exactitude du développement d'Edgeworth d'ordre 2 de la longueur moyenne

Le développement d'Edgeworth d'ordre 2 approxime la vraie longueur avec une erreur de $O(n^{-2})$. Lorsque p n'est pas proche des frontières 0 et 1, cette approximation est très exacte même pour un n petit à modéré. Par contre, si p est proche de ces frontières, le n doit être plus grand pour avoir une bonne approximation. Le tableau 2.2.8 ci-dessous présente l'erreur maximale commise entre la longueur moyenne

Tableau 2.2.8 Erreur maximale commise entre la longueur moyenne avec et sans développement d'Edgeworth d'ordre 2 pour les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj. Le p varie de 0.10 à 0.90, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$

n	20	30	40	100	150	200
ICs	0.0104	0.0034	0.0013	0.0001	0.0000	0.0000
ICw	0.0075	0.0029	0.0014	0.0002	0.0001	0.0000
ICac	0.0179	0.0070	0.0035	0.0004	0.0001	0.0001
ICrv	0.0016	0.0005	0.0002	0.0000	0.0000	0.0000
ICj	0.0024	0.0009	0.0004	0.0000	0.0000	0.0000

avec et sans développement d'Edgeworth d'ordre 2 pour les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj. Le p varie de 0.10 à 0.90, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$. (Prog. A.6 et Prog. A.7, Annexe A)

2.2.8 Comparaison des longueurs moyennes

Le coefficient du terme $n^{-1/2}$ est le même pour toutes les expressions du développement d'Edgeworth d'ordre 2 des longueurs espérées des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj, c'est-à-dire $2z(pq)^{1/2}$. Donc, la comparaison entre ces longueurs espérées se fait sur la base du coefficient du terme $n^{-3/2}$. Puisque ce coefficient est une fonction à 2 variables z et p, alors, en premier, on fixe α, disons $\alpha = 0.05$ $(z = 1.96)$, pour avoir une fonction à une seule variable p. Ensuite, on compare les fonctions obtenues selon les valeurs de p. Ces fonctions sont :

$$\begin{aligned}
\eta_s(z,p) &= 1 \qquad (ICs);\\
\eta_w(z,p) &= 1 + z^2(8pq - 1) \qquad (ICw);\\
\eta_{ac}(z,p) &= 1 + z^2(12pq - 2) \qquad (ICac);\\
\eta_{rv}(z,p) &= 1 + z^2(\frac{26}{9}pq - \frac{2}{9}) \qquad (ICrv);\\
\eta_j(z,p) &= 1 + z^2(\frac{26}{9}pq - \frac{2}{9}) + (\frac{34}{9}pq - \frac{4}{9}) \qquad (ICj).
\end{aligned}$$

Le tableau 2.2.9 ci-dessous présente toutes les comparaisons possibles entre les longueurs

Tableau 2.2.9 Toutes les comparaisons possibles entre les longueurs moyennes des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj, où $0 \leq p \leq 1$ et $\alpha = 0.05\,(z \approx 1.96)$

interv.	ICs	ICw	ICac	ICrv	ICj
ICs contient	NA	$0.146 \leq p \leq 0.854$	$0.211 \leq p \leq 0.789$	$0.084 \leq p \leq 0.916$	$0.061 \leq p \leq 0.939$
ICw contient	$0.854 \leq p$ ou $p \leq 0.146$	NA	jamais	$0.779 \leq p$ ou $p \leq 0.179$	$0.799 \leq p$ ou $p \leq 0.201$
ICac contient	$0.789 \leq p$ ou $p \leq 0.211$	$0 \leq p \leq 1$	NA	$0.734 \leq p$ ou $p \leq 0.266$	$0.713 \leq p$ ou $p \leq 0.287$
ICrv contient	$0.916 \leq p$ ou $p \leq 0.084$	$0.179 \leq p \leq 0.779$	$0.266 \leq p \leq 0.734$	NA	$0.136 \leq p \leq 0.864$
ICj contient	$0.939 \leq p$ ou $p \leq 0.061$	$0.201 \leq p \leq 0.799$	$0.287 \leq p \leq 0.713$	$0.864 \leq p$ ou $p \leq 0.136$	NA

Tableau 2.2.10 Domaines de p, où les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj sont les plus courts ou les plus longs. Le p varie de 0 à 1 et $\alpha = 0.05$

Intervalles	Plus court	Plus long
ICs	$0 \leq p \leq 0.084$ ou $0.916 \leq p \leq 1$	$0.211 \leq p \leq 0.789$
ICw	$0.200 \leq p \leq 0.800$	jamais
ICac	jamais	$0.286 \leq p \leq 0.714$
ICrv	$0.084 \leq p \leq 0.136$ ou $0.864 \leq 0.916$	jamais
ICj	$0.136 \leq p \leq 0.200$ ou $0.800 \leq p \leq 0.864$	jamais

moyennes des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj, où p varie de 0 à 1 et $\alpha = 0.05\,(z \approx 1.96)$.

D'après ce tableau, on voit par exemple que ICs contient ICw, si $0.146 \leq p \leq 0.854$ et ICw contient ICs, si $p \leq 0.146$ ou $p \geq 0.854$.

Toujours dans le cas où p varie de 0 à 1 et $\alpha = 0.05$, on détermine les domaines de p où les intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj sont les plus courts ou les plus longs. Le tableau 2.2.10 ci-dessus résume tous les cas possibles.

D'après ce tableau, on voit que ICw est plus court sur un domaine de p plus large : $(0.200 \leq p \leq 0.800)$ et que les intervalles $ICac$ et ICs sont plus longs sur des domaines de p plus larges : $(0.286 \leq p \leq 0.714)$ pour $ICac$ et $(0.211 \leq p \leq 0.789)$ pour ICs.

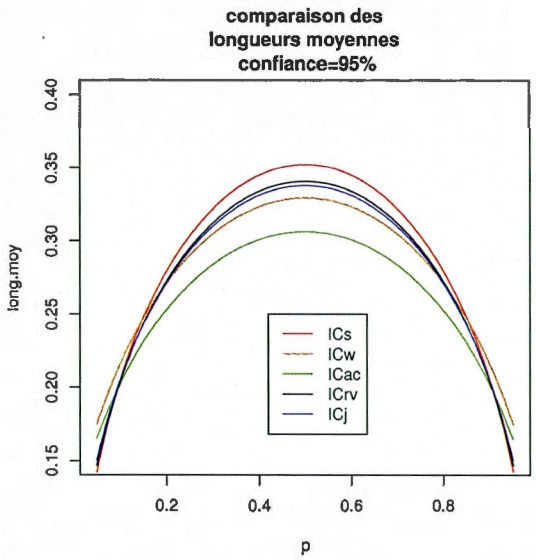

Figure 2.8 Comparaison graphique entre les longueurs moyennes des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj, où $0 \leq p \leq 1$, $\alpha = 0.05$ et $n = 30$

Avec p variant de 0 à 1, $\alpha = 0.05$ et $n = 30$, la figure 2.8 ci-dessus montre une comparaison graphique entre les longueurs moyennes des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj. (Prog. A.8, Annexe A)

2.3 Application

Dans cette section, on commence par la détermination des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj de la moyenne au bâton de chaque frappeur de l'ensemble S_1 en se basant sur la statistique H_i/N_i, où $H_i \sim B(N_i, p_i)$. Ensuite, on calcule la vraie moyenne au bâton de chaque frappeur de l'ensemble $S_1 \cap S_2$ et le nombre de fois où ces vraies moyennes au bâton sont à l'intérieur de leurs intervalles de confiance correspondants. Enfin, on calcule la probabilité de couverture moyenne et la longueur moyenne de chaque intervalle.

Remarque : Lors du calcul de la longueur moyenne, les intervalles ayant une borne inférieure négative sont tronqués à 0.

Tableau 2.3.11 Longueur moyenne et probabilité de couverture moyenne des intervalles ICs, ICw, $ICAc$, $ICrv$, et ICj. Proportion des vraies moyennes au bâton des frappeurs de l'ensemble $S_1 \cap S_2$ qui sont à l'intérieur de leurs intervalles correspondants. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs ayant au moins 11 présences au bâton. La prévision est basée sur les 2 premiers mois de la saison. Le $\alpha = 0.05$

	Tous			Non-lanceurs		
	Long.moy.	Prop.(%)	Couv.moy.(%)	Long.moy.	Prop.(%)	Couv.moy.(%)
ICs	0.1921	87.13	93.49	0.1795	88.05	94.35
ICw	0.1941	87.55	94.76	0.1772	87.81	94.98
ICac	0.1983	89.03	95.08	0.1792	88.78	95.07
ICrv	0.1932	86.08	94.27	0.1780	86.83	94.88
ICj	0.1928	85.87	93.72	0.1776	86.59	94.54

2.3.1 Prévision basée sur les 2 premiers mois de la saison

Les calculs se font, comme lors de l'estimation ponctuelle, sur tous les frappeurs (lanceurs et non-lanceurs) et sur les non-lanceurs ayant au moins 11 présences au bâton. Dans cette situation, le cardinal de S_1 pour l'estimation est de 518 pour tous les frappeurs et de 444 pour les non-lanceurs et le cardinal de $S_1 \cap S_2$ pour la validation est de 474 pour tous les frappeurs et de 410 pour les non-lanceurs. Voir le tableau 2.3.11 ci-dessus pour le résumé des résultats.

2.3.2 Prévision basée sur les 4 premiers mois de la saison

Les calculs sont restreints, comme lors de l'estimation ponctuelle, sur tous les frappeurs et sur les non-lanceurs ayant au moins 11 présences au bâton. Dans cette nouvelle situation, le cardinal de S_1 pour l'estimation est de 602 pour tous les frappeurs et de 509 pour les non-lanceurs et le cardinal de $S_1 \cap S_2$ pour la validation est de 490 pour tous les frappeurs et de 427 pour les non-lanceurs. Voir le résumé des résultats dans le tableau 2.3.12 ci-dessous. (Prog. A.9, Annexe A)

Maintenant, on analyse le cas où S_1 est composé des 12 frappeurs sélectionnés lors de l'estimation ponctuelle (Section 1.4.2). Dans ce cas, les cardinaux des ensembles S_1 et $S_1 \cap S_2$ sont égaux à 12. Le tableau 2.3.13 ci-dessous résume les résultats.

Tableau 2.3.12 Longueur moyenne et probabilité de couverture moyenne des intervalles ICs, ICw, $ICAc$, $ICrv$, et ICj. Proportion des vraies moyennes au bâton des frappeurs de l'ensemble $S_1 \cap S_2$ qui sont à l'intérieur de leurs intervalles correspondants. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs ayant au moins 11 présences au bâton. La prévision est basée sur les 4 premiers mois de la saison. Le $\alpha = 0.05$

	Tous			Non-lanceurs		
	Long.moy.	Prop.(%)	Couv.moy.(%)	Long.moy.	Prop.(%)	Couv.moy.(%)
ICs	0.1496	73.27	92.69	0.1380	74.24	93.68
ICw	0.1492	72.86	94.37	0.1361	73.77	94.87
ICac	0.1516	73.47	95.18	0.1372	74.00	95.21
ICrv	0.1491	72.25	93.99	0.1367	73.77	94.60
ICj	0.1489	72.25	92.95	0.1365	73.77	93.78

Tableau 2.3.13 Longueur moyenne et probabilité de couverture moyenne des intervalles ICs, ICw, $ICAc$, $ICrv$ et ICj. Proportion des vraies moyennes au bâton des frappeurs de l'ensemble $S_1 \cap S_2$ qui sont à l'intérieur de leurs intervalles correspondants. L'analyse est faite sur 12 frappeurs ayant au moins 11 présences au bâton. La prévision est basée sur les 4 premiers mois de la saison. Le $\alpha = 0.05$

	Long.moy.	Prop.(%)	Couv.moy.(%)
ICs	0.2163	83.33	90.68
ICw	0.2144	66.67	93.32
ICac	0.2196	66.67	95.41
ICrv	0.2115	66.67	93.44
ICj	0.2145	66.67	92.05

2.3.3 Prévision basée sur les 3 premiers mois de la saison

Les calculs se font sur tous les frappeurs et les non-lanceurs ayant au moins 11 présences au bâton. Dans cette situation, le cardinal de S_1 pour l'estimation est de 567 pour tous les frappeurs et de 486 pour les non-lanceurs et le cardinal de $S_1 \cap S_2$ pour la validation est de 499 pour tous les frappeurs et de 435 pour les non-lanceurs. Voir les résultats dans le tableau 2.3.14 ci-dessous.

Tableau 2.3.14 Longueur moyenne et probabilité de couverture moyenne des intervalles ICs, ICw, $ICAc$, $ICrv$ et ICj. Proportion des vraies moyennes au bâton des frappeurs de l'ensemble $S_1 \cap S_2$ qui sont à l'intérieur de leurs intervalles correspondants. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs ayant au moins 11 présences au bâton. La prévision est basée sur la première moitié de la saison. Le $\alpha = 0.05$

	Tous			Non-lanceurs		
	Long.moy.	Prop.(%)	Couv.moy.(%)	Long.moy.	Prop.(%)	Couv.moy.(%)
ICs	0.1668	83.17	93.00	0.1551	82.76	93.93
ICw	0.1675	81.76	94.51	0.1535	81.38	94.86
ICac	0.1707	81.76	95.12	0.1551	81.38	95.15
ICrv	0.1671	81.76	94.07	0.1541	81.61	94.66
ICj	0.1669	81.76	93.16	0.1538	81.38	94.04

2.4 Analyse des résultats

Dans cette section, on analyse la performance des intervalles de confiance ICs, ICw, $ICac$, $ICrv$ et ICj, selon leurs probabilités de couverture et leurs longueurs moyennes.

2.4.1 Performance des cinq intervalles de confiance selon leurs probabilités de couverture moyennes

Pour n'importe quelle période de prévision et pour n'importe quel groupe de frappeurs, les cinq intervalles sont classés, selon leurs probabilités de couverture moyennes, dans l'ordre décroissant suivant : $ICac$, ICw, $ICrv$, ICj et ICs. On note aussi que les probabilités de couverture moyennes d'un intervalle quelconque sont comparables, indépendamment de la période de prévision et du groupe de frappeurs utilisés dans l'étude.

2.4.2 Performance des cinq intervalles de confiance selon leurs longueurs moyennes

Dans le cas des non-lanceurs, les cinq intervalles de confiance sont classés, selon leurs longueurs moyennes et indépendamment de la période de prévision utilisée dans l'étude,

dans l'ordre croissant suivant : ICw, ICj, $ICrv$, $ICac$ et ICs. Par contre, dans le cas de tous les frappeurs, le classement dépend de la période de prévision. On distingue les deux cas suivants :

- Si la période de prévision est basée sur les 4 premiers mois de la saison, les cinq intervalles sont classés, selon leurs longueurs moyennes, dans l'ordre croissant suivant : ICj, $ICrv$, ICw, ICs et $ICac$.
- Si la période de prévision est basée sur les 2 ou les 3 premiers mois de la saison, ces intervalles sont classés, selon leurs longueurs moyennes, dans l'ordre croissant suivant : ICs, ICj, $ICrv$, ICw et $ICac$.

Aussi, on note que la longueur moyenne d'un intervalle quelconque est inversement proportionnelle à la largeur de la période de prévision utilisée et ceci indépendamment du groupe de frappeurs étudié.

CONCLUSION

Soit la variable aléatoire $X_i = \arcsin\sqrt{\frac{H_i+1/4}{N_i+1/2}}$, où H_i et N_i (≥ 11) sont respectivement le nombre de coups sûrs et le nombre de présences au bâton du joueur i à l'intérieur d'une période de 1 à 5 mois. Les valeurs de X_i ne sont pas bien ajustées à une distribution normale et la corrélation entre les valeurs de N_i et celles de X_i est assez élevée et ceci pour n'importe quelle période de prévision et groupe de frappeurs analysés. À cause de ces déviations des hypothèses idéales utilisées pour la modélisation des estimateurs de Bayes empiriques et de James-Stein, les performances de ces estimateurs sont affectées mais avec des degrés différents. Selon nous, dans de telles circonstances, l'estimateur de Bayes empirique non paramétrique et de James-Stein sont les plus adéquats.

La simplicité du calcul et le balancement entre la probabilité de couverture et la longueur moyenne forment, ensemble, le moyen pour évaluer la performance d'un intervalle de confiance. Dans le cas où seule la simplicité du calcul est prise comme critère d'évaluation, l'intervalle d'Agresti-Coull $ICac$ sera l'idéal. Cependant, la probabilité de couverture de cet intervalle est au moins égale au niveau de confiance nominal $1-\alpha$ (trop conservateur) et sa longueur moyenne est plus grande que celles de tous les intervalles alternatifs étudiés. Donc, on ne peut suggérer cet intervalle comme un intervalle alternatif à celui de Wald. L'intervalle de Jeffreys bilatéral ICj est un bon compétiteur, mais son calcul nécessite l'utilisation d'un logiciel. Maintenant, si le besoin de simplicité du calcul n'est pas très important, alors, selon nous, l'intervalle de Wilson ICw est le candidat préféré. En effet, il n'est pas trop difficile à calculer et sa performance, du point de vue compromis probabilité de couverture-longueur moyenne, est meilleure que celles de tous les autres intervalles alternatifs étudiés.

ANNEXE A

PROGRAMMES

Dans cette annexe, on présente quelques programmes en R (version 2.11.1) parmi ceux qui ont été développés dans ce mémoire pour effectuer les calculs et traçer les graphiques.

Dans les programmes Prog. A.1 et Prog. A.9 développés ci dessous, on a utilisé la base de données de baseball 2005 de Brown (2008b). Cette base est composée de 929 lignes (joueurs) et de 16 colonnes. Voici les noms de ces colonnes :

[1]	"First.Name"	"Last.Name"	"Pitcher"	"Season.AB"	"AB.4."
[6]	"AB.5."	"AB.6."	"AB.7."	"AB.8."	"AB.9.10."
[11]	"H.4."	"H.5."	"H.6."	"H.7."	"H.8."
[16]	"H.9.10."				

- "First.Name" et "Last.Name" sont les prénoms et les noms des joueurs ;
- "Pitcher" = 0 ou 1 selon que le frappeur est non-lanceur ou lanceur ;
- "Season.AB" est le nombre de présences au bâton d'un joueur à l'intérieur de toute la saison ;
- "AB.i." est le nombre de présences au bâton d'un joueur à l'intérieur du mois i ;
- "AB.i.j." est le nombre de présences au bâton d'un joueur à l'intérieur des mois i et j ;
- "H.i." est le nombre de coups sûrs d'un joueur à l'intérieur du mois i ;
- "H.i.j." est le nombre de coups sûrs d'un joueur à l'intérieur des mois i et j.

Prog. A.1 : Estimation de l'erreur quadratique totale normalisée $\widehat{EEQT}^{(n)}$ pour la moyenne générale $\hat{\theta}_{mg}$, Bayes empirique avec la méthode des moments $\hat{\theta}_{bemm}$, Bayes empirique avec la méthode du maximum de vraisemblance $\hat{\theta}_{bemv}$ et Bayes empirique non paramétrique $\hat{\theta}_{benp}$. L'analyse est faite sur tous les frappeurs et sur les non-lanceurs ayant au moins 11 présences au bâton et la prévision est basée sur les 2 premiers mois de la saison. (Tableau 1.4.3)

```
# Dans ce programme, on a la notation suivante: nL = Non-lanceurs et
# A = tous les frappeurs.

# Importation des données.
  base<-read.csv(file.choose())

# Présence au bâton >10.
  SnL<-base[-which(base$Pitcher==1),]

# Présence au bâton >10 pour la période d'estimation et pour le restant
# de la saison.
  SnL1_1<-SnL[-which((SnL$AB.4.+SnL$AB.5.)<11),]
  SnL1_2<-SnL[-which((SnL$AB.6.+SnL$AB.7.+ SnL$AB.8.
  +SnL$AB.9.10.)<11),]
  SA1_1<-base[-which((base$AB.4.+base$AB.5.)<11),]
  SA1_2<-base[-which((base$AB.6.+base$AB.7.+ base$AB.8.
  +base$AB.9.10.)<11),]

# Joueurs en commun pour la validation.
  SnL12_1<-SnL1_2[which((SnL1_2$AB.4.+SnL1_2$AB.5.)>10),]
  SA12_1<-SA1_2[which((SA1_2$AB.4.+SA1_2$AB.5.)>10),]
```

```
# Nombre de coups sûrs pour la période d'estimation.
  HnL1_1<-SnL1_1$H.4.+SnL1_1$H.5.
  HA1_1<-SA1_1$H.4.+SA1_1$H.5.

# Nombre de coups sûrs pour la validation.
  HnL12_1<-SnL12_1$H.6.+SnL12_1$H.7.+ SnL12_1$H.8.+SnL12_1$H.9.10.
  HA12_1<-SA12_1$H.6.+SA12_1$H.7.+ SA12_1$H.8.+SA12_1$H.9.10

# Nombre de présences au bâton pour la période d'estimation.
  NnL1_1<-SnL1_1$AB.4.+SnL1_1$AB.5.
  NA1_1<-SA1_1$AB.4.+SA1_1$AB.5.

# Nombre de présences au bâton pour la validation.
  NnL12_1<-SnL12_1$AB.6.+SnL12_1$AB.7.+ SnL12_1$AB.8.+SnL12_1$AB.9.10.
  NA12_1<-SA12_1$AB.6.+SA12_1$AB.7.+ SA12_1$AB.8.+SA12_1$AB.9.10

# Nombre de coups sûrs pour le restant de la saison.
  HnL1_2<-SnL1_2$H.6.+SnL1_2$H.7.+SnL1_2$H.8.+SnL1_2$H.9.10.
  HA1_2<-SA1_2$H.6.+SA1_2$H.7.+SA1_2$H.8.+SA1_2$H.9.10.

# Nombre de présences au bâton pour le restant de la saison.
  NnL1_2<-SnL1_2$AB.6.+SnL1_2$AB.7.+ SnL1_2$AB.8.+SnL1_2$AB.9.10.
  NA1_2<-SA1_2$AB.6.+SA1_2$AB.7.+ SA1_2$AB.8.+SA1_2$AB.9.10.

# Vecteurs relatifs à la période d'estimation.
  RnL1<-(HnL1_1+1/4)/(NnL1_1+1/2)
  XnL1<-asin(sqrt(RnL1))
  VarnL1<-1/(4*NnL1_1)
  RA1<-(HA1_1+1/4)/(NA1_1+1/2)
```

```
  XA1<-asin(sqrt(RA1))
  VarA1<-1/(4*NA1_1)

# Vecteurs relatifs à la validation.
  RnL12<-(HnL12_1+1/4)/(NnL12_1+1/2)
  XnL12<-asin(sqrt(RnL12))
  VarnL12<-1/(4*NnL12_1)
  RA12<-(HA12_1+1/4)/(NA12_1+1/2)
  XA12<-asin(sqrt(RA12))
  VarA12<-1/(4*NA12_1)

# Estimateur de la moyenne générale.
  enL1_MG<-(sin(mean(XnL1)))^2
  eA1_MG<-(sin(mean(XA1)))^2

# Vecteurs servant à la détermination de l'estimateur trivial.
  XnLc1<-asin(sqrt(((SnL12_1$H.4.+SnL12_1$H.5.)+1/4)
  /((SnL12_1$AB.4.+SnL12_1$AB.5.)+1/2)))
  XAc1<-asin(sqrt(((SA12_1$H.4.+SA12_1$H.5.)+1/4)
  /((SA12_1$AB.4.+SA12_1$AB.5.)+1/2)))

# Estimation de l'erreur quadratique totale pour le cas trivial.
  TEQnL1_T<-sum((XnL12-XnLc1)^2)-sum(VarnL12)
  TEQA1_T<-sum((XA12-XAc1)^2)-sum(VarA12)

# Somme de l'erreur quadratique pour le cas de la moyenne générale.
  SEQnL1_MG<-sum((XnL12-mean(XnL1))^2)
  SEQA1_MG<-sum((XA12-mean(XA1))^2)
```

```
# Estimation de l'erreur quadratique totale pour le cas de la moyenne
# générale.
  TEQnL1_MG<-SEQnL1_MG-sum(VarnL12)
  TEQA1_MG<-SEQA1_MG-sum(VarA12)

# Estimation de l'erreur quadratique totale normalisée pour le cas de
# la moyenne générale.
  TEQnL1_MG/TEQnL1_T
  TEQA1_MG/TEQA1_T

# Estimation de tau au carré initial servant à la détermination de l'
# estimateur de Bayes empirique par la méthode des moments.
  tau0nL1<-(1/(length(XnL1)-1))* sum((XnL1-mean(XnL1))^2)
  -(1/(length(XnL1)))* sum(VarnL1)
  tau0A1<-(1/(length(XA1)-1))* sum((XA1-mean(XA1))^2)
  -(1/(length(XA1)))* sum(VarA1)

# Estimation de mu.
  munL1<-(sum(XnL1/(tau0nL1+VarnL1)))/sum(1/(tau0nL1+VarnL1))
  muA1<-(sum(XA1/(tau0A1+VarA1)))/sum(1/(tau0A1+VarA1))

# Estimation de tau au carré.
  tau2nL1<-(1/(length(XnL1)-1))* sum((XnL1-munL1)^2)
  -(1/(length(XnL1)))*sum(VarnL1)
  tau2A1<-(1/(length(XA1)-1))* sum((XA1-muA1)^2)
  -(1/(length(XA1)))*sum(VarA1)

# Estimateur de Bayes empirique avec la méthode des moments.
  EBMnL1<-munL1+(tau2nL1/(tau2nL1+VarnL1))* XnL1-munL1)
```

```
  enl1_BM<-(sin(EBMnL1))^2
  EBMA1<-muA1+(tau2A1/(tau2A1+VarA1))* (XA1-muA1)
  eA1_BM<-(sin(EBMA1))^2

# Joueurs de la  période d'estimation avec des présences au bâton >10
# et somme de l'erreur quadratique pour le cas de Bayes empirique avec
# la méthode des moments.
  u2<-which((SnL1_1$AB.6.+SnL1_1$AB.7.
  +SnL1_1$AB.8.+SnL1_1$AB.9.10.)>10)
  SEQnL1_BM<-sum((XnL12-EBMnL1[u2])^2)
  u3<-which((SA1_1$AB.6.+SA1_1$AB.7.
  +SA1_1$AB.8.+SA1_1$AB.9.10.)>10)
  SEQA1_BM<-sum((XA12-EBMA1[u3])^2)

# Estimation de l'erreur quadratique totale pour le cas de Bayes
# empirique avec la méthode des moments.
  TEQnL1_BM<-SEQnL1_BM-sum(VarnL12)
  TEQA1_BM<-SEQA1_BM-sum(VarA12)

# Estimation de l'erreur quadratique totale normalisée pour le cas de
# Bayes empirique avec la méthode des moments.
  TEQnL1_BM/TEQnL1_T
  TEQA1_BM/TEQA1_T

# Estimation de tau au carré et de mu pour le cas de Bayes empirique
# avec la méthode du maximum de vraisemblance.
  tnL1<-uniroot(f=function(x){sum((1/(x+VarnL1)^2)*(XnL1
  -(sum(XnL1/(x+VarnL1)))/(sum(1/(x+VarnL1))))^2)
  -sum(1/(x+VarnL1))},lower=0,upper=100,tol=0.00001)
```

```
mnL1<-(sum(XnL1/(tnL1$root+VarnL1)))/(sum(1/(tnL1$root+VarnL1)))
tnL1$root
mnL1
tA1<-uniroot(f=function(x){sum((1/(x+VarA1)^2)*(XA1
-(sum(XA1/(x+VarA1)))/(sum(1/(x+VarA1))))^2)
-sum(1/(x+VarA1))},lower=0,upper=100,tol=0.00001)
mA1<-(sum(XA1/(tA1$root+VarA1)))/(sum(1/(tA1$root+VarA1)))
tA1$root
mA1

# Estimateur de Bayes empirique avec la méthode du maximum de
# vraisemblance.
EBVnL1<-mnL1+(tnL1$root/(tnL1$root+VarnL1))*(XnL1-mnL1)
enl1_BV<-(sin(EBVnL1))^2
EBVA1<-mA1+(tA1$root/(tA1$root+VarA1))*(XA1-mA1)
eA1_BV<-(sin(EBVA1))^2

# Somme de l'erreur quadratique pour le cas de Bayes empirique avec
# la méthode du maximum de vraisemblance.
SEQnL1_BV<-sum((XnL12-EBVnL1[u2])^2)
SEQA1_BV<-sum((XA12-EBVA1[u3])^2)

# Estimation de l'erreur quadratique totale pour le cas de Bayes
# empirique avec la méthode du maximum de vraisemblance.
TEQnL1_BV<-SEQnL1_BV-sum(VarnL12)
TEQA1_BV<-SEQA1_BV-sum(VarA12)

# Estimation de l'erreur quadratique totale normalisée pour le cas
# de Bayes empirique avec la méthode du maximum de vraisemblance.
```

```
TEQnL1_BV/TEQnL1_T
TEQA1_BV/TEQA1_T

# Estimateur de Bayes empirique non paramétrique. Faire aussi le
# calcul avec h=0.25
npnL1<-function(h){
gnL1<-rep(0,length(XnL1))
dgnL1<-rep(0,length(XnL1))
ENPnL1<-rep(0,length(XnL1))
for(i in 1:length(XnL1)){
k<-which((VarnL1-(1+h)*VarnL1[i])<0)
k1<-which((VarnL1[i]-VarnL1[k])>0)
k2<-which((VarnL1[k]-VarnL1[i])>0)
r<-1/length(k)
a<-VarnL1[k[k1]]
b<-VarnL1[k[k2]]
a1<-XnL1[k[k1]]
b1<-XnL1[k[k2]]
gnL1[i]<-r*(sum((1/sqrt((1+h)*VarnL1[i]-a))
*dnorm((XnL1[i]-a1)/sqrt((1+h)*VarnL1[i]-a)))
+sum((1/sqrt(h*b))*dnorm((XnL1[i]-b1)/sqrt(h*b))))
dgnL1[i]<-r*(sum(((a1-XnL1[i])/((1+h)*VarnL1[i]-a)^(3/2))
*dnorm((XnL1[i]-a1)/sqrt((1+h)*VarnL1[i]-a)))
+sum(((b1-XnL1[i])/(h*b)^(3/2))*dnorm((XnL1[i]-b1)/sqrt(h*b))))
ENPnL1[i]<-XnL1[i]+VarnL1[i]*(dgnL1[i])/(gnL1[i])
}
return(ENPnL1)
}
ENPnL1<-npnL1(1/log(444))
```

```
  enL_NP<-(sin(ENPnL1))^2

# Somme de l'erreur quadratique et estimation de l'erreur quadratique
# totale pour le cas de Bayes empirique non paramétrique.
  SEQnL1_NP<-sum((XnL12-ENPnL1[u2])^2)
  TEQnL1_NP<-SEQnL1_NP-sum(VarnL12)

# Estimation de l'erreur quadratique totale normalisée pour le cas de
# Bayes empirique non paramétrique.
  TEQnL1_NP/TEQnL1_T

# Estimateur de Bayes empirique non paramétrique. Faire aussi le
# calcul avec h=0.25
  npA1<-function(h){
  gA1<-rep(0,length(XA1))
  dgA1<-rep(0,length(XA1))
  ENPA1<-rep(0,length(XA1))
  for(i in 1:length(XA1)){
  k<-which((VarA1-(1+h)*VarA1[i])<0)
  k1<-which((VarA1[i]-VarA1[k])>0)
  k2<-which((VarA1[k]-VarA1[i])>0)
  r<-1/length(k)
  a<-VarA1[k[k1]]
  b<-VarA1[k[k2]]
  a1<-XA1[k[k1]]
  b1<-XA1[k[k2]]
  gA1[i]<-r*(sum((1/sqrt((1+h)*VarA1[i]-a))
  *dnorm((XA1[i]-a1)/sqrt((1+h)*VarA1[i]-a)))
  +sum((1/sqrt(h*b))*dnorm((XA1[i]-b1)/sqrt(h*b))))
```

```
  dgA1[i]<-r*(sum(((a1-XA1[i])/((1+h)*VarA1[i]-a)^(3/2))
  *dnorm((XA1[i]-a1)/sqrt((1+h)*VarA1[i]-a)))
  +sum(((b1-XA1[i])/(h*b)^(3/2))*dnorm((XA1[i]-b1)/sqrt(h*b))))
  ENPA1[i]<-XA1[i]+VarA1[i]*(dgA1[i])/(gA1[i])
  }
  return(ENPA1)
  }
  ENPA1<-npA1(1/log(518))
  eA_NP<-(sin(ENPA1))^2

# Somme de l'erreur quadratique et estimation de l'erreur quadratique
# totale pour le cas de Bayes empirique non paramétrique.
  SEQA1_NP<-sum((XA12-ENPA1[u3])^2)
  TEQA1_NP<-SEQA1_NP-sum(VarA12)

# Estimation de l'erreur quadratique totale normalisée pour le cas de
# Bayes empirique non paramétrique.
  TEQA1_NP/TEQA1_T

# Estimation de mu pour le cas de James-Stein.
  mu_nL1<-(sum(XnL1/VarnL1))/sum(1/VarnL1)
  mu_A1<-(sum(XA1/VarA1))/sum(1/VarA1)

# Estimateur de James-Stein.
  EJSnL1<-mu_nL1+(1-(length(XnL1)-3)/sum((XnL1-mu_nL1)^2/VarnL1))
  *(XnL1-mu_nL1)
  enL1_JS<-(sin(EJSnL1))^2
  EJSA1<-mu_A1+(1-(length(XA1)-3)/sum((XA1-mu_A1)^2/VarA1))
  *(XA1-mu_A1)
```

```
  eA1_JS<-(sin(EJSA1))^2

# Somme de l'erreur quadratique pour le cas de James-Stein.
  SEQnL1_JS<-sum((XnL12-EJSnL1[u2])^2)
  SEQA1_JS<-sum((XA12-EJSA1[u3])^2)

# Estimation de l'erreur quadratique totale pour le cas de
# James-Stein.
  TEQnL1_JS<-SEQnL1_JS-sum(VarnL12)
  TEQA1_JS<-SEQA1_JS-sum(VarA12)

# Estimation de l'erreur quadratique totale normalisée pour le cas
# de James-Stein.
  TEQnL1_JS/TEQnL1_T
  TEQA1_JS/TEQA1_T
```

Prog. A.2 : Calculs et graphiques du biais, de la variance et des coefficients d'asymétrie et d'applatissement de $W_n = \frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}\hat{q}}}$, lorsque $p \in \{0.10, 0.25, 0.50\}$ et n varie de 20 à 200. (Tableau 2.2.1 et Figure 2.1)

```
# Fonction pour le calcul et les graphiques du biais, de la variance et
# des coefficients d'asymétrie et d'applatissement de Wn avec p=0.10,
# 0.25 et 0.50 et n varie de 20 à 200.
  vsk<-function(p){
  variance<-rep(0,19)
  skewness<-rep(0,19)
  kurtosis<-rep(0,19)
  espw1<-rep(0,19);espw2<-rep(0,19);espw3<-rep(0,19);espw4<-rep(0,19)
```

```
  mc3<-rep(0,19)
  mc4<-rep(0,19)
  for(i in 1:19){
  n<-10*(i+1)
  b<-(1-2*p)/sqrt(n*p*(1-p))
  k<-(1-6*p*(1-p))/(n*p*(1-p))
  espw1[i]<--b/2
  espw2[i]<-1+3/n+2*(b^2)
  espw3[i]<--(7/2)*b
  espw4[i]<-3+25*(b^2)+k+30/n
  mc3[i]<--2*b
  mc4[i]<-3+(39/2)*(b^2)+k+30/n
  variance[i]<-1+(7/4)*(b^2)+3/n
  skewness[i]<--2*b
  kurtosis[i]<-3+9*(b^2)+k+12/n
  }
  liste<-list(variance,skewness,kurtosis,espw1)
  return(liste)
  }

# Biais.
  b1<-vsk(0.25)[[4]]
  b2<-vsk(0.10)[[4]]
  biais<-vsk(0.50)[[4]]

# Variances.
  var1<-vsk(0.25)[[1]]
  var2<-vsk(0.10)[[1]]
  variance<-vsk(0.50)[[1]]
```

```
# Coefficients d'asymétrie.
  s1<-vsk(0.25)[[2]]
  s2<-vsk(0.10)[[2]]
  skewness<-vsk(0.50)[[2]]

# Coefficients d'applatissement.
  k1<-vsk(0.25)[[3]]
  k2<-vsk(0.10)[[3]]
  kurtosis<-vsk(0.50)[[3]]

# Les quatre paramètres arrondis.
  Biais<-round(c(b1,b2,biais),rep(2,3))
  Variance<-round(c(var1,var2,variance),rep(2,3))
  Skewness<-round(c(s1,s2,skewness),rep(2,3))
  Kurtosis<-round(c(k1,k2,kurtosis),rep(2,3))
  Biais
  Variance
  Skewness
  Kurtosis

# Graphique du biais avec p=0.10, 0.25 et 0.50 et n varie de 20 à 200.
  par(mfrow=c(2,2))
  n<-seq(20,200,10)
  plot(biais~n,ylim=c(-0.30,0),type="l",col="green",main="Biais de Wn
  avec p=0.1 , 0.25  et  0.5")
  axis(side=1,at=20,labels="20")
  legend(x=130,y=-0.205,c("p=0.1","p=0.25","p=0.5"),col=c("red",
  "blue","green"),lty=1)
```

```
  lines(b1~seq(20,200,10),ylim=c(-0.30,0),type="l",col="blue")
  lines(b2~seq(20,200,10),ylim=c(-0.30,0),type="l",col="red")
  abline(v=20)

# Graphique de la variance avec p=0.10, 0.25 et 0.50 et n varie de
# 20 à 200.
  plot(variance~n,ylim=c(1,1.8),type="l",col="green",main="Variance
  de Wn avec p=0.1 , 0.25  et  0.5")
  axis(side=1,at=20,labels="20")
  legend(x=130,y=1.85,c("p=0.1","p=0.25","p=0.5"),col=c("red",
  "blue","green"),lty=1)
  lines(var1~seq(20,200,10),ylim=c(1,1.8),type="l",col="blue")
  lines(var2~seq(20,200,10),ylim=c(1,1.8),type="l",col="red")
  abline(v=20)
  abline(h=1)

# Graphique du coefficient d'asymétrie avec p=0.10, 0.25 et 0.50 et
# n varie de 20 à 200.
  plot(skewness~n,ylim=c(-1.20,0),type="l",col="green",main=
  "Coeff. d'asym. de Wn avec p=0.1 , 0.25  et  0.5")
  axis(side=1,at=20,labels="20")
  legend(x=130,y=-0.82,c("p=0.1","p=0.25","p=0.5"),col=c("red",
  "blue","green"),lty=1)
  lines(s1~seq(20,200,10),ylim=c(-1.20,0),type="l",col="blue")
  lines(s2~seq(20,200,10),ylim=c(-1.20,0),type="l",col="red")
  abline(v=20)

# Graphique du coefficient d'applatissement avec p=0.10, 0.25 et 0.50
# et n varie de 20 à 200.
```

```
plot(kurtosis~n,ylim=c(3.05,7.06),type="l",col="green",main=
"Coeff. d'appl. de Wn avec p=0.1 , 0.25  et  0.5")
axis(side=1,at=20,labels="20")
legend(x=130,y=7.25,c("p=0.1","p=0.25","p=0.5"),col=c("red",
"blue","green"),lty=1)
lines(k1~seq(20,200,10),ylim=c(3.05,7.06),type="l",col="blue")
lines(k2~seq(20,200,10),ylim=c(3.05,7.06),type="l",col="red")
abline(v=20)
abline(h=3)
```

Prog. A.3 : Calcul de la probabilité de couverture $C(p,n)$ de l'intervalle standard ICs et de son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, où $p = 0.2$, $\alpha = 0.05$ et n varie de 20 à 200. (Taleau 2.2.2)

```
# Fonction pour le calcul des probabilités de couverture C(n,p) de
# l'intervalle standard ICs et leurs approximations e(p,n) par un
# développement d'Edgeworth d'ordre 1. Le p=0.2, alpha=0.05 et n
# varie de 20 à 200.
  stand1<-function(z,p,alpha){
  ls<-rep(0,19)
  us<-rep(0,19)
  hl<-rep(0,19)
  hu<-rep(0,19)
  gl<-rep(0,19)
  gu<-rep(0,19)
  cnp<-rep(0,19)
  enp<-rep(0,19)
  for(i in 1:19){
```

```
  n<-20+10*(i-1)
  ls[i]<-((0.5-p)*(z^2)*(n^(0.5))-z*n*sqrt(p*(1-p)+(z^2)/(4*n)))
  /((n+(z^2))*sqrt(p*(1-p)))
  us[i]<-((0.5-p)*(z^2)*(n^(0.5))+z*n*sqrt(p*(1-p)+(z^2)/(4*n)))
  /((n+(z^2))*sqrt(p*(1-p)))
  hl[i]<-n*p+ls[i]*sqrt(n*p*(1-p))
  hu[i]<-n*p+us[i]*sqrt(n*p*(1-p))
  gl[i]<-hl[i]-trunc(hl[i])
  gu[i]<-hu[i]-trunc(hu[i])
  cnp[i]<-pbinom(hu[i],n,p)-pbinom(hl[i],n,p)
  enp[i]<-1-alpha+(gl[i]-gu[i])*dnorm(z)*(1/sqrt(n*p*(1-p)))
  }
  liste<-list(cnp=cnp,enp=enp)
  return(liste)
  }

# Ensemble des probabilités de couverture C(n,p).
  cnp<-stand1(qnorm(0.975),0.2,0.05)[[1]]

# Ensemble des probabilités de couverture e(n,p).
  enp<-stand1(qnorm(0.975),0.2,0.05)[[2]]

# Ensemble C(n,p), e(n,p) et C(n,p)-e(n,p) des probabilités de
# couverture arrondies.
  round(cnp,3)
  round(enp,3)
  round(cnp-enp,3)
```

Prog. A.4 : Graphiques de la probabilité de couverture $C(p, n)$ de l'intervalle standard

ICs et de son approximation $e(p,n)$ par un développement d'Edgeworth d'ordre 1, lorsque $0.05 \leq p \leq 0.95$, $\alpha = 0.05$ et $n = 100$. (Figure 2.2)

```
# Fonction pour traçer le graphique de la probabilité de couverture
# C(n,p) de l'intervalle standard ICs et le graphique de son
# approximation e(p,n) par un développement d'Edgeworth d'ordre 1.
# Le p varie de 0.05 à 0.95, alpha=0.05 et n=100.
  graph_stand1<-function(z,n,alpha){
  ls<-rep(0,901)
  us<-rep(0,901)
  hl<-rep(0,901)
  hu<-rep(0,901)
  gl<-rep(0,901)
  gu<-rep(0,901)
  cnp<-rep(0,901)
  enp<-rep(0,901)
  for(i in 1:901){
  p<-0.05+0.001*(i-1)
  ls[i]<-((0.5-p)*(z^2)*(n^(0.5))-z*n*sqrt(p*(1-p)+(z^2)/(4*n)))
  /((n+(z^2))*sqrt(p*(1-p)))
  us[i]<-((0.5-p)*(z^2)*(n^(0.5))+z*n*sqrt(p*(1-p)+(z^2)/(4*n)))
  /((n+(z^2))*sqrt(p*(1-p)))
  hl[i]<-n*p+ls[i]*sqrt(n*p*(1-p))
  hu[i]<-n*p+us[i]*sqrt(n*p*(1-p))
  gl[i]<-hl[i]-trunc(hl[i])
  gu[i]<-hu[i]-trunc(hu[i])
  cnp[i]<-pbinom(hu[i],n,p)-pbinom(hl[i],n,p)
  enp[i]<-1-alpha+(gl[i]-gu[i])*dnorm(z)*(1/sqrt(n*p*(1-p)))
  }
```

```
  liste<-list(cnp=cnp,enp=enp)
  return(liste)
  }

# Ensemble des probabilités de couverture C(n,p) de l'intervalle
# standard ICs.
  cnp<-graph_stand1(qnorm(0.975),100,0.05)[[1]]

# Ensemble des probabilités de couverture e(n,p) de l'intervalle
# standard ICs.
  enp<-graph_stand1(qnorm(0.975),100,0.05)[[2]]

# Graphique de C(n,p).
  plot(cnp,type="l",xaxt="n",ylim=c(0.87,0.97),main="n=100,confiance=
  95% et 0.05<=p<=0.95 \nIntervalle standard",xlab="p",ylab=
  "probabilité de couverture")
  axis(side=1,at=c(0,200,400,600,800,1000),labels= c(0.0,0.2,0.4,0.6,
  0.8,1.0))
  axis(side=2,at=0.95,labels=0.95)

# Ajout du graphique de e(n,p) à celui de C(n,p).
  lines(enp,type="l",col="red")

# Ajout de la droite horizontale C(p,n)=0.95
  abline(h=0.95)

# Identification des graphiques de C(p,n) et de e(p,n).
  legend(x=356,y=0.92,c("C(p,n)","e(p,n)"),col=c("black","red"),lty=1)
```

Prog. A.5 : Courbes des termes non-oscillatoires d'ordre n^{-1} de la probabilité de couverture des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj, lorsque $0.05 \leq p \leq 0.95$ et $\alpha = 0.05$. (Figure 2.7)

```
# Fonction pour traçer les Courbes des termes non-oscillatoires d'ordre
# n^(-1) de la probabilité de couverture des intervalles ICs, ICw, ICac,
# ICrv et ICj. Le p varie de 0.05 à 0.95 et alpha=0.05.
  comp<-function(z){
  Ps<-rep(0,901)
  Pw<-rep(0,901)
  Pac<-rep(0,901)
  Prv<-rep(0,901)
  Pj<-rep(0,901)
  for(i in 1:901){
  p<-0.05+0.001*(i-1)
# ième probabilité de couverture de ICs.
  Ps[i]<-((4/9-1/(9*p*(1-p)))*(z^5)-(1/(18*p*(1-p))+11/18)*(z^3)
  +(1/6-1/(6*p*(1-p)))*z)*dnorm(z)
# ième probabilité de couverture de ICw.
  Pw[i]<-((1/9-1/(36*p*(1-p)))*(z^5)+(7/(36*p*(1-p))
  -11/18)*(z^3)+(1/6-1/(6*p*(1-p)))*z)*dnorm(z)
# ième probabilité de couverture de ICac.
  Pac[i]<-((1/9-1/(36*p*(1-p)))*(z^5)+(4/(9*p*(1-p))
  -29/18)*(z^3)+(1/6-1/(6*p*(1-p)))*z)*dnorm(z)
# ième probabilité de couverture de ICrv.
  Prv[i]<-(1/6-1/(6*p*(1-p)))*z*dnorm(z)
# ième probabilité de couverture de ICj.
  Pj[i]<--(z/(12*p*(1-p)))*dnorm(z)
  }
```

```
  liste<-list(Ps,Pw,Pac,Prv,Pj)
  return(liste)
  }

# Ensemble des probabilités de couverture des intervalles ICs, ICw,
# ICac, ICrv et ICj.
  Ps<-comp(qnorm(0.975))[[1]]
  Pw<-comp(qnorm(0.975))[[2]]
  Pac<-comp(qnorm(0.975))[[3]]
  Prv<-comp(qnorm(0.975))[[4]]
  Pj<-comp(qnorm(0.975))[[5]]

# Courbe relative à ICac.
  curve(((1/9-1/(36*x*(1-x)))*(1.96^5)+(4/(9*x*(1-x))-29/18)*(1.96^3)
  +(1/6-1/(6*x*(1-x)))*1.96)*dnorm(1.96),xlim=c(0.1,0.90),ylim=
  c(-2,1),col="green",xlab="p",ylab="n*(terme oscillatoire)",main=
  comparaison des probabilités de couverture\n confiance=95%)

# Ajout de la courbe relative à ICj.
  curve(-(1.96/(12*x*(1-x)))*dnorm(1.96),col="blue",add=TRUE)

# Ajout de la courbe relative à ICs.
  curve(((4/9-1/(9*x*(1-x)))*(1.96^5)-(1/(18*x*(1-x))+11/18)*(1.96^3)
  +(1/6-1/(6*x*(1-x)))*1.96)*dnorm(1.96),col="red",add=TRUE)

# Ajout de la droite horizontale pointillée.
  abline(h=0,lty=2,lwd=2)

# Ajout de la courbe relative à ICrv.
```

```
  curve((1/6-1/(6*x*(1-x)))*1.96*dnorm(1.96),col="black",add=TRUE)

# Ajout de la courbe relative à ICw.
  curve(((1/9-1/(36*x*(1-x)))*(1.96^5)+(7/(36*x*(1-x))-11/18)*(1.96^3)
  +(1/6-1/(6*x*(1-x)))*1.96)*dnorm(1.96),col="brown",add=TRUE)

# Identification des courbes.
  legend(x=0.42,y=-1,c("ICac","ICw","ICj","ICrv","ICs"),col=c("green",
  "brown","blue","black","red"),lty=1)
```

Prog. A.6 : Calcul de l'erreur commise entre la longueur moyenne avec et sans développement d'Edgeworth d'ordre 2 pour l'intervalle de Wilson ICw, lorsque $0.10 \leq p \leq 0.90$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$. (Tableau 2.2.8)

```
# Fonction pour le calcul des longueurs moyennes avec et sans
# développement d'Edgeworth d'ordre 2 pour l'intervalle ICw. Le
# p varie de 0.1 à 0.90, alpha=0.05 et n appartient à {20,30,40,
# 100,150,200}.
# Remplacer les valeurs de n dans les objets mLw et L.
  Lwilson<-function(z,n){
  mLw<-rep(0,801)
  for(i in 1:801){
  p<-0.1+0.001*(i-1)
# ième longueur moyenne de ICw avec un développement d'Edgeworth
# d'ordre 2.
  mLw[i]<-2*z*sqrt(p*(1-p))*n^(-0.5)*(1-(1+z^2*(8*p*(1-p)-1))
  /(8*n*p*(1-p)))
  }
```

```
  return(mLw)
  }

# Ensemble des longueurs moyennes de ICw avec un développement
# d'Edgeworth d'ordre 2.
  mLw<-Lwilson(qnorm(0.975),n)

# Ensemble des longueurs moyennes de ICw sans développement
# d'Edgeworth.
# Charger le package "binom".
  library("binom")
  L<-binom.length(seq(0.1,0.90,0.001),n, conf.level=0.95,methods=
  "wilson")

# Erreur maximale arrondie.
  max(round(max(L$length-mLw),4),round(max(mLw-L$length),4))
```

Prog. A.7 : Calcul de l'erreur commise entre la longueur moyenne avec et sans développement d'Edgeworth d'ordre 2 pour l'intervalle de Jeffreys bilatéral ICj, lorsque $0.10 \leq p \leq 0.90$, $\alpha = 0.05$ et $n \in \{20, 30, 40, 100, 150, 200\}$. (Tableau 2.2.8)

```
# Fonction pour le calcul des longueurs moyennes avec et sans
# développement d'Edgeworth d'ordre 2 pour l'intervalle ICj. Le
# p varie de 0.1 à 0.90, alpha=0.05 et n appartient à {20,30,40,
# 100,150,200}.
# Remplacer les valeurs de n dans les objets mLj et L.
  Lbayes<-function(z,n){
  mLj<-rep(0,801)
```

```
  for(i in 1:801){
  p<-0.1+0.001*(i-1)
# ième longueur moyenne de ICj avec un développement d'Edgeworth
# d'ordre 2.
  mLj[i]<-2*z*sqrt(p*(1-p))*n^(-0.5)*(1-(1+z^2*((26/9)*p*(1-p)-2/9)
  +((34/9)*p*(1-p)-4/9))/(8*n*p*(1-p)))
  }
  return(mLj)
  }

# Ensemble des longueurs moyennes de ICj avec un développement
# d'Edgeworth d'ordre 2.
  mLj<-Lbayes(qnorm(0.975),200)

# Ensemble des longueurs moyennes de ICj sans développement
# d'Edgeworth.
# Charger le package "binom".
  library("binom")
  L<-binom.length(seq(0.1,0.90,0.001),200,conf.level=0.95,methods=
  "bayes")

# Erreur maximale arrondie.
  max(round(max(L$length-mLj),4),round(max(mLj-L$length),4))
```

Prog. A.8 : Graphiques des longueurs moyennes des intervalles ICs, ICw, $ICac$, $ICrv$ et ICj, lorsque $0 \leq p \leq 1$, $\alpha = 0.05$ et $n = 30$. (Figure 2.8)

```
# Courbes des longueurs moyennes des intervalles ICs, ICw, ICac,
```

```
# ICrv et ICj. Le p varie de 0 à 1, alpha=0.05 et n=30.
# Courbe relative à ICs.
  curve(((2*1.96/sqrt(30))*sqrt(x-x^2))*(1-1/(240*x-240*x^2)),
  from=0.05,to=0.95,ylim=c(0.15,0.4),type="l",col="red", main=
  "comparaison des longueurs moyennes\n confiance=95%",xlab="p",
  xlab="p",ylab="long.moy")

# Ajout de la courbe relative à ICw.
  curve(((2*1.96/sqrt(30))*sqrt(x-x^2))*(1-(1+(1.96)^2
  *(8*x-8*x^2-1))/(240*x-240*x^2)),col="chocolate",add=T)

# Ajout de la courbe relative à ICac.
  curve(((2*1.96/sqrt(30))*sqrt(x-x^2))*(1-(1+(1.96)^2
  *(12*x-12*x^2-1))/(240*x-240*x^2)),col="green",add=T)

# Ajout de la courbe relative à ICrv.
  curve(((2*1.96/sqrt(30))*sqrt(x-x^2))*(1-(1+(1.96)^2*((26/9)*x
  -(26/9)*x^2-2/9))/(240*x-240*x^2)),col="black",add=T)

# Ajout de la courbe relative à ICj.
  curve(((2*1.96/sqrt(30))*sqrt(x-x^2))*(1-(1+(1.96)^2*((26/9)*x
  -(26/9)*x^2-2/9)+(34/9)*x-(34/9)*x^2-4/9)/(240*x-240*x^2)),
  col="blue",add=T)

# Identification des courbes.
  legend(x=0.419,y=0.25,c("ICs","ICw","ICac","ICrv","ICj"),
  c("red","chocolate","green","black","blue"),lty=1)
```

Prog. A.9 : Calcul de la longueur moyenne et de la probabilité de couverture moyenne

de l'intervalle ICj et de la proportion des vraies moyennes au bâton des frappeurs de l'ensemble $S_1 \cap S_2$ qui sont à l'intérieur de cet intervalle. L'analyse est faite sur tous les frappeurs ayant au moins 11 présences au bâton et la prévision est basée sur les 4 premiers mois de la saison, où $\alpha = 0.05$. (Tableau 2.3.12)

```
# Dans ce programme, on a la notation suivante: nL = Non-lanceurs et
# A = tous les frappeurs

# Importation des données.
  base<-read.csv(file.choose())

# Présence au bâton >10 pour la période d'estimation et le restant de
# la saison.
  SA2_1<-base[-which((base$AB.4.+base$AB.5.+base$AB.6.
  +base$AB.7.)<11),]
  SA2_2<-base[-which((base$AB.8.+base$AB.9.10.)<11),]

# Joueurs en commun pour la validation.
  SA12_2<-SA2_2[which((SA2_2$AB.4.+SA2_2$AB.5.+SA2_2$AB.6.
  +SA2_2$AB.7.)>10),]

# Nombre de coups sûrs pour la période d'estimation.
  HA2_1<-SA2_1$H.4.+SA2_1$H.5.+SA2_1$H.6.+SA2_1$H.7.

# Nombre de coups sûrs pour la validation.
  HA12_2<-SA12_2$H.8.+SA12_2$H.9.10.

# Nombre de présences au bâton pour la période d'estimation.
  NA2_1<-SA2_1$AB.4.+SA2_1$AB.5.+SA2_1$AB.6.+SA2_1$AB.7.
```

```
# Nombre de coups sûrs pour le restant de la saison.
  HA2_2<-SA2_2$H.8.+SA2_2$H.9.10.

# Nombre de présences au bâton pour le restant de la saison.
  NA2_2<-SA2_2$AB.8.+SA2_2$AB.9.10.

# Nombre de présences au bâton pour la validation.
  NA12_2<-SA12_2$AB.8.+SA12_2$AB.9.10.

# Joueurs de la période d'estimation avec des présences au bâton >10.
  v3<-which((SA2_1$AB.8.+SA2_1$AB.9.10.)>10)

# Intervalles de confiance de ICj.
# Charger le package ("binom").
  library("binom")
  bayesA<-binom.bayes(HA2_1,NA2_1,conf.level=0.95,prior.shape1=0.5,
  prior.shape2=0.5,type=c("highest","central"),tol=0.000001,maxit=
  1000)

# Bornes inférieures et supérieures.
  binf_bA<-bayesA[,7]
  bsup_bA<-bayesA[,8]

# Vraies moyennes au bâton.
  estim_restA<-HA12_2/NA12_2

# Matrice à 3 colonnes et 490 lignes.
  int_bayesA<-cbind(binf_bA[v3],bsup_bA[v3],estim_restA)
```

```
# Nombre des vraies moyennes qui sont à l'intérieur de leurs
# intervalles ICj correspondants.
  interieur_bA<-length(which(int_bayesA[,3]<=int_bayesA[,2]&
  int_bayesA[,3]>=int_bayesA[,1]))

# Nombre des vraies moyennes qui sont à droite des bornes
# supérieures de leurs intervalles ICj correspondants.
  droite_bA<-length(which((int_bayesA[,3]
  -int_bayesA[,2])>0))
  droite_bA

# Nombre des vraies moyennes qui sont à gauche des bornes
# inférieures de leurs intervalles ICj correspondants.
  gauche_bA<-length(which((int_bayesA[,3]
  -int_bayesA[,1])<0))
  gauche_bA

# Nombre des bornes inférieures négatives.
  bneg_bA<-length(which(int_bayesA[,1]<0))
  bneg_bA

# Proportion des vraies moyennes au bâton qui sont à l'intérieur
# de leurs intervalles ICj correspondants.
  prop_bA<-interieur_bA/length(estim_restA)

# Proportion en %.
  100*prop_bA
```

```
# Fonction donnant la longueur moyenne de l'intervalle ICj.
  bA<-function(){
  binf_tr_bA<-rep(0,length(binf_bA[v3]))
  for(i in 1:length(binf_bA[v3])){
  if(binf_bA[v3][i]<0){
  binf_tr_bA[i]<-0}else{
  binf_tr_bA[i]<-binf_bA[v3][i]
  }
  }
  return(binf_tr_bA)
  }
  binf_tr_bA<-bA()
  long_bA<-bsup_bA[v3]-binf_tr_bA
  long_moy_bA<-mean(long_bA)

# Longueur moyenne.
  long_moy_bA

# Fonction donnant la probabilité de couverture moyenne de
# l'intervalle ICj.
  jeffA<-function(z,alpha){
  lj<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  uj<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  t1<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  t2<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  hl<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  hu<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  gl<-rep(0,length(NA12_2[-which(HA12_2==0)]))
  gu<-rep(0,length(NA12_2[-which(HA12_2==0)]))
```

```
enp<-rep(0,length(NA12_2[-which(HA12_2==0)]))
for(i in 1:length(NA12_2[-which(HA12_2==0)])){
p<-(HA12_2/NA12_2)[-which(HA12_2==0)][i]
n<-NA12_2[-which(HA12_2==0)][i]
t1[i]<-(1/6)*(2*(z^2)+1)*(1-2*p)
t2[i]<-(1/36)*(((z^2)+2)/(p*(1-p))-(13*(z^2)+17))
*z/(sqrt(p*(1-p)))
lj[i]<--z+(1/6)*((z^2)-1)*(1-2*p)/sqrt(n*p*(1-p))
-(1/n)*((8/(p*(1-p))-1/3)*(z^3)+z*(1/3-(1-2*p)*t1[i]
/(2*p*(1-p)))+t2[i]/sqrt(p*(1-p)))
uj[i]<-z+(1/6)*((z^2)-1)*(1-2*p)/sqrt(n*p*(1-p))+(1/n)
*((8/(p*(1-p))-1/3)*(z^3)+z*(1/3-(1-2*p)*t1[i]
/(2*p*(1-p)))+t2[i]/sqrt(p*(1-p)))
hl[i]<-n*p+lj[i]*sqrt(n*p*(1-p))
hu[i]<-n*p+uj[i]*sqrt(n*p*(1-p))
gl[i]<-hl[i]-trunc(hl[i])
gu[i]<-hu[i]-trunc(hu[i])
enp[i]<-1-alpha+(gl[i]-gu[i])*dnorm(z)*(1/sqrt(n*p*(1-p)))
-(z/(12*p*(1-p)))*(1/n)*dnorm(z)+(((2*p-1)/3)*(1-gl[i]-gu[i])
-(1/2)*((gl[i])^2)-(1/2)*((gu[i])^2)+(1/2)*gl[i]+(1/2)*gu[i]-1/6)
*(z/(n*p*(1-p)))*dnorm(z)
}
return(enp)
}
enp<-jeffA(qnorm(0.975),0.05)

# Probabilité de couverture moyenne en %.
100*mean(enp)
```

RÉFÉRENCES

AGRESTI, A. et COULL, B.A. 1998. « Approximate is better than « exact » for interval estimation of binomial proportion », Amer. Statist. **52** 119-126.

BHATTACHARYA, R.N. et RANGA RAO, R. 1976. « Normal approximation and asymptotic expansions », Wiley, New York.

BROWN, L. D., CAI, T. et DASGUPTA, A. 1999. « Confidence intervals for a binomial proportion and asymptotic expansions », Ann. Statist. **30** 160-201.

BROWN, L.D., CAI, T. et DASGUPTA, A. 2001. « Interval estimation for a binomial proportion (with discussion) », Statist. Sci. **16** 101-133.

BROWN, L. D. 2008a. « In-season prediction of batting averages : A field test of empirical Bayes and Bayes methodologies », Annals of Applied Statistics **2** 113-152.

BROWN, L. D. 2008b. « Supplement to In-season prediction of batting averages : A field test of empirical Bayes and Bayes methodologies ». En ligne. < `http://www-stat.wharton.upenn.edu/lbrown/publications.html` >.

HALL, P. 1992. « The Bootstrap and Edgeworth Expansion », Springer, New York.

JOHNSON, N. L., KOTZ, S. et BALAKRISHNAN, N. 1995. « Continuous Univariate Distributions, 2nd edition », Wiley, New York.

NEWCOMBE, R.G. 1998. « Two-sided confidence intervals for the single proportion ; comparison of several methods », Statist. Med. **17** 857-872.

STEIN, C. 1981. « Estimation of the mean of a multivariate normal distribution », Ann. Statist. **9** 1135-1151.

STRAWDERMAN, R.L. et WELLS, M.T. 1998. « Approximately exact inference for the common odds ratio in several 2×2 tables (with discussion) », J. Amer. Statist. Assoc. **93** 1294-1320.