Stockholms universitet

Konfidensintervall för binomiella proportioner

Markus Lidström

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2019:5
Matematisk statistik
Augusti 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2019:5**
http://www.math.su.se/matstat

Konfidensintervall för binomiella proportioner

Markus Lidström*

Augusti 2019

Sammanfattning

Konfidensintervall för binomiella proportioner tillhör grunderna i statistiskt arbete. Dock kan de skapas på flera olika sätt. Detta arbete undersöker sex olika sätt att göra intervallen under vissa omständigheter. Den nominella täckningsgraden jämförs mot den faktiska täckningsgraden och även mot olika typer av intervall. Det framkommer att Wilson-Score, Agresti-Coull och Jeffreys liksvansade intervall är de som lämpar sig bäst för de fall som undersökts i detta arbete.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: markus.lidstrom@gmail.com. Handledare: Ola Hössjer.

Förord

Jag vill tacka min handledare Ola G H Hössjer för vägledning i detta kandidatarbete. Stödet från min familj har varit ovärderligt.

Innehåll

1 Inledning **4**
1.1 Bakgrund 4
1.2 Syfte 4

2 Metod **5**
2.1 Problemformulering 5
2.2 Teori 5
2.2.1 Konfidensintervall 5
2.2.2 Täckningsgraden för konfidensintervall 7
2.2.3 Mått för approximationens noggrannhet 7

3 Genomgång av konfidensintervall **8**
3.1 Frekventistiska Intervall 8
3.1.1 Wald-intervall 8
3.1.2 Wilson-Score Interval 11
3.1.3 Agresti-Coull Interval 15
3.1.4 Arcsin Transform 18
3.2 Bayesianskt Interval 21
3.2.1 Jeffreys Highest Posterior Density Interval 21
3.2.2 Jeffreys liksvansade intervall 22
3.3 Utvärdering av täckningsgraden 26

4 Undersökande av normalitetsantagande under $\mathcal{H}_0$ **27**
4.1 Samband mellan faktisk täckningsgrad och $\mathcal{D}_n$ 27
4.2 Numerisk beräkning av $\mathcal{D}_n$ 28
4.3 Asymptotisk teori för $\mathcal{D}_n$ 29

5 Diskussion **31**

6 Referenslista **33**

1 Inledning

1.1 Bakgrund

Den binomiala fördelningen är mycket använd och spridd modell inom alla olika forskningsområden och olika industriella applikationer, till exempel då man vill skatta proportionen av människor med en viss sjukdom, proportionen av produkter som går sönder efter 1 år eller proportionen av X-åriga människor som dör innan nästa år. Det är en modell där man ofta kan vara mycket säker på att själva modellvalet stämmer men denna säkerhet i modellens riktighet kommer på bekostnad av säkerheten för den sannolikhet (eller proportion) π som man vill skatta, se exempelvis Thulin (2012). Det är mycket viktigt att konfidensintervallen är så nära den nominella täckningsgraden som möjligt, men så är sällan fallet. Detta reser frågan om det finns konfidensintervall som är bättre än andra och kanske under olika förutsättningar. Wald-intervallet är det mest spridda och mest använda konfidensintervallet men det besitter egenskaper som gör det olämplig under flera situationer, exempelvis om π är mycket litet eller mycket stort. Det är dock det mest kända intervallet för binomiella proportioner till stor del för att det ofta är det enda som lärs ut och till en viss del så är det dess simplicitet som gör det vida använt i situationer där det verkligen inte ska användas. Flera av intervallen är frekventistiska och konstruerade utifrån inversionen av test-statistikor som bygger på ett antagande av normalapproximation. En viktig fråga är då om dessa olika test-statistikor uppnår normalitet olika snabbt då datamaterialets storlek ökar.

1.2 Syfte

Denna uppsats ämnar till att undersöka flera olika konfidensintervall, hur deras egenskaper ser ut, hur snabbt de test-statistikor som används i konstruktionen av konfidensintervallen konvergerar mot den fördelning de avser att approximera. I denna studie kommer vi undersöka situationer då antalet obserbationer $n = 10, 25, 50$ och ofta så får den sökta sannolikheten $\pi \in (0, 1)$ variera fritt. Givet dessa förutsättningar är ambitionen att utvärdera flera olika konfidensintervall emot varandra och se om det finns något konfidensintervall man kan rekommendera generellt för användning.

2 Metod

2.1 Problemformulering

Den binomiella fördelningen uppkommer när vi tittar på summor av n oberoende och identiskt fördelade Bernoullivariabler, så varje n utfall av en binomiell variabel kan ses som ett stickprov från dessa Bernoullivariabler. Den binomiella fördelningens sannolikhetsfunktion är

$$P_\pi(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \tag{1}$$

för $x = 1, \ldots, n$, och genom standardberäkningar får vi att väntevärdet och variansen för X ges av $E[X] = n\pi$ och $Var(X) = n\pi(1-\pi)$. För att skatta parametern π används ofta Maximum Likelihood-skattningen där vi maximerar Likelihoodfunktionen och får att $\hat{\pi}_{MLE} = x/n$. Problemet är sedan att i många fall så vill man inte bara ha en punktskattning utan man vill ha ett intervall eller region av möjliga parametervärden som man är säker, till en viss grad av säkerhet, att det sanna värdet av π ligger i. Detta är huvudproblemet som kommer uppta en stor del av denna uppsats.

2.2 Teori

I denna sektion gås all relevant teori igenom, med vilka mått vi ska utvärdera konfidensintervallen och hur vi räknar ut täckningsgraden.

2.2.1 Konfidensintervall

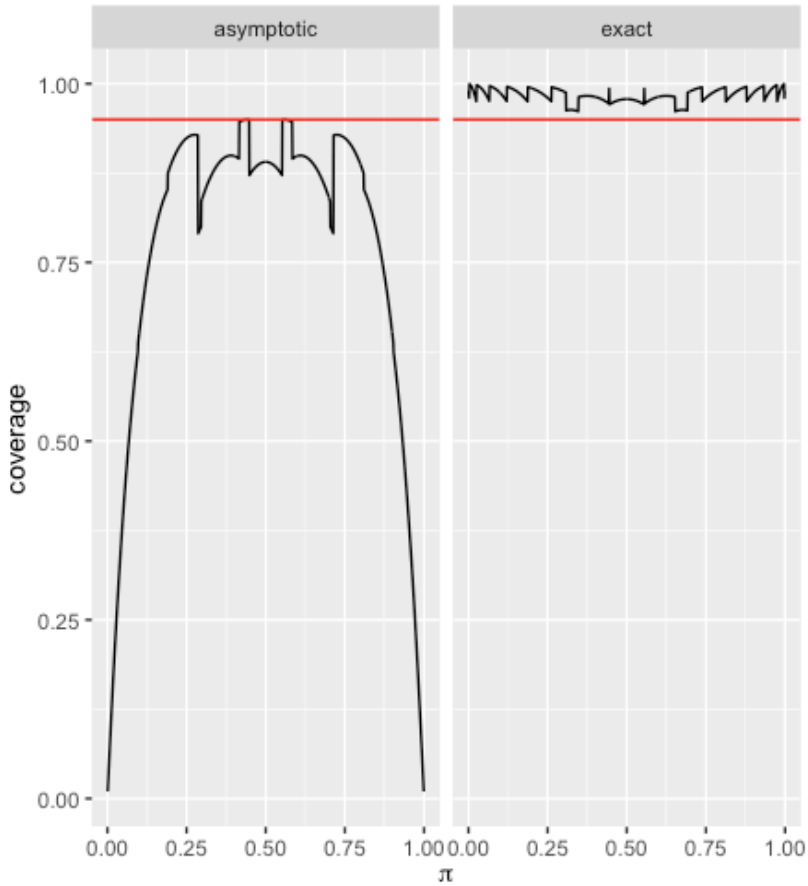
Vi börjar med att definiera vad ett konfidensintervall är. För ett fixt $\alpha \in (0,1)$, så är ett exakt $(1-\alpha) \cdot 100\%$ konfidensintervall för parametern θ definierat av två test-statistikor $L(X_{1:n})$ och $U(X_{1:n})$ baserat på ett stickprov $X_{1:n} = (X_1, \ldots, X_n)$, som uppfyller

$$\inf_{\theta \in \Theta} P_\theta(L(X_{1:n}) < \theta < U(X_{1:n})) \geq (1-\alpha).$$

Ett konfidensintervall kallas approximativt om

$$P_\theta(L(X_{1:n}) < \theta < U(X_{1:n})) \approx 1-\alpha$$

Vid en första anblick så kan det te sig konstigt att vi definierar exakthet som infimum över parameterrummet Θ ska vara minst lika stor som den nominella konfidensgraden $1-\alpha$. Detta är dock på grund av osäkerheten som gäller konfidensintervall för diskreta fördelningar. I diagrammen nedan så visas skillnaden i täckningsgraden för ett exakt och ett approximativt konfidensintervall där den nominella konfidensgraden är 95% och $n = 10$.



Figur 1: Nominell (röd) and exakt täckningsgrad (svart) för ett asymptotiskt (vänster) och exakt (höger) konfidensintervall då $n = 10$ och då π varierar mellan 0 och 1.

Vi ser i figur 1 en graf över de två olika typerna av konfidensintervall och att för det exakta konfidensintervallet så faller aldrig täckningsgraden under den nominella medan detta inte gäller för det approximativa konfidensintervallet.

I denna uppsats är $X_1, \ldots, X_n$ oberoende och Bernoullifördelade variabler med sannolikhetsfunktion $\pi = P_\theta(X_i = 1) = 1 - P_\theta(X_i = 0)$, och den parameter vi vill uppskatta är $\theta = \pi$. Vidare har $X = \sum_{i=1}^n X_i$ en binomialfördelning enligt ekvation (1). Man kan visa att X är en tillräcklig statistika för att skatta θ, och därför kommer vi fortsättningsvis att ersätta stickprovet $X_{1:n}$ med X.

2.2.2 Täckningsgraden för konfidensintervall

Låt $\mathcal{X} = \{0, 1, \ldots, n\}$ vara vår utfallsrymd och $CI_\alpha(X) = (L_\alpha(X), U_\alpha(X))$ vara vårt konfidensintervall, av nominell konfidensgrad $1 - \alpha$, för vår parameter θ. Då är den exakta eller faktiska täckningsgraden

$$\beta_n(\theta) = P_\theta(\theta \in CI_\alpha(X)) = \sum_{x \in \mathcal{X}} \mathbb{1}_{CI_\alpha(x)}(\theta) \cdot P_\theta(X = x),$$

där $\mathbb{1}_{CI_\alpha(x)}(\theta)$ är indikatorfunktionen för om θ täcks av konfidensintervallet $CI_\alpha(x)$.

Vi kan skriva om definitionen av $\beta_n(\theta)$ som

$$\beta_n(\theta) = \sum_{x \in S_\theta} P_\theta(X = x)$$

där $S_\theta = \{x \in \mathcal{X} | \mathbb{1}_{I_\alpha(x)}(\theta) = 1\}$. Om vi till exempel väljer två parametervärden $\theta_1 < \theta_2$ så att $S_{\theta_1} = S_{\theta_2}$ då kommer vi i allmänhet att ha att

$$P_{\theta_1}(\theta_1 \in CI_\alpha(X)) \neq P_{\theta_2}(\theta_2 \in CI_\alpha(X))$$

när $P_{\theta_1}(X = x) \neq P_{\theta_2}(X = x)$. Dock, om exempelvis $\theta_1 \to \theta_2$ och $S_{\theta_2} \setminus S_{\theta_1} = \{x\}$, så kommer vi att få ett diskontinuerligt hopp i $\beta_n(\theta)$ i punkten θ_2 genom den tillkomna punktmassan $P_\theta(X = x)$.

2.2.3 Mått för approximationens noggrannhet

Många konfidensintervall konstrueras med hjälp av test-statistikor vars fördelning är asymptotiskt lik en annan fördelning, oftast normalfördelningen. Därför är det av intresse att mäta hur långt ifrån fördelningen för vår teststatistika är från den riktiga fördelningen. Detta kommer vi mäta med hjälp

av Kolmogorovdistansen som definieras enligt följande. Givet två kumulativa sannolikhetsfördelningar F och G så är Kolmogorovdistansen

$$\mathcal{D} = \sup_{x \in R} |G(x) - F(x)|$$

och i vårt fall när vi undersöker hur bra teststatistikor är normalapproximerade så beror detta så klart på n. Så därför kommer vi att studera $\mathcal{D}_n = \sup_{x \in R} |G(x) - F_n(x)|$, där G är fördelningsfunktionen för en standard normalfördelning med väntevärde 0 och varians 1, medan F_n är fördelningsfunktionen för teststorheten när stickprovet har storlek n.

3 Genomgång av konfidensintervall

3.1 Frekventistiska Intervall

Vi kommer basera våra frekventistiska konfidensintervall på en teststorhet $T(x, \theta_0)$ för att testa $\mathcal{H}_0 : \theta = \theta_0$ mot $\mathcal{H}_1 : \theta \neq \theta_0$. Fördelningen för T kommer sedan approximeras av en kontinuerlig fördelning G under $\mathcal{H}_0$. Det ger ett tvåsidigt konfidensintervall

$$CI(x) = \{\theta | T(x, \theta) \in (G^{-1}(\alpha/2), G^{-1}(1 - \alpha/2)\}.$$

Om exempelvis G är fördelningsfunktionen för en standard normalfördelning får vi

$$CI(x) = \{\theta; T(x, \theta) \in (-z_{\alpha/2}, z_{\alpha/2})\},$$

där z_p är p-kvantilen för en standard normalfördelning.

3.1.1 Wald-intervall

Detta är det mest använda konfidensintervallet då det är lätt att applicera och inte kräver någon avancerad kunskap. Det konstrueras genom att invertera test-statistikan

$$T = \frac{\hat{\theta} - \theta}{se(\hat{\theta})} \approx N(0, 1)$$

där $\hat{\theta}$ är Maximum Likelihood-estimatorn av parametern av intresse och $N(0, 1)$ betecknar en standard normalfördelning. Så för att konstruera ett intervall med konfidensgrad $1-\alpha$ av testen så inverteras $z_{1-\alpha/2} < \hat{\theta} - \theta / se(\hat{\theta}) < z_{\alpha/2}$. När $\theta = \pi$ svarar mot sannolikheten i en binomialfördelning så gäller

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.7358667	-0.1641333
0.90	25	0.8217163	-0.0782837
0.90	50	0.8560693	-0.0439307
0.95	10	0.7701116	-0.1798884
0.95	25	0.8636894	-0.0863106
0.95	50	0.9013773	-0.0486227
0.99	10	0.8012754	-0.1887246
0.99	25	0.9018591	-0.0881409
0.99	50	0.9413603	-0.0486397

Tabell 1: Nominell täckningsgrad $1-\alpha$, genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett Waldbaserat konfidensintervall baserat på olika stickprovsstorlekar n.

att medelfelet $se(\hat{\pi}) = \hat{\pi}(1-\hat{\pi})/n$. Efter några algebraiska manipulationer så får vi konfidensintervallet:

$$CI_W = \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

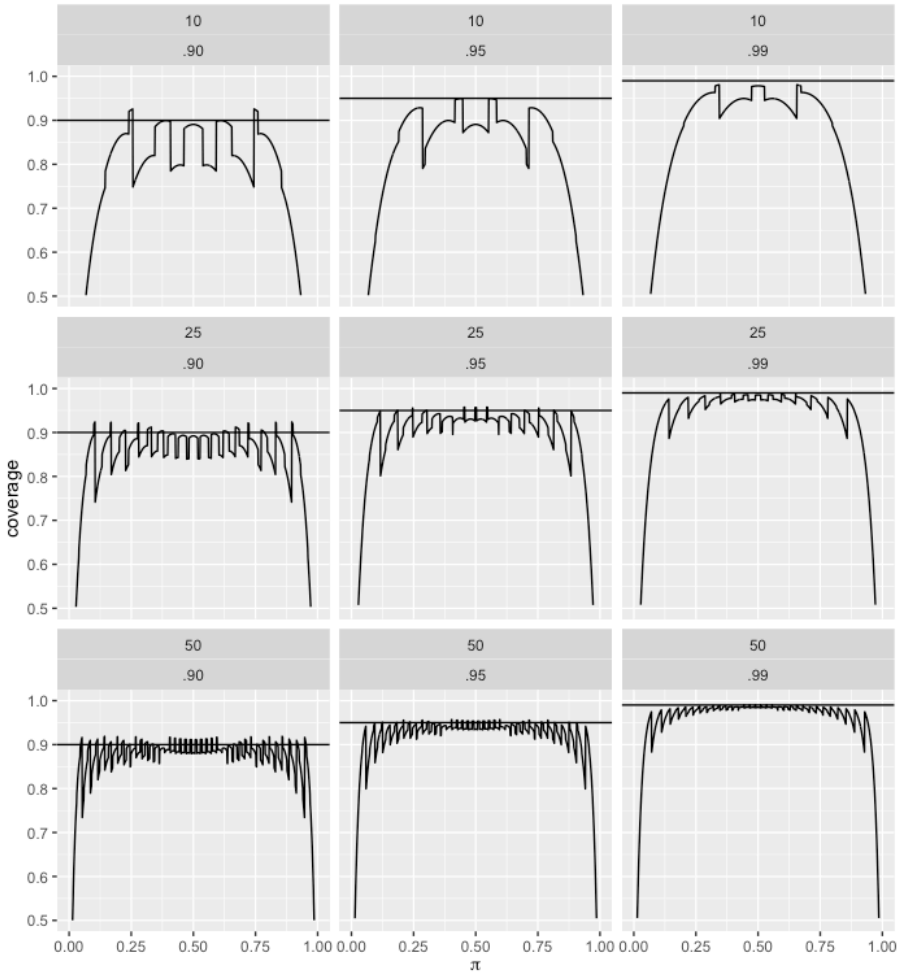
där $\hat{\pi} \pm d$ är ett förkortat skrivsätt för intervallet $(\hat{\pi} - d, \hat{\pi} + d)$. Man kan se att $\hat{\pi}(1-\hat{\pi}) \to 0$ när $\hat{\pi} \to 0$ och $\hat{\pi}(1-\hat{\pi}) \to 0$ när $\hat{\pi} \to 1$ och därför kollapsar konfidensintervallet till en punkt i dess två extrema fall av $\hat{\pi}$.

Vi ser i figur 2 att den faktiska täckningsgraden för Waldintervallet går mot 0 då π är nära 0 eller 1. Vi ser i tabell 1 att den genomsnittliga faktiska täckningsgraden ligger under den nominella med mellan 4 till 18 procentenheter i de olika fallen. Vi ser även att avståndet till nominella täckningsgraden minskar då konfidensgraden ökar.

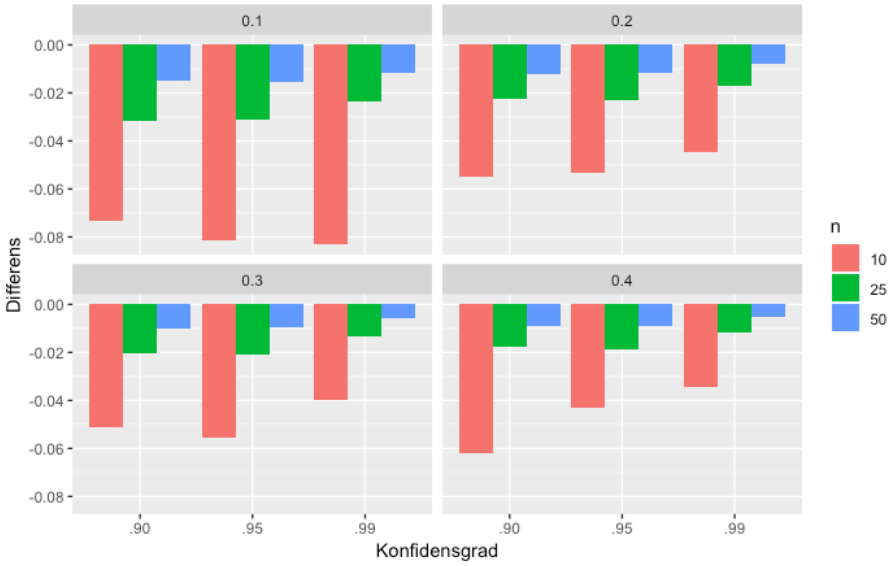
I figur 3 ser vi genomsnittliga faktiska täckningsgraden

$$\bar{\beta}_n(b) = \frac{1}{1-2b}\int_b^{1-b} \beta_n(\pi)d\pi, \tag{2}$$

när vi bara låter $b \leq \pi \leq 1-b$ där b är en vald konstant. Vi ser att Wald intervallet når det nominella konfidensnivån när $b = 0.1, 0.2, 0.3, 0.4$. Vi ser att skillnaden mellan den nominella täckningsgraden och den genomsnittliga faktiska täckningsgraden generellt sett minskar ju större b är och att genomsnittliga faktiska täckningsgraden för Waldintervallet är konsekvent under den nominella täckningsgraden. Vi ser samma beteende för $b = 0.1$ och $n = 10$ som det var när när vi lät $\pi \in (0, 1)$ i tabell 1, att differensen ökar ju högre konfidensgraden är.



Figur 2: Nominell och faktiskt täckningsgrad för Waldintervallet som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



Figur 3: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Waldintervallet som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1-\alpha$ och $b = 0.1, 0.2, 0.3, 0.4$.

3.1.2 Wilson-Score Interval

Detta konfidensintervall föreslogs av Edwin B. Wilson och är ett Score-test baserat intervall som använder den förväntade Fisherinformationen. Man konstruerar konfidensintervallet genom att invertera det nämnda testet och det leder till att lösa en kvadratiskt ekvation och kommer inte visas här, se Bové (2013). Låt $L(\pi|x)$ vara log likelihood-funktionen för en binomialfördelning, definierad som logaritmen av (1). Scorefunktionen är

$$U(\pi) = \frac{\partial\ L(\pi|x)}{\partial \pi} = \frac{x - n\pi}{\pi(1-\pi)}$$

och den väntade Fisherinformationen är

$$\mathcal{J}_n(\pi) = -E\left[\frac{\partial^2 L(\pi|x)}{\partial \pi^2}\right] = \frac{n}{\pi(1-\pi)}.$$

Scoretestet baseras på teststorheten $T = (\hat{\pi} - \pi)/\sqrt{\pi(1-\pi)/n} \approx N(0,1)$ och genom att invertera enligt ovan nämnda referens så får vi:

$$CI_{WS} = \frac{\hat{\pi} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.9106017	0.0106017
0.90	25	0.9060895	0.0060895
0.90	50	0.9035492	0.0035492
0.95	10	0.9542671	0.0042671
0.95	25	0.9523885	0.0023885
0.95	50	0.9517583	0.0017583
0.99	10	0.9875425	-0.0024575
0.99	25	0.9887836	-0.0012164
0.99	50	0.9891493	-0.0008507

Tabell 2: Nominell täckningsgrad $1-\alpha$, faktisk genomsnittlig täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett Wilson-Score baserat konfidensintervall baserat på olika stickprovsstorlekar n.

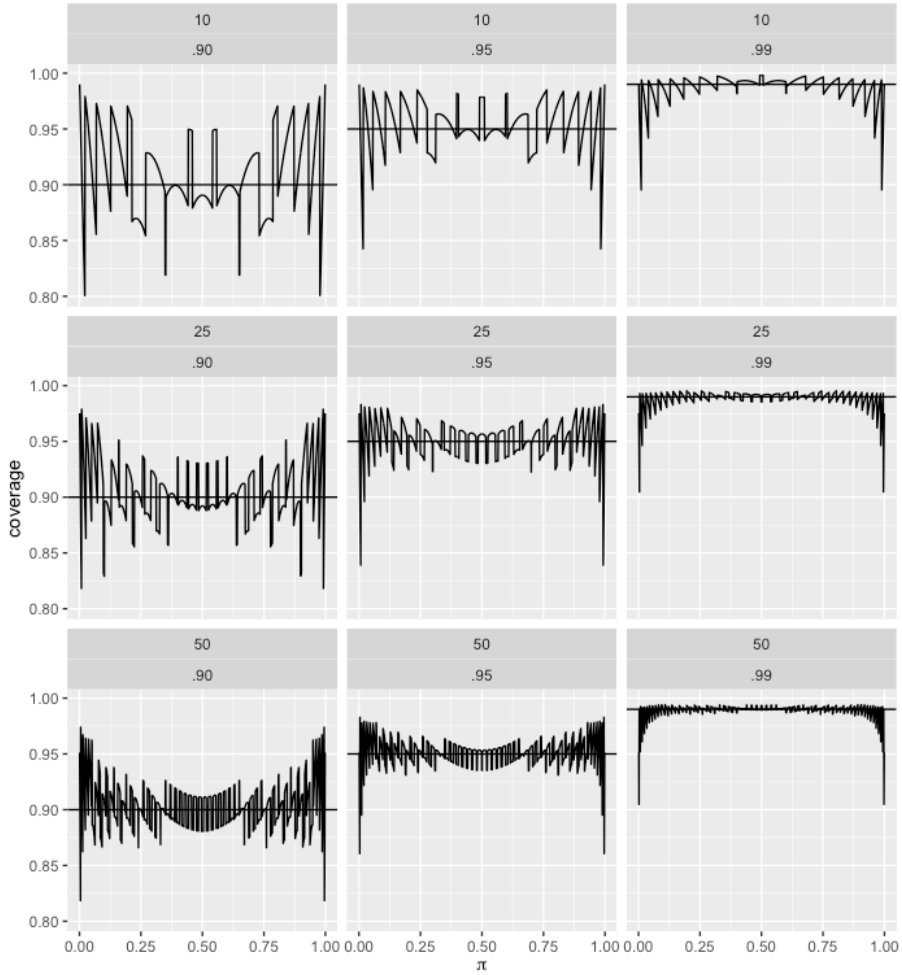
Genom samma argument som för Waldintervallet så kan man se att detta interval inte kollapsar till en punkt när $\hat{\pi}$ är nära 0 eller 1 och detta är för att mittpunktenpå detta intervall är viktat mellan Maximum Likelihood-skattningen och 1/2. Detta kan ses om vi skriver om mittpunkten för intervallet som

$$\hat{\pi}\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}$$

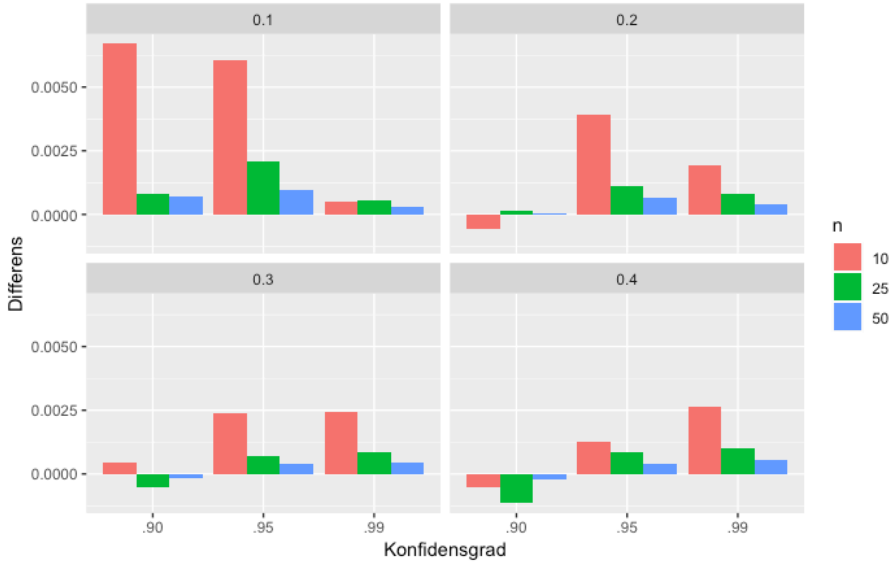
och vi kan se att $\hat{\pi}$ ges mer vikt när n ökar se Agresti och Coull (1998).

I figur 4 ser vi att Wilson-Score intervallet har en faktisk täckningsgrad som inte sjunker drastiskt när vi är nära 0 eller 1 dock blir de diskontinuerliga hoppen större. I tabell 2 visas den genomsnittliga faktiska täckningsgraden för Wilson-Score intervallet. Vi kan se en förbättring jämfört med Waldintervallet då den faktiska täckningsgraden inte är nästintill under den nominella som det var för Waldintervallet. Vi ser också att vi får större osäkerhet i täckningsgraden då π är litet eller väldigt stort men det degenererar inte, som för Waldintervallet. Vi ser också att den genomsnittliga täckningsgraden är väldigt nära den nominella.

I figur 5 ser vi den genomsnittliga täckningsgraden när vi infört samma restriktioner för π som förut och beräknat dem enligt ekvation (2). Vi ser att för t.ex. för $b = 0.4$ och konfidensgrad 0.90 så blir den faktiska täckningsgraden inte nödvändigtvis bättre när n går från 10 till 25, vilket går emot det vanliga tankesättet att ju större stickprov desto bättre täckningsgrad. Vi ser också att skillnaden mellan den genomsnittliga faktiska och nominella täckningsgraden ibland ökar och ibland minskar när konfidensgraden ökar



Figur 4: Nominell och faktisk täckningsgrad för Wilson-Score intervallet som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



Figur 5: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Wilson-Score intervallet som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1-\alpha$ och $b = 0.1$, 0.2, 0.3, 0.4.

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.9263937	0.0263937
0.90	25	0.9173824	0.0173824
0.90	50	0.9117997	0.0117997
0.95	10	0.9646403	0.0146403
0.95	25	0.9610300	0.0110300
0.95	50	0.9580296	0.0080296
0.99	10	0.9916967	0.0016967
0.99	25	0.9921540	0.0021540
0.99	50	0.9919113	0.0019113

Tabell 3: Nominell täckningsgrad $1-\alpha$, genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett Agresti-Coull konfidensintervall baserat på olika stickprovsstorlekar n.

för $b = 0.2, 0.3$ och 0.4. Vi ser även att i de flesta fallen så är Wilson-Score intervallet konservativt.

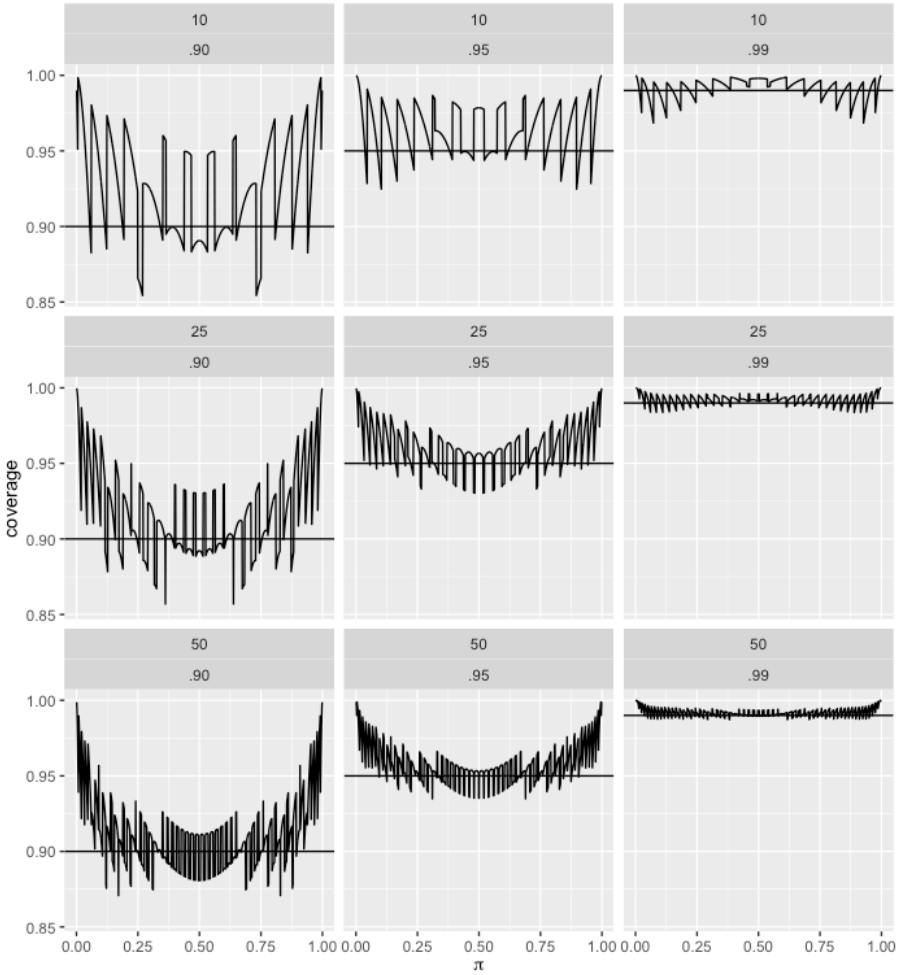
3.1.3 Agresti-Coull Interval

Agresti-Coull intervallet föreslogs 1998 (Agresti and Coull, 1998). Det är baserat på mittpunkten av Wilson Score-intervallet och att göra en normalapproximation runt den punkten. Med införande av beteckningar $\tilde{n} = n + z^2_{\alpha/2}$ och $\tilde{\pi} = \frac{1}{\tilde{n}}(X + \frac{z^2_{\alpha/2}}{2})$ så kan vi skriva detta intervall enligt

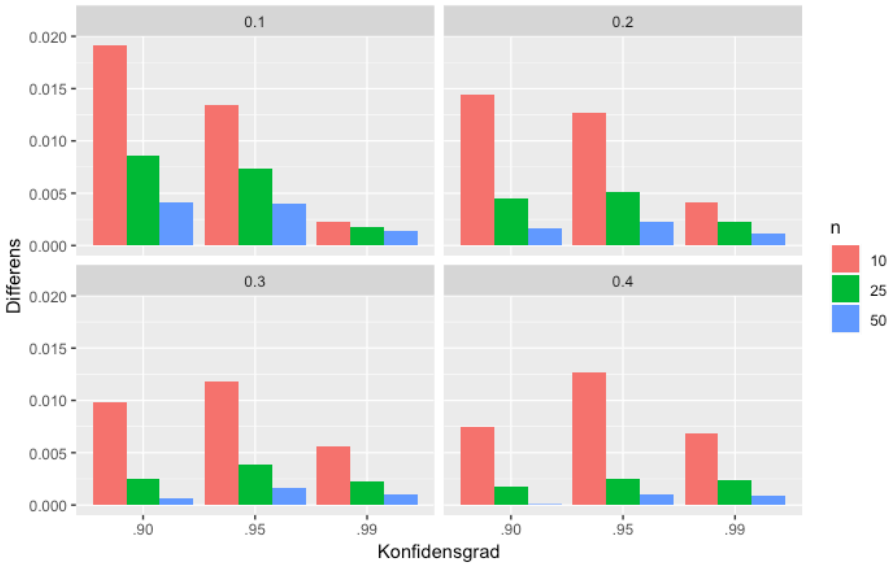
$$CI_{AC} = \tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}}$$

I figur 6 ser att Agresti-Coull intervallet är väldigt konservativt då täckningsgraden går mot 1 när π nära 0 eller 1 vilket kan vara en positiv egenskap eller en negativ egenskap, beroende på hur man ser det. Vi ser även att vi inte har lika stor ökning av storleken av de diskontinuerliga hoppen nära 0 och 1 som vi ser för Wilson-Scoreintervallet i figur 3. I tabell 4 ser vi den faktiska genomsnittliga täckningsgraden ligger lite över den nominella, med cirka 1-2 procentenheter men detta försvinner nästan helt för konfidensgraden 0.99

I figur 7 ser vi den genomsnittliga faktiska täckningsgraden när vi återigen inför restriktioner på π enligt ekvation (2). Vi ser att Agresti-Coull intervallet är genomgående konservativt och att vi inte får någon stor förbättring i täckningsgraden.



Figur 6: Nominell och faktisk täckningsgrad för Agresti-Coull intervallet som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



Figur 7: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Agresti-Coull intervallet som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1 - \alpha$ och $b = 0.1$, 0.2, 0.3, 0.4.

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.7504019	-0.1495981
0.90	25	0.8350320	-0.0649680
0.90	50	0.8658353	-0.0341647
0.95	10	0.7860895	-0.1639105
0.95	25	0.8790715	-0.0709285
0.95	50	0.9130551	-0.0369449
0.99	10	0.8123833	-0.1776167
0.99	25	0.9142173	-0.0757827
0.99	50	0.9512010	-0.0387990

Tabell 4: Nominell täckningsgrad $1-\alpha$, genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett arcsin-transformbaserat konfidensintervall baserat på olika stickprovsstorlekar n.

3.1.4 Arcsin Transform

Transformationen $h(\hat{\pi}) = \arcsin\left(\sqrt{\hat{\pi}}\right)$ är en variansstabliserande transformation. Låt $\hat{\pi}$ vara ML-skattningen av π och om vi gör en första ordningens Taylorutveckling av $h(\hat{\pi})$ kring punkten π så ser vi att

$$Var(h(\hat{\pi})) \approx h'(\pi)^2 Var(\hat{\pi}) = \frac{1}{4n},$$

eftersom

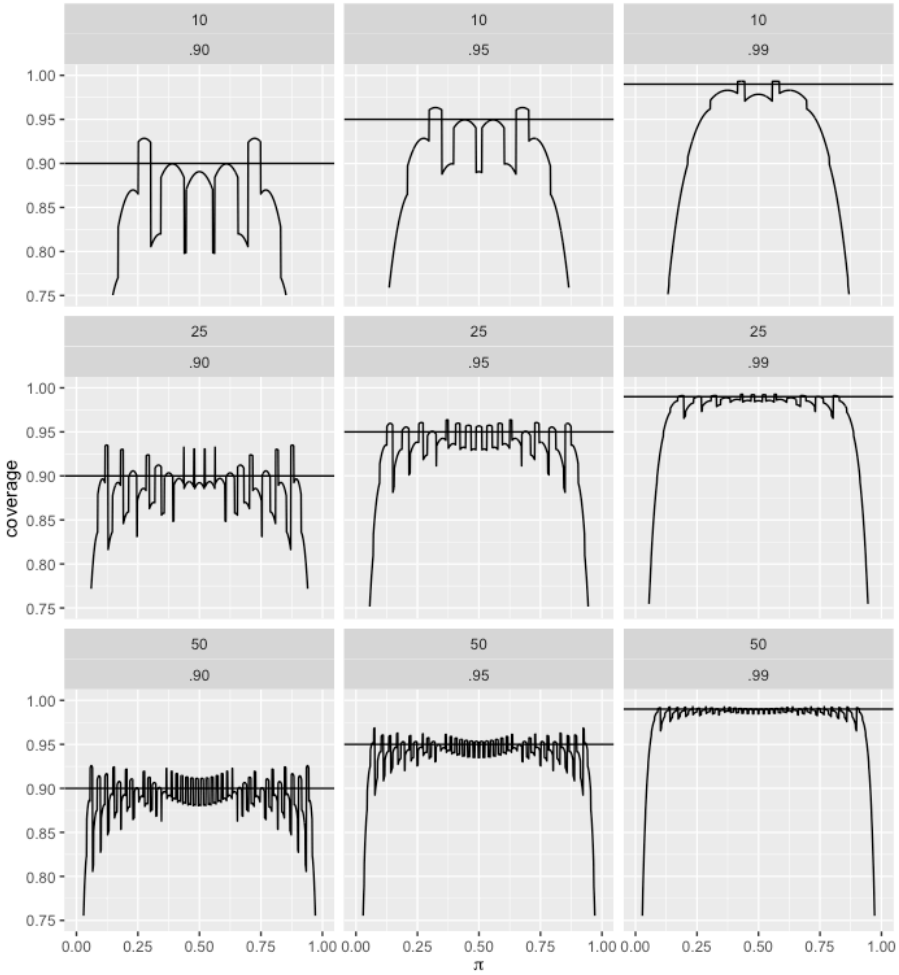
$$h'(\pi) = \frac{1}{2\sqrt{\pi(1-\pi)}}.$$

Då får vi ett konfidensinterval av formen

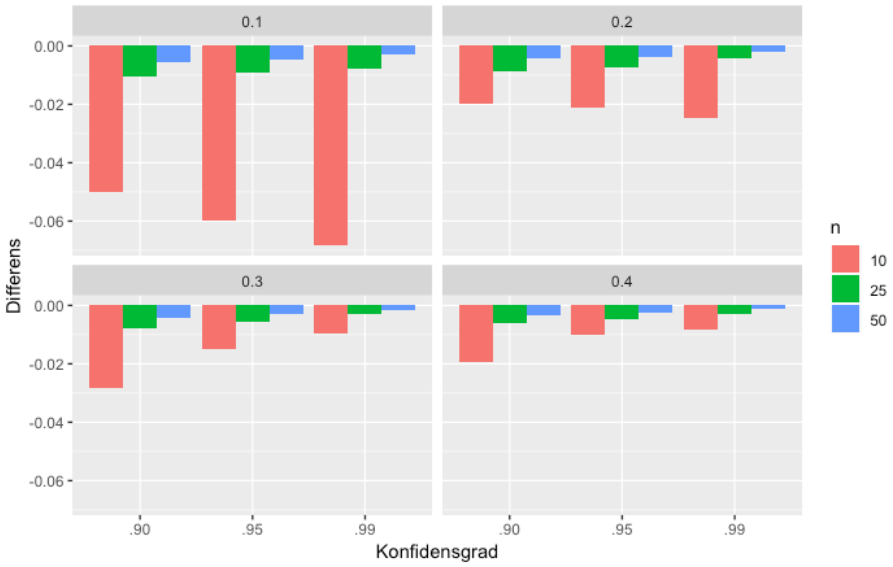
$$CI_{AT} = \sin^2\left(\arcsin\left(\sqrt{\hat{\pi}}\right) - \frac{z_{\alpha/2}}{\sqrt{4n}}\right) < \pi < \sin^2\left(\arcsin\left(\sqrt{\hat{\pi}}\right) + \frac{z_{\alpha/2}}{\sqrt{4n}}\right).$$

I figur 8 ser vi att arcsin-transformen lider av samma problem som Waldintervallet då faktiskta täckningsgraden sjunker drastiskt när vi är nära 0 eller 1. I tabell 4 ser att den genomsnittliga faktiska täckningsgraden ligger mest under den nominella och att det ofta är ganska långt under, med ett avstånd på 3-17 procentenheter.

I figur 9 ser vi att den genomsnittliga faktiska täckningsgraden för arcsin-intervallet konsekvent ligger under den nominella täckningsgraden. När vi låter $b = 0.2, 0.3$ och 0.4 så är den genomsnittliga faktiska täckningsgraden faktiskt väldigt nära den nominella. När $b = 0.1$ och $n = 10$ så ser vi att skillnaden mellan genomsnittliga faktiska täckningsgraden och den nominella



Figur 8: Nominell och faktisk täckningsgrad för arcsin-transform intervallet som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



Figur 9: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Arcsin-transformintervallet som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1 - \alpha$ och $b = 0.1$, 0.2, 0.3, 0.4.

täckningsgraden ökar när konfidensgraden ökar, vilket var samma beteende som för Waldintervallet.

3.2 Bayesianskt Interval

Inom den Bayesianska statistiken så använder man kredibilitetsintervall istället för konfidensintervall då parametern har en sannolikhetsfördelning. Den Bayesianska metoden bygger på att använda information som man har sedan tidigare om parametern, aprioritätheten $f(\pi)$, och den nya informationen, i form av sannolikhetsfördelningen $f(x|\pi) = P_\pi(X = x)$ hos X från ekvation (1). Genom Bayes sats får vi aposterioritätheten

$$f(\pi|x) = \frac{f(x|\pi)f(\pi)}{\int f(x|\pi)f(\pi)d\pi}.$$

Från aposteriorifördelningen får vi kredibilitetsintervallet för parametern i fråga som $\int_a^b f(\pi|x)d\pi = 1 - \alpha$ där a, b är ändpunkterna på vårt kredibilitetsintervall och α den nominella kredibilitetsnivån. Aposteriorifördelning är beroende av aprioritfördelningen och därav måste den väljas med stor omsorg om vi inte har erhållit kunskap från tidigare undersökningar. Ett populärt val är att välja aprioritfördelningen för parametern π i binomialfördelningen som en $Beta(\alpha, \beta)$-fördelning med täthetsfunktion

$$f(\pi) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1 - \pi)^{\beta-1}.$$

Ett annat val är en så kallad Jeffreys prior som är proportionell mot roten ur det förväntade Fisherinformationen så i vårat fall $p(\pi) \propto \sqrt{\mathcal{J}(\pi)}$. Då vi vet att $\sqrt{\mathcal{J}(\pi)} = \pi^{-1/2}(1 - \pi)^{-1/2}$ så ser vi att detta är kärnan i den nämnda Beta fördelningen med parametrar $\alpha = 1/2$ och $\beta = 1/2$. Med Jeffreys prior som aprioritfördelning så ser vi att

$$f(\pi|x) \propto \pi^x(1 - \pi)^{n-x}\pi^{-1/2}(1 - \pi)^{-1/2} = \pi^{x-1/2}(1 - \pi)^{n-x-1/2}$$

så aposteriorifördelningen är $Beta(x - 1/2, n - x - 1/2)$ och därav är aprioritfördelningen och aposteriorifördelningen medlemmar av samma familj av fördelningar.

3.2.1 Jeffreys Highest Posterior Density Interval

Ett HDP-interval CI är så att $f(\pi|x) \geq f(\bar{\pi}|x)$ för alla $\pi \in CI$ och $\bar{\pi} \notin CI$. Då Beta-fördelningen är unimodal så är detta det samma som

$$\int_a^b f(\pi|x)d\pi = 1 - \alpha$$

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.8682989	-0.0317011
0.90	25	0.8796037	-0.0203963
0.90	50	0.8868092	-0.0131908
0.95	10	0.9324250	-0.0175750
0.95	25	0.9380075	-0.0119925
0.95	50	0.9418577	-0.0081423
0.99	10	0.9861621	-0.0038379
0.99	25	0.9869483	-0.0030517
0.99	50	0.9877785	-0.0022215

Tabell 5: Nominell täckningsgrad $1-\alpha$, genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett HDP-baserat konfidensintervall baserat på olika stickprovsstorlekar n.

så att $b-a$ minimeras.

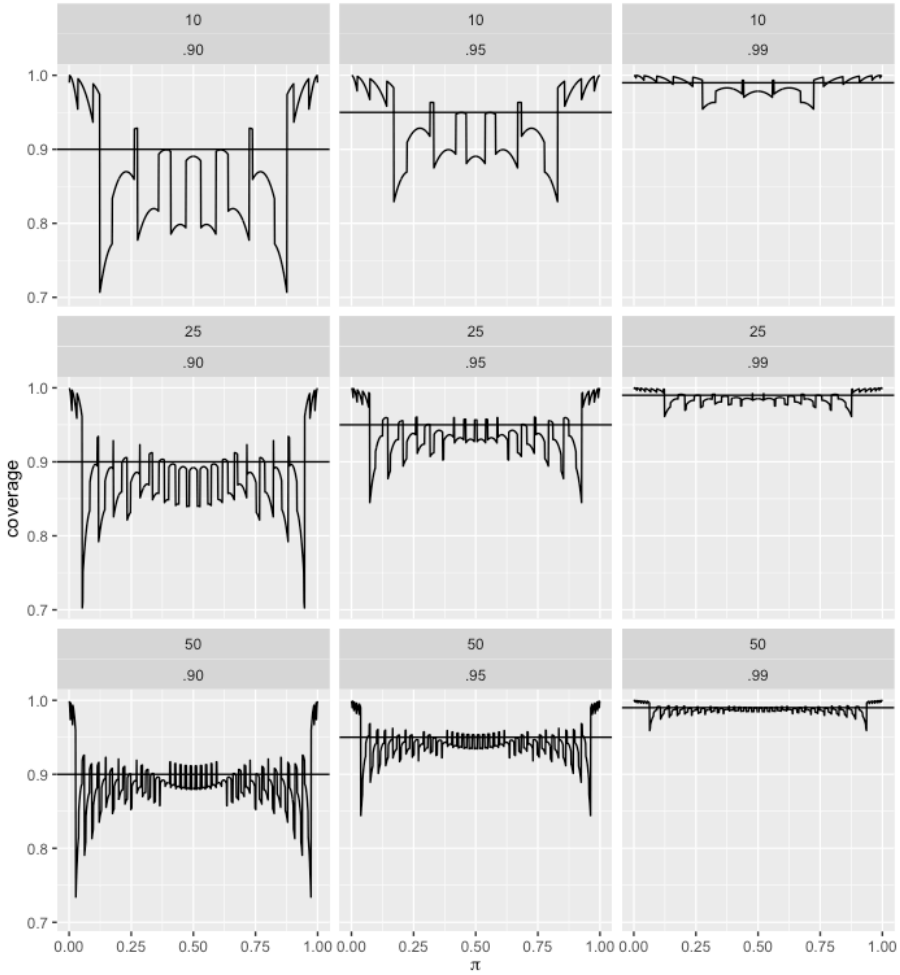
Vi ser i figur 10 att även om osäkerheten ökar i täckningsgraden då π nära 0 och 1 så är det inte lika dåligt som Wald-intervallet eller arcsin-transformen. Den är också konservativ, för täckningsgraden ligger nära 1 då π väldigt nära 0 och 1. I tabell 5 ser vi den genomsnittliga faktiska täckningsgraden och vi ser att när konfidensgraden är 0.90 eller 0.95 så ligger den genomsnittliga faktiska täckningsgraden under den nominella med några procentenheter men när konfidensgraden är 0.99 så är vi väldigt nära den nominella täckningsgraden.

I figur 11 ser vi att intervallet har ett konsekvent beteende i den mening att differensen minskar när vi går till högre konfidensgrad eller ökar stickprovsstorleken och låter b vara fixt. Vi ser även att den genomsnittliga faktiska täckningsgraden är längre ifrån den nominella i många fall när vi har färre restriktioner på π till exempel när $b = 0.1$, konfidensgrad 0.9 och $n = 10$ så är differensen ungefär -0.06 medan utan restriktioner så är differensen ungefär -0.03 då antingen stickprovsstorleken ökas till 25 eller konfiidensgraden till 0.95. .

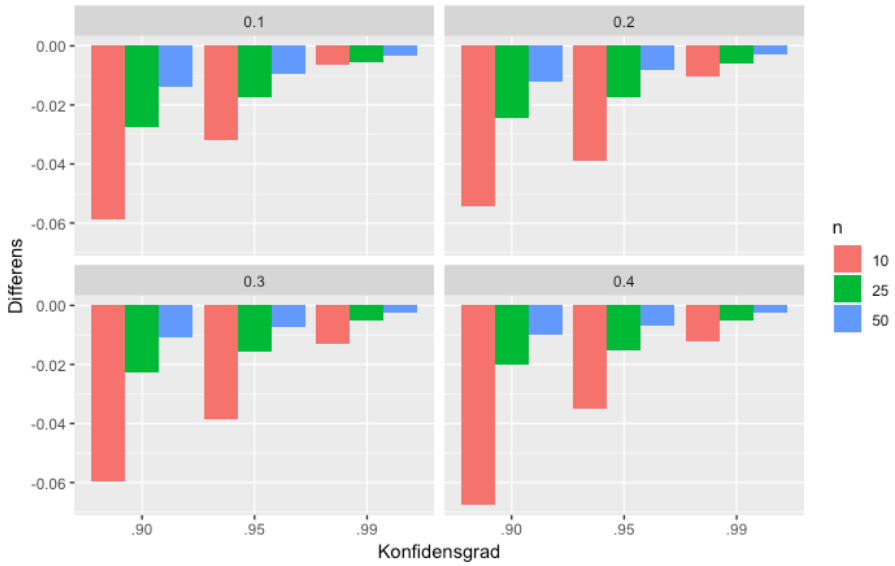
3.2.2 Jeffreys liksvansade intervall

För ett liksvansat intervall så använder vi $\alpha/2$ och $1-\alpha/2$-kvantilerna från $Beta(x+1/2, n-x+1/2)$ som våra nedre och övre gränser.

I figur 12 ser vi att täckningsgraden inte sjunker drastiskt när π är nära 0 eller 1. Den ligger ofta över den nominella täckningsgraden då och därav kan även detta intervall ses som lite konservativt. I tabell 6 ser vi att den



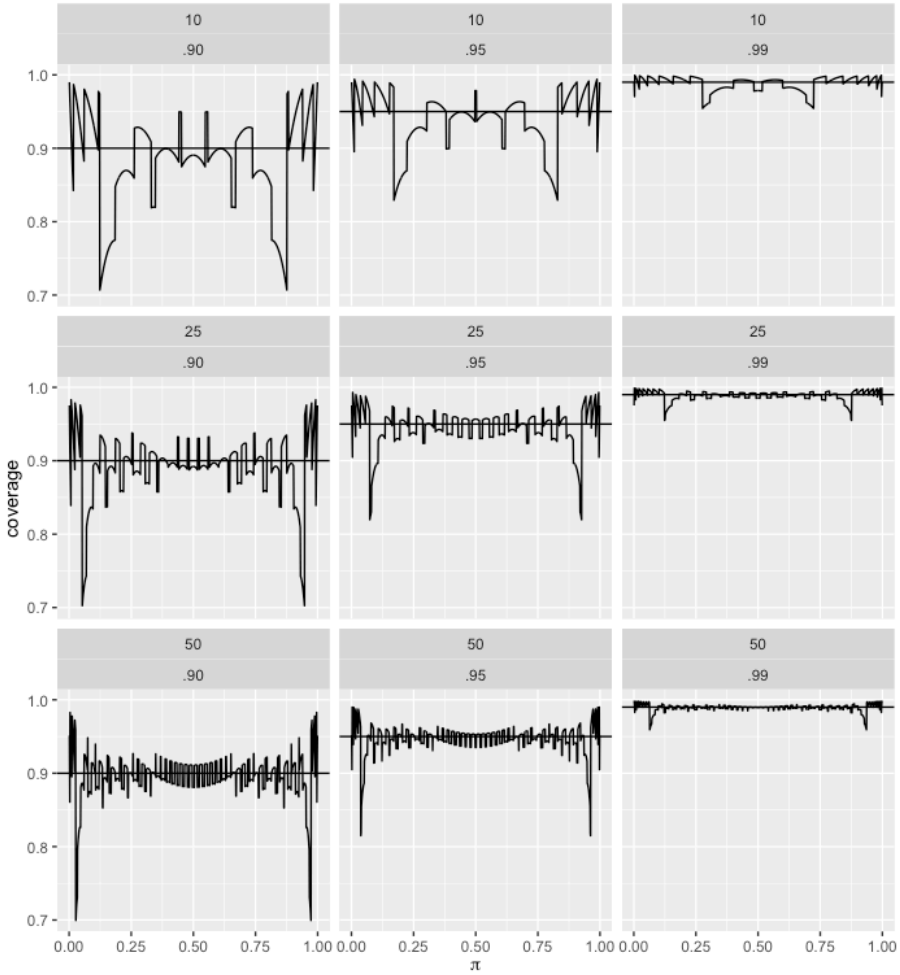
Figur 10: Nominell och faktisk täckningsgrad för Jeffreys HDP intervallet som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



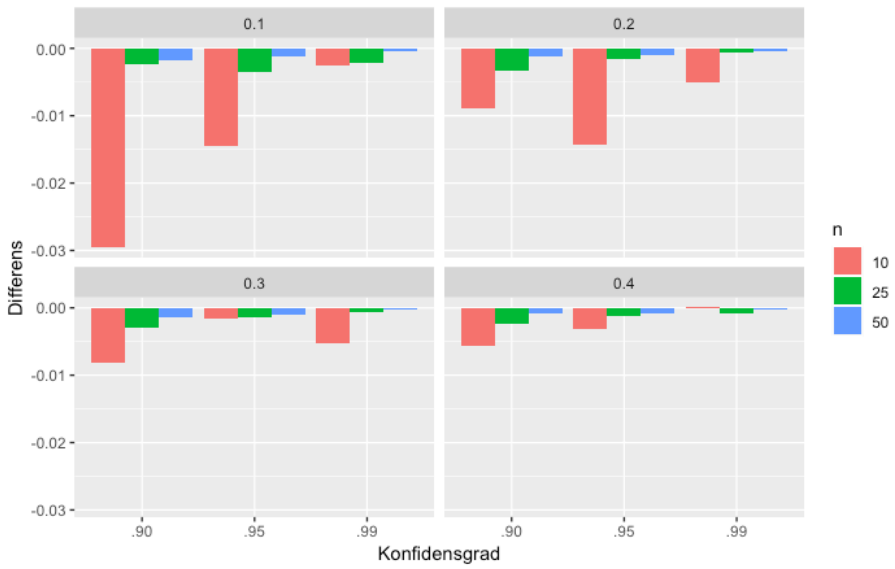
Figur 11: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Jeffreys HDP-intervall som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1-\alpha$ och $b = 0.1$, 0.2, 0.3, 0.4.

$1-\alpha$	n	$\bar{\beta}_n(0)$	differens
0.90	10	0.8850155	-0.0149845
0.90	25	0.8927157	-0.0072843
0.90	50	0.8957693	-0.0042307
0.95	10	0.9420954	-0.0079046
0.95	25	0.9457750	-0.0042250
0.95	50	0.9473973	-0.0026027
0.99	10	0.9885920	-0.0014080
0.99	25	0.9891343	-0.0008657
0.99	50	0.9893841	-0.0006159

Tabell 6: Nominell täckningsgrad $1-\alpha$, faktisk genomsnittlig täckningsgrad $\bar{\beta}_n(0)$ mellan 0 och 1, samt differensen dem emellan, för ett Jeffreys liksvansat konfidensintervall baserat på olika stickprovsstorlekar n.



Figur 12: Nominell och faktisk täckningsgrad för Jeffreys liksvansade intervall som funktion av den sökta sannolikheten π för olika val av stickprovsstorleken n och den nominella täckningsgraden $1 - \alpha$.



Figur 13: Skillnad mellan nominell täckningsgrad och genomsnittlig faktisk täckningsgrad $\bar{\beta}_n(b)$ för Jeffreys liksvansade intervall som funktion av olika val av stickprovsstorleken n, nominella täckningsgraden $1 - \alpha$ och $b = 0.1$, 0.2, 0.3, 0.4.

genomsnittliga faktiska täckningsgraden ligger väldigt nära den nominella förutom när konfidensgraden är 0.90 och $n = 10$.

I figur 13 ser vi den genomsnittliga faktiska täckningsgraden när vi infört restriktioner på π. Vi ser en försämring av den faktiska täckningsgraden när vi har mer restriktioner på π, till exempel när $n = 10$, konfidensgrad 0.9 och $b = 0.1$ är differensen mellan faktiska och nominella täckningsgraden -0.03 medans utan restriktioner på π så är differensen ungefär -0.015 om konfidensgraden ökas till 0.95.

3.3 Utvärdering av täckningsgraden

För att utvärdera konfidensintervallen så kommer vi först jämföra de frekventiska intervallen emot varandra och sedan de Bayesianska intervallen emot varandra. Vi ser bland de frekventistiska intervallen att Wald-intervallet har den klart sämsta faktiska genomsnittliga täckningsgraden följt av arcsin-transformationen. Skillnaden mellan dessa två intervall beror dock på hur stor differensen är mellan den faktiska täckningsgraden och den nominella täckningsgraden när vi inför restriktioner på π. Vi ser då att Waldintervallet har fortfarande väldigt stor differens för alla testade värden av b medan

för arcsin-transformationen så blir differensen liten när $b = 0.2, 0.3$ och 0.4 så man kan anse att intervallet approximativt uppnår sin nominella täckningsgrad. Dessa två intervall delar båda egenskapen att den faktiska täckningsgraden går mot 0 för π nära 0 eller 1 och det är detta som gör att den genomsnittliga faktiska täckningsgraden är mycket mindre än den nominella täckningsgraden. Wilson-Score intervallet och Agresti-Coull intervallet har inte så kraftiga avvikelser från den nominella täckningsgraden som Wald och arcsin-transformationen. Deras faktiskta täckningsgrad skiljer sig aldrig kraftigt åt från varandra men Agresti-Coullintervallet har alltid högre faktiskt täckningsgrad än den nominella och är på så sätt ett mer konservativt intervall. Dock så är den genomsnittliga faktiska täckningsgraden för Wilson-Score intervallet under den nominella endast när vi har en nominell konfidensgrad av 0.90. För alla våra kombinationer av n och α så är den genomsnittliga faktiska konfidensgraden för Wilson-Score alltid närmre den nominella täckningsgraden. När vi inför restriktioner på π så ser vi att differensen mellan den genomsnittliga faktiska täckningsgraden och den nominella täckningsgraden är mindre hos Wilson-Score intervallet än hos Agresti-Coull intervallet för alla olika värden av b. Med detta i åtanke så är då Wilson-Score intervallet ett bättre konfidensintervall om man inte vill ha ett konfidensintervall vars genomsnittliga faktiska täckningsgrad alltid är högre än den nominella. Vid en jämförelse mellan de två Bayesianska intervallen så är det liksvansade intervallet så ser vi att båda intervallen har en genomsnittlig faktisk täckningsgrad som ligger under den nominella täckningsgraden. Vi ser även att differensen är större för HDP-intervallet när vi har en konfidensgrad av 0.9 (för alla olika n) och när vi har en konfidensgrad av 0.95 när vi låter $n = 10$ och 25 annars har det liksvansade intervallet en större differens. När vi inför restriktioner på π så ser vi att det liksvansade intervallet har en lägre differens för alla olika värdena av b, n och nominell konfidensgrad.

4 Undersökande av normalitetsantagande under $\mathcal{H}_0$

4.1 Samband mellan faktisk täckningsgrad och $\mathcal{D}_n$

Det finns ett samband mellan den faktiska täckningsgraden, $\beta_n(\pi)$, och Kolmogorovavståndet, $\mathcal{D}_n$. Utifrån den allmänna frekventistiska definitionen av konfidensintervall i avsnitt 3.1 får man nämligen

$$|\beta_n(\pi) - (1 - \alpha)| \leq 2\mathcal{D}_n,$$

då man kan utnyttja att

$$\beta_n(\pi) - (1-\alpha) = F_n(b) - G(b) - (F_n(a) - G(a))$$

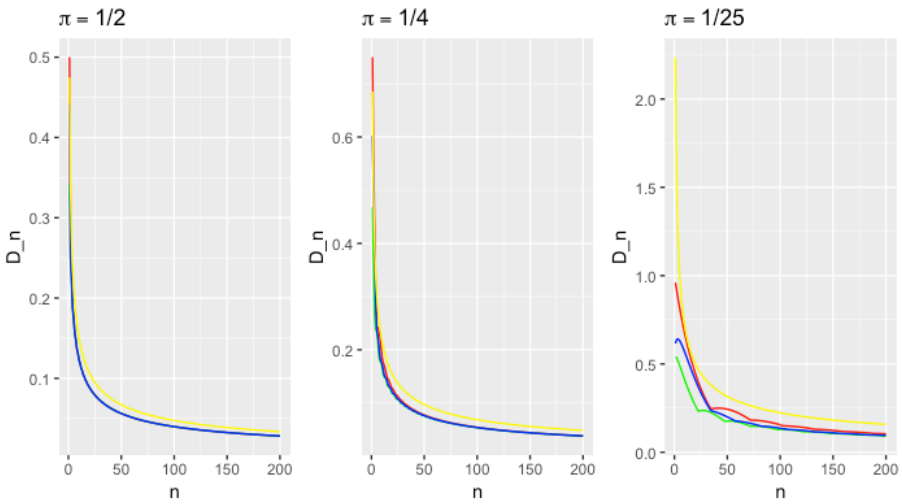
där $a = G^{-1}(\alpha/2)$ och $b = G^{-1}(1-\alpha/2)$, och sedan använda triangelolikheten för att få uppskattningen ovan. Detta samband ger ett mer explicit bekräftelse av att $\mathcal{D}_n$ faktiskt mäter hur bra den nominella täckningsgraden approximerar den nominella.

4.2 Numerisk beräkning av $\mathcal{D}_n$

I detta avsnitt ska vi undersöka hur snabbt test-statistikorna som ligger används för Wald-, Score- respektive arcsin-intervallen går mot normalfördelning. Vi ska göra detta med hjälp av Kolmogorovdistansen $\mathcal{D}_n = \sup_{x \in R} |G(x) - F_n(x)|$ där G är fördelningsfunktionen för standard normalfördelningen och F_n är fördelningsfunktionen för den valda test-statistikan T_n som kommer bero på π. Då utfallsrummet för $X \sim Bin(n, \pi)$ är diskret och finit så är mängderna av möjliga värden för test-statistikorna också finita och har samma kardinalitet som utfallsrummet. Om vi låter $t = \{t_0, t_1, ..., t_n\}$ så att $P(T_n = t_i) = P(X = i)$ så ser vi att det enda test-statistikorna gör är att flytta runt sannolikhetspunktmassorna och frågan är ju då om dessa omflyttningar av test-statistikorna är olika bra på att närma sig normalitet. Vi kan då bilda den nya kumulativa fördelningsfunktionen för T_n som $F_n(x)$. Man ser att F_n blir en trappstegsfunktion som tar steg vid $t_0, t_1, ..., t_n$. Vi vill då undersöka supremumavståndet till en annan kumulativ fördelningsfunktion (den standardiserade normala i vårt fall) $G(x)$. Vi avgränsar oss alltså till att $G(x)$ är kontinuerlig och vi kan då se att på ett intervall $[t_{i-1}, t_i)$ så är F_n konstant medan $G(x)$ är monotont växande och då kommer vi antingen ha supremumavståndet i t_{i-1} eller i t_i. Man ser det enklast om man delar in i tre olika fall.

1. Fall 1: $F_n \leq G$ på $[t_{i-1}, t_i)$ så är supremumavståndet i t_i.
2. Fall 2: $F_n \geq G$ på $[t_{i-1}, t_i)$ då är supremumavståndet i t_{i-1}.
3. Fall 3 $F_n \leq G$ på $[t_{i-1}, c)$ och $F_n \geq G$ på $[c, t_i)$ för någon konstant c, vi har då supremumavståndet i antingen t_{i-1} eller t_i.

Det finns bara ett litet problem med detta upplägg, för Wald-teststatistikan kan det inträffa att $\hat{\pi}(1-\hat{\pi}) = 0$. Vi kommer då att dividera med 0 i definitionen av T_n är så klart odefinierat i de flesta fall. Här låter vi då teststorheten T_n anta värdet $-\infty$ eller ∞, det vill säga om utfallet är $X = 0$, så är $t_0 = -\infty$ och om $X = n$ så sätter vi $t_n = \infty$. Tyvärr kommer detta



Figur 14: Kolmogorovavståndet $\mathcal{D}_n$ mellan teststorheten och en standard normalfördelning som funktion av stickprovsstorleken n, för Waldstatistikan (röd), Score-statistiken (grön), arcsin-teststatistikan (blå) och $d_n = C/\sqrt{n}$ (gul).

göra att F_n i en formell mening inte är uppfyller ett av kraven för kumulativa fördelningsfunktioner då $\lim_{x\to-\infty} F_n(x) = P(T_n = t_0) \neq 0$ och $\lim_{x\to\infty}$ $\mathrm{F}_n(x) = P(T_n < t_n) \neq 1$.

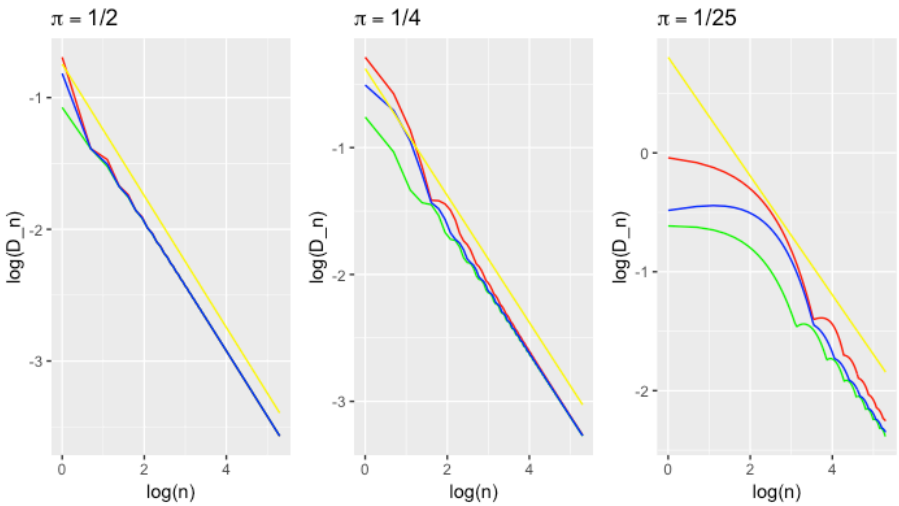
Här i figur 8 ser vi $\mathcal{D}_n$ för Wald-teststatistikan, Score-teststatistikan, arcsin-teststatistikan och $d_n = C/\sqrt{n}$, där konstanten C definieras i avsnitt 4.3. Vi ser att $\mathcal{D}_n$ approximativt har samma form som d_n i fallen då $\pi = 1/2, 1/4$ men när $\pi = 1/25$.

I figur 9 ser vi en log-log plot av sambanden från figur 8 och vi ser ännu tydligare att de olika teststatistikorna går mot $-1/2$ i de två tidigare fallen och vi ser att för $\pi = 1/25$ verkar lutningen också närma sig -1/2. Detta ger oss en misstanke att $\mathcal{D}_n \propto 1/\sqrt{n}$.

4.3 Asymptotisk teori för $\mathcal{D}_n$

Enligt Berry-Esseens sats så existerar ett C så att givet $Z_1, Z_2, \ldots$ som är oberoende och likafördelade med $E[Z_1] = 0$, $E[Z_1^2] = \sigma^2 < \infty$, $E[|Z_1|^3] = \rho < \infty$ och om F_n är kumulativa fördelningsfunktionen för

$$\frac{Z_1 + \ldots + Z_n}{\sigma\sqrt{n}}$$



Figur 15: Logaritmen $\log[\mathcal{D}_n]$ av Kolmogorovavståndet mellan teststorheten och en standard normalfördelning som funktion av logaritmen $\log(n)$ av stickprovsstorleken. De olika kurvorna svarar mot Waldstatistikan (röd), Score-statistiken (grön), arcsin-teststatistikan (blå) och $d_n = C/\sqrt{n}$ (gul).

så är

$$\mathcal{D}_n = \sup_{x \in R} |F_n(x) - G(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

där G är fördelningsfunktionen för en standard normalfördelning.

Den nuvarande bästa estimeringen av C är 0.4748 framtagen av Shevtsova (2011). Vi kan då undersöka hur Kolmogorovavståndet är för den standardiserade binomiella variabeln

$$T_n = \frac{\sum_{i=1}^{n}(X_i - \pi)}{\sqrt{n\pi(1-\pi)}},$$

vilket vi känner igen som T_n för Score-teststatistikan, och använder att $Z_i = X_i - \pi$, där X_i är Bernoullifördelad. Det ger att $E[Z_i] = 0$,$E[Z_i^2] = \pi(1-\pi)$ och att $E[|Z_i|^3] = \pi^3(1-\pi) + (1-\pi)\pi^3 = \pi(1-\pi)[\pi^2 + (1-\pi)^2]$. Det ger i sin tur att den övre gränsen för Kolmogorovavståndet blir

$$\mathcal{D}_n \leq \frac{C\pi(1-\pi)[\pi^2 + (1-\pi)^2]}{\sqrt{(\pi(1-\pi))}^3\sqrt{n}} = \frac{C(\pi^2 + (1-\pi)^2)}{\sqrt{\pi(1-\pi)}\sqrt{n}}.$$

Vidare ser vi att den lägsta övre gränsen för Kolmogorovavståndet är när $\pi = 1/2$ vilket stämmer överens med intuitionen då den binomiella fördelningen då är symmetrisk. Om vi då anser att något är approximativt normalfördelat

när $\mathcal{D}_n < \varepsilon$ för något $\varepsilon > 0$ så kan vi få en undre gräns för hur många observationer vi måste ta med i vårt stickprov för att uppnå den nivån av normalapproximation. Med hjälp av Berry-Esseens sats då vi får att

$$n = \left(\frac{C(\pi^2 + (1-\pi)^2)}{\sqrt{\pi(1-\pi)}\varepsilon} \right)^2$$

blir vårt n för att garantera att $\mathcal{D}_n \leq \varepsilon$.

5 Diskussion

Vi har nu sett hur de olika intervallen presterar under våra valda omständigheter och att Wald och arcsin-transform intervallen presterar klart sämst när det gäller genomsnittlig faktisk täckningsgrad på grund av deras dåliga egenskaper när π är nära 0 eller 1. Vi ser dock att när vi inför restriktioner på π och inte tar med områdena nära 0 eller 1 så uppnår arcsin-transform intervallet en genomsnittlig faktisk täckningsgrad som är väldigt nära den nominella när π är inom (0.2,0.8) medan Wald-intervallets täckningsgrad inte ligger nära det nominella. Så arcsin-transform intervallet är att föredra över Wald-intervallet men de båda är sämre än alla andra undersökta intervall. Om vi jämför Wilson-Score och Agresti-Coull så är det svårt att säga vilken metod som är bättre då de båda har en faktisk genomsnittlig täckningsgrad som är väldigt nära den nominella i alla undersökta fall, dock så är Wilson-Score intervallet närmre den nominella täckningsgraden i de undersökta fallen. En stor skillnad mellan dem är hur de beter sig när π är nära 0 eller 1, då Agresti-Coull blir väldigt konservativ medan täckningsgraden för Wilson-Score gör hopp mellan att vara under och över den nominella täckningsgraden. Man kan se fördelar och nackdelar med båda beroende på omständigheterna bakom skattningen av proportionen. En fördel Agresti-Coull har är att den är väldigt lätt att lära ut till personer som är vana att använda Wald-intervallet i fallet $\alpha = 0.05$ då $z_{(\alpha/2)} \approx 2$ och då kan läras ut som att man adderar 2 lyckade observationer och 2 misslyckade observationer och sedan gör ett Wald-intervall på den modifierade datan. Att Agresti-Coull intervallet är väldigt lätt att lära ut till personer som använder Wald-intervallet och inte har så stor skillnad i faktisk täckningsgrad gentemot Wilson-Score gör att man kan rekommendera denna metod över Wilson-Score. Vi såg också hur Jeffreys liksvansade intervall hade en genomsnittlig faktisk täckningsgrad som var närmare den nominella täckningsgraden än Jeffreys HDP-intervall. Om vi jämför figur 10 och 12 så ser vi att den faktiska täckningsgraden för HDP-intervallet ser ut att ligga antingen under eller över det nominella värdet för

större sammanhängande mängder av värden på π medan det liksvansade intervallet ofta gör diskontinuerliga hopp mellan att vara över och under den nominella täckningsgraden då π varierar.

Vi har också sett hur avståndet till den standardiserade normalfördelningen inte påverkas så mycket av vilken test-statistika man använder för stora n, dock för de utvalda värdena för π så ser vi att Score-teststatistikan alltid var den som hade lägst avstånd till standardiserade normalfördelningen. Vidare studier kan göras där man undersöker andra metoder att göra konfidensintervall, till exempel så kan man ta med exakta konfidensintervall. En studie i varför $\mathcal{D}_n$ får större oscillationer ju närmare π är 0 eller 1 skulle också vara av intresse.

6 Referenslista

Agresti, A. och Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*. 52: 119126. doi:10.2307/2685469. JSTOR 2685469. MR 1628435

Held, L. and Bové, D. (2013) *Applied Statistical Inference*. Springer, Berlin. ISBN: 9783642378867

Shevtsova, I. (2011). On the absolute constants in the Berry Esseen type inequalitites for identically distributed summands. arXiv:1111.6554.

Thulin, M. (2014). On Confidence Intervals and Two-Sided Hypothesis Testing. DiVA, id: diva2:736874