

Kandidatuppsats

Statistiska institutionen

Bachelor thesis, Department of Statistics

Nr 2015:3

Utveckling och testning av ett program för imputering vid longitudinella kliniska studier

*Development and testing of a computer program for
imputation in longitudinal clinical trials*

Michael Schenkenberg van Mierop

Självständigt arbete 15 högskolepoäng inom Statistik III, HT2015
Handledare: Hans Nyquist

Sammanfattning

I uppsatsen jämförs tre imputeringsmetoder för att avgöra vilken av metoderna som är lämpligast att tillämpa i ett dataprogram som ska imputera vid bortfall i longitudinella kliniska prövningar.

Jämförelsen görs dels genom att mäta metodernas träffsäkerhet vid imputeringar (kvantitativ jämförelse) och dels genom att titta på hur väl metoderna hanterar ett antal problemområden (kvalitativ jämförelse). Den metod som framstår som lämpligast är tänkt att senare föras in i ett dataprogram som ska imputera data vid bortfall i longitudinella kliniska prövningar.

För ändamålet jämförs tre olika metoder i Last object carried forward familjen: Last Value Carried Forward, Last Percentile Carried Forward och Last Grade Carried Forward vilken är en ny metod som baserar imputeringarna på ordningsstatistik.

Den kvantitativa bedömningen görs i ett testprogram, programmerad i Python, som simulerar ett antal tidsserier med simulerade bortfall enligt olika bortfallsmekanismer och jämför sedan förmågan att imputera värden med de olika metoderna genom att mäta träffsäkerheten i imputeringarna utifrån tre olika avståndsmått. De kvantitativa jämförelserna visar i uppsatsen på att resultatet inte är entydigt utan att olika metoder är bra på olika typer av bortfallsmekanismer och datatyper.

I den kvalitativa bedömningen jämförs metodernas förmåga att hantera de brister som de olika varianterna av Last Object Carried Forward för med sig enligt punktlistan nedan.

- Hur bör mätvärden extrapoleras då de högst och lägst rankade försökspersonerna saknas?
- Hur interpoleras mätvärden då flera individer intill varandra har fallit bort?
- Vilken rang ska tas bort då en av två personer med lika mätvärden hoppar av.

Sammantaget visar metodjämförelserna att valet av metod måste utgå från hur data ser ut men om en metod behöver väljas för att hantera all data framstår LGCF som den bästa kompromissen.

Nyckelord: bortfall, bortfallsmekanism, Imputering, kliniska försök, Last Grade Carried Forward, Last Percentile carried forward.

Innehållsförteckning

1	Introduktion	3
1.1	Bakgrund	3
1.2	Syfte.....	3
1.3	Disposition.....	4
2	Grundläggande begrepp	5
2.1	Bortfallsmönster och bortfallsmekanik	5
2.1.1	Bortfallsmönster	5
2.1.2	Ignorerbar och inte ignorerbar bortfallsmekanisk	6
2.1.3	Bortfallsmekanismer	6
3	Konstruktion av datorprogram för imputering av bortfall	10
3.1	Vilken variant av LOCF lämpar sig bäst för programmet?	10
3.1.1	Metoden Last Value Carried Forward.....	11
3.1.2	Metoden Last Rank Carried forward.....	11
3.1.3	Metoden Last Percentile Carried Forward	12
3.1.4	Vilka problem finns med LPCF och LVCF metoderna?.....	13
3.1.5	Metoden Last Grade Carried Forward.....	14
3.2	Jämförelse av de starkaste kandidaterna.....	16
3.2.1	Simulerade tidsserier	17
3.2.2	Simulerade bortfall	19
3.2.3	Kriterier för val av imputeringsmetod.....	19
3.2.4	Resultat.....	20
3.2.5	Analys.....	24
3.2.6	Slutsats	29
3.3	Fortsatta studier	30
3.3.1	Metodjämförelserna	30
3.3.2	Färdigställa programmet	31
	Litteraturförteckning	32
	Bilagor	34
	Bilaga 1, Lämpliga metoder beroende på bortfallsantaganden.	34
	Bilaga 2, Multipel imputering (MI).....	35

1 Introduktion

Ett mål i det här arbetet är att skriva ett program i programmeringsspråket Python som ska imputera data vid longitudinella kliniska provningar. Metoden för imputering ska, förutom att imputera med god träffsäkerhet även hantera ett antal vanliga problem som imputeringsmetoder medför. Jämförelser mellan ett antal metoder kommer göras på ett antal tidsserier som simulerar longitudinella studier med olika egenskaper och med olika typer av bortfall. Även om programmet inte kommer att färdigställas i samband med att uppsatsen skrivs så kommer grundläggande programkod för imputering att skrivas i samband med det här arbetet.

1.1 Bakgrund

Uppsatsen skrivs på uppdrag av ett statistikanalysföretag som är i behov av ett dataprogram som imputerar data vid bortfall i longitudinella kliniska studier¹. Programmet ska kunna hantera trendande data och ett antal problematiska specialfall som en del imputeringsmetoder inte kan hantera. De metoder som jämförs i programmet är Last Percentile Carried Forward (LPCF), Last Value Carried Forward (LVCF)² och Last Grade Carried Forward (LGCF). LPCF och LVCF har utsatts för en del kritik till exempel för att de inte alltid imputerar konservativt, saknar tillräcklig statistisk grund (Deng (2012)), och ger systematiska fel i imputeringen (Fitzmaurice, G.M., (2011)). För att hantera de problemområden som företaget har lyft upp och också för att göra en ansats till att skapa en statistisk grund för den här metodfamiljen så introduceras dessutom en tredje ny metod som benämns Last Grade Carried Forward (LGCF) och som enkelt uttryckt är en variant av metoden LPCF men som utgår från ordningsstatistik vid imputeringarna.

Innan avgränsningen av vilka metoder som ska ingå i jämförelsen togs ytterligare en vanlig metod under övervägande: Multipel imputation³. Metoden är intressant på grund av dess popularitet men har inte tagits med i jämförelsen då det skulle medföra omfattande programmeringsarbete och då metoden istället kan tillämpas i befintliga statistikverktyg som till exempel SAS, SPSS och R. Dessutom anses denna metod inte alltid vara lämplig vid alla bortfallsmekanismer⁴.

1.2 Syfte

Syftet är att jämföra imputeringsmetoderna Last Value Carried Forward, Last Percentile Carried Forward och Last Grade Carried Forward⁵ för att avgöra vilken av dessa som är lämpligast för ett dataprogram som ska imputera vid bortfall i data från longitudinella kliniska provningar.

¹ En klinisk studie eller klinisk provning ”är en undersökning på friska eller sjuka människor för att studera effekten av ett läkemedel eller behandlingsmetod.” En longitudinell klinisk studie är en klinisk studie som genomförs över tid (kort eller lång) då upprepade mätningar görs”, wikipedia (2015)

² LPCF och LRCF är två likartade metoder som används framförallt för imputering vid longitudinella data. Se avsnitt 3.1 för en mer ingående beskrivning.

³ Se bilaga 2.

⁴ Se bilaga 1, bilaga 2 och avsnitt 2.1.3.

⁵ Metoderna beskrivs mer ingående i avsnitt 3.1.

1.3 Disposition

I nästa kapitel, Kapitel 2, ges en beskrivning av de grundläggande begrepp som används i uppsatsen vilka är viktiga för att förstå tanken bakom de tester och jämförelser som görs i Kapitel 3. I Kapitel 3 beskrivs valet av metod som innebär att titta mer ingående på de alternativa metoderna och titta på hur jämförelserna utförs avseende hur testdata tas fram. I Kapitel 3 presenteras även jämförelsens resultat, analys och slutsats och slutligen beskrivs i korta ordalag nästa steg avseende fortsatt analys och hur dataprogrammet ska färdigställas.

2 Grundläggande begrepp

I det här kapitlet beskrivs begrepp som behövs för att förstå resonemanget i uppsatsen: Dels beskrivs de viktiga begreppen bortfallsmönster och bortfallsmekanik vilka är viktiga för att förstå de jämförande testernas utformning och dels beskrivs mer översiktligt longitudinella kliniska försök.

2.1 Bortfallsmönster och bortfallsmekanik

Bortfallsmönster och bortfallsmekanik är viktiga begrepp för analys av bortfall i syfte att till exempel hitta lämplig bortfallshanteringsmetod. ”Att titta på bortfallsmönstret svarar på frågan: Vilka data saknas? Att titta på datas bortfallsmekanik svarar på frågan: Hur ser relationen mellan bortfallet och övriga variabler ut?” (Little and Rubin. (2002).

2.1.1 Bortfallsmönster

Bortfallsmönstret beskriver vilka värden som kan observeras i en datamatrix och vilka som fallit bort och kan ibland vara till hjälp för att välja imputeringsmetod eller för att förstå bortfallsmekanismer. En typ av bortfallsmönster som är intressant för det här uppsatsarbetet är monotont bortfallsmönster vilket är typiskt för bortfall vid longitudinella kliniska prövningar. Monotont bortfallsmönster (*Tabell 2.1*) innebär att ett bortfall för en individ följs av upprepade bortfall för samma individ fram till den kliniska prövningens slut. Orsaken till det här mönstret kan vara att individen har valt att avbryta försöket eller har avlidit.

Monotont bortfall

	År 1	År 2	År 3	År 4	År 5	År 6
Individ 1	0	0	0	0	0	0
Individ 2	0	0	0	0	1	1
Individ 3	0	0	0	1	1	1
Individ 4	0	0	1	1	1	1

0 = Observerat data

1 = Bortfall

Tabell 2.1 Monotont bortfall kännetecknas av att bortfall vid en tidpunkt åtföljs av ytterligare bortfall fram till och med senaste tidpunkten.

Analys av bortfallsmönstret görs genom att visuellt fastställa vilka värden i matrisen som har fallit bort och vilka värden som har observerats. Det medför att data måste sorteras rätt innan mönstret kan upptäckas och det är särskilt viktigt vid stora datamatriser. I *Tabell 2.1* har data sorterats över bortfall och mättidpunkt vilket gör att det monotona bortfallsmönstret enkelt kan identifieras.

Förutom ”monotont” bortfallsmönster finns även ”univariat” (när endast en variabel innehåller bortfall), ”fil-matchning” (när bortfall aldrig inträffar samtidigt för de olika variablerna), ”generell” (när bortfall inte uppvisar ett mönster utan är slumpmässigt) med flera (Little and Rubin (2002).

Då den här uppsatsen fokuserar på longitudinella kliniska försök där monotont bortfall är ett typiskt bortfallsmönster behöver den metod som väljs för imputeringsprogrammet åtminstone kunna hantera monotont bortfallsmönster.

2.1.2 Ignorerbar och inte ignorerbar bortfallsmekanism

Som ett led i att finna en lämplig metod för att hantera bortfall kan begreppet ignorerbarhet⁶ användas.

När man talar om att ett bortfall är ignorerbart eller inte syftar man på bortfallsmekanismen⁷ det vill säga om bortfallsmekanismen behöver modelleras för att det inte ska uppstå systematiska fel vid imputering.

När ett bortfall inte är ignorerbart beror sannolikheten för bortfall på själva bortfallsvariabeln och då kan bortfallsmekanismen inte ignoreras utan måste modelleras för att hantera bortfallet. När ett bortfall *är* ignorerbart beror bortfallet på andra variabler än bortfallsvariabeln vilket troligen innebär att dessa variabler istället behöver analyseras för att hantera bortfallet.

Det finns en relation mellan ignorerbarhet och bortfallsmekanismer enligt *Tabell 2.2* nedan. Här framgår att bortfallsmekanismen inte behöver modelleras då bortfallet är ignorerbart. Det förekommer dock sällsynta fall då bortfall behöver modelleras trots att det klassas som Missing At Random (Allison 2001, s.5). Bilaga 1 beskriver ett förslag på hur relationen mellan bortfallsmönster, bortfallsmekanismer och lämpliga imputeringsmetoder ser ut.

Bortfallsmekanism	Ignorerbar	Att beakta vid imputering
Missing Not At Random	Ignorerbar	Bortfallsmekanismen behöver inte modelleras.
Missing At Random	ignorerbar	Bortfallsmekanismen behöver inte modelleras.
Missing Not At Random	inte ignorerbar	Bortfallsmekanismen kan inte ignoreras utan måste modelleras.

Tabell 2.2: Tabellen visar relationen mellan bortfallsmekanism och huruvida bortfallsmekanismen är ignorerbar (eng. ignorable) eller inte.

2.1.3 Bortfallsmekanismer

Teorin om bortfallsmekanismer formulerades av Rubin (1976) i tidsskriften Biometrika. Här beskrev Rubin tre olika mekanismer, var och en med olika beroenden mellan bortfall och övriga variabler. Missing At Random (MAR), Missing Completely At Random (MCAR) och Missing Not At Random (MNAR).

Bortfallsmekanismerna, som även benämns bortfallsklasser eller bortfallsfördelningar är ofta ett viktigt verktyg för att finna en lämplig metod för att hantera bortfall. I bilaga 1 finns ett exempel på hur lämpliga metoder kan väljas utifrån bortfallsmekanism.

⁶ Eng. "Ingorability"

⁷ Se kapitel 2.1.3

Bortfallsmekanismerna nämns ofta i litteraturen och utgör idag ett etablerat klassificeringssystem för beskrivning av bortfall, och är därmed ett av underlagen när LPCF, LVCF och LGCR senare i uppsatsen jämförs. Nedan ges en kort beskrivning av respektive bortfallsklass. Beskrivningarna innehåller exempelbortfall baserade på följande fiktiva exempeltest. Exemplet är hämtat från McKnight, P.E., et. al. (2007).

Exempeltestet består av 10 individer vilka ska uppskatta sina matematikkunskaper på en 10-gradig skala. Därefter ska individerna genomföra två matematiktester med en veckas mellanrum vilka i tabellen nedan benämns Math test 1 och Math test 2 men i uppsatstexten hädanefter benämns *test 1* och *test 2*. I Tabell 2.3 nedan visas tidsserierna för exempeltestet. Tidsserien innehåller följande data: 'Case' som är Individerna i testet, 'Percieved ability' som är den beroende variabeln där individerna bedömer sig själva, Test 1' är oberoende variabel nr. 1 och 'test 2' är oberoende variabel nr 2. Under rubriken 'Mechanisms on Math Test 2' visas i en matris de bortfall som skapats för de tre bortfallsmekanismerna (bortfall representeras här av täckande svarta fyrkanter). Under rubriken 'Dummy codes' finns dummyvariabler som, i variabler, visar var bortfall uppstår. Observera att bortfallen endast drabbar variabel 2, 'test 2'.

Case	Original data			Mechanism on Math Test 2			Dummy codes ^a		
	Perceived Ability	Math Test 1	Math Test 2	MCAR	MAR	MNAR	D _{MCAR}	D _{MAR}	D _{MNAR}
1	3	25	28		28	28	1	0	0
2	2	22	19	19	19	19	0	0	0
3	4	23	26	26	26	26	0	0	0
4	5	27	32	32	32		0	0	1
5	1	15	16	16		16	0	1	0
6	3	16	20	20		20	0	1	0
7	7	22	25	25	25	25	0	0	0
8	8	28	26		26	26	1	0	0
9	9	30	35	35	35		0	0	1
10	5	26	31	31	31	31	0	0	0

Tabell 2.3: Under rubriken 'Original data' finns: 'Case' (individer), 'Percieved ability' den beroende variabeln där individerna bedömer sig själva, 'Math Test 1' är oberoende variabel nr. 1 och 'Math test 2' är oberoende variabel nr 2, 'MCAR', 'MAR', 'MNAR' som är de bortfall som vi skapas för de tre bortfallsmekanismerna (bortfall representeras här av täckande svarta fyrkanter), D_{MCAR}, D_{MAR}, D_{MNAR} är bortfallsindikatorer. Observera att bortfallen endast drabbar variable 2 – 'Math test 2'.

MCAR mekanismen är enklast att förstå och ger många möjligheter att välja metod för hantering av bortfall. Här uppstår bortfall oberoende av både observerat och icke-observerat data. *Exempel: En individ singlar slant för att besluta sig för att genomföra test 2 eller inte. Om individen inte genomför testet uppstår ett bortfall.*

Då bortfallsmekanismen här inte beror av variabler i eller utanför försöket medför den här bortfallsklassen att bortfallsmekanismen kan ignoreras och behöver inte modelleras i estimeringsprocessen. Val av metod för hantering av bortfall medför därför inte heller en risk för systematiska fel.

Little, R. och Rubin, D. (2002) beskriver bortfallsmekanismen som den betingade fördelningen för M givet Y:

$$f(R|Y, \phi) = f(R|\phi) \quad \text{för alla } Y, \phi, \quad (2.1)$$

där:

Φ = okända parametrar,

$Y = (y_{ij})$ = det fullständiga datat,

$R = (R_{ij})$ = bortfalls-indikatormatrisen.

MAR är en mer begränsad form av bortfallsmekanik än MCAR och ger färre möjligheter vid val av metod. Till skillnad från vid MCAR finns här ett beroende mellan bortfall och något av eller flera av testets variabler men inte variabeln som har drabbats av bortfall⁸. *Exempel: Sannolikheten för bortfall i variabeln test 2 beror på test 1 eller 'Perceived Ability' men inte på test 2 (variabeln med bortfall).*

Liksom vid MCAR gäller att bortfallsmekanismen bakom bortfallet kan ignoreras eftersom variabeln som styr bortfallsmekanismen inte är relaterad till variabeln som ska estimeras (det vill säga bortfallet) och val av metod för hantering av bortfall påverkar inte heller här risken för systematiska fel.

$$f(R|Y, \phi) = f(R|Y_{obs}, \phi) \quad \text{för alla } Y, \phi \quad (2.2)$$

där:

Φ = okända parametrar,

Y_{obs} = är observerat data (Y)

$R = (R_{ij})$ är bortfallsindikationsmatrisen.

MNAR gäller när MCAR eller MAR inte gäller. Vid MNAR beror bortfall på bortfallsvariabeln själv eller på icke-observerat data. *Exempel: Sannolikheten för bortfall i variabeln test 2 beror på test 2 eller att försöksledaren tappar bort testsvaren för en individ.* Vid MNAR säger man att bortfallsmekanismen inte kan ignoreras utan att systematiska fel uppstår. Här krävs snarare god kunskap om orsaken till bortfallet och väl genomtänkt val av metod där bortfallsmekanismen modelleras vid estimering av bortfallet (Allison 2001, s.5).

$$f(Y, R|\Theta, \phi) = f(Y|\Theta)f(R|Y, \phi) \quad \text{för alla } Y, \phi \quad (2.3)$$

där:

Φ = okända parametrar,

$f(y_i|\Theta)$ = täthet för y_i ,

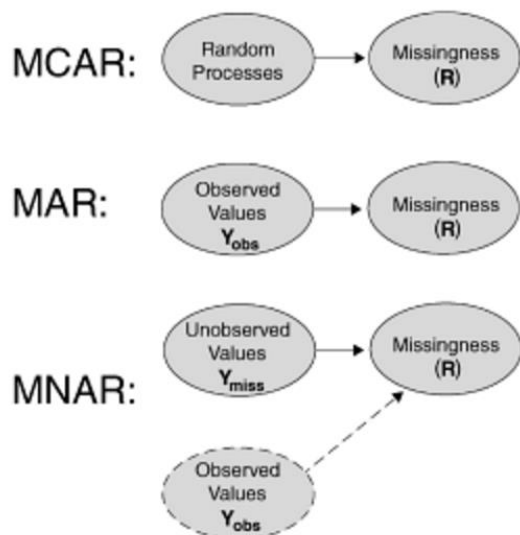
Θ = Okända parametrar,

Y_{obs} = är observerat data (Y),

$R = (R_{ij})$ är bortfallsindikationsmatrisen.

Beroendeförhållanden mellan bortfall och observerade eller icke-observerade värden sammanfattas i bilden nedan. Här är Y_{obs} observerade variabler, Y_{miss} variabler som inte är observerade som till exempel bortfallet, R är en bortfallsindikator och pilarna står för beroende förhållanden.

⁸ Ursprungligen skrev Rubin att bortfall enligt MAR var beroende av andra variabler än bortfallsvariabeln givet andra variabler än bortfallsdata men förenklade senare villkoren enligt ovan. Den nya tolkningen av MAR har blivit vanlig i modern statistik och är den som Rubin själv vill lyfta fram i Rubin (2002) s. 11.



Figur 2.1: Bilden beskriver hur relationen mellan bortfall och kända eller okända variabler ser ut vid Rubins tre bortfallsmekanismer (Mcknight, P.E., et.al.(2007)). Där: R =En dummyvariabel som representerar bortfall, Y_{miss} = ickeobserverad variabel och Y_{obs} =observerad variabel. Pilarna står för beroendeförhållanden.

3 Konstruktion av datorprogram för imputering av bortfall

Longitudinella kliniska studier (eller prövningar) är statistiska undersökningar som genomförs i medicinska sammanhang där mätningar utförs upprepade gånger över en tidsperiod, ofta över flera år. Bortfall i longitudinella kliniska prövningar utgörs ofta av bortfall med bortfallsmönstret monotont bortfall. Statistikanalysföretaget kräver inte att datorprogrammet ska imputera enligt en specifik bortfallshanteringsmetod men förordar LVCF eller en variant av LVCF. Denna grupp av imputeringsmetoder är relativt vanliga vid longitudinella kliniska försök eftersom de lämpar sig för hantering av bortfall i longitudinellt data med monotont bortfall (Furberg C. et.al. (2010), kap. 17). Men de kräver att försöket förbereds för LVCF metoden till exempel genom randomiseringar av försöksunderlaget etc. Två varianter av LVCF är⁹ LPCF och Last Rank Carried Forward¹⁰ (LRCF). LPCF kommer att vara en av de metoder som ska jämföras i den här uppsatsen.

LVCF-liknande metoder benämns i uppsatsen härnäst Last Object Carried Forward LOCF. Jag vill uppmärksamma läsaren på att LOCF inte ska blandas ihop med Last Observation Carried Forward som i litteraturen vanligen också förkortas LOCF och har då samma innebörd som LVCF har i den här uppsatsen.

LOCF är metoder som är enkla att förstå och tillämpa men de har också fått kritik för att vara behäftade med ett antal brister. Till exempel anses de medföra systematiska fel men trots det tycks de vara relativt vanliga och omnämns specifikt i dokumentet ICH E9 "Statistical principles for clinical trials" som har tagits fram av internationella organet The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)¹¹ som är en viktig och normsättande aktör avseende hur kliniska försök bör genomföras. För att hantera några av LOCF metodernas brister görs i uppsatsen en ansats att åtgärda några av bristerna genom att ta fram en ny LOCF variant som är baserad på ordningsstatistik och benämns Last Grade Carried Forward (LGCF).

Inför konstruktionen av datorprogrammet jämförs några av de olika LOCF varianterna som nämns ovan genom att tillämpa dem på fem uppsättningar simulerat longitudinellt kliniskt data som skapade med olika modeller för att testa metoderna under olika förutsättningar. Även olika typer av bortfall enligt de olika bortfallsmekanismerna skapas i tidsserierna.

3.1 Vilken variant av LOCF lämpar sig bäst för programmet?

De olika LOCF varianterna lämpar sig olika väl för olika typer av longitudinellt data och möjligen lämpar de sig också olika väl för olika typer av bortfallsmekanismer. Vid valet av metod för datorprogrammet jämförs de tre varianterna LVCF, LPCF och den nya varianten LGCF genom att mäta avstånd vid imputering på de fem simulerade tidsserierna och de tre bortfallsmekanismerna.

⁹ Vid denna LOCF variant låter man senaste percentilen bibehållas framåt vid varje kommande mättillfälle fram till försökets slut.

¹⁰ Vid denna LOCF variant låter man senaste rank bibehållas framåt vid varje kommande mättillfälle fram till försökets slut.

¹¹ Citat från ICH E9 där LVCF nämns: "[...] Imputation techniques, ranging from the carrying forward of the last observation to the use of complex mathematical models, may also be used in an attempt to compensate for missing data [...]"

En kort beskrivning av olika LOCF varianterna ges i avsnitt 3.1.1. nedan.

3.1.1 Metoden Last Value Carried Forward

LVCF¹² beräknas enligt nedan:

1. Då en patient hoppar av ersätts det saknade värdet med det senast uppmätta värdet, som följer med till sista mättillfället (*Tabell 3.4*).

Individ	1	2	3
Mättillfälle 1	3.8	4.1	2.7
Mättillfälle 2	3.1	3.5	2.4
Mättillfälle 3	2.0	3.8	2.9
Mättillfälle 4	? -> 2.0	2.4	3.5
Mättillfälle 5	? -> 2.0	2.8	? -> 3.5
Mättillfälle 6	? -> 2.0	3.0	? -> 3.5

Tabell 3.4: Exmpel på Last Value Carried Forward (LVCF)

3.1.2 Metoden Last Rank Carried forward

LRCF jämförs inte i programmet men beskrivs här av pedagogiska skäl. LRCF genomförs enligt de tre stegen nedan:

1. Vid varje mättillfälle rangordnas försökspersonerna baserat på deras uppmätta mätvärden (*Tabell 3.8*).

2. Då en patient hoppar av noteras dess rang vilken stryks från listan över möjliga ranger de kvarvarande försökspersonerna kan få vid kommande mättillfällen. Hädanefter har den avhoppade patienten alltid sin sist uppmätta rang.

3. Vid sista mättillfället noteras vilka absoluta värden som mäts upp för de kvarvarande patienterna samt rangordnar dem som innan och interpolerar de saknade försökspersonernas värden utifrån värden för de närmast högre och lägre rankade individerna.

¹² En annan vanlig benämning för metoden är Last observation Carried Forward, LOCF

Individ		1	2	3	4	5	6
Mättillfälle 1	värde	89	76	64	45	32	11
	rang	1	2	3	4	5	6
Mättillfälle 2	värde	70	80	76	55	43	50
	rang	3	1	2	4	6	5
Mättillfälle 3	värde	?	60	?	43	55	20
	rang	3	1	2	5	4	6
Mättillfälle 4	värde	?	70	?	40	49	15
	rang	3	1	2	5	4	6
Mättillfälle 5	värde	?	55	?	35	?	25
	rang	3	1	2	5	4	6
Mättillfälle 6	värde	Tas bort	40	34	28	29	30
	rang	3	1	2	6	4	5

Tabell 3.8: Imputering vid LRCF. Här beräknas rangen för individerna vid varje mättillfälle. Vid bortfall behåller individen sin senaste rang och värde beräknas då utifrån omkringliggande observationer. I detta exempel beräknas inte värdet eftersom tillvägagångssätt i detta fall är oklart (se avsnitt 3.1.4).

3.1.3 Metoden Last Percentile Carried Forward

Imputering enligt LPCF utförs i följande tre steg:

1. Vid varje mättillfälle rangordnas försökspersonerna i en $[0,1]$ värdemängd baserat på deras och övriga individers uppmätta mätvärden (Tabell 3.6).
2. Då en patient hoppar av behåller den avhoppade patienten sin sist uppmätta percentil.
3. De framräknade och estimerade percentilerna skalas i sista steget om till den ursprungliga värdemängden (Tabell 3.7).

Exemplet nedan visar hur LPCF stegen ser ut uppdelat på tre tabeller för vardera: observerade värden, rangen för ett värde i värdemängden $[0,1]$ samt värdet som motsvarar percentilrangen men i den ursprungliga värdemängden. Kontakta författaren för att få en kopia av Excel-dokumentet som har använts i exemplet.

Observerade värden och bortfall						
Individ	1	2	3	4	5	6
Mättillfälle 1	3,8	4,1	2,7	3,2	4	4
Mättillfälle 2	3,1	3,5	2,4	3,4	4,3	3,6
Mättillfälle 3	2	3,8	2,9	4	?	3,8
Mättillfälle 4	?	2,4	3,5	4,3	?	4,3
Mättillfälle 5	?	2,8	?	5	?	5,3
Mättillfälle 6	?	3	?	5,2	?	6

Tabell 3.5: Tabellen innehåller exempel på observationer och bortfall. Bortfall indikeras som "?".

Värden och percentilranger i värdemängden [0,1]						
Individ	1	2	3	4	5	6
Mättillfälle 1	0,786	1,000	0,000	0,357	0,929	0,929
Mättillfälle 2	0,368	0,579	0,000	0,526	1,000	0,632
Mättillfälle 3	0,000	0,900	0,450	1,000	1,000	0,900
Mättillfälle 4	0,000	0,000	0,579	1,000	1,000	1,000
Mättillfälle 5	0,000	0,000	0,579	0,880	1,000	1,000
Mättillfälle 6	0,000	0,000	0,579	0,733	1,000	1,000

Tabell 3.6: Tabellen innehåller ranger uttryckta som ett värde mellan 0 och 1. Bortfall ersätts med den senast uppmätta rangen.

Värden och percentilranger i den ursprungliga värdemängden						
Individ	1	2	3	4	5	6
Mättillfälle 1	3,8	4,1	2,7	3,2	4	4
Mättillfälle 2	3,1	3,5	2,4	3,4	4,3	3,6
Mättillfälle 3	2	3,8	2,9	4	4	3,8
Mättillfälle 4	2,4	2,4	3,5	4,3	4,3	4,3
Mättillfälle 5	2,8	2,8	4,2	5	5,3	5,3
Mättillfälle 6	3	3	4,7	5,2	6	6

Tabell 3.7: Tabellen innehåller de värden som erhålls när rangerna i Tabell 3.3 överförs från värdemängden [0,1] till den ursprungliga värdemängden i Tabell 3.2.

3.1.4 Vilka problem finns med LPCF och LVCF metoderna?

Vid imputering med LVCF, LPCF och LRCF uppstår ett antal problem:

- Högsta och lägsta mätvärden hanteras inte.
- Två eller flera intilliggande mätvärden hanteras inte.
- Det är oklart vilken rang som ska tas bort då en av två personer med samma mätvärden hoppar av¹³.

Vissa av dessa problem har enkla men också ganska grova lösningar som kan medföra olika typer av statistiska fel. Ett exempel på en sådan enkel lösning är att radera de högsta och lägsta värdena då individen är högst eller lägst rankad.

Ett alternativt sätt att hantera dessa problem kan vara en ny metod¹⁴ som innebär att ordningsstatistik appliceras på LOCF principen. Denna ansats benämns härmed "Last Grade Carried Forward" (LGCF) och den kommer senare i uppsatsen jämföras med LVCF och LPCF i datorprogrammet med tester på simulerat data och med bortfall från de tre olika bortfallsmekanismerna. LRCF ingår inte i jämförelsen då metoden är lika eller mindre intressant att jämföra än LPCF.

¹³ Eftersom ordningsberäkningen görs om vid varje mättillfälle.

¹⁴ Såvitt känt presenteras idén för första gången i denna uppsats.

3.1.5 Metoden Last Grade Carried Forward

De problem som LPCF och LVCF ger upphov till och som nämns i avsnitt 3.1.4 hanteras möjligen på ett bättre sätt LGCF då metoden baserar imputeringarna på ordningsstatistik och utgår därför från matematiskt definierade imputeringar och vilket i det här gör att LGCF undkommer problemområdena från 3.1.4. Metoden benämns Last Grade Carried Forward (LGCF) där Grade står för betyg. Benämningen 'betyg' ska spegla metodens förmåga att justera det imputerade värdet för antalet individer i urvalet.

Vid LGCF baseras beräkningen av en imputering vid ett enskilt bortfall på individens ordning bland övriga individer. Det värde som ett bortfall imputeras med är väntevärdet $E(Y_i)$ där i är observationen för en individ vid ett mättillfälle.

Betygen beräknas med ordningsstatistik enligt följande: Observationerna antas vara ett slumpmässigt urval från en standardiserad uniform fördelning, $U(0,1)$ där $f(y)=1$ och $0 < y < 1$. Sannolikhetsfördelningen för en observation med ordningsnummer (rang) i är då enligt (3.1) (McKnight, P.E., et.al. (2007)).

$$Y_{(i)}(y) = \frac{n!}{(i-1)!(n-i)!} F(y)^{i-1} [1 - F(y)]^{n-i} f(y), \text{ där } 0 < y < 1 \quad (3.1)$$

Beräkning vid imputeringar kan göras utifrån betafördelningen.

Låt $Y_{(i)}$ vara den i :te ordningsstatistikan från det uniforma utfallsrummet $U(0,1)$. Då är sannolikhetsfördelningen för $Y_{(i)}$ enligt formel 3.1 som konstaterades i förra stycket men även enligt betafördelningen med parametrarna i och $n-i+1$ (Wikipedia (2016)). Här beräknas väntevärdet för den i :te observationen och tillika det imputerade värdet för en observation enligt (3.2).

$$E(Y_{(i)}) = \frac{i}{n+1}, \text{ där } 1 \leq i \leq n \quad (3.2)$$

i är ordningsnumret (rangen) och n det totala antalet individer, inklusive bortfall, vid ett mättillfälle.

För att åsätta en rang för ett bortfall tillämpas här samma metod som vid LPCF, att utgå från den rang som gällde vid det senaste mättillfället då en observation registrerades. Det är troligt att det är bättre att räkna om ordningarna innan aktuellt mättillfälle men det tillvägagångssättet testas inte i den här uppsatsen. Beräkningen av väntevärdet utifrån betafördelningen innebär en förenklad beräkning och snabbare exekvering i datorprogrammet då formeln är enkel.

Utöver hanteringen av problempunkterna i 3.1.4 tillgodoser LGCF genom ordningsstatistiken även följande:

- Ett betyg som erhållits av en individ med högst rang i en population av storlek n bör ha lägre betyg än en individ som nått högst rang i en population av storlek $n+m$, där $n > 0$ och $m > 0$ ¹⁵.
- Betyget har större varians när det skattats från en liten population och mindre varians när det skattas från större population.

Nedan skildras i tre steg hur imputering enligt LGCF metoden går till vid ett fiktivt fall med sex individer och sex mättillfällen. Exemplet baseras på samma observerade data som i LPCF- och LVCF-exemplen i 3.1.2.

Individ	1	2	3	4	5	6
Mättillfälle 1	3,8	4,1	2,7	3,2	4	4
Mättillfälle 2	3,1	3,5	2,4	3,4	4,3	3,6
Mättillfälle 3	2	3,8	2,9	4	?	3,8
Mättillfälle 4	?	2,4	3,5	4,3	?	4,3
Mättillfälle 5	?	2,8	?	5	?	5,3
Mättillfälle 6	?	3	?	5,2	?	6

Tabell 3.9: Tabellen innehåller exempel på observationer och bortfall. Bortfall indikeras som "?".

Observationer och imputeringar i (0,1) värdemängden.						
Individ	1	2	3	4	5	6
Mättillfälle 1	0,786	1,000	0,000	0,357	0,929	0,929
Mättillfälle 2	0,368	0,579	0,000	0,526	1,000	0,632
Mättillfälle 3	0,000	0,900	0,450	1,000	0,857	0,900
Mättillfälle 4	0,143	0,000	0,579	1,000	0,857	1,000
Mättillfälle 5	0,143	0,000	0,571	0,880	0,857	1,000
Mättillfälle 6	0,143	0,000	0,571	0,733	0,857	1,000

Tabell 3.10: Tabellen innehåller observationer och ordningsstatistik uttryckta som ett värde mellan 0 och 1. Bortfall ersätts med den k :te ordningsstatistikan där k motsvarar ordningen vid det senaste observerade värdet för en individ. Exempel: för individ 1 är $k=1$ för samtliga ordningsstatistikor (mättillfälle 4-6).

¹⁵ En simmare som vinner VM bör ha bättre betyg än simmare som vinner SM då VM innefattar en större population än SM.

Observationer och imputeringar i den ursprungliga värdemängden.						
Individ	1	2	3	4	5	6
Mättillfälle 1	3,8	4,1	2,7	3,2	4,0	4,0
Mättillfälle 2	3,1	3,5	2,4	3,4	4,3	3,6
Mättillfälle 3	2,0	3,8	2,9	4,0	3,7	3,8
Mättillfälle 4	2,7	2,4	3,5	4,3	4,0	4,3
Mättillfälle 5	3,2	2,8	4,2	5,0	4,9	5,3
Mättillfälle 6	3,4	3,0	4,7	5,2	5,6	6,0

Tabell 3.11: Tabellen innehåller de värden som erhålls när ordningsstatistikorna i Tabell 3.5 överförs från värdemängden [0,1] till den ursprungliga värdemängden i Tabell 3.4. Exempel: för individ 1 och mättillfälle 1 är observationen 3,8 som är samma värde som i Tabell 3.4.

3.2 Jämförelse av de starkaste kandidaterna

När LGCF, LVCF och LPCF jämförs i dataprogrammet görs jämförelsen

kvantitativt genom att beräkna hur mycket metodernas imputeringar avviker från de värden som fanns innan bortfallen skapades. Dataprogrammet exekveras på fem olika typer av tidsserier¹⁶: autoregressiv process av första ordningen där $\Phi=0$ (AR(1) $\Phi=0$), autoregressiv process av första ordningen där $\Phi=0,9$ (AR(1) $\Phi=0,9$), autoregressiv process av första ordningen där $\Phi=-0,9$ (AR(1) $\Phi=-0,9$), upprepade Bernoulliförsök (Bernoulli) och fraktionerade upprepade Bernoulliförsök (f. Bernoulli). Dessa exekveringar utförs för att testa hur väl metoderna imputerar på tidsserier som uppträder på skilda sätt som till exempel trendande data eller data med hög standardavvikelse¹⁷. Tidsserierna är indata i programmet i form av dataset som består av 10 individer och 15 mättillfällen för att likna data från longitudinella försök. För att upptäcka skillnader i metodernas förmåga att imputera vid olika bortfallsmekanismer skapas bortfall utifrån MCAR, MAR och MNAR i tidsserierna. Databeräkningarna upprepas med slumpmässiga urval 10000 replikat i syfte att få ett tillräckligt stort underlag för att fånga upp avgörande resultatskillnader som kan uppstå för de olika kombinationerna.

De simulerade bortfallen beskrivs mer utförligt i 3.2.2. Tabell 3.12 nedan visar de 15 olika typerna av tester som utförs av programmet:

Exekveringskonfigurationer			
	MCAR	MAR	MNAR
AR(1) , $\Phi=0$	10000	10000	10000
AR(1) , $\Phi=0,9$	10000	10000	10000
AR(1) , $\Phi=-0,9$	10000	10000	10000
Bernoulli	10000	10000	10000
f. Bernoulli	10000	10000	10000

Tabell 3.12: visar de 15 olika exekveringskonfigurationer som utgörs av alla kombinationer av datatyp och bortfallsmekanism. Varje kombination exekveras 10000 gånger.

¹⁶ Tidsserierna beskrivs i avsnitt 3.2.1

¹⁷ Se figur 3.1

De kriterier som avgör hur väl de olika metoderna presterar är dels förmågan att hantera de olika problemen som nämndes i avsnitt 3.1.4 och dels träffsäkerheten i imputeringarna det vill säga hur kort avstånd det är mellan imputeringarna och de ursprungliga värdena. Avstånden mäts på tre olika sätt: bias, absolut avstånd och kvadrerat avstånd. Kriterier för val av imputeringsmetod beskrivs mer utförligt i avsnitt 3.2.4.

3.2.1 Simulerade tidsserier

De tidsserier som testas är tre varianter av AR(1) (A(1) data) och två varianter av upprepade Bernoulliförsök (Bernoulli data) enligt nedan.

De fem tidsserierna:

- AR(1) $\phi=0$, AR(1) $\phi=0,9$ och AR(1) $\phi=-0,9$: ger en autoregressiv tidsserie där varje individs mätvärde x vid ett mättillfälle t är en linjär kombination av det senaste värdet av processen självt plus en felterm enligt:

$$x_t = c + \phi x_{t-1} + u_t, \quad t = +1, +2, \dots \quad (3.3)$$

Där feltermen, $u_t \sim N(0,1)$. Konstanten c är i jämförelsen fixerad vid 0. ϕ som anger beroendet till x_{t-1} provas med 0, 0,9 och -0,9. Vid $\phi = 0$ fås så kallat 'vitt brus'¹⁸ data som är tidsseriedata utan prediktivt värde. Se figur 3.1 för ett diagram av tidsserierna. Samtliga autoregressiva processer ovan är stationära.

- Upprepade Bernoulli-försök (Bernoulli): Ger en tidsserie där varje individs mätvärde x vid ett mättillfälle t är en linjär kombination av det senaste värdet av processen självt plus en dragning från ett Bernoulliförsök enligt:

$$x_t = x_{t-1} + b_t, \quad t = +1, +2, \dots \quad (3.4)$$

där b_t är en dragning från ett Bernoulli-försök¹⁹ med sannolikhet för lyckat utfall $= p$. Här är $p = 0,6$. Denna tidsserie är den enda av de fem tidsserierna som har en tydlig trend. Se figur 3.1 för ett diagram av tidsserien. Processen är den enda av de fem processerna som inte är stationär utan har en trend.

- Fraktionella upprepade Bernoulli-försök (f. Bernoulli): En datatyp där det finns ett visst beroende mellan tidpunkter även för värden långt ifrån varandra. Här ges varje individs mätvärde x vid ett mättillfälle t av fördelningen:

$$x_t = \frac{(t-1)x_{t-1} + b_t}{t}, \quad t = +1, +2, \dots \quad (3.5)$$

där b_t är en dragning från ett Bernoulli-försök där sannolikheten för lyckat utfall p här är satt till 0,6. Se figur 3.1 för ett diagram av tidsserien. Processen är stationär.

¹⁸ Eng. white noise

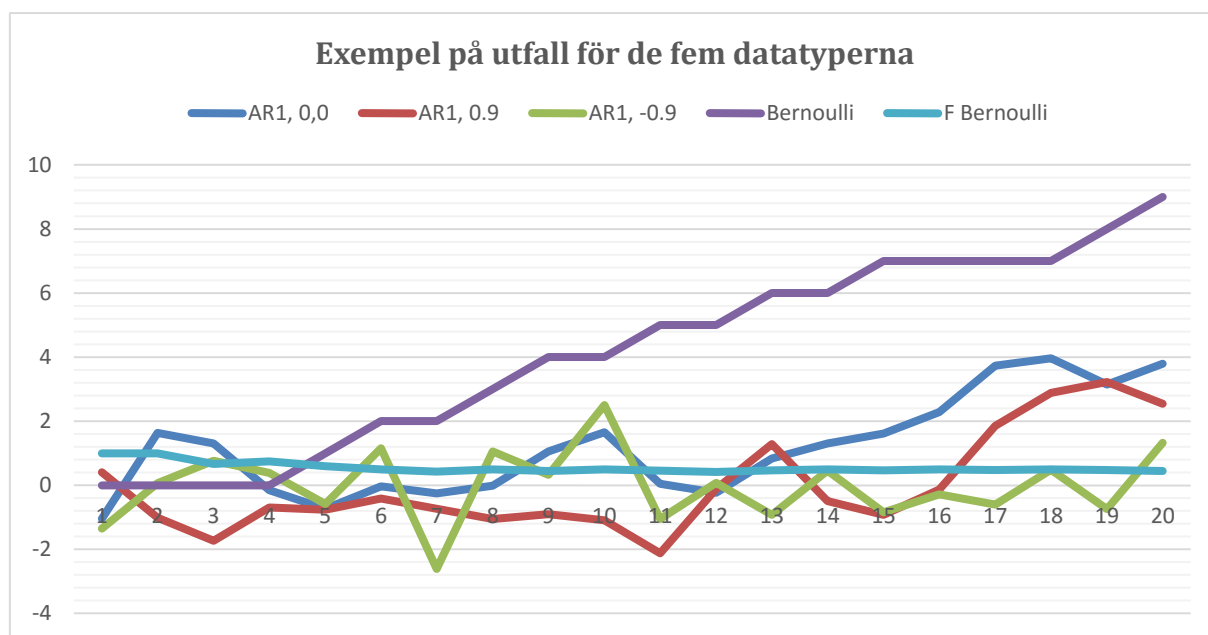
¹⁹ $q=1-p$

För att läsaren ska få en känsla för tidsserierna visas de olika tidsserierna grafiskt i *Figur 3.1* nedan.

Bernoulli ger tydligt trendande data som bör fungera bäst med LGCF och LPCF men sämst med LVCF. F. Bernoulli motsvarar i princip Bernoulli men ger inte upphov till någon trend. Gemensamt för dessa två metoder är att de ger relativt stabila dataserier (*Tabell 3.13 och figur 3.1*). Tidsserier beräknade utifrån AR(1) processerna varierar som synes mer. Hur de olika metoderna hanterar dessa kommer att analyseras i avsnitt 3.2.6. om analys.

Datotyp	Standard avvikelse
AR(1), Φ 0,0	1,51
AR(1), Φ 0.9	1,50
AR(1), Φ -0.9	1,11
Bernoulli	0,17 ²⁰
f. Bernoulli	0,17

Tabell 3.13: Visar standardavvikelse för de olika tidsserierna i diagrammet. Bernoulli och F. Bernoulli visar på lägst variation.



Figur 3.1: Diagrammet visar de fem tidsserierna vilka metoderna testas på. Här framgår att Bernoulli är trendande och att övriga inte är det men uppvisar vissa skillnader i framåtskridandet. Framför allt varierar Bernoulli och f. Bernoulli mindre medan de olika AR1-processerna varierar mer.

²⁰ Här har volatilitet beräknats istället för standardavvikelse för att fånga upp variationen i trenden.

3.2.2 Simulerade bortfall

I jämförelsen mellan de tre LOCF metoderna jämförs även hur stor avvikelse blir vid Rubins olika bortfallsmekanismer: MCAR, MAR och MNAR. I *Tabell 3.14* nedan ges en kort beskrivning av mekanismerna och hur bortfallen utförs för att motsvara mekanismerna.

Benämning	Bortfallsmekanismen	Algoritm för att skapa bortfall
MCAR	Bortfall uppstår oberoende av både observerat och icke-observerat data.	<p>För att bortfall ska vara oberoende allt annat skapas bortfall helt och hållet utifrån slumpen enligt:</p> <p>För varje individ görs ett Bernoulli försök. Om försöket lyckas väljs ett måttillfälle slumpmässigt av alla måttillfällen efter det första. Varje måttillfälle från och med det valda måttillfället blir bortfall.</p>
MAR	Bortfall beror på observerat data men inte på variabeln där bortfallet skett.	<p>För att bortfall endast ska vara beroende av variabler i data men inte variabeln med bortfall så skapas bortfall enligt:</p> <p>För varje individ testas vid varje måttillfälle om en individs rang är högst eller näst högst i det ursprungliga tidsserien (alltså inte efter bortfall). Om så är fallet görs ett Bernoulliförsök. Om försöket lyckas blir alla efterföljande måttillfällen bortfall.</p>
MNAR	Bortfall beror på bortfallsvariabeln själv eller på icke-observerat data.	<p>För att bortfall ska vara beroende av variabeln med bortfallet i eller icke-observerat data så skapas bortfall enligt:</p> <p>Varje individs sista måttillfälle observeras och för varje individ görs ett Bernoulli försök. Ju lägre rang den sista observationen i bortfallsvariabeln har desto större sannolikhet har Bernoulliförsöket att lyckas. Om Bernoulliförsöket lyckas väljs ett slumpmässigt måttillfälle ut. Samtliga måttillfällen efter detta blir bortfall.</p>

Tabell 3.14: beskriver procedurena för att skapa bortfall enligt Rubins bortfallsmekanismer MCAR, MAR och MNAR.

3.2.3 Kriterier för val av imputeringsmetod

Kriterierna för val av metod utgörs dels av de kvantitativa avståndsmätningarna och dels av de kvalitativa problemområdena som har lyfts upp av statistikanalyset. För att mäta avvikelse vid imputeringarna används tre distansmått som mäter skillnaden mellan det ursprungliga värdet som fanns i tidsserierna innan bortfallet och imputeringarna som produceras av de olika metoderna i jämförelseprogrammet för att ersätta bortfallen:

Avståndsmått	Formel ,formel (3.5), (3.6), (3.7)	Kommentar
Bias	$B = \sum_{t=imp1}^{t=impN} \frac{(I_t - S_t)}{n}$ <p>Där B=bias I=imputerat värde t=tid S= ursprungligt värde imp1=första imputeringen impN=sista imputeringen n=antal individer</p>	Om det finns en övervikt negativa alternativt positiva värden finns ett systematiskt fel.
Absolut avstånd ²¹	$A = \sum_{t=imp1}^{t=impN} \frac{ABS(I_t - S_t)}{n}$ <p>Där A=avstånd I=imputerat värde t=tid S= ursprungligt värde imp1=första imputeringen impN=sista imputeringen n=antal individer</p>	Summerar felen för imputeringarna utan hänsyn till om felen var positiva eller negativa.
Kvadratisk avstånd	$KA = \sqrt{\sum_{t=imp1}^{t=impN} \frac{(I_t - S_t)^2}{n}}$ <p>Där KA=kvadratisk avstånd I=imputerat värde t=tid S= ursprungligt värde imp1=första imputeringen impN=sista imputeringen n=antal individer</p>	Visar felens storlek uttryckt som kvadrerat avstånd vilken indikerar variationen i imputeringarna. Inte heller kvadratisk avstånd tar hänsyn till tecken. Kvadratisk avstånd är mer känslig för extrema avstånd mellan imputering och värde innan bortfall.

Tabell 3.16: visar formler för hur de tre avståndsmåten beräknas med kommentar.

Avvikelserna mäts för varje kombination av: Avståndsmått, imputeringsmetod, tidsserie och bortfallsmekanik. Inom varje tidsserie presenteras resultaten i tre grupper, en för vardera bortfallsmekanism: MCAR, MAR, MNAR vilket framgår av resultattabellerna 3.17 till 3.21 nedan.

3.2.4 Resultat

I Tabell.3.17 – 3.21 nedan presenteras resultaten från simuleringarna. Resultaten består av avståndet som imputeringarna ger relativt de ursprungliga värdena. Avstånden mäts i tre olika

²¹ Benämns även MAE, mean absolute error.

avståndsmått: bias, absolut avstånd och kvadratisk avstånd²² enligt Tabell 3.16. i avsnitt 3.2.3. Mätningarna har utförts vid 10000 försök vilka var och en består av 15 individer och 10 mättillfällen.

Resultat 1. Data från en AR(1) process där , $\phi = 0,9$.

MCAR	LPCF	LGCF	LVCF
bias	-0.0127	-0.08	-0.01
bias std. dev.	0.0046	0.0046	0.0034
square distance	1.0551	1.0251	0.9107
square distance std. dev.	0.0068	0.0065	0.0057
absolute distance	0.4926	0.4829	0.4231
absolute distance std. dev.	0.0037	0.0037	0.0031

MAR	LPCF	LGCF	LVCF
bias	0.2596	0.1613	0.1575
bias std. dev.	0.0041	0.0036	0.003
square distance	0.9371	0.8511	0.732
square distance std. dev.	0.0079	0.007	0.0061
absolute distance	0.3889	0.3509	0.3026
absolute distance std. dev.	0.0037	0.0033	0.0028

MNAR	LPCF	LGCF	LVCF
bias	0.126	0.1234	0.0922
bias std. dev.	0.003	0.0028	0.0024
square distance	0.6427	0.6029	0.5373
square distance std. dev.	0.0064	0.006	0.0054
absolute distance	0.2507	0.235	0.2086
absolute distance std. dev.	0.0028	0.0026	0.0023

Tabell 3.17, Resultatet är från simulering där data skapas med en AR(1) process med autokorrelation, $\phi = 0,9$.

Resultat 2. Data från en AR(1) process där , $\phi = 0$.

MCAR	LPCF	LGCF	LVCF
bias	0.0016	-0.0278	0.0003
bias std. dev.	0.0025	0.0023	0.0022
square distance	0.6045	0.5483	0.6006
square distance std. dev.	0.0037	0.0033	0.0037
absolute distance	0.2829	0.2563	0.2804
absolute distance std. dev.	0.002	0.0018	0.002

²² Eng. squared distance

MAR	LPCF	LGCF	LVCF
bias	0.2698	0.1781	0.3082
bias std. dev.	0.0027	0.0022	0.0025
square distance	0.6892	0.5817	0.7341
square distance std. dev.	0.0043	0.0036	0.0043
absolute distance	0.3255	0.269	0.3459
absolute distance std. dev.	0.0024	0.0019	0.0024

MNAR	LPCF	LGCF	LVCF
bias	0.0883	0.0813	0.0716
bias std. dev.	0.0017	0.0015	0.0015
square distance	0.3772	0.3407	0.3617
square distance std. dev.	0.0036	0.0032	0.0035
absolute distance	0.1476	0.1324	0.1408
absolute distance std. dev.	0.0016	0.0014	0.0015

Tabell 3.18, Resultatet är från simulering där data skapas med en AR(1) process med autokorrelation, $\phi = 0,0$ (white noise).

Resultat 3. Data från en AR(1) process där , $\phi = -0,9$.

MCAR	LPCF	LGCF	LVCF
bias	-0.0031	-0.0691	-0.0003
bias std. dev.	0.0057	0.0053	0.0047
square distance	1.3623	1.2395	1.255
square distance std. dev.	0.0086	0.0077	0.0084
absolute distance	0.6338	0.5788	0.5779
absolute distance std. dev.	0.0047	0.0042	0.0044

MAR	LPCF	LGCF	LVCF
bias	0.5623	0.3899	0.4299
bias std. dev.	0.0059	0.005	0.0045
square distance	1.4938	1.2884	1.2562
square distance std. dev.	0.0096	0.0081	0.0081
absolute distance	0.6916	0.5869	0.5715
absolute distance std. dev.	0.0054	0.0043	0.0044

MNAR	LPCF	LGCF	LVCF
bias	0.2005	0.1817	0.1602
bias std. dev.	0.0039	0.0034	0.0033
square distance	0.8362	0.7579	0.7366
square distance std. dev.	0.0083	0.0074	0.0075
absolute distance	0.3272	0.2956	0.2847

absolute distance std. dev.	0.0036	0.0032	0.0031
-----------------------------	--------	--------	--------

Tabell 3.19, Resultatet är från simulering där data skapas med en AR(1) process med autokorrelation, $\phi = -0,9$.

Resultat 4. Data från upprepade Bernoulliförsök där $p=0,6$.

MCAR

	LPCF	LGCF	LVCF
bias	0.0238	-0.3317	-1.1279
bias std. dev.	0.0034	0.0042	0.0076
square distance	0.7611	0.9937	2.3328
square distance std. dev.	0.0052	0.0062	0.0123
absolute distance	0.3489	0.4766	1.1279
absolute distance std. dev.	0.0028	0.0036	0.0076

MAR

	LPCF	LGCF	LVCF
bias	0.1123	-0.4515	-2.1101
bias std. dev.	0.0052	0.0049	0.0118
square distance	1.2831	1.215	4.0264
square distance std. dev.	0.0075	0.0074	0.0153
absolute distance	0.603	0.5824	2.1101
absolute distance std. dev.	0.0043	0.0043	0.0118

MNAR

	LPCF	LGCF	LVCF
bias	0.0927	-0.0268	-0.4824
bias std. dev.	0.0023	0.0021	0.0049
square distance	0.4715	0.4131	1.2127
square distance std. dev.	0.005	0.0041	0.0107
absolute distance	0.1818	0.162	0.4824
absolute distance std. dev.	0.0021	0.0018	0.0049

Tabell 3.20, Resultatet är från simulering där data skapas med upprepade Bernoulliförsök, $p=0,6$.

Resultat 5. Data från fraktionella upprepade Bernoulliförsök där $p=0,6$.

MCAR

	LPCF	LGCF	LVCF
bias	0.0016	-0.0221	0.0004
bias std. dev.	0.0002	0.0003	0.0003
square distance	0.0507	0.0662	0.0773
square distance std. dev.	0.0003	0.0004	0.0006
absolute distance	0.0233	0.0318	0.034
absolute distance std. dev.	0.0002	0.0002	0.0003

MAR

	LPCF	LGCF	LVCF
bias	0.0075	-0.0301	-0.0001
bias std. dev.	0.0003	0.0003	0.0007
square distance	0.0855	0.081	0.2181
square distance std. dev.	0.0005	0.0005	0.0009
absolute distance	0.0402	0.0388	0.1122
absolute distance std. dev.	0.0003	0.0003	0.0007

MNAR

	LPCF	LGCF	LVCF
bias	0.0062	-0.0018	0.0
bias std. dev.	0.0002	0.0001	0.0002
square distance	0.0314	0.0275	0.0452
square distance std. dev.	0.0003	0.0003	0.0005
absolute distance	0.0121	0.0108	0.017
absolute distance std. dev.	0.0001	0.0001	0.0002

Tabell 3.21, Resultatet är från simulering där data skapas med fraktionella upprepade Bernoulliförsök, $p=0,6$.

3.2.5 Analys

Målet med analysen är att ta reda på vilken av LOCF varianterna som är lämpligast för imputeringsdataprogrammet utifrån de kvantitativa resultaten som presenterades i förra avsnittet och de kvalitativa problemområden som beskrevs i avsnitt 3.1.4.

Kvantitativ jämförelse utifrån bortfallsmekanism:

Diagram 3.2 till 3.4 nedan visar kvoten mellan de förväntade avstånden, mätt i absolut avstånd och kvadrerat avstånd för LGCF/LPCF respektive LGCF/LVCF. Värden större än ett indikerar alltså att LGCF ger större avstånd (sämre träffsäkerhet) än LPCF respektive LVCF. I diagrammen framgår att metoderna presterar olika bra avseende bortfallsmekanism.

Eftersom kurvorna varierar markant mellan diagrammen kan fastställas att bortfallsmekanism påverkar metodernas träffsäkerhet. Tre exempel på skillnad: LVCF ger sämre träffsäkerhet än LGCF med Bernoulli och f. Bernoulli data vid samtliga bortfallsmekanismer; LGCF ger sämre träffsäkerhet än LPCF med Bernoulli och f. Bernoulli data vid MCAR och LGCF ger något bättre träffsäkerhet än LPCF vid MNAR.

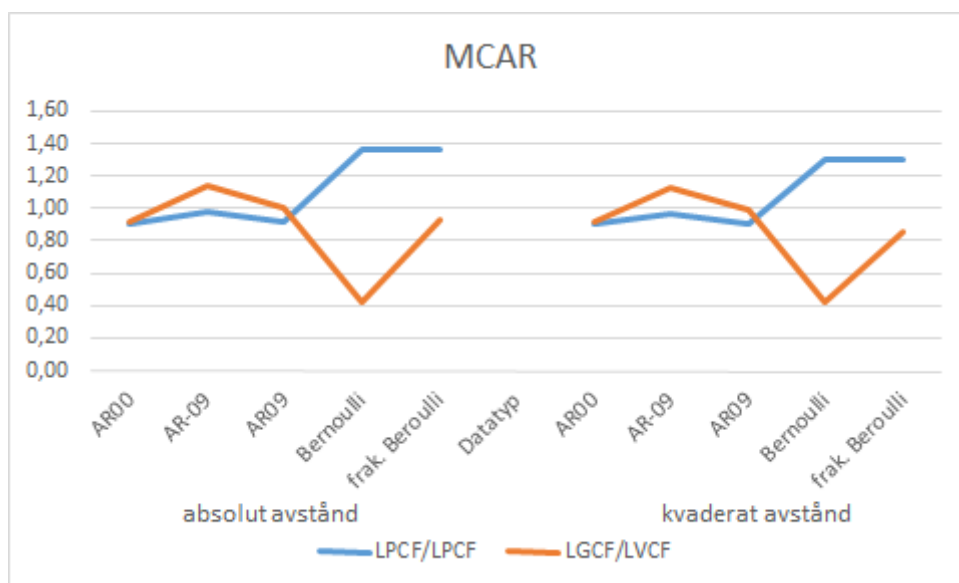


Diagram 3.2: Avstånd för metoderna LGCF och LPCF respektive LVCF över datatyper. Y-axeln visar kvoten mellan metodernas träffsäkerhet vid resultat från de olika datatyperna vid MCAR-bortfall.

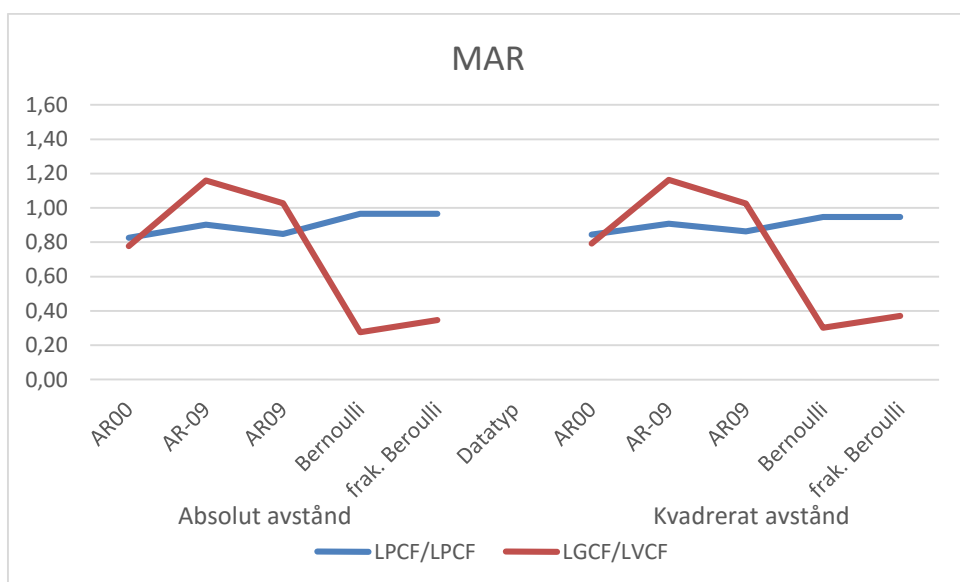


Diagram 3.3: Motsvarande jämförelse som i Diagram 3.2 med vid MAR-bortfall. kvoterna liknar de vid MCAR-bortfall.

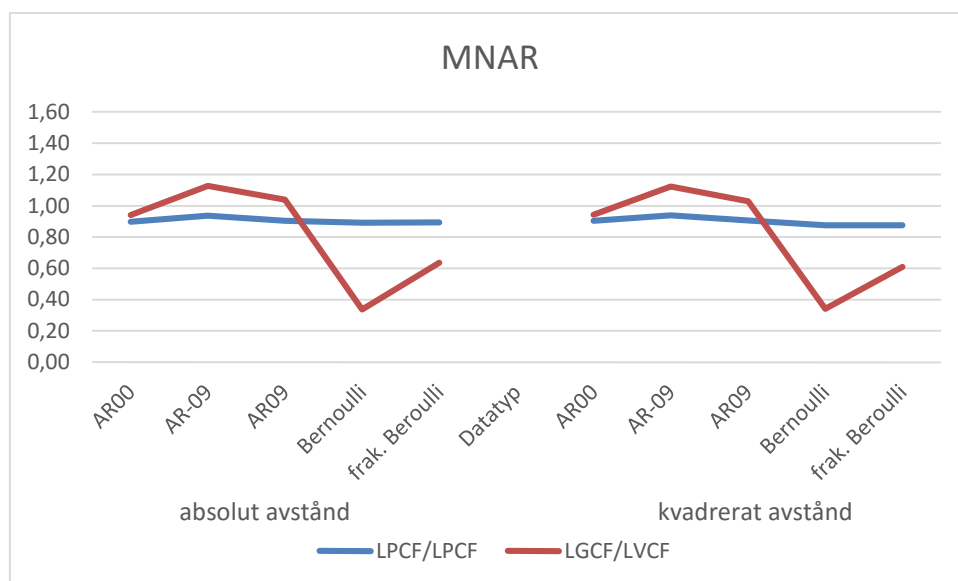


Diagram 3.4: Motsvarande jämförelse som i Diagram 3.2 men vid MNAR-bortfall. Här ger jämförelsen liknande skillnader mellan metoderna som vid MCAR och MAR-bortfall men är inte lika markanta.

Kvantitativ jämförelse utifrån datatyp:

Tabellerna 3.22-3.24 nedan visar genomsnittliga avstånd uppmätt som bias, absolut avstånd och kvadrerat avstånd vid kombinationerna datatyp och bortfallsmekanik.

Mätningen av bias, eller systematiska fel, i Tabell 3.22 visar att metoderna är relativt jämbördiga avseende bias men både LVCF och LPCF kan ge relativt extrema resultat till skillnad från LGCF som inte vid något fall ger mest bias.

Bias sammanställningen visar också att samtliga metoder ger lägst bias vid MCAR. Skälet till det är att MCAR-bortfall är helt slumpmässig och därför inte borde ge upphov till bias .

Bias

Avståndsmått	Datatyp	Bortfallsmekanism	LPCF	LGCF	LVCf
bias	AR(1),0,0	MCAR	0,0016	-0,0278	0,0003
bias	AR(1),0,0	MAR	0,2698	0,1781	0,3082
bias	AR(1),0,0	MNAR	0,0883	0,0813	0,0716
bias	AR(1),0,9	MCAR	-0,0127	-0,08	-0,01
bias	AR(1),0,9	MAR	0,2596	0,1613	0,1575
bias	AR(1),0,9	MNAR	0,126	0,1234	0,0922
bias	AR(1),-0,9	MCAR	-0,0031	-0,0691	-0,0003
bias	AR(1),-0,9	MAR	0,5623	0,3899	0,4299
bias	AR(1),-0,9	MNAR	0,2005	0,1817	0,1602
bias	Bernoulli	MCAR	0,0238	-0,3317	-1,1279
bias	Bernoulli	MAR	0,1123	-0,4515	-2,1101
bias	Bernoulli	MNAR	0,0927	-0,0268	-0,4824
bias	f. Bernoulli	MCAR	0,0016	-0,0221	0,0004
bias	f. Bernoulli	MAR	0,0075	-0,0301	-0,0001
bias	f. Bernoulli	MNAR	0,0062	-0,0018	0

Tabell 3.22: Tabellen visar bias uppdelat på bortfallsmekanism, Datatyp och imputeringsmetoderna. Starkt positiva eller starkt negativa värden för bias kan bero på systematiskt fel i imputeringen.

Mätningen av absolut avstånd och kvadrerat avstånd analyseras här samtidigt då måtten är likartade. Skillnaden mellan dessa avståndsmått är att kvadrerat avstånd straffar extrema värden mer än absolut avstånd eftersom skillnader mellan imputerat värde och ursprungliga värden kvadreras.

För att upptäcka fler eller färre förekomster av extrema värden mäts här kvoten mellan absolut avstånd och kvadrerat avstånd (Tabell 3.25). Genomsnittliga kvoten är här 0.44. Avvikelser upp från 0.44 tyder på färre fall av extremt dålig träffsäkerhet och avvikelser ner från 0.44 tyder på fler fall av extremt dålig träffsäkerhet. Resultatet i Tabell 3.15 är färgkodad så att mörkare grön visar på starkare avvikelser upp och mörkare röd visar på starkare avvikelser ner. Kvoterna varierar inte mycket men man kan ändå konstatera att antalet extrema avvikelser tycks förekomma oftare vid data med MNAR-bortfall än vid andra bortfallsmekanismer. LVCf tycks särskilt drabbat av extrema avvikelser då de dessutom förekommer vid Bernoullidata och f. Bernoullidata.

Förekomst av extrema avstånd mellan imputering och ursprungligt värde.

Kvot	Datatyp	Bortfallsmekanism	LPCF	LGCF	LVCF
absolute distance / squared dis	AR(1),0,0	MAR	0,472287	0,462438	0,4711892
absolute distance / squared dis	AR(1),0,0	MCAR	0,46799	0,467445	0,4668665
absolute distance / squared dis	AR(1),0,0	MNAR	0,391304	0,388612	0,3892729
absolute distance / squared dis	AR(1),0,9	MAR	0,415004	0,41229	0,413388
absolute distance / squared dis	AR(1),0,9	MCAR	0,466875	0,471076	0,4645877
absolute distance / squared dis	AR(1),0,9	MNAR	0,390073	0,389783	0,3882375
absolute distance / squared dis	AR(1),-0,9	MAR	0,46298	0,455526	0,4549435
absolute distance / squared dis	AR(1),-0,9	MCAR	0,465243	0,466962	0,4604781
absolute distance / squared dis	AR(1),-0,9	MNAR	0,391294	0,390025	0,3865056
absolute distance / squared dis	Bernoulli	MAR	0,469956	0,479342	0,5240662
absolute distance / squared dis	Bernoulli	MCAR	0,458415	0,479622	0,4834962
absolute distance / squared dis	Bernoulli	MNAR	0,385578	0,392157	0,3977901
absolute distance / squared dis	f. Bernoulli	MAR	0,470175	0,479012	0,5144429
absolute distance / squared dis	f. Bernoulli	MCAR	0,459566	0,480363	0,4398448
absolute distance / squared dis	f. Bernoulli	MNAR	0,38535	0,392727	0,3761062

Tabell 3.25: Tabellen visar på förekomster av extrema avstånd genom att presentera kvoten mellan absolut avstånd och kvadrerat avstånd. Extrema avstånd förekommer i högre grad vid mörkröda celler som visar lägre värden än genomsnittet (0.45) beroende på högre värde på kvadrerat avstånd relativt absolut avstånd.

Efter analysen av skillnaden mellan absolut avstånd och kvadrerat avstånd räcker det med att analysera det ena av de två avståndsmåtten. Valet faller på absolut avstånd vars sammanställda resultat visas i *Tabell 3.23*. Inte oväntat ger LVCF stora avstånd vid Bernoulli data beroende på att LVCF imputeringarna görs med det senast uppmätta värdet det vill säga om de observerade värdena skulle trendat upp eller ner så ökar felen i imputeringarna vid varje ytterligare mättillfälle. Det kan även konstateras att LVCF och LGCF tenderar imputera med större träffsäkerhet vid AR(1) data än LPCF. Lägsta avstånden återfinns för körningar med f. Bernoulli data. Det gäller för alla bortfallsmekanismer. I övrigt ger metoderna likartade resultat med avseende på datatyp.

Absolut avstånd

Avståndsmått	Datatyp	Bortfallsmekanism	LPCF	LGCF	LVCF
absolute distance	AR(1),0,0	MAR	0,3255	0,269	0,3459
absolute distance	AR(1),0,0	MCAR	0,2829	0,2563	0,2804
absolute distance	AR(1),0,0	MNAR	0,1476	0,1324	0,1408
absolute distance	AR(1),0,9	MAR	0,3889	0,3509	0,3026
absolute distance	AR(1),0,9	MCAR	0,4926	0,4829	0,4231
absolute distance	AR(1),0,9	MNAR	0,2507	0,235	0,2086
absolute distance	AR(1),-0,9	MAR	0,6916	0,5869	0,5715
absolute distance	AR(1),-0,9	MCAR	0,6338	0,5788	0,5779
absolute distance	AR(1),-0,9	MNAR	0,3272	0,2956	0,2847
absolute distance	Bernoulli	MAR	0,603	0,5824	2,1101
absolute distance	Bernoulli	MCAR	0,3489	0,4766	1,1279
absolute distance	Bernoulli	MNAR	0,1818	0,162	0,4824
absolute distance	f. Bernoulli	MAR	0,0402	0,0388	0,1122
absolute distance	f. Bernoulli	MCAR	0,0233	0,0318	0,034
absolute distance	f. Bernoulli	MNAR	0,0121	0,0108	0,017

Tabell 3.23: Tabellen visar absolut avstånd uppdelat på bortfallsmekanism, datatyp och de olika imputeringsmetoderna. Stora värden för absolut absolute distance betyder dålig träffsäkerhet vid imputeringarna. Absolut avstånd påverkas i mindre grad än kvadrerat avstånd av extrema värden.

Avstånd i kvadrat

Avståndsmått	Datatyp	Bortfallsmekanism	LPCF	LGCF	LVCF
squared distance	AR(1),0,0	MAR	0,6892	0,5817	0,7341
squared distance	AR(1),0,0	MCAR	0,6045	0,5483	0,6006
squared distance	AR(1),0,0	MNAR	0,3772	0,3407	0,3617
squared distance	AR(1),0,9	MAR	0,9371	0,8511	0,732
squared distance	AR(1),0,9	MCAR	1,0551	1,0251	0,9107
squared distance	AR(1),0,9	MNAR	0,6427	0,6029	0,5373
squared distance	AR(1),-0,9	MAR	1,4938	1,2884	1,2562
squared distance	AR(1),-0,9	MCAR	1,3623	1,2395	1,255
squared distance	AR(1),-0,9	MNAR	0,8362	0,7579	0,7366
squared distance	Bernoulli	MAR	1,2831	1,215	4,0264
squared distance	Bernoulli	MCAR	0,7611	0,9937	2,3328
squared distance	Bernoulli	MNAR	0,4715	0,4131	1,2127
squared distance	f. Bernoulli	MAR	0,0855	0,081	0,2181
squared distance	f. Bernoulli	MCAR	0,0507	0,0662	0,0773
squared distance	f. Bernoulli	MNAR	0,0314	0,0275	0,0452

Tabell 3.24: Tabellen visar kvadrerat avstånd uppdelat på bortfallsmekanism, datatyp och de olika imputeringsmetoderna. Stora värden för kvadrerat avstånd betyder dålig träffsäkerhet vid imputeringarna. Kvadrerat avstånd påverkas i högre grad av extrema värden än absolut värde.

3.2.6 Slutsats

Slutsatserna nedan är indikativa och inte statistiskt säkerställda.

Den kvantitativa jämförelsen visar att ingen av imputeringsmetoderna bör användas för alla typer av data eller alla bortfallsmekanismer. Istället framgår av jämförelsen att metod ska väljas utifrån hur data ser ut. Innan val av metod bör frågor liknande dessa ställas: Är det longitudinella kliniska data volatilt eller inte? Finns en trend eller inte? Vilken eller vilka bortfallsmekanismer förekommer? Nedan visas en sammanställning över några relativt tydliga samband som kan utläsas ur Diagram 3.2 till 3.4 och Tabell 3.22 till 3.24. Vid ett faktiskt fall skulle fokus ligga på de samband som är intressanta vilket gör en analys enklare.

- LGCF ger relativt låg bias då imputering utförs på data med bortfall som uppstått enligt MAR.
- LVCF ger relativt hög bias då imputering utförs på data med låg volatilitet (simulering från en AR(1) process).
- LVCF och LGCF ger låg bias då imputering utförs på data med hög volatilitet (simulering från en AR(1) process).
- LPCF och LGCF bättre resultat än LVCF då imputering utförs på data med en trend.
- LGCF är bättre än LPCF då imputering utförs på data med bortfall som uppstått enligt MNAR.
- Extremt felaktiga imputeringar förekommer mest vid data med bortfall som uppstått enligt MNAR. Det gäller samtliga metoder.

I den kvalitativa bedömningen framgår att endast LVCF kan hantera samtliga problemområden i avsnitt 3.1.4 utan besvärande approximationer eller genom att offra data. LGCF och LPCF hanterar två av de tre problemen det vill säga fallen där flera individer intill varandra fallit bort och då en av två individer med samma observerade värden hoppar av.

Utifrån jämförelsen ovan och de kriterier som gäller för val av metod enligt avsnitt 3.2.3 bör metod väljas från fall till fall utifrån hur det longitudinella kliniska datat uppför sig avseende volatilitet, trender och bortfallsmekanismer. Om endast ett av alternativen måste väljas framstår LGCF som den bästa kompromissen då LGCF inte vid någon av jämförelserna ger sämst träffsäkerhet och ofta ger bäst träffsäkerhet. Dessutom hanterar LGCF två av de tre kvantitativa problemområdena i avsnitt 3.1.4.

3.3 Fortsatta studier

Studien av den nya metoden LGCF i det här arbetet har visat att LGCF står sig väl i jämförelse med LPCF och LVCF vilket gör att fortsatta studier av LGCF kan vara befogad. I 3.3.1 nedan ges exempel på fortsatta studier som kan vara intressanta. I 3.3.2 ges en kort beskrivning av de kommande momenten rörande programmet som ska programmeras åt statistikanalysföretaget.

3.3.1 Metodjämförelserna

- Även titta på andra sätt att vid LGCF applicera ett ordningsnummer för ett bortfall, till exempel: istället för att låta individen ärva rangen från förra mättillfället så kan individerna, baserat på observationer och imputeringar, rangordnas på nytt vid varje mättillfälle.

- Jämföra med andra metoder än LOCF som till exempel multipel imputering.
- En djupare analys är lämplig för att svara på frågor som inte besvaras i här. Till exempel: Varför uppstår större bias med metoderna LPCF och LVCF än med LGCF vid MAR-bortfall?
- Fastställa imputeringskvalitén statistiskt eventuellt i jämförelse med alternativa metoder.

3.3.2 Färdigställa programmet

Färdigställa ett program enligt:

- koden felhanteras och granskas ur ett prestandaperspektiv.
- Fastställa användar- och datagränssnittet.
- Paketera programmet som ett exekverbart program.

Litteraturförteckning

- [1] Allison, P.D. (2001). Missing data.
- [2] Birmingham, J. Fitzmaurice, G.M. Rotnitzky, A. (2003). Pattern: Mixture and Selection Models for Analysing Longitudinal Data with Monotone Missing Patterns.
- [3] Deng. (2012). <http://onbiostatistics.blogspot.se/2012/05/is-last-observation-carried-forward.html>
- [3] Deng. (2015). <http://onbiostatistics.blogspot.se/2015/08/missing-data-assumptions-and.html>.
- [4] Figueredo, A.J. McKnight, P.E. McKnight, K.M. Sidani, S. (2007). Missing data: a gentle introduction.
- [5] Fitzmaurice, G.M., James H., Nan M. Laird, N.M. (2011). Applied Longitudinal Analysis
- [6] Friedman, L. Furberg, C. DeMets, D. (2010). Fundamentals of clinical trials, 4th edition, kap. 17.
- [7] International conference on harmonization of technical requirements for registration of pharmaceuticals for human use, expert working group (1998). ICH Harmonised tripartite guideline, Statistical principles for clinical trials E9.
- [8] Little, R. Rubin, D.B. (2002). Statistical Analysis with missing data.
- [9] Mallinckrodt, C.H. (2013). Preventing and Treating Missing Data in Longitudinal Clinical Trials, A Practical Guide.
- [10] Rubin, D.B. (1976). Inference and missing data. Biometrika 63, 581-592.
- [11] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys.
- [12] Van Buuren, S. (2012). Flexible imputation of missing data.
- [13] Wackerly D., Mendenhall, W., Scheaffer, R.L. (2008). Mathematical statistics with applications, 7:th edition, kap 6.7.
- [14] Wikipedia (2015). https://sv.wikipedia.org/wiki/Klinisk_pr%C3%B6vning
- [15] Wikipedia (2016). [https://en.wikipedia.org/wiki/Uniform_distribution_\(continuous\)](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)).
- [16] Xiao-Hua, Z., Chuan. Z., Danping, L. (2014). Statistics in Practice : Applied Missing Data Analysis in the Health Sciences”.

Deng. (2012). <http://onbiostatistics.blogspot.se/2012/05/is-last-observation-carried-forward.html>

Bilagor

Bilaga 1, Lämpliga metoder beroende på bortfallsantaganden²³.

No assumption	MCAR	MAR	MNAR
	Missing Complete at Random	Missing at Random – ignorability assumption	Missing Not at Random
	<p>The missingness is independent of both unobserved and observed data.</p> <p>The probability of missingness is the same for all units.</p>	<p>Conditional on the observed data, the missingness is independent of the unobserved measurements.</p> <p>The probability a variable is missing depends only on available information.</p>	<p>Not MCAR or MAR.</p> <p>Missingness that depends on unobserved predictors.</p> <p>Missingness is no longer at random if it depends on information that has not been recorded and this information also predicts the missing values.</p> <p>Missingness that depends on the missing value itself</p>
	Ignorable	Ignorable	Non-ignorable
<p>LOCF (last observation carried forward)</p> <p>BOCF (baseline value carried forward)</p> <p>WOCF (worst observation carried forward)</p> <p>Imputation based on logical rules</p>	<p>CC (Complete-case Analysis) - listwise deletion</p> <p>Pairwise Deletion</p> <p>Available Case analysis</p> <p>Single-value Imputation (for example, mean replacement, regression prediction (conditional mean imputation), regression prediction plus error (stochastic regression imputation))</p> <p>– under MCAR, throwing out cases with missing data does not bias your inferences. However, there are many drawbacks</p>	<p>Maximum Likelihood using the EM algorithm – FIML (full information maximum likelihood)</p> <p>MMRM (mixed model repeated measurement) – REML (restricted maximum likelihood)</p> <p>Multiple Imputation</p> <p>Two assumptions: the joint distribution of the data is multivariate normal and the missing data mechanism is ignorable</p> <p>Under MAR, it is acceptable to exclude the missing cases, as long as the regression controls for all the variables that affect the probability of missingness</p>	<p>PMM (Pattern-mixture modeling)</p> <p>Jump to Reference</p> <p>Last Mean Carried Forward.</p> <p>Copy Differences in Reference Copy Reference</p> <p>Tipping Point Approach</p> <p>Selection model (Heckman)</p>

²³ Källa: Dr. Deng (2015)

Bilaga 2, Multipel imputering (MI)

Multipel imputering har många fördelar som att den kan, om den implementeras korrekt, användas med i princip all typ av data. Den ger också estimat som är konsistenta, effektiva och asymptotiskt normala när bortfall är MAR²⁴. Det finns även ett stort utbud av mjukvara vilket i vissa fall kan reducera de möjliga problemen eller nackdelarna med MI vilka är att metoden kan vara komplicerad att tillämpa och att det är lätt implementera MI felaktigt.

Metoden togs fram av Donald B. Rubin²⁵ som beskriver MI som en process i tre steg enligt:

1. *Imputering*: Här simuleras imputerade värden n genom att imputeringar dras från en lämplig fördelning n gånger så att n dataset med möjliga imputerade värden skapas.
2. *Analys*: De simulerade dataset som skapades i steg 1 analyseras med så kallad complete-data analysis för att beräkna statistiska för de imputerade dataseten.
3. *Pooling*: De n analyserna kombineras för att åter få *ett* dataset men som nu innehåller imputerat data. De resulterande imputeringarna ger med denna metod möjligheten att mäta till exempel imputeringarnas standardfel och varians.

Steg 1 är det mest utmanande och samtidigt viktigaste steget vid multipel imputering. Donald B. Rubin har fastställt att metoden är statistiskt korrekt förutsatt att steg 1, det vill säga metoden för att skapa de n imputerade tidsserierna är korrekta (Rubin 1987).

Även om Multipel imputering är lämplig för alla typer av missing data kräver MAR och MNAR att steg 1 genomförs med hänsyn till datats och bortfallens beskaffenhet. Om det till exempel finns ett beroende mellan orsaken till bortfallet och det som undersöks behöver imputeringssteget, det vill säga steg 1 ta hänsyn till det.

De imputerade värdena hämtas ur en lämplig fördelning som ofta skapas i en Bayesiansk process. De imputerade värdena kan vara olika vid olika mättillfällen och för olika försökspersoner.

Steg 2 innebär ett antal analyser eftersom varje dataset ska analyseras i avsikt att kombinera de n tidsserierna statistisk korrekt. Detta steg genomförs normalt med standardmetoder i de mjukvaruprogram som kan hantera multipel imputering.

I Steg steg 3 beräknas statistikor som medelvärde och varians för att finna de imputerade värdena och är det enklaste steget i multipel imputering.

Det är viktigt att modellen för multipel imputering kontrolleras i efterhand för att undvika felaktiga modeller och för att eventuellt förbättra modellen.

²⁴ Allison 2001

²⁵ Rubin D.B. 1987