

Université de Montréal

**Utilisation de l'estimateur d'Agresti-Coull dans la
construction d'intervalles de confiance bootstrap pour
une proportion**

par

Mikaël Pilotte

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

23 octobre 2020

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Utilisation de l'estimateur d'Agresti-Coull dans la construction d'intervalles de confiance bootstrap pour une proportion

présenté par

Mikaël Pilotte

a été évalué par un jury composé des personnes suivantes :

Mylène Bédard

(président-rapporteur)

Christian Léger

(directeur de recherche)

Pierre Duchesne

(membre du jury)

Mémoire accepté le :

5 janvier 2021

Sommaire

Pour construire des intervalles de confiance, nous pouvons utiliser diverses approches bootstrap. Nous avons un problème pour le contexte spécifique d'un paramètre de proportion lorsque l'estimateur usuel, la proportion de succès dans l'échantillon \hat{p} , est nul. Dans un contexte classique d'observations indépendantes et identiquement distribuées (i.i.d.) de la distribution Bernoulli, les échantillons bootstrap générés ne contiennent que des échecs avec probabilité 1 et les intervalles de confiance bootstrap deviennent dégénérés en un seul point, soit le point 0. En contexte de population finie, nous sommes confrontés aux mêmes problèmes lorsqu'on applique une méthode bootstrap à un échantillon de la population ne contenant que des échecs.

Une solution possible s'inspire de l'estimateur utilisé dans les méthodes de [Wilson, 1927] et [Agresti et Coull, 1998] où ceux-ci considèrent \tilde{p} l'estimateur qui prend la proportion de succès d'un échantillon augmenté auquel on a ajouté deux succès et deux échecs. La solution que nous introduisons consiste à effectuer le bootstrap de la distribution de \hat{p} mais en appliquant les méthodes bootstrap à l'échantillon augmenté de deux succès et deux échecs, tant en statistique classique que pour une population finie.

Les résultats ont démontré qu'une version de la méthode percentile est la méthode bootstrap la plus efficace afin d'estimer par intervalle de confiance un paramètre de proportion autant dans un contexte i.i.d. que dans un contexte d'échantillonnage avec le plan aléatoire simple sans remise. Nos simulations ont également démontré que cette méthode percentile pouvait compétitionner avantageusement avec les meilleures méthodes traditionnelles.

Mots-clés : bootstrap, estimateur de proportion, échantillonnage, pseudo-population, poids bootstrap, intervalle de confiance bootstrap.

Summary

A few bootstrap approaches exist to create confidence intervals. Some difficulties appear for the specific case of a proportion when the usual estimator, the proportion of success in a sample, is 0. In the classical case where the observations are independently and identically distributed (i.i.d.) from a Bernoulli distribution, the bootstrap samples only contain zeros with probability 1 and the resulting bootstrap confidence intervals are degenerate at the value 0. We are facing the same problem in the survey sampling case when we apply the bootstrap method to a sample with all observations equal to 0.

A possible solution is suggested by the estimator found in the confidence intervals of [Wilson, 1927] and [Agresti et Coull, 1998] where they use \tilde{p} the proportion of success in an augmented sample consisting of adding two successes and two failures to the original sample. The proposed solution is to use the bootstrap method on \hat{p} but where the bootstrap is based on the augmented sample with two additional successes and failures, whether the sample comes from i.i.d. Bernoulli variables or from a simple random sample.

Results show that a version of the percentile method is the most efficient bootstrap method to construct confidence intervals for a proportion both in the classical setting or in the case of a simple random sample. Our results also show that this percentile interval can compete with the best traditional methods.

Key words: bootstrap, binomial proportion estimate, sampling, pseudo-population, bootstrap weights, bootstrap confidence interval.

Table des matières

Sommaire	v
Summary	vii
Liste des tableaux	xi
Table des figures	xiii
Remerciements	xix
Introduction	1
Chapitre 1. Méthodes bootstrap dans un contexte i.i.d.	7
1.1. Contexte i.i.d.	7
1.2. Problématique et solution proposée	10
1.3. Intervalles de confiance bootstrap	11
Chapitre 2. Méthodes bootstrap en échantillonnage	15
2.1. Échantillonnage	15
2.2. Méthode bootstrap par pseudo-population	19
2.3. Méthode des poids bootstrap	20
2.4. Modifications aux méthodes bootstrap	23
2.5. Intervalles de confiance bootstrap	30
2.5.1. Estimation	31
Chapitre 3. Méthodes traditionnelles considérées	35

3.1.	Intervalle de Wald	35
3.2.	Intervalle Logit	36
3.3.	Intervalle de Wilson.....	36
3.4.	Intervalle Agresti-Coull	37
3.5.	Intervalle Clopper-Pearson.....	37
3.6.	Intervalle Arcsine Square Root.....	38
3.7.	Changements à apporter dans un contexte d'échantillonnage	38
Chapitre 4.	Résultats	39
4.1.	Méthodologie de la simulation.....	39
4.2.	Résultats sous un contexte i.i.d.....	41
4.3.	Résultats sous un contexte d'échantillonnage	58
4.3.1.	Méthode par pseudo-population.....	59
4.3.2.	Méthode des poids.....	73
4.3.3.	Méthodes traditionnelles	84
4.3.4.	Comparaisons des meilleures méthodes entres elles	90
Conclusion	93
Bibliographie	97
Annexe A.	Longueurs d'intervalle bootstrap-t dans le contexte d'échantillonnage	A-i
A.1.	Longueurs d'intervalles pour la méthode bootstrap- t	A-i

Liste des tableaux

- A.1 Quantiles $\alpha/2$ de la distribution de $K_{n_s}^*$ utilisés dans la construction de la borne supérieure de l'intervalle bootstrap- t pour la méthode par les poids avec un bootstrap à la Agresti-Coull.....A-iii
- A.2 Quantiles $\alpha/2$ de la distribution de $K_{n_s}^*$ utilisés dans la construction de la borne supérieure de l'intervalle bootstrap- t pour la méthode par pseudo-population avec un bootstrap à la Agresti-Coull.....A-v

Table des figures

4.1	Couverture des méthodes bootstrap avec l'utilisation du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 40$	42
4.2	Longueur moyenne des intervalles bootstrap à l'aide du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 40$	43
4.3	Couverture des méthodes bootstrap avec l'utilisation du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 200$	44
4.4	Longueur moyenne des intervalles bootstrap à l'aide du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 200$	45
4.5	Couverture des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 40$	46
4.6	Couverture et longueur moyenne des intervalles des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 40$	47
4.7	Couverture des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 200$	49
4.8	Couverture et longueur moyenne des intervalles des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 200$	50
4.9	Couverture des méthodes traditionnelles dans un contexte i.i.d. avec $n = 40$	51
4.10	Longueur moyenne des méthodes traditionnelles dans un contexte i.i.d. avec $n = 40$	53
4.11	Couverture des méthodes traditionnelles dans un contexte i.i.d. avec $n = 200$	54
4.12	Couverture et longueur moyenne des intervalles des méthodes Logit, Wilson et ASSQR dans un contexte i.i.d. avec $n = 200$	55

4.13	Couverture et longueur moyenne des intervalles des méthodes percentile sous le bootstrap à la Agresti-Coull, sous le bootstrap adaptatif, de la méthode ASSQR et de la méthode de Wilson avec $n = 40$.	56
4.14	Couverture et longueur moyenne des intervalles des méthodes percentile sous le bootstrap à la Agresti-Coull, sous le bootstrap adaptatif, de la méthode ASSQR et la méthode de Wilson avec $n = 200$.	58
4.15	Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 40$.	59
4.16	Longueur moyenne des intervalles bootstrap pour la méthode par pseudo-population à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.	60
4.17	Longueur moyenne des intervalles bootstrap pour la méthode par pseudo-population à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.	61
4.18	Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$.	63
4.19	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.	64
4.20	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.	65
4.21	Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$.	66

4.22	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.	67
4.23	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.	68
4.24	Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$.	69
4.25	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.	70
4.26	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.	71
4.27	Couvertures de l'intervalle percentile sous la méthode par pseudo-population . . .	72
4.28	Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 40$.	73
4.29	Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.	74
4.30	Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.	75
4.31	Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$.	76
4.32	Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.	77
4.33	Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$.	78

4.34	Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap adaptatif sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$..	79
4.35	Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap adaptatif sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$..	80
4.36	Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$	81
4.37	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.	82
4.38	Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.	83
4.39	Couvertures de l'intervalle percentile sous la méthode des poids.....	85
4.40	Couverture des méthodes traditionnelles dans un contexte d'échantillonnage avec $n = 40$	86
4.41	Couverture et longueur moyenne des intervalles des méthodes de Wilson, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$	87
4.42	Couverture et longueur moyenne des intervalles des méthodes de Wilson, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$	88
4.43	Couverture des méthodes traditionnelles dans un contexte d'échantillonnage avec $n = 200$	89
4.44	Couverture et longueur moyenne des intervalles des méthodes de Wilson, Logit, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$	90
4.45	Couverture et longueur moyenne des intervalles percentile avec un bootstrap à la Agresti-Coull sous la méthode des poids et par pseudo-population avec la méthode ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$	91

4.46	Couverture et longueur moyenne des intervalles percentile avec un bootstrap à la Agresti-Coull sous la méthode des poids et par pseudo-population avec la méthode ASSQR dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$	92
------	---	----

Remerciements

L'année 2020 restera l'une des plus historiques de tous les temps en ayant apporté son lot d'épreuves et de difficultés au niveau mondial. Elle gardera, dans ce sens, une signification toute spéciale pour moi, car elle marque la finalisation de ce travail. Plusieurs personnes m'ont aidé de près ou de loin et je m'en voudrais de ne pas les remercier.

Mes premiers remerciements s'adressent tout naturellement à mon directeur de recherche, M. Christian Léger. Vous avez su être un directeur plus qu'exemplaire pour moi tout au long de ce processus qui vient aujourd'hui de s'achever. Vos explications et vos remarques étaient toujours très justes et pertinentes. Je n'oublierai jamais votre soutien ainsi que la patience et la compréhension que vous avez démontrés à mon endroit. Sachez que vous êtes la personne qui m'a permis d'accomplir ce travail colossal et de l'amener à un niveau supérieur grâce à vos conseils d'une valeur inestimable. Je vous serai toujours reconnaissant pour tout ce que vous avez fait pour moi au cours des dernières années.

C'est avec beaucoup d'émotions que j'adresse en second lieu mes remerciements à mes parents. Un simple remerciement n'est pas suffisant à mes yeux pour vous démontrer la gratitude et la reconnaissance que j'éprouve à votre égard. Vous m'avez toujours soutenu, financièrement et mentalement, mais surtout encouragé à travers toutes mes études. Vous savez à quel point ce travail représente beaucoup pour moi. Il représente surtout le dénouement et l'achèvement de toutes ces études qui, comme vous le savez, m'ont demandé tellement d'efforts, de travail et de sacrifices. Les valeurs de persévérance et d'abnégation que vous m'avez inculquées sont inestimables à mes yeux et font de moi la personne que je suis aujourd'hui. De la première à la dernière seconde, vous avez été présents. Merci infiniment ! Je remercie également ma sœur, qui m'a toujours soutenue et

encouragée à continuer mes études dans le domaine des mathématiques. Sache que je t'en suis extrêmement reconnaissant.

Merci également à mes amis qui sont, pour la plupart, dans le monde du soccer, et qui, sans trop vraiment comprendre sur quoi je travaillais, m'ont permis plus souvent qu'autrement de me libérer l'esprit. Sachez qu'à votre façon, vous m'avez permis de passer à travers ce travail qui m'a demandé beaucoup. Merci à Alex, Dave, David, Nicholas, Nicolas, David et William, pour ne nommer que ceux-là.

Un merci tout spécial à mes collègues avec qui j'ai passé de très belles années universitaires. Ça toujours été un réel plaisir de pouvoir vous côtoyer et d'échanger. Merci à toute l'équipe d'administration et du secrétariat mais aussi et surtout à l'équipe du soutien informatique. Vous avez rendu mes études universitaires beaucoup plus simples. Un dernier remerciement s'adresse aux membres du jury qui ont pris le temps de lire ce travail qui représente beaucoup pour moi.

Introduction

Plusieurs méthodes existent autant dans un contexte classique d'une distribution binomiale que dans un contexte d'échantillonnage avec une population finie afin d'estimer par intervalle de confiance un paramètre de proportion. Il y a d'abord une méthode générale qui est utilisée dans plusieurs contextes différents, la méthode de Wald. Cette méthode, qui est aussi appelée l'intervalle standard, additionne et soustrait à l'estimateur la marge d'erreur qui est habituellement le quantile $\alpha/2$ d'une loi normale de moyenne 0 et de variance égale à 1 multiplié par l'estimation de l'erreur standard. Plusieurs autres méthodes spécialisées ont été introduites pour tenir compte des spécificités de l'estimation d'une proportion. La méthode logit consiste essentiellement à inverser un intervalle de type Wald pour la transformation logit de p avant de remettre les bornes de l'intervalle sur l'échelle originale. La méthode de Wilson utilise l'erreur standard exacte basée sur p au lieu de l'erreur standard estimée basée sur \hat{p} utilisée pour standardiser la variable aléatoire $\hat{p} - p$ dans la construction de l'intervalle de Wald. L'intervalle Agresti-Coull utilise un intervalle de type Wald, mais en utilisant un estimateur inspiré de celui sur lequel est centré l'intervalle de confiance de Wilson, soit la proportion de succès dans l'échantillon où l'on a ajouté deux succès et deux échecs. Il y a également la méthode Clopper-Pearson qui considère, dans la construction des bornes de l'intervalle, une fonction incluant les degrés de liberté d'une loi de Fisher ainsi que ses quantiles. Finalement, la méthode Arcsine Square Root utilise essentiellement $\arcsin(\sqrt{p})$ comme transformation stabilisatrice de la variance de p (voir [Brown *et al.*, 2001]). Ces différentes méthodes ont été introduites parce qu'il avait été remarqué que la méthode générale de Wald était moins efficace pour p prenant des valeurs près de 0 et 1 ([Brown *et al.*, 2001]). En effet, lorsque la proportion p est près des extrémités, l'approximation de la binomiale par une loi normale n'est pas très bonne. Comme l'intervalle de Wald, une autre approche générale pour construire un intervalle de confiance est la méthode bootstrap.

Le bootstrap a été introduit par [Efron, 1979] dans un contexte de statistique classique et est essentiellement basé sur la réplication des observations de l'échantillon initial, basé sur la simulation, dans le but d'estimer la distribution d'un estimateur, notamment dans le but de construire des intervalles de confiance. Dans un contexte de statistique classique, l'échantillon initial provient de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) d'une loi de probabilité F inconnue. Dans le cas qui nous concerne, lorsque le paramètre d'intérêt est un paramètre de proportion, chacune des observations de l'échantillon initial est générée indépendamment selon une loi de Bernoulli(p) qui est la loi F inconnue. Le bootstrap consiste à estimer la distribution inconnue F par un estimateur \hat{F} et à calculer les estimateurs sur des échantillons bootstrap générés à partir de la distribution connue \hat{F} . Pour estimer F , il est possible d'utiliser une approche paramétrique qui consiste essentiellement à supposer que la loi F fait partie d'une famille de lois paramétriques et donc que $F \equiv F_p$, avec p comme paramètre de la loi F . Il est commun d'estimer un paramètre de proportion p par son estimateur naturel $\hat{p} = X/n$ qui représente la proportion du nombre de succès X parmi l'échantillon initial de taille n . De ce fait, il est habituel d'estimer la loi de F_p par $F_{\hat{p}}$. La méthode bootstrap suggère de générer ensuite des échantillons bootstrap selon la loi de \hat{F} qui dans ce cas-ci est égale à la loi de $F_{\hat{p}}$. À partir de ces échantillons bootstrap, nous sommes en mesure d'estimer la distribution de l'estimateur \hat{p} . Il nous est ensuite possible de créer des intervalles de confiance bootstrap pour p . Ces intervalles sont essentiellement basés sur la distribution de $\hat{p}^* - \hat{p}$ où \hat{p}^* représente l'estimateur bootstrap soit la proportion de succès dans l'échantillon bootstrap. Bien que nous soyons en mesure de calculer des intervalles de confiance bootstrap à partir de cette méthode, nous rencontrons des difficultés lorsque \hat{p} , l'estimateur calculé à partir de l'échantillon initial, est nul ou prend la valeur de 1.

Prenons le cas où \hat{p} est nul puisque le cas où \hat{p} est égal à 1 mène à la même problématique. Lorsque cette situation survient, comme nous estimons la loi de F_p par $F_{\hat{p}}$, les échantillons bootstrap deviennent inévitablement des échantillons ne contenant que des échecs puisque $F_{\hat{p}}$ a une probabilité de succès qui est nulle. L'estimateur bootstrap, \hat{p}^* , devient nul avec probabilité 1 et les intervalles de confiance bootstrap basés sur la distribution de $\hat{p}^* - \hat{p}$ ne deviennent qu'un seul point, soit le point 0. Ceci suggère donc qu'on est certain de la valeur de p . L'intervalle de Wald, qui est une méthode générale elle aussi, devient également dégénérée lorsque l'estimateur \hat{p} devient nul ou égal à 1. Nous devons donc trouver une

alternative à la méthode bootstrap classique pour tenir compte de la spécificité du problème lorsqu'il est question d'un paramètre de proportion. La solution nous servira afin d'estimer la loi de F et ainsi pouvoir contourner les difficultés mentionnées. Ainsi, nous serons en mesure de créer des intervalles de confiance qui seront autre chose qu'un point unique. Il va de soi donc de trouver une solution qui permettra la construction d'intervalles de confiance bootstrap lorsque \hat{p} devient 0 ou 1 et qui permettra de compétitionner avantageusement avec les méthodes traditionnelles utilisées pour répondre à la spécificité d'un estimateur de proportion.

L'une des solutions possibles qu'on introduit dans le cadre de ce mémoire consiste à estimer la loi de F_p par une loi $F_{\tilde{p}}$ où \tilde{p} est défini comme un estimateur qui ne peut en aucun cas être nul ou égal à 1. L'estimateur qu'on utilise est la proportion de succès calculée sur un échantillon augmenté, soit l'échantillon initial auquel on ajoute deux succès et deux échecs. On peut définir cet estimateur par $\tilde{p} = (X + 2)/(n + 4)$ où X est le nombre de succès dans l'échantillon initial de taille n . Cet estimateur est très comparable à la forme de l'estimateur utilisé dans l'intervalle de confiance des méthodes traditionnelles de Wilson et d'Agresti-Coull. [Agresti et Coull, 1998] ont employé cet estimateur dans leur proposition car ils avaient remarqué l'inefficacité de l'intervalle standard surtout lorsque p prend des valeurs près de 0 ou 1. [Agresti et Coull, 1998] se sont inspirés de la méthode de construction de l'intervalle de confiance de [Wilson, 1927]. Celui-ci a introduit cet estimateur dans un contexte de statistique classique. L'estimateur de proportion qu'a introduit Wilson est défini comme $(X + z_{\alpha/2}^2/2)/(n + z_{\alpha/2}^2)$ où $z_{\alpha/2}$ est le quantile $\alpha/2$ d'une loi normale de moyenne 0 et variance égale à 1. Dans le cas où nous désirons construire des intervalles de confiance bilatéraux de niveau 95%, la quantité $z_{\alpha/2}^2 \approx 4$.

Dans un contexte de statistique classique, la solution présentée dans ce mémoire consiste donc à créer des intervalles de confiance bootstrap basés sur l'estimation de la distribution de l'estimateur \hat{p} mais en effectuant le bootstrap en générant les observations à partir de la distribution $F_{\tilde{p}}$ où $\tilde{p} = (X + 2)/(n + 4)$. L'estimateur \tilde{p} qui consiste à calculer la proportion de succès après avoir ajouté deux succès et deux échecs à l'échantillon initial permet donc de ne plus avoir d'échantillons bootstrap nuls avec probabilité 1.

Nous traiterons également d'un contexte considérant une population finie où l'échantillon initial est tiré de manière aléatoire simple sans remise à partir de cette population

finie. La plupart des méthodes traditionnelles présentées précédemment ont été adaptées pour un contexte de population finie. En effet, il est important de prendre en considération le facteur de la fraction de sondage lorsque vient le temps de construire des intervalles de confiance en contexte d'échantillonnage. L'une des approches utilisées entre autres par [Franco *et al.*, 2019] est d'utiliser la taille effective à la place de la taille échantillonnale dans la construction des intervalles. [Korn et Graubard, 1998] a aussi introduit cette notion de taille effective. En effet, ils ont étudié et constaté le comportement assez décevant de la méthode de Wald dans un contexte d'échantillonnage lorsque p prend des valeurs près de 0 et 1.

Pour utiliser le bootstrap dans un contexte de population finie, des modifications ont dû être apportées afin de tenir compte de l'impact du plan de sondage sur la distribution des estimateurs. [Mashreghi *et al.*, 2016] ont d'ailleurs introduit plusieurs méthodes bootstrap pour un contexte d'échantillonnage. Dans un contexte de population finie, c'est la population qui joue le rôle de la distribution inconnue F_p à estimer et c'est à partir de l'échantillon que l'on considérera qu'on effectue le bootstrap. Si nous n'avons essentiellement qu'une seule manière de faire le bootstrap dans le cas i.i.d., dans un contexte d'échantillonnage, plusieurs méthodes existent.

Dans le cadre de ce mémoire, nous nous concentrerons sur la méthode utilisant une pseudo-population de [Booth *et al.*, 1994] ainsi que sur la méthode utilisant des poids bootstrap de [Rao *et al.*, 1992]. La méthode par pseudo-population de [Booth *et al.*, 1994] consiste essentiellement à répliquer l'échantillon initial afin d'y créer une pseudo-population à partir de laquelle nous tirons un échantillon bootstrap selon le plan de sondage. Dans la méthode basée sur les poids de [Rao *et al.*, 1992] on utilise le fait que l'estimateur étudié soit représenté par une somme pondérée des observations de l'échantillon et au lieu de créer de nouveaux échantillons, on crée plutôt de nouveaux poids, les poids bootstrap, qu'on applique aux observations de l'échantillon original. Il est facile de se convaincre qu'un échantillon ne contenant que des succès ou des échecs mènera à la même problématique que celle décrite dans le cas des échantillons i.i.d. lorsque nous appliquerons l'une ou l'autre de ces formes de bootstrap. Prenons l'exemple d'un échantillon s de la population qui ne contient que des échecs. Si nous décidons d'utiliser la méthode par pseudo-population de [Booth *et al.*, 1994], la pseudo-population étant composée d'un

certain nombre de répétitions de chaque observation de l'échantillon original, ne contiendra que des échecs. Ainsi tout échantillon bootstrap ne sera composé que d'échecs. L'estimateur bootstrap \hat{p}^* calculé sera inévitablement nul avec probabilité 1 et nous nous retrouverons avec les mêmes problèmes que ceux mentionnés précédemment dans un contexte de statistique classique, c'est-à-dire que les intervalles bootstrap ne contiendront qu'un seul point, soit le point 0. Avec un raisonnement similaire, il est facile de comprendre que peu importe comment sont générés les poids bootstrap de la méthode de [Rao *et al.*, 1992], si nous considérons l'échantillon s et que celui-ci ne contient que des échecs, peu importe les poids bootstrap générés, l'estimateur bootstrap calculé sera nul avec probabilité 1 et les intervalles bootstrap basés sur la distribution de $\hat{p}^* - \hat{p}$ ne contiennent que le point 0.

Cela étant, si dans un contexte i.i.d., la solution proposée consiste à estimer la loi de F_p par $F_{\hat{p}}$, dans un contexte de population finie, cette même solution consiste à considérer un échantillon modifié où on ajoute 4 observations à l'échantillon initial s soit 2 succès et 2 échecs. En effet, en créant une pseudo-population en se basant sur un échantillon modifié, dénoté s^+ , et qui a les propriétés mentionnées précédemment, la pseudo-population créée ne contiendra jamais que des échecs ou que des succès et donc l'échantillon bootstrap que l'on tirera de cette pseudo-population ne sera pas nul avec probabilité 1. De plus, pour la méthode par les poids, en choisissant la taille de l'échantillon bootstrap telle qu'une inégalité soit respectée (ce qui est toujours possible), nous pouvons nous assurer que l'estimateur bootstrap ne soit jamais nul en considérant un échantillon de type s^+ plutôt que l'échantillon original s .

La solution proposée autant dans un contexte de statistique classique que dans un contexte de population finie, consiste donc à utiliser la loi de $F_{\hat{p}}$ ou de considérer un échantillon modifié s^+ en tout temps. Toutefois, comme les probabilités de générer des échantillons bootstrap n'ayant que des échecs ou que des succès sont plus fortes lorsque la vraie valeur de p se situe très près de 0 ou 1 respectivement, il pourrait être envisageable de ne considérer l'application de cette solution que lorsque l'estimateur \hat{p} est nul ou égal à 1. Lorsque notre estimateur est autre chose que 0 ou 1, nous pourrions penser qu'il est possible de conserver la méthodologie initiale, c'est-à-dire d'estimer la loi de F_p par $F_{\hat{p}}$ dans un contexte de statistique classique ou de considérer l'échantillon s dans un cadre d'échantillonnage avec une population finie dans l'utilisation des méthodes par pseudo-population

et par les poids. Comme l'estimateur \tilde{p} prend la forme proposée par l'estimateur utilisé dans la méthode d'intervalle de confiance d'Agresti-Coull, nous utiliserons la nomenclature d'un bootstrap à la Agresti-Coull lorsqu'on appliquera la solution mentionnée en tout temps. Par contre, lorsqu'on utilisera cette solution que pour les situations où \hat{p} est 0 ou 1, nous nommerons cette approche par un bootstrap adaptatif.

Au chapitre 1, nous présenterons la manière dont sont générées les données selon une loi binomiale avant de présenter le fonctionnement général de la méthode bootstrap d'[Efron, 1979]. Nous discuterons des intervalles de confiance bootstrap et de quelle manière ceux-ci sont modifiés en développant la solution mentionnée ci-dessus. Nous nous intéresserons au contexte d'une population finie au chapitre suivant où l'on débutera par présenter quelques concepts reliés à l'échantillonnage avant de poursuivre en présentant en détail le fonctionnement des méthodes par pseudo-population de [Booth *et al.*, 1994] et la méthode par les poids de [Rao *et al.*, 1992]. Nous verrons finalement comment ces méthodes sont ajustées afin de considérer la nouvelle solution présentée ainsi que les intervalles de confiance bootstrap à utiliser pour les deux méthodes, soit le bootstrap à la Agresti-Coull et le bootstrap adaptatif. Il sera important de comparer l'efficacité de ces méthodes bootstrap à celles des méthodes plus traditionnelles qui n'utilisent pas le bootstrap. Au chapitre 3, nous présenterons donc les six méthodes traditionnelles que nous considérons dans le cadre de ce mémoire soit les méthodes de Wald, de Logit, de Wilson, d'Agresti-Coull, de Clopper-Pearson et Arcsine Square Root. C'est au chapitre 4 que nous présenterons les résultats de nos différentes simulations tout d'abord dans un contexte i.i.d. puis dans un contexte d'échantillonnage avec l'utilisation du plan aléatoire simple sans remise. Il sera intéressant de voir à ce chapitre l'efficacité du bootstrap à la Agresti-Coull, du bootstrap adaptatif et des meilleures méthodes bootstrap face aux méthodes traditionnelles les plus efficaces avant de conclure.

Chapitre 1

Méthodes bootstrap dans un contexte i.i.d.

Dans ce chapitre, il sera question du contexte où les données sont générées indépendamment les unes des autres selon une même loi. Nous présenterons tout d'abord de quelle manière les données sont générées en plus de décrire le fonctionnement de la méthode bootstrap de [Efron, 1979]. Nous énoncerons ensuite la problématique rencontrée lors de la création d'intervalles de confiance bootstrap lorsque l'estimateur de proportion devient nul tout en présentant la solution afin de contourner ces difficultés. Nous verrons finalement comment cette solution s'applique dans la construction des intervalles de confiance bootstrap.

1.1. Contexte i.i.d.

Afin d'estimer un paramètre de proportion p , l'estimateur habituel est \hat{p} , la proportion de succès de l'échantillon. En plus d'être un estimateur sans biais du paramètre de proportion, il est l'estimateur avec la plus petite variance parmi tous les estimateurs de proportion sans biais existants. C'est pourquoi cet estimateur est aussi populaire et utilisé. Considérons le contexte où les indicatrices de succès Z_i sont indépendantes et identiquement distribuées (i.i.d.) selon une loi Bernoulli(p) de telle sorte que nous pouvons écrire

$$\mathbf{Z} = (Z_1, Z_2, \dots, Z_n) \sim i.i.d \text{ Bernoulli}(p) \quad (1.1.1)$$

qui représente l'échantillon initial dans un contexte i.i.d. Les valeurs Z_i sont donc définies ainsi :

$$Z_i = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } 1 - p. \end{cases} \quad (1.1.2)$$

Afin de simplifier les choses, nous noterons dans un contexte i.i.d. que

$$X = \sum_{i=1}^n Z_i \quad (1.1.3)$$

qui représente le nombre de succès dans l'échantillon initial \mathbf{Z} . Nous pouvons donc noter

$$\hat{p} = \frac{X}{n}. \quad (1.1.4)$$

Comme la variance de \hat{p} est égale à $V(\hat{p}) = p(1-p)/n$, la variance de \hat{p} peut être estimée par l'estimateur habituel de sa variance soit

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}. \quad (1.1.5)$$

Nous désirons estimer la distribution de \hat{p} , entre autres pour construire un intervalle de confiance pour le paramètre p . Afin de construire des intervalles de confiance pour le paramètre p , nous pouvons soit utiliser la distribution de $\hat{p} - p$ ou employer la distribution de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$, où $\check{V}(\hat{p})$ est un estimateur quelconque de la variance de \hat{p} . Nous utiliserons le bootstrap afin de créer ces intervalles de confiance qui seront définis plus loin. Définissons tout d'abord comment fonctionne la méthode bootstrap de [Efron, 1979].

Supposons maintenant une loi de probabilité inconnue F avec θ comme paramètre telle que nous obtenons l'échantillon $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n) \sim \text{i.i.d. } F$. L'estimateur $\hat{\theta}$, calculé à partir de \mathbf{Z} , est utilisé pour estimer θ . De plus, nous estimons la loi de F par \hat{F} , un estimateur quelconque de F et nous tirons ensuite des échantillons $\mathbf{Z}^* = (Z_1^*, Z_2^*, \dots, Z_n^*)$ i.i.d. de la distribution \hat{F} . Nous verrons bientôt deux approches pour estimer F . Nous répétons ensuite le tirage de \mathbf{Z}^* un grand nombre de fois, B . Nous dénotons par \mathbf{Z}_j^* le j^e échantillon bootstrap, $j = 1, \dots, B$ où $\mathbf{Z}_j^* = (Z_{j1}^*, Z_{j2}^*, \dots, Z_{jn}^*)$ et Z_{ji}^* représente la i^e observation bootstrap. C'est donc à partir des B échantillons bootstrap \mathbf{Z}^* que nous sommes en mesure de calculer nos différents estimateurs bootstrap, $\hat{\theta}^*$, calculés à partir de \mathbf{Z}^* . De plus, il est possible de calculer $\check{V}^*(\hat{\theta}^*)$, un estimateur de la variance du j^e estimateur bootstrap. L'une des manières est d'utiliser un second bootstrap afin de calculer cet estimateur. Nous générons B échantillons \mathbf{Z}_j^{**} i.i.d. \hat{F}_j^* , l'estimateur de \hat{F} . À partir des B échantillons \mathbf{Z}_{jk}^{**} , $k = 1, \dots, B$, nous calculons l'estimateur bootstrap dénoté $\hat{\theta}_{jk}^{**}$ où la moyenne des B estimateurs bootstrap est dénoté $\bar{\theta}^{**}$. Ainsi, un estimateur de la variance de l'estimateur bootstrap $\hat{\theta}_j^*$ peut prendre la forme

$$\hat{V}^*(\hat{\theta}_j^*) = V^{**}(\hat{\theta}^{**}) = \frac{1}{B-1} \sum_{k=1}^B (\hat{\theta}_{jk}^{**} - \bar{\theta}^{**})^2. \quad (1.1.6)$$

Il est aussi possible d'estimer la variance de l'estimateur bootstrap en employant la formule d'estimation de variance du paramètre d'intérêt. Un estimateur de la variance de $\hat{\theta}$ est défini par $\check{V}(\hat{\theta})$ et peut prendre la forme

$$\hat{V}(\hat{\theta}) = V^*(\hat{\theta}^*) = \frac{1}{B-1} \sum_{i=1}^B (\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2 \quad (1.1.7)$$

où $\bar{\hat{\theta}}^* = B^{-1} \sum_i \hat{\theta}_i^*$.

Voici comment fonctionne l'algorithme bootstrap de [Efron, 1979] :

- (1) Soit notre échantillon initial $\mathbf{Z} \sim \text{i.i.d. } F$ avec \hat{F} son estimateur. On calcule $\hat{\theta}$, l'estimateur calculé au moyen de \mathbf{Z} .
- (2) Pour $j = 1, \dots, B$:
 - On génère $\mathbf{Z}_j^* = (Z_{j1}^*, Z_{j2}^*, \dots, Z_{jn}^*) \sim \text{i.i.d. } \hat{F}$.
 - On calcule l'estimateur bootstrap $\hat{\theta}_j^*$.
 - On calcule l'estimateur de la variance sur l'échantillon bootstrap, $\check{V}^*(\hat{\theta}_j^*)$.
- (3) On calcule ensuite la distribution de $(\hat{\theta}_j^* - \hat{\theta}) / \sqrt{\check{V}^*(\hat{\theta}_j^*)}$ nécessaire à la construction d'intervalles de confiance bootstrap- t .
- (4) On calcule $\check{V}(\hat{\theta})$, un estimateur de la variance de $\hat{\theta}$.

Il existe deux méthodes permettant de construire un estimateur \hat{F} de F . La première consiste à utiliser la fonction de répartition expérimentale \hat{F} , celle qui donne un poids de $1/n$ à chacune des observations Z_i . Tirer un échantillon de taille n à partir de cette méthode non paramétrique équivaut à piger avec remise n observations à partir de $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$. La seconde méthode possible est une méthode paramétrique. Cette méthode suppose que la loi F fait partie d'une famille de lois paramétriques et donc que $F \equiv F_p$ avec p comme paramètre de la loi F . Cette dernière est estimée par $F_{\hat{p}}$ où \hat{p} est un estimateur de p . Dans le cas où F_p est une loi de Bernoulli(p), la fonction de répartition expérimentale et la distribution $F_{\hat{p}}$ avec $\hat{p} = n^{-1} \sum_{i=1}^n Z_i$ sont identiques, ce qui implique donc que les deux méthodes bootstrap sont les mêmes. Nous désirons estimer la distribution de $\hat{p} - p$ et de $(\hat{p} - p) / \sqrt{\check{V}(\hat{p})}$. Rappelons qu'ici, le paramètre p représente la vraie proportion (inconnue) de succès de la distribution F . Avec l'emploi d'un bootstrap paramétrique avec $F_{\hat{p}}$, on estimera ces distributions par $\hat{p}^* - \check{p}$ ou $(\hat{p}^* - \check{p}) / \sqrt{\check{V}^*(\hat{p}^*)}$ où \hat{p}^* représente l'estimateur bootstrap, soit la proportion de succès dans l'échantillon bootstrap, et $\check{V}^*(\hat{p}^*)$ l'estimateur de la variance de \hat{p} calculé à partir

de l'échantillon bootstrap. Comme p représente la vraie proportion associée à la distribution F_p dans la distribution de $\hat{p} - p$, il va donc de soi d'utiliser la vraie proportion \check{p} associée à la distribution $F_{\check{p}}$ dans la version bootstrap pour la distribution de $\hat{p}^* - \check{p}$. De plus, $\check{V}^*(\hat{p}^*)$ est l'estimateur de la variance bootstrap de l'estimateur \hat{p}^* , calculé au moyen de l'échantillon bootstrap \mathbf{Z}^* . Par exemple, si nous utilisons $\hat{V}(\hat{p}) = \hat{p}(1 - \hat{p})/n$ alors $\hat{V}^*(\hat{p}^*) = \hat{p}^*(1 - \hat{p}^*)/n$.

1.2. Problématique et solution proposée

Transportons-nous dans un cas particulier, mais non improbable et qui peut nous occasionner quelques difficultés dans la création d'intervalles de confiance pour p . Imaginons le cas où toutes les observations de l'échantillon initial $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ n'ont que des échecs c'est-à-dire que tous les Z_i sont 0 et donc que $\hat{p} = 0$, ou que tous les Z_i sont égaux à 1 et donc que $\hat{p} = 1$. Sans perte de généralité, nous allons à partir de maintenant seulement parler du cas $\hat{p} = 0$, l'autre cas menant aux mêmes problèmes. Dans cette situation, le bootstrap paramétrique avec $F_{\hat{p}}$ ferait en sorte que Z_i^* serait 0 avec probabilité 1 ce qui impliquerait que $\hat{p}^* = n^{-1} \sum_{i=1}^n Z_i^* = 0$ avec probabilité 1 et ainsi l'estimation de la variance de \hat{p}^* dénotée par $\hat{V}^*(\hat{p}^*) = \hat{p}^*(1 - \hat{p}^*)/n = 0$. Par conséquent, la distribution de la variable aléatoire bootstrap $\hat{p}^* - \hat{p}$ nécessaire à la création d'intervalles de confiance serait concentrée sur la valeur 0 avec probabilité 1, alors que la variable aléatoire bootstrap $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p}^*)}$ ne serait pas définie car autant le numérateur que le dénominateur prennent la valeur 0 avec probabilité 1. Dans ces conditions, les intervalles de confiance basés sur la distribution de $\hat{p}^* - \hat{p}$ sont dégénérés dans le sens où ils ne prennent qu'un point et ceux basés sur la distribution de $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p}^*)}$ ne sont tout simplement pas définis.

L'une des solutions possibles et étudiée ici est l'emploi d'un estimateur ayant la caractéristique de prendre des valeurs différentes de 0 (et par conséquent de 1) avec probabilité 1. Un estimateur satisfaisant cette caractéristique est l'estimateur présenté par [Agresti et Coull, 1998] que l'on dénotera par \tilde{p} et qui prend la forme

$$\tilde{p} = \frac{X + 2}{n + 4}. \quad (1.2.1)$$

Cet estimateur est équivalent à ajouter 4 individus dans l'échantillon, dont 2 d'entre eux possèdent la caractéristique étudiée, et de considérer la proportion de succès de cet échantillon modifié. Même s'il est biaisé, cet estimateur ne peut en aucun moment être nul, ce qui ne nous mènera pas à des valeurs de \hat{p}^* égales à 0 avec probabilité 1, et nous aidera donc dans

la création d'intervalles de confiance. Nous considérerons le bootstrap paramétrique avec la distribution de $F_{\tilde{p}}$. Il peut aussi être utile de définir un estimateur de la variance de \hat{p} basé sur \tilde{p} avec $\tilde{V}(\hat{p}) = \tilde{p}(1 - \tilde{p})/n$. Il sera donc intéressant de constater, à l'aide de simulations, à quel point le biais de cet estimateur nous pose problème ou non comparativement à l'efficacité qu'il démontre en terme de couverture du vrai paramètre de proportion.

1.3. Intervalles de confiance bootstrap

Afin de présenter des intervalles de confiance bootstrap, nous allons d'abord introduire des intervalles de confiance potentiellement exacts basés sur les distributions de $\hat{p} - p$ et $(\hat{p} - p)/\sqrt{\tilde{V}(\hat{p})}$ où $\tilde{V}(\hat{p})$ est un estimateur de la variance de \hat{p} . Soit

$$J_n(x, F_p) = Prob(\hat{p} - p \leq x) \quad (1.3.1)$$

et $J_n^{-1}(\alpha, F_p)$ la fonction de répartition et le quantile α de la distribution de $\hat{p} - p$ lorsque la distribution des n observations i.i.d. est F_p . Nous pouvons montrer que si la distribution de $J_n(\cdot, F_p)$ était connue (et donc ses quantiles) alors

$$[\hat{p} - J_n^{-1}(1 - \alpha/2, F_p), \hat{p} - J_n^{-1}(\alpha/2, F_p)] \quad (1.3.2)$$

serait un intervalle de confiance exact de niveau $1 - \alpha$ pour la proportion inconnue p . (À cause du caractère discret de \hat{p} , l'intervalle ne peut être exact que pour certaines valeurs de α , alors que pour les autres valeurs, l'intervalle sera approximativement de niveau $1 - \alpha$. Ce commentaire s'applique également aux prochains intervalles « exacts »). Une alternative au dernier intervalle de confiance consiste à inverser les quantiles et le signe devant ceux-ci menant à l'intervalle

$$[\hat{p} + J_n^{-1}(\alpha/2, F_p), \hat{p} + J_n^{-1}(1 - \alpha/2, F_p)]. \quad (1.3.3)$$

Si la distribution $J_n(\cdot, F_p)$ n'est pas symétrique par rapport à 0, ce qui est le cas ici (sauf pour $p = 1/2$), l'intervalle de confiance n'est pas exact. Mais comme la distribution de $\hat{p} - p$ converge vers une distribution symétrique autour de 0 lorsque n tend vers l'infini, les quantiles $J_n^{-1}(\alpha/2, F_p)$ et $-J_n^{-1}(1 - \alpha/2, F_p)$ se rapprochent de telle sorte qu'asymptotiquement cet intervalle aura une probabilité de couverture de $1 - \alpha$. De manière similaire, nous pouvons poser

$$K_n(x, F_p) = Prob\left([\hat{p} - p]/\sqrt{\tilde{V}(\hat{p})} \leq x\right) \quad (1.3.4)$$

et $K_n^{-1}(\alpha, F_p)$ la fonction de répartition et le quantile α de la distribution de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$. Nous faisons l'hypothèse ici que $\check{V}(\hat{p})$ est non nul afin d'assurer que K_n soit toujours défini. On peut aussi montrer que si $K_n(\cdot, F_p)$ était connue (et donc ses quantiles) alors

$$\left[\hat{p} - K_n^{-1}(1 - \alpha/2, F_p)\sqrt{\check{V}(\hat{p})}, \hat{p} - K_n^{-1}(\alpha/2, F_p)\sqrt{\check{V}(\hat{p})} \right] \quad (1.3.5)$$

serait un intervalle de confiance exact de niveau $1 - \alpha$ pour la proportion inconnue p . Comme les distributions $J_n(\cdot, F_p)$ ainsi que $K_n(\cdot, F_p)$ dépendent de la vraie proportion p , inconnue, nous ne pouvons pas bâtir ces intervalles de confiance en pratique. Il existe des solutions bootstrap afin de pouvoir bâtir ces intervalles de confiance. Comme nous estimons la loi de F par \hat{F} , par exemple par $F_{\hat{p}}$ ou $F_{\check{p}}$, nous pouvons estimer les quantiles $J_n^{-1}(\alpha, F_p)$ et $K_n^{-1}(\alpha, F_p)$ par les quantiles $J_n^{-1}(\alpha, \hat{F})$ et $K_n^{-1}(\alpha, \hat{F})$. De plus, il faudra déterminer l'estimateur de la variance de \hat{p} , utilisé dans le calcul des quantiles de la distribution de K_n relié à son intervalle de confiance, soit quelle version de $\check{V}(\hat{p})$ sera utilisée. Le premier intervalle de confiance présenté soit l'intervalle (1.3.2) devient

$$[\hat{p} - J_n^{-1}(1 - \alpha/2, \hat{F}), \hat{p} - J_n^{-1}(\alpha/2, \hat{F})]. \quad (1.3.6)$$

C'est l'intervalle de confiance bootstrap de base de niveau $1 - \alpha$. Le second intervalle est

$$[\hat{p} + J_n^{-1}(\alpha/2, \hat{F}), \hat{p} + J_n^{-1}(1 - \alpha/2, \hat{F})] \quad (1.3.7)$$

et est l'estimation de l'intervalle exact défini en (1.3.3). C'est ce qu'on appelle l'intervalle de confiance bootstrap percentile de niveau $1 - \alpha$. Notons que si on utilise un bootstrap paramétrique et qu'on utilise la loi $F_{\check{p}}$ avec $\check{p} = \hat{p}$, alors $J_n^{-1}(\alpha, F_{\hat{p}}) = L_n^{-1}(\alpha, F_{\hat{p}}) - \hat{p}$ où $L_n(x, F_{\hat{p}}) = Prob(\hat{p}^* \leq x)$ de telle sorte que la borne $\hat{p} + J_n^{-1}(\alpha, F_{\hat{p}})$ est égale à $L_n^{-1}(\alpha, F_{\hat{p}})$ et ce second intervalle de confiance est équivalent à

$$[L_n^{-1}(\alpha/2, F_{\hat{p}}), L_n^{-1}(1 - \alpha/2, F_{\hat{p}})]. \quad (1.3.8)$$

Cette équivalence entre les deux intervalles ne tient que pour le bootstrap paramétrique fait à partir de $F_{\hat{p}}$. Comme les usagers habitués aux intervalles de confiance bootstrap utilisent souvent l'intervalle de confiance percentile défini par l'intervalle (1.3.8), nous allons également considérer cet intervalle de confiance, même dans le cas où nous utiliserons un estimateur \tilde{p} différent de \hat{p} et nous l'appellerons l'intervalle de confiance bootstrap percentile *classique*. C'est dans leur revue des méthodes bootstrap en échantillonnage où le paramètre de la

distribution bootstrap est parfois différent de l'estimateur calculé sur les données originales, que [Mashreghi et al., 2016] ont distingué l'intervalle percentile calculé via la formule (1.3.7) de l'intervalle percentile classique calculé via la formule (1.3.8). Le quatrième type d'intervalle est

$$\left[\hat{p} - K_n^{-1}(1 - \alpha/2, \hat{F})\sqrt{\check{V}(\hat{p})}, \hat{p} - K_n^{-1}(\alpha/2, \hat{F})\sqrt{\check{V}(\hat{p})} \right] \quad (1.3.9)$$

où $K_n(x, \hat{F}) = Prob([\hat{p}^* - \check{p}]/\sqrt{\check{V}^*(\hat{p}^*)} \leq x)$ et $\check{V}^*(\hat{p}^*)$ est l'estimation de la variance bootstrap de \hat{p}^* où \check{p} est l'estimateur utilisé dans le bootstrap paramétrique via $F_{\check{p}}$ ou encore \hat{p} si \hat{F} est la fonction de répartition expérimentale. L'intervalle de confiance présenté en (1.3.9) est ce qu'on appelle l'intervalle de confiance bootstrap- t de niveau $1 - \alpha$.

Tel qu'indiqué précédemment, si nous considérons un bootstrap paramétrique avec $F_{\check{p}}$ et que l'on prend $\check{p} = \hat{p}$ menant à $F_{\hat{p}}$ et que \hat{p} est égal à 0 alors la distribution $J_n(\cdot, F_{\hat{p}})$ est la distribution de masse 1 sur la valeur 0 et les intervalles bootstrap de base (1.3.6) percentile (1.3.7) et percentile classique (1.3.8) ne contiennent que le point 0, ce qui suggère que nous sommes certains de la valeur de p . Également, si on utilisait $\check{V}(\hat{p}) = \hat{V}(\hat{p}) = \hat{p}(1 - \hat{p})/n$ alors la variable aléatoire $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p})}$ dans la définition de $K_n(\cdot, \hat{F})$ ne serait pas définie avec probabilité 1 puisque tant le numérateur que le dénominateur seraient 0 avec probabilité 1. L'intervalle de confiance bootstrap- t n'est donc pas calculable. Si nous faisons le bootstrap paramétrique à partir de $F_{\hat{p}}$ en utilisant plutôt $\check{V}(\hat{p}) = \tilde{V}(\hat{p}) = \tilde{p}(1 - \tilde{p})/n$ alors lorsque $\hat{p} = 0$, $(\hat{p}^* - \hat{p})/\sqrt{\tilde{V}^*(\hat{p})}$ serait 0 avec probabilité 1 puisque le numérateur serait 0 avec probabilité 1 alors que le dénominateur serait différent de 0 avec probabilité 1. Mais comme tous les quantiles de $K_n(\cdot, F_{\hat{p}})$ seraient égaux à 0, l'intervalle de confiance bootstrap- t ne contiendrait à nouveau que le point 0.

Nous allons considérer deux estimateurs pour p dans le bootstrap pour la création d'intervalles de confiance. Le premier a été préalablement défini en (1.2.1) et est l'estimateur de [Agresti et Coull, 1998] dénoté précédemment par \tilde{p} . Nous considérerons un second estimateur de p qui sera dénoté par \hat{p} (à ne pas confondre avec \hat{p}) et qui prendra la forme

$$\hat{p} = \begin{cases} \tilde{p} & \text{si } \hat{p} = 0 \text{ ou } \hat{p} = 1 \\ \hat{p} & \text{sinon.} \end{cases} \quad (1.3.10)$$

L'estimateur \hat{p} permettra de distinguer deux situations différentes. Si l'estimateur \tilde{p} s'applique à tout échantillon, peu importe la valeur de l'estimateur habituel \hat{p} , l'estimateur \hat{p} distingue deux situations, soit lorsque \hat{p} est 0 (ou \hat{p} est 1) et lorsqu'il prend toute autre valeur. L'intérêt d'utiliser l'estimateur \tilde{p} est de ne pas générer des valeurs \hat{p}^* égales à 0 avec probabilité 1. Comme les risques de se confronter à cette situation sont plus élevés lorsque l'estimateur habituel \hat{p} est 0, il pourrait être correct d'utiliser \hat{p} si sa valeur n'est pas nulle. Par contre, peu importe que les données soient générées selon la distribution de $F_{\tilde{p}}$ ou $F_{\hat{p}}$, nous définissons l'estimateur de variance de \hat{p} par $\tilde{V}(\hat{p}) = \tilde{p}(1 - \tilde{p})/n$ dans la construction des intervalles bootstrap- t , et ce tant dans la définition de la variable aléatoire dont K_n représente la distribution que dans la construction de l'erreur standard calculée sur l'échantillon original qui multiplie le quantile bootstrap de K_n . Chacun de ces deux estimateurs mène donc à 4 intervalles de confiance bootstrap pour p . Pour simplifier la nomenclature des approches bootstrap dans la simulation, nous allons dénoter le bootstrap effectué avec la distribution de $F_{\tilde{p}}$ par le bootstrap à la Agresti-Coull, voir [**Brown et al., 2001**], puis nous dénoterons par bootstrap adaptatif, le bootstrap effectué selon la distribution de $F_{\hat{p}}$.

Chapitre 2

Méthodes bootstrap en échantillonnage

Nous avons présenté jusqu'ici les méthodes bootstrap reliées à un contexte i.i.d., c'est-à-dire où les données sont générées de manière indépendante les unes des autres selon une même loi. Comme les méthodes présentées au chapitre précédent ne sont pas applicables pour des échantillons pigés sans remise à partir d'une population finie, nous discuterons dans ce chapitre des méthodes bootstrap utilisées dans le cas où un échantillon est pris d'une population finie à partir d'un plan de sondage. Si dans un contexte i.i.d. le bootstrap ne s'effectue essentiellement que d'une seule manière, il y a plusieurs approches possibles afin d'appliquer le bootstrap dans le cas d'une population finie. En plus de présenter quelques concepts utiles sur la théorie de l'échantillonnage, nous présenterons la méthode bootstrap utilisant une pseudo-population de [Booth *et al.*, 1994] et la méthode bootstrap de [Rao *et al.*, 1992] qui utilise les poids des observations dans l'estimation du paramètre d'intérêt. Au chapitre précédent, nous avons introduit une solution dans le contexte i.i.d. pour le problème de la construction d'un intervalle de confiance pour la proportion p lorsque la proportion dans l'échantillon \hat{p} est nulle. Nous verrons comment modifier les méthodes de [Booth *et al.*, 1994] et [Rao *et al.*, 1992] afin d'obtenir une solution similaire dans le cas d'un échantillon d'une population finie. Finalement, nous verrons comment ces modifications se reflètent au niveau des intervalles de confiance bootstrap.

2.1. Échantillonnage

Si au chapitre précédent, c'était l'échantillon \mathbf{Z} (1.1.1) qui était généré selon une loi de Bernoulli(p), dans un contexte d'échantillonnage, c'est à partir de la population finie que nous tirons un échantillon selon un plan de sondage.

Définissons tout d'abord la population finie considérée, qui est dénotée par \mathbf{U} et qui contient N unités (ou individus). La population finie \mathbf{U} est définie de la manière suivante :

$$\mathbf{U} = \{1, \dots, i, \dots, N\} \quad (2.1.1)$$

où le i^e individu est représenté par l'étiquette i avec $i = 1, \dots, N$. Même si dans l'étude qui nous concerne nous connaissons N , en pratique, pour plusieurs populations la valeur de N est inconnue.

Dans ce mémoire, nous nous intéressons à la proportion des individus de la population finie possédant la caractéristique étudiée. Définissons donc par Z une variable réponse de type dichotomique et soit Z_i la valeur de Z associée à l'individu i dans la population. La valeur Z_i sera donc égale à 1 pour les individus possédant la caractéristique étudiée et 0 pour ceux qui ne possèdent pas cette même caractéristique.

Comme nous l'avons mentionné plus tôt, nous désirons estimer la proportion d'individus de la population, dénotée p , possédant la caractéristique étudiée. Par conséquent, la population \mathbf{U} a une proportion $p = N^{-1} \sum_{i \in \mathbf{U}} Z_i$ d'individus possédant la caractéristique étudiée. C'est donc dire que pN individus de \mathbf{U} possèdent cette caractéristique et que les $N - pN$ autres individus de \mathbf{U} ne la possèdent pas.

C'est donc à partir de la population finie \mathbf{U} que nous tirons un échantillon, dénoté \mathbf{s} , selon un plan de sondage. L'appartenance des individus à l'échantillon \mathbf{s} peut être défini par le vecteur des indicatrices

$$(I_1, \dots, I_i, \dots, I_N) \quad (2.1.2)$$

où

$$I_i = \begin{cases} 1 & \text{si } i \in \mathbf{s} \\ 0 & \text{sinon.} \end{cases} \quad (2.1.3)$$

Le fait d'appartenir ou non à \mathbf{s} dépend de la réalisation du processus aléatoire servant à tirer l'échantillon. Ce processus aléatoire permettant la sélection de l'échantillon est le plan de sondage. Ce dernier peut être défini comme une fonction $p(\cdot)$ qui assigne à tout échantillon \mathbf{s} , une probabilité de sélectionner l'échantillon en question, $p(\mathbf{s})$. Supposons que l'on désire tirer un échantillon \mathbf{s} de taille n_s selon un plan aléatoire simple sans remise tel que nous

avons $\sum_{i \in \mathbf{U}} I_i = n_s$. On dénombre $\binom{N}{n_s}$ échantillons possibles et nous avons

$$p(\mathbf{s}) = \begin{cases} 1 / \binom{N}{n_s} & \text{si } \mathbf{s} \text{ a } n_s \text{ unités} \\ 0 & \text{sinon.} \end{cases} \quad (2.1.4)$$

Nous pouvons ainsi définir π_i la probabilité que l'individu i soit sélectionné dans l'échantillon \mathbf{s} tel que nous pouvons écrire

$$\pi_i = P(i \in \mathbf{s}) = \sum_{\mathbf{s} \ni i} p(\mathbf{s}). \quad (2.1.5)$$

Dans le cas d'un plan aléatoire simple sans remise où l'on tire un échantillon de taille n_s , chaque unité (ou individu) a une probabilité

$$\pi_i = \frac{n_s}{N} \quad (2.1.6)$$

d'être sélectionnée dans l'échantillon \mathbf{s} . Si nous référons à \mathbf{s} l'ensemble contenant les indices i correspondant aux éléments ou individus de la population présents dans l'échantillon, nous dénoterons par s , de taille n_s , l'ensemble des variables réponses des individus dans l'échantillon :

$$s = \{Z_i\}_{i \in \mathbf{s}}. \quad (2.1.7)$$

Si plusieurs estimateurs sont construits comme une fonction des observations, il est possible de voir l'estimation comme une fonction des poids de chacune des observations de l'échantillon. Plusieurs estimateurs utilisent l'approche des poids de chacune des observations de l'échantillon. L'estimateur le plus commun est sans doute l'estimateur [Horvitz et Thompson, 1952]. Cet estimateur utilise le poids

$$w_i = \frac{1}{\pi_i} = \frac{N}{n_s} \quad (2.1.8)$$

associé à l'individu i au niveau de la population lorsque le plan de sondage utilisé est un plan aléatoire simple sans remise où \mathbf{s} est de taille n_s . Cet estimateur est défini pour toutes les statistiques qui s'expriment comme une fonction du total, comme la moyenne. De plus, il est facile de montrer qu'un estimateur de proportion est un estimateur pour la moyenne, où les valeurs Z_i sont dichotomiques. L'estimateur de proportion Horvitz-Thompson est donc défini ainsi :

$$\hat{p}_{HT} = \frac{1}{N} \sum_{i \in \mathbf{s}} w_i Z_i = \frac{1}{n_s} \sum_{i \in \mathbf{s}} Z_i \quad (2.1.9)$$

si \mathbf{s} , de taille n_s , est tiré de manière aléatoire simple sans remise à partir de \mathbf{U} . Nous utiliserons la notation \hat{p} à partir de maintenant afin de décrire l'estimateur de proportion utilisé dans un cas de population finie et nous le définirons par

$$\hat{p} = \frac{X}{n_s} \quad (2.1.10)$$

où $X = \sum_{i \in \mathbf{s}} Z_i$.

Nous tirons un échantillon selon le plan de sondage $p(\cdot)$ à partir de notre population finie dans le but d'estimer p . Soit $f_s = n_s/N$, la fraction de sondage. À partir de la variance de l'estimateur d'une moyenne pour un plan aléatoire simple sans remise (voir [Särndal *et al.*, 1992]), on peut montrer que la variance de \hat{p} pour ce même plan de sondage est

$$V(\hat{p}) = (1 - f_s) \left(\frac{N}{N - 1} \right) \frac{p(1 - p)}{n_s}. \quad (2.1.11)$$

L'un des estimateurs de la variance de \hat{p} est

$$\hat{V}(\hat{p}) = (1 - f_s) \frac{\hat{p}(1 - \hat{p})}{n_s - 1} \quad (2.1.12)$$

qui est l'estimation de la variance de \hat{p} la plus populaire et utilisée, lorsque le plan de sondage est le plan aléatoire simple sans remise.

Il est nécessaire d'estimer la distribution de \hat{p} dans le but de créer des intervalles de confiance pour p . Dans le contexte i.i.d., le bootstrap nous aide à construire ces intervalles de confiance et consiste à estimer la distribution inconnue F_p des observations, puis à prendre des échantillons i.i.d. de cette distribution estimée avant de calculer l'estimateur bootstrap sur ces données. Bien que l'on puisse utiliser une méthode paramétrique ou non paramétrique afin d'estimer F_p , il n'y a essentiellement qu'une manière d'utiliser le bootstrap. Dans le cas d'une population finie, il est toujours possible de bâtir des intervalles de confiance en utilisant le bootstrap, mais en employant des approches qui diffèrent de celle présentée au chapitre précédent. De plus, contrairement au cas précédent, c'est maintenant la population qui joue le rôle de la distribution inconnue F_p et dont on ne connaît pas la proportion p et à partir de laquelle nous tirons un échantillon \mathbf{s} à partir d'un plan de sondage. Dans le cas d'une population finie, plusieurs approches sont possibles afin d'appliquer le bootstrap dont des méthodes par pseudo-population et des méthodes utilisant les poids bootstrap.

2.2. Méthode bootstrap par pseudo-population

Une approche du bootstrap par pseudo-population en population finie pour un plan aléatoire simple sans remise a été amenée par [Booth *et al.*, 1994] et consiste à créer une pseudo-population en répliquant chaque observation de l'échantillon s (de taille n_s) un certain nombre de fois puis à tirer des échantillons bootstrap à partir de cette pseudo-population en utilisant le plan de sondage. On dénote par \mathbf{U}^* la pseudo-population où l'on distingue deux parties :

$$\mathbf{U}^* = \mathbf{U}^f \cup \mathbf{U}^{c*} \quad (2.2.1)$$

où \mathbf{U}^f représente la partie fixe de la pseudo-population tandis que \mathbf{U}^{c*} représente la partie aléatoire. Soit $\lfloor x \rfloor$, la fonction qui renvoie pour tout nombre réel x le plus grand entier inférieur ou égal à x . La méthode de pseudo-population consiste tout d'abord à répéter chacune des observations de l'échantillon initial s , $k = \lfloor N/n_s \rfloor$ fois afin de créer \mathbf{U}^f . Si N/n_s n'est pas un entier, on complète notre pseudo-population en sélectionnant dans s un échantillon sans remise de taille $N - n_s k$ afin de créer la partie aléatoire \mathbf{U}^{c*} . Dans le cas où N/n_s est un entier, \mathbf{U}^{c*} représente l'ensemble vide. On dénote p^* , la proportion de succès de la pseudo-population \mathbf{U}^* de taille N . Nous sélectionnons ensuite de manière sans remise notre échantillon bootstrap s^* de taille n_s à partir de \mathbf{U}^* . Nous dénotons par Z_i^* la i^e donnée présente dans l'échantillon bootstrap. Nous sommes donc en mesure de calculer l'estimateur bootstrap

$$\hat{p}^* = \frac{1}{n_s} \sum_{i \in s^*} Z_i^*. \quad (2.2.2)$$

Notons que s^* correspond à l'ensemble $\{Z_1^*, \dots, Z_{n_s}^*\}$ et non pas les indices des unités de la pseudo-population qui ont été choisies de telle sorte que c'est par abus de notation qu'on dénote la somme des Z_i^* par $\sum_{i \in s^*} Z_i^*$. Cet abus de notation reviendra à quelques reprises avec d'autres ensembles, notamment s^+ et s^{*+} qui seront définis plus loin. De manière similaire à l'estimateur de la variance de \hat{p} défini en (2.1.12) pour un plan aléatoire simple sans remise, un estimateur de la variance de \hat{p}^* peut être

$$\hat{V}^*(\hat{p}^*) = (1 - f_s) \frac{\hat{p}^*(1 - \hat{p}^*)}{n_s - 1}. \quad (2.2.3)$$

De manière générale, dénotons par $\check{V}^*(\hat{p}^*)$ un estimateur de la variance de l'estimateur bootstrap. Ce dernier pourrait être l'estimateur défini en (2.2.3) ou autre chose. Nous répétons

ensuite un grand nombre de fois, B , la création de la partie aléatoire \mathbf{U}^{c*} afin de tirer nos différents échantillons bootstrap à partir de \mathbf{U}^* . En résumé, l'algorithme de pseudo-population de [Booth *et al.*, 1994] se définit ainsi :

- (1) On répète k fois chaque unité de notre échantillon initial s pour créer \mathbf{U}^f où $k = \lfloor N/n_s \rfloor$.
- (2) On crée \mathbf{U}^{c*} en sélectionnant sans remise $N - n_s k$ unités à partir de s . On a donc notre pseudo-population $\mathbf{U}^* = \mathbf{U}^f \cup \mathbf{U}^{c*}$. Soit p^* , la proportion de succès dans la pseudo-population \mathbf{U}^* .
- (3) On sélectionne sans remise un échantillon bootstrap s^* de taille n_s à partir de \mathbf{U}^* .
- (4) On calcule l'estimateur bootstrap \hat{p}^* et un estimateur de sa variance $\check{V}^*(\hat{p}^*)$.
- (5) On répète les étapes (2) à (4) un grand nombre de fois, B , pour avoir les différents estimateurs bootstrap, \hat{p}^* , paramètres de proportion p^* associés à \mathbf{U}^* ainsi que les estimateurs de variance bootstrap, $\check{V}^*(\hat{p}^*)$.

Une approche bootstrap qui considère l'estimation comme une fonction des poids associés aux observations de l'échantillon est possible dans un contexte de population finie.

2.3. Méthode des poids bootstrap

Dans cette approche, les observations sont considérées comme étant fixes et ce sont les poids qui sont générés selon une certaine méthode. La méthode par poids bootstrap qui fera l'objet de ce mémoire est une méthode modifiée d'une méthode bootstrap directe. Cette dernière, développée par [Rao et Wu, 1988], consiste à modifier chacune des données de l'échantillon en les redimensionnant. Nous dénotons $\bar{Z} = n_s^{-1} \sum_{i \in s} Z_i$, la moyenne des valeurs de Z dans s . Pour cette méthode, la taille de l'échantillon bootstrap peut être différente de n_s , la taille de l'échantillon s . C'est pourquoi nous définissons par n' la taille de l'échantillon bootstrap. Nous définirons plus loin la valeur de n' . Soit $Z'_i = \bar{Z} + C(Z_i - \bar{Z})$ la valeur redimensionnée de Z_i associée à l'unité i où

$$C = \sqrt{\frac{n'(1 - f_s)}{n_s - 1}}. \quad (2.3.1)$$

Nous avons donc l'échantillon $s' = \{Z'_i\}_{i \in s}$ de taille n_s , qui correspond aux unités redimensionnées de l'échantillon initial. On sélectionne ensuite avec remise un échantillon de taille

n' à partir de s' pour former notre échantillon bootstrap s^* . Nous tirons un grand nombre, B , d'échantillons bootstrap où nous calculons, entre autres, l'estimateur bootstrap et un estimateur de la variance de l'estimateur bootstrap pour chacun des B échantillons.

Quatre années plus tard, [Rao *et al.*, 1992] amène une méthode similaire mais qui met en œuvre des poids bootstrap au lieu de redimensionner chacune des valeurs Z_i de l'échantillon initial. En effet, on tire un échantillon bootstrap s^* de taille n' avec remise de s avant d'associer à chacune des valeurs Z_i un poids bootstrap

$$w_i^* = a_i^* w_i \quad (2.3.2)$$

où w_i est défini comme en (2.1.8) et a_i^* est défini de la manière suivante :

$$a_i^* = 1 + C \left(\frac{n_s m_i^*}{n'} - 1 \right) \quad (2.3.3)$$

avec m_i^* qui représente le nombre de fois que l'unité i fait partie de l'échantillon bootstrap s^* de taille n' . Nous avons donc que $\sum_{i \in s} m_i^* = n'$. Par définition, le vecteur $m^* = (m_1^*, \dots, m_n^*)$ suit une loi multinomiale de taille n' et de probabilités $(1/n_s, \dots, 1/n_s)$. Nous écrivons donc $m^* = (m_1^*, \dots, m_n^*) \sim MN(n'; 1/n_s, \dots, 1/n_s)$ d'où l'espérance et la variance de m_i^* s'écrivent respectivement comme

$$E^*(m_i^*) = \frac{n'}{n_s} \quad (2.3.4)$$

et

$$V^*(m_i^*) = \frac{n'}{n_s} \left(1 - \frac{1}{n_s} \right). \quad (2.3.5)$$

Il est important de noter que la covariance entre m_i^* et m_j^* est

$$cov^*(m_i^*, m_j^*) = \frac{-n'}{n_s^2}. \quad (2.3.6)$$

Comme mentionné précédemment, l'estimateur bootstrap s'exprime comme une fonction des poids des observations dans l'échantillon contrairement à l'estimateur bootstrap sous la méthode de pseudo-population présentée où l'estimateur est calculé au moyen de l'échantillon bootstrap. L'estimateur bootstrap pour une approche par poids bootstrap prend donc la forme

$$\hat{p}^* = \frac{1}{N} \sum_{i \in s} w_i^* Z_i = \frac{1}{N} \sum_{i \in s} a_i^* w_i Z_i = \frac{1}{n_s} \sum_{i \in s} a_i^* Z_i \quad (2.3.7)$$

si le plan de sondage utilisé est un plan aléatoire simple sans remise. Un estimateur de la variance de l'estimateur bootstrap pourrait prendre la même forme que celle définie pour une

approche par pseudo-population (2.2.3) mais en employant un estimateur calculé au moyen de l'approche par les poids :

$$\hat{V}^*(\hat{p}^*) = (1 - f_s) \frac{\hat{p}^*(1 - \hat{p}^*)}{n_s - 1}. \quad (2.3.8)$$

Encore une fois ici, dénotons par $\check{V}^*(\hat{p}^*)$ un estimateur de la variance de l'estimateur bootstrap. Il est ensuite possible de répéter un grand nombre de fois, B , l'échantillonnage de s^* afin d'avoir les différents estimateurs bootstrap et estimateurs de variance bootstrap.

Comme montré ci-dessous, la constante C définie dans le terme a_i^* (2.3.3) est choisie telle que la variance de \hat{p}^* , que l'on dénote par $V^*(\hat{p}^*)$, soit égale à l'estimation de la variance de \hat{p} sous le plan aléatoire simple sans remise défini en (2.1.12).

$$\begin{aligned} V^*(\hat{p}^*) &= V^* \left[\frac{1}{n_s} \sum_{i \in s} \left(1 + C \left(\frac{n_s m_i^*}{n'} - 1 \right) \right) Z_i \right] \\ &= \frac{1}{n_s^2} \left[\sum_{i \in s} C^2 \frac{n_s^2}{n'^2} V^*(m_i^*) Z_i^2 + 2 \sum_{\substack{i < j \\ i, j \in s}} C^2 \frac{n_s^2}{n'^2} \text{cov}^*(m_i^*, m_j^*) Z_i Z_j \right] \\ &= \frac{C^2}{n_s n'} \left[\sum_{i \in s} Z_i - \frac{1}{n_s} \sum_{i \in s} Z_i - \frac{2}{n_s} \sum_{\substack{i < j \\ i, j \in s}} Z_i Z_j \right] \\ &= \frac{C^2}{n_s n'} \left[\sum_{i \in s} Z_i - \frac{1}{n_s} \sum_{i \in s} Z_i - \frac{1}{n_s} \sum_{\substack{i \neq j \\ i, j \in s}} Z_i Z_j \right] \\ &= \frac{C^2}{n_s n'} \left[\sum_{i \in s} Z_i - \frac{1}{n_s} \sum_{i \in s} Z_i - \frac{1}{n_s} \left(\sum_{i, j \in s} Z_i Z_j - \sum_{\substack{i=j \\ i, j \in s}} Z_i Z_j \right) \right] \\ &= \frac{C^2}{n_s n'} \left[n_s \hat{p} - \hat{p} - \frac{1}{n_s} \left(\sum_{i \in s} Z_i \sum_{j \in s} Z_j - \sum_{i \in s} Z_i \right) \right] \\ &= \frac{C^2}{n_s n'} \left[n_s \hat{p} - \hat{p} - n_s \hat{p}^2 + \hat{p} \right] \\ &= \frac{C^2}{n'} [\hat{p}(1 - \hat{p})] \\ &= (1 - f_s) \frac{\hat{p}(1 - \hat{p})}{n_s - 1}. \end{aligned}$$

Nous avons donc que

$$V^*(\hat{p}^*) = \hat{V}(\hat{p}) = (1 - f_s) \frac{\hat{p}(1 - \hat{p})}{n_s - 1}. \quad (2.3.9)$$

Il est aussi possible de démontrer que l'estimateur bootstrap \hat{p}^* est sans biais pour l'estimateur au niveau de l'échantillon, \hat{p} , c'est-à-dire que $E^*(\hat{p}^*) = \hat{p}$.

Nous avons soulevé une problématique dans le contexte i.i.d. au chapitre précédent et celle-ci survient lorsque l'estimateur \hat{p} est 0. La solution envisagée dans ce contexte était d'estimer la distribution inconnue F_p par $F_{\hat{p}}$ afin de générer les observations bootstrap. Voyons comment cette solution peut être adaptée dans un contexte de population finie.

2.4. Modifications aux méthodes bootstrap

Lorsque l'estimateur \hat{p} est nul, l'estimateur bootstrap \hat{p}^* est 0 avec probabilité 1 et l'estimation de la variance de \hat{p}^* comme on la connaît, $\hat{V}^*(\hat{p}^*)$, devient alors nulle avec probabilité 1 également. Les intervalles de confiance basés sur la distribution de $\hat{p}^* - \hat{p}$ sont dégénérés alors qu'ils ne contiennent qu'un point, le point 0, et la distribution de $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p}^*)}$ n'est tout simplement pas définie car autant le numérateur que le dénominateur valent 0 avec probabilité 1.

Il est important de constater dans un contexte de population finie que tant pour l'approche par pseudo-population que celle par les poids, le bootstrap repose sur l'échantillon observé et non tout simplement sur un estimateur de la distribution qui génère un échantillon, comme dans un contexte i.i.d. À nouveau, dans un contexte de population finie, si toutes les observations de l'échantillon sont égales à 0 de telle sorte que \hat{p} est 0, alors la pseudo-population créée ne contiendra que des observations égales à 0 et l'échantillon bootstrap ne contiendra que des 0 aussi. L'estimateur bootstrap \hat{p}^* sera nécessairement 0 avec probabilité 1, comme dans le cas i.i.d. De la même façon, si on prend une approche basée sur les poids, quels que soient les poids générés, si on les applique à un échantillon qui ne contient que des 0, alors à nouveau \hat{p}^* sera nécessairement 0. Dans ce cas, il devient impossible de créer des intervalles de confiance basés sur les distributions de $\hat{p}^* - \hat{p}$ et $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p}^*)}$.

La solution qui a été apportée dans le cas i.i.d. afin de contourner les difficultés occasionnées par cette problématique consiste à faire le rééchantillonnage à partir de la distribution $F_{\hat{p}}$. Or, l'estimateur \tilde{p} correspond au nombre de succès dans un échantillon modifié où l'on a ajouté 4 observations, soit 2 succès et 2 échecs. Ainsi c'est comme si on modifiait l'échantillon original $s = \{Z_i\}_{i \in \mathcal{S}}$ de taille n_s en l'échantillon modifié de taille $n_s + 4$,

$s^+ = \{Z_i\}_{i \in s} \cup \{0,0,1,1\}$. La solution proposée pour un échantillon tiré d'une population finie consiste donc à appliquer une approche bootstrap de type pseudo-population ou de type poids bootstrap à l'échantillon modifié s^+ plutôt qu'à l'échantillon original s . Comme l'échantillon modifié est de taille $n_s + 4$ plutôt que n_s , des ajustements aux méthodes devront être apportés.

Dénotons tout d'abord l'échantillon

$$s^+ = \{Z_i\}_{i \in s} \cup \{0,0,1,1\} \quad (2.4.1)$$

soit l'échantillon initial s auquel nous avons ajouté 2 succès et 2 échecs. Si \hat{p} (2.1.10) dénote la proportion de succès dans l'échantillon s , de son côté, la proportion de succès de l'échantillon s^+ est dénotée par

$$\tilde{p} = \frac{1}{n_{s^+}} \sum_{i \in s^+} Z_i \quad (2.4.2)$$

où n_{s^+} est la taille de l'échantillon s^+ .

Au niveau de l'approche par pseudo-population, nous avons vu à la section 2.2 la méthodologie utilisée lorsque l'échantillon considéré est l'échantillon s . Dans le cas où l'on considère maintenant l'échantillon s^+ , quelques changements devront être apportés.

On répète tout d'abord chaque unité de l'échantillon s^+ , $k^+ = \lfloor N/n_{s^+} \rfloor$ fois afin de créer \mathbf{U}^f . Nous sélectionnons ensuite sans remise dans s^+ un échantillon de taille $N - n_{s^+}k^+$ pour créer \mathbf{U}^{c*} . Notre pseudo-population est ainsi créée. Comme la pseudo-population est créée à partir de s^+ , nous utiliserons la notation $\mathbf{U}^*(s^+)$ afin de dénoter la pseudo-population qui a été créée à partir de s^+ . C'est donc dire que la pseudo-population $\mathbf{U}^*(s)$ est équivalente à la notation \mathbf{U}^* (2.2.1) présentée à la section 2.2 où l'échantillon s est utilisé afin de créer la pseudo-population.

Soit $p_{s^+}^*$, la proportion de succès dans $\mathbf{U}^*(s^+)$. Encore une fois, à la section 2.2, nous avons dénoté la proportion de succès dans la pseudo-population \mathbf{U}^* par p^* . Cette dernière forme est équivalente ici à écrire p_s^* et qui dénote la proportion de succès de la pseudo-population créée à partir de l'échantillon s .

À partir de la pseudo-population $\mathbf{U}^*(s^+)$, nous tirons un échantillon bootstrap sans remise s^* de taille n_s . L'estimateur bootstrap représente la proportion de succès dans l'échantillon bootstrap et il est calculé de la même manière que précédemment (2.2.2), mais on va le

dénoter par

$$\hat{p}_{s^+}^* = \frac{1}{n_s} \sum_{i \in s^*} Z_i^*. \quad (2.4.3)$$

L'indice s^+ dans (2.4.3) signifie que l'estimateur bootstrap est calculé sur un échantillon bootstrap de taille n_s , mais dont la pseudo-population a été créée à partir de l'échantillon s^+ . L'estimateur bootstrap (2.2.2) présenté à la section 2.2 pourrait donc désormais être dénoté par

$$\hat{p}_s^* = \frac{1}{n_s} \sum_{i \in s^*} Z_i^* \quad (2.4.4)$$

compte tenu que la pseudo-population a été créée à partir de l'échantillon s dans ce cas-ci. Lorsque l'échantillon s^+ est considéré pour la création de la pseudo-population, le paramètre de proportion associé à la pseudo-population, $p_{s^+}^*$, sera différent de 0 avec probabilité 1. De plus, l'estimateur bootstrap $\hat{p}_{s^+}^*$ n'est plus nécessairement nul. Comme nous le verrons à la section 2.5.1, nous sommes donc en mesure de bâtir les intervalles de confiance basés sur la distribution de $\hat{p}_{s^+}^* - p_{s^+}^*$ sans problème.

Comme mentionné plus tôt, nous sommes confrontés à des difficultés lors de la création d'intervalles de confiance basés sur la distribution de $(\hat{p}^* - \hat{p})/\sqrt{\hat{V}^*(\hat{p}^*)}$ qui n'est pas défini à cause du dénominateur qui devient nul lorsque \hat{p} est 0. Afin d'estimer la variance de l'estimateur bootstrap $\hat{p}_{s^+}^*$, nous pourrions penser qu'il est suffisant d'utiliser la même forme que celle présentée en (2.2.3), mais en remplaçant l'estimateur \hat{p}^* par $\hat{p}_{s^+}^*$. Toutefois, ce dernier peut encore être nul et nous pourrions nous retrouver face aux mêmes difficultés, soit avec un estimateur de la variance de l'estimateur bootstrap égal à 0. C'est pourquoi il pourrait être intéressant de définir un échantillon bootstrap de taille $n_{s^{*+}} = n_{s^*} + 4$ dénoté

$$s^{*+} = \{Z_i^*\}_{i \in s^*} \cup \{0,0,1,1\} \quad (2.4.5)$$

qui est l'échantillon bootstrap s^* auquel nous avons ajouté 4 observations soit 2 succès et 2 échecs afin de considérer l'estimateur bootstrap dénoté

$$\tilde{p}_{s^+}^* = \frac{1}{n_{s^{*+}}} \sum_{i \in s^{*+}} Z_i^* \quad (2.4.6)$$

qui lui, ne peut jamais être nul. Encore une fois ici, il est important de souligner que l'indice s^+ dans le terme à gauche dans l'égalité (2.4.6) signifie que la pseudo-population utilisée est

$\mathbf{U}^*(s^+)$. C'est pourquoi nous pouvons également définir l'estimateur

$$\hat{p}_s^* = \frac{1}{n_{s^{*+}}} \sum_{i \in s^{*+}} Z_i^* \quad (2.4.7)$$

et qui prend la même forme que l'estimateur précédent mais dont la pseudo-population utilisée est $\mathbf{U}^*(s)$. Lorsque l'échantillon considéré afin de créer la pseudo-population est s^+ , nous allons estimer la variance de l'estimateur bootstrap, $\hat{p}_{s^+}^*$, par

$$\tilde{V}^*(\hat{p}_{s^+}^*) = (1 - f_s) \frac{\hat{p}_{s^+}^* (1 - \hat{p}_{s^+}^*)}{n_s - 1} \quad (2.4.8)$$

tandis que nous utiliserons la forme

$$\tilde{V}^*(\hat{p}_s^*) = (1 - f_s) \frac{\hat{p}_s^* (1 - \hat{p}_s^*)}{n_s - 1} \quad (2.4.9)$$

afin d'estimer la variance de l'estimateur bootstrap lorsque l'échantillon considéré pour construire la pseudo-population est s , contrairement à la forme définie en (2.2.3). Comme nous le verrons à la section 2.5.1, cette forme nous sera préférable afin d'estimer la variance de l'estimateur bootstrap lorsqu'on considère l'échantillon s . Nous verrons également dans quelles circonstances nous utiliserons l'échantillon s .

Voici l'algorithme modifié de [Booth *et al.*, 1994] lorsque l'échantillon considéré pour la création de la pseudo-population est s^+ :

- (1) On répète k^+ fois chaque unité de notre échantillon s^+ pour créer \mathbf{U}^f où $k^+ = \lfloor N/n_{s^+} \rfloor$.
- (2) On crée \mathbf{U}^{c^*} en sélectionnant sans remise $N - n_{s^+}k^+$ unités dans s^+ . On a donc notre pseudo-population $\mathbf{U}^*(s^+) = \mathbf{U}^f \cup \mathbf{U}^{c^*}$. Soit $p_{s^+}^*$, la proportion de succès de la pseudo-population $\mathbf{U}^*(s^+)$.
- (3) On sélectionne sans remise un échantillon bootstrap s^* de taille n_s à partir de \mathbf{U}^* .
- (4) On calcule l'estimateur bootstrap $\hat{p}_{s^+}^*$ et l'estimateur de sa variance $\tilde{V}^*(\hat{p}_{s^+}^*)$.
- (5) On répète les étapes (2) à (4) un grand nombre de fois, B , pour avoir les différents estimateurs bootstrap, $\hat{p}_{s^+}^*$, paramètres de proportion $p_{s^+}^*$ associés à $\mathbf{U}^*(s^+)$ ainsi que les estimateurs de variance bootstrap, $\tilde{V}^*(\hat{p}_{s^+}^*)$.

Si à la section 2.3 nous avons présenté la méthodologie entourant l'utilisation de la méthode des poids bootstrap de [Rao *et al.*, 1992] lorsque l'échantillon considéré était

l'échantillon s , présentons maintenant les modifications à apporter à cette méthode lorsque l'échantillon s^+ est considéré.

Puisque dans la méthode des poids bootstrap, l'estimateur bootstrap est calculé non seulement à partir de poids générés par un échantillon bootstrap mais également à partir des données de l'échantillon observé, l'estimateur bootstrap et les poids bootstrap associés à l'échantillon devront être modifiés. Si à la section 2.3 on tirait un échantillon bootstrap s^* de taille n' avec remise de s , dans le cas qui nous concerne, on tire dorénavant un échantillon bootstrap s^* de taille n' avec remise de s^+ . L'estimateur bootstrap est donc maintenant calculé sur l'échantillon s^+ . Soit s^+ l'échantillon considéré, nous pouvons définir l'estimateur bootstrap par

$$\tilde{p}^* = \frac{1}{n_{s^+}} \sum_{i \in s^+} a_{i,s^+}^* Z_i, \quad (2.4.10)$$

où le terme a_{i,s^+}^* est défini comme

$$a_{i,s^+}^* = 1 + C_{s^+} \left(\frac{n_{s^+} m_i^*}{n'} - 1 \right) \quad (2.4.11)$$

et m_i^* représente encore le nombre de fois que l'observation i , de s^+ , fait partie de l'échantillon bootstrap s^* de taille n' . Nous avons maintenant que $E^*(m_i^*) = n'/n_{s^+}$, $V^*(m_i^*) = n'n_{s^+}^{-1}(1 - 1/n_{s^+})$ et $cov^*(m_i^*, m_j^*) = -n'/n_{s^+}^2$. Nous verrons plus loin de quelle manière est définie la constante C_{s^+} .

Ceci est à contraster avec l'estimateur (2.3.7) de la section 2.3 qu'on peut réécrire sous la forme

$$\hat{p}^* = \frac{1}{n_s} \sum_{i \in s} a_{i,s}^* Z_i \quad (2.4.12)$$

où $a_{i,s}^*$ remplace a_i^* mais prend la même forme que celle définie en (2.3.3) et où la constante C est maintenant définie par la notation C_s mais calculée de la même manière que précédemment.

C'est donc dire qu'un estimateur bootstrap calculé avec la méthode par pseudo-population utilise un indice afin d'indiquer de quelle manière est créée la pseudo-population comme en (2.4.3) tandis qu'un estimateur bootstrap calculé au moyen de la méthode des poids bootstrap est défini sans indice, mais l'utilisation du tilde ($\tilde{\cdot}$) et du chapeau ($\hat{\cdot}$) nous permet de différencier les estimateurs calculés lorsque l'échantillon considéré est s^+ ou s comme en (2.4.10) et (2.4.12) respectivement. La solution dans un contexte i.i.d. était d'ajouter deux succès et deux échecs à l'estimateur de proportion qui était défini par \tilde{p} , ce

qui est identique à considérer la proportion d'un échantillon s^+ dans ce contexte-ci. C'est donc pourquoi nous considérons le tilde lorsque l'échantillon considéré est s^+ et le chapeau lorsque l'échantillon considéré est plutôt s .

Soit $w_{i,s^+} = N/n_{s^+}$,

$$w_{i,s^+}^* = a_{i,s^+}^* w_{i,s^+} \quad (2.4.13)$$

est le poids bootstrap associé à l'unité i où l'échantillon considéré est s^+ . Le poids w_i^* défini en (2.3.2) prend désormais la forme $w_{i,s}^*$ où le poids w_i défini en (2.1.8) prend maintenant la forme $w_{i,s}$.

Nous verrons plus loin une condition afin de garantir que l'estimateur bootstrap, \tilde{p}^* , ne soit jamais nul. Notons que l'estimateur \tilde{p} (2.4.2) ne peut en aucun cas être nul. Comme nous le verrons à la section 2.5.1, en s'assurant que l'estimateur bootstrap ne soit jamais nul, il sera possible de créer des intervalles de confiance basés sur la distribution de $\tilde{p}^* - \tilde{p}$ lorsque l'échantillon considéré est s^+ .

Faisons donc l'hypothèse qu'une condition (qu'on précisera plus tard) soit respectée afin que \tilde{p}^* ne puisse en aucun cas être nul. Nous pouvons utiliser ce dernier directement afin d'estimer la variance de l'estimateur bootstrap. Nous allons donc utiliser la forme (2.3.8) pour estimer la variance bootstrap de la proportion lorsque l'échantillon considéré est l'échantillon s et nous utiliserons la forme

$$\tilde{V}^*(\tilde{p}^*) = (1 - f_s) \frac{\tilde{p}^*(1 - \tilde{p}^*)}{n_s - 1} \quad (2.4.14)$$

afin d'estimer la variance de l'estimateur bootstrap lorsque l'échantillon considéré est s^+ .

À la section 2.3, nous avons mentionné que la constante C (2.3.1) assurait que la variance de l'estimateur bootstrap soit la même que l'estimation usuelle de la variance de \hat{p} . En d'autres mots, lorsque s est l'échantillon considéré, en utilisant la constante C définie en (2.3.1), on a que $V^*(\hat{p}^*)$ est égal à $\hat{V}(\hat{p})$. Comme mentionné plus tôt, cette constante est désormais dénotée par C_s . Si l'échantillon considéré est l'échantillon s^+ , nous devons trouver la nouvelle constante dénotée C_{s^+} qui assure l'égalité entre la variance de l'estimateur bootstrap et l'estimation de la variance de \hat{p} . Il faut donc spécifier l'estimateur que nous utiliserons pour estimer la variance de \hat{p} . Comme nous le verrons à la section 2.5.1, il sera important que l'estimation de la variance de \hat{p} ne soit pas nulle, afin d'assurer que l'intervalle bootstrap- t soit autre chose que le point 0 lorsque \hat{p} est 0 car il est possible que l'estimateur

de variance $\hat{V}(\hat{p})$ (2.1.12) soit nul si \hat{p} est nul. C'est pourquoi, peu importe que l'échantillon considéré soit l'échantillon s ou s^+ , nous considérerons

$$\tilde{V}(\hat{p}) = (1 - f_s) \frac{\tilde{p}(1 - \tilde{p})}{n_s - 1} \quad (2.4.15)$$

pour estimer la variance de \hat{p} . Il est important de noter qu'ici \tilde{p} est basé sur l'échantillon s^+ de taille n_{s^+} , mais que f_s et n_s sont associés à l'échantillon original s afin de refléter la variabilité de l'estimateur \hat{p} qui est également basé sur l'échantillon s . Il est aussi utile de noter ici qu'on utilisera également cet estimateur (2.4.15) afin d'estimer la variance de \hat{p} lors de la création des intervalles de confiance bootstrap- t avec l'approche par pseudo-population.

Avec un raisonnement similaire à celui effectué à la section 2.3, il est possible de calculer la variance de l'estimateur bootstrap lorsque l'échantillon considéré est s^+ :

$$\begin{aligned} V^*(\tilde{p}^*) &= V^* \left[\frac{1}{n_{s^+}} \sum_{i \in s^+} \left(1 + C_{s^+} \left(\frac{n_{s^+} m_i^*}{n'} - 1 \right) \right) Z_i \right] \\ &= \frac{1}{n_{s^+}^2} \left[\sum_{i \in s^+} C_{s^+}^2 \frac{n_{s^+}^2}{n'^2} V^*(m_i^*) Z_i^2 + 2 \sum_{\substack{i < j \\ i, j \in s^+}} C_{s^+}^2 \frac{n_{s^+}^2}{n'^2} \text{cov}^*(m_i^*, m_j^*) Z_i Z_j \right] \\ &= \frac{C_{s^+}^2}{n_{s^+} n'} \left[\sum_{i \in s^+} Z_i - \frac{1}{n_{s^+}} \sum_{i \in s^+} Z_i - \frac{2}{n_{s^+}} \sum_{\substack{i < j \\ i, j \in s^+}} Z_i Z_j \right] \\ &= \frac{C_{s^+}^2}{n_{s^+} n'} \left[\sum_{i \in s^+} Z_i - \frac{1}{n_{s^+}} \sum_{i \in s^+} Z_i - \frac{1}{n_{s^+}} \sum_{\substack{i \neq j \\ i, j \in s^+}} Z_i Z_j \right] \\ &= \frac{C_{s^+}^2}{n_{s^+} n'} \left[\sum_{i \in s^+} Z_i - \frac{1}{n_{s^+}} \sum_{i \in s^+} Z_i - \frac{1}{n_{s^+}} \left(\sum_{i, j \in s^+} Z_i Z_j - \sum_{\substack{i=j \\ i, j \in s^+}} Z_i Z_j \right) \right] \\ &= \frac{C_{s^+}^2}{n_{s^+} n'} \left[n_{s^+} \tilde{p} - \tilde{p} - \frac{1}{n_{s^+}} \left(\sum_{i \in s^+} Z_i \sum_{j \in s^+} Z_j - \sum_{i \in s^+} Z_i \right) \right] \\ &= \frac{C_{s^+}^2}{n_{s^+} n'} \left[n_{s^+} \tilde{p} - \tilde{p} - n_{s^+} \tilde{p}^2 + \tilde{p} \right] \\ &= \frac{C_{s^+}^2}{n'} [\tilde{p}(1 - \tilde{p})]. \end{aligned}$$

Il se trouve que, lorsque l'échantillon considéré est s^+ , la variance de l'estimateur bootstrap, $V^*(\hat{p}^*)$ est égale à l'estimation de la variance de \hat{p} définie par $\check{V}(\hat{p})$ (2.4.15) lorsque

$$C_{s^+} = C_s = \sqrt{\frac{n'(1-f_s)}{n_s-1}}. \quad (2.4.16)$$

Si dans le cadre de la méthode de pseudo-population, nous avons créé un échantillon bootstrap de la forme s^{*+} afin de pouvoir bâtir un estimateur qui ne peut jamais être nul et qui prend la forme $\hat{p}_{s^+}^*$ (2.4.6) ou encore \tilde{p}_s^* (2.4.7), dans le cadre de la méthode des poids bootstrap, une condition nous permet d'avoir un estimateur bootstrap qui n'est jamais nul. Afin d'avoir un estimateur bootstrap toujours plus grand que 0, nous devons avoir n' qui satisfait l'inégalité suivante :

$$n' \leq \frac{n_s - 1}{1 - f_s}. \quad (2.4.17)$$

En effet, l'inégalité (2.4.17) assure que C_{s^+} (et donc C_s) soit toujours inférieur à 1. Dans ce cas, les poids bootstrap w_{i,s^+}^* (et donc aussi $w_{i,s}^*$) sont strictement positifs et assurent que l'estimateur bootstrap \hat{p}^* ne soit jamais nul même si m_i^* , le nombre de fois que la i^e valeur de s^+ est dans l'échantillon bootstrap, est nul.

L'inégalité (2.4.17) n'assure toutefois pas que \hat{p}^* ne sera pas nul. En effet, si s ne contient que des échecs, l'estimateur bootstrap \hat{p}^* sera nul même si l'inégalité (2.4.17) est respectée, ce qui nous causerait des ennuis dans la construction des intervalles bootstrap- t puisque l'estimation de la variance de l'estimateur bootstrap serait nulle. Par contre, comme nous le verrons, nous n'utiliserons \hat{p}^* , et donc l'échantillon s , que lorsque \hat{p} est non nul. C'est pourquoi on considère que \hat{p}^* ne peut être nul si l'inégalité (2.4.17) est respectée.

2.5. Intervalles de confiance bootstrap

Comme mentionné précédemment, pour construire des intervalles de confiance nous pouvons utiliser les distributions de $\hat{p} - p$ ou bien de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$ où $\check{V}(\hat{p})$ est un estimateur de la variance de \hat{p} . Dans un contexte de population finie, c'est la population qui joue le rôle de la distribution des observations. C'est pourquoi, si au chapitre 1 nous avons défini les quantités $J_n(x, F_p)$ et ses quantiles $J_n^{-1}(\alpha, F_p)$, nous allons désormais définir

$$J_{n_s}(x) = Prob(\hat{p} - p \leq x) \quad (2.5.1)$$

et $J_{n_s}^{-1}(\alpha)$, la fonction de répartition et le quantile α de la distribution de $\hat{p} - p$ lorsque les n_s observations sont tirées à partir de \mathbf{U} selon le plan de sondage. Précédemment, nous avons

également défini la quantité $K_n(x, F_p)$ dans un contexte i.i.d. Nous définissons maintenant

$$K_{n_s}(x) = Prob \left([\hat{p} - p] / \sqrt{\check{V}(\hat{p})} \leq x \right) \quad (2.5.2)$$

et $K_{n_s}^{-1}(\alpha)$, la fonction de répartition et le quantile α de la distribution de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$ lorsque les n_s observations sont tirées à partir de \mathbf{U} selon le plan de sondage. Nous faisons l'hypothèse ici que $\check{V}(\hat{p})$ est non nul afin d'assurer que la variable aléatoire définissant K_{n_s} soit toujours définie. Des intervalles de confiance exacts basés sur les distributions de $\hat{p} - p$ et de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$ pourraient être définis. De façon similaire à l'intervalle de confiance défini en (1.3.2) au chapitre 1, nous pouvons montrer que si J_{n_s} était connue que l'intervalle

$$[\hat{p} - J_{n_s}^{-1}(1 - \alpha/2), \hat{p} - J_{n_s}^{-1}(\alpha/2)] \quad (2.5.3)$$

serait un intervalle de confiance exact de niveau $1 - \alpha$ pour p . La même alternative à l'intervalle (1.3.3) au chapitre précédent nous donne l'intervalle de confiance

$$[\hat{p} + J_{n_s}^{-1}(\alpha/2), \hat{p} + J_{n_s}^{-1}(1 - \alpha/2)]. \quad (2.5.4)$$

Nous pouvons également montrer qu'en connaissant K_{n_s} , et ses quantiles, l'intervalle

$$\left[\hat{p} - K_{n_s}^{-1}(1 - \alpha/2)\sqrt{\check{V}(\hat{p})}, \hat{p} - K_{n_s}^{-1}(\alpha/2)\sqrt{\check{V}(\hat{p})} \right] \quad (2.5.5)$$

serait un intervalle de confiance de niveau exact $1 - \alpha$. Malgré que ces intervalles de confiance soient bel et bien définis, il est impossible de pouvoir les bâtir directement car les distributions de $\hat{p} - p$ et de $(\hat{p} - p)/\sqrt{\check{V}(\hat{p})}$ dépendent directement de la vraie population \mathbf{U} . Les deux approches bootstrap par pseudo-population et par les poids bootstrap nous permettent d'estimer ces distributions et d'ainsi bâtir ces intervalles de confiance.

2.5.1. Estimation

L'estimation des différentes distributions s'effectue différemment selon la méthode bootstrap utilisée, mais la façon d'estimer dépendra également de l'échantillon considéré, soit s ou s^+ . C'est pourquoi nous dénoterons par \check{s} l'échantillon considéré dans l'estimation des différentes distributions utiles à la création d'intervalles de confiance bootstrap.

On estime la distribution de $J_{n_s}(x)$ par $J_{n_s,*}(x; m, \check{s})$ où \check{s} est l'échantillon considéré et

$$m = \begin{cases} PP & \text{si la méthode par pseudo-population est utilisée} \\ W & \text{si la méthode par les poids est utilisée.} \end{cases} \quad (2.5.6)$$

Comme la proportion p représente la proportion de succès associée à la population \mathbf{U} , il va de soi d'utiliser la proportion p_s^* associée à la pseudo-population $\mathbf{U}^*(\check{s})$ dans la version bootstrap lorsque la pseudo-population est créée à partir de l'échantillon \check{s} . Dans le cas d'une approche par pseudo-population, on estime la distribution de $J_{n_s}(x)$ par

$$J_{n_s; *}(x; PP, \check{s}) = Prob^*(\hat{p}_s^* - p_s^* \leq x) \quad (2.5.7)$$

et donc $J_{n_s; *}^{-1}(\alpha; PP, \check{s})$ est le quantile α de la distribution bootstrap $\hat{p}_s^* - p_s^*$. Si une approche par les poids bootstrap est utilisée, on estime la distribution de $J_{n_s}(x)$ par

$$J_{n_s; *}(x; W, \check{s}) = Prob^*(\check{p}^* - \check{p} \leq x) \quad (2.5.8)$$

et donc $J_{n_s; *}^{-1}(\alpha; W, \check{s})$ est le quantile α de la distribution bootstrap $\check{p}^* - \check{p}$. Si on considère l'échantillon s^+ dans l'estimation, \check{p}^* prend donc la forme \tilde{p}^* (2.4.10) et l'estimateur \check{p} devient \tilde{p} (2.4.2). Si, au contraire, on considère l'échantillon s , l'estimateur bootstrap défini par \check{p}^* prend la forme \hat{p}^* (2.3.7) et \check{p} devient \hat{p} (2.1.10).

L'intervalle de confiance bootstrap de base de niveau $1 - \alpha$ prend la forme

$$[\hat{p} - J_{n_s; *}^{-1}(1 - \alpha/2; m, \check{s}), \hat{p} - J_{n_s; *}^{-1}(\alpha/2; m, \check{s})]. \quad (2.5.9)$$

Le second intervalle est l'intervalle de confiance percentile de niveau $1 - \alpha$ et est défini par

$$[\hat{p} + J_{n_s; *}^{-1}(\alpha/2; m, \check{s}), \hat{p} + J_{n_s; *}^{-1}(1 - \alpha/2; m, \check{s})]. \quad (2.5.10)$$

Définissons $L_{n_s}(x) = Prob(\hat{p} \leq x)$ et $L_{n_s}^{-1}(\alpha)$, la fonction de répartition et le quantile α de la distribution de \hat{p} lorsque les n_s observations sont tirées à partir de \mathbf{U} selon le plan de sondage. Nous estimons $L_{n_s}(x)$ par $L_{n_s; *}(x; m, \check{s})$ où $L_{n_s; *}(x; PP, \check{s}) = Prob^*(\hat{p}_s^* \leq x)$ et $L_{n_s; *}(x; W, \check{s}) = Prob^*(\check{p}^* \leq x)$. Nous pouvons donc définir l'intervalle de confiance percentile *classique* de niveau $1 - \alpha$ défini par

$$[L_{n_s; *}^{-1}(\alpha/2; m, \check{s}), L_{n_s; *}^{-1}(1 - \alpha/2; m, \check{s})]. \quad (2.5.11)$$

À nouveau, voir [Mashreghi *et al.*, 2016] pour la distinction entre l'intervalle percentile et l'intervalle percentile classique. Afin de bâtir le dernier intervalle, nous estimons la distribution de $K_{n_s}(x)$ par $K_{n_s; *}(x; m, \check{s})$ où l'échantillon \check{s} est considéré dans l'estimation. Dans le cas d'une approche par pseudo-population, on estime $K_{n_s}(x)$ par

$$K_{n_s; *}(x; PP, \check{s}) = Prob^*\left([\hat{p}_s^* - p_s^*] / \sqrt{\check{V}^*(\hat{p}_s^*)} \leq x\right) \quad (2.5.12)$$

où $\check{V}^*(\hat{p}_s^*)$ est un estimateur de la variance de l'estimateur bootstrap. Si l'échantillon s^+ est considéré, $\check{V}^*(\hat{p}_s^*)$ prend la forme définie en (2.4.8) et si l'échantillon s est l'échantillon considéré, $\check{V}^*(\hat{p}_s^*)$ prend la forme définie en (2.4.9). C'est donc dire que $K_{n_s; *}^{-1}(\alpha; PP, \check{s})$ est le quantile α de la distribution de $(\hat{p}_s^* - p_s^*)/\sqrt{\check{V}^*(\hat{p}_s^*)}$.

Si une approche par les poids bootstrap est utilisée, on estime $K_{n_s}(x)$ par

$$K_{n_s; *}(x; W, \check{s}) = Prob^* \left([\check{p}^* - \check{p}] / \sqrt{\check{V}^*(\check{p}^*)} \leq x \right) \quad (2.5.13)$$

où $\check{V}^*(\check{p}^*)$ est un estimateur de la variance de l'estimateur bootstrap. Si l'échantillon s^+ est considéré, $\check{V}^*(\check{p}^*)$ prend la forme définie en (2.4.14) tandis que si l'échantillon s est l'échantillon considéré, $\check{V}^*(\check{p}^*)$ prend la forme définie en (2.3.8). C'est donc dire que $K_{n_s; *}^{-1}(\alpha; W, \check{s})$ est le quantile α de la distribution de $(\check{p}^* - \check{p})/\sqrt{\check{V}^*(\check{p}^*)}$.

Nous pouvons donc bâtir l'intervalle bootstrap- t de niveau $1 - \alpha$ qui prend la forme

$$\left[\hat{p} - K_{n_s; *}^{-1}(1 - \alpha/2; m, \check{s})\sqrt{\check{V}(\hat{p})}, \hat{p} - K_{n_s; *}^{-1}(\alpha/2; m, \check{s})\sqrt{\check{V}(\hat{p})} \right] \quad (2.5.14)$$

où $\check{V}(\hat{p})$ est un estimateur de la variance de \hat{p} . Comme nous l'avons mentionné plus tôt, l'une des conditions afin de s'assurer que l'intervalle (2.5.14) soit défini ou qu'il soit autre chose qu'un seul point, est que l'estimateur de la variance de l'estimateur bootstrap ne soit jamais nul de manière à assurer que le quantile α de la distribution K_{n_s} soit toujours calculable. La seconde condition afin d'assurer que l'intervalle de confiance précédent soit autre chose qu'un seul point, le point 0, est que l'estimateur de la variance de \hat{p} se doit d'être toujours différent de 0. Peu importe l'échantillon utilisé, s ou s^+ , et peu importe la méthode bootstrap utilisée, pseudo-population ou par les poids, l'estimateur de la variance de \hat{p} utilisé sera celui défini en (2.4.15) et qui utilise l'estimateur \tilde{p} .

Nous avons mentionné plus tôt la problématique concernant la création des intervalles de confiance bootstrap lorsque l'estimateur \hat{p} est égal à 0 et qui implique donc, entres autres, que notre estimateur bootstrap \hat{p}^* devienne 0 avec probabilité 1. La solution envisagée jusqu'à maintenant consiste à utiliser un échantillon modifié dénoté s^+ , de manière à ce que l'estimateur bootstrap ne soit pas 0 avec probabilité 1. Au chapitre précédent, nous avons également mentionné qu'il pourrait être correct d'utiliser l'estimateur dénoté \tilde{p} seulement lorsque \hat{p} est égal à 0 et non pour toutes les situations. Entre autres, dans le cas où \hat{p} n'est pas nul, la probabilité que l'estimateur bootstrap soit 0 est inférieure à 1. Nous avons donc

introduit l'estimateur \hat{p} défini en (1.3.10) qui est égal à \tilde{p} seulement si \hat{p} est nul. Il est égal à \hat{p} dans le cas contraire. L'équivalent à l'estimateur \hat{p} dans un contexte de population finie est d'utiliser l'échantillon s^+ seulement lorsque \hat{p} est égal à 0 et d'utiliser s si ce n'est pas le cas. Nous définissons donc l'échantillon

$$\hat{s} = \begin{cases} s^+ & \text{si } \hat{p} = 0 \text{ ou } \hat{p} = 1 \\ s & \text{sinon.} \end{cases} \quad (2.5.15)$$

En considérant ce nouvel échantillon, nous avons donc 2 solutions afin de contourner la problématique soulevée en considérant un échantillon s^+ ou encore l'échantillon \hat{s} . Afin de faciliter la lecture des résultats des simulations, nous ferons référence à la méthode bootstrap par pseudo-population ou par les poids à la Agresti-Coull lorsque ces méthodes seront basées sur l'échantillon s^+ alors que nous parlerons de méthode bootstrap adaptative par pseudo-population ou par les poids lorsque l'échantillon \hat{s} sera utilisé.

Chapitre 3

Méthodes traditionnelles considérées

Comme nous l'avons vu, il est possible de bâtir des intervalles de confiance en utilisant le bootstrap. Plusieurs autres méthodes ont été introduites pour construire des intervalles de confiance pour une proportion, soit dans un contexte i.i.d. ou soit dans un contexte d'enquête. Dans ce chapitre, nous introduisons brièvement les méthodes que nous comparerons au bootstrap dans les simulations. Il s'agit des méthodes de Wald, Logit, Wilson, Agresti-Coull, Clopper-Pearson et Arcsine Square Root.

Dans ce chapitre, nous décrivons les différentes méthodes applicables dans un contexte i.i.d., c'est-à-dire où les données de l'échantillon sont générées selon une loi de Bernoulli avec n essais et X succès. Nous verrons par la suite les modifications à considérer lorsqu'on tire un échantillon selon un plan de sondage. La description des intervalles de confiance présentée est en partie celle qui est utilisée par [Franco *et al.*, 2019].

3.1. Intervalle de Wald

La méthode de Wald est la méthode la plus courante dans la construction d'intervalles de confiance pour un paramètre de proportion p . L'intervalle de confiance de Wald de niveau $1 - \alpha$ pour p prend la forme :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{p})} \quad (3.1.1)$$

où $z_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale de moyenne 0 et de variance égale à 1 dénotée $N(0,1)$ et $\hat{V}(\hat{p})$ dénote l'estimation de la variance de \hat{p} . Dans un contexte i.i.d., nous avons que

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n}. \quad (3.1.2)$$

L'intervalle de confiance de Wald est basé sur le fait que lorsque $n \rightarrow \infty$, la distribution de $(\hat{p} - p)/\sqrt{\hat{p}(1 - \hat{p})/n} \rightarrow N(0,1)$.

3.2. Intervalle Logit

La méthode Logit consiste à inverser un intervalle de confiance de type Wald pour la transformation logit de p , $\lambda = \log(p/(1 - p))$, et d'ensuite remettre les bornes de l'intervalle sur l'échelle originale. Il serait possible d'estimer λ par $\hat{\lambda} = \log(\hat{p}/(1 - \hat{p}))$. Par contre, l'estimateur ne serait pas défini lorsque \hat{p} serait nul ou égal à 1. L'une des solutions possibles a été amenée par [Anscombe, 1956] qui suggère l'estimateur

$$\hat{\lambda} = \log\left(\frac{X + \frac{1}{2}}{n - X + \frac{1}{2}}\right) \quad (3.2.1)$$

avec X représentant le nombre de succès dans l'échantillon de taille n . L'intervalle de confiance de niveau $1 - \alpha$ prend la forme

$$\left[\frac{\exp(\lambda_l)}{1 + \exp(\lambda_l)} ; \frac{\exp(\lambda_u)}{1 + \exp(\lambda_u)} \right] \quad (3.2.2)$$

où $\lambda_l = \hat{\lambda} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\lambda})}$, $\lambda_u = \hat{\lambda} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\lambda})}$ et où

$$\hat{V}(\hat{\lambda}) = \frac{(n + 1)(n + 2)}{n(X + 1)(n - X + 1)}. \quad (3.2.3)$$

3.3. Intervalle de Wilson

L'intervalle de confiance de Wald est très simple à construire, mais il tend à donner des résultats décevants sauf si n est grand. L'une des solutions à ce problème a été discutée par [Wilson, 1927] où il utilise l'erreur standard exacte basée sur la proportion inconnue p au lieu de l'erreur standard estimée basée sur la proportion estimée \hat{p} utilisée pour standardiser la variable aléatoire $(\hat{p} - p)$ dans la construction de l'intervalle de Wald. C'est donc dire que les bornes de l'intervalle de confiance de Wilson sont les solutions aux équations :

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2} \quad (3.3.1)$$

ce qui mène à l'intervalle de confiance de Wilson de niveau $1 - \alpha$:

$$\hat{p}_W \pm \frac{z_{\alpha/2}\sqrt{n}}{n_W} \sqrt{\hat{p}(1 - \hat{p}) + \frac{z_{\alpha/2}^2}{4n}} \quad (3.3.2)$$

où

$$\hat{p}_W = \frac{X + z_{\alpha/2}^2/2}{n_W} \quad (3.3.3)$$

et $n_W = n + z_{\alpha/2}^2$ respectivement l'estimateur de proportion et la « taille de l'échantillon » utilisés dans la construction de l'intervalle de confiance de Wilson. Notez que si l'on désire construire un intervalle de confiance bilatéral de niveau 95%, comme c'est le cas ici, le quantile d'ordre 95% d'une loi normale est $z_{0,025} = 1,96$. En arrondissant ce quantile, c'est-à-dire en considérant par exemple $z_{\alpha/2} = 2$, l'estimateur \hat{p}_W devient l'estimateur \tilde{p} , l'estimateur introduit par Agresti-Coull défini en (1.2.1). C'est d'ailleurs la forme de cet intervalle qui a suggéré l'utilisation de l'estimateur \tilde{p} dans la construction d'intervalles de confiance bootstrap.

3.4. Intervalle Agresti-Coull

L'intervalle de confiance Agresti-Coull a été introduit par [Agresti et Coull, 1998] et il a la même forme que l'intervalle de confiance standard de Wald, mais il est basé sur l'estimateur de la méthode de Wilson soit \hat{p}_W . En d'autres mots, l'intervalle Agresti-Coull est un intervalle de Wald où l'on remplace \hat{p} par \hat{p}_W et n par n_W . L'intervalle de confiance d'Agresti-Coull de niveau $1 - \alpha$ prend donc la forme :

$$\hat{p}_W \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_W(1 - \hat{p}_W)}{n_W}}. \quad (3.4.1)$$

Les intervalles de Wilson et d'Agresti-Coull sont centrés au même point \hat{p}_W . Il est possible de démontrer que les intervalles de confiance de Agresti-Coull ne sont toutefois jamais plus courts que ceux de Wilson.

3.5. Intervalle Clopper-Pearson

[Clopper et Pearson, 1934] suggère un intervalle de confiance qui a des propriétés très conservatrices alors que la probabilité de couverture n'est jamais plus petite que la valeur nominale de couverture (voir [Brown *et al.*, 2001]). L'intervalle de confiance de niveau $1 - \alpha$ est défini avec les bornes inférieure $L(X, n)$ et supérieure $U(X, n)$:

$$L(X, n) = \frac{\nu_1 F_{\nu_1, \nu_2}(\alpha/2)}{\nu_2 + \nu_1 F_{\nu_1, \nu_2}(\alpha/2)}; \quad (3.5.1)$$

$$U(X,n) = \frac{\nu_3 F_{\nu_3, \nu_4}(1 - \alpha/2)}{\nu_4 + \nu_3 F_{\nu_3, \nu_4}(1 - \alpha/2)} \quad (3.5.2)$$

où $\nu_1 = 2X$, $\nu_2 = 2(n - X + 1)$, $\nu_3 = 2(X + 1)$, $\nu_4 = 2(n - X)$ et $F_{d_1, d_2}(\alpha)$ est le quantile d'ordre α de la loi de Fisher avec degrés de liberté d_1 et d_2 . Il est important de noter ici que $L(X,n) = 0$ si $\hat{p} = 0$ et $U(X,n) = 1$ si $\hat{p} = 1$.

3.6. Intervalle Arcsine Square Root

L'intervalle de confiance Arcsine Square Root utilise $\arcsine\sqrt{p}$ comme une transformation stabilisatrice de la variance de (dans ce cas-ci) $\hat{p} = (X + 0,5)/(n + 1)$ afin de construire des intervalles plus conservateurs que ceux provenant de la méthode de Wald [Gilary *et al.*, 2012]. L'intervalle de confiance Arcsine Square Root de niveau $1 - \alpha$ est $[L; U]$ avec respectivement la borne inférieure (L) et supérieure (U) données par

$$L = \sin^2 \left\{ \max \left(0, \arcsin \sqrt{\frac{X + 0,5}{n + 1} - \frac{z}{\sqrt{4n}}} \right) \right\}, \quad (3.6.1)$$

$$U = \sin^2 \left\{ \min \left(\frac{\pi}{2}, \arcsin \sqrt{\frac{X + 0,5}{n + 1} + \frac{z}{\sqrt{4n}}} \right) \right\}.$$

3.7. Changements à apporter dans un contexte d'échantillonnage

Les intervalles de confiance présentés précédemment sont ceux utilisés lorsque l'échantillon est généré selon une loi Bernoulli avec n essais dont X succès. Par contre, il est important de considérer le facteur $(1 - f_s)$ dans les termes de variance lorsqu'on utilise le plan aléatoire simple sans remise afin de tirer notre échantillon. C'est pourquoi, en s'inspirant de ce que [Korn et Graubard, 1998] ainsi que [Franco *et al.*, 2019] proposent afin d'obtenir les intervalles de confiance de ces méthodes sous le plan aléatoire simple sans remise, il est possible de montrer qu'il suffit de remplacer n par $\hat{n}_{eff} = (n_s - 1)/(1 - f_s)$ où n_s est la taille de l'échantillon s et $f_s = n_s/N$, la fraction de sondage et de remplacer X par $\hat{p} \cdot \hat{n}_{eff}$ afin d'obtenir les intervalles de confiance lorsque le plan aléatoire simple sans remise est utilisé.

Chapitre 4

Résultats

Afin d'évaluer la performance des différentes méthodes, nous avons effectué des simulations. Nous utiliserons la couverture ainsi que la longueur moyenne des intervalles de confiance afin d'évaluer leur efficacité. Nous comparerons les différentes méthodes entre elles avec différentes tailles échantillonnales et pour différentes valeurs de p . Dans un premier temps, nous présenterons les détails de la simulation puis, nous présenterons les résultats des méthodes dans un contexte i.i.d. où les données de l'échantillon sont générées de manière indépendante les unes des autres selon une loi Bernoulli(p). Finalement, nous nous pencherons sur les résultats des méthodes dans un contexte d'échantillonnage où les données de l'échantillon ont été prélevées de la population à partir d'un plan aléatoire simple sans remise.

4.1. Méthodologie de la simulation

Il était important de tout d'abord fixer la taille des échantillons que nous utilisons pour nos simulations. Nous avons décidé d'utiliser deux tailles échantillonnales soit de 40 et de 200 observations. Par contre, dans un contexte d'échantillonnage, il est nécessaire et habituel de considérer un paramètre additionnel, soit la fraction de sondage $f_s = n_s/N$ lorsqu'on établit nos différents scénarios. En effet, le fait d'avoir une fraction de sondage qui est négligeable, c'est-à-dire près de 0 ou non, a définitivement un impact sur les propriétés des estimateurs. Nous définissons donc 2 fractions de sondage, soit 0,05 et 0,20. Il est important de noter que la taille de la population créée dépendra de la taille de l'échantillon, n , ainsi que de la fraction de sondage, f . Nous avons donc 2 scénarios pour un contexte i.i.d. et 4 scénarios,

soit la combinaison de nos tailles échantillonales et de nos fractions de sondage, pour un contexte d'échantillonnage.

Comme les propriétés des méthodes dépendent de la proportion de succès, chaque scénario sera repris pour chaque valeur de p . Comme plusieurs études l'ont fait précédemment dont celle de [Mantel et Neusy, 2015], nous avons utilisé les proportions de 1% jusqu'à la proportion de 99% par saut de 1% pour les valeurs de p . Par conséquent, pour chaque valeur de p , un échantillon de taille n a été généré selon une Bernoulli(p) dans un contexte i.i.d. puis dans un cadre d'échantillonnage, un échantillon de taille n a été tiré à partir d'un plan aléatoire simple sans remise d'une population de taille N où l'on y retrouve exactement pN succès. Notons que les valeurs de f et n considérées font en sorte que les valeurs de N sont des multiples de 100 de telle sorte que pN est toujours un entier.

Pour les méthodes bootstrap, nous avons répété $B = 1\ 000$ fois la génération ou le tirage d'échantillons bootstrap, dépendamment que nous soyons respectivement dans un contexte i.i.d. ou dans un cadre d'échantillonnage avec un plan aléatoire simple sans remise. Dans la méthode des poids, nous utilisons comme taille d'échantillon bootstrap $n' = n - 1$. Ce choix satisfait l'inégalité (2.4.17) qui garantit un estimateur de variance non nul.

Nous avons répété 10 000 fois la génération ou le tirage de nos échantillons, selon le contexte, pour chaque valeur de p . Un intervalle de confiance pour chaque méthode est ainsi créé pour chacun des 10 000 échantillons.

Le but ultime étant d'évaluer la qualité des méthodes de construction d'intervalles de confiance, les deux critères que nous utilisons sont la couverture de l'intervalle de confiance dont la valeur nominale est de 95% ainsi que la longueur moyenne des intervalles de confiance. La couverture est la proportion des 10 000 échantillons pour laquelle l'intervalle de confiance contient la valeur de p tandis que la longueur moyenne représente la moyenne des longueurs des 10 000 intervalles de confiance créés pour chaque p . Comme certains intervalles mènent à des bornes inférieures à 0 ou supérieures à 1 alors qu'une proportion est limitée à l'intervalle $[0,1]$, en pratique on ne rapporterait pas ces bornes. On rapporterait plutôt l'intersection de l'intervalle calculé et l'intervalle $[0,1]$ qui respecte les limites d'une proportion. En conséquence, la longueur moyenne calculée est celle des intervalles tronqués.

C'est donc en se servant de la couverture et de la longueur moyenne des intervalles de confiance que nous jugerons de leur performance. Pour la couverture, le meilleur comportement aura généralement lieu pour les valeurs centrales de p . Avoir une couverture près de 95% pour ces valeurs plutôt que constamment au-dessus ou constamment au-dessous sera privilégié. En effet, pour ces valeurs de p , la couverture et la longueur moyenne des intervalles sont très liées. Ainsi, on ne privilégiera pas un intervalle dont la couverture est trop petite, pas plus qu'un intervalle qui couvre trop souvent parce que dans ce dernier cas, ce sera associé à un intervalle trop large. Dans les extrémités, la couverture des intervalles aura souvent de grandes variations. Des intervalles avec une couverture minimale faible seront désavantagés. Si la couverture est bien plus grande que 95%, il faut regarder la longueur moyenne de l'intervalle. Contrairement aux valeurs centrales de p , dans les extrémités, certains intervalles ont à peu près la même longueur avec certains qui couvrent la vraie valeur beaucoup plus souvent (incluant des couvertures supérieures à 95%). Ainsi, à longueur égale, il est préférable que l'intervalle couvre plus souvent que moins souvent.

4.2. Résultats sous un contexte i.i.d.

Débutons notre analyse du contexte i.i.d. en considérant les résultats des 4 méthodes bootstrap avec l'utilisation de la distribution de $F_{\hat{p}}$ afin de générer les échantillons bootstrap, ce qu'on appelle la méthode bootstrap à la Agresti-Coull.

Chacun des 4 graphiques de la figure 4.1 montre la couverture, c'est-à-dire la proportion des 10 000 échantillons dont l'intervalle de confiance créé contient la valeur de p , et ce, pour les différentes valeurs de p . Nous avons donc les courbes de couverture pour chacune des méthodes bootstrap où le bootstrap est effectué à la Agresti-Coull et où la taille de l'échantillon initial est de 40 observations. Chacun de ces graphiques contient également une ligne horizontale pointillée noire qui représente la valeur nominale de 95%. Deux lignes pointillées grises plus fines sont ajoutées un peu au-dessus et au-dessous de cette dernière et représentent les bornes de l'intervalle de confiance créé pour la valeur nominale. Avec 10 000 répétitions, la couverture d'un intervalle de confiance de niveau 95% devrait se situer entre 94,57% et 95,43%.

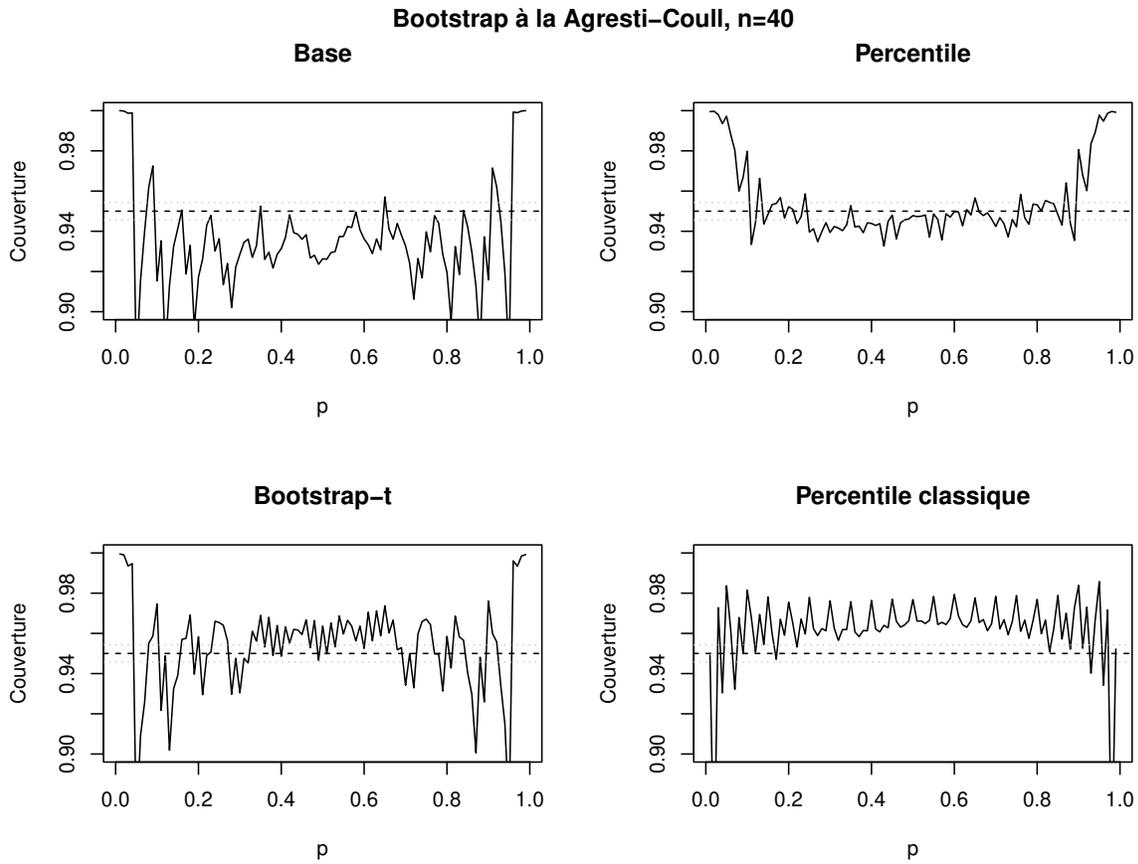


Figure 4.1. Couverture des méthodes bootstrap avec l’utilisation du bootstrap à la Agresti–Coull sous un contexte i.i.d. avec $n = 40$.

L’une des premières choses que l’on remarque en regardant la figure 4.1 est que la méthode bootstrap percentile semble être la meilleure des 4 méthodes bootstrap en terme de couverture du paramètre de proportion puisqu’elle est la seule dont la couverture se rapproche de la valeur nominale de 95% surtout lorsque p prend des valeurs entre 0,1 et 0,9. On peut voir que la couverture remonte un peu dans les extrémités avant d’atteindre une couverture de 100%. De plus, les résultats de la simulation nous montrent que la méthode percentile a une couverture minimale de 93,27%, comparativement à 86,02% pour la méthode bootstrap- t , 87,05% pour la méthode bootstrap de base et 83,71% pour la méthode percentile classique. Comme nous le verrons, toutes les méthodes ont une probabilité de couverture qui oscille en fonction de la vraie valeur de p à cause du caractère discret du nombre de succès. Pour plus de détails sur le phénomène d’oscillation, consultez [Brown *et al.*, 2001]. La pire méthode bootstrap semble être la méthode bootstrap de base qui, en plus d’avoir une courbe

de couverture qui oscille énormément, n'atteint que très rarement la valeur nominale de 95% en étant généralement en-dessous. L'intervalle de confiance bootstrap- t fait un peu mieux en terme de couverture que l'intervalle bootstrap de base surtout lorsque p prend des valeurs entre 0,37 et 0,63, mais demeure inefficace pour couvrir p de manière générale. L'intervalle de type percentile classique a une couverture la plupart du temps au-dessus de la valeur nominale, mais sa couverture redescend vraiment plus bas que la valeur nominale pour p près de 0 et 1 avant de remonter. L'intervalle percentile offre une couverture qui oscille moins que les 3 autres méthodes bootstrap en plus d'être très près du 95% de couverture, surtout lorsque p prend des valeurs entre 0,1 et 0,9.

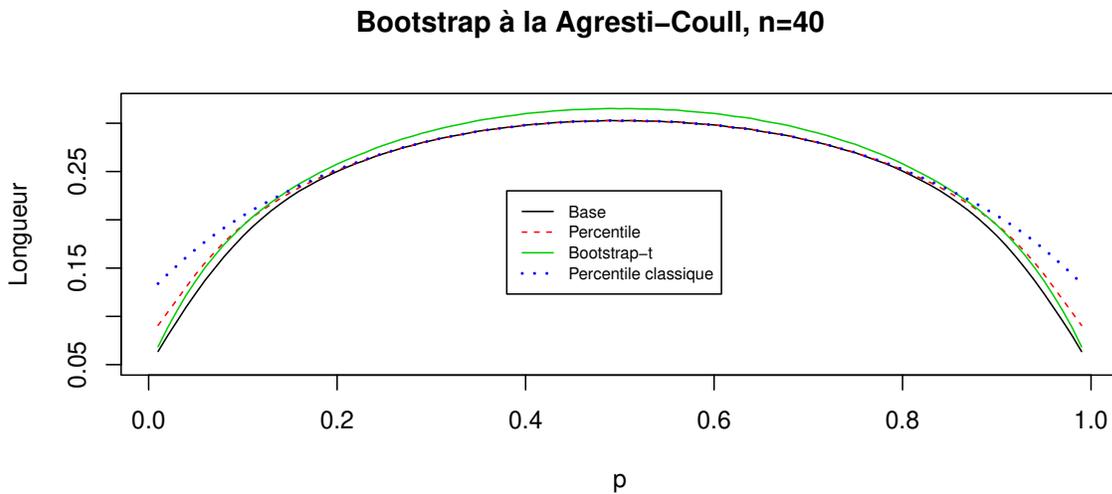


Figure 4.2. Longueur moyenne des intervalles bootstrap à l'aide du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 40$.

La figure 4.2 nous permet de voir la longueur moyenne des 10 000 intervalles de confiance pour chaque méthode en fonction de p . Ce qui est important de noter sur cette figure est que, malgré le fait que la méthode percentile a une couverture bien supérieure à 95% dans les extrémités, cette dernière offre tout de même des intervalles de confiance presque aussi courts que ceux des méthodes de bootstrap de base et bootstrap- t .

En résumé, la méthode percentile a une couverture très près de la valeur nominale pour beaucoup de valeurs de p en plus d'avoir une couverture minimale bien supérieure aux 3 autres méthodes bootstrap (93,27%). En plus, dans les extrémités, la méthode percentile offre une couverture bien supérieure à 95% (qui atteint 100% près de 0 et 1), mais en ayant des

intervalles d'une longueur qui se compare avantageusement aux 3 autres méthodes bootstrap.

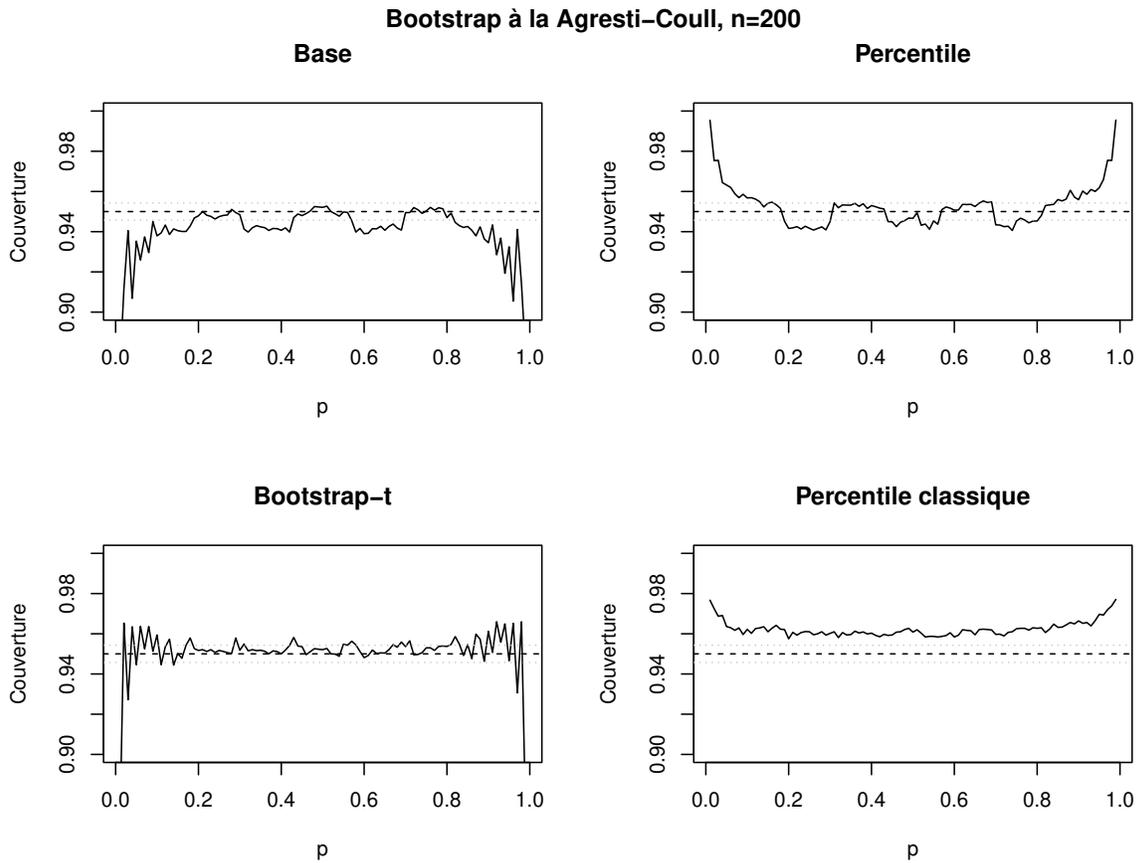


Figure 4.3. Couverture des méthodes bootstrap avec l'utilisation du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 200$.

Intéressons-nous maintenant au cas où l'on conserve la même distribution afin de générer les échantillons bootstrap, mais qu'on considère une taille échantillonnale de 200 observations. Rappelons qu'idéalement, une bonne méthode nous donnerait une couverture qui s'approche le plus de la valeur nominale de 95% sans descendre trop bas surtout pour des valeurs centrales de p et que dans les extrémités, une couverture bien supérieure à 95% ne soit pas conséquence d'un intervalle de confiance trop long. Sur cette base, on peut voir que les méthodes bootstrap de base et bootstrap- t sont à éviter, puisque pour des valeurs de p près de 0 et 1, la couverture est bien en deçà de 95% (voir figure 4.3). La simulation nous indique que les méthodes bootstrap de base et bootstrap- t ont une couverture minimale de 87,02% et de 86,03% respectivement comparativement à 94,06% pour la méthode percentile et 95,75%

pour la méthode percentile classique. La méthode percentile classique est toujours supérieure à la valeur nominale surtout dans les extrémités. La méthode percentile a une couverture qui est la plupart du temps très près de la valeur nominale (souvent pas statistiquement différente de 95%) pour p compris entre 0,1 et 0,9.

À l'aide de la figure 4.4, on peut voir que malgré que la méthode percentile ait une couverture qui monte jusqu'à 100% dans les extrémités (bien plus haut que la couverture de la méthode percentile classique), elle offre tout de même des intervalles de confiance en moyenne plus courts que ceux de la méthode percentile classique. Rappelons qu'en plus d'être le candidat idéal pour un échantillon de taille 200, l'intervalle percentile est celui qui performe le mieux lorsque la taille échantillonnale est de 40 observations. C'est pourquoi cette dernière méthode serait un bon candidat dans un contexte i.i.d. avec l'utilisation d'un bootstrap à la Agresti-Coull pour estimer via intervalle de confiance un paramètre de proportion peu importe la taille échantillonnale.

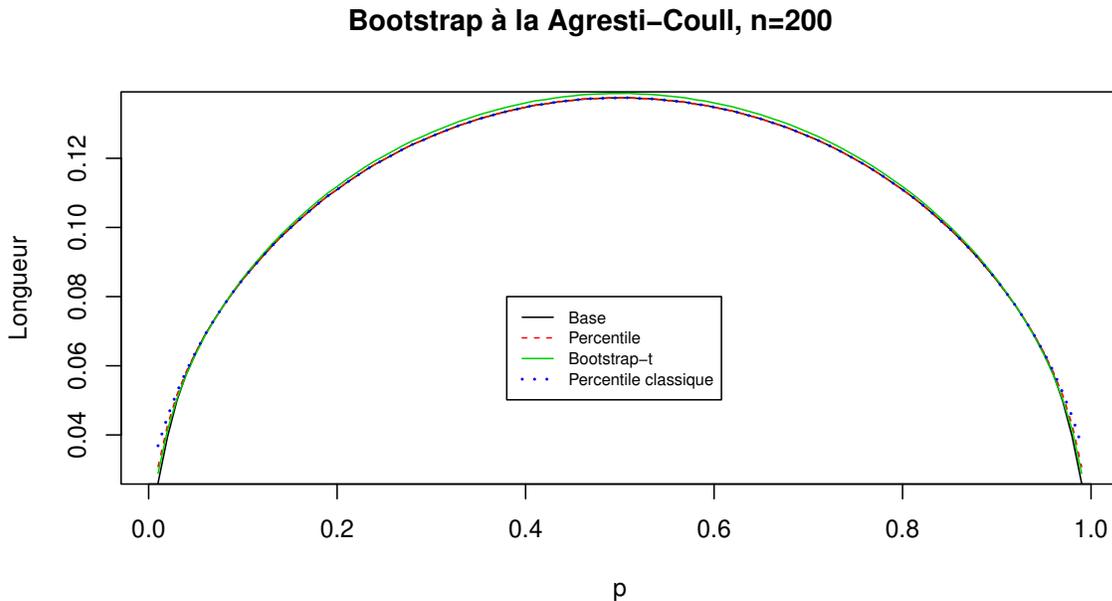


Figure 4.4. Longueur moyenne des intervalles bootstrap à l'aide du bootstrap à la Agresti-Coull sous un contexte i.i.d. avec $n = 200$.

En consultant la figure 4.5, on constate que le bootstrap adaptatif, qui consiste à considérer un bootstrap selon les distribution de $F_{\hat{p}}$ ou $F_{\hat{p}}$ selon qu'il y ait respectivement aucun

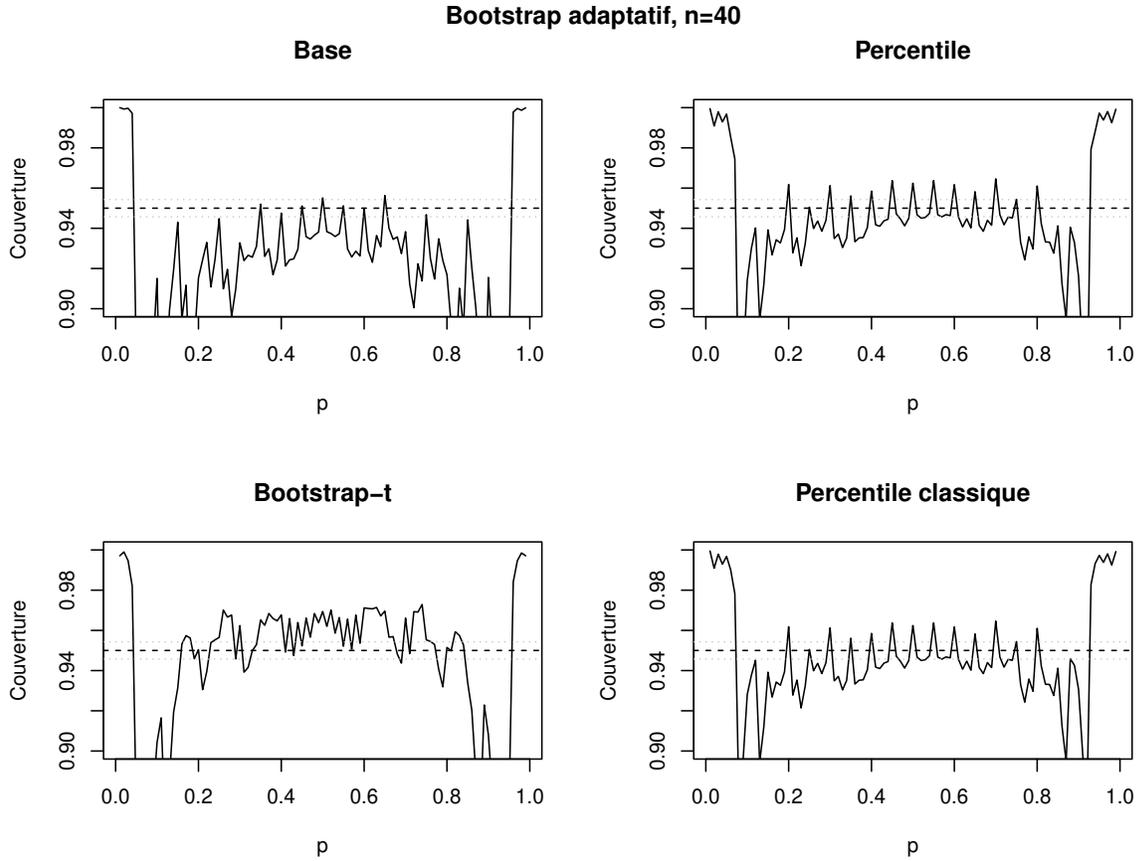


Figure 4.5. Couverture des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 40$.

succès (ou aucun échec) ou au moins un succès (et au moins un échec), ne réussit pas à bien s'adapter lorsque la valeur de p est près de 0 et 1 pour $n = 40$. Toutes les méthodes bootstrap démontrent beaucoup de difficultés à bien couvrir le paramètre d'intérêt dans ces circonstances. En effet, les 4 méthodes ont une couverture qui descend très bas avant de remonter vers une couverture de 100% dans les extrémités. Les résultats nous indiquent que les méthodes percentile et percentile classique ont une couverture minimale de 83,57% et de 87,13% respectivement et qu'elles sont bien supérieures aux couvertures minimales des méthodes bootstrap de base, 69,99% et bootstrap- t , 68,75%. Sinon, les intervalles bootstrap de base semblent trop souvent en-dessous de la valeur nominale de 95% tandis que l'intervalle bootstrap- t couvre trop le paramètre d'intérêt lorsque p prend des valeurs entre 0,3 et 0,7.

Sur la figure 4.6, on peut y voir 2 graphiques superposés où le graphique du haut représente les couvertures des différentes méthodes avec la longueur moyenne des intervalles

sur le graphique du bas. En y jettant un coup d'oeil, on peut voir que la couverture des méthodes percentile ainsi que percentile classique sont identiques pour beaucoup de valeurs de p (les lignes pointillées bleue et rouge de la figure 4.6). En effet, les deux couvertures sont identiques pour tous les p sauf pour une proportion comprise entre 0,06 et 0,12 puis entre 0,88 et 0,94. Rappelons que l'intervalle de confiance percentile est égal à l'intervalle de confiance percentile classique seulement lorsque l'on effectue un bootstrap paramétrique en utilisant la distribution de $F_{\hat{p}}$. En effet, comme l'estimateur \hat{p} est l'estimateur \hat{p} lorsque ce dernier n'est ni 0 ni 1, il va de soi que les méthodes percentile et percentile classique donnent les mêmes intervalles de confiance lorsque p est suffisamment plus grand que 0 (plus grand que 0,12 dans ce cas-ci) ou lorsqu'il est suffisamment plus petit que 1 (plus petit que 0,88 dans ce cas-ci). L'estimateur calculé à partir de l'échantillon généré dans ces cas est différent de 0 ou 1 avec une très grande probabilité.

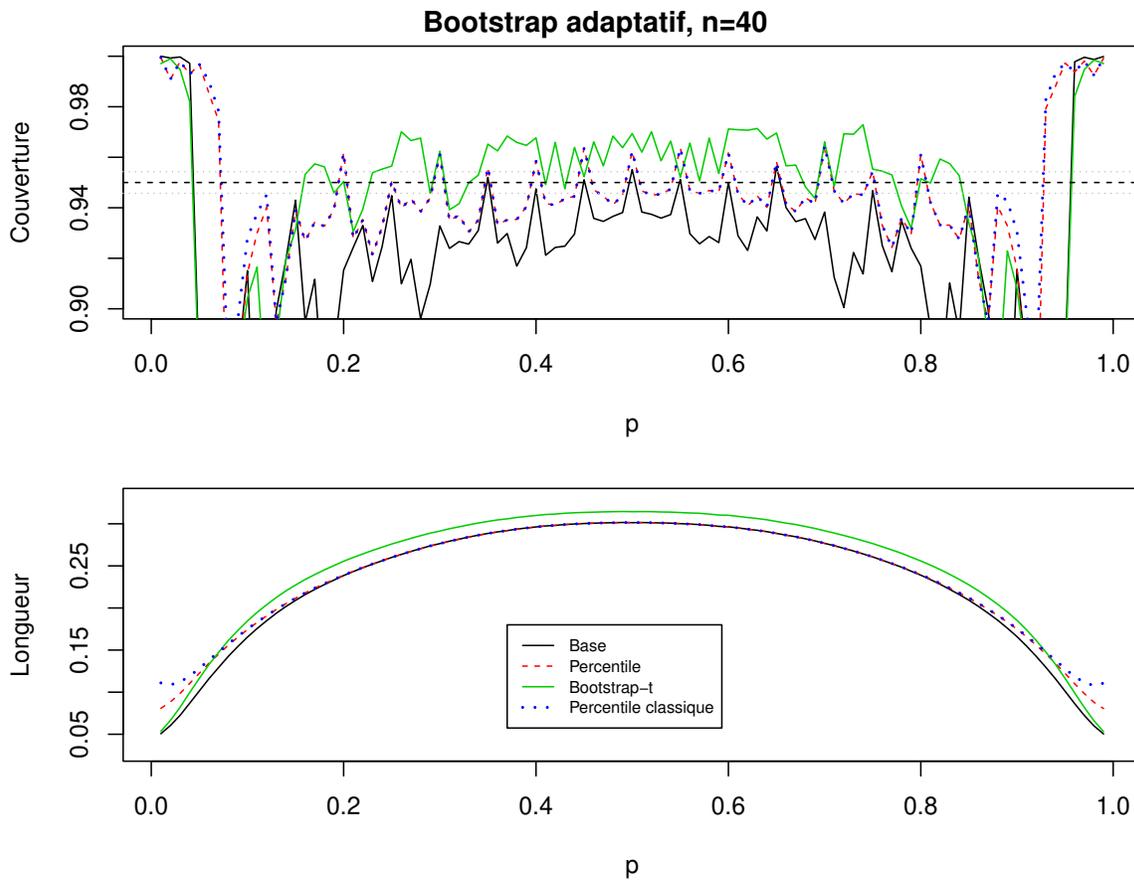


Figure 4.6. Couverture et longueur moyenne des intervalles des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 40$.

Autre fait à noter concernant les intervalles percentile et percentile classique est la longueur moyenne des intervalles. Les résultats de la simulation nous montrent que la longueur de ces deux intervalles est égale pour p entre 0,24 et 0,76, mais la méthode percentile offre des intervalles plus courts que la méthode percentile classique pour des proportions plus petites que 0,24 et plus grandes que 0,76. Rappelons que nous avons décidé de considérer les intervalles de confiance tronqués. En effet, comme ces intervalles sont ceux pour un paramètre de proportion et qu'il est naturel d'avoir un estimateur prenant une valeur entre 0 et 1, l'intervalle tronqué correspond donc à l'intersection entre l'intervalle de confiance obtenu et l'intervalle $[0,1]$. Si l'intervalle percentile classique n'est jamais tronqué, dû à la forme de son intervalle, l'intervalle percentile de son côté peut l'être.

Prenons l'exemple où \hat{p} est 0. Nous générons nos échantillons bootstrap selon l'estimateur $\tilde{p} = 2/(n + 4)$, qui lui, est nécessairement non nul. Puisque pour $n = 40$, la probabilité de succès est de $2/44$, la probabilité de n'avoir aucun succès parmi 40 essais est de $(1 - 2/44)^{40} = 0,16$ et il ne serait pas surprenant d'avoir plusieurs échantillons bootstrap n'ayant aucun succès, d'où \hat{p}^* serait nul. C'est donc dire que le quantile $\alpha/2$ de la distribution estimée de $J_n(x, F_{\tilde{p}}) = Prob(\hat{p}^* - \tilde{p} \leq x)$ peut être plus petit que 0. Par conséquent, la borne inférieure de l'intervalle percentile $[\hat{p} + J_n^{-1}(\alpha/2, \hat{F}), \hat{p} + J_n^{-1}(1 - \alpha/2, \hat{F})]$ (1.3.7) serait donc plus petite que 0. Maintenant, comme nous l'avons mentionné plus tôt, cet intervalle serait tronqué de telle sorte que la borne inférieure serait égale à 0. Puis comme nous tronquons cet intervalle dans ce cas précis, la longueur de celui-ci serait plus courte que l'intervalle percentile classique, où les bornes de ce dernier sont toujours entre 0 et 1. En effet, si l'intervalle percentile n'était pas tronqué alors la longueur des deux intervalles serait identique, soit $J_n(1 - \alpha/2, \hat{F}) - J_n(\alpha/2, \hat{F})$. Le même raisonnement existe avec la borne supérieure de l'intervalle de confiance percentile lorsque \hat{p} est égal à 1. Il en demeure pas moins que l'on peut conclure que les méthodes percentile et percentile classique ont une efficacité similaire puisqu'elles offrent la même couverture sauf pour quelques valeurs de p où pour ces valeurs, la différence en terme de couverture est minime.

Lorsqu'on considère des échantillons de taille 200, nous nous apercevons (voir figure 4.7) que la couverture des intervalles percentile ainsi que percentile classique est un peu plus haute que la valeur nominale de 95%, en étant près de la borne supérieure de l'intervalle de confiance pour la valeur nominale, pour beaucoup de p entre 0,05 et 0,95. Les intervalles

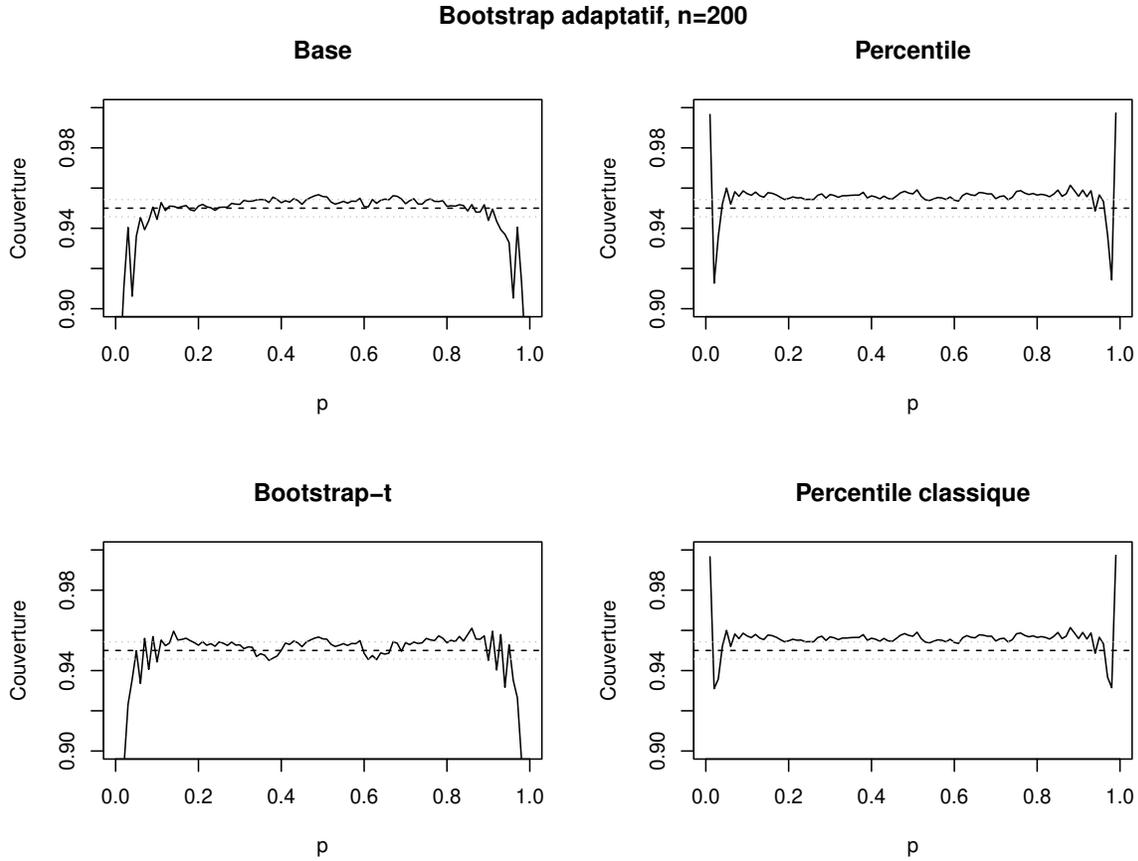


Figure 4.7. Couverture des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 200$.

bootstrap- t et de base performent un peu mieux que les deux intervalles percentiles pour des valeurs de p comprises entre 0,1 et 0,9 avec une légère préférence pour la méthode de base dont la couverture est plus stable. Ces deux méthodes performent bien pour ces valeurs de p , mais démontrent de la difficulté à bien couvrir le paramètre d'intérêt dans les extrémités, ce que percentile et percentile classique font mieux. En effet, les couvertures minimales des méthodes bootstrap de base et bootstrap- t sont de 87,02% et 85,45% respectivement en comparaison aux méthodes percentile et percentile classique qui offrent une couverture minimale de 91,27% et 93,10% respectivement.

Lorsqu'on prête attention aux couvertures des intervalles de confiance percentile et percentile classique sur la figure 4.8, on s'aperçoit que lorsque p prend des valeurs près de 0 et 1, les deux courbes descendent un peu avant de remonter vers une couverture près de 100%. Toutefois, en regardant la figure 4.7, on peut voir que la méthode percentile classique

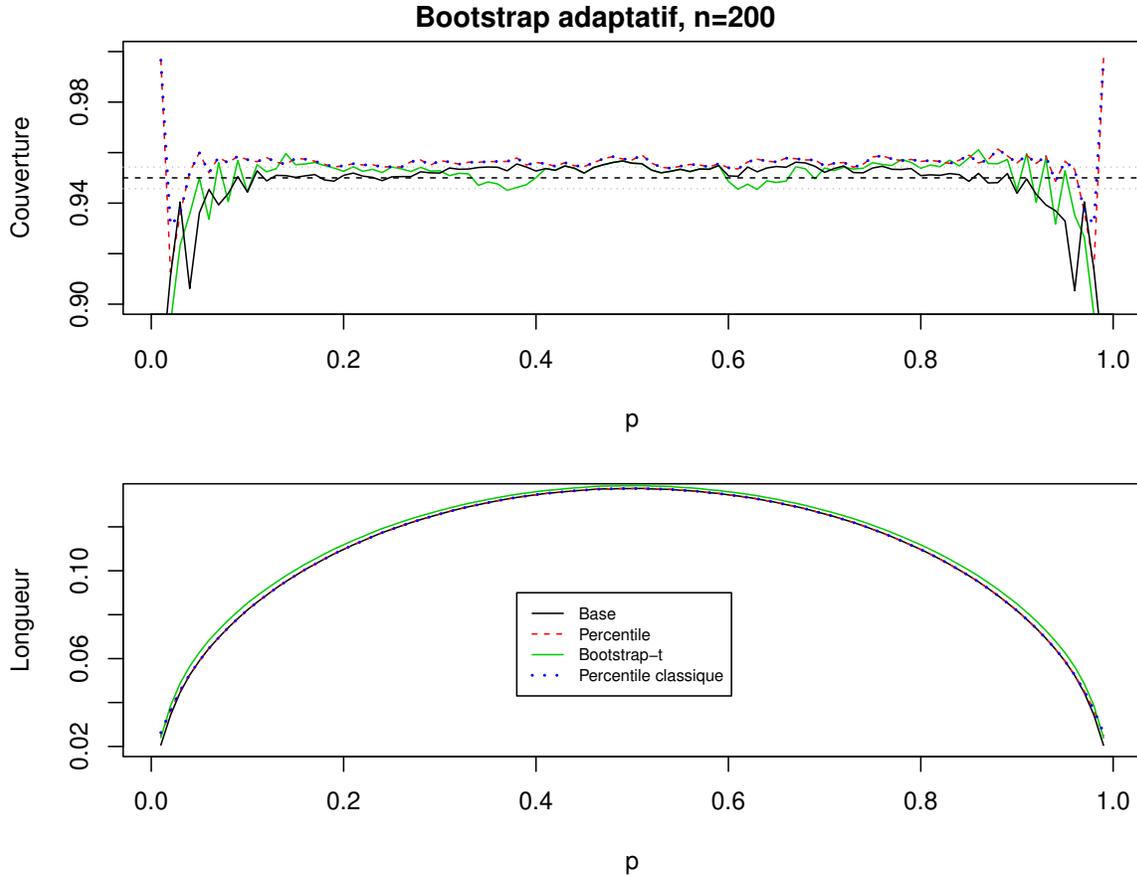


Figure 4.8. Couverture et longueur moyenne des intervalles des méthodes bootstrap avec l'utilisation du bootstrap adaptatif sous un contexte i.i.d. avec $n = 200$.

performe un peu mieux puisque la courbe de cette dernière descend moins bas que celle de la méthode percentile. On peut toutefois affirmer sans se tromper que les deux méthodes offrent des performances extrêmement similaires si ce n'est que pour 2 valeurs de p . Même si cette dernière performe moins bien que les méthodes bootstrap de base et bootstrap- t pour des proportions allant de 0,1 à 0,9, on note à nouveau que la couverture des méthodes percentile et percentile classique demeurent à la limite de la borne supérieure de l'intervalle de confiance créé pour la valeur nominale pour ces mêmes valeurs de p . En d'autres mots, même si percentile et percentile classique ont une couverture supérieure à la valeur nominale de 95%, leurs couvertures demeurent très près de celle-ci et font bien près de 0 et 1.

Si en considérant le bootstrap à la Agresti-Coull il était clair que la méthode percentile était la meilleure méthode, dans le cas d'un bootstrap adaptatif, les méthodes percentile et percentile classique offrent des performances très similaires. En effet, comme nous l'avons

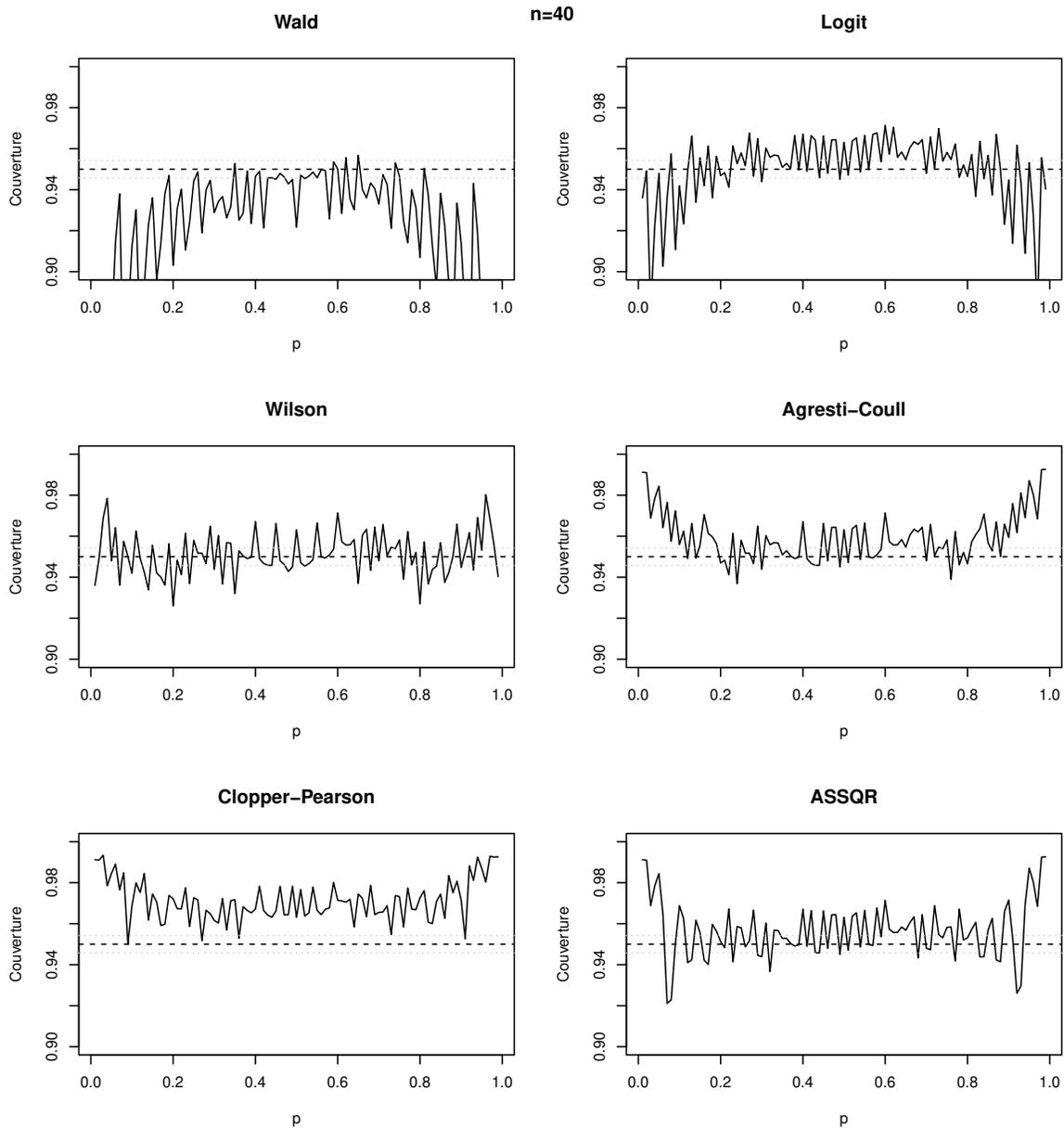


Figure 4.9. Couverture des méthodes traditionnelles dans un contexte i.i.d. avec $n = 40$.

vu, lorsque la taille de l'échantillon est de 40 observations, les deux méthodes offrent la même couverture pour beaucoup de valeurs de p . La méthode percentile classique offre une couverture minimale un peu plus élevée que la méthode percentile (de 3,56 % de différence), mais c'est celle-ci qui offre les intervalles les plus courts près des extrémités. Pour ces raisons, un petit avantage va à la méthode percentile. Lorsque $n = 200$, il est encore plus difficile de déterminer une méthode plus performante entre les deux. En effet, pour toutes les valeurs de

p , sauf pour 2 valeurs, les deux méthodes offrent la même couverture. Encore ici, la méthode percentile classique offre une couverture minimale un peu plus élevée (différence que de 1,83%) avec des intervalles de longueurs extrêmement comparables. À cause du fait que la méthode percentile est notre candidat final pour estimer un paramètre de proportion pour un bootstrap adaptatif avec une taille échantillonnale de 40 observations, nous considérerons la méthode percentile aussi lorsque la taille devient de 200 observations puisque dans tous les cas, la méthode percentile performe bien.

Après avoir analysé les résultats des méthodes bootstrap, jetons un coup d'oeil sur la performance des méthodes traditionnelles. On peut remarquer sur la figure 4.9 que la méthode de Wald ne performe pas très bien puisqu'elle a une couverture qui n'atteint que très rarement la valeur nominale de 95%. Cette constatation a été faite dans plusieurs études dans le passé dont dans celle de [Brown *et al.*, 2001]. En effet, l'intervalle de Wald est basé sur le théorème de la limite centrale et pour n petit, comme ici, avec en plus une proportion près de 0 et 1, la distribution normale n'est pas une bonne estimation de la distribution de \hat{p} . De son côté, la couverture des intervalles de confiance de Clopper-Pearson est toujours au-dessus de la valeur nominale, il n'est donc pas surprenant que cette méthode donne des intervalles de confiance très longs (figure 4.10). Une autre méthode qui construit visiblement des intervalles de confiance très longs dans les extrémités est la méthode Agresti-Coull et qui offre une couverture souvent supérieure à 95% lorsque p prend des valeurs près de 0 et 1.

La meilleure des méthodes semble être la méthode de Wilson qui a une couverture très près de 95% et ce peu importe la valeur de p (avec une couverture minimale de 92,60%) tout en offrant des intervalles de confiance de longueurs comparables à ceux des méthodes Logit et ASSQR. Ces deux méthodes nous donnent des couvertures assez similaires lorsque p prend des valeurs entre 0,3 et 0,7. Pour la méthode Logit, la couverture minimale est de 88,03%. De son côté, la méthode ASSQR assure une couverture minimale de 92,11%. La couverture de la méthode ASSQR remonte dans les extrémités pour atteindre une couverture de près de 100%, tout en ayant des intervalles de confiance plus courts que la méthode de Wilson pour ces valeurs de p . Par conséquent, pour des petites ou grandes valeurs de p , la méthode ASSQR serait préférable à la méthode de Wilson puisqu'elle couvre davantage en ayant des intervalles plus courts. La méthode de Wilson demeure tout de même la méthode à favoriser

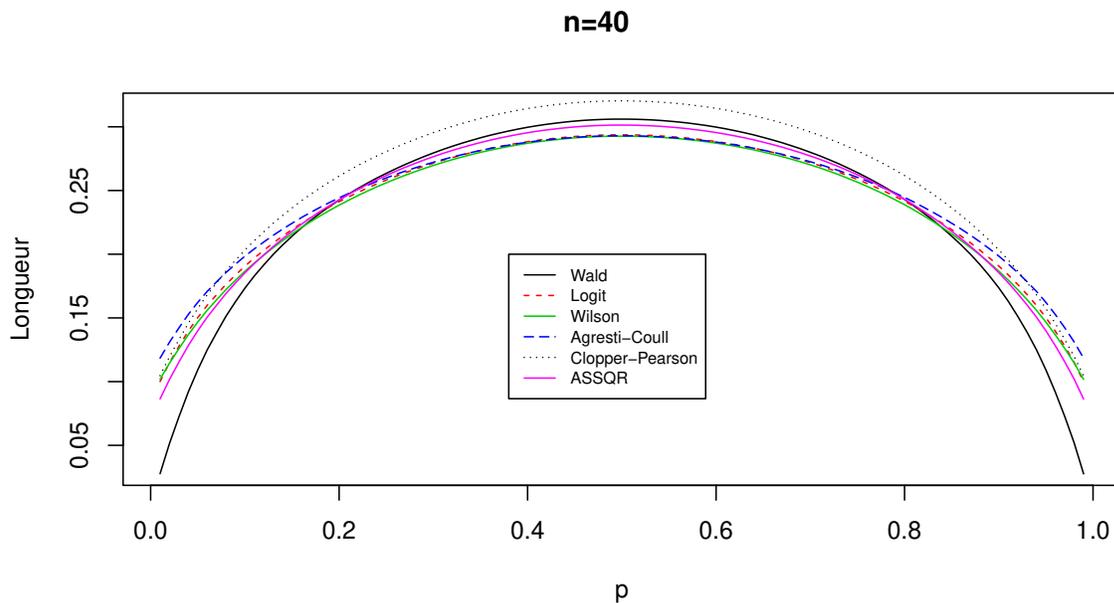


Figure 4.10. Longueur moyenne des méthodes traditionnelles dans un contexte i.i.d. avec $n = 40$.

dans ce contexte grâce à une meilleure stabilité et à une couverture minimale supérieure à celle de la méthode ASSQR.

Même en considérant un échantillon initial de taille 200 au lieu de seulement 40 observations, nous avons le même phénomène qui se réalise pour la méthode de Clopper-Pearson dont la couverture dépasse 95% pour tous les p tel que le prévoyait la théorie (voir figure 4.11). Remarquons que pour Wald, la méthode performe bien pour p entre 0,05 et 0,95 dans ce contexte, mais la couverture redescend très bas dans les extrémités.

On s’attendait à voir les courbes de couverture osciller beaucoup moins pour un échantillon de taille 200 que pour un échantillon de taille 40. C’est effectivement ce que nous constatons pour la méthode Logit ainsi que ASSQR. On voit le même phénomène pour la méthode de Wilson, mais pour cette dernière, la différence n’est pas aussi importante que pour les deux autres. D’ailleurs, en comparant ces trois méthodes sur un même graphique (figure 4.12), on s’aperçoit justement que les méthodes Logit et ASSQR ont une couverture un peu plus constante autour de la valeur nominale.

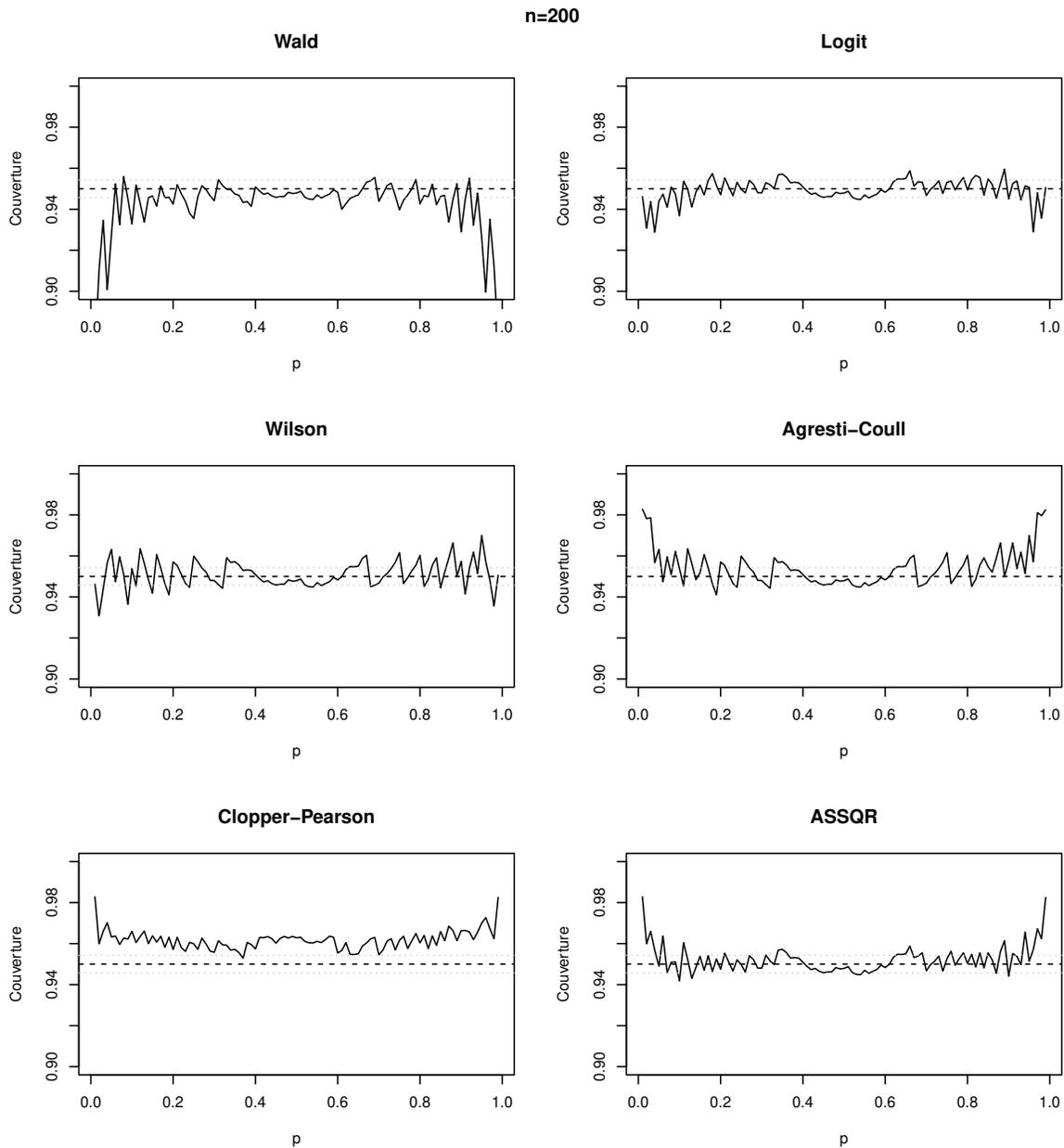


Figure 4.11. Couverture des méthodes traditionnelles dans un contexte i.i.d. avec $n = 200$.

Lorsque p prend des valeurs près de 0 et 1, ASSQR couvre plus souvent le paramètre d'intérêt que les deux autres méthodes en ayant des intervalles de confiance de longueurs comparables. Cette dernière serait la méthode à considérer parmi les méthodes traditionnelles pour $n = 200$ surtout pour des valeurs de p près de 0 et 1. La méthode de Wilson a quant à elle une couverture plus constante autour de la valeur nominale pour ces mêmes valeurs de p .

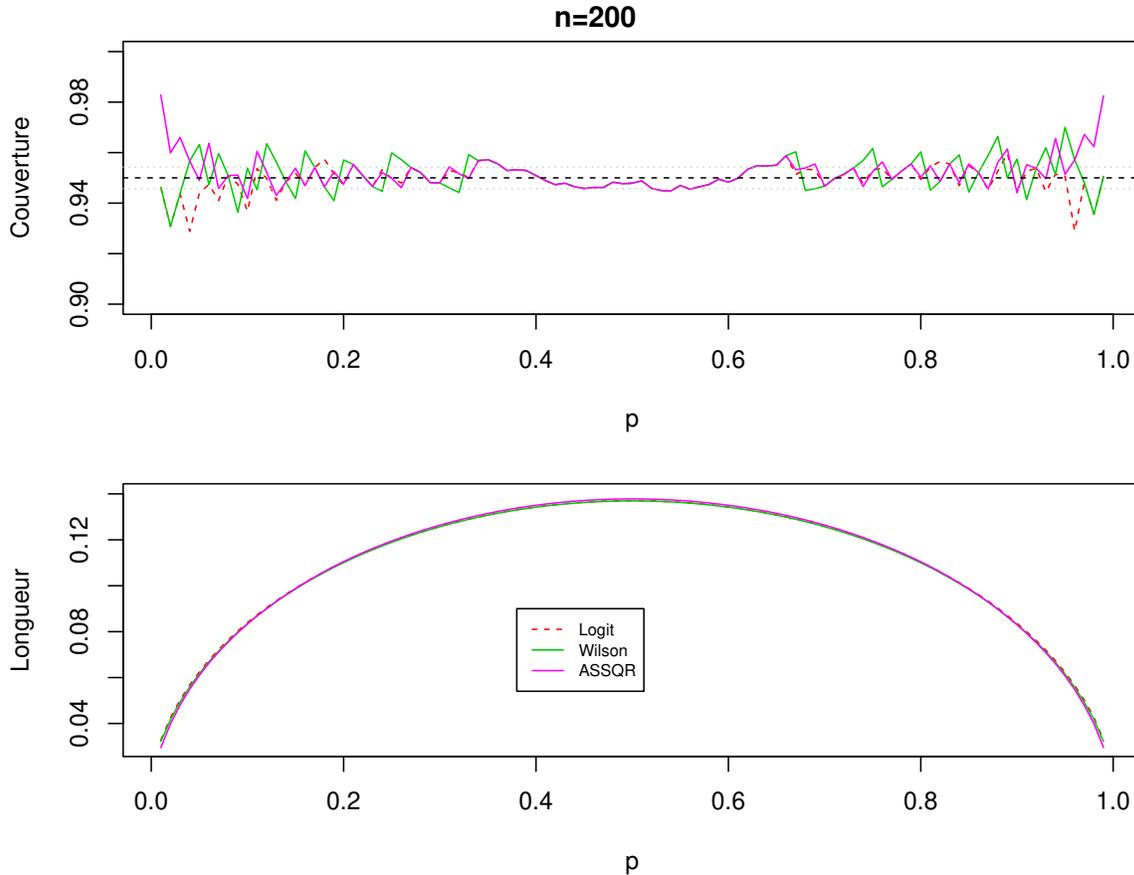


Figure 4.12. Couverture et longueur moyenne des intervalles des méthodes Logit, Wilson et ASSQR dans un contexte i.i.d. avec $n = 200$.

Nous avons conclu plus tôt que la méthode bootstrap percentile est la méthode la plus performante parmi les 4 méthodes bootstrap afin d'estimer un paramètre de proportion autant lorsqu'on fait le bootstrap à la Agresti-Coull que lorsqu'on utilise le bootstrap adaptatif. Comparons donc les résultats de la méthode percentile sous ces deux distributions bootstrap avec les deux méthodes traditionnelles les plus efficaces, soit la méthode de Wilson et la méthode ASSQR.

Pour $n = 40$, en comparant la méthode percentile pour le bootstrap à la Agresti-Coull et le bootstrap adaptatif, on se rend compte qu'autant la méthode percentile sous le bootstrap adaptatif que la méthode percentile sous le bootstrap à la Agresti-Coull donnent des résultats similaires lorsque p prend des valeurs entre 0,2 et 0,8 (voir figure 4.13). Toutefois, la méthode percentile avec le bootstrap à la Agresti-Coull a une courbe de couverture plus près de la valeur nominale qu'en utilisant le bootstrap adaptatif lorsque p prend des valeurs entre 0,08

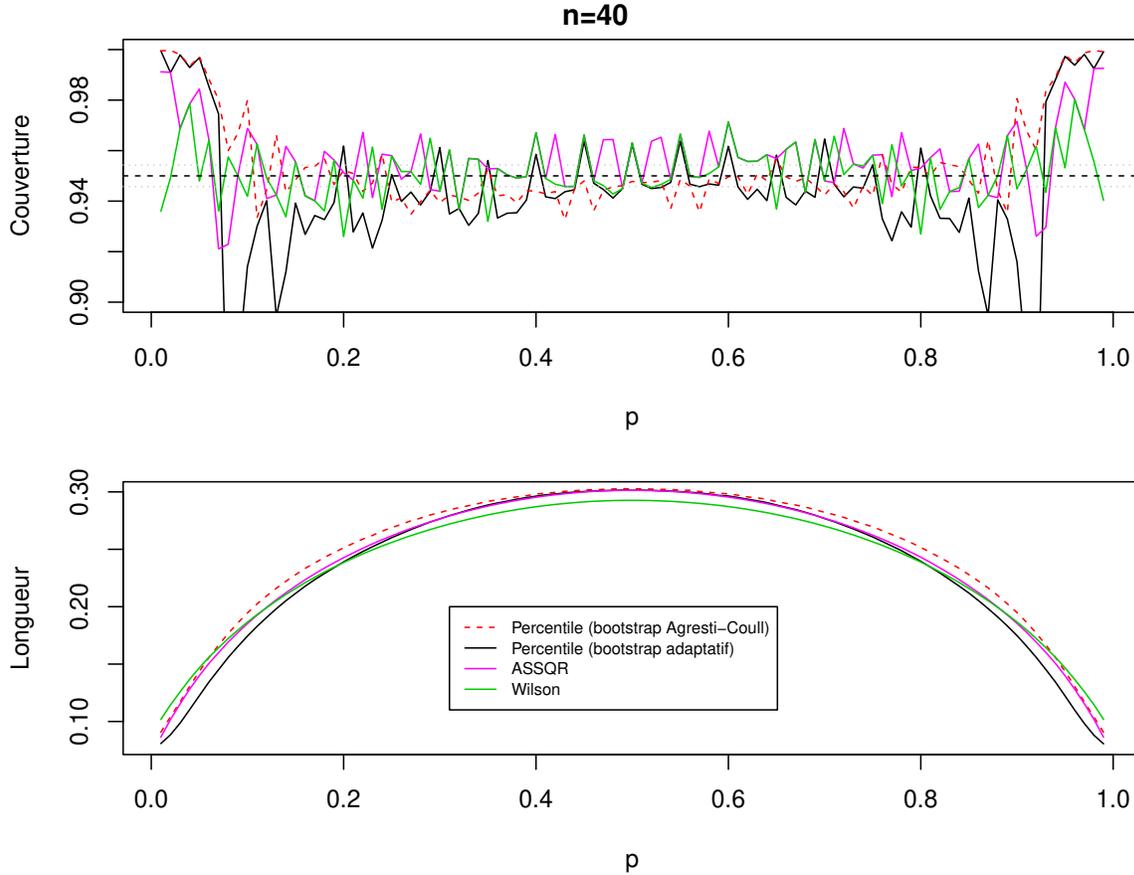


Figure 4.13. Couverture et longueur moyenne des intervalles des méthodes percentile sous le bootstrap à la Agresti-Coull, sous le bootstrap adaptatif, de la méthode ASSQR et de la méthode de Wilson avec $n = 40$.

et 0,2 ainsi qu'entre 0,8 et 0,92 pour finalement voir les deux méthodes avoir une couverture près du 100% d'efficacité pour p près de 0 ou près de 1. La méthode de Wilson est toujours très constante autour de la valeur nominale. La méthode ASSQR démontre également beaucoup de constance autour de la valeur nominale. Notons à nouveau que la couverture minimale de la méthode ASSQR (92,1%) est légèrement inférieure à la couverture minimale de la méthode de Wilson (92,6%).

Lorsqu'on prête attention à la longueur moyenne de ces intervalles de confiance, on peut remarquer que lorsque p prend des valeurs près de 0 et 1, Wilson est la méthode qui donne les intervalles de confiance les plus longs et la méthode percentile avec le bootstrap adaptatif donne les intervalles les plus courts. Comme nous l'avons vu plus haut, la méthode percentile avec l'utilisation du bootstrap adaptatif a beaucoup de difficultés à bien s'adapter pour

des proportions entre 0,08 et 0,2 puis entre 0,8 et 0,92. Un fait notable est que la méthode percentile avec l'utilisation du bootstrap à la Agresti-Coull donne des intervalles de confiance en moyenne plus courts que la méthode de Wilson lorsque p est près de 0 ou de 1, mais qu'en plus, elle couvre plus souvent le paramètre d'intérêt. Toutefois, la méthode ASSQR a plutôt un comportement similaire à la méthode percentile à la Agresti-Coull lorsque p prend de petites ou grandes valeurs. Donc si la méthode percentile à la Agresti-Coull performe mieux que la méthode de Wilson dans les extrémités, elle compétitionne bien avec la méthode ASSQR alors que les deux ont des comportements similaires, c'est-à-dire des couvertures qui montent près de 100% avec des longueurs d'intervalles comparables. Nous pouvons donc affirmer que la méthode percentile avec l'utilisation du bootstrap à la Agresti-Coull fait mieux que la méthode de Wilson dans les extrémités et peut facilement compétitionner avec la méthode ASSQR lorsque l'échantillon initial est de taille 40.

Pour $n = 200$, la méthode de Wilson, la méthode ASSQR et la méthode percentile avec l'utilisation du bootstrap à la Agresti-Coull performent bien de manière générale (voir figure 4.14). En effet, lorsque p est entre 0,1 et 0,9, les trois couvertures sont très près de la valeur nominale de 95% tout en ayant des intervalles de longueurs très comparables. On peut même mentionner que dans les extrémités, la méthode percentile avec l'utilisation du bootstrap à la Agresti-Coull et la méthode ASSQR sont préférables à la méthode de Wilson puisqu'ils ont une couverture supérieure à cette dernière tout en ayant des intervalles de confiance de même longueur que celle-ci. La méthode percentile avec l'utilisation du bootstrap adaptatif a une couverture trop souvent supérieure à la valeur nominale. On peut aussi voir le même phénomène qu'avec $n = 40$ concernant l'utilisation du bootstrap à la Agresti-Coull et le bootstrap adaptatif lorsque p prend des valeurs un peu plus grandes que 0 et un peu plus petites que 1. En effet, dans ces cas précis, la méthode percentile avec l'utilisation du bootstrap adaptatif a une couverture minimale de 91,27% tandis que la méthode percentile avec l'utilisation du bootstrap à la Agresti-Coull offre une couverture minimale de 94,06%.

En regardant la longueur moyenne des intervalles de confiance de ces 4 méthodes, on peut remarquer que les intervalles sont de longueurs très comparables. De manière générale pour $n = 200$, si la méthode de Wilson et la méthode percentile offrent des couvertures similaires lorsque p est entre 0,1 et 0,9, il semble que dans les extrémités, la méthode percentile avec

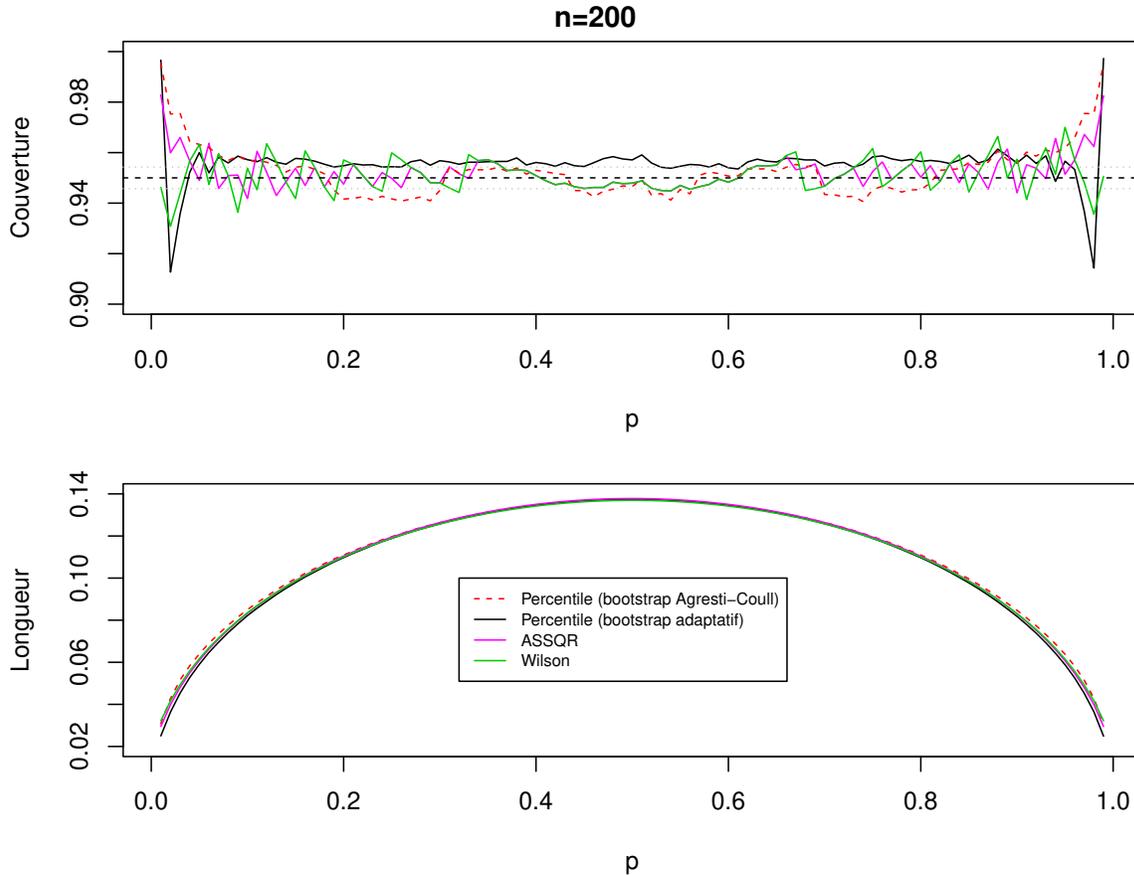


Figure 4.14. Couverture et longueur moyenne des intervalles des méthodes percentile sous le bootstrap à la Agresti-Coull, sous le bootstrap adaptatif, de la méthode ASSQR et la méthode de Wilson avec $n = 200$.

l'utilisation du bootstrap à la Agresti-Coull et la méthode ASSQR sont préférables aux deux autres à cause du fait qu'ils couvrent beaucoup plus en ayant des intervalles de confiance de mêmes longueurs.

4.3. Résultats sous un contexte d'échantillonnage

Poursuivons l'analyse des résultats de la simulation en considérant le contexte d'échantillonnage, c'est-à-dire où un échantillon de taille n est tiré à l'aide d'un plan aléatoire simple sans remise à partir de la population. Deux méthodes ont été utilisées pour effectuer le bootstrap soit la méthode par pseudo-population de [Booth *et al.*, 1994] et la méthode des poids de [Rao *et al.*, 1992]. Concentrons-nous d'abord sur la méthode par pseudo-population.

4.3.1. Méthode par pseudo-population

Débutons l'analyse des résultats avec ceux obtenus en utilisant le bootstrap à la Agresti-Coull.

À la figure 4.15, nous avons la couverture des 4 méthodes bootstrap lorsqu'un échantillon de taille 40 est tiré à l'aide du plan de sondage à partir de la population. On peut remarquer une courbe noire et une autre courbe pointillée rouge qui représentent respectivement la couverture des méthodes lorsque la fraction de sondage est de 0,05 et de 0,20.

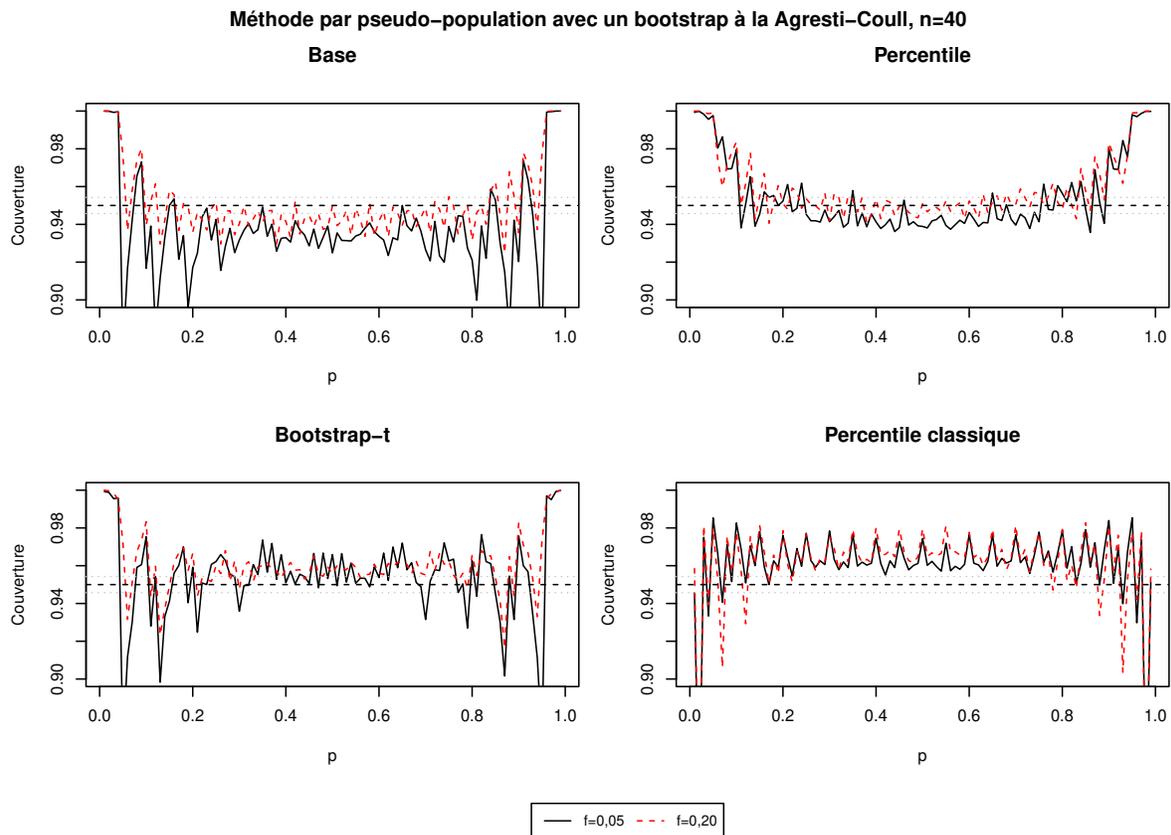


Figure 4.15. Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 40$.

Pour une fraction de sondage de 5% (lignes noires sur les graphiques de la figure 4.15), nous pouvons voir que la méthode percentile offre la meilleure couverture générale en assurant

une couverture d'au moins 93,57% contrairement à 87,29% pour la méthode bootstrap de base, à 86,54% pour la méthode bootstrap- t et à 82,06% pour la méthode percentile classique.

De plus les méthodes bootstrap de base et bootstrap- t démontrent toutes les deux beaucoup de difficultés à bien couvrir le paramètre d'intérêt lorsque p prend des valeurs inférieures à 0,2 et supérieures à 0,8 d'où leurs couvertures minimales assez petites comparativement à la méthode percentile, malgré que ces deux méthodes atteignent jusqu'à 100% de couverture dans les extrémités. De son côté, la méthode percentile classique a une couverture supérieure à la valeur nominale lorsque p prend des valeurs entre 0,1 et 0,9, puis elle redescend dans les extrémités avant de remonter près du 95% de couverture. Toutefois, sa couverture est très instable dans ces circonstances. En effet, sa couverture oscille beaucoup en étant très inconstante en plus d'avoir des intervalles plus longs que la méthode percentile (figure 4.16). La méthode percentile couvre plus dans les extrémités, mais en ayant des intervalles plus courts que ceux de la méthode percentile classique. Puis comme les méthodes bootstrap de base et bootstrap- t ont des couvertures qui oscillent beaucoup dans les extrémités, elles ne sont pas préférables à la méthode percentile, malgré leurs intervalles de confiance plus courts que cette dernière.

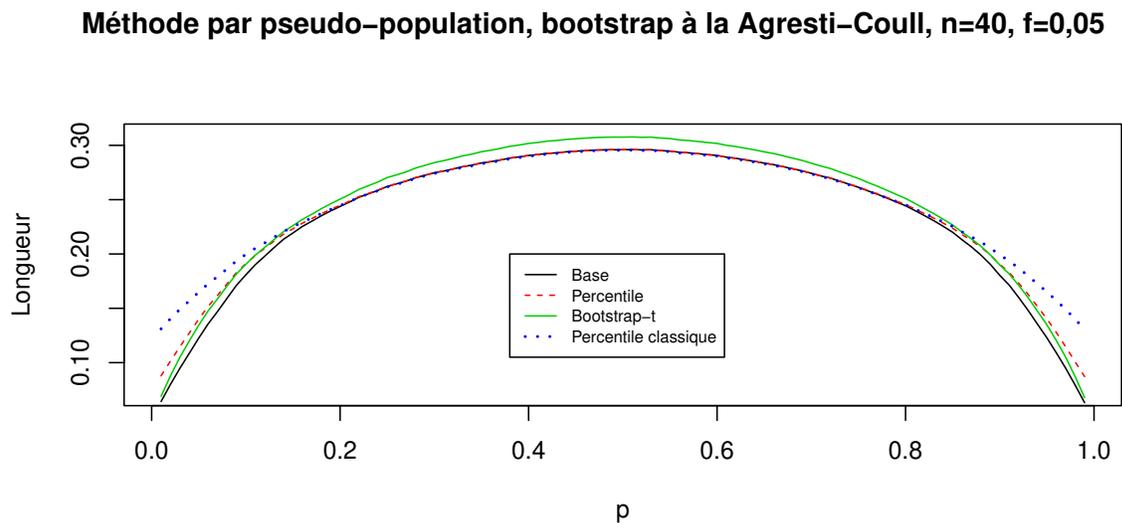


Figure 4.16. Longueur moyenne des intervalles bootstrap pour la méthode par pseudo-population à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

Lorsqu'on s'intéresse maintenant aux couvertures obtenues avec l'utilisation d'une fraction de sondage de 20% (ligne pointillée rouge), on peut voir un meilleur comportement pour les méthodes bootstrap de base et bootstrap- t pour des valeurs de p inférieures à 0,2 et supérieures à 0,8. En effet, leurs couvertures sont beaucoup plus près de la valeur nominale de 95% en étant plus stables autour de cette valeur. La couverture minimale de la méthode bootstrap de base est de 92,47% puis celle de la méthode bootstrap- t est désormais de 91,63%. Malgré cette amélioration notable comparativement aux résultats de ces méthodes pour une fraction de sondage de 5%, la méthode percentile semble encore une fois être la meilleure parce qu'elle assure une couverture d'au moins 94,04%. De plus cette dernière a maintenant des intervalles de confiance de longueurs très comparables à la méthode bootstrap- t et légèrement supérieures à la méthode bootstrap de base dans les extrémités (figure 4.17).

Méthode par pseudo-population, bootstrap à la Agresti-Coull, $n=40$, $f=0,20$

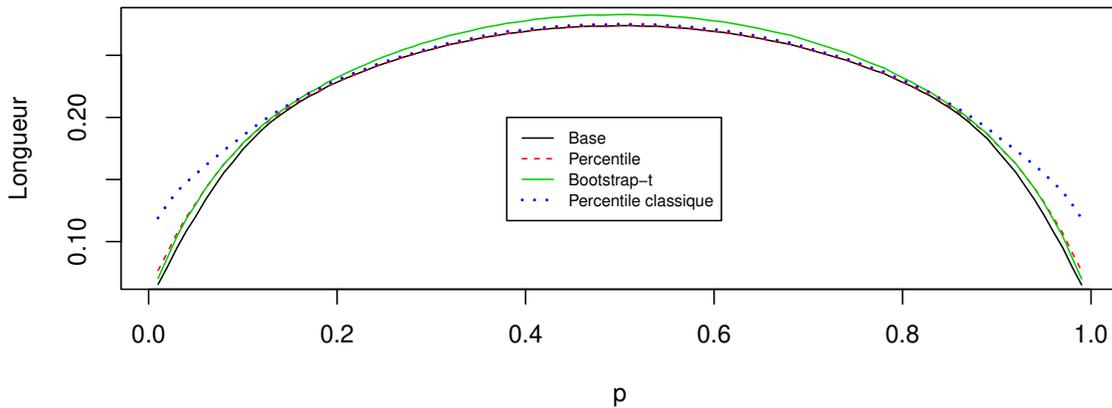


Figure 4.17. Longueur moyenne des intervalles bootstrap pour la méthode par pseudo-population à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.

La méthode percentile classique démontre encore une fois beaucoup de difficultés dans ce contexte en ayant un comportement comparable à celui observé lorsque la fraction de sondage est de 5%. En effet, pour des valeurs centrales de p , sa couverture est toujours supérieure à la valeur nominale puis dans les extrémités elle oscille beaucoup (couverture minimale de 82,10%). De plus, ses intervalles de confiance sont beaucoup plus longs que

ceux de la méthode percentile dans ces circonstances.

Lorsqu'on considère maintenant un échantillon de 200 observations, on peut voir que les couvertures sont beaucoup plus stables autour de la valeur nominale en oscillant beaucoup moins (figure 4.18). En considérant une fraction de sondage de 5% (ligne noire sur les graphiques de la figure 4.18), on peut voir que même si la couverture de la méthode percentile classique est toujours supérieure à la valeur de 95% (couverture minimale de 95,55%), la méthode bootstrap de base semble faire mieux en atteignant la valeur nominale à quelques reprises surtout pour les valeurs centrales de p , mais ne fait pas très bien dans les extrémités en ayant une couverture bien en deçà du 95% (couverture minimale de 91,02%). En revanche, les deux meilleures méthodes pour estimer p semblent être les méthodes percentile et bootstrap- t . En effet, ces deux dernières ont une couverture très près de la valeur nominale sans descendre trop bas en plus d'atteindre jusqu'à près 100% dans les extrémités. La couverture minimale de la méthode percentile est de 93,92% puis de 93,25% pour la méthode bootstrap- t .

C'est davantage à l'aide de la figure 4.19 que nous pouvons voir le comportement assez similaire des méthodes percentile et bootstrap- t . Notons que les intervalles de confiance de ces méthodes sont de longueurs plutôt semblables. On peut voir que pour des valeurs de p près de 0,08 et 0,92, la méthode bootstrap- t est beaucoup plus instable que la méthode percentile puisque sa couverture oscille beaucoup. Puis, en se rapprochant des extrémités, la méthode percentile semble avoir une couverture bien supérieure, donc être préférable vu la similarité de leurs longueurs.

Si on s'intéresse maintenant aux résultats des simulations avec une taille échantillonnale de 200 mais une fraction de sondage qui est plutôt de 20%, on peut voir sur la figure 4.18 que les couvertures des méthodes bootstrap de base, percentile et bootstrap- t offrent des résultats assez satisfaisants (lignes pointillées rouges). Seule la méthode percentile classique offre une couverture qui est encore et toujours supérieure à la valeur nominale en ayant une couverture minimale de 95,54%. Les méthodes bootstrap- t ainsi que percentile démontrent toujours beaucoup d'efficacité comme nous l'avons vu avec une fraction de sondage de 0,05. En effet, elles assurent respectivement des couvertures d'au moins 93,79% et 94,37% pour

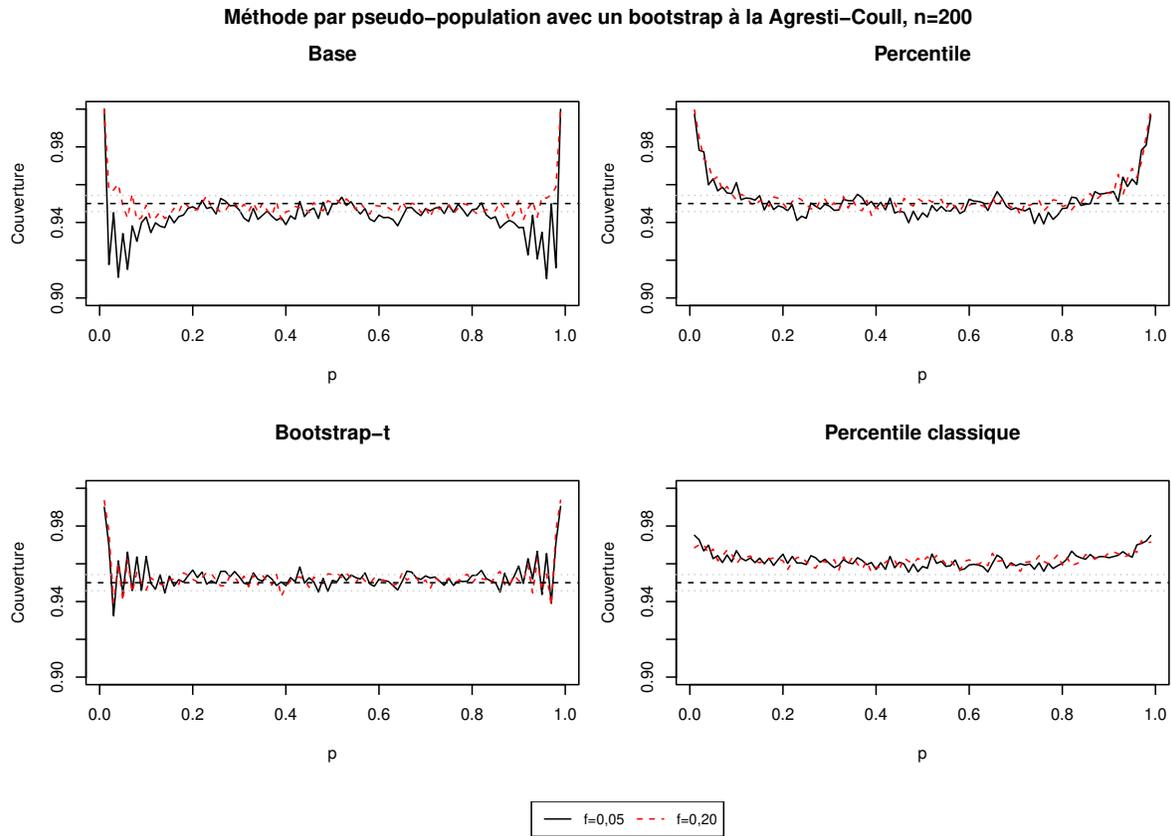


Figure 4.18. Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$.

toutes les valeurs de p . Si la méthode bootstrap de base démontrait beaucoup d'inconstance surtout dans les extrémités lorsque la fraction de sondage était de 5%, il semble que sa couverture soit beaucoup plus près de la valeur nominale en descendant beaucoup moins bas et assurant une couverture minimale de 94,10%.

À l'aide de la figure 4.20, on peut encore constater que les intervalles sont de longueurs très comparables entre elles. Par contre, nous avons aussi que la méthode percentile semble être la méthode la plus efficace des méthodes bootstrap afin d'estimer p . Si les méthodes percentile, bootstrap- t et bootstrap de base ont tous des couvertures très près du 95% pour des valeurs centrales de p , la méthode percentile offre une couverture supérieure aux deux autres méthodes lorsque p prend des valeurs près de 0,05 et 0,95 en ayant des intervalles de confiance de longueurs comparables.

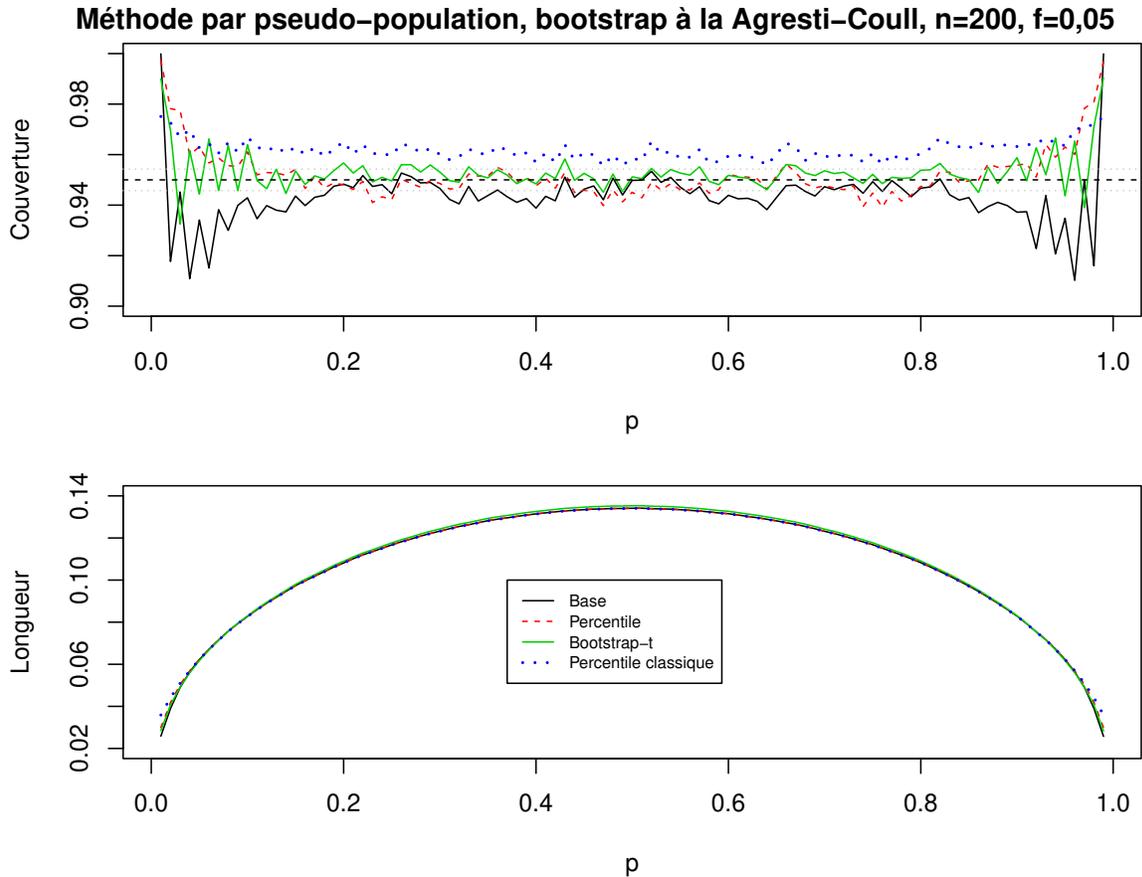


Figure 4.19. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

C'est donc dire que peu importe le taille échantillonnale et la fraction de sondage, la méthode percentile semble être la méthode bootstrap la plus efficace pour estimer par intervalle de confiance un paramètre de proportion lorsqu'on utilise un bootstrap à la Agresti-Coull avec la méthode par pseudo-population.

Penchons-nous maintenant sur les résultats des simulations effectuées avec la méthode par pseudo-population mais en utilisant un bootstrap adaptatif.

Sur la figure 4.21 on peut apercevoir les couvertures des 4 méthodes bootstrap lorsque la taille échantillonnale est de 40 observations. Concentrons-nous pour débuter sur les couvertures avec l'utilisation d'une fraction de sondage égale à 5% (ligne noire sur les graphiques de la figure 4.21).

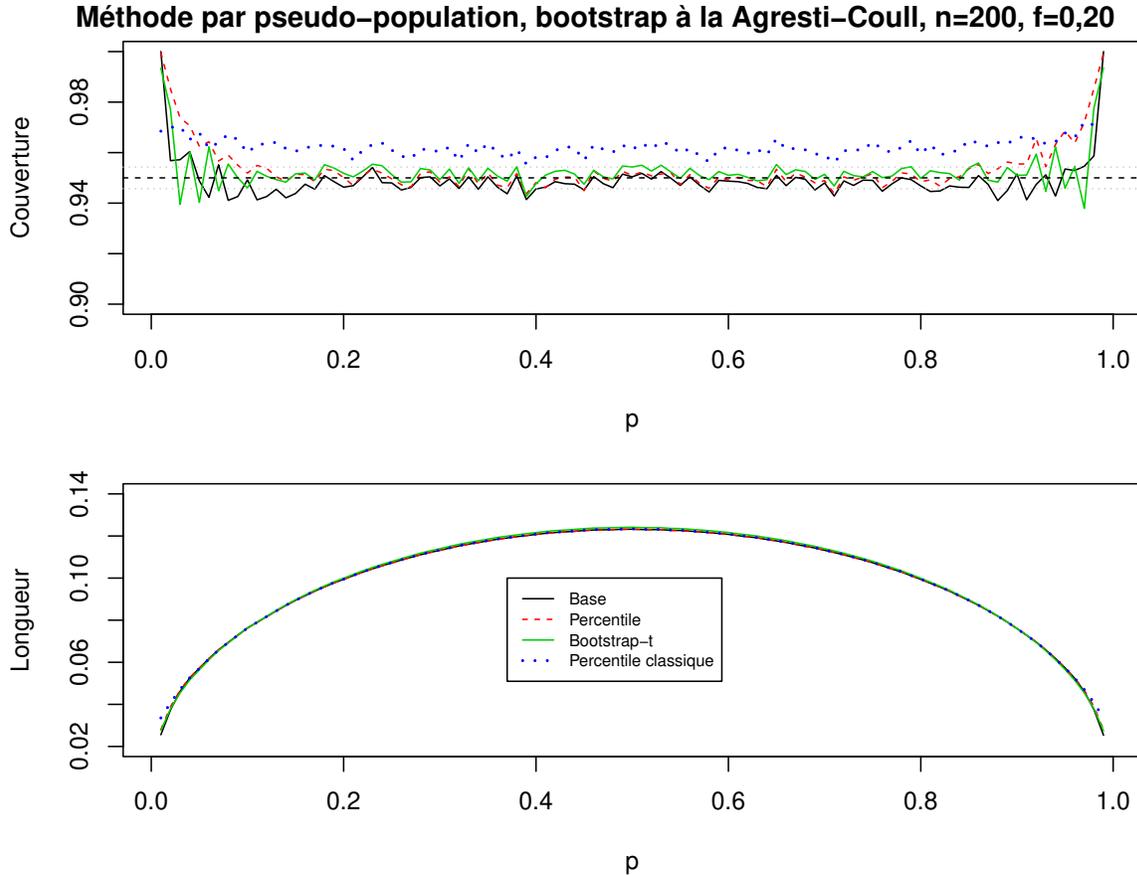


Figure 4.20. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.

Tout d'abord, si quelques méthodes se débrouillent bien pour des valeurs centrales de p , on peut remarquer que, de manière générale, les méthodes ne performent pas très bien avec l'utilisation du bootstrap adaptatif, surtout dans les extrémités. Il est à noter que pour des valeurs de p près de 0,1 et 0,9, les méthodes bootstrap sous le bootstrap adaptatif ont beaucoup de difficultés à estimer p . C'est entre autres, un comportement qui a déjà été aperçu dans un contexte i.i.d. lorsque vient le temps d'utiliser le bootstrap adaptatif. Rappelons que le bootstrap adaptatif consiste à considérer l'échantillon s^+ ou s selon la valeur de \hat{p} . Comme pour de petites et grandes valeurs de p , l'estimateur alterne souvent en étant nul ou non nul, l'échantillon considéré change également de sorte que les méthodes semblent avoir de la difficulté à s'adapter afin d'estimer p .

Malgré le fait que pour chacune de ces méthodes, elles atteignent jusqu'à 100% de couverture dans les extrémités, elles n'assurent pas une couverture minimale très près de la

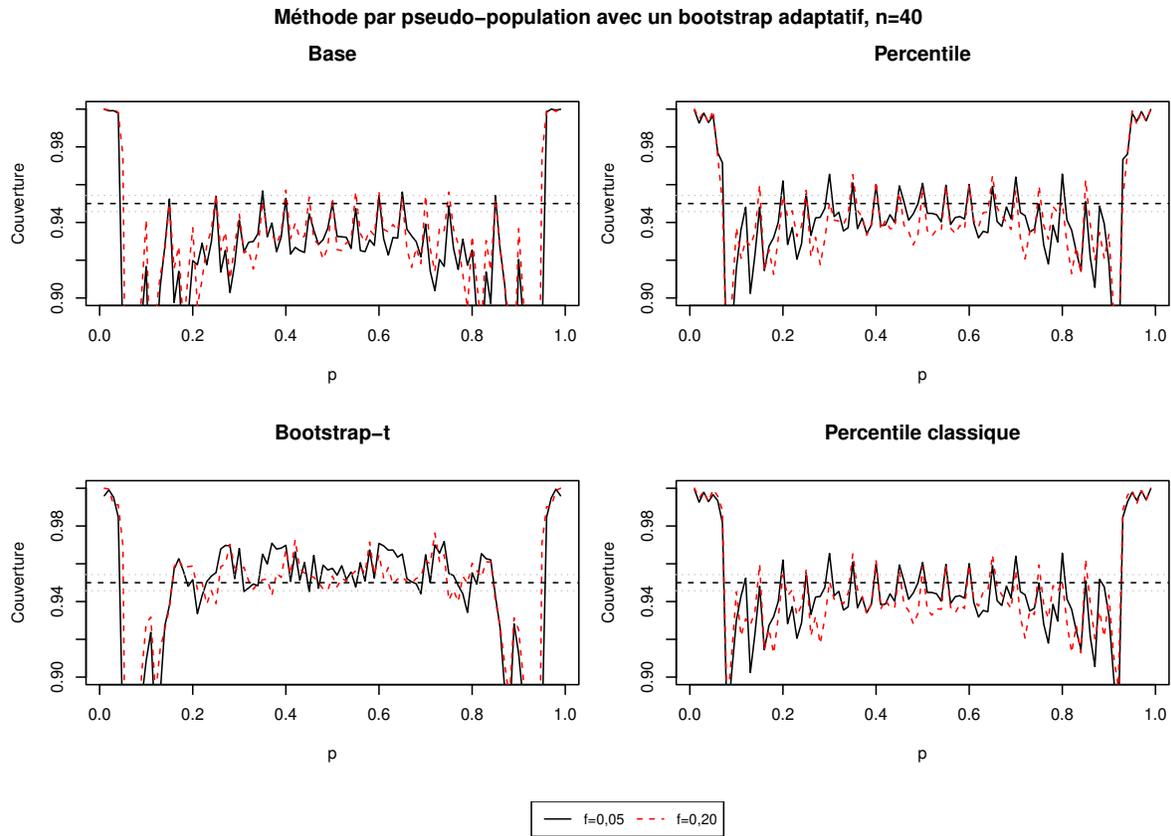


Figure 4.21. Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$.

valeur nominale, surtout pour les méthodes bootstrap de base et bootstrap- t . En effet, pour ces deux méthodes, elles ont une couverture minimale de 70,09% et de 68,72% respectivement. La méthode percentile a une couverture minimale de 84,48% tandis que la méthode percentile classique a une couverture minimale de 87,11%.

Les résultats de la simulation nous ont montré que les couvertures des méthodes percentile et percentile classique sont les mêmes pour toutes les valeurs de p qui ne sont pas entre 0,06 et 0,12 et entre 0,88 et 0,94. Il est possible d'observer ce fait à l'aide de la figure 4.22. On peut dire que les méthodes percentile et percentile classique offrent autant d'efficacité l'une que l'autre. La méthode percentile classique a une couverture minimale plus élevée (d'à peine 3%), mais la méthode percentile offre les intervalles les plus courts dans les extrémités. C'est pourquoi on conclut que les deux méthodes sont très similaires et performant aussi bien dans ce contexte pour estimer par intervalle de confiance le paramètre de proportion.

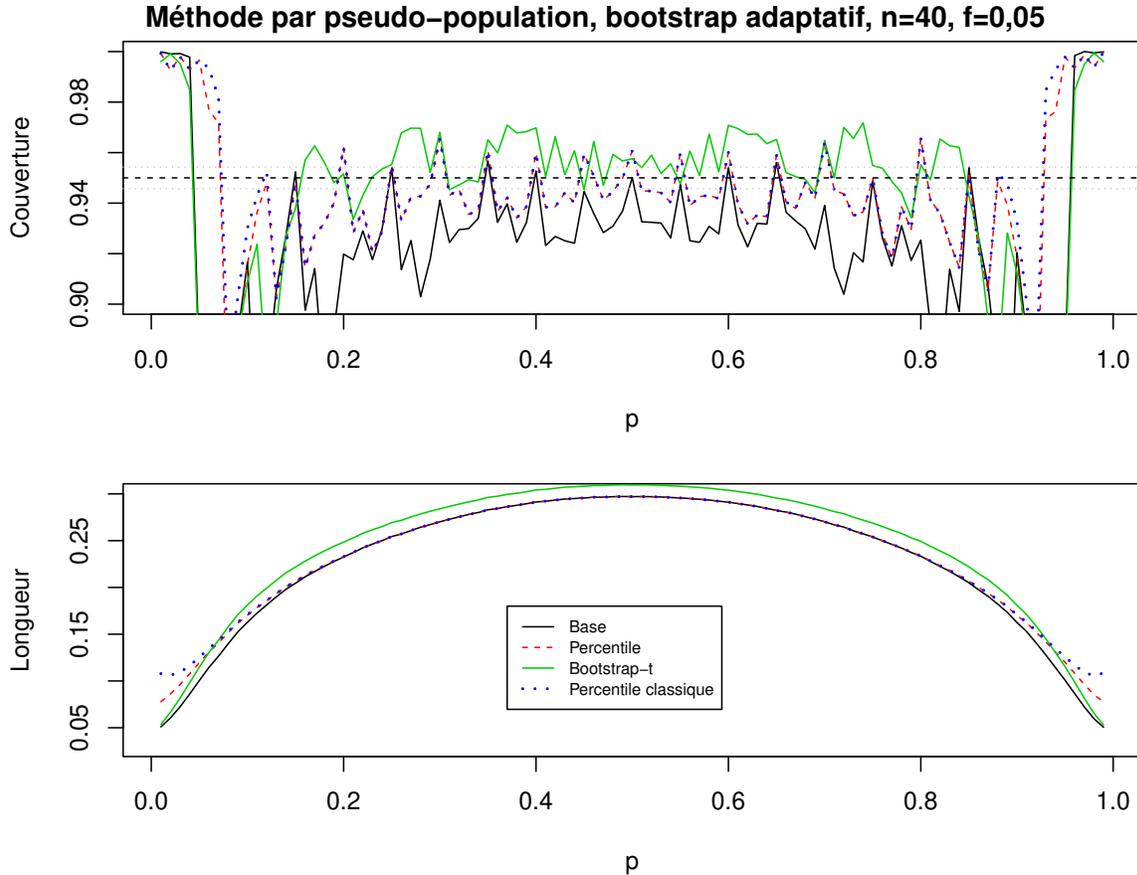


Figure 4.22. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

Lorsque la fraction de sondage est égale à 20%, on observe peu de différences dans le sens où les couvertures sont assez similaires à ce que l'on a observé pour une fraction de sondage de 0,05 puis les méthodes démontrent encore beaucoup de difficultés à estimer p pour des valeurs de p près de 0,1 et 0,9 (lignes pointillées rouges sur la figure 4.21).

Les couvertures minimales qu'assurent les différentes méthodes sont de 72,74% pour la méthode bootstrap de base, 71,44% pour la méthode bootstrap- t puis 86,07% et 88,37% pour les méthodes percentile et percentile classique respectivement. Ces deux dernières sont encore les deux meilleures méthodes afin d'estimer un paramètre de proportion puisqu'elles garantissent une couverture minimale supérieure aux deux autres méthodes bootstrap.

La couverture des méthodes percentile et percentile classique sont les mêmes pour des valeurs de p qui ne sont pas comprises entre 0,06 et 0,12 puis entre 0,88 et 0,94 (figure 4.23). On observe donc le même comportement que nous avons observé pour ces deux méthodes

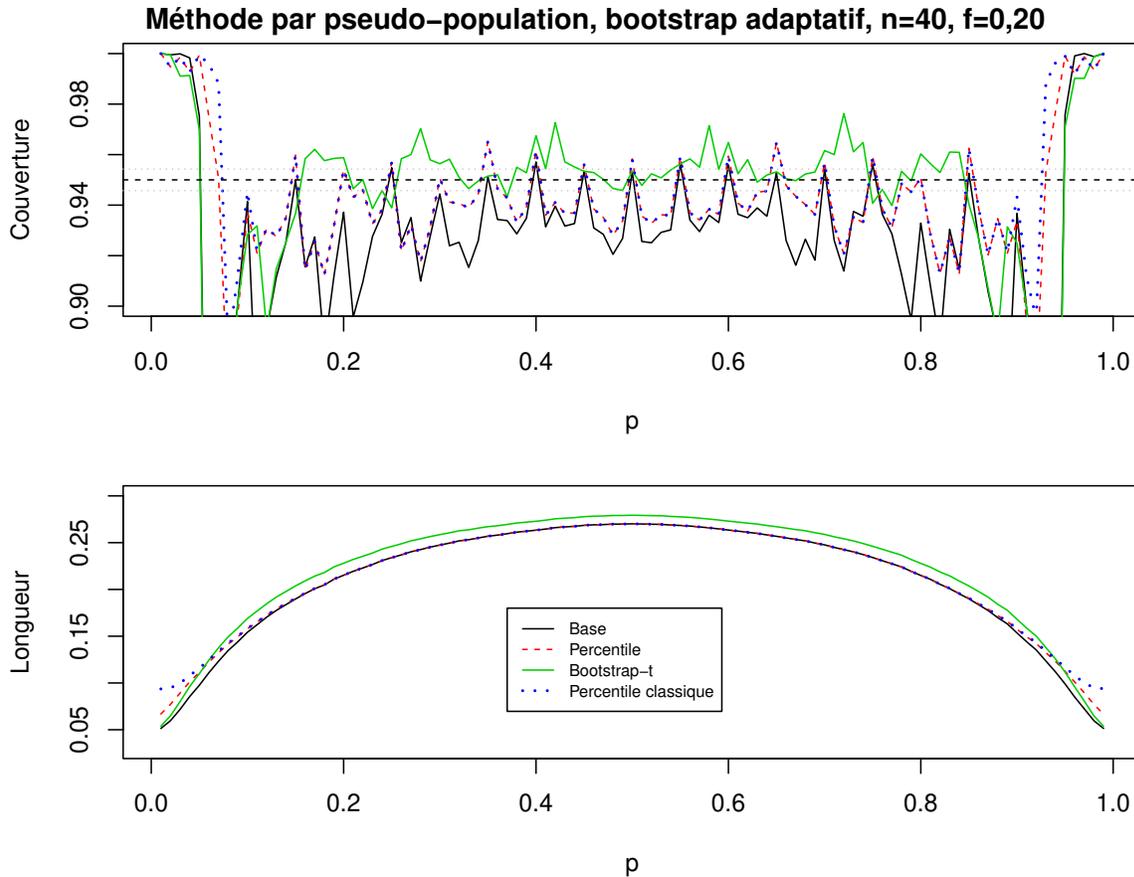


Figure 4.23. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.

pour une fraction de sondage de 5%. La performance des méthodes percentile et percentile classique est plutôt similaire encore une fois puisque même si la méthode percentile classique a une couverture minimale supérieure à celle de la méthode percentile, la différence entre ces deux couvertures n'est que d'un peu plus de 2%. Puis dans les extrémités, les deux méthodes vont jusqu'à atteindre 100% d'efficacité mais la méthode percentile est préférable puisqu'elle a les intervalles les plus courts.

Lorsqu'on considère un échantillon de taille 200 pour l'utilisation du bootstrap adaptatif, on peut voir que les couvertures sont bien meilleures que celles ayant été observées lorsqu'un échantillon de taille 40 était considéré (figure 4.24). Considérons pour débuter les résultats avec l'utilisation d'une fraction de sondage de 5%, soit les courbes noires sur les graphiques de la figure 4.24.

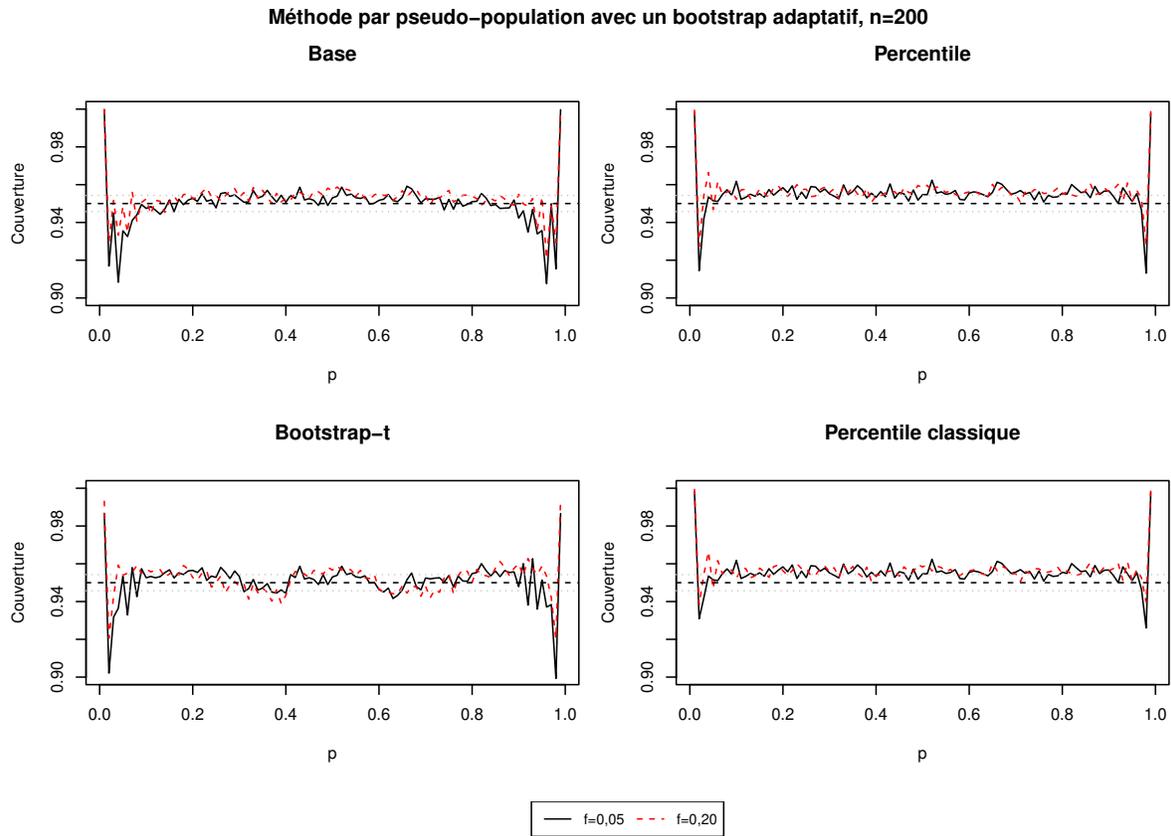


Figure 4.24. Couverture des méthodes bootstrap sous la méthode par pseudo-population avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$.

Si pour des valeurs centrales de p , les méthodes percentile et percentile classique ont des couvertures qui semblent légèrement supérieures à 95%, c'est la méthode bootstrap de base qui semble performer le mieux puisqu'elle a toujours une couverture très près de la valeur nominale dans ces circonstances. De son côté, pour ces mêmes valeurs de p , la méthode bootstrap- t ne performe pas très bien en oscillant beaucoup autour de la valeur de 95%. Par contre, dans les extrémités, la méthode bootstrap de base n'a pas la même efficacité. En effet, comme la méthode bootstrap- t , elle semble avoir beaucoup de mal à estimer p puisque sa couverture est bien en deçà de la valeur nominale comparativement aux méthodes percentile et percentile classique. Par ailleurs, la méthode bootstrap de base a une couverture minimale de 90,76%, la méthode bootstrap- t a une couverture minimale de 89,93% tandis que la méthode percentile et la méthode percentile classique ont des couvertures minimales respectives de 91,31% et 92,60%.

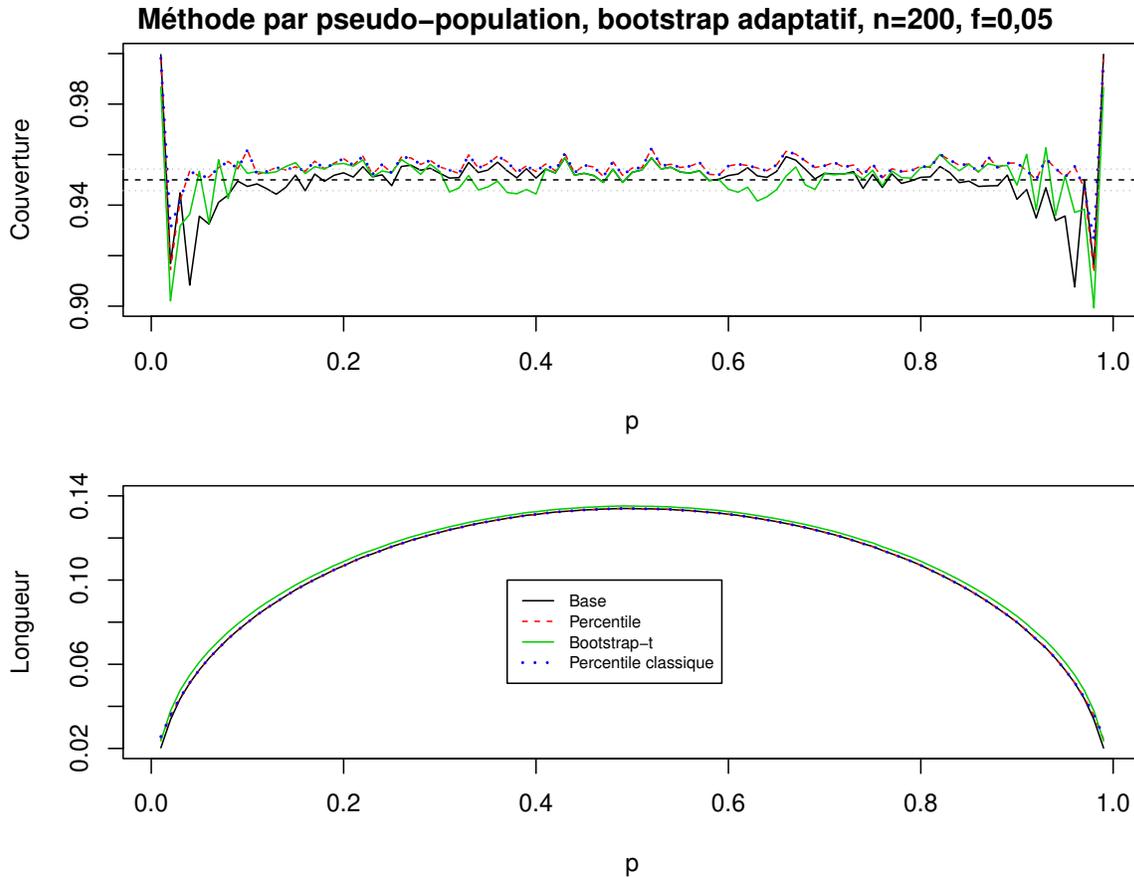


Figure 4.25. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

Ces deux dernières offrent des couvertures identiques pour des valeurs de p qui ne sont pas égales à 0,02 ou 0,98 comme on peut le voir à la figure 4.25. On peut donc dire que les deux méthodes sont très similaires en terme d'efficacité. La seule différence de couverture entre les deux méthodes, pour p égal à 0,02 et 0,98, est d'à peine 1%.

Pour une taille échantillonnale de 200 observations ainsi qu'une fraction de sondage de 20%, on peut voir que toutes les méthodes bootstrap font plutôt bien en étant toutes très près de la valeur nominale (lignes pointillées rouges sur la figure 4.24). En effet, chacune des 4 méthodes assure une couverture minimale pas trop basse en n'oscillant pas beaucoup pour des valeurs centrales de p , en étant très près de 95%, et en atteignant jusqu'à 100% dans les extrémités. La couverture minimale de la méthode bootstrap de base est de 92,07%, la couverture minimale de la méthode bootstrap- t est de 92,03% puis pour les méthodes

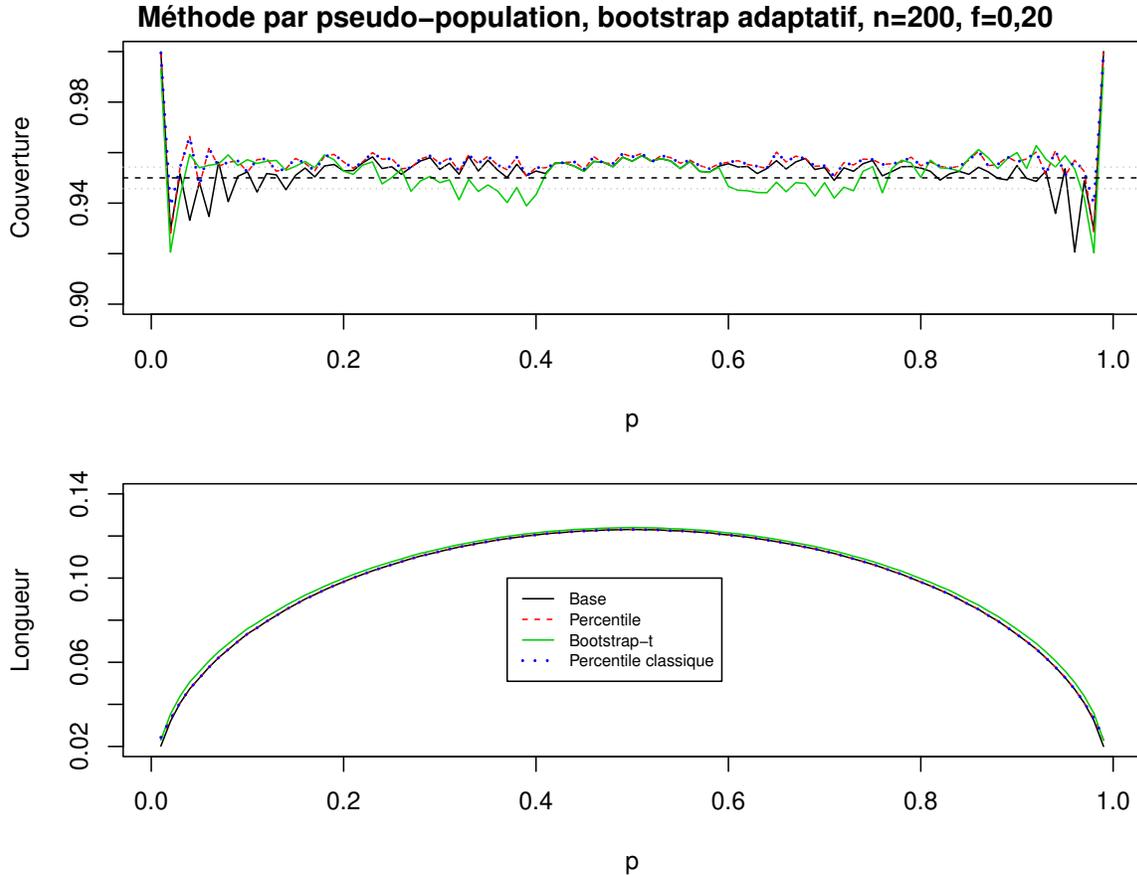


Figure 4.26. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.

percentile et percentile classique les couvertures minimales sont respectivement de 92,73% et de 93,85%.

Encore ici, comme nous l'avons vu avec l'utilisation d'une fraction de sondage de 20%, les couvertures des méthodes percentile et percentile classique sont les mêmes pour tous les p sauf pour 0,02 et 0,98 où la couverture de la méthode percentile classique est légèrement supérieure. C'est ce que la figure 4.26 nous aide à voir en plus de nous montrer que les intervalles des différentes méthodes sont de longueurs très comparables entre elles. Les méthodes percentile et percentile classique offrent une efficacité comparable dans ce contexte.

Pour résumer, dans le cas du bootstrap à la Agresti-Coull, on a conclu que la meilleure méthode était la méthode percentile. Pour la méthode bootstrap adaptatif, les méthodes percentile et percentile classique ont des comportements identiques sauf pour deux valeurs de p lorsque $n = 200$ et quelques valeurs de p lorsque $n = 40$. Dans ce dernier cas, la

longueur dans les extrémités avantagait la méthode percentile. Dans le cas de $n = 200$, pour les deux seules valeurs de p où les méthodes diffèrent, la méthode percentile classique faisait mieux. En conséquence, pour comparer les meilleures méthodes entre elles, nous avons décidé de comparer aux autres la méthode percentile pour le bootstrap adaptatif par pseudo-population.

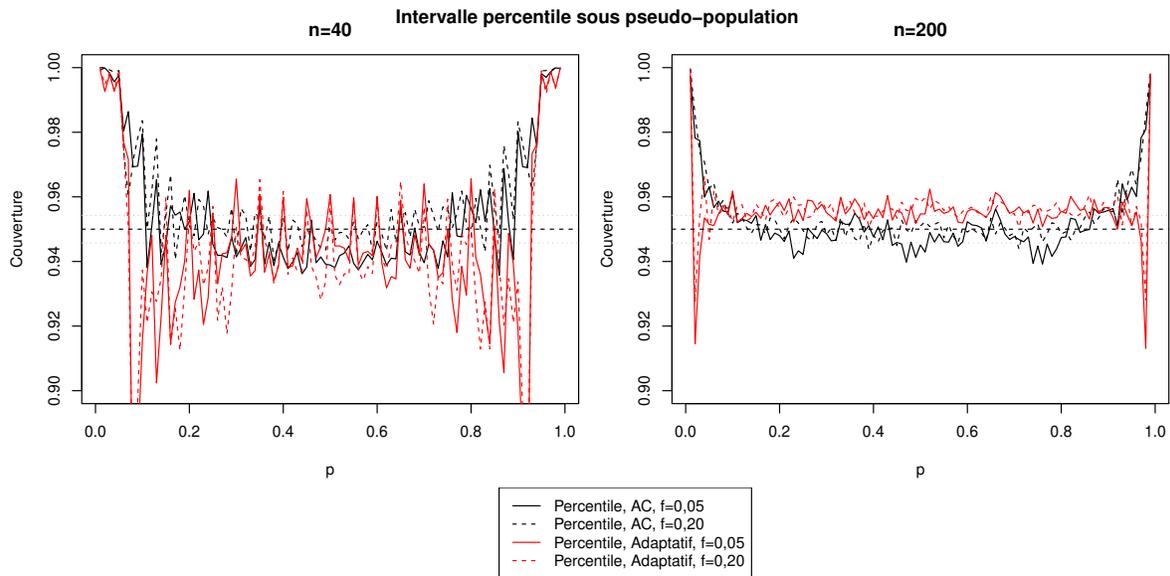


Figure 4.27. Couvertures de l'intervalle percentile sous la méthode par pseudo-population

Sur la figure 4.27, on peut voir les différentes couvertures obtenues de l'intervalle percentile lorsqu'on utilise la méthode par pseudo-population. Le graphique à gauche contient les couvertures de l'intervalle percentile lorsque la taille échantillonnale est de 40 observations et que le bootstrap à la Agresti-Coull est utilisé (lignes complètes et pointillées noires représentant respectivement une fraction de sondage de 5 et 20%) ainsi que lorsque le bootstrap adaptatif est considéré (lignes complètes et pointillées rouges représentant respectivement une fraction de sondage de 5 et de 20%). Sur le graphique à droite, nous y retrouvons exactement la même chose, avec une taille échantillonnale de 200 observations.

En regardant cette figure, il est clair que peu importe la taille échantillonnale et la fraction de sondage, le bootstrap à la Agresti-Coull est préférable au bootstrap adaptatif lorsqu'on considère l'intervalle percentile. C'est donc pourquoi nous ne considérerons que la méthode percentile sous un bootstrap à la Agresti-Coull lorsque viendra le temps de comparer

l'efficacité des méthodes bootstrap par pseudo-population aux méthodes bootstrap basées sur les poids et aux méthodes traditionnelles.

4.3.2. Méthode des poids

Poursuivons l'analyse des résultats lorsqu'on considère la méthode des poids avec un bootstrap à la Agresti-Coull.

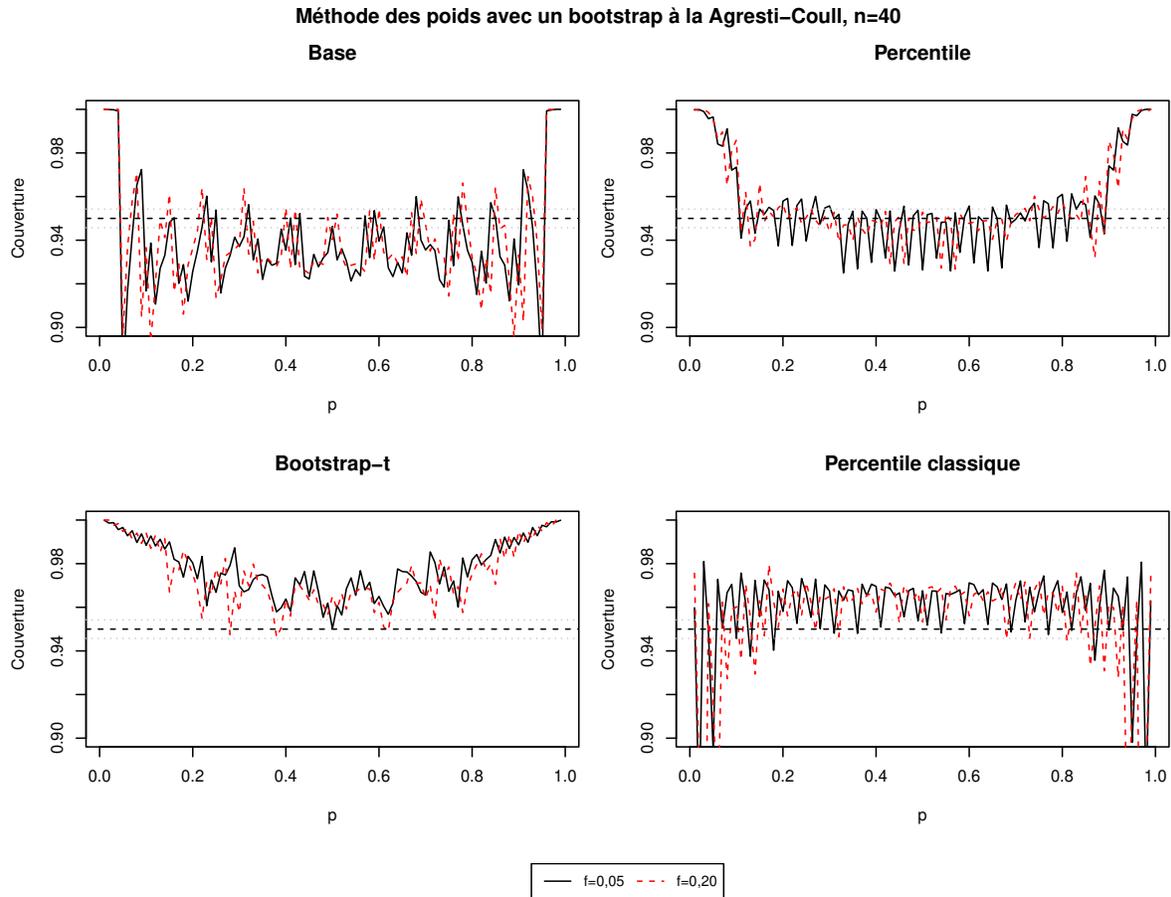


Figure 4.28. Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 40$.

À l'aide de la figure 4.28, on peut voir que lorsque la fraction de sondage est de 0,05 (ligne noire), la méthode qui semble être la plus efficace est la méthode percentile puisqu'elle a une couverture qui est très souvent autour de la valeur nominale pour des valeurs de p entre 0,1 et 0,9 contrairement aux méthodes bootstrap de base et percentile classique qui, de leur côté, n'atteignent que rarement la valeur de 95%. La couverture minimale de la méthode

percentile est de 92,5% tandis que celle de la méthode bootstrap de base est de 87,53% et la couverture minimale de la méthode percentile classique est de 86,33%. En effet, la couverture de la méthode bootstrap de base est plus souvent sous la valeur nominale lorsque p prend des valeurs entre 0,1 et 0,9 puis la méthode percentile classique a un comportement contraire à cette dernière en ayant plus souvent qu'autrement une couverture au-dessus de 95% en atteignant rarement la valeur nominale pour ces mêmes valeurs de p . De son côté, la méthode bootstrap- t a une couverture presque toujours supérieure à la valeur nominale (couverture minimale de 95,02%) pour tous les p . En regardant la longueur des intervalles de confiance à l'aide de la figure 4.29, nous pouvons conclure que c'est en raison de ses intervalles de confiance trop longs que nous avons ce comportement pour la méthode bootstrap- t . Il est important de noter la forme différente de la courbe de longueur d'intervalle de la méthode bootstrap- t . Nous pouvons observer que pour de petites et grandes valeurs de p , la courbe remonte avant de redescendre et cette dernière est très loin des 3 autres courbes. Ce comportement très différent de la méthode par pseudo-population est expliqué plus en détails dans l'annexe A.1.

Méthode des poids, bootstrap à la Agresti-Coull, $n=40$, $f=0,05$

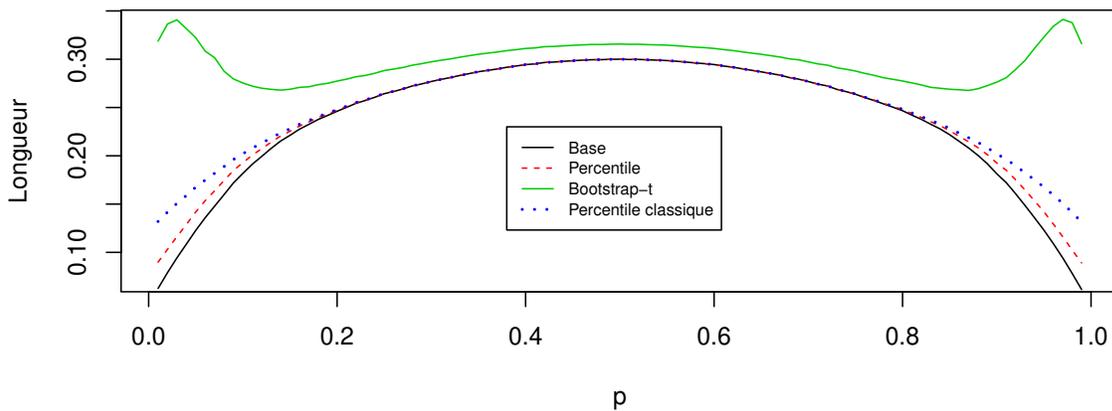


Figure 4.29. Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

La méthode percentile serait donc la méthode à considérer pour l'utilisation du bootstrap à la Agresti-Coull avec la méthode des poids pour un échantillon de taille 40 et une fraction

de sondage de 5%.

Méthode des poids, bootstrap à la Agresti-Coull, $n=40$, $f=0,20$

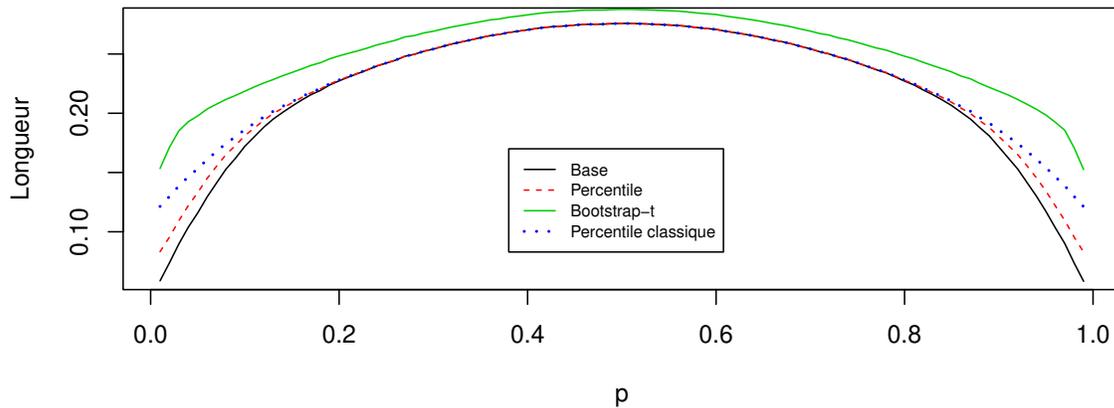


Figure 4.30. Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.

Malgré de très légères différences entre l'utilisation d'une fraction de sondage de 5% et de 20% pour quelques méthodes, ce n'est pas suffisant pour avoir des conclusions différentes de celles effectuées avec une fraction de sondage de 0,05.

À l'aide des figures 4.28 et 4.30 on peut y tirer les mêmes conclusions qu'avec une fraction de sondage de 0,05 concernant l'efficacité des méthodes bootstrap. La méthode bootstrap de base demeure la méthode qui offre les intervalles de confiance les plus courts, mais a une couverture qui oscille énormément dans les extrémités. De son côté, la méthode percentile oscille moins dans les extrémités tout en ayant une couverture qui tend vers 100%, comme la méthode bootstrap- t , mais cette dernière offre des intervalles de confiance beaucoup plus longs que la méthode percentile. C'est pourquoi la méthode percentile semble encore une fois être la méthode la plus efficace dans ce contexte. La méthode percentile est donc la méthode à considérer pour estimer par intervalle de confiance un paramètre de proportion lorsqu'on utilise un bootstrap à la Agresti-Coull pour un échantillon de taille 40, peu importe la fraction de sondage.

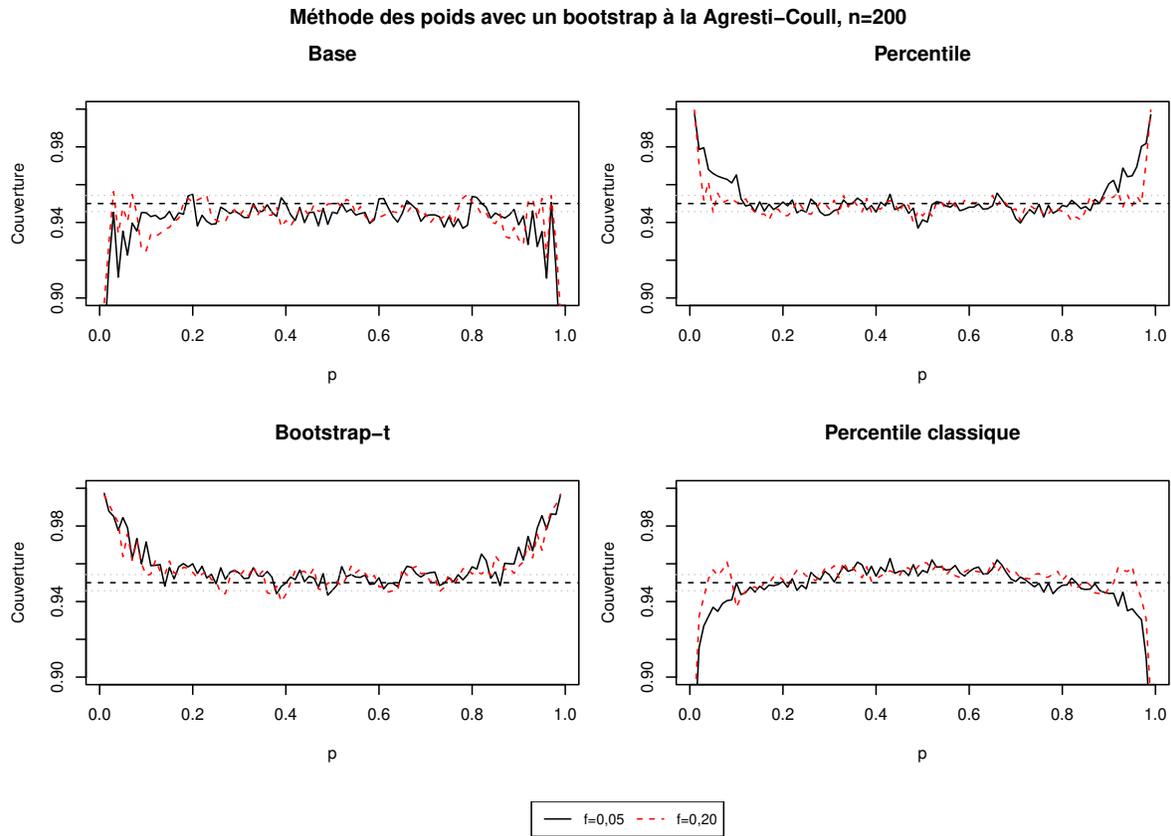


Figure 4.31. Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap à la Agresti-Coull dans un contexte d'échantillonnage avec $n = 200$.

Lorsqu'on considère maintenant une taille échantillonnale de 200 observations, on pouvait s'attendre à des couvertures qui oscillent moins, et ce pour toutes les méthodes bootstrap. C'est effectivement ce que nous remarquons à la figure 4.31 où toutes les méthodes ont des couvertures beaucoup plus constantes. Un fait à remarquer ici est que toutes les méthodes performant plutôt bien pour des valeurs de p entre 0,1 et 0,9. En effet, même si les méthodes percentile et bootstrap- t semblent démontrer plus d'efficacité à être autour de la valeur nominale, les méthodes bootstrap de base ainsi que percentile classique démontrent tout de même beaucoup d'efficacité en ayant une couverture très près de celle-ci. Par contre, ces deux dernières méthodes ne performant pas très bien dans les extrémités en ayant une couverture qui redescend très bas. En effet, pour $f = 0,05$, la couverture minimale de la méthode bootstrap de base est de 86,81% puis celle de la méthode percentile classique est de 86,73%. Au contraire, les méthodes percentile et bootstrap- t ont des couvertures qui

atteignent jusqu'à 100% dans les extrémités en assurant une couverture minimale de 93,7% et de 94,34% respectivement. Si les méthodes percentile et bootstrap- t démontrent autant d'efficacité l'une que l'autre, la méthode percentile semble être la méthode la plus efficace pour de petites et grandes valeurs de p . En effet, malgré que la méthode bootstrap- t tende vers une couverture de 100% pour ces valeurs de p , c'est entre autre dû à ses intervalles de confiance trop longs (figure 4.32). De son côté, la méthode percentile tend vers une couverture de 100% dans les extrémités mais en ayant des intervalles de confiance de longueurs comparables aux méthodes bootstrap de base et percentile classique.

Méthode des poids, bootstrap à la Agresti-Coull, $n=200$, $f=0,05$

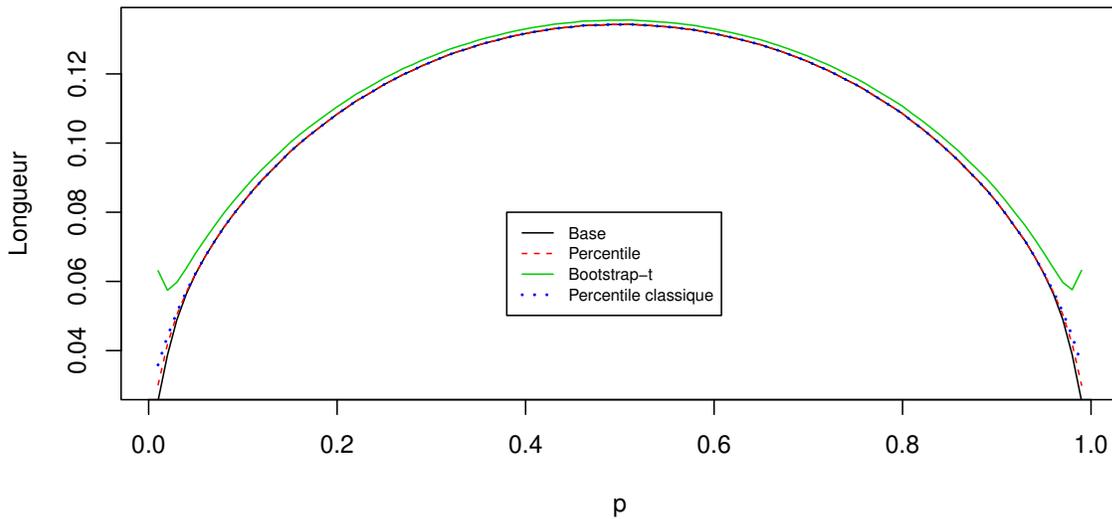


Figure 4.32. Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap à la Agresti-Coull sous un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

En considérant une fraction de sondage de 0,20 (pointillé rouge sur la figure 4.31), que de légères différences sont observables et les conclusions sur l'efficacité des méthodes demeurent les mêmes. En effet, la méthode percentile semble être la meilleure des 4 méthodes afin d'estimer p grâce à sa stabilité autour de la valeur nominale pour des valeurs de p comprises entre 0,05 et 0,95 puis grâce à sa grande couverture aux extrémités sans avoir des intervalles de confiance trop grands.

C'est donc dire que, comme nous l'avons vu dans le cas de la méthode par pseudo-population, la méthode percentile est la meilleure des méthodes bootstrap lorsqu'on utilise un bootstrap à la Agresti-Coull avec la méthode des poids, peu importe la taille échantillonnale et la fraction de sondage.

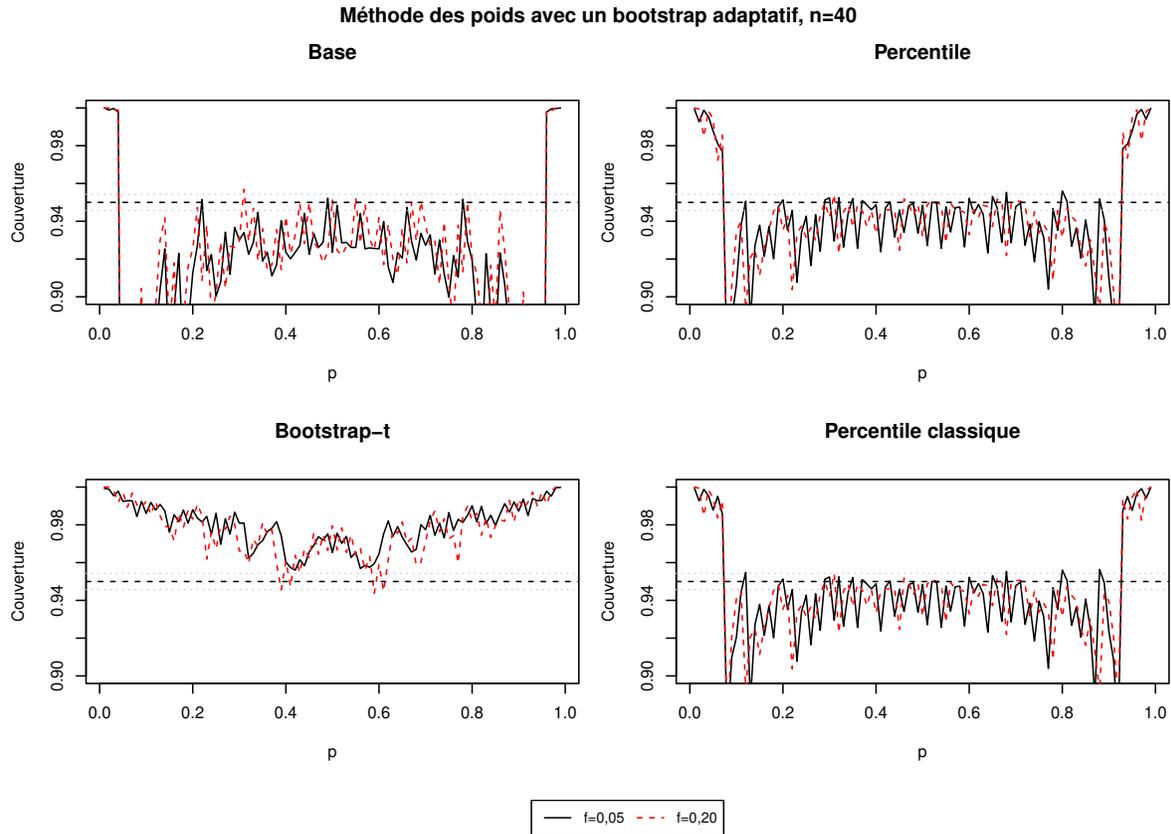


Figure 4.33. Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 40$.

Après avoir analysé les résultats de la simulation lorsqu'on utilise un bootstrap à la Agresti-Coull, intéressons-nous maintenant aux résultats lorsqu'on effectue le bootstrap en considérant l'échantillon \hat{s} , soit le bootstrap adaptatif.

La première chose que nous pouvons remarquer en regardant les couvertures des méthodes bootstrap à l'aide de la figure 4.33 est que pour $n = 40$ et $f = 0,05$ (courbe noire), la méthode bootstrap de base ne fait pas très bien en étant trop souvent en-dessous de la valeur nominale surtout pour des valeurs centrales de p , puis la méthode bootstrap- t a une couverture qui

Méthode des poids, bootstrap adaptatif, $n=40$, $f=0,05$

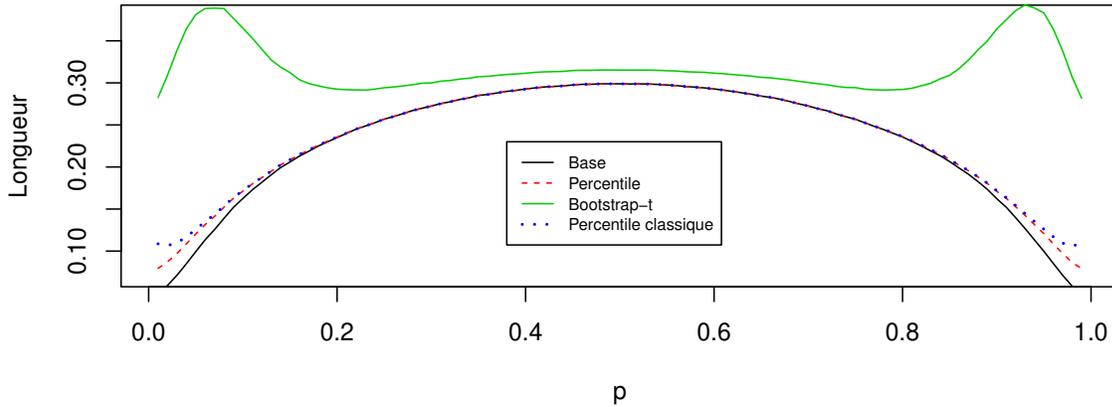


Figure 4.34. Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap adaptatif sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

est toujours au-dessus de la valeur de 95%. En effet, la méthode bootstrap de base a une couverture minimale de 60,23% tandis que celle de la méthode bootstrap- t est de 95,60%. De leur côté, les méthodes percentile et percentile classique ont des couvertures très semblables surtout pour des valeurs de p entre 0,07 et 0,93 en ayant une couverture qui oscille, mais sans être trop en-dessous de la valeur nominale. Ces deux méthodes arrivent régulièrement à atteindre le 95% de couverture en ayant des couvertures minimales de 86,16% pour la méthode percentile et de 86,88% pour la méthode percentile classique. Un fait à remarquer est que ces deux méthodes semblent avoir de la difficulté à couvrir le paramètre d'intérêt pour des valeurs de p entre 0,1 et 0,25 puis entre 0,75 et 0,9. Ce comportement a déjà été aperçu par les méthodes bootstrap dans un contexte i.i.d. puis dans l'utilisation de la méthode par pseudo-population avec l'emploi du bootstrap adaptatif où ces dernières semblaient avoir beaucoup de difficultés à construire des intervalles contenant le paramètre d'intérêt pour ces mêmes valeurs de p .

Si les méthodes percentile et percentile classique démontrent toutes les deux la meilleure efficacité parmi cette classe de méthodes, il semble que la méthode percentile soit préférable puisque dans les extrémités, elle a, comme la méthode percentile classique, une couverture qui atteint jusqu'à 100%, mais en ayant des intervalles de confiance plus courts que cette dernière (figure 4.34). De son côté, la méthode bootstrap- t a des intervalles de confiance très

longs, d'où le fait qu'elle couvre toujours bien au-delà de la valeur nominale pour toutes les valeurs de p . Pour toutes les raisons mentionnées auparavant, la méthode percentile semble être la méthode à considérer pour estimer p par intervalle de confiance avec l'emploi du bootstrap adaptatif lorsque la taille échantillonnale est de 40 observations et la fraction de sondage est égale à 0,05.

Méthode des poids, bootstrap adaptatif, $n=40$, $f=0,20$

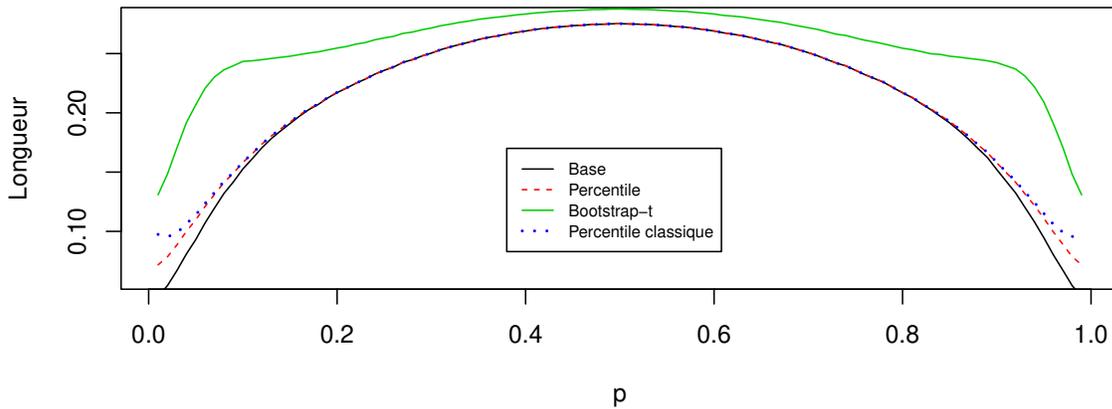


Figure 4.35. Longueur moyenne des intervalles bootstrap pour la méthode des poids à l'aide du bootstrap adaptatif sous un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.

Lorsqu'on considère maintenant une fraction de sondage égale à 0,20 (courbe pointillée rouge sur la figure 4.33), nous pouvons voir que pour les méthodes percentile et percentile classique, les couvertures sont très similaires à celles observées lorsque la fraction de sondage est de 0,05. Même si de légers changements s'observent, ces différences sont toutefois minimales et les constatations semblent vouloir demeurer les mêmes que celles que nous avons faites lorsque la fraction de sondage est de 0,05 puisqu'en regardant la figure 4.35, nous avons toujours que les intervalles bootstrap- t sont très longs, puis que les intervalles percentile sont plus courts que ceux de la méthode percentile classique et que pour ces raisons, la méthode percentile semble être celle qui performe le mieux.

En observant maintenant les couvertures des méthodes bootstrap avec $n = 200$ et $f = 0,05$, nous pouvons constater que pour des valeurs de p entre 0,3 et 0,7, toutes les méthodes performant bien en ayant des couvertures très près de 95%. Par contre, la méthode bootstrap

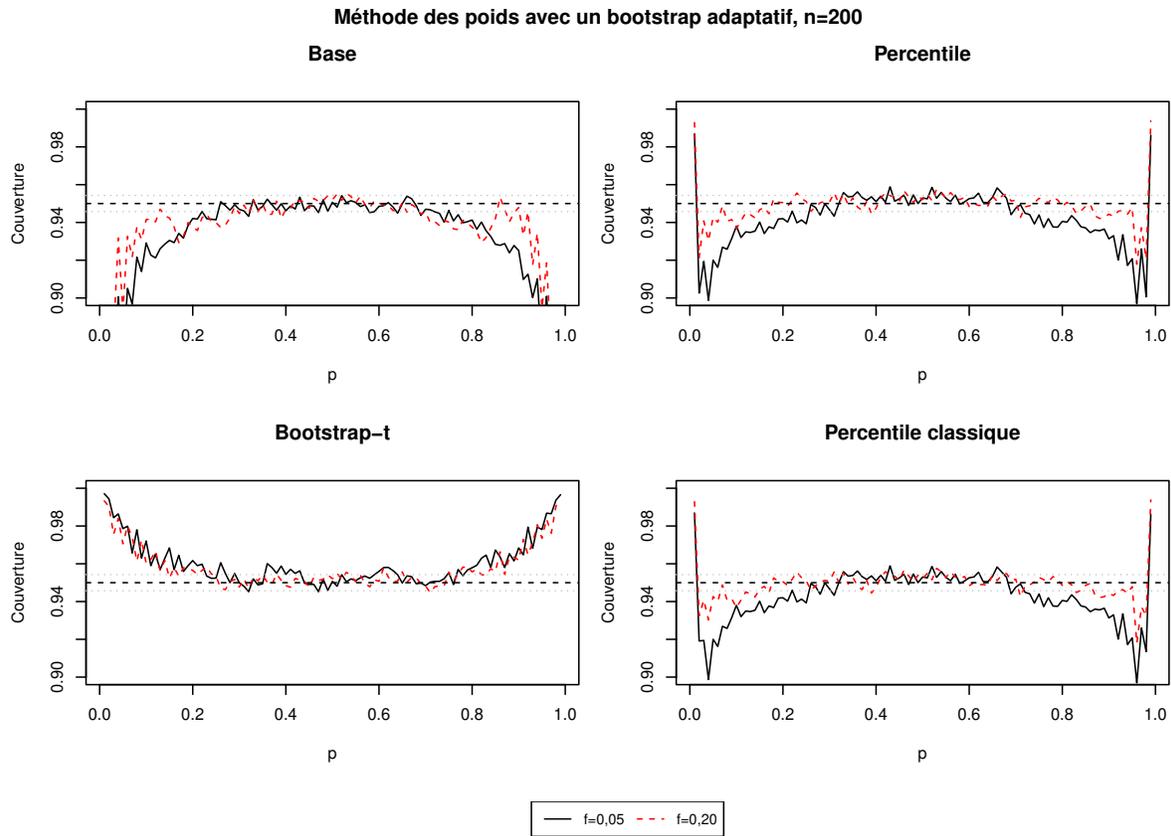


Figure 4.36. Couverture des méthodes bootstrap sous la méthode des poids avec l'utilisation du bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$.

de base ne performe pas très bien dans les extrémités en ayant une couverture minimale de 59,74%. Encore ici, la méthode bootstrap- t a de son côté une couverture qui tend vers 100% pour de petites ou grandes valeurs de p , mais ceci vient avec des intervalles de confiance trop longs (figure 4.37). À l'aide de la figure 4.37, on peut constater que les couvertures des méthodes percentile et percentile classique sont les mêmes sauf pour p égal à 0,02 et 0,98 où il semble que la courbe de la méthode percentile classique redescende moins bas que la courbe de la méthode percentile en étant un peu plus près de la valeur nominale. Nous avons noté un comportement identique entre les méthodes percentile et percentile classique pour la méthode par pseudo-population avec l'utilisation du bootstrap adaptatif pour une taille échantillonnale de 200 observations comme ici.

C'est plutôt intéressant de voir ce qui se passe avec l'emploi du bootstrap adaptatif avec $n = 200$ et une fraction de sondage égale à 0,20 (courbe pointillée rouge sur la figure 4.36).

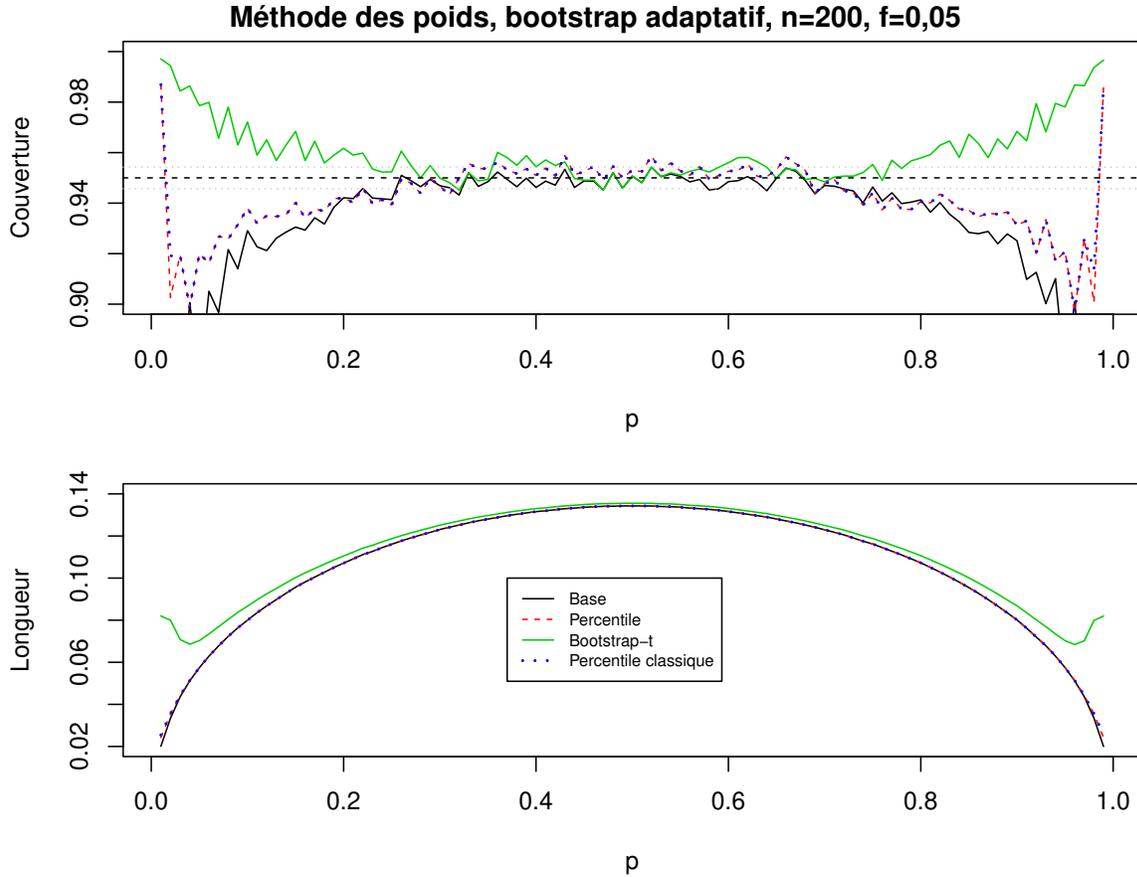


Figure 4.37. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

Si pour $n = 40$, des différences minimes étaient détectables entre l'utilisation des 2 fractions de sondage, il semble qu'ici les différences soient un peu plus marquées. En effet, mis à part la méthode bootstrap- t qui offre toujours une couverture la plupart du temps au-dessus de la valeur nominale, il semble que les 3 autres méthodes démontrent toutes une meilleure efficacité en terme de couverture pour des valeurs de p entre 0,05 et 0,3 puis entre 0,7 et 0,95. La méthode bootstrap de base a une couverture qui semble vouloir s'approcher de la valeur nominale pour ces mêmes valeurs de p en n'étant toutefois pas très efficace pour couvrir le paramètre d'intérêt de manière générale. De leur côté, les méthodes percentile et percentile classique sont beaucoup plus efficaces pour ces valeurs de p en assurant une couverture minimale supérieure à celle lorsque la fraction de sondage est de 0,05. En effet, ces dernières assurent des couvertures minimales de 91,79% lorsque la fraction de sondage

est de 0,20. Les deux méthodes ont une couverture minimale de 89,7% lorsque la fraction de sondage est plutôt de 0,05.

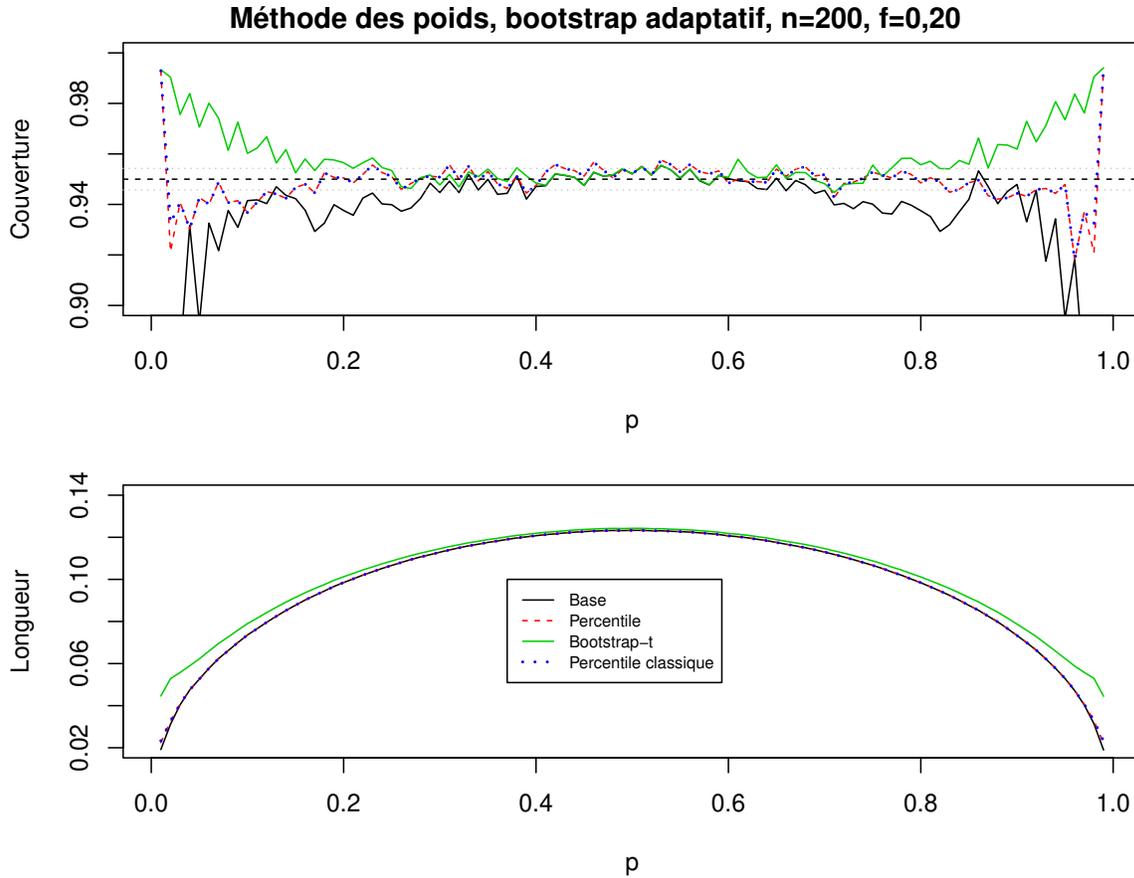


Figure 4.38. Couverture et longueur moyenne des intervalles des méthodes bootstrap sous le bootstrap adaptatif dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,20$.

Encore ici, même pour une fraction de sondage de 0,20, il est possible de remarquer que les deux couvertures des méthodes percentile et percentile classique sont identiques pour beaucoup de valeurs de p (figure 4.38). En effet, comme nous l'avons remarqué sous l'utilisation du bootstrap adaptatif sous la méthode par pseudo-population, les seules différences remarquables sont lorsque p vaut 0,02 (1,12% de différence) et 0,98 (1,16% de différence) où la couverture de la méthode percentile classique semble encore une fois supérieure à la couverture de la méthode percentile. Comme les deux méthodes ont des intervalles de longueurs comparables puis que les méthodes bootstrap de base et bootstrap- t sont trop souvent respectivement en-dessous et au-dessus de la valeur nominale en ayant des intervalles très

longs pour cette dernière, les méthodes percentile et percentile classique semblent être les deux meilleures méthodes afin d'estimer p par intervalle de confiance dans ce contexte.

Rappelons que la méthode percentile est le candidat idéal afin d'estimer le paramètre de proportion lorsqu'on effectue un bootstrap adaptatif avec une taille échantillonnale de 40 observations. En effet, dans ce contexte, peu importe la fraction de sondage, même si les couvertures des méthodes percentile et percentile classique étaient très similaires, la méthode percentile offrait toujours des intervalles de confiance plus courts que cette dernière dans les extrémités. En considérant des échantillons de taille 200, nous n'observons plus une différence dans la longueur des intervalles percentile et percentile classique dans les extrémités. Donc à longueur comparable, il va de soi de considérer la méthode couvrant le mieux, et c'était le cas de la méthode percentile classique, mais où elle faisait mieux que la méthode percentile pour seulement 2 valeurs de p soit pour p égal à 0,02 et 0,98. Par contre, comme nous l'avons noté plus haut, les différences sont négligeables. On peut donc considérer la méthode percentile comme la meilleure méthode parmi les méthodes bootstrap afin d'estimer p lorsqu'on emploie un bootstrap adaptatif. Cette dernière offre la meilleure couverture en plus d'avoir des intervalles plus courts que la méthode percentile classique pour $n = 40$ et démontre exactement la même efficacité que cette dernière, à deux valeurs de p près, lorsque $n = 200$.

Comme nous l'avons fait lorsqu'on était dans le cadre de la méthode par pseudo-population, la figure 4.39 nous aide à comparer l'efficacité entre l'utilisation du bootstrap à la Agresti-Coull et du bootstrap adaptatif pour l'intervalle percentile sous la méthode des poids. Encore ici, nous pouvons conclure la même chose que lorsqu'on utilise la méthode par pseudo-population, c'est-à-dire que l'utilisation du bootstrap à la Agresti-Coull est préférable à l'utilisation du bootstrap adaptatif lorsqu'on considère l'intervalle percentile sous la méthode des poids.

4.3.3. Méthodes traditionnelles

Lorsqu'on s'intéresse désormais au comportement des méthodes traditionnelles lorsque la taille échantillonnale est de 40 observations et la fraction de sondage est de 5% (lignes noires sur les graphiques de la figure 4.40), on peut voir que les méthodes qui performant le

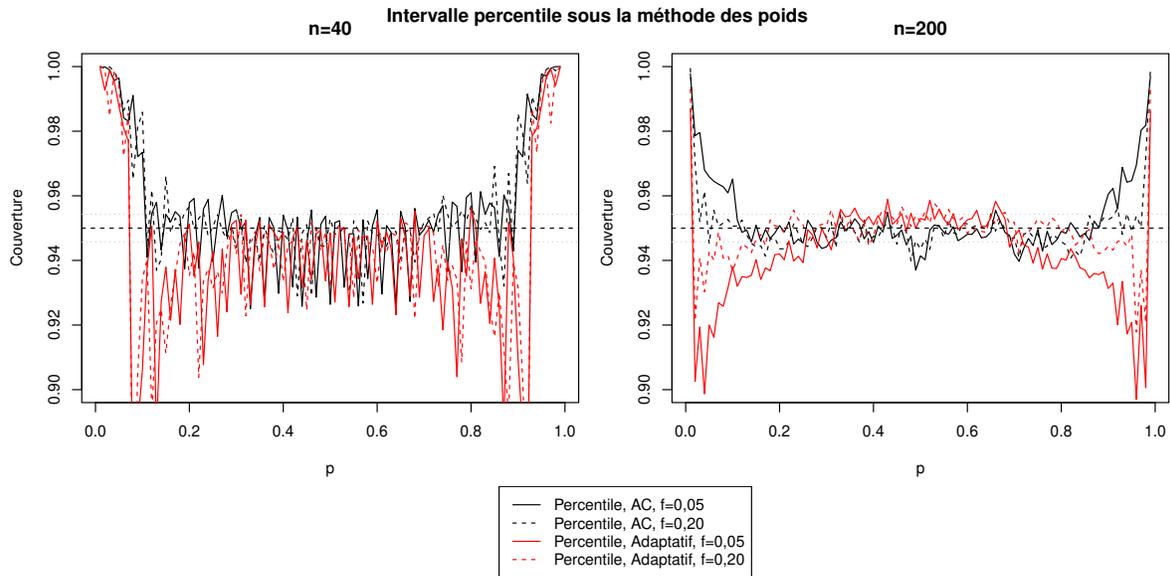


Figure 4.39. Couvertures de l'intervalle percentile sous la méthode des poids.

mieux sont les méthodes de Wilson, Agresti-Coull et ASSQR. En effet, les trois méthodes assurent des couvertures minimales relativement près de la valeur nominale de 95%. La couverture minimale de la méthode de Wilson est de 92,67% tandis que celle de la méthode Agresti-Coull est de 93,95% et la couverture minimale qu'assure la méthode ASSQR est de 92,85%. Si on s'intéresse maintenant au comportement des méthodes de Wald, Logit et Clopper-Pearson, on peut voir ici que les deux premières démontrent beaucoup de difficultés à couvrir le paramètre d'intérêt dans les extrémités surtout pour la méthode de Wald qui assure une couverture minimale très faible de 32,37%, la couverture minimale de la méthode Logit est de 88,49%, puis la méthode Clopper-Pearson a toujours une couverture supérieure à la valeur nominale par construction ; voir la section 3.5. Les résultats de la simulation nous montrent qu'à cause de ce comportement, ses intervalles de confiance sont très longs.

Lorsqu'on s'intéresse particulièrement aux 3 meilleures méthodes parmi cette classe de méthodes, on peut voir que, malgré qu'elle a la couverture minimale la plus élevée, la méthode Agresti-Coull a beaucoup de difficultés à être près de la valeur nominale dans les extrémités et elle est la méthode qui offre les intervalles les plus longs dans ce contexte (figure 4.41). De leur côté, les méthodes de Wilson et ASSQR sont très comparables en terme d'efficacité. On a d'un côté la méthode de Wilson qui est toujours très près du 95% et qui a une couverture minimale

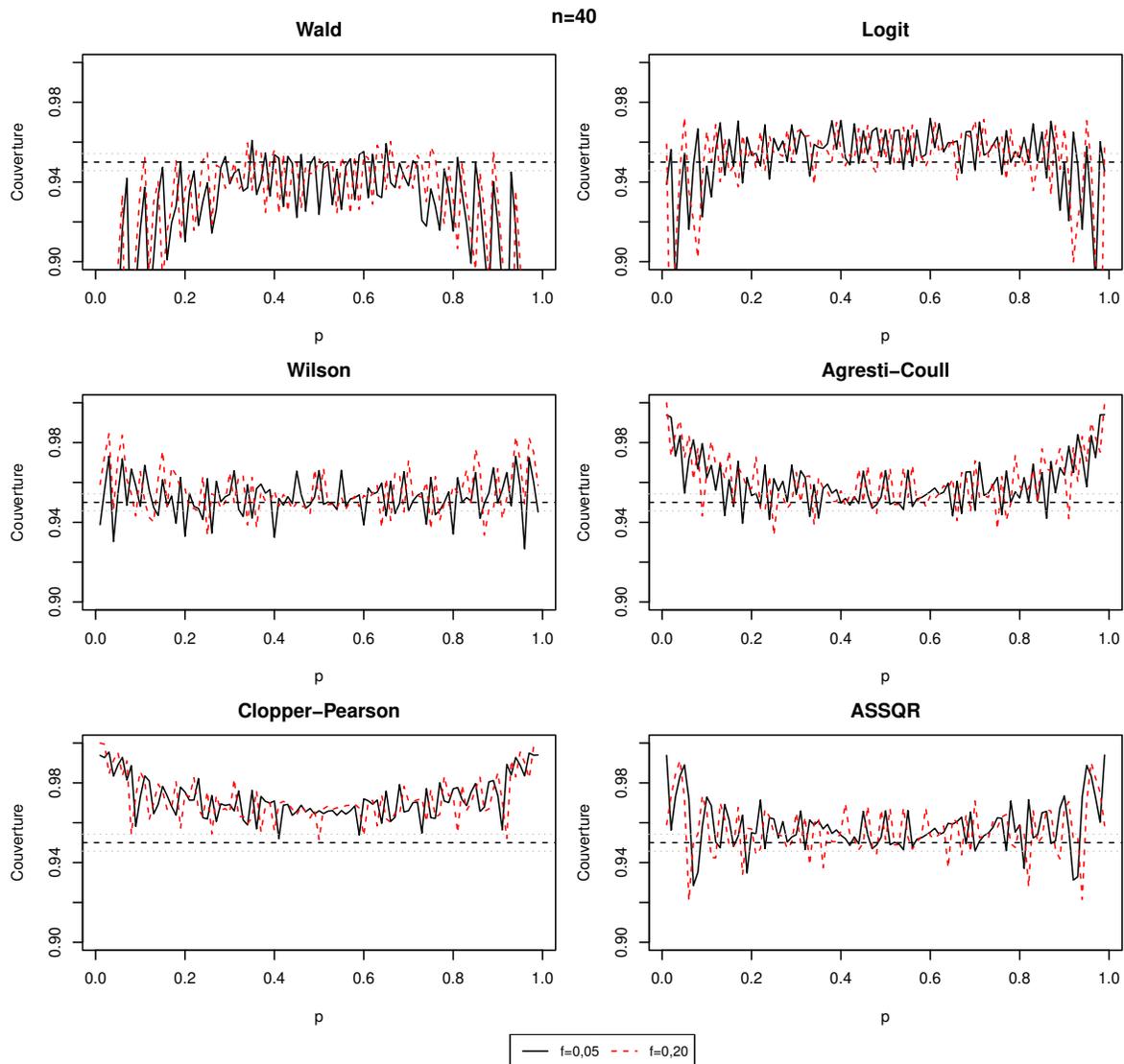


Figure 4.40. Couverture des méthodes traditionnelles dans un contexte d'échantillonnage avec $n = 40$.

extrêmement comparable à celle de la méthode ASSQR. D'un autre côté, la méthode ASSQR performe un peu mieux pour de petites et grandes valeurs de p . En effet, elle a des intervalles plus courts que la méthode de Wilson en couvrant plus. Reste que les intervalles sont de longueurs très comparables, d'où le fait que les deux méthodes performent très bien dans ces circonstances.

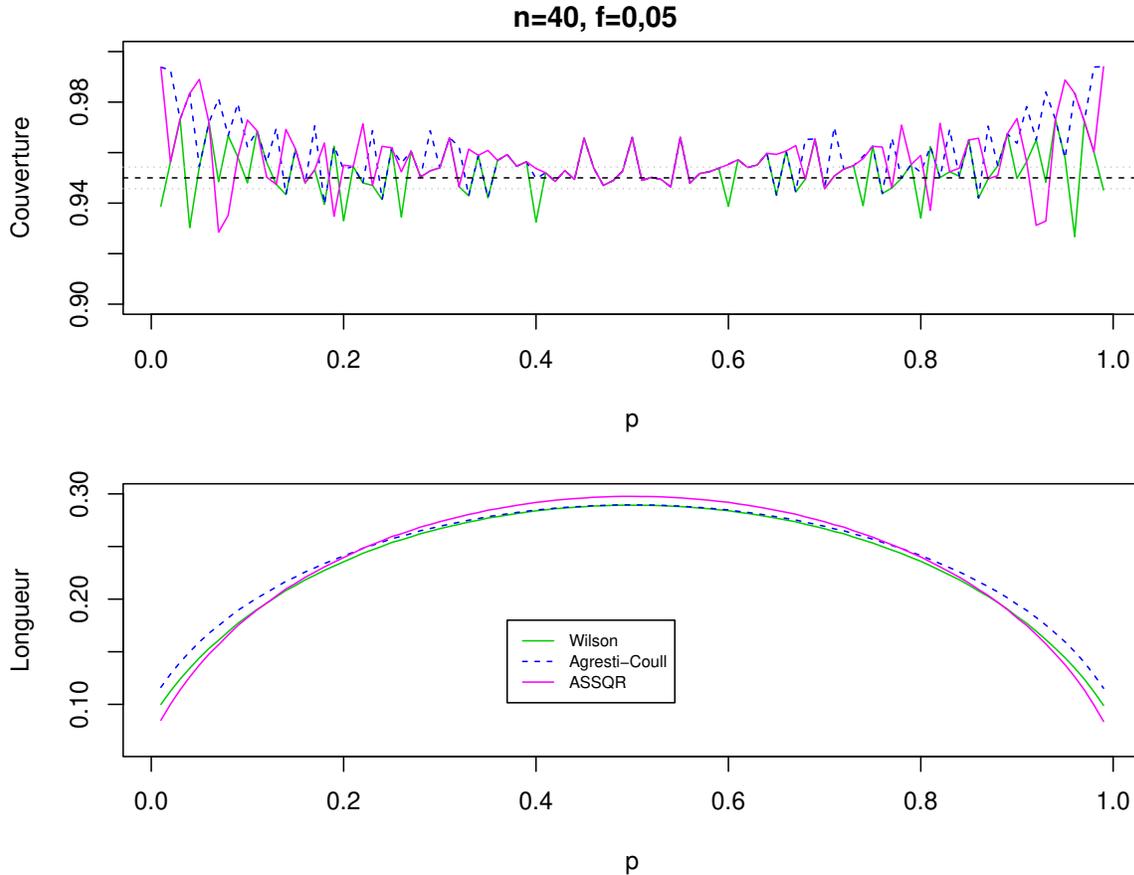


Figure 4.41. Couverture et longueur moyenne des intervalles des méthodes de Wilson, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

Les mêmes conclusions sont tirées lorsqu'on considère plutôt une fraction de sondage de 20% (figure 4.42). En effet, la méthode Agresti-Coull a les intervalles les plus longs comparativement aux deux autres méthodes qui couvrent toujours très bien. Les méthodes de Wilson et ASSQR effectuent tous les deux un excellent travail afin d'estimer p .

Avec une taille échantillonnale de 200 observations, on peut voir que beaucoup de méthodes performant très bien (figure 4.43). En effet, mis à part la méthode de Wald qui a toujours beaucoup de difficultés à avoir des intervalles de confiance incluant p lorsque celui-ci prend des valeurs très petites ou très grandes, ainsi que la méthode Clopper-Pearson qui est encore une fois ici trop souvent au-dessus de la valeur nominale, les quatre autres méthodes ont des couvertures assez stables autour du 95% en assurant des couvertures minimales assez élevées. En effet, pour $f = 0,05$, la couverture minimale des méthodes Logit et Wilson sont de

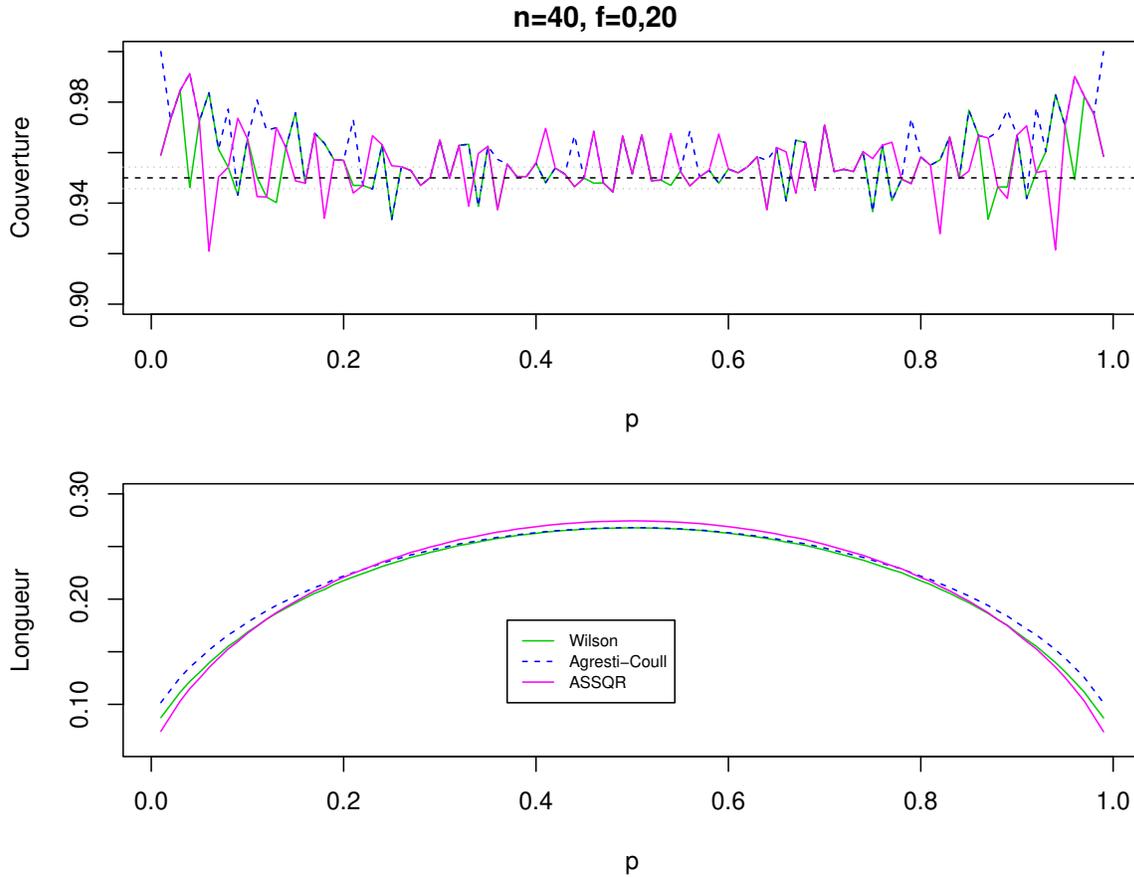


Figure 4.42. Couverture et longueur moyenne des intervalles des méthodes de Wilson, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,20$.

93,39% et 93,13% respectivement puis de 94,00% et 93,42% pour les méthodes Agresti-Coull et ASSQR.

En observant de plus près le comportement de ces 4 meilleures méthodes traditionnelles dans ce contexte (figure 4.44), la première chose à remarquer est que toutes les 4 méthodes ont des intervalles de confiance de longueurs très comparables. C'est un comportement que nous avons déjà vu concernant les différentes méthodes, traditionnelles ou bootstrap lorsque la taille échantillonnale est de 200. Les 4 méthodes sont toutes très près de la valeur nominale pour des valeurs centrales de p . Dans les extrémités, les méthodes Agresti-Coull et ASSQR seraient les méthodes à considérer, à cause du fait qu'elles ont les couvertures les plus élevées, malgré leurs intervalles de longueurs comparables aux autres méthodes. Un comportement similaire est noté pour une fraction de sondage de 20% et les conclusions demeurent les mêmes.

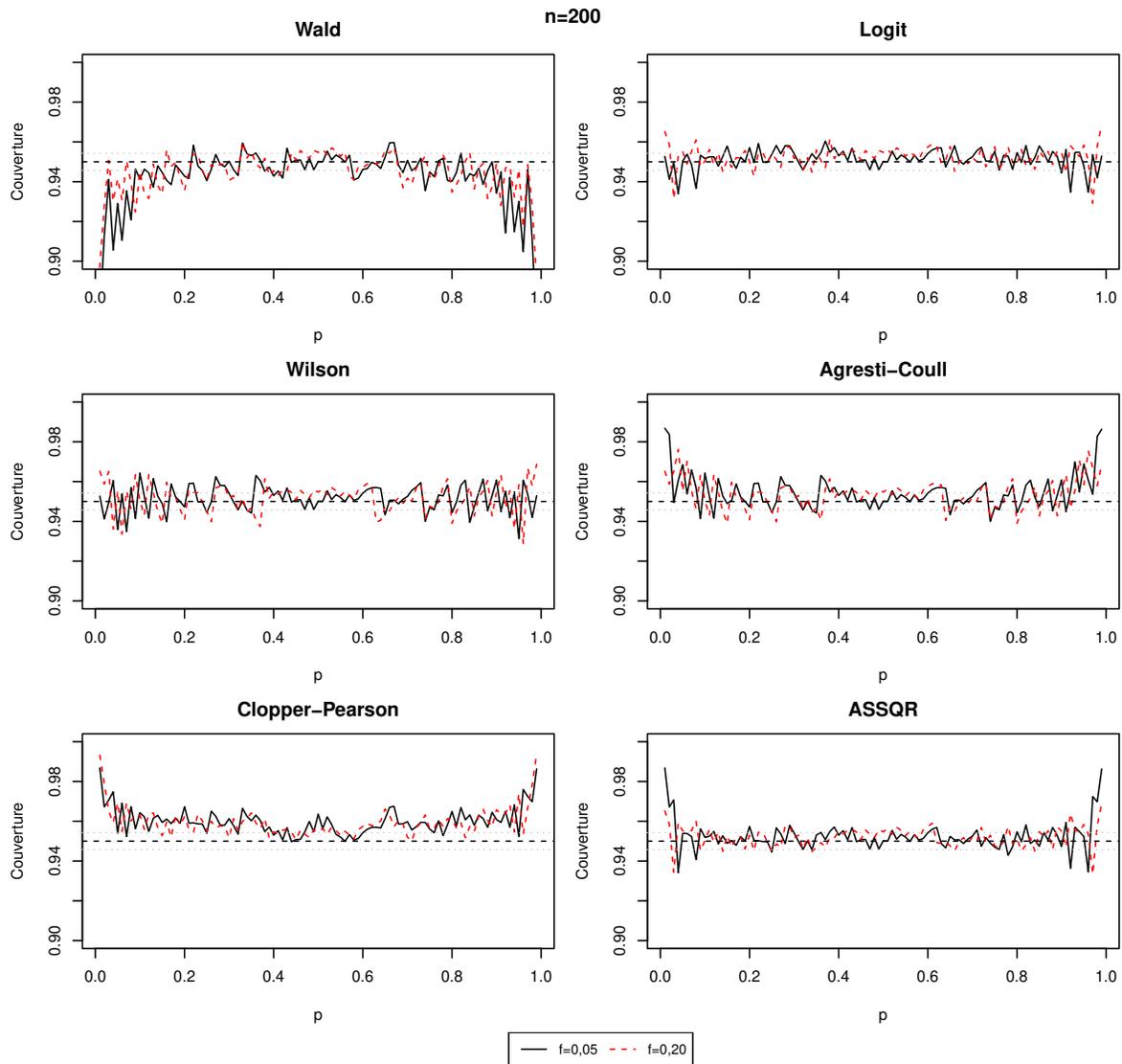


Figure 4.43. Couverture des méthodes traditionnelles dans un contexte d'échantillonnage avec $n = 200$.

À cause du fait que la méthode ASSQR était la méthode à considérer pour une taille échantillonnale de 40 observations et qu'elle est l'une des deux méthodes les plus efficaces afin d'estimer p par intervalle de confiance pour une taille échantillonnale de 200 observations, on peut la considérer comme la meilleure méthode traditionnelle afin d'estimer par intervalle

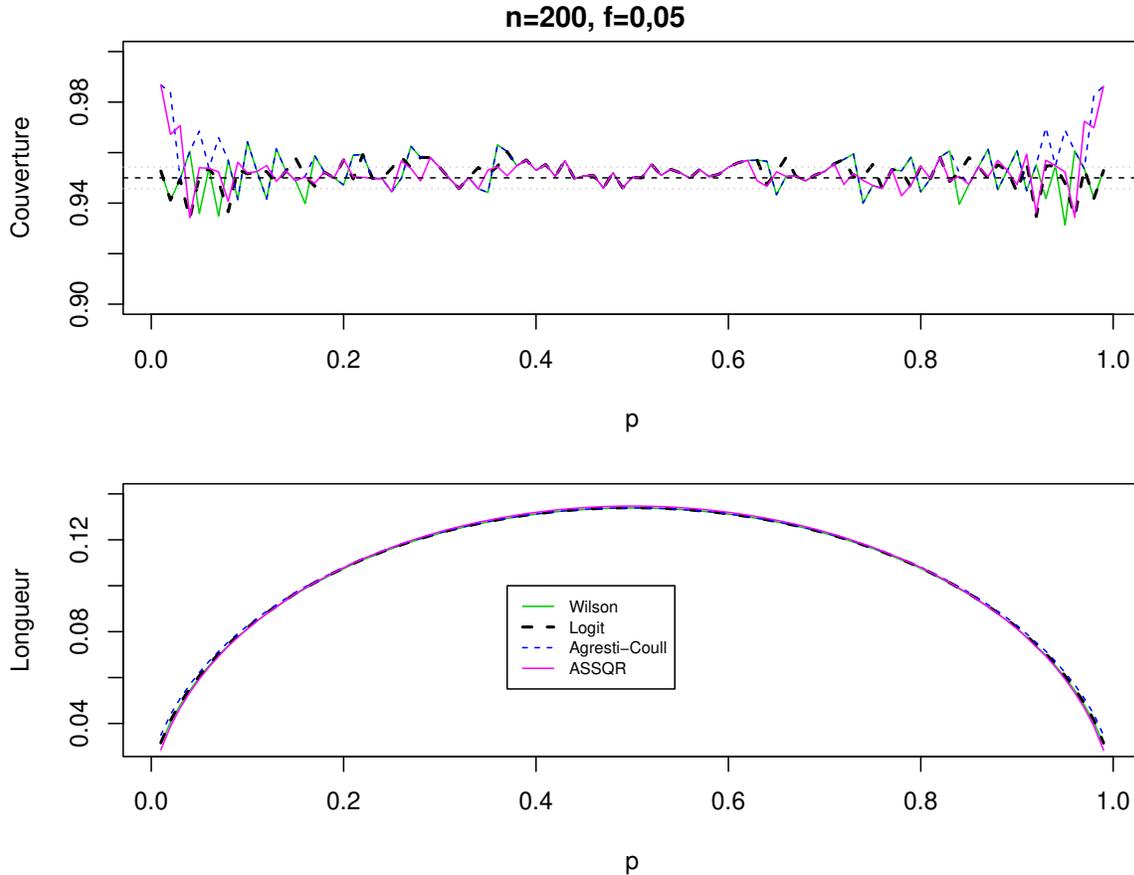


Figure 4.44. Couverture et longueur moyenne des intervalles des méthodes de Wilson, Logit, Agresti-Coull et ASSQR dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

de confiance un paramètre de proportion lorsqu'on tire un échantillon sans remise à partir de la population.

4.3.4. Comparaisons des meilleures méthodes entres elles

L'un des points à retenir concernant les résultats de notre simulation est que le bootstrap adaptatif n'effectue pas un bon travail de manière générale. En effet, peu importe que la méthode bootstrap utilisée soit celle des poids ou par pseudo-population, le bootstrap adaptatif démontre beaucoup de difficultés à estimer p lorsque celui-ci prend des valeurs près de 0,1 et 0,9. Le bootstrap à la Agresti-Coull est de loin préférable où la meilleure des 4 méthodes bootstrap dans ce contexte est la méthode percentile, et ce, peu importe qu'on utilise la méthode par les poids ou par pseudo-population.

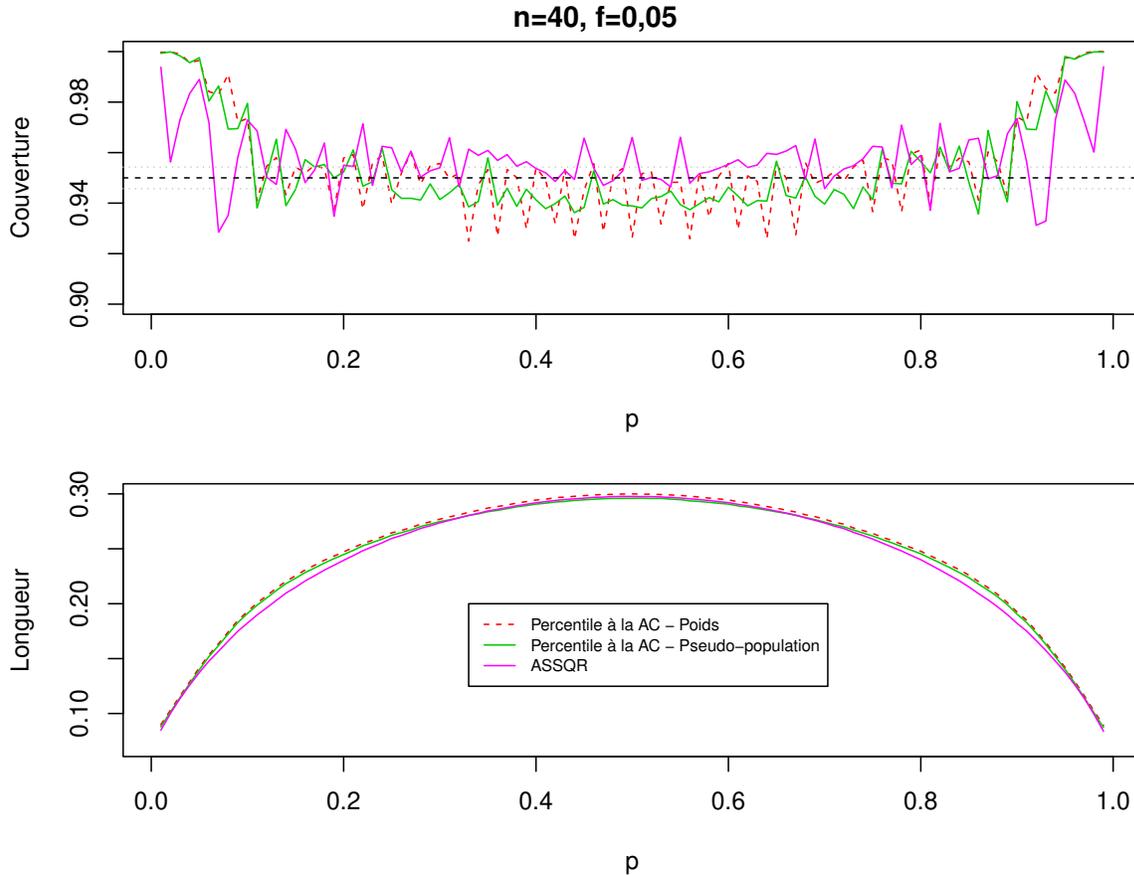


Figure 4.45. Couverture et longueur moyenne des intervalles percentile avec un bootstrap à la Agresti-Coull sous la méthode des poids et par pseudo-population avec la méthode ASSQR dans un contexte d'échantillonnage avec $n = 40$ et $f = 0,05$.

À la figure 4.45, nous avons réuni sur le même graphique les méthodes percentile avec un bootstrap à la Agresti-Coull (noté AC sur le graphique) sous la méthode des poids et sous la méthode par pseudo-population avec la meilleure des méthodes traditionnelles soit la méthode ASSQR lorsque la taille échantillonnale est de 40 observations avec une fraction de sondage de 5%. En plus d'avoir des couvertures minimales très comparables, les 3 méthodes sont toutes très près de la valeur nominale en ayant une couverture très élevée dans les extrémités. On peut justement voir dans ce contexte que les méthodes bootstrap sont préférables puisqu'elles ont des intervalles de même longueur que la méthode ASSQR en couvrant davantage. Comme on constate le même comportement général pour une fraction de sondage de 20% on ne présente pas de nouveau graphique.

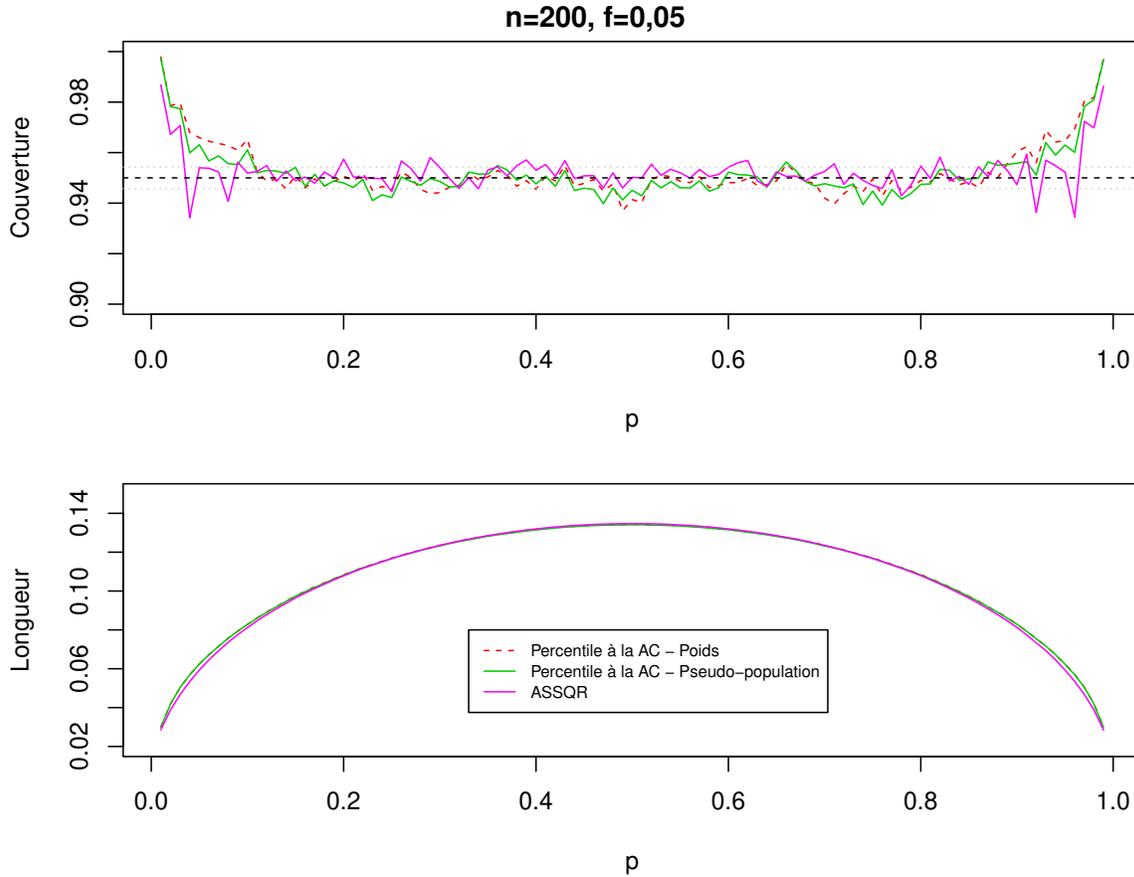


Figure 4.46. Couverture et longueur moyenne des intervalles percentile avec un bootstrap à la Agresti-Coull sous la méthode des poids et par pseudo-population avec la méthode ASSQR dans un contexte d'échantillonnage avec $n = 200$ et $f = 0,05$.

À l'aide de la figure 4.46, on peut voir le comportement des 3 méthodes avec une taille échantillonnale de 200 observations et une fraction de sondage de 0,05. Encore ici, il est possible de constater que les méthodes percentile sous la méthode des poids et par pseudo-population à la Agresti-Coull permettent de rivaliser avec la meilleure méthode traditionnelle dans ce contexte. En plus d'avoir des intervalles de longueurs comparables entre elles, les méthodes bootstrap ont une couverture minimale légèrement supérieure à la méthode ASSQR en plus de couvrir un peu plus dans les extrémités. Un comportement similaire se reflète avec une fraction de sondage de 20%.

Force est de constater que les méthodes percentile avec un bootstrap à la Agresti-Coull sous la méthode des poids et par pseudo-population réussissent à bien compétitionner avec la meilleure méthode traditionnelle dans un contexte d'échantillonnage.

Conclusion

Lorsqu'on utilise la méthode générale du bootstrap pour construire un intervalle de confiance pour un paramètre de proportion, nous nous retrouvons en difficulté lorsque l'estimateur usuel \hat{p} prend la valeur 0 ou 1. En effet, les intervalles de confiance bootstrap deviennent dégénérés en ne contenant qu'un point unique soit 0 ou 1. L'intervalle de confiance perd alors tout son sens.

Nous avons donc proposé essentiellement 2 approches bootstrap utilisant le même type d'estimateur considéré dans les méthodes de [Wilson, 1927] et [Agresti et Coull, 1998], soit basé sur l'estimateur $\tilde{p} = (X + 2)/(n + 4)$, où X est le nombre de succès parmi un échantillon de taille n . Si l'une des deux approches consiste à utiliser dans tous les cas cet estimateur lors du bootstrap, méthode que l'on nomme bootstrap à la Agresti-Coull, la seconde utilise cet estimateur seulement lorsque \hat{p} est nul ou égal à 1, sinon elle utilise l'estimateur \hat{p} . Nous avons nommé cette dernière méthode le bootstrap adaptatif.

De manière générale, les résultats de nos simulations ont démontré que peu importe que nous soyons dans un contexte de statistique classique avec la distribution de la loi binomiale ou dans un contexte d'échantillonnage avec un plan aléatoire simple sans remise, l'approche du bootstrap adaptatif est beaucoup moins efficace qu'une approche par un bootstrap à la Agresti-Coull. En effet, pour un intervalle critique autour d'une faible proportion qui dépend de la taille de l'échantillon, le bootstrap adaptatif a de la difficulté à bien choisir entre utiliser \tilde{p} ou \hat{p} pour estimer p afin de construire les intervalles bootstrap ce qui mène à des taux de couverture beaucoup plus éloignés de la valeur nominale qu'avec le bootstrap à la Agresti-Coull. Au final, la méthode percentile semble être la méthode bootstrap la plus efficace afin d'estimer par intervalle un paramètre de proportion dans un contexte bootstrap. En effet, elle fait mieux que les méthodes bootstrap de base,

bootstrap- t ainsi que bootstrap percentile classique de manière générale. C'est pourquoi, à la suite de nos simulations, il est clair qu'un intervalle de confiance de type bootstrap percentile en utilisant un bootstrap à la Agresti-Coull est préférable afin d'estimer par intervalle de confiance un paramètre de proportion. Rappelons que la borne inférieure de l'intervalle percentile consiste à additionner à l'estimateur \hat{p} le quantile $\alpha/2$ de la distribution de \hat{p}^* moins le paramètre bootstrap et non pas seulement de considérer le quantile $\alpha/2$ de la distribution de \hat{p}^* . Le paramètre bootstrap dépend de la méthode utilisée mais peut être, par exemple dans un contexte i.i.d., l'estimateur \tilde{p} calculé sur l'échantillon original.

Dans le cas où les données sont générées d'une manière indépendante suivant une même loi de Bernoulli(p), la méthode Arcsine Square Root demeure la méthode traditionnelle la plus efficace selon nos simulations. Nous avons conclu que la méthode bootstrap percentile sous un bootstrap à la Agresti-Coull peut facilement compétitionner avec la méthode Arcsine Square Root peu importe la taille échantillonnale et la vraie valeur de p .

Dans le cas où nous tirons un échantillon de taille n sans remise de la population, nous avons distingué 2 méthodes bootstrap soit celle par pseudo-population de [Booth *et al.*, 1994] ainsi que celle utilisant les poids de [Rao *et al.*, 1992]. Lorsqu'on choisit une approche par une pseudo-population, nos simulations ont démontré encore ici que la méthode percentile était la méthode bootstrap la plus efficace lors de l'utilisation d'un bootstrap à la Agresti-Coull, et ce, peu importe la taille échantillonnale et la fraction de sondage. La méthode percentile offre dans tous les cas la couverture minimale la plus élevée en plus d'être extrêmement comparable à l'une des meilleures méthodes traditionnelles dans ce contexte, soit la méthode Arcsine Square Root. La méthode percentile est très efficace en terme de couverture même lorsque p est près de 0 ou 1 en plus d'offrir des intervalles plus courts que les méthodes de Wilson et Agresti-Coull en ayant une couverture supérieure. Les conclusions sont similaires lorsqu'on utilise plutôt la méthode par les poids. La méthode bootstrap qui semble être la meilleure peu importe la taille échantillonnale et la fraction de sondage est la méthode percentile encore ici. Elle offre le meilleur compromis entre bonne couverture générale et longueur d'intervalle. Son comportement et son efficacité dans les extrémités sont comparables à ceux de la méthode percentile sous la méthode par pseudo-population. Par le fait même, elle représente une alternative efficace à la meilleure

méthode traditionnelle ici soit la méthode Arcsine Square Root.

Il pourrait être intéressant de voir si l'utilisation d'un bootstrap à la Agresti-Coull demeure efficace dans un contexte d'échantillonnage avec des plans de sondage plus complexes comme un plan de sondage stratifié, par grappes ou encore un plan de sondage combinant les deux. D'un côté, généraliser ce qui a été présenté dans le cas d'un plan stratifié serait assez trivial. En effet, nous n'aurions qu'à effectuer le même processus et les mêmes modifications mais pour chacune des strates. D'un autre côté, si ajouter deux succès et deux échecs à un échantillon où les probabilités de sélection sont identiques est chose simple, de considérer la même solution dans un plan de sondage avec des probabilités de sélection qui diffèrent entre les individus en est une autre et il n'est pas clair comment les méthodes par pseudo-population et par les poids pourraient être modifiées. Finalement, en constatant l'efficacité de la méthode Arcsine Square Root, qui utilise essentiellement un estimateur comparable à celui d'Agresti-Coull mais qui est plutôt basé sur $(X + 0,5)/(n + 1)$, il pourrait être envisageable de considérer la même solution que nous avons présentée, mais en considérant plutôt cet estimateur à la place de l'estimateur $\tilde{p} = (X + 2)/(n + 4)$. On pourrait penser que d'ajouter moins d'observations à l'échantillon initial nous permettrait d'être plus efficace car ce dernier serait moins « différent » de l'estimateur habituel \hat{p} et nous permettrait encore de pouvoir contourner la problématique lorsque \hat{p} est nul ou égal à 1. D'un autre côté, on pourrait également penser que cette solution serait moins bonne, considérant que l'estimateur d'Arcsine Square Root devient plus petit que celui d'Agresti-Coull lorsque \hat{p} est nul générant donc plus souvent des échantillons nuls.

Bibliographie

- [Agresti et Coull, 1998] AGRESTI, A. et COULL, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician Vol.52(2)*, pages 119–126.
- [Anscombe, 1956] ANSCOMBE, F. (1956). On estimating binomial response relations. *Biometrika Vol.43(3/4)*, pages 461–464.
- [Booth et al., 1994] BOOTH, J., BUTLER, R. et HALL, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association Vol.89(428)*, pages 1282–1289.
- [Brown et al., 2001] BROWN, L. D., CAI, T. et DASGUPTA, A. (2001). Interval estimation for a binomial proportion. *Statistical Science Vol.16(2)*, pages 101–133.
- [Clopper et Pearson, 1934] CLOPPER, C. et PEARSON, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika Vol.26(4)*, pages 404–413.
- [Efron, 1979] EFRON, B. (1979). Bootstrap methods : Another look at the jackknife. *The Annals of Statistics Vol.7(1)*, pages 1–26.
- [Franco et al., 2019] FRANCO, C., LITTLE, R., LOUIS, T. et SLUD, E. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology Vol.7(3)*, pages 334–364.
- [Gilary et al., 2012] GILARY, A., MAPLES, G. et SLUD, E. (2012). Small area confidence bounds on small cell proportions in survey populations. *JSM proceedings, Survey Research Section, American Statistical Association*, pages 3541–3555.
- [Horvitz et Thompson, 1952] HORVITZ, D. et THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association Vol.47(260)*, pages 663–685.
- [Korn et Graubard, 1998] KORN, E. L. et GRAUBARD, B. I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology Vol.24(2)*, pages 1030–1039.
- [Mantel et Neusy, 2015] MANTEL, H. et NEUSY, E. (2015). *Reporting the sampling variability of proportions*. Technical Report Presented at Statistics Canada’s Advisory Committee on Statistical Methods.
- [Mashreghi et al., 2016] MASHREGHI, Z., HAZIZA, D. et LÉGER, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys Vol.10*, pages 1–52.

- [Rao et Wu, 1988] RAO, J. et WU, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* Vol.83(401), pages 231–241.
- [Rao et al., 1992] RAO, J., WU, C. et YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* Vol.18(2), pages 209–217.
- [Särndal et al., 1992] SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New-York : Springer-Verlag Inc.
- [Wilson, 1927] WILSON, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* Vol.22(158), pages 209–212.

Annexe A

Longueurs d'intervalle bootstrap- t dans le contexte d'échantillonnage

A.1. Longueurs d'intervalles pour la méthode bootstrap- t

Cette section explique la différence importante visible entre la longueur moyenne des intervalles bootstrap- t sous la méthode par les poids lorsqu'un bootstrap à la Agresti-Coull est utilisé en comparaison avec ce même intervalle mais où on utilise la méthode par pseudo-population avec un bootstrap à la Agresti-Coull. En effet, nous observons des intervalles beaucoup plus longs dans les extrémités lorsqu'on utilise la méthode par les poids à la place de la méthode par pseudo-population.

Dans le cas de la méthode par les poids avec l'utilisation d'un bootstrap à la Agresti-Coull, rappelons qu'on tire avec remise un échantillon bootstrap s^* de taille $n' = n_s - 1$ à partir de l'échantillon s^+ . Soit la variable m_i^* , le nombre de fois que l'élément i est choisi dans l'échantillon bootstrap, le terme a_{i,s^+}^* est défini comme

$$a_{i,s^+}^* = 1 + C_{s^+} \left(\frac{n_{s^+} m_i^*}{n_s - 1} - 1 \right)$$

avec

$$C_{s^+} = \sqrt{1 - f_s}$$

de telle sorte que l'estimateur

$$\tilde{p}^* = \frac{1}{n_{s^+}} \sum_{i \in s^+} a_{i,s^+}^* Z_i.$$

Soit $\tilde{V}(\hat{p}) = (n_s - 1)^{-1}(1 - f_s)\tilde{p}(1 - \tilde{p})$, rappelons qu'avec l'utilisation de la méthode par les poids en utilisant un bootstrap à la Agresti-Coull que l'intervalle bootstrap- t prend la forme

$$\left[\hat{p} - K_{n_s; *}^{-1}(1 - \alpha/2; W, s^+) \sqrt{\tilde{V}(\hat{p})}, \hat{p} - K_{n_s; *}^{-1}(\alpha/2; W, s^+) \sqrt{\tilde{V}(\hat{p})} \right] \quad (\text{A.1.1})$$

où

$$K_{n_s; *}(x; W, s^+) = \text{Prob}^*(K_W^* \leq x)$$

avec $K_W^* = (\tilde{p}^* - \tilde{p})/\sqrt{\tilde{V}^*(\tilde{p}^*)}$ où $\tilde{V}^*(\tilde{p}^*) = (n_s - 1)^{-1}(1 - f_s)\tilde{p}^*(1 - \tilde{p}^*)$.

Comme nous l'avons mentionné précédemment, nous observons le comportement distinctif de la longueur moyenne d'intervalle de l'intervalle bootstrap- t lorsque \hat{p} est près de 0 ou 1. Lorsque $\hat{p} = 0$, la variable aléatoire K_W^* va prendre sa valeur la plus petite lorsque l'estimateur bootstrap \tilde{p}^* va lui aussi prendre sa valeur minimale. Cela se produira lorsque les deux succès de l'échantillon s^+ ne sont pas choisis dans l'échantillon bootstrap s^* . C'est donc dire que $m_i^* = 0$ pour les deux Z_i égaux à 1. Dans ce cas, nous avons que :

$$\begin{aligned} \tilde{p}^* &= \frac{1}{n+4} \sum_{i \in s^+} a_{i,s^+}^* Z_i \\ &= \frac{1}{n+4} \sum_{i \in s^+} a_{i,s^+}^* \mathbf{1}(Z_i = 1) \\ &= \frac{1}{n+4} \sum_{i \in s^+} \left(1 + \sqrt{1 - f_s} \left(\frac{(n+4)m_i^*}{n-1} - 1 \right) \right) \mathbf{1}(Z_i = 1) \\ &= \frac{2}{n+4} (1 - \sqrt{1 - f_s}). \end{aligned}$$

Nous retrouvons cette situation avec probabilité :

$$\left(1 - \frac{2}{n+4} \right)^{n-1} \cong \exp(-2) = 0,135.$$

Dans le cas où nous sommes dans la situation où $n = 40$ et $f = 0,05$, $\tilde{p}^* = 0,0012$ avec probabilité 0,163 et donc l'estimateur de la variance de l'estimateur bootstrap $\tilde{V}^*(\tilde{p}^*) = 2,8 \times 10^{-5}$. La plus petite valeur de K_W^* est donc :

$$K_W^* = \frac{\tilde{p}^* - \tilde{p}}{\sqrt{\tilde{V}^*(\tilde{p}^*)}} = -8,3721.$$

Comme nous avons cette valeur de K_W^* avec une probabilité 0,163 et que nous désirons, pour la borne supérieure de l'intervalle bootstrap- t , le quantile 0,025 de la distribution de $K_{n_s}^*$, c'est donc dire que ce quantile prend exactement la valeur $-8,3721$ lorsqu'on se trouve dans la situation où $n = 40$ et $f = 0,05$.

Vous retrouverez au tableau A.1 les quantiles 2,5% de la distribution de $K_{n_s}^*$ pour les 3 autres situations de notre simulation en considérant les différentes tailles échantillonales et fractions de sondage lorsqu'on considère la méthode des poids avec un bootstrap à la Agresti-Coull.

	f = 0,05	f = 0,20
	$K_{n_s; *}^{-1}(0,025; W, s^+)$	$K_{n_s; *}^{-1}(0,025; W, s^+)$
n = 40	-8,3721	-4,1076
n = 200	-8,7790	-4,3011

Tableau A.1. Quantiles $\alpha/2$ de la distribution de $K_{n_s}^*$ utilisés dans la construction de la borne supérieure de l'intervalle bootstrap- t pour la méthode par les poids avec un bootstrap à la Agresti-Coull.

Faisons le même exercice avec la méthode par pseudo-population. Dans le cas de la méthode par pseudo-population, afin de créer la pseudo-population, nous répétons l'échantillon s^+ , $k^+ = \lfloor N/n_{s^+} \rfloor$ fois avant de compléter la pseudo-population en tirant sans remise $N - n_s^+ k^+$ unités dans s^+ . Soit $p_{s^+}^*$, la proportion de succès dans la pseudo-population créée et $\hat{p}_{s^+}^* = n_s^{-1} \sum_{i \in s^*} Z_i^*$, l'estimateur bootstrap, l'intervalle bootstrap- t est défini comme (A.1.1) mais où la fonction de répartition $K_{n_s; *}$ est plutôt définie comme

$$K_{n_s; *}(x; PP, s^+) = Prob^*(K_{PP}^* \leq x)$$

avec $K_{PP}^* = (\hat{p}_{s^+}^* - p_{s^+}^*) / \sqrt{\tilde{V}^*(\hat{p}_{s^+}^*)}$ et $\tilde{V}^*(\hat{p}_{s^+}^*) = (1 - f_s)(n_s - 1)^{-1} \tilde{p}_{s^+}^* (1 - \tilde{p}_{s^+}^*)$ et où

$$\tilde{p}_{s^+}^* = \frac{1}{n_{s^{*+}}} \sum_{i \in s^{*+}} Z_i^*.$$

Considérons encore une fois le cas où $\hat{p} = 0$, la variable aléatoire K_{PP}^* prendra sa valeur la plus petite lorsque tous les éléments de l'échantillon bootstrap, s^* , sont nuls. À titre de comparaison avec la méthode par les poids, considérons la situation où la taille échantillonnale est de 40 observations et la fraction de sondage égale à 5%. Dans ce cas, $\hat{p}_{s^+}^*$ devient nul et nous avons que $\tilde{p}_{s^+}^* = 2/44$. Ainsi, $\tilde{V}^*(\hat{p}_{s^+}^*) = 0,0010$. La valeur de la proportion de succès dans la pseudo-population, $p_{s^+}^*$, variera entre 3 valeurs, considérant les cas où les $N - n_{s^+} k^+$ unités à tirer parmi l'échantillon s^+ contiennent soit aucun succès, 1 succès ou 2 succès. Dans la situation qui nous concerne, c'est-à-dire où $n = 40$ et $f = 0,05$, la pseudo-population doit avoir une taille de 800 unités. Après avoir répété l'échantillon s^+ , il restera 8 unités à tirer parmi s^+ afin de compléter la pseudo-population. La valeur de $p_{s^+}^*$ varie entre 36/800 et 38/800.

La plus petite valeur pour la variable K_{PP}^* se produira lorsque la pseudo-population aura une proportion de succès de 38/800. En effet, dans ce cas, K_{PP}^* prendra la valeur de $-1,4611$. Grâce aux propriétés de la loi hypergéométrique, il est possible de calculer la probabilité d'obtenir cette valeur, qui est essentiellement la probabilité que l'on tire aucun succès dans notre échantillon bootstrap alors que la pseudo-population contient 38 succès. Cette probabilité est égale à 0,40%. La seconde plus petite valeur pour K_{PP}^* est de $-1,4226$ et elle prend cette valeur dans 4,36% des cas. Comme nous recherchons le quantile 2,5% de la distribution de $K_{n_s}^*$ dans la construction de la borne supérieure de l'intervalle bootstrap- t , celui-ci prendra exactement la valeur de $-1,4226$. Il est possible de voir les différents quantiles de niveau 2,5% de la distribution de $K_{n_s}^*$ selon les différentes situations de notre simulation en considérant les différentes tailles échantillonnales et fractions de sondage dans le cas de la méthode par pseudo-population à l'aide du tableau (A.2).

C'est donc dire que sous les mêmes conditions, soit avec $n = 40$ et $f = 0,05$, le quantile de niveau 2,5% de la distribution de $K_{n_s}^*$ sous la méthode par les poids est de $-8,3721$ tandis que

	f = 0,05	f = 0,20
	$K_{n_s; *}^{-1}(0,025; PP, s^+)$	$K_{n_s; *}^{-1}(0,025; PP, s^+)$
n = 40	-1,4226	-1,3408
n = 200	-1,3955	-1,4407

Tableau A.2. Quantiles $\alpha/2$ de la distribution de $K_{n_s}^*$ utilisés dans la construction de la borne supérieure de l'intervalle bootstrap- t pour la méthode par pseudo-population avec un bootstrap à la Agresti-Coull.

sous la méthode par pseudo-population il est de $-1,4226$. Ces deux quantiles sont ceux utilisés dans la construction de la borne supérieure de l'intervalle bootstrap- t pour chacune des deux méthodes. Comme ces deux valeurs sont très petites négativement, la borne supérieure de l'intervalle se voit être très grande positivement car nous soustrayons à l'estimateur $\hat{\rho}$, nul, ce quantile négatif. Comme le quantile en question est beaucoup plus petit sous la méthode des poids que sous la méthode par pseudo-population, la borne supérieure de l'intervalle bootstrap- t est de loin supérieure pour la méthode par les poids comparativement à la borne supérieure du même intervalle sous la méthode par pseudo-population. Ceci explique donc que la longueur de l'intervalle sous la méthode des poids soit plus grande que sous la méthode par pseudo-population, d'où la forme bizarre de la courbe de longueurs moyennes d'intervalle pour l'intervalle bootstrap- t sous la méthode des poids.

