Stockholms universitet

Statistisk modellering av stora och små skador: En vägledning för beslut om återförsäkring för XL-kontrakt

Esma Özdolap

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2025:5
Försäkringsmatematik
Juni 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2025:5**
http://www.math.su.se/matstat

Statistisk modellering av stora och små skador: En vägledning för beslut om återförsäkring för XL-kontrakt

Esma Özdolap*

Juni 2025

Sammanfattning

Storskador inträffar sällan, men när de väl gör det kan de innebära ekonomiska konsekvenser för försäkringsbolaget. Därför spelar valet av självbehållsnivå i ett återförsäkringsavtal en central roll för hur försäkringsbolaget på sikt vill hantera sina risker. Syftet med denna uppsats är att analysera och studera hur försäkringsbolag kan resonera vid beslut om självbehållsnivå för att hantera risker och kostnader. Genom att modellera skadebelopp baserat på verklig egendomsförsäkringsdata tillhandahållet av Länsförsäkringar AB, med särskild uppdelning mellan stora och små skador, möjliggörs en mer nyanserad analys av skadeutfallen. De stora skadorna modelleras med hjälp av en Paretofördelning och genom att tillämpa extremvärdesteori, främst genom den generaliserade Paretofördelningen (GPD) via Peaks Over Threshold (POT)-metoden. Antalet stora skador modelleras separat från skadebeloppen med hjälp av Poisson- eller negativ binomialfördelning. Skadebeloppen under brytpunkten, det vill säga de mindre skadebeloppen aggregeras årsvis och modelleras med klassiska fördelningar såsom normal-, lognormal- och gammafördelning. Uppsatsen använder både maximum likelihood- och momentmetoden för att skatta parametrar i olika fördelningar. Modellernas lämplighet bedöms med hjälp av goodness-of-fit-tester såsom Kolmogorov–Smirnov-och Anderson–Darling-test, samt jämförelsemått som Akaikes informationskriterium (AIC) och Bayesianskt informationskriterium (BIC). En årlig total skadekostnad estimeras genom simuleringar. Baserat på simulerat data utvärderas hur olika självbehållsnivåer påverkar försäkringsbolagets egna skadekostnader under ett skadeår. Resultaten visar att det är viktigt var gränsen dras mellan stora och små skador, eftersom det påverkar hur tillförlitliga modellerna blir. Resultaten visar även att den genomsnittliga kostnaden är relativt opåverkad av de självbehållsnivåer som har undersökts, även vid ett större antal simuleringar. Däremot syns en viss ökning i variation, särskilt i de högre percentilerna (95% och 99%), när självbehållen ökar. Det tyder på att ett högre självbehåll kan bidra till ett mer oförutsägbart skadeutfall vid extrema skadeår, även om den genomsnittliga skadekostnaden förblir i stort sett oförändrad. Sammanfattningsvis ger uppsatsen ett praktiskt och teoretiskt verktyg som kan hjälpa ett försäkringsbolag att förstå och fatta välgrundade beslut om återförsäkring, särskilt vad gäller valet av självbehållsnivå.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: esma2001@hotmail.se. Handledare: Kristofer Lindensjö.

Abstract

Large claims occur infrequently, but when they do, they can have a significant financial impact on an insurance company. Therefore, the choice of retention level in a reinsurance agreement plays a crucial role in how the company manages its risks over time. The purpose of this thesis is to analyze and explore how insurance companies may reason when determining an appropriate retention level to manage both risk and cost. By analyzing claim outcomes and their associated costs using real property insurance data provided by Länsförsäkringar AB, the claims are modeled with a specific distinction between large and small claims.

Large claims are modeled using a Pareto distribution and by applying extreme value theory, primarily through the Generalized Pareto Distribution (GPD) via the Peaks Over Threshold (POT) method. The number of large claims is modeled separately from the claim amounts using Poisson and negative binomial distributions. Smaller claims are aggregated on an annual basis and modeled using classical distributions such as the normal, lognormal, and gamma distributions.

The thesis employs both the maximum likelihood method and the method of moments to estimate model parameters. The adequacy of the models is assessed using goodness-of-fit tests such as the Kolmogorov–Smirnov and Anderson–Darling tests, as well as evaluation criteria like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). An annual total claims cost is estimated through simulations.

Based on the simulated data, the impact of different retention levels on the insurer's net claim costs over a claims year is evaluated. The results indicate that the threshold between large and small claims is crucial, as it influences the reliability of the models. The findings also show that the average cost is relatively unaffected by the retention levels studied, even with a large number of simulations. However, an increase in variability, particularly in the higher percentiles (95% and 99%), is observed as the retention increases. This suggests that a higher retention level may lead to more unpredictable claim outcomes in extreme claims years, even if the average cost remains largely unchanged.

In summary, the thesis provides a practically and theoretically grounded framework that can help an insurance company understand and make informed decisions about reinsurance, especially regarding the choice of retention level.

Keywords: Reinsurance, Excess of loss, Extreme value theory, Pareto distribution, Generalized Pareto Distribution, Peaks Over Threshold model, Anderson–Darling test, Kolmogorov–Smirnov test, Lognormal distribution, Maximum likelihood.

Förord

Jag vill börja med att tacka mina kära föräldrar för deras stöd under hela arbetets gång. Jag vill även uttrycka min tacksamhet till mina externa handledare, David Lam och Henrik Vallgren (Länsförsäkringar AB), för deras engagemang, vägledning och värdefulla synpunkter. Ett särskilt tack riktas också till Louise Lindström (Guy Carpenter) för hennes tid, stöd och kloka råd. Till sist vill jag rikta ett stort tack till Kristofer Lindensjö för all support och hjälp kring upplägget av uppsatsen.

Användning av AI

I denna uppsats har ChatGPT använts som stöd vid felsökning och rättning av `R`. Dessutom har den använts som stöd vid visuella förbättringar av både grafer och tabeller. För andra ändamål har ingen AI-robot kommit till användning.

Innehållsförteckning

1 Introduktion

Detta avsnitt syftar till att introducera bakgrunden till återförsäkring, identifiera centrala problemställningar och mål samt klargöra uppsatsens syfte.

1.1 Bakgrund

När en försäkring tecknas förbinder sig försäkringsbolaget att bistå ekonomiskt till försäkringstagaren om en oväntad skadehändelse inträffar, till exempel en brand, en olycka eller en annan skada som omfattas av försäkringsskyddet. För detta betalar försäkringstagaren en avgift, en så kallad premie.

Ibland inträffar större händelser, antingen i form av en enskild stor skada eller en händelse som påverkar många försäkringar samtidigt, såsom omfattande bränder, stormar eller översvämningar. Då kan försäkringsbolaget behöva ersätta ett stort antal skador på en gång, eller en enskild stor skada, vilket kan leda till mycket höga kostnader. I värsta fall kan den ekonomiska belastningen bli så stor att ett enskilt försäkringsbolag inte klarar av att bära den ensam.

För att skydda sig mot sådana situationer använder sig försäkringsbolag av återförsäkring. Det innebär att försäkringsbolaget tecknar en form av försäkring hos ett återförsäkringsbolag, som åtar sig att täcka en del av kostnaderna vid en inträffad skadehändelse. Genom att teckna återförsäkring sprider försäkringsbolaget sina risker vidare, så att det inte behöver täcka hela kostnaden självt.

Återförsäkring är således ett sätt för försäkringsbolag att vara bättre förberedda vid stora och kostsamma händelser, så att de kan upprätthålla sin betalningsförmåga och fullfölja sina åtaganden gentemot kunderna.

1.2 Problem och mål

En av de centrala övervägningarna som försäkringsbolag ställs inför när de väljer nivå på sina självbehåll, vilket kan ses som en tröskel där återförsäkraren endast ansvarar för den del av skadan som överstiger denna nivå, är balansen mellan kostnad och risk. Ett försäkringsbolag kan antingen välja att betala en högre återförsäkringspremie för att minska variationen i framtida kostnader, eller acceptera större osäkerhet i skadeutfallet i utbyte mot en lägre premie. Denna avvägning mellan finansiell stabilitet och kostnadsbesparing utgör en central del av försäkringsbolagens riskhantering.

Syftet med denna uppsats är att skapa förståelse för hur försäkringsbolag bör tänka när de fattar beslut gällande valet av självbehållsnivå. Ett ytterligare syfte är att analysera hur olika nivåer på självbehållet påverkar det förväntade värdet och variationen i de totala skadebeloppen. Genom att undersöka dessa samband dras slutsatser kring hur försäkringsbolag kan optimera sitt skydd utifrån sina behov.

1.3 Syfte

Syftet med denna uppsats är att undersöka hur ett försäkringsbolag bör resonera kring valet av självbehåll ur ett akademiskt och teoretiskt perspektiv.

1.4 Uppsatsens struktur

Uppsatsen består av fem kapitel, exklusive appendix. Uppsatsen inleds med en introduktion samt en beskrivning av det datamaterial som används. Därefter presenteras relevant teori, följt av metod- och resultatanalys. Uppsatsen avslutas med en diskussion av resultaten och deras implikationer.

2 Databeskrivning

Detta avsnitt presenterar datamaterialet som ligger till grund för analysen samt motiverar hur och varför skadebeloppen har inflationsjusterats.

2.1 Data

Denna uppsats baseras på ett datamaterial tillhandahållet av Länsförsäkringar AB som fokuserar på analyser av egendomsskador, såsom brand- och vattenskador. Materialet omfattar samtliga skador rapporterade av de 23 olika länsförsäkringsbolagen och innehåller bland annat skadenummer, datum för när skadan inträffade och avslutades, utbetalda skadebelopp samt reserverade belopp för framtida utbetalningar. För analyserna i denna uppsats används data från ett enskilt länsförsäkringsbolag. Eftersom varje skada kan ge upphov till flera utbetalningar har dessa summerats för att beräkna det totala skadebeloppet per skada, oavsett utbetalningsår. Det totala skadebeloppet har därefter hänförts till det skadeår då den första utbetalningen inträffade.

I analyserna används främst de totala skadebeloppen per skada, X_i, där i avser en enskild skada. Dessutom används antalet skador, N, under åren 1999–2024. I tabell 1 presenteras en sammanställning av fördelningen av skadebeloppen samt motsvarande antal skador som ligger till grund för analyserna.

Tabell 1: Fördelning av antalet skador per skadebeloppsintervall för försäkringsbolaget.

Skadebelopp (MSEK)	**Antalet skador**
$X_i < 9$	$400\,000 \leq \mathrm{N} < 500\,000$
$9 \leq X_i < 35$	$300 \leq \mathrm{N} < 400$
$35 \leq X_i < 70$	$200 \leq \mathrm{N} < 400$
$70 \leq X_i < 115$	$100 \leq \mathrm{N} < 200$
$X_i \geq 115$	$0 \leq \mathrm{N} < 100$

Av sekretesskäl anges antalet skador i intervall och skadebeloppen per enskild skada redovisas i skalad form.

2.2 Inflationsjusteringar

I de flesta fall delar försäkringsbolag endast med sig av den del av skadestatistiken som är relevant för den aktuella återförsäkringen, det vill säga inte hela informationen om inträffade skador, skadebelopp och frekvens. Detta leder till att återförsäkraren måste hantera begränsad information och göra analyser utifrån det som finns tillgängligt. Exempelvis kan försäkringsbolagen välja att bara rapportera skador som varit större än halva självrisken under de senaste åren. För att kunna jämföra dessa äldre skador och dra korrekta slutsatser är det vanligt att uttrycka dem i dagens penningvärde. Genom att exempelvis

inflationsjustera skadebeloppen kan en uppfattning om vad en skada för ett par år sedan skulle vara värd i dagens penningvärde fås. [3, ss.232-233]

2.2.1 Inflationsjustering

För att uppnå ett tillräckligt omfattande datamaterial krävs det ofta att skador från flera år inkluderas. Under en sådan period kan de genomsnittliga skadekostnaderna förändras, till exempel på grund av inflation eller ökade byggkostnader. Det innebär att datamaterialet behöver inflationsjusteras med hjälp av lämpliga index för att möjliggöra jämförelser över tid. Genom att justera skadebelopp för dessa prisförändringar säkerställs att materialet är representativt och användbart för vidare analys [19, ss.34-35].

I denna uppsats baseras analyserna på historisk skadedata avseende skadebelopp för åren 1999–2024. För att möjliggöra jämförelser över tid har beloppen inflationsjusterats.

Den främsta anledningen till att inflationsjustera skadebelopp är att samtliga skadekostnader ska uttryckas i dagens penningvärde, det vill säga vad skadorna skulle ha kostat om de hade inträffat idag.

Inflationsjusteringen baseras på två index, konsumentprisindex (KPI) och byggnadskostnadsindex (BKI) som båda tillhandahålls av Statistiska centralbyrån (SCB) [33].

I vissa situationer kan en enskild skadehändelse bestå av flera olika typer av skador. Ett exempel är om en fabrik brinner ner, då uppstår en skada på själva egendomen, vilket innebär att kostnaden för återuppbyggnad bör justeras med hjälp av byggnadskostnadsindex (BKI).

Om samma fabrik däremot står still under en period och förlorar intäkter på grund av avbrottet i produktionen, är det istället mer relevant att justera skadan med konsumentprisindex (KPI), eftersom det rör sig om en ekonomisk förlust snarare än en fysisk återuppbyggnad.

Eftersom analysperioden inleds år 1999 används detta år som basår, det vill säga ett referensår som används för att kunna jämföra prisnivåer över tid. Ett basår sätts vanligtvis till 100, och genom att jämföra andra års indexvärden med detta år blir det möjligt att se hur priser har förändrats.

För samtliga efterföljande år beräknas ett index genom att dividera respektive års medelvärde med värdet från basåret, och därefter multiplicera med 100.

Vi betecknar I_i^K som årsmedelvärdet för KPI för år i. Det justerade indexet, $I_{K,i}^{\text{just}}$, beräknas då enligt följande formel

$$I_{K,i}^{\text{just}} = \frac{I_i^K}{I_{1999}^K} \cdot 100. \tag{1}$$

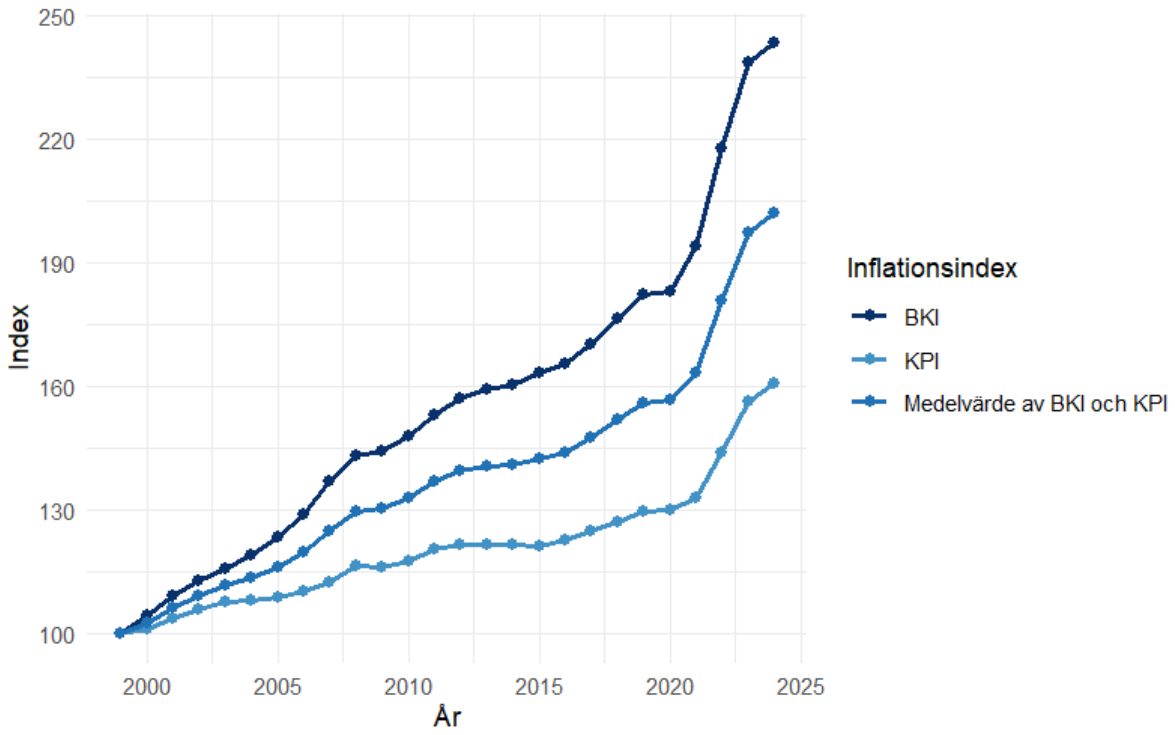
Analoga beräkningar används för det justerade indexet BKI, där $B_{K,i}^{\text{just}}$ motsvarar det justerade indexet för byggnadskostnadsindex.

I det tillgängliga datamaterialet ingår det totala skadebeloppet per enskild skada, men information saknas om i vilken utsträckning skadorna bör inflationsjusteras med byggnadskostnadsindex (BKI) respektive konsumentprisindex (KPI).

Mot denna bakgrund används ett medelvärde av KPI och BKI som inflationsjusteringsfaktor som beräknas enligt

$$I_{i,\text{medel}} = \frac{I_{K,i}^{\text{just}} + B_{K,i}^{\text{just}}}{2}. \tag{2}$$

I figur 1 nedan visas utvecklingen av det svenska BKI, KPI samt deras genomsnitt för åren 1999–2024.



Figur 1: Utveckling av byggnadskostnadsindex (BKI), konsumentprisindex (KPI) samt deras genomsnitt under åren 1999–2024.

3 Teori

Detta avsnitt presenterar den teori som är mest relevant för uppsatsen. Inledningsvis definieras begreppet återförsäkring. Därefter introduceras modeller för såväl antalet skador som skadebelopp. Vidare behandlas metoder för parameterskattning. Avsnittet avslutas med en genomgång av olika goodness-of-fit-tester och deras egenskaper.

3.1 Definition av återförsäkring

Återförsäkring är en form av försäkring där en återförsäkrare åtar sig att ersätta försäkringsgivaren (cedenten) för en viss andel av de skador som annars skulle ha belastat det ursprungliga försäkringsbolaget. Skyddet kan avse enskilda försäkringar eller en hel försäkringsportfölj. Återförsäkraren tar därmed över en del av den risk som försäkringsgivaren annars själv skulle bära. I utbyte mot detta skydd betalar återförsäkringstagaren en premie till återförsäkraren. Den del som täcks av återförsäkraren kan antingen utgöra en procentandel av en portfölj eller en del av varje enskild skada som överstiger ett förutbestämt belopp.

Sammanfattningsvis innebär återförsäkring att ett försäkringsbolag skyddar sig mot både större skador och höga kostnader, exempelvis vid omfattande bränder eller naturkatastrofer, genom att överföra en del av sitt åtagande till ett återförsäkringsbolag. Återförsäkring utgör därmed ett centralt verktyg för riskhantering och riskspridning inom försäkringsbranschen [9].

3.1.1 Olika former av återförsäkring

Det finns olika former av återförsäkring, men de två vanligaste är treatyåterförsäkring och fakultativ återförsäkring. Treatyåterförsäkring omfattar en hel portfölj av risker enligt ett avtal mellan parterna, medan fakultativ återförsäkring tecknas för enskilda, särskilt stora eller riskfyllda försäkringshändelser.

Återförsäkringsavtal kan även delas in i två typer beroende på hur riskerna fördelas, proportionell återförsäkring och icke-proportionell återförsäkring. Vid proportionell återförsäkring delas risk och därmed både premier och skadeutbetalningar mellan försäkringsgivaren och återförsäkraren enligt en på förhand fastställd andel. Vid icke-proportionell återförsäkring, även benämnd som excess of loss-återförsäkring, täcker återförsäkraren endast den del av varje enskild skada som överstiger ett förutbestämt belopp, oftast upp till en överenskommen gräns [14, ss. 343–356].

Det belopp som försäkringsbolaget står för innan återförsäkringen träder in kallas självbehåll. Självbehållet fungerar som en tröskel, där återförsäkraren endast ansvarar för den del av skadan som överstiger denna nivå. En mer detaljerad genomgång av excess of loss-försäkring följer i nästa avsnitt.

3.1.2 Excess of loss

Detta avsnitt behandlar excess of loss-återförsäkring (XL-skydd) mer ingående. Denna typ av återförsäkring innebär att återförsäkraren täcker den del av varje enskild skada som överstiger ett förutbestämt belopp. Ersättningen gäller oftast endast upp till ett i förväg avtalat tak. Detta innebär att återförsäkraren inte delar på samtliga skador proportionellt, utan träder in först när en skada överstiger den angivna gränsen. För att försäkringstagaren (cedenten) ska få ersättning från återförsäkringsgivaren så måste skadebeloppet överstiga ett visst belopp, det vill säga självbehållet, U. Återförsäkraren har dock en övre gräns, C, vilket är det i förväg avtalade taket. Den totala skadekostnaden, X, som försäkringsbolaget behöver stå för innan återförsäkringen träder in, kan därmed delas upp i två delar enligt

$$X = X_U + X_R. \tag{3}$$

Där X_U är den del som försäkringsbolaget behöver stå för på egen bekostnad, det vill säga försäkringsbolagets självbehåll, och definieras enligt

$$X_U = X - X_R. \tag{4}$$

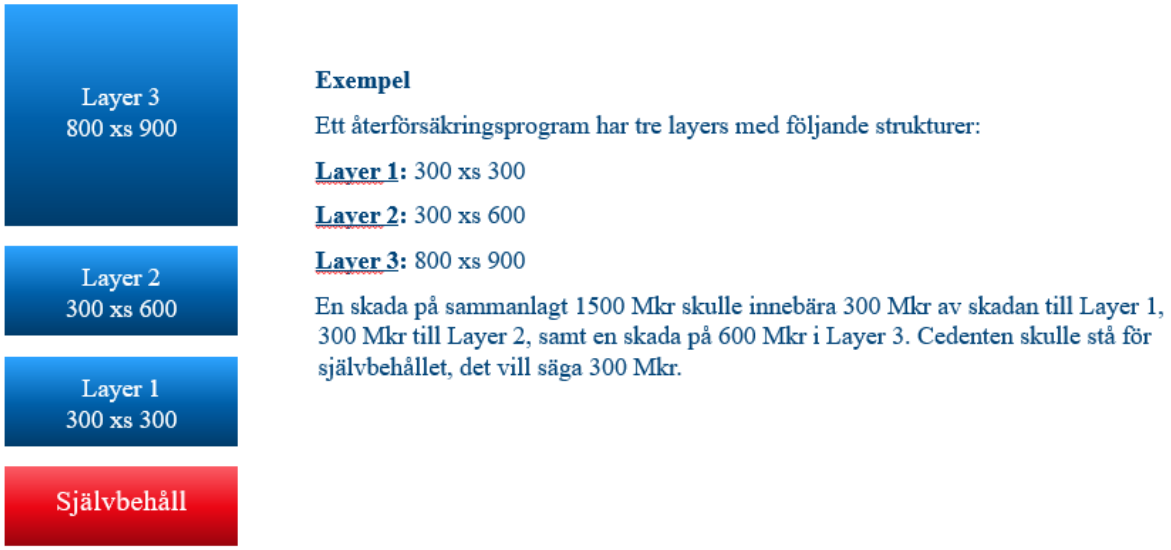
Där X_R är det belopp som återförsäkraren betalar och definieras enligt

$$X_R = (X - U)^+ - (X - C)^+. \tag{5}$$

Här betecknar $(\cdot)^+$ den positiva delen, det vill säga $\max(0, \cdot)$. Försäkringsbolaget får ersättning från återförsäkringsgivaren när skadebeloppet X överstiger självbehållet U, det vill säga om $X > U$.

Excess of loss-kontrakt är ofta uppbyggda i flera nivåer, så kallade lager (layers), vilket illustreras i figur 2. Varje layer definieras av ett intervall mellan en nedre och en övre gräns.

Ett excess of loss-kontrakt uttrycks som $Z \,\text{xs}\, U$, vilket utläses som ”Z excess U”, där $Z = C - U$ och innebär att återförsäkringsgivaren täcker skador upp till C förutsatt att skadan överstiger U. Självbehållet är den del av skadekostnaden som försäkringsbolaget ansvarar för innan återförsäkringen träder in. Genom att dela upp skyddet i olika layers sprids risken mellan flera återförsäkringsgivare. Exemplet i figur 2 är Layer 1 definierad som 300 xs 300, vilket innebär att skador mellan 300 och 600 miljoner kronor omfattas. Layer 2 omfattar skador mellan 600 och 900 miljoner kronor, och Layer 3 omfattar skador mellan 900 och 1700 miljoner kronor. För att en återförsäkrare på en viss layer ska behöva ersätta en skada krävs alltså att skadan först överstiger självbehållet. Alla detaljer och information kring excess of loss-återförsäkring går att hitta i [5, ss.1-5].



Figur 2: Struktur för en excess of loss-återförsäkring med flera lager (layers).

3.2 Modeller för antalet skador

Syftet med denna uppsats är att modellera de totala skadebeloppen för ett försäkringsbolag genom att separat modellera antalet skador och skadebeloppen. Två klassiska fördelningar som används för att modellera antalet skador är Poisson- och negativa binomialfördelningen.

Den huvudsakliga skillnaden mellan dessa två fördelningar är att Poissonfördelningen förutsätter att skadorna inträffar oberoende av varandra och med konstant intensitet över tid. Den negativa binomialfördelningen används däremot när det finns variation eller överspridning i data, det vill säga när variansen är större än väntevärdet. Detta kan indikera att skadorna uppträder med någon form av struktur eller beroende.

I nästa avsnitt presenteras Poisson- och negativa binomialfördelningen mer ingående, tillsammans med deras respektive egenskaper.

3.2.1 Poissonfördelning

En av de vanligaste fördelningarna som används vid modellering av antalet skador är Poissonfördelningen. Det är en diskret sannolikhetsfördelning som anger sannolikheten för att ett visst antal händelser inträffar inom ett givet tidsintervall.

Enligt [19, ss. 43–47] är Poissonfördelningen lämplig för att modellera antalet skador, förutsatt att följande antaganden är uppfyllda:

1. Antalet händelser som inträffar inom ett tidsintervall är oberoende av varandra

2. Endast en händelse kan inträffa åt gången

3. Sannolikheten för att en händelse inträffar under ett mycket kort tidsintervall är proportionell mot längden på det tidsintervallet.

Om dessa villkor är uppfyllda kan antalet händelser modelleras med en Poissonfördelning. Enligt [4, ss. 86-87] ges sannolikhetsfunktionen för en stokastisk variabel X som är Poissonfördelad av

$$p_X(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{6}$$

Där λ representerar det förväntade antalet händelser inom tidsintervallet. En viktig egenskap hos Poissonfördelningen är att både väntevärdet och variansen är lika med λ. Det vill säga

$$\mathbb{E}(X) = \mathrm{Var}(X) = \lambda. \tag{7}$$

3.2.2 Negativ binomialfördelning

En alternativ modell för antalet skador är den negativa binomialfördelningen. Denna fördelning är att föredra när datamaterialet uppvisar överspridning, det vill säga när variansen är större än väntevärdet.

Den negativa binomialfördelningen bygger på en serie oberoende försök, där varje försök ger antingen ett lyckat eller ett misslyckat utfall. Dessa kallas Bernoulliförsök.

För varje försök finns det en sannolikhet p för ett lyckat utfall, och sannolikheten för ett misslyckat utfall är då $1 - p$.

En negativ binomialfördelning beskriver hur många försök som behöver göras innan ett visst antal lyckade utfall uppnås.

Låt X vara antalet försök som krävs för att uppnå exakt r lyckade försök. För att ett visst utfall x ska inträffa krävs att de första $x - 1$ försöken innehåller exakt $r - 1$ lyckade utfall, och att det x:te försöket är ett lyckat utfall. Enligt [4, ss. 92–93] sägs en diskret slumpvariabel X vara negativ binomialfördelad med parametrar $r \geq 1$ och p ($p < 0 < 1$) om sannolikhetsfunktionen ges av

$$p_X(x) = P(X = x) = \binom{x-1}{r-1} \cdot p^r \cdot (1-p)^{x-r}, \quad x = r, r+1, r+2, \ldots \tag{8}$$

Vilket innebär att det n:te försöket resulterar i ett lyckat utfall med sannolikheten p. Här är $\binom{x-1}{r-1}$ en binomialkoefficient. En binomialkoefficient anger antalet sätt att välja ett visst antal objekt från en mängd utan hänsyn till ordningen. Antalet sätt att ordna n element, ges av $n! = n(n-1)(n-2), \cdots, 1$ och antalet sätt att välja ut k element bland n utan hänsyn till ordningen ges av binomialkoefficienten

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}, \tag{9}$$

och utläses ”n över k”, vilket representerar antalet sätt att välja k objekt ur en mängd med n objekt [4, ss. 20–21]. I ekvation 8 anger binomialkoefficienten $\binom{x-1}{r-1}$ alltså på hur många sätt de $r-1$ första utfallen kan placeras bland de första $x-1$ försöken utan hänsyn till ordningen. Det sista försöket, det x:te, måste då vara det r:te lyckade utfallet.

Om X är en stokastisk variabel som följer en negativ binomialfördelning, uttrycks detta som

$$X \sim \text{NegBin}(r, p), \tag{10}$$

och sannolikheten att få exakt x försök innan r lyckade utfall uppnås ges av ekvation 8. För antalet försök ges det förväntade värdet, det vill säga väntevärdet, av

$$\mathbb{E}[X] = \frac{r}{p}, \tag{11}$$

och variansen ges av

$$\text{Var}(X) = \frac{r(1-p)}{p^2}. \tag{12}$$

Om Y istället definieras som antalet misslyckade försök innan den r:te framgången, där $Y = X - r$, erhålls den alternativa formen av en negativ binomialfördelning. Sannolikheten i detta fall ges enligt [23] av

$$P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y. \tag{13}$$

Då ges istället väntevärdet och variansen enligt

$$\mathbb{E}[Y] = \frac{r(1-p)}{p}, \tag{14}$$

och

$$\text{Var}(Y) = \text{Var}(X) = \frac{r(1-p)}{p^2}. \tag{15}$$

3.3 Modeller för skadebeloppen

Efter att modellerna för antalet skador har behandlats riktas nu fokus mot de totala skadebeloppen. Två vanliga fördelningar som ofta används vid analys och modellering av skadebelopp är Pareto- och lognormalfördelningen. Dessa är dock inte de enda fördelningarna som förekommer i praktiken.

I detta avsnitt presenteras Pareto- och lognormalfördelningen samt deras egenskaper, tillsammans med andra relevanta fördelningar som är av betydelse för modellering av skadebeloppen.

3.3.1 Den generaliserade Paretofördelningen

Den generaliserade Paretofördelningen (GPD) används inom extremvärdesteori för att modellera skadebelopp som överstiger ett tröskelvärde μ, det vill säga observationer $X > \mu$, vilka betraktas som extrema [18].

En av de främsta fördelarna med den generaliserade Paretofördelningen är att den inkluderar en formparameter, ξ, som styr svansens beteende. Hur snabbt sannolikheten i svansen avtar, beror på värdet av denna parameter.

Om $\xi > 0$ har fördelningen en tung svans, vilket innebär att sannolikheten $\mathbb{P}(X > x) = 1 - F(x)$ avtar långsammare än för en exponentialfördelning. Tungsvans innebär att extremvärden är mer sannolika. För en mer detaljerad genomgång, se [13, ss. 7–11]. Om $\xi < 0$ har fördelningen en tunn svans, vilket innebär att sannolikheten för extrema värden är mindre. I fallet då $\xi = 0$ och $\mu = 0$ motsvarar den generaliserade Paretofördelningen en exponentialfördelning. Denna flexibilitet gör att den generaliserade Paretofördelningen kan anpassas till en rad olika extrembeteenden (se delkapitel 3.4 i [12]).

För en stokastisk variabel X som följer en generaliserad Paretofördelning ges fördelningsfunktionen av

$$F_X(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-\mu)}{\beta}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\frac{x-\mu}{\beta}\right), & \xi = 0, \end{cases} \tag{16}$$

där $x \geqslant \mu$ när $\xi \geqslant 0$, och $\mu \leqslant x \leqslant \mu - \beta/\xi$ när $\xi < 0$. Den motsvarande täthetsfunktionen ges av

$$f_X(x) = \frac{1}{\beta}\left(1 + \frac{\xi(x-\mu)}{\beta}\right)^{-(1/\xi+1)}, \tag{17}$$

täthetsfunktionen är definierad för $x \geq \mu$ om $\xi \geq 0$, och för $\mu \leq x \leq \mu - \beta/\xi$ om $\xi < 0$.

Parametern μ är lägesparametern som bestämmer var fördelningen börjar. Parametern β är en skalparameter som styr spridningen, och ξ är formparametern som styr svansens tjocklek. Väntevärdet och variansen för en stokastisk variabel X som är generaliserad Paretofördelad ges av

$$\mathbb{E}[X] = \begin{cases} \mu + \dfrac{\beta}{1-\xi}, & \text{om } \xi < 1 \\ \infty, & \text{om } \xi \geq 1 \end{cases} \tag{18}$$

$$\mathrm{Var}(X) = \begin{cases} \dfrac{\beta^2}{(1-\xi)^2(1-2\xi)}, & \text{om } \xi < \dfrac{1}{2} \\ \infty, & \text{om } \xi \geq \dfrac{1}{2} \end{cases} \tag{19}$$

Detta innebär att om $\xi \geq 1$ existerar inte väntevärdet, och om $\xi \geq \frac{1}{2}$ är variansen oändlig. I denna uppsats används den generaliserade Paretofördelningen för att modellera skadebelopp som överstiger ett tröskelvärde, enligt peaks over threshold-metoden (se avsnitt 3.4.1). För mer information om den generaliserade Paretofördelningen och dess egenskaper hänvisas läsaren till [18].

3.3.2 Paretofördelning

Paretofördelningen är en vanlig fördelning som används för att modellera tungsvansad data. Till skillnad från den generaliserade Paretofördelningen har Paretofördelningen endast två parametrar, α och x_m, vilka är form- respektive skalparameter. Denna fördelning inkluderar alltså inte någon lägesparameter μ.

Fördelningsfunktionen för en stokastisk variabel X som är Paretofördelad ges av

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m, \end{cases} \tag{20}$$

och täthetsfunktionen ges av

$$f_X(x) = \begin{cases} \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} & x \geq x_m \\ 0 & x < x_m. \end{cases} \tag{21}$$

Eftersom Paretofördelningen är ett specialfall av den generaliserade Paretofördelningen finns det ett tydligt samband mellan deras parametrar. För att erhålla Paretofördelningen som ett specialfall av den generaliserade Paretofördelningen sätts lägesparametern $\mu = x_m$, formparametern $\xi = \frac{1}{\alpha}$ och skalparametern $\beta = \frac{x_m}{\alpha}$ i ekvation 16, vilket då blir:

$$F_X(x) = 1 - \left(1 + \frac{1}{\alpha} \cdot \frac{x - x_m}{x_m/\alpha}\right)^{-\alpha} = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}. \tag{22}$$

Detta motsvarar exakt fördelningsfunktionen för en stokastisk variabel som är Paretofördelad.

Väntevärdet och variansen för en Paretofördelning ges av

$$\mathbb{E}[X] = \begin{cases} \frac{\alpha x_m}{\alpha - 1}, & \alpha > 1 \\ \infty, & \alpha \leq 1 \end{cases} \tag{23}$$

$$\mathrm{Var}(X) = \begin{cases} \frac{\alpha x_m^2}{(\alpha-1)^2(\alpha-2)}, & \alpha > 2 \\ \infty, & \alpha \leq 2 \end{cases} \tag{24}$$

Notera att väntevärdet inte existerar för $\alpha \leq 1$, och om $\alpha \leq 2$ är variansen oändlig. All information som presenteras ovan finns att läsa i [36].

3.3.3 Normalfördelning

En fördelning som ibland används för att modellera mindre skadebelopp är normalfördelningen. Täthetsfunktionen för en stokastisk variabel X som är normalfördelad med parametrarna μ och σ^2, $X \sim N(\mu, \sigma^2)$, definieras enligt [4, ss.100-107]

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty \tag{25}$$

och fördelningsfunktionen är

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt. \tag{26}$$

Väntevärde och varians för en stokastisk variabel X som är normalfördelad ges av

$$\mathbb{E}[X] = \mu, \tag{27}$$

$$\mathrm{Var}(X) = \sigma^2. \tag{28}$$

För att kunna modellera skadebelopp med en normalfördelning krävs det att beloppen är relativt små. Dessutom kan normalfördelningen ge upphov till negativa värden, vilket inte är rimligt vid modellering av skadebelopp.

Vid modellering av mindre skadebelopp rekommenderas därför ofta andra fördelningar, såsom gamma- eller lognormalfördelningen, eftersom de är definierade enbart för positiva värden.

3.3.4 Lognormalfördelning

Lognormalfördelningen används ofta, liksom Paretofördelningen, vid modellering av skadebelopp. En stokastisk variabel som är lognormalfördelad kan uttryckas som $Y = e^X$, där $X \sim N(\mu, \sigma^2)$. Standardiseringen $(X - \mu)/\sigma$ är därmed standardnormalfördelad, och dess fördelningsfunktion betecknas vanligtvis som $\Phi(\cdot)$. Enligt [4, ss.111-112] ges fördelningsfunktionen för en lognormalfördelad variabel av följande uttryck:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\ln(Y) \leq \ln(y)) = \mathbb{P}\left(\tfrac{\ln(Y)-\mu}{\sigma} \leq \tfrac{\ln(y)-\mu}{\sigma}\right) = \Phi\left(\tfrac{\ln(y)-\mu}{\sigma}\right). \tag{29}$$

Täthetsfunktionen är

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right), \quad y > 0. \tag{30}$$

För en stokastisk variabel Y som är lognormalfördelad ges väntevärdet och variansen

av följande uttryck

$$\mathbb{E}[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{31}$$

$$\mathrm{Var}(Y) = \left(\exp(\sigma^2) - 1\right)\exp\left(2\mu + \sigma^2\right). \tag{32}$$

3.3.5 Gammafördelning

En annan fördelning som kan användas vid modellering av skadebelopp är gammafördelningen. Antag att den stokastiska variabeln X är gammafördelad, $X \sim \mathrm{Gamma}(\alpha, \beta)$, där $\alpha > 0$ är formparametern och $\beta > 0$ är skalparametern. Täthetsfunktionen för gammafördelningen ges av

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{33}$$

jämför med [4, ss. 113–114], där β tolkas som en intensitetsparameter. $\Gamma(\cdot)$ betecknar gammafunktionen, som definieras som

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0. \tag{34}$$

Fördelningsfunktionen uttrycks enligt

$$F_X(x) = P(X \leq x) = \int_0^x f_X(t)\, dt = \frac{\gamma\left(\alpha, \frac{x}{\beta}\right)}{\Gamma(\alpha)}, \tag{35}$$

där $\gamma(\alpha, \frac{x}{\beta})$ är den nedre inkompletta gammafunktionen, definierad som

$$\gamma\left(\alpha, \frac{x}{\beta}\right) = \int_0^{x/\beta} t^{\alpha-1} e^{-t}\, dt. \tag{36}$$

Den nedre inkompletta gammafunktionen representerar samma typ av integral som gammafunktionen (ekvation 34), men med integrationsgränser från 0 till $\frac{x}{\beta}$ i stället för till oändligheten. Det innebär att den beskriver hur stor andel av den totala fördelningen som har observerats upp till ett visst värde.

Väntevärdet och variansen för en gammafördelad variabel, X, ges av

$$\mathbb{E}[X] = \alpha\beta \tag{37}$$

$$\text{Var}(X) = \alpha\beta^2. \tag{38}$$

3.4 Extremvärdesteori

Vid modellering av ett stickprov kan olika statistiska modeller ofta fånga upp de vanliga händelserna på ett tillfredsställande sätt, särskilt de som ligger nära fördelningens median. De största utmaningarna ligger däremot i att modellera de extrema observationerna i svansarna.

Inom försäkringsbranschen är det vanligt att försäkringsbolag återförsäkrar sig mot extrema händelser som inträffar med låg sannolikhet men som kan ge upphov till mycket stora kostnader. För att kunna hantera dessa situationer krävs metoder som fokuserar på svansskattning, vilket är just vad extremvärdesteori gör. Givet ett stickprov av tidigare observerade värden, till exempel skadekostnader, kan extremvärdesteori användas för att analysera extremt stora skadehändelser, det vill säga svansbeteenden. Extremvärdesteori är således användbar när sannolikheten för sällsynta och extrema händelser ska uppskattas.

Sammanfattningsvis är extremvärdesteori ett tillvägagångssätt för att analysera observationer som ligger långt från medianen i en sannolikhetsfördelning. Dessa extrema händelser kan utgöras av både mycket små och mycket stora värden, men i ett återförsäkringssammanhang är det framför allt de stora, sällsynta skadorna i den högra ändpunkten som är av intresse.

Enligt [10, ss.167-181] finns det två huvudsakliga metoder inom extremvärdesteori, block maxima- och peaks over threshold (POT)-modellen. I denna uppsats tillämpas POT-modellen för att modellera händelser som genererat höga skadekostnader. Till skillnad från block maxima-modellen är POT-modellen mer praktisk vid arbete med små stickprov, vilket ofta är fallet inom återförsäkring. Därav kommer block maxima-modellen inte att hanteras i denna uppsats, de intresserade hänvisas till [22, ss. 291–295] för en mer detaljerad genomgång av denna metod.

3.4.1 Peaks Over Threshold (POT)

Peaks over threshold (POT)-metoden används för att skatta svanssannolikheter i en fördelning. Den är särskilt användbar vid modellering av observationer som överskrider ett visst tröskelvärde u.

Enligt [28] är POT-metoden en av de vanligaste metoderna inom extremvärdesteorin. Den är särskilt effektiv för att modellera svansen i en fördelning och bygger på användning av den generaliserade Paretofördelningen (GPD).

Det är naturligt att betrakta observationerna X_i som överstiger ett högt tröskelvärde u som extrema skadehändelser och därmed ligger nära den högra ändpunkten i fördelningen.

I POT-modellen studeras skadebelopp som överskrider ett tröskelvärde u. För ett givet högt tröskelvärde $u \in \mathbb{R}$ definieras de överskridande beloppen som

$$Y_i = X_i - u \mid X_i > u, \tag{39}$$

under antaganden enligt [8, ss. 74–76] att $X_1, X_2, \ldots, X_n$ är en följd av oberoende och likafördelade stokastiska variabler som har samma okända fördelningsfunktion. De överskridande skadebeloppens fördelning kan då beskrivas enligt

$$\begin{aligned} F_u(y) = \mathbb{P}(Y_i \leq y) &= \mathbb{P}(X_i - u \leq y \mid X_i > u) \\ &= \mathbb{P}(X_i \leq y + u \mid X_i > u) \\ &= \frac{\mathbb{P}(u < X \leq y + u)}{\mathbb{P}(X > u)} \\ &= \frac{F(y+u) - F(u)}{1 - F(u)} \quad \text{för } 0 \leq y \leq x_F - u, \end{aligned} \tag{40}$$

där $x_F \leq \infty$ är den högra ändpunkten för den underliggande fördelningen F.

Om den underliggande fördelningsfunktionen F vore känd, skulle sannolikheten att en skada överskrider ett givet tröskelvärde u kunna beräknas exakt. I praktiken är dock F oftast okänd. Eftersom den exakta fördelningsfunktionen för F inte är känd, används en approximation som fungerar bra när tröskelvärdet u är tillräckligt högt.

Denna approximation baseras på extremvärdesteori, närmare bestämt Pickands-Balkema-de Haan-satsen, som säger att för höga tröskelvärden u kan överskridande fördelningen $F_u(y)$ approximeras väl med en generaliserad Paretofördelning enligt

$$F_u(y) \approx G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\beta}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\beta}\right), & \xi = 0, \end{cases} \tag{41}$$

där $\beta > 0$ samt $y \geq 0$ då $\xi \geq 0$ och $0 \leq y \leq -\beta/\xi$ då $\xi < 0$. Denna approximation används ofta i praktiska tillämpningar. Pickands-Balkema-de Haan-satsen säger i stora drag att, för ett tillräckligt högt tröskelvärde u kan överskridande fördelningen approximeras av en generaliserad Paretofördelning under vissa villkor och antaganden. För dessa villkor, antaganden och en djupare förklaring av Pickands-Balkema-de Haan-satsen se [22, ss. 295–298]. Dessutom, för en mer detaljerad genomgång av antagandena bakom POT-modellen, se [29].

3.4.2 Val av tröskelvärden

När ett lämpligt tröskelvärde u ska väljas vid tillämpning av POT-modellen, uppstår en viktig avvägning mellan bias och varians i skattningen av form- och skalparametrarna i den generaliserade Paretofördelningen (GPD). Bias syftar på den systematiska avvikelsen mellan den skattade parametern och det sanna parametervärdet. Ett för lågt valt tröskelvärde innebär att många observationer inkluderas i analysen, vilket visserligen minskar variansen men samtidigt ökar risken för bias. Detta beror på att observationer nära medianen av fördelningen inte följer GPD-antagandet lika väl och därmed kan leda till en mindre tillförlitlig skattning. Å andra sidan leder ett för högt tröskelvärde till att endast ett fåtal extremvärden används i skattningen. Detta minskar bias men ökar variansen, eftersom skattningen då baseras på färre datapunkter och därmed blir mer osäker [17, ss. 223-228]. Syftet är att hitta ett tröskelvärde u som uppnår en bra balans mellan låg bias och låg varians, vilket är ett klassiskt exempel på avvägning mellan bias och varians inom statistisk modellering. Teoretiskt sett skulle det kunna finnas ett optimalt tröskelvärde som minimerar både bias och varians, men i praktiken är detta sällan möjligt att identifiera exakt. Enligt [12, ss. 350- 360] rekommenderas därför att grafiska metoder, såsom QQ-plotar, används för att undersöka hur parametrarna i en GPD-fördelning förändras vid olika val av tröskelvärden.

3.5 Parameterskattning

En viktig del av denna uppsats är att skatta parametrar för de fördelningar som har testats utifrån det empiriska datamaterialet. Detta eftersom de skattade parametrarna används för att anpassa fördelnings- och täthetsfunktionerna. Två klassiska metoder som ofta används för parameterskattning är maximum likelihood-metoden (MLE) och momentmetoden (MoM). I denna uppsats används momentmetoden för att skatta parametrarna för en Poisson- och negativ binomialfördelning. De resterande fördelningarnas parametrar skattas med maximum likelihood-metoden.

3.5.1 Maximum likelihood-metoden

För att bestämma maximum likelihood-skattningen antas att X_k, där $k = 1, \ldots, n$, är n oberoende och lika fördelade stokastiska variabler med en gemensam täthetsfunktion. Enligt [25] definieras likelihoodfunktionen som följande

$$\mathcal{L}(x;\theta) = \prod_{k=1}^{n} f(x_k;\theta) = f(x_1;\theta)f(x_2;\theta)\ldots f(x_k;\theta), \tag{42}$$

där θ är en parameter i fördelningsfunktionen. För att förenkla beräkningarna logaritmeras likelihoodfunktionen, vilket resulterar i log-likelihoodfunktionen given av

$$\log(\mathcal{L}(x;\theta)) = \sum_{k=1}^{n} \log(f(x_k;\theta)). \tag{43}$$

För att hitta det värde på parametern θ som maximerar log-likelihoodfunktionen deriveras uttrycket i ekvation 43 med avseende på θ, varefter derivatan sätts lika med noll och ekvationen löses med avseende på θ.

I praktiken kan det ofta vara svårt att räkna ut derivatan, särskilt när fördelningen innehåller många parametrar eller om log-likelihoodfunktionen inte går att lösa på ett enkelt sätt. I sådana fall kan det vara lämpligt att använda sig av numeriska optimeringsmetoder, såsom Newton-Raphsons metod.

Exempel 3.1. *Normalfördelning*

Antag att den stokastiska variabeln X är normalfördelad, $X \sim \mathcal{N}(\mu, \sigma^2)$. Täthetsfunktionen för normalfördelningen ges av

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \mu \in \mathbb{R},\ \sigma^2 > 0. \tag{44}$$

Likelihoodfunktionen uttrycks som

$$\mathcal{L}(x;\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right). \tag{45}$$

Genom att logaritmera uttrycket i ekvation 45 erhålls log-likelihoodfunktionen enligt följande

$$\ln(\mathcal{L}(x;\mu,\sigma^2)) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{46}$$

För att maximera log-likelihoodfunktionen tas den partiella derivatan med avseende på parametern μ

$$\frac{\partial}{\partial\mu}\ln(\mathcal{L}(x;\mu,\sigma^2)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu). \tag{47}$$

Eftersom den andra derivatan med avseende på μ är negativ, är maximum likelihoodskattningen av μ lika med stickprovsmedelvärdet, det vill säga

$$\hat{\mu}(x) = \bar{x} := \sum_{i=1}^{n}\frac{x_i}{n}. \tag{48}$$

Den partiella derivatan med avseende på σ^2 ges av följande

$$\frac{\partial}{\partial\sigma^2}\ln(\mathcal{L}(x;\mu,\sigma^2)) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{49}$$

Då $\hat{\mu}(x) = \bar{x}$, ges maximum likelihood-skattningen av variansen enligt

$$\hat{\sigma}^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{50}$$

En motsvarande genomgång för lognormalfördelningen kan ses i [35].

3.5.2 Momentmetoden

En alternativ metod för parameterskattning av fördelningar såsom Poissonfördelningen eller den negativa binomialfördelningen är momentmetoden (MoM). Metoden bygger på att sätta de empiriska momenten med den teoretiska fördelningens moment. För att bestämma momentmetodens skattningar för en viss fördelning definieras $X_1, \ldots, X_n$ som oberoende och lika fördelade stokastiska variabler, där n är antalet observationer från en fördelning som beror på en eller flera parametrar θ.

Det teoretiska momentet för en fördelning definieras som följande

$$\mu^k = \mathbb{E}\left[X^k\right]. \tag{51}$$

Vidare, låt θ beteckna parametern som fördelningen beror på. Då definieras det k:te teoretiska momentet enligt

$$\mu^k(\theta) = \mathbb{E}\left[X^k; \theta\right], \quad k = 1, 2, \ldots \tag{52}$$

Där μ^k motsvarar det k:te teoretiska momentet. Vidare kan det k:te empiriska momentet från stickprovet definieras som

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \bar{X}^k, \tag{53}$$

vilket blir stickprovsmedelvärdet. Vid parameterskattning med momentmetoden sätts de empiriska momenten lika med de teoretiska. För det första ordningens moment ($k = 1$) innebär detta att det empiriska momentet M_1 sätts lika med väntevärdet

$$M_1 = \mathbb{E}[X; \theta] = \bar{X} = \mu_1(\theta). \tag{54}$$

Om fördelningen istället har en tvådimensionell parametervektor $\theta = (\theta_1, \theta_2)$, räcker det inte med ekvation 54 för parameterskattningen. Då tillämpas även motsvarande ekvation för det andra momentet vilket ges av följande

$$M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}[X^2; \theta]. \tag{55}$$

Tillsammans med ekvation 54 erhålls därmed två ekvationer, utifrån vilka moment-

skattningarna av θ_1 och θ_2 kan bestämmas. På motsvarande sätt kan ytterligare momentekvationer inkluderas om θ är flerdimensionell [4, ss. 275–276].

Vidare påminns om definitionen av variansen, som kan uttryckas i termer av det första och det andra teoretiska momenten på följande sätt

$$\mathrm{Var}(X;\theta) = \mathbb{E}[X^2;\theta] - (\mathbb{E}[X;\theta])^2. \tag{56}$$

Variansen kan därmed uttryckas som skillnaden mellan det andra momentet och kvadraten av det första momentet. En utförlig genomgång av detta återfinns i [34, ss. 261–265] samt i [4, ss. 275–276].

Momentmetoden har i denna uppsats använts för att skatta parametrarna för både Poisson- och negativa binomialfördelningen.

Exempel 3.2. *Poissonfördelning*

Antag att den stokastiska variabeln X är Poissonfördelad med parameter λ, $X \sim \mathrm{Po}(\lambda)$. Det första teoretiska momentet (för $k = 1$) är väntevärdet av den stokastiska variabeln X och definieras enligt

$$\mu(\lambda) = \mathbb{E}[X;\lambda] = \lambda. \tag{57}$$

Det första empiriska momentet från stickprovet blir stickprovsmedelvärdet enligt följande

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{58}$$

Genom att sätta det teoretiska momentet lika med det empiriska kan parametern λ enkelt lösas ut på följande sätt

$$M_1 = \lambda \Leftrightarrow \hat{\lambda}_{\mathrm{MoM}} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{59}$$

Momentmetodens skattning av parametern λ är alltså stickprovsmedelvärdet. För en tillämpning av momentmetoden på andra fördelningar, exempelvis den negativa binomialfördelningen, se [30].

3.6 Modellutvärdering

För att utvärdera och avgöra vilken sannolikhetsfördelning som är mest lämplig för ett datamaterial med exempelvis skadekostnader kan olika typer av goodness-of-fit-tester (GOF-tester) användas. De två klassiska goodness-of-fit-testerna är Anderson-Darling- och Kolmogorov-Smirnov-test, vilka båda utgör hypotesprövningar mot en nollhypotes, H_0, [2]. Nollhypotesen, H_0, säger att datat följer en specifik fördelning, medan den alternativa hypotesen, H_1, säger att datat inte följer den specifika fördelningen. Båda fallen genererar en teststatistika som kan användas för att avgöra ifall H_0 ska förkastas eller inte. Om sannolikheten att observera teststatistikan, givet att H_0 är sann, är lägre än en förutbestämd signifikansnivå (vanligtvis 5%), förkastas nollhypotesen. P-värdet definieras som sannolikheten att, under antagandet att H_0 är sann, observera en teststatistika som är minst lika extrem som den som erhållits från det observerade datamaterialet. Om p-värdet understiger signifikansnivån förkastas nollhypotesen. Det klassiska valet av signifikansnivå är 5% och denna nivå används även i denna uppsats [26]. För att ytterligare jämföra modellerna kommer både Akaikes informationskriterium (AIC) och Bayesianskt informationskriterium (BIC) att användas. Dessa mått bedömer hur väl en modell passar data, samtidigt som de tar hänsyn till modellens komplexitet.

3.6.1 Kolmogorov-Smirnov-test

Ett av de vanligaste testen för att se ifall data följer en viss fördelning, exempelvis normal- eller lognormalfördelning, är Kolmogorov-Smirnov-test (KS-test). Detta test är ett icke-parametriskt test som används för att jämföra en empirisk fördelning med en teoretisk fördelning, eller för att jämföra två empiriska fördelningar med varandra. Enligt [31] bygger KS-testet på att hitta det maximala avståndet D_n mellan en given teoretisk fördelningsfunktion $F(x)$, som ska testas, och den empiriska fördelningsfunktionen $\hat{F}_n(x)$, som baseras på det observerade datamaterialet. För KS-testet finns det ett viktigt krav, vilket är att den teoretiska fördelningsfunktionen $F(x)$ måste vara en kontinuerlig fördelning. Givet ett sorterat stickprov $x_1, x_2, \ldots, x_n$ och en teoretisk fördelningsfunktion $F(x)$, definieras den empiriska fördelningsfunktionen som andelen observationer som är mindre än eller lika med x. Den uttrycks enligt följande

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{x_i \leq x}, \tag{60}$$

där I är indikatorfunktionen som är 1 om $x_i \leq x$, annars 0. Teststatistikan definieras som

$$D_n = \sup_x \left| F(x) - \hat{F}_n(x) \right|. \tag{61}$$

Det vill säga D_n är det största avståndet mellan den empiriska och teoretiska för-

delningsfunktionen. För att kunna genomföra ett KS-test behöver den teoretiska fördelningsfunktionen $F(x)$ specificeras, vilket kräver att fördelningens parametrar är kända. Dessa parametrar, till exempel medelvärde och standardavvikelse i fallet med en normalfördelning, skattas utifrån det observerade datamaterialet. Två vanliga metoder för att skatta dessa parametrar är maximum likelihood- och momentmetoden, vilka beskrivs närmare i avsnitt 3.5.1 respektive avsnitt 3.5.2. Exempel på teoretiska fördelningar är den generaliserade Paretofördelningen och lognormalfördelningen, samt andra relevanta fördelningar av intresse inom återförsäkring.

Låt $F_n(x)$ beteckna fördelningsfunktionen för D_n under nollhypotesen H_0 och att de n observationerna är oberoende samt har fördelningsfunktionen F, enligt [31] kan då $F_n(x)$ definieras som

$$F_n(x) = \mathbb{P}(D_n \leq x \mid H_0), \quad \text{för } x \in [0, 1], \tag{62}$$

där F_n informellt kallas för KS-fördelningen. Enligt [31] finns det ett flertal metoder för att beräkna den exakta KS-fördelningen. Dock är dessa metoder långsamma och beräkningsmässigt kostsamma för stora n.

Enligt [20] kan därmed KS-fördelningen approximeras med hjälp av gränsvärdesfördelningen

$$\lim_{n\to\infty} \Pr\left(\sqrt{n}D_n \leq x\right) = L(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{0}^{\infty}e^{-(2i-1)^2\pi^2/(8x^2)} \tag{63}$$

För att avgöra om H_0 ska förkastas jämförs det observerade värdet av $\sqrt{n}D_n$ med kritiska värden som erhålls från en tabell, se [32]. Alternativt kan ett p-värde beräknas enligt gränsvärdet i ekvation 63, det vill säga baserat på sannolikheten att $\sqrt{n}D_n$ är minst lika stor som det observerade värdet under H_0. Om p-värdet är mindre än signifikansnivån α, förkastas H_0. I detta arbete används den alternativa metoden, det vill säga p-värdet beräknas med hjälp av den inbyggda funktionen *ks.test()* i `R`, och undersöks om p-värdet är mindre än den klassiska signifikansnivån 5%.

3.6.2 Anderson-Darling-test

Ett annat statistiskt test som används för att bedöma om en given datamängd följer en specifik fördelning är Anderson-Darling-test (AD-test). Till skillnad från KS-testet lägger AD-testet större vikt vid avvikelser i fördelningens svansar, vilket gör det väl lämpat för användning inom området där extremvärden har stor betydelse [32]. Teststatistikan A_n bygger på att vikta den kvadratiska skillnaden mellan den teoretiska fördelningsfunktionen $F(x)$ och den empiriska fördelningsfunktionen $\hat{F}_n(x)$ med en viktfunktion $w(x) = [F(x)(1 - F(x))]^{-1}$. Den generella formeln för teststatistikan ges av

$$A_n = n \int_{-\infty}^{\infty} (\hat{F}_n(x) - F(x))^2 w(x) dF(x), \tag{64}$$

där n är antalet datapunkter, $\hat{F}_n(x)$ är den empiriska fördelningsfunktionen och $F(x)$ är den specificerade fördelningsfunktionen under nollhypotesen, H_0, exempelvis en normalfördelning med angivna parametrar. I praktiken används en diskret approximation av teststatistikan, baserad på sorterade observationer. Givet ett ordnat stickprov $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ definieras teststatistikan som

$$A_n = -n - S, \tag{65}$$

där

$$S = \sum_{i=1}^{n} \frac{(2i-1)}{n} (ln(F(Y_i)) + ln(1 - F(Y_{n+1-i}))). \tag{66}$$

Likt KS-fördelningen är det svårt att hitta fördelningen för teststatistikan A_n. Enligt [21] kan AD-fördelningen approximeras med hjälp av gränsvärdesfördelningen

$$\lim_{n\to\infty} \mathbb{P}(A_n < x) = \frac{\sqrt{2\pi}}{x} \sum_{j=0}^{\infty} \binom{-\frac{1}{2}}{j} (4j+1) e^{-(4j+1)^2\pi^2/(8x)} \int_0^{\infty} e^{-\frac{x}{(1+w)^2}} w^2 (4j+1)^2 \frac{\pi^2}{8x}\, dw. \tag{67}$$

Det är dock viktigt att poängtera att de kritiska värdena för Anderson-Darling-testet endast gäller när fördelningen $F(x)$ är helt specificerad, det vill säga när dess parametrar är kända. I praktiken skattas dock ofta parametrarna till $F(x)$ från datat, vilket påverkar testets fördelning och därmed kräver justerade kritiska värden. I exempelvis `R` finns funktionen *ad.test()* som använder justerade kritiska värden för ett antal vanliga fördelningar med skattade parametrar.

Teststatistikan A_n kan sedan jämföras med kritiska värden för Anderson-Darling-testet, vilket beror på den specificerade fördelningen som testas och valda signifikansnivån α. Dessa kritiska värden finns för vissa fördelningar, såsom normal-, lognormal- och exponentialfördelning, att hitta i [32]. Om teststatistikan A_n överstiger det kritiska värdet för vald signifikansnivå α, förkastas nollhypotesen H_0, vilket innebär att den empiriska fördelningen skiljer sig signifikant från den teoretiska. Om A_n däremot är mindre än det kritiska värdet finns det inte tillräckligt med stöd för att förkasta H_0, och det kan då antas att data följer den specificerade fördelningen. Alternativt kan p-värden användas för att avgöra om H_0 ska förkastas. Om det uträknade p-värdet är mindre än den valda

signifikansnivån, indikerar detta en statistiskt signifikant avvikelse från den teoretiska fördelningen. Även för Anderson–Darling-testet beräknas p-värden med hjälp av den inbyggda funktionen `ad.test()` i `R` och undersöks med avseende på om de är mindre än den klassiska signifikansnivån 5%. All information ovan går att hitta i [6] och [32].

3.6.3 Akaikes informationskriterium (AIC)

För att utvärdera och jämföra olika modeller används Akaikes informationskriterium (AIC). AIC är ett mått på hur väl en modell passar data i förhållande till andra modeller.

Kriteriet bygger på sannolikhetsfunktionen och väger samman modellens anpassning och komplexitet. För att undvika överanpassning införs ett straff för varje extra parameter som ingår i modellen. Detta beror på att modeller med flera parametrar ofta förbättrar anpassningen till data, men risken för överanpassning ökar. AIC beräknas enligt följande

$$\text{AIC} = 2k - 2\log(\hat{\mathcal{L}}(x;\theta)), \tag{68}$$

där k är antalet parametrar i modellen och $\hat{\mathcal{L}}(x;\theta)$ betecknar det maximala värdet av likelihoodfunktionen (se ekvation 42 för likelihoodfunktionen).

Ett lägre AIC-värde indikerar en bättre modell vid jämförelse med andra modeller, då det representerar en bättre balans mellan god anpassning och modellkomplexitet [1, ss.212-213].

3.6.4 Bayesianskt informationskriterium (BIC)

Det Bayesianska informationskriteriet (BIC) liknar AIC men skiljer sig genom att det även tar hänsyn till antalet observationer i datamängden. BIC straffar modeller med fler parametrar hårdare än AIC, vilket gör det särskilt användbart när syftet är att undvika överanpassning. BIC beräknas enligt följande formel

$$\text{BIC} = k\log(n) - 2\log(\hat{\mathcal{L}}(x;\theta)), \tag{69}$$

där k är antalet parametrar i modellen, n är antalet observationer och $\hat{\mathcal{L}}(x;\theta)$ är det maximala värdet av likelihoodfunktionen (se ekvation 42 för likelihoodfunktionen).

Likt AIC är den modell med lägst BIC-värde den modell som föredras, eftersom lägst värde indikerar en bättre passform för modellen.

4 Modellering och resultat

Inledningsvis redogörs för metod och tillvägagångssätt i arbetet, varefter resultatet presenteras.

4.1 Introduktion till modellering

Målet med denna uppsats är att undersöka hur olika självbehållsval påverkar ett försäkringsbolags skadekostnader i en försäkringsportfölj. För detta ändamål utvecklas och analyseras modeller för både antalet skador och skadebeloppen. Eftersom detta arbete baseras på ett försäkringsbolags totala skadekostnader för varje enskild skada, förekommer variation i skadebeloppen inom försäkringsportföljen. Majoriteten av skadorna är relativt små medan några få skadebelopp är mycket stora. Därav görs i modelleringen en uppdelning mellan skadebeloppen, där de större och de mindre skadebeloppen modelleras separat. Eftersom antalet skador antas vara oberoende av skadebeloppen, modelleras dessutom antalet separat från skadebeloppen med hjälp av en Poisson- eller negativ binomialfördelning. Till sist summeras skadebeloppen, eftersom även dessa antas vara oberoende. För att få en uppfattning om hur ett skadeår kan se ut byggs en simuleringsmodell, varefter olika självbehåll appliceras för att utvärdera hur dessa påverkar försäkringsbolagets skadekostnader.

De större skadebeloppen modelleras med både Paretofördelningen, som är en klassisk modell för tungsvansad data, och med extremvärdesteori genom tillämpning av POT-modellen med den generaliserade Paretofördelningen (GPD). Därefter utvärderas modellerna för att avgöra vilken som bäst beskriver data. För de mindre skadebeloppen kommer modelleringen ske för hela portföljen för ett år, det vill säga att skadebeloppen aggregeras årligen och modelleras separat från de större skadorna med klassiska fördelningar såsom normal-, lognormal- och gammafördelning.

Valet av en lämplig brytpunkt mellan stora och små skadebelopp bestäms genom att flera potentiella brytpunkter testas. För varje brytpunkt jämförs de empiriska datapunkterna med den skattade fördelningsfunktionen visuellt, samtidigt som de statistiska testerna Kolmogorov-Smirnov- och Anderson-Darling-test tillämpas för att utvärdera modellernas anpassning.

Vid modellering med hjälp av POT-metoden kommer respektive brytpunkt i tabell 2 ses som ett tröskelvärde. För den generaliserade Paretofördelningen utförs valet av ett lämpligt tröskelvärde genom avvägning mellan bias och varians. Ett lägre tröskelvärde kan leda till systematiska avvikelser, medan ett för högt minskar datamängden och ökar osäkerheten i skattningen.

För den estimerade fördelningsfunktionen behöver parametrar för respektive fördelning skattas. Detta kommer att göras med hjälp av maximum likelihood- eller momentmetoden. För att skatta parametrarna i en Poisson- och negativ binomialfördelning är

momentmetoden ett klassiskt val. Dessutom medför metoden beräkningar som är relativt enkla att genomföra. Med denna bakgrund kommer parametrarna för dessa två fördelningar att skattas med hjälp av momentmetoden. För de resterande fördelningar som används i denna uppsats kommer parametrarna att skattas med hjälp av maximum likelihood-metoden. Modellernas lämplighet utvärderas med både visuella analyser och statistiska tester, där de visuella analyserna består av att estimera parametrar för fördelningarna och jämföra deras fördelningsfunktioner med de empiriska datapunkterna. Utöver Kolmogorov-Smirnov- och Anderson-Darling-test används även AIC- och BIC-värden för att jämföra modellerna sinsemellan.

Till sist utförs 10 000 respektive 100 000 simuleringar för att få en förståelse för hur ett skadeutfall för ett typiskt skadeår kan se ut för ett försäkringsbolag.

Programspråket `R` används för datahantering, beräkningar, visualiseringar och simuleringar med hjälp av färdiga funktioner och paket. Excel används för inflationsjustering samt för nerladdning av inflationsindex.

De metoder som har presenterats ovan utgör grunden för den fortsatta analysen, där resultaten från respektive modell redovisas och utvärderas. Nedan ges en mer detaljerad genomgång av uppdelningen mellan större och mindre skadebelopp, den årsvisa aggregeringen av de mindre skadebeloppen samt hur simuleringarna genomförs.

4.2 Uppdelning av skadebelopp och metod för simulering

Av tabell 1 i avsnitt 2.1 framgår att det finns stora variationer mellan de individuella skadebeloppen. Mot denna bakgrund delas skadebeloppen upp i två delar, mindre och större skadebelopp, vilka modelleras separat.

Det första steget är att finna en lämplig brytpunkt som skiljer de mindre skadebeloppen från de större. Målet är att hitta en optimal gräns så att respektive del kan modelleras med en så passande sannolikhetsfördelning som möjligt. Detta görs genom att systematiskt testa olika brytpunkter, vilket kan ses i tabell 2 nedan.

Tabell 2: Brytpunkter (i miljoner kr).

12	17	21	26	30	35	38	43	47	52

Vi börjar med att definiera X_i som skadebeloppet för en enskild skada. Om Z_i betecknar de skadebelopp som överstiger en given brytpunkt u som ses i tabell 2, kan detta uttryckas med hjälp av en indikatorvariabel enligt följande

$$Z_i = X_i \cdot \mathbb{I}_{\{X_i > u\}}, \tag{70}$$

där X_i är det ursprungliga skadebeloppet för en skada i, och $\mathbb{I}$ är en indikatorfunktion som antar värdet 1 om $X_i > u$ och annars 0. För samtliga observationer Z_i som är större

än 0, skattas parametrar för både en Paretofördelning och en generaliserad Paretofördelning. Syftet är att utvärdera hur väl skadebeloppen över brytpunkterna följer respektive fördelning.

För att uppnå detta syfte härleds fördelningsfunktionerna utifrån de skattade parametrarna och jämförs med de empiriska skadebeloppen. Utöver den visuella jämförelsen används även statistiska goodness-of-fit-tester för att stärka analysen och dra slutsatser om modellernas lämplighet.

För modelleringen av de mindre skadebeloppen definieras Y_i som det individuella skadebeloppet under brytpunkten u, det vill säga

$$Y_i = X_i \cdot \mathbb{I}_{\{X_i \leq u\}}, \tag{71}$$

där X_i är det ursprungliga skadebeloppet för en skada i, och $\mathbb{I}$ är en indikatorfunktion som antar värdet 1 om $X_i \leq u$ och annars 0. På så sätt inkluderas endast skadebeloppen som understiger brytpunkten u.

För mindre skadebelopp fokuserar försäkringsbolagen inte på antalet skador, utan enbart på den totala kostnaden som förväntas utbetalas till kunderna. Dessutom har de mindre skadebeloppen ingen inverkan på hur mycket som återförsäkras så länge självbehållet är större än u, det vill säga de mindre skadorna försäkras inre bort. Av denna anledning, samt att beräkningarna blir enklare, modelleras hela portföljen av de mindre skadebeloppen genom att aggregera ihop dessa per år. Till skillnad från de större beloppen modelleras summan av de mindre beloppen (det vill säga summan av alla Y_i för ett år) direkt, utan att antalet och de individuella skadebeloppen modelleras separat.

Den aggregerade årliga skadekostnaden betecknas då som Y_k där k representerar respektive år mellan 1999-2024. Även för de mindre skadebeloppen kommer olika goodness-of-fit-test och visuella analyser att undersökas och utvärderas.

Vidare antas att de mindre och de större skadebeloppen är oberoende av varandra, och därav kan modellen för den totala skadekostnaden per år, S_k, uttryckas som

$$\mathrm{S}_k = \sum_{i=1}^{N_k} Z_i + Y_k, \tag{72}$$

där k representerar ett år mellan 1999-2024, och N_k är antalet skador per år k som överstiger brytpunkten. Antalet stora skador per år N_k kommer att modelleras med en av de två klassiska fördelningarna, Poisson- eller negativ binomialfördelning. Den främsta skillnaden mellan dessa fördelningar är att negativ binomialfördelningen tillåter överspridning, det vill säga att variansen överstiger väntevärdet. Därför undersöks om det finns tecken på överspridning, i syfte att kunna välja en lämplig modell för att beskriva antalet skador som överstiger brytpunkten u.

Till sist genomförs, baserat på de valda modellerna för skadebelopp och antalet stora

skador, 10 000 respektive 100 000 simuleringar, där varje simulering motsvarar ett möjligt utfall av den totala skadekostnaden per år för försäkringsbolaget. För dessa simuleringar bestäms antalet stora skador slumpmässigt utifrån den fördelning och dess skattade parametrar som bäst överensstämmer med data. Både stora och små skadebelopp kommer att simuleras utifrån olika fördelningar, med skattade parametrar som valts utifrån vad som är mest lämpligt baserat på datamaterialet. Observera att för de mindre skadebeloppen simuleras inte antalet skador utan enbart de årsvis aggregerade skadebeloppen. Varje simulerat utfall genereras alltså genom att slumpa fram antalet stora skador samt både de stora och små skadebeloppen. Därefter summeras skadekostnaderna per simulering för att möjliggöra analys av nyckeltal såsom medelvärde, median och standardavvikelse. Detta tillvägagångssätt för att simulera skador och tillhörande skadebelopp är en välkänd metod som kallas Monte Carlo-simulering.

Enligt Bühlmann (1970) [7, ss.131-144] så användes Monte Carlo-metoden inom försäkringssammanhang ofta för att simulera den totala summan av skadekostnaden under en viss period. Antalet skador modelleras då med en diskret sannolikhetsfördelning, såsom Poisson- eller negativ binomialfördelning, medan enskilda skadebelopp simuleras från en kontinuerlig fördelning, exempelvis lognormal- eller Paretofördelning. Den aggregerade totala skadan beräknas genom att summera de simulerade skadebeloppen för respektive simulering. Denna typ av modellstruktur kallas för den traditionella sammanlagda riskmodellen och passar väl för Monte Carlo-simulering.

4.3 Resultat

Efter att ha introducerat modelleringen är det nu dags att tillämpa modellerna och presentera resultaten.

4.3.1 Modellering av större skadebelopp

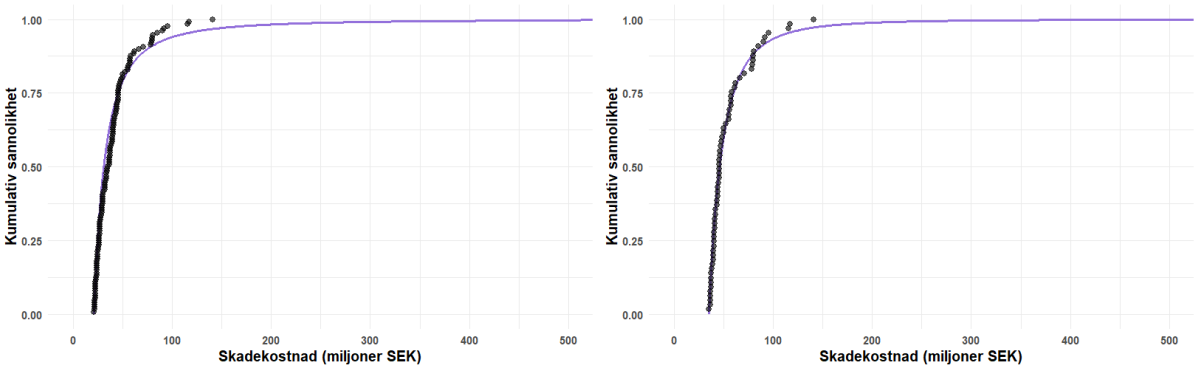
Analysen inleds med att undersöka om Paretofördelningen, som är en klassisk modell för data med tunga svansar och ofta används inom försäkringsbranschen, är lämplig för att modellera större skadebelopp. Vid parameterestimering av en Paretofördelning finns två huvudsakliga angreppssätt. Det första innebär att både formparametern x_m och skalparametern α skattas med en lämplig metod. Det andra bygger på att formparametern x_m sätts lika med den förutbestämda brytpunkten u, varpå endast skalparametern α skattas. Detta är möjligt eftersom formparametern även fungerar som den lägsta möjliga observationen, vilket gör att den i praktiken fungerar som en lägesparameter. Vid analys av överskott är det vanligt att sätta $x_m = u$, medan det vid modellering av hela populationen generellt är mer lämpligt att skatta även formparametern. Följande avsnitt analyserar hur modellresultaten påverkas beroende på om formparametern sätts till brytpunkten eller skattas utifrån data.

Analysen inleds med fallet där $x_m = u$, där u motsvarar varje brytpunkt som redovisas i tabell 2, och endast α skattas. Skattningen av skalparametern sker med hjälp av maximum likelihood-metoden som beskrivs närmare i avsnitt 3.5.1. Därefter har en fördelningsfunktion och QQ-plot, baserad på den skattade α-parametern och de angivna brytpunkterna i tabell 2, anpassats till skadebeloppen som överstiger respektive brytpunkt i syfte att grafiskt undersöka hur väl dessa överensstämmer med fördelningen.

I tabell 3 visas de parametrar som används för att anpassa de fördelningsfunktionerna för respektive brytpunkt.

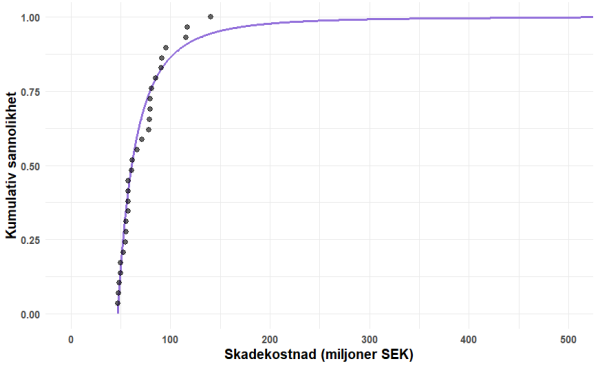
Tabell 3: Parametrarna för Paretofördelningen då formparametern $x_m = u$

Skador	**Parameter**	**Värde**
Över 21 MSEK	α	1,8028
	x_m	21 000 000
Över 35 MSEK	α	2,5830
	x_m	35 000 000
Över 47 MSEK	α	2,6373
	x_m	47 000 000



(a) Brytpunkt 21 MSEK

(b) Brytpunkt 35 MSEK



(c) Brytpunkt 47 MSEK

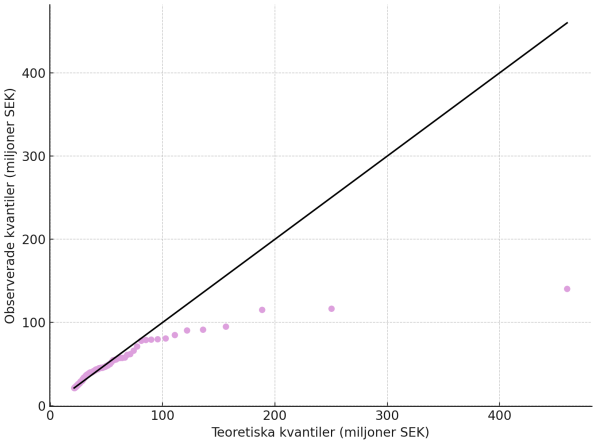
Figur 3: Fördelningsfunktioner för olika brytpunkter för Paretofördelningen med $x_m = u$.

I figur 3 visas den skattade fördelningsfunktionen (den lila kurvan) för en Paretofördelning utifrån parametrarna i tabell 3 samt de empiriska datapunkterna (de svarta punkterna). På x-axeln visas skadekostnaden i miljoner kronor för varje enskild skada, och på y-axeln återges den kumulativa sannolikheten, vilken visar andelen observationer som är mindre än eller lika med ett visst skadebelopp.

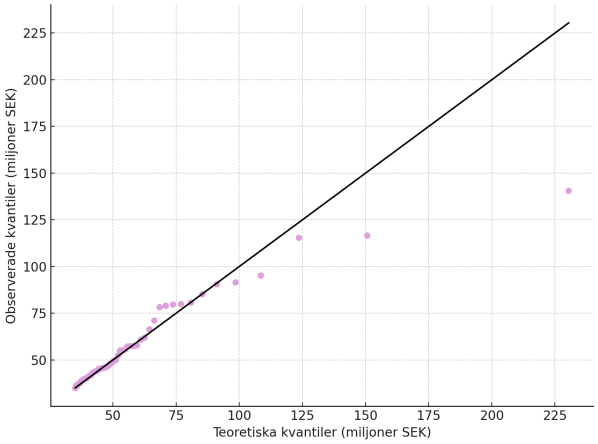
Varje svart punkt representerar observerad skadekostnad från det empiriska datamaterialet, tillsammans med dess empiriska kumulativa sannolikhet, det vill säga vilken kvantil det tillhör bland de observerade skadorna. Exempelvis framgår det av figur 3b att ett skadebelopp strax under 100 miljoner kronor motsvarar ungefär 0,95-kvantil på y-axeln, vilket innebär att cirka 95% av skadorna i datamängden som är större än 35 miljoner kronor är mindre än eller lika med 100 miljoner kronor.

En viss avvikelse mellan den skattade fördelningsfunktionen och de empiriska datapunkterna kan observeras för de största skadebeloppen. Detta är dock inte oväntat, eftersom extremt stora skador är sällsynta och därför representeras av få observationer i datamängden. Vid simulering från en fördelning är det vanligt att just de största värdena uppvisar störst avvikelse från den teoretiska fördelningen, vilket kan förklaras av deras natur som sällsynta men stora utfall.

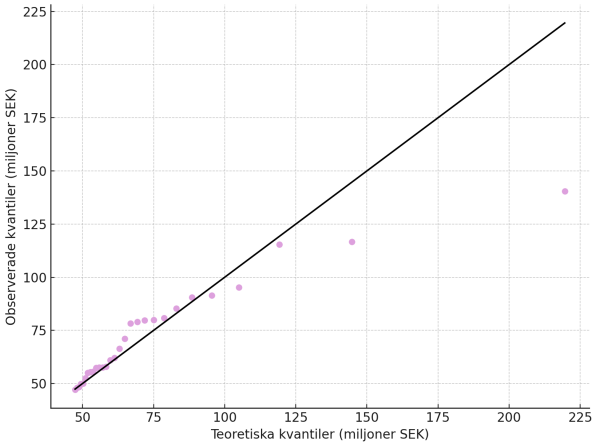
Som illustreras i figur 3c, ökar avvikelsen mellan de empiriska datapunkterna och den skattade Paretofördelningens fördelningsfunktion vid högre brytpunkter. Samtidigt minskar antalet datapunkter som ingår i analysen, vilket leder till en ökad osäkerhet i parameterskattningen. Kombinationen av större avvikelser och färre observationer påverkar modellens stabilitet negativt och minskar dess tillförlitlighet.



(a) Brytpunkt 21 MSEK



(b) Brytpunkt 35 MSEK



(c) Brytpunkt 47 MSEK

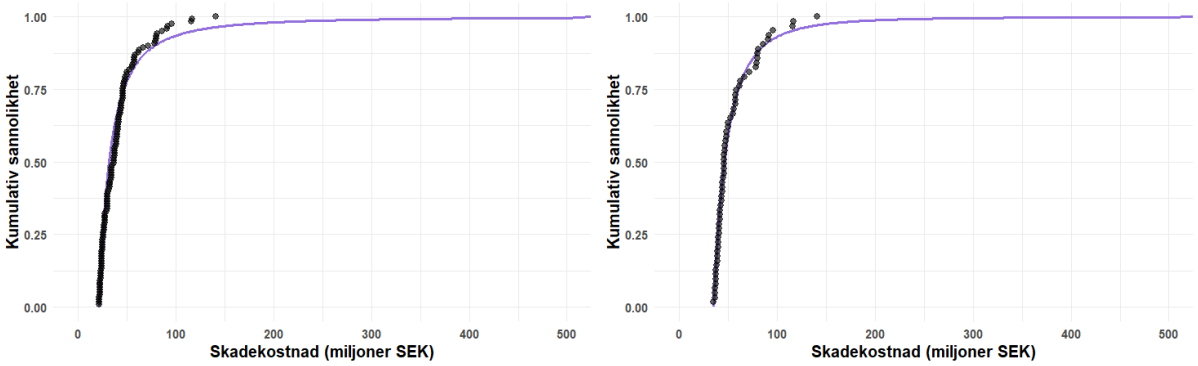
Figur 4: QQ-plot för olika brytpunkter för Paretofördelningen med $x_m = u$.

Figur 4 visar QQ-plottar för olika brytpunkter för Paretofördelningen. I QQ-plotten jämförs de teoretiska kvantilerna med de observerade kvantilerna. Generellt ligger de högre och mer extrema skadebeloppen under linjen, vilket indikerar att Paretofördelningen tenderar att överskatta de mest extrema värdena. För skadebelopp som ligger strax över brytpunkten tycks Paretofördelningen däremot ge en god passning. Det kan även noteras att passningen för de extrema värdena försämras något när en högre brytpunkt används.

För att se fördelningsfunktionerna för andra brytpunkter hänvisas läsaren till appendix. Vidare skattas både formparametern x_m och skalparametern α med hjälp av maximum likelihood-metoden för att undersöka om en bättre anpassning kan uppnås. De skattade parametrarna kan ses i tabell 4.

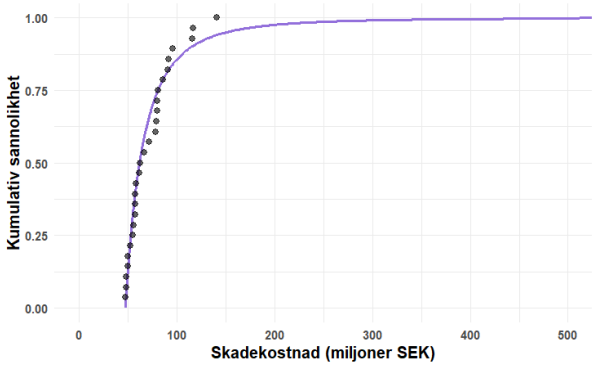
Tabell 4: Parametrarna för Paretofördelningen då formparameterm x_m skattas.

Skador	Parameter	Värde
Över 21 MSEK	α	1,8028
	x_m	21 236 039
Över 35 MSEK	α	2,5830
	x_m	35 003 742
Över 47 MSEK	α	2,6373
	x_m	47 088 187



(a) Brytpunkt 21 MSEK

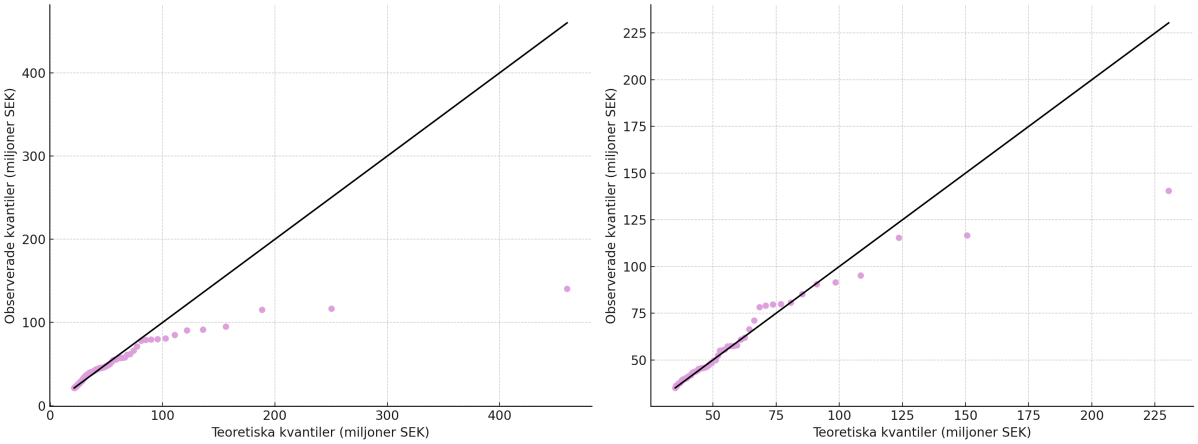
(b) Brytpunkt 35 MSEK



(c) Brytpunkt 47 MSEK

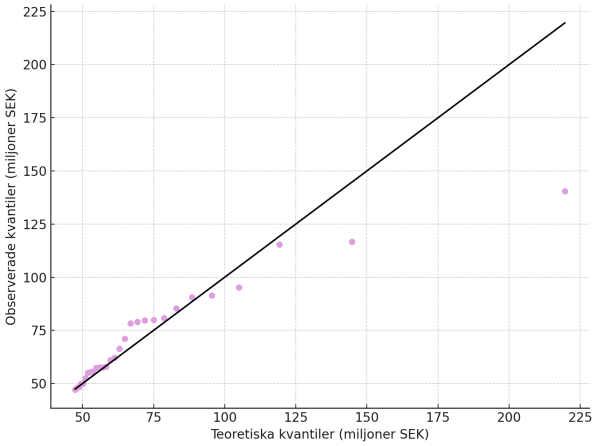
Figur 5: Fördelningsfunktioner för olika brytpunkter för Paretofördelningen då formparametern x_m skattas.

I figur 5 visas den skattade fördelningsfunktionen (den lila kurvan) för en Paretofördelning utifrån de skattade parametrarna i tabell 4 samt de empiriska datapunkterna (de svarta punkterna). Även i detta fall representerar varje svart punkt en observerad skadekostnad från det empiriska datamaterialet, tillsammans med dess empiriska kumulativa sannolikhet, alltså vilken kvantil det tillhör bland de observerade skadorna.



(a) Brytpunkt 21 MSEK

(b) Brytpunkt 35 MSEK



(c) Brytpunkt 47 MSEK

Figur 6: QQ-plot för olika brytpunkter för Paretofördelningen då formparametern x_m skattas.

Figur 6 visar QQ-plottar för olika brytpunkter för Paretofördelningen när formparametern x_m skattas. För skadebelopp som ligger strax över brytpunkten tycks Paretofördelningen ge en god passning. För högre och mer extrema skadebelopp ligger de observerade värdena generellt under linjen, vilket indikerar att fördelningen tenderar att överskatta de mest extrema skadebeloppen.

Det kan konstateras att resultaten inte skiljer sig nämnvärt beroende på om formparametern x_m skattas eller om den sätts lika med respektive brytpunkt. Detta kan motiveras med att maximum likelihood-skattningen av formparametern x_m blir det minsta observerade värdet, vilket vanligtvis ligger mycket nära den valda brytpunkten. Därför kan liknande slutsatser dras oavsett metod, och vid modellering av skadebelopp som överstiger en brytpunkt (så kallat överskott) framstår det som en enkel och rimlig strategi att sätta formparametern lika med den valda brytpunkten, det vill säga $x_m = u$.

Mot denna bakgrund tillämpas extremvärdesteorin, som beskrevs i avsnitt 3.4, med särskilt fokus på POT-modellen, för att undersöka om de större skadebeloppen istället kan modelleras med en generaliserad Paretofördelning. Observera att varje brytpunkt i tabell 2 nu kommer att representera respektive tröskelvärde vid modellering av den generaliserade Paretofördelningen. Därav likställs tröskelvärdena med brypunkterna.

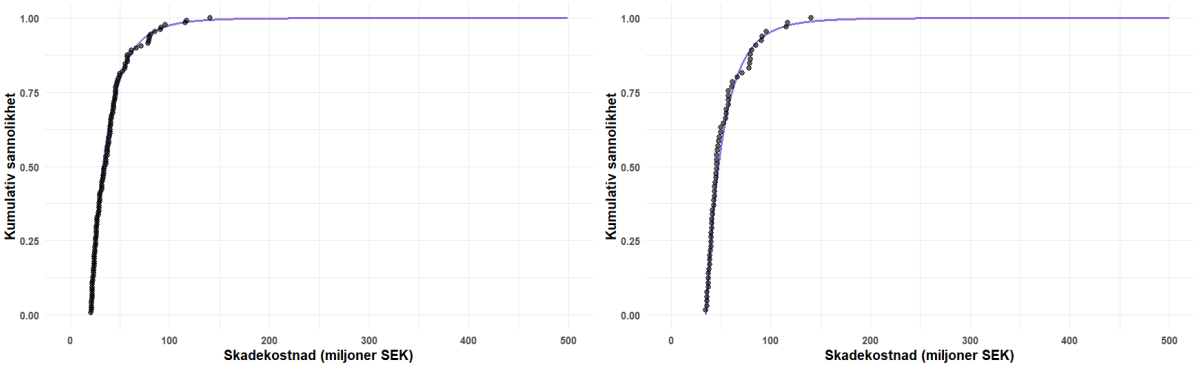
Som påpekas i avsnitt 3.4.2 nämner Embrechts med flera att det är lämpligt att använda grafiska metoder för att undersöka hur de skattade parametrarna i en generaliserad Paretofördelning påverkas av olika val av brytpunkter, vilket här efterföljs.

För varje vald brytpunkt skattas formparametern ξ och skalparametern β i den generaliserade Paretofördelningen. Lägesparametern μ sätts lika med den aktuella brytpunkten u. Skattningarna för parametrarna ξ och β genomförs för varje lämplig brytpunkt som ses i tabell 2 med hjälp av maximum likelihood-metoden. De skattade parametrarna för brytpunkterna 21, 35 och 47 miljoner kan ses i tabell 5.

De estimerade parametrarna används därefter för att konstruera den skattade fördelningsfunktionen. Syftet är att grafiskt utvärdera hur väl de överskjutande skadebeloppen följer en generaliserad Paretofördelning givet olika brytpunkter.

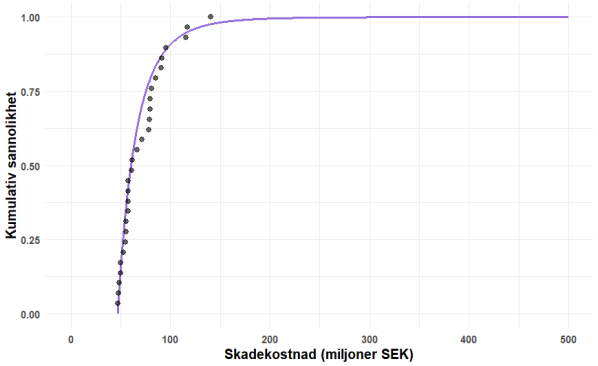
Tabell 5: Parametrar fö generaliserad Paretofördelningen.

Parameter	**Värde**
μ	21 MSEK
ξ	0,1299
β	16 611 205
μ	35 MSEK
ξ	0,1360
β	17 195 132
μ	47 MSEK
ξ	0,1631
β	18 249 748



(a) Brytpunkt 21 MSEK

(b) Brytpunkt 35 MSEK

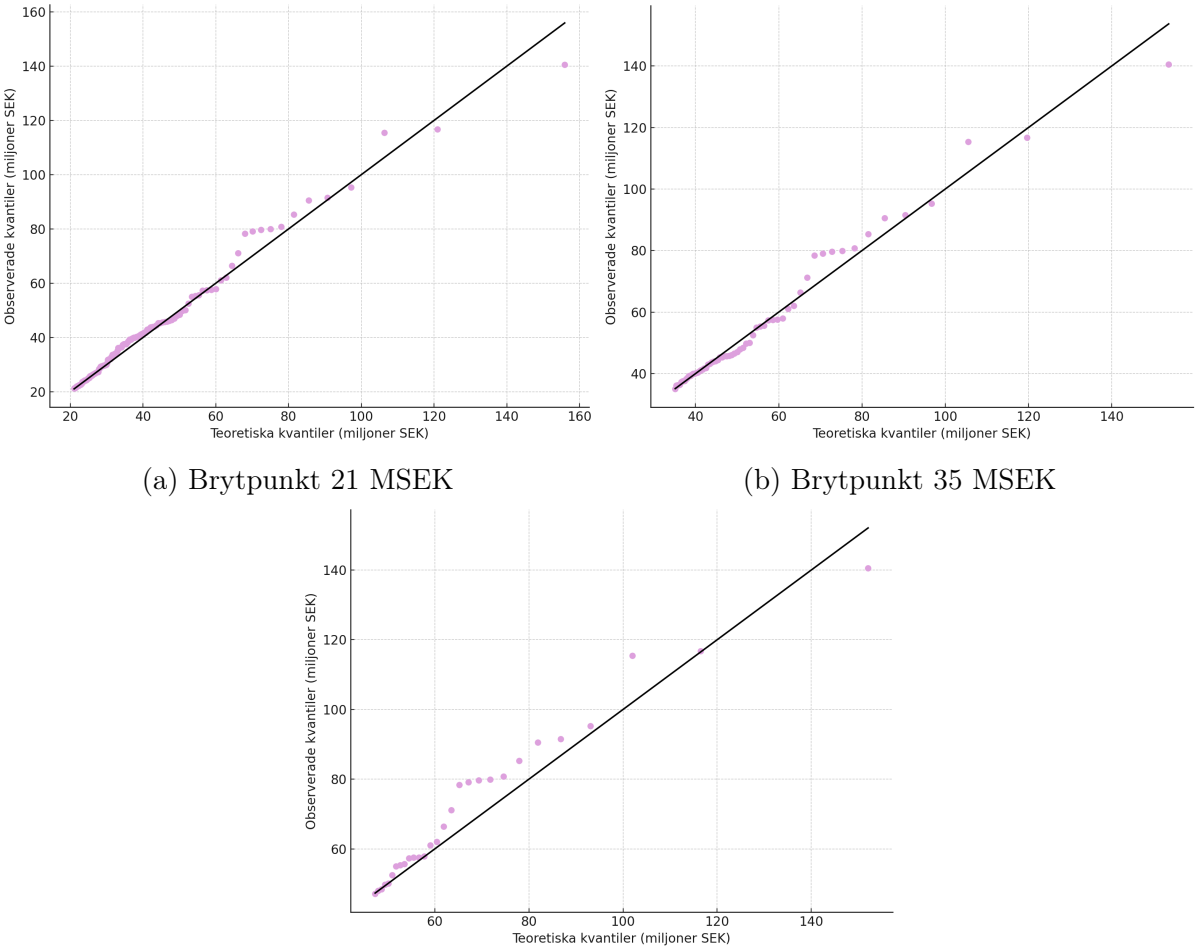


(c) Brytpunkt 47 MSEK

Figur 7: Fördelningsfunktioner för olika brytpunkter för den generaliserade Paretofördelningen.

I figur 7 visas den skattade fördelningsfunktionen (den lila kurvan) baserad på de skattade parametrarna i tabell 5, tillsammans med svarta punkter som representerar skadebelopp från de empiriska datapunkterna. Observera att figur 7 visar de skattade fördelningsfunktionerna där brytpunkterna är inkluderade, vilket medför en förskjutning av kurvan. Detta görs främst för att kunna få figuren i jämförbar form som i figur 3 och figur 5. Även i detta fall representerar en svart punkt den observerade skadekostnaden från det empiriska datamaterialet, tillsammans med dess empiriska kumulativa sannolikhet, det vill säga vilken kvantil de tillhör bland de observerade skadorna.

Det framgår att de överskjutande skadebeloppen över 21 respektive 35 miljoner kronor överensstämmer väl med den skattade fördelningsfunktionen för en generaliserad Paretofördelning. Däremot visar figur 7c att de empiriska skadebeloppen börjar avvika från den skattade fördelningsfunktionen vid högre brytpunkter, såsom 47 miljoner kronor. Detta är dock väntat, med samma argumentation som för Paretofördelningen, extremt stora skador är sällsynta och representeras därför av få observationer i datamängden. Det begränsade antalet observationer kan i sin tur påverka modellens tillförlitlighet.



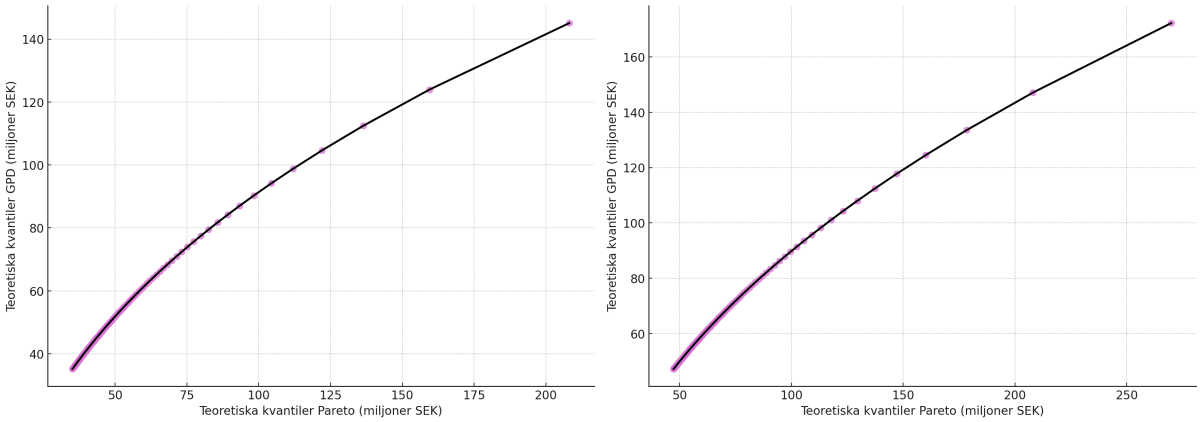
(a) Brytpunkt 21 MSEK

(b) Brytpunkt 35 MSEK

(c) Brytpunkt 47 MSEK

Figur 8: QQ-plot för olika brytpunkter för generaliserade Paretofördelningen.

Figur 8 visar QQ-plottar för olika brytpunkter för den generaliserade Paretofördelningen. I QQ-plotten jämförs de teoretiska kvantilerna med de observerade kvantilerna. Från QQ-plottarna ser vi att den generaliserade Paretofördelningen ger klart bättre anpassning till de teoretiska kvantilerna jämfört med Paretofördelningen. Den förbättrade passningen gäller både för skadebelopp strax över brytpunkten och för högre skadebelopp. Vidare i figur 9 jämförs kvantilerna från den anpassade Paretomodellen med kvantilerna från den anpassade generaliserade Paretomodellen.



(a) Brytpunkt 35 MSEK (b) Brytpunkt 47 MSEK

Figur 9: Jämförelse av kvantiler mellan GPD och Pareto-modellerna.

I figur 9 plottas kvantilerna för den skattade generaliserade Paretofördelningen mot kvantilerna för den skattade Paretofördelningen. Brytpunkten vid 21 miljoner utesluts från jämförelsen eftersom den avviker markant för en Paretofördelning. I de lägre och mellersta kvantilerna följer punkterna referenslinjen väl, vilket visar att båda modellerna beskriver de vanliga utfallen väl. Observera att varje lila punkt är ett par av kvantiler från en Pareto- och generaliserad Pareto-modell. Men i den högra svansen, där de största och mest kritiska skadorna finns, avviker Pareto-kvantilen ofta från GPD-kvantilen och tenderar att ligga högre. Utifrån dessa QQ-plottar ser vi att Pareto-kvantilerna tenderar att överskatta de extrema händelserna jämfört med den generaliserade Pareto-kvantilen.

Vid en visuell jämförelse mellan Paretofördelningen och den generaliserade Paretofördelningen framgår det att avvikelserna är större för Paretofördelningen, både när formparametern skattas och när den sätts lika med den valda brytpunkten, särskilt vid höga skadebelopp. En visuell analys av figur 3 – figur 9 visar att den generaliserade Paretofördelningen ger en bättre anpassning till de empiriska skadebeloppen, och framstår därmed som den mest lämpliga modellen för att beskriva de större skadebeloppen i portföljen. För att stärka denna slutsats används därefter olika statistiska mått för modellutvärdering.

För att se de estimerade fördelningsfunktionerna samt de skattade parametrarna för övriga brytpunkter hänvisas läsaren till appendix. Notera att appendix inte innehåller resultat för samtliga brytpunkter som presenteras i tabell 2, eftersom resultaten förändras marginellt vid små justeringar av brytpunkter, exempelvis från 12 till 17 miljoner kronor eller från 47 till 52 miljoner kronor.

Modellutvärdering

För att få en bättre förståelse genomförs goodness-of-fit-tester såsom Anderson-Darling- och Kolmogorov-Smirnov-test. Dessa tester har utförts i `R` genom att använda de inbyggda funktionerna *ad.test* och *ks.test*. Båda dessa tester har använts i syfte att undersöka

huruvida skadebeloppen som överstiger brytpunkterna följer en generaliserad Paretofördelning eller en Paretofördelning. I båda testerna är nollhypotesen, H_0, att data följer den givna fördelningen, medan alternativhypotesen, H_1, innebär att detta inte är fallet.

Till skillnad från KS-testet lägger AD-testet större vikt vid fördelningens svansar och kommer därför att vara avgörande vid valet av brytpunkt för både Pareto- och generaliserade Paretofördelningen. Teststatistikan och p-värdet för respektive test redovisas i tabell 6, tabell 7 och tabell 8. Som första steg genomförs AD- och KS-test för en Paretofördelning där formparametern $x_m = u$. Resultatet presenteras i tabell 6.

Tabell 6: Teststatistika och p-värde från Anderson–Darling- och Kolmogorov–Smirnov-test för olika brytpunkter, under antagandet att $x_m = u$ för Paretofördelning.

Paretofördelning då $x_m = u$					
Brytpunkt (u)	**Anderson-Darling-test**		**Kolmogorov-Smirnov-test**		**Slutsats**
	Teststatistika	**p-värde**	**Teststatistika**	**p-värde**	
12 MSEK	3,1909	0,0220	0,1795	0,0129	Förkasta H_0
17 MSEK	2,5661	0,0458	0,1253	0,0359	Förkasta H_0
21 MSEK	2,5310	0,0478	0,1496	0,0244	Förkasta H_0
26 MSEK	1,8933	0,1053	0,0877	0,1333	Förkasta ej H_0
30 MSEK	1,9319	0,1002	0,0604	0,2581	Förkasta ej H_0
35 MSEK	0,7541	0,5186	0,0873	0,6716	Förkasta ej H_0
38 MSEK	0,6847	0,4515	0,0706	0,4862	Förkasta ej H_0
43 MSEK	0,7629	0,4913	0,1223	0,5318	Förkasta ej H_0
47 MSEK	0,9143	0,4212	0,1541	0,4518	Förkasta ej H_0
52 MSEK	1,2306	0,3171	0,1979	0,2671	Förkasta ej H_0

Med hjälp av de färdiga programvarorna i `R` beräknas både p-värdet och teststatistikan för respektive test som kan ses i tabell 6. I programvaran beräknas teststatistikan enligt ekvation 61 för KS-testet och enligt ekvation 65 för AD-testet. För att avgöra huruvida nollhypotesen, H_0, ska förkastas tillämpas en signifikansnivå på 5%, och p-värdet för respektive test jämförs med denna nivå. Från tabell 6 framgår att p-värdena för både Anderson–Darling- och Kolmogorov–Smirnov-testerna är statistiskt signifikanta på en 5%-nivå för brytpunkterna 12, 17 och 21 miljoner. Detta innebär att nollhypotesen om att skadebeloppen följer en Paretofördelning för dessa brytpunkter förkastas. För samtliga övriga brytpunkter finns däremot inte tillräckligt stöd för att förkasta nollhypotesen. Vidare undersöks om testresultaten påverkas av valet att skatta formparametern x_m, samt hur resultaten ser ut för en generaliserad Paretofördelning. Resultatet presenteras i tabell 7 och tabell 8.

Tabell 7: Teststatistika och p-värde från Anderson–Darling- och Kolmogorov–Smirnov-tester för olika brytpunkter, då form- och skalparametern skattas.

Paretofördelning då x_m skattas					
Brytpunkt (u)	**Anderson-Darling-test**		**Kolmogorov-Smirnov-test**		**Slutsats**
	Teststatistika	**p-värde**	**Teststatistika**	**p-värde**	
12 MSEK	2,4532	0,0479	0,1524	0,0272	Förkasta H_0
17 MSEK	1,6956	0,1360	0,0889	0,1472	Förkasta ej H_0
21 MSEK	1,1156	0,0746	0,1267	0,0519	Förkasta ej H_0
26 MSEK	1,6926	0,1366	0,0590	0,3324	Förkasta ej H_0
30 MSEK	2,0906	0,0821	0,1469	0,0738	Förkasta ej H_0
35 MSEK	0,5156	0,7302	0,0826	0,8092	Förkasta ej H_0
38 MSEK	0,9142	0,4051	0,0867	0,7699	Förkasta ej H_0
43 MSEK	0,9447	0,3871	0,1190	0,5666	Förkasta ej H_0
47 MSEK	0,9404	0,3666	0,1605	0,4226	Förkasta ej H_0
52 MSEK	1,0674	0,3230	0,1894	0,3145	Förkasta ej H_0

Tabell 8: Teststatistika och p-värde från Anderson–Darling- och Kolmogorov–Smirnov-tester för olika brytpunkter.

Generaliserad Paretofördelning					
Brytpunkt (u)	**Anderson-Darling-test**		**Kolmogorov-Smirnov-test**		**Slutsats**
	Teststatistika	**p-värde**	**Teststatistika**	**p-värde**	
12 MSEK	5,2936	0,0021	0,1005	0,0111	Förkasta H_0
17 MSEK	1,9316	0,1003	0,1417	0,0856	Förkasta ej H_0
21 MSEK	1,8253	0,1150	0,1350	0,1229	Förkasta ej H_0
26 MSEK	1,5007	0,1763	0,1259	0,1057	Förkasta ej H_0
30 MSEK	0,8437	0,3871	0,0853	0,3418	Förkasta ej H_0
35 MSEK	0,4754	0,7714	0,0808	0,7589	Förkasta ej H_0
38 MSEK	0,5572	0,6886	0,1071	0,5076	Förkasta ej H_0
43 MSEK	0,8433	0,4361	0,1279	0,4921	Förkasta ej H_0
47 MSEK	0,8923	0,4181	0,1930	0,1976	Förkasta ej H_0
52 MSEK	0,9795	0,3673	0,2332	0,1246	Förkasta ej H_0

Utifrån tabell 7 och tabell 8 framgår att p-värdena för både Anderson-Darling- och Kolmogorov-Smirnov-testerna är statistiskt signifikanta på en 5%-nivå för brytpunkten 12 miljoner. Detta innebär att nollhypotesen om att skadebeloppen följer en generaliserad Paretofördelning eller en Paretofördelning för just denna brytpunkt förkastas. För samtliga övriga brytpunkter finns däremot inte tillräckligt stöd för att förkasta nollhypotesen. Vid en första anblick kan en brytpunkt på 12 miljoner kronor ge intryck av att det finns tillräckligt många datapunkter för att modellera skadebeloppen på ett tillförlitligt sätt. Det är dock viktigt att komma ihåg att ett stort antal observationer inte alltid innebär att

en exempelvis generaliserad Paretofördelning eller Paretofördelning passar bra. Det som spelar störst roll är om data verkligen är representativ för de extrema skadehändelserna som modelleras. Om en stor andel av observationerna ligger precis ovanför brytpunkten, men inte fångar in de riktigt stora skadorna, finns det en risk att parametrarna blir missvisande och att modellen underskattar hur stora de mest extrema utfallen kan bli.

Vidare visar tabell 7 och tabell 8 att p-värdena tenderar att vara lägre både vid mycket låga och mycket höga brytpunkter, för såväl Paretofördelningen som den generaliserade Paretofördelningen. Den brytpunkt som genererar det högsta p-värdet i tabell 8 är 35 miljoner kronor för den generaliserade Paretofördelningen, vilket tyder på att skadebelopp som överskrider denna nivå uppvisar bäst överensstämmelse med en generaliserad Paretofördelning. Liksom för den generaliserade Paretofördelningen framgår det av tabell 7 att den brytpunkt som genererar det högsta p-värdet är 35 miljoner kronor. Detta tyder på att skadebelopp som överstiger denna nivå uppvisar bäst överensstämmelse med en Paretofördelning.

Utifrån tabell 7 och tabell 8 kan det även noteras att p-värdena generellt sett är något högre för den generaliserade Paretofördelningen än för Paretofördelningen. Även om p-värdena för de flesta brytpunkter inte är statistiskt signifikanta på 5%-nivån, tyder dessa resultat troligen på att den generaliserade Paretofördelningen ger en något bättre anpassning till data. För en ytterligare jämförelse av modellerna med brytpunkt 35 miljoner har AIC- och BIC-värden beräknats enligt ekvation 68 respektive ekvation 69, vilket redovisas i tabell 9.

Tabell 9: AIC- och BIC-värden för Pareto- och generaliserade Paretofördelningen med brytpunkt 35 miljoner.

Fördelning	**AIC**	**BIC**
Paretofördelning	2428,44	2434,96
Generaliserade Paretofördelning	2321,19	2327,71

Av tabell 9 framgår det att den generaliserade Paretofördelningen ger något lägre AIC- och BIC-värden jämfört med Paretofördelningen. Detta indikerar att den generaliserade Paretofördelningen förmodligen ger en bättre anpassning till skadebeloppen som överstiger 35 miljoner kronor. Baserat på dessa resultat modelleras därför samtliga skadebelopp över 35 miljoner kronor med en generaliserad Paretofördelning.

4.3.2 Modellering av antalet stora skador

Eftersom de mindre skadebeloppen modelleras på aggregerad nivå per år, det vill säga genom att hela portföljen summeras årsvis, analyseras inte enskilda skador separat. Därför modelleras inte heller antalet skador för dessa skadebelopp.

För de större skadebeloppen, som hanteras individuellt, behöver även antalet skador modelleras. Detta eftersom varje enskild stor skada har en påverkan på det totala utfallet och därmed på hur mycket försäkringsbolaget själv måste betala innan återförsäkringen träder in.

Vid modellering av antalet stora skador, N, är den klassiska utgångspunkten att anta att antalet följer en Poissonfördelning. Johansson (2014) nämner [19, ss.70-71] att det inom POT-modellen ofta antas att N är Poissonfördelad med parameter λ. För att denna modell ska vara lämplig krävs det dock att skadorna inträffar oberoende av varandra och att variansen är lika med väntevärdet, det vill säga $\mathbb{E}[N] = \text{Var}[N] = \lambda$.

Om datat däremot uppvisar överspridning, det vill säga att variansen överstiger väntevärdet, är en alternativ modell att föredra. Ett vanligt val i sådana fall är den negativa binomialfördelningen, som tillåter att $\text{Var}[N] > \mathbb{E}[N]$.

För att undersöka vilken modell som är mest lämpligt beräknas väntevärde och varians för antalet skador som överstiger 35 miljoner kronor. I tabell 10 redovisas parametrarna för Poissonfördelningen. I tabell 11 presenteras motsvarande parametrar, väntevärde och varians för den negativa binomialfördelningen. Väntevärdet och variansen för den negativa binomialfördelningen har beräknats enligt ekvation 14 och ekvation 15.

Tabell 10: Estimerat väntevärde och varians för Poissonfördelningen.

Parametrar	**Värde**
$\mathbb{E}[N] = \text{Var}[N] = \lambda$	2,5

Tabell 11: Estimerade parametrar, väntevärde och varians för den negativa binomialfördelningen.

Parametrar	**Värde**
r	5,208
p	0,676
$\mathbb{E}[N]$	2,5
$\text{Var}[N]$	3,7
$\text{Var}[N] > \mathbb{E}[N]$	Ja
Lämplig fördelning	**Neg Bin**

Av tabell 11 framgår att $\text{Var}[N] > \mathbb{E}[N]$, vilket tyder på överspridning i data. Med överspridning avses att variansen överstiger väntevärdet, något som Poissonfördelningen inte kan hantera, eftersom varians och väntevärde alltid är lika för denna fördelning. Den negativa binomialfördelningen har en extra spridningsparameter, r, som gör det möjligt att modellera överspridning. Därmed är den negativa binomialfördelningen en

mer lämplig modell i detta fall för att modellera antalet skador över 35 miljoner kronor. Parametrarna i tabell 10 respektive tabell 11 beräknas med hjälp av momentmetoden som beskrivs närmare i avsnitt 3.5.2.

För att ytterligare bedöma vilken fördelning som bäst modellerar antalet stora skador används AIC- och BIC-värden som jämförelsemått. I tabell 12 framgår dessutom att den negativa binomialfördelningen uppvisar något lägre AIC- och BIC-värden, vilket gör den till en mer lämplig modell för antalet skador.

Tabell 12: AIC- och BIC-värden för Poisson- och negativ binomialfördelningen.

Modell	**AIC**	**BIC**
Poissonfördelning	105,18	107,70
Negativ binomialfördelning	104,69	105,95

Med hänsyn till att variansen i datamaterialet överstiger medelvärdet, samt att AIC- och BIC-testet indikerar en något bättre modellanpassning för den negativa binomialfördelningen, väljs denna som modell för antalet stora skador.

4.3.3 Modellering av mindre skadebelopp

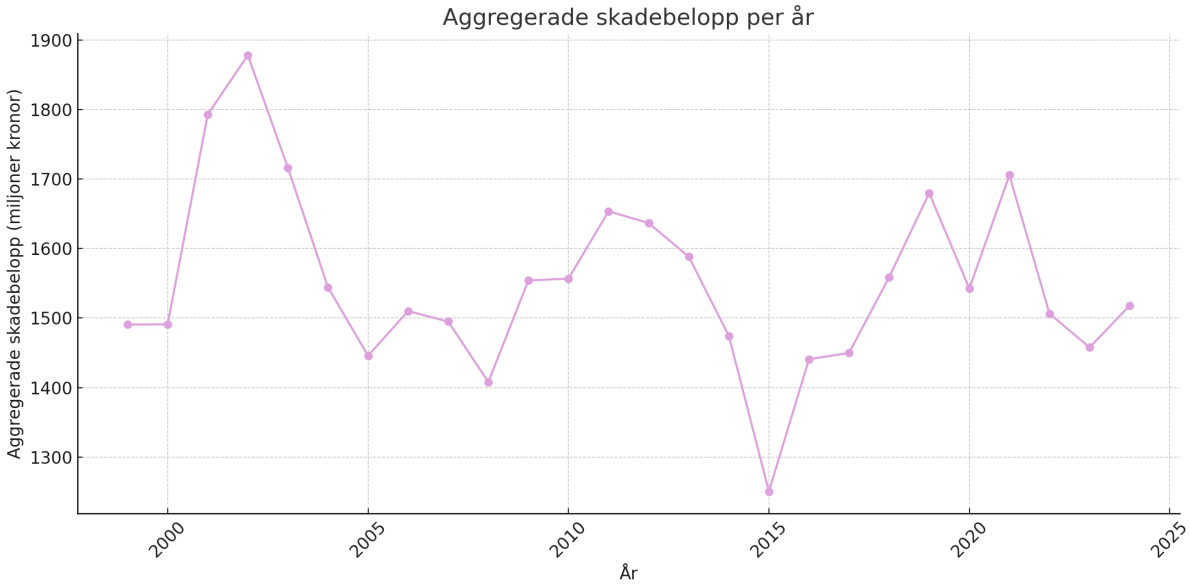
För modellering av de mindre skadebeloppen används tre klassiska fördelningar, nämligen normal-, lognormal- och gammafördelning. Efter olika analyser i avsnitt 4.3.1 identifierades 35 miljoner kronor som en rimlig brytpunkt, och alla skadebelopp över detta värde modelleras med en generaliserad Paretofördelning. Nästa steg är att undersöka om skadebeloppen under 35 miljoner kronor kan modelleras med någon av de tre klassiska fördelningarna.

Modelleringen av de mindre skadebeloppen, definierade som individuella skador under 35 miljoner kronor, har utförts på årlig aggregerad nivå för perioden 1999–2024. Modelleringen har genomförts på ihopaggregerad nivå eftersom figur 10 inte visar någon tydlig trend. De årliga, aggregerade skadebeloppen framgår av tabell 13.

För de två senaste åren, 2023 och 2024, finns det vissa skadehändelser under 35 miljoner kronor som ännu inte har rapporterats. Detta beror oftast på att det förekommer en fördröjning mellan att en skada inträffar och att den faktiskt rapporteras till försäkringsbolaget. För att få en mer representativ bild av skadebeloppen för dessa år har de aggregerade värdena justerats med hänsyn till *IBNR* (Incurred But Not Reported), vilket avser skador som har inträffat men ännu inte har rapporterats till försäkringsbolaget. Detta innebär att en uppskattning har lagts till för de skador som har inträffat men ännu inte hunnit rapporteras. Syftet med att inkludera IBNR är att de aggregerade skadebeloppen för 2023 och 2024 ska bli jämförbara med tidigare år, för vilka mer fullständig data finns tillgänglig.

Tabell 13: Aggregerade skadebelopp per år.

Årtal	Aggregerade skadebelopp
1999	1 490 520 687
2000	1 490 924 888
2001	1 792 613 006
2002	1 878 000 188
2003	1 715 959 459
2004	1 544 102 493
2005	1 445 880 985
2006	1 510 028 655
2007	1 494 800 862
2008	1 407 932 341
2009	1 554 052 571
2010	1 556 542 243
2011	1 653 540 774
2012	1 636 473 429
2013	1 588 269 896
2014	1 474 374 256
2015	1 250 634 884
2016	1 440 718 184
2017	1 449 747 105
2018	1 558 845 716
2019	1 679 634 257
2020	1 542 349 380
2021	1 705 429 919
2022	1 506 283 705
2023	1 457 710 701
2024	1 517 889 536



Figur 10: Aggregerade skadebelopp per år.

Av figur 10 och tabell 13 framgår det att skadebeloppen är relativt jämnt fördelade över åren utan någon tydlig trend, vilket indikerar att försäkringsbolagets utfall har varit stabilt över tid. Om det däremot hade funnits en uppåt- eller nedåtgående trend i skadebeloppen, hade det varit viktigt att ta hänsyn till detta i modellen. Att bortse från en sådan trend riskerar annars att leda till missvisande slutsatser.

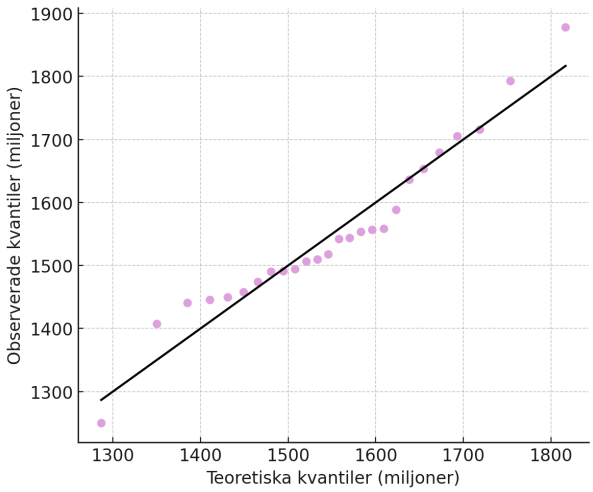
Vid modellering av mindre skadebelopp finns flera möjliga sannolikhetsfördelningar att använda. De klassiska valen är normal-, lognormal- och gammafördelningen.

Enligt Ghasemi och Zahediasl (2012) [15] finns det ett flertal metoder för att bedöma om ett datamaterial kan antas vara normalfördelat. Ett första steg i analysen är vanligtvis att genomföra visuella analyser, exempelvis med hjälp av QQ-plottar, histogram eller boxplottar. Författarna betonar dock att sådana grafiska metoder bör kompletteras med statistiska tester, såsom Anderson–Darling-test och Kolmogorov–Smirnov-test, för att ge en mer tillförlitlig bedömning. Rekommendationerna följs, först genomförs en visuell analys av normalfördelningen. Denna omfattar analys av QQ-plott, den skattade fördelningsfunktionen samt ett histogram med en skattad täthetsfunktion. I tabell 14 presenteras de parameterskattningar som har tagits fram utifrån datamaterialet med hjälp av maximum likelihood-metoden.

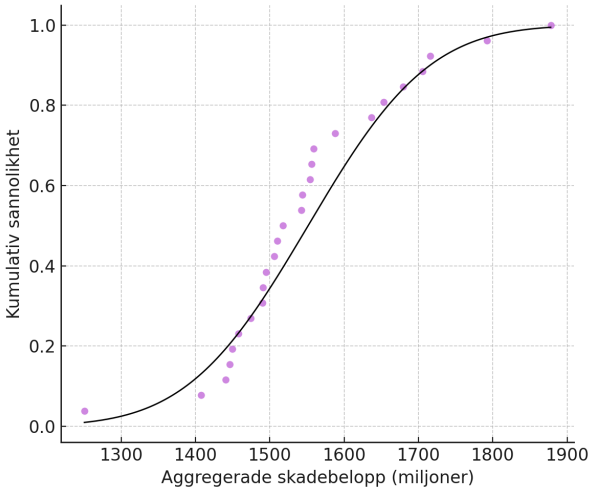
Tabell 14: Estimerat väntevärde och standardavvikelse för den anpassade normalfördelning.

Parametrar	**Värde**
μ	1 551 663 851
σ	130 577 282

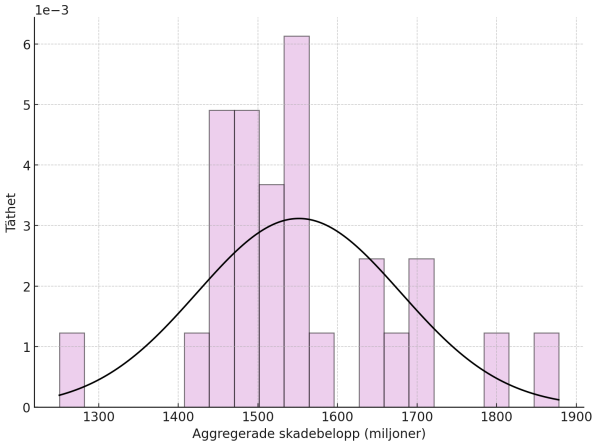
Efter att parametrarna har estimerats används dessa för att generera en QQ-plott, skattade fördelningsfunktioner samt histogram med en skattad täthetsfunktion, vilka presenteras i figur 11.



(a) QQ-plot



(b) Skattad fördelningsfunktion



(c) Histogram med skattad täthetsfunktion

Figur 11: Visuella analyser för normalfördelningen.

I figur 11 framgår att de empiriska datapunkterna överlag stämmer väl överens med den skattade normalfördelningen, särskilt i området där det finns mest data. Dock kan mindre avvikelser observeras i fördelningens svansar, vilket antyder att modellen inte helt fångar extremvärdena. Detta indikerar att normalfördelningen i viss mån kan vara en lämplig modell för de mindre, aggregerade skadebeloppen.

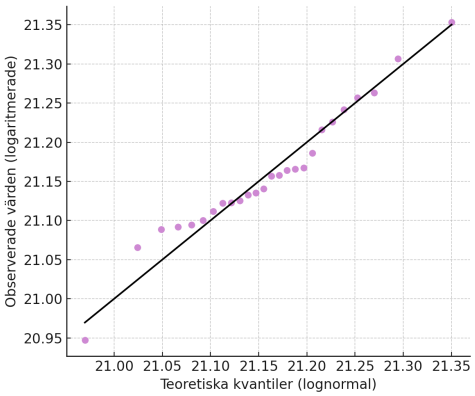
Det är dock viktigt att beakta att normalfördelningen tillåter negativa värden, vilket inte är förenligt med skadebelopp, då dessa per definition inte kan vara negativa. Med hänsyn till detta, samt till avvikelserna i svansarna, kan det vara motiverat att överväga alternativa fördelningar, såsom lognormal- eller gammafördelningen.

Att studera QQ-plottar, histogram och andra grafiska metoder är ett effektivt sätt att undersöka om data överensstämmer med en specifik fördelning. I denna uppsats an-

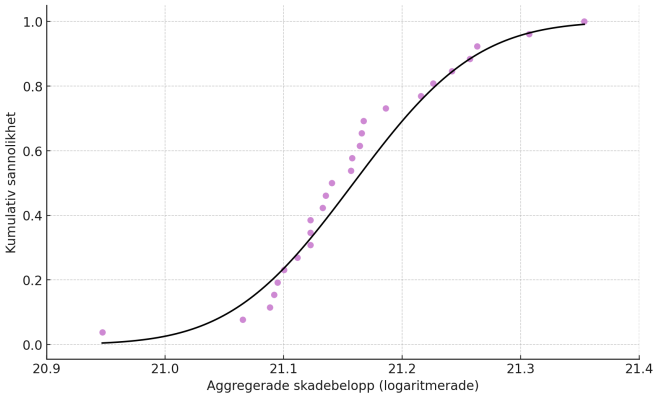
vänds därför sådana metoder även för att utvärdera hur väl en lognormalfördelning passar datamaterialet. Parametrarna för lognormalfördelningen bestäms genom att tillämpa maximum likelihood-metoden, enligt beskrivningen i avsnitt 3.5.1. Parametrarna μ och σ kan ses i tabell 15.

Tabell 15: Estimerat väntevärde och standardavvikelse för den anpassade lognormalfördelning.

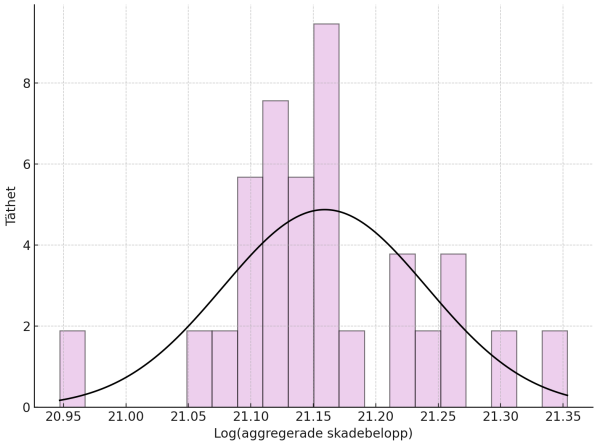
Parametrar	**Värde**
μ	21,1592
σ	0,0835



(a) QQ-plot



(b) Skattad fördelningsfunktion



(c) Histogram med skattad täthetsfunktion

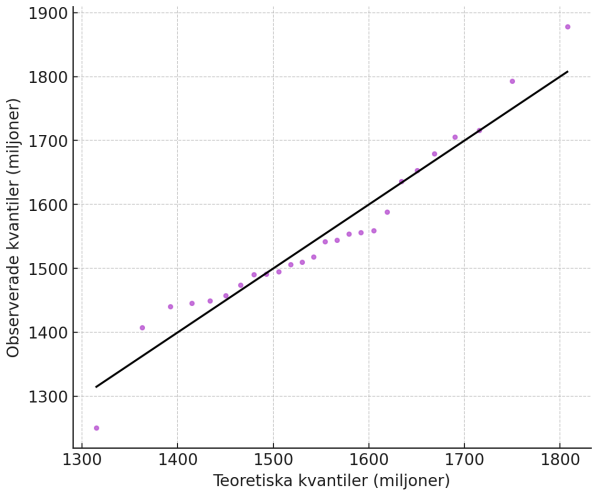
Figur 12: Visuella analyser för lognormalfördelningen.

Figur 12 visar att den skattade lognormalfördelningen stämmer bra överens med de observerade skadebeloppen. I QQ-plotten ligger punkterna nära diagonalen, vilket tyder på att fördelningens kvantiler passar bra mot datan. Den skattade fördelningsfunktionen följer den empiriska fördelningen väl. Histogrammet visar att täthetsfunktionen fångar formen på datan på ett rimligt sätt. Sammantaget tyder detta på att lognormalfördelningen är en lämplig modell för att beskriva de aggregerade skadebeloppen.

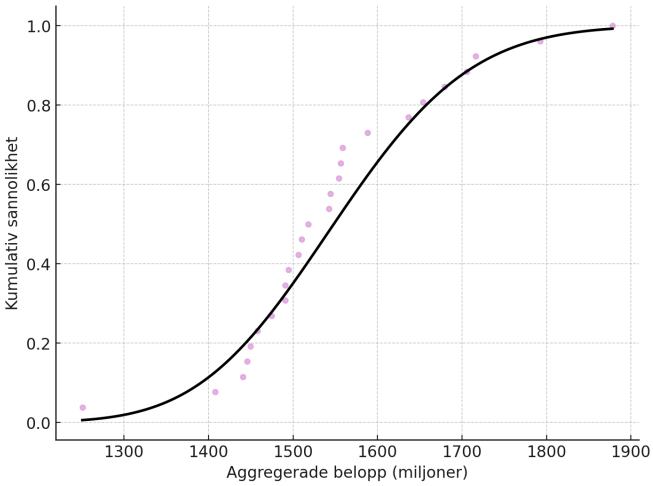
Samma analys genomförs för en gammafördelning, där maximum likelihood-metoden även här har använts för att skatta parametrarna, vilka redovisas i tabell 16.

Tabell 16: Estimerade skal- och formparametern för den anpassade gammafördelningen.

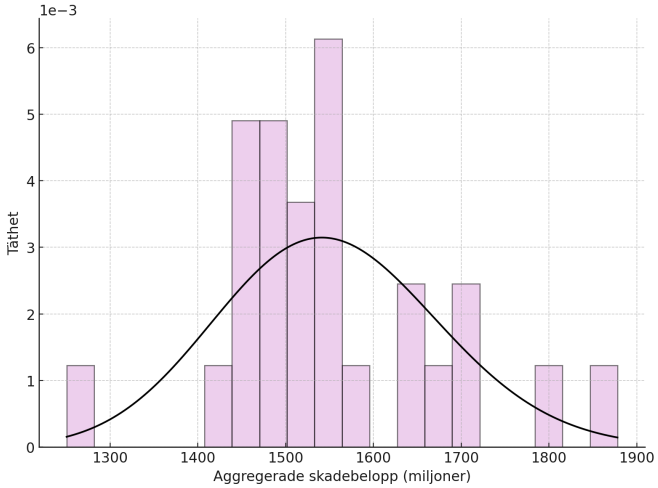
Parametrar	**Värde**
α	148,899
β	10 420 915,802



(a) QQ-plot



(b) Skattad fördelningsfunktion



(c) Histogram med skattad täthetsfunktion

Figur 13: Visuella analyser för gammafördelningen.

Utifrån figur 13 framgår det att de empiriska datapunkterna uppvisar god överensstämmelse med en gammafördelning. Varken QQ-plotten, den skattade fördelningsfunktionen eller histogrammet med skattad täthetsfunktion indikerar några systematiska avvikelser. Särskilt i QQ-plotten ligger punkterna nära diagonalen, vilket tyder på att modellens kvantiler stämmer överens med de observerade. Även den kumulativa fördelnings-

funktionen följer datapunkterna väl. Sammantaget tyder dessa resultat på att gammafördelningen är en rimlig modell för att beskriva de aggregerade skadebeloppen.

Modellutvärdering

För att ytterligare utvärdera vilken av de tre fördelningarna som bäst lämpar sig för att modellera de aggregerade skadebeloppen tillämpas både Anderson–Darling-och Kolmogorov–Smirnov-testet. Testresultaten för respektive fördelning redovisas i tabell 17.

Tabell 17: Teststatistika och p-värde för Anderson-Darling- och Kolmogorov-Smirnov-test för de tre klassiska fördelningarna.

Fördelning	AD-teststatistika	p-värde	KS-teststatistika	p-värde
Normalfördelning	0,6042	0,6424	0,1704	0,3929
Lognormalfördelning	**0,5121**	**0,7328**	**0,1541**	**0,5181**
Gammafördelning	0,5295	0,7153	0,1590	0,4789

Av resultaten i tabell 17 framgår att det samtliga tre fördelningar genererar höga p-värden, vilket innebär att nollhypotesen, att data följer respektive fördelning, inte kan förkastas för skadebelopp under 35 miljoner kronor. Detta indikerar att 35 miljoner kronor kan utgöra en lämplig brytpunkt för att särskilja de mindre skadebeloppen i modelleringen. Vidare visar tabell 17 att lognormalfördelningen genererar något högre p-värden i både Anderson–Darling- och Kolmogorov–Smirnov-testen. För att jämföra dessa fördelningar sinsemellan används jämförelsemåtten AIC och BIC, vars resultat redovisas i tabell 18.

Tabell 18: AIC- och BIC-värden för de tre klassiska fördelningarna.

Fördelning	AIC	BIC
Normalfördelning	1048,5138	1051,0300
Lognormalfördelning	**1047,9061**	**1050,4223**
Gammafördelning	1048,0381	1050,5543

Av tabell 18 framgår det att de olika fördelningarna uppvisar relativt liknande värden för både AIC och BIC. Den fördelning som däremot ger de lägsta värdena för både AIC och BIC är lognormalfördelningen. Eftersom lägre värden betyder att modellen passar datat bättre, tyder det på att lognormalfördelningen är den mest lämpliga modellen i det här fallet.

För analysen av de mindre skadebeloppen har rekommendationerna från Ghasemi och Zahediasl (2012) [15] följts, genom att inledningsvis genomföra en visuell analys av olika fördelningar för de mindre skadebeloppen. Dessa har därefter kompletterats med statistiska tester, såsom Anderson-Darling- och Kolmogorov-Smirnov-test.

Utöver dessa tester har även Akaikes informationskriterium (AIC) och Bayesianskt informationskriterium (BIC) använts för att jämföra hur väl modellerna passar data. Sammantaget tyder resultaten på att lognormalfördelningen ger den bästa anpassningen till de aggregerade skadebeloppen under 35 miljoner kronor, jämfört med övriga utvärderade modeller.

4.4 Modell för totala skadebelopp

Det är nu dags att sammanfoga de olika modellerna som utvecklats för de mindre respektive de större skadebeloppen och definiera slutmodellen S_k, som tidigare introducerats i ekvation 72.

Vi har tidigare konstaterat att skadebelopp som överstiger 35 miljoner kronor modelleras med en generaliserad Paretofördelning enligt

$$Z_i \sim \text{GPD}(\mu, \beta, \xi), \tag{73}$$

där parametrarna β och ξ har skattats med hjälp av maximum likelihood-metoden och redovisas i tabell 5. Lägesparametern μ sätts lika med den valda brytpunkten, det vill säga 35 miljoner.

Antalet skador som överskrider 35 miljoner kronor modelleras separat från skadebeloppen med en negativ binomialfördelning enligt

$$N_k \sim \text{NegBin}(r, p), \tag{74}$$

med skattade parametrar r och p angivna i tabell 11.

Det totala småskadebeloppet, som utgörs av en summa av enskilda skador understigande eller lika med brytpunkten 35 miljoner kronor, modelleras däremot med en lognormalfördelning enligt följande

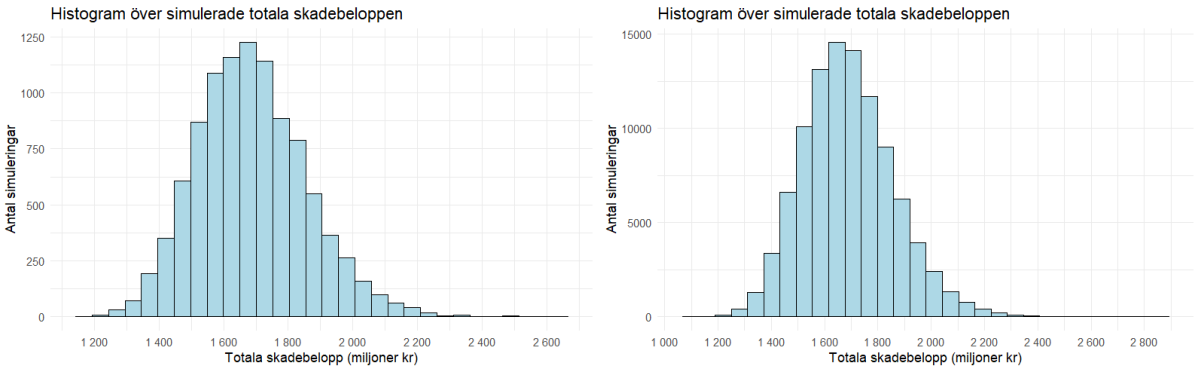
$$Y_k \sim \text{LogN}(\mu, \sigma), \tag{75}$$

och de tillhörande parametrarna framgår i tabell 15. Slutmodellen S_k, som definieras i ekvation 72, bygger på de tre ovan nämnda fördelningarna och används för att undersöka hur olika val av självbehåll påverkar försäkringsbolaget.

Från och med nu fokuserar analysen enbart på att modellera och undersöka de totala skadorna. För detta ändamål har 10 000 respektive 100 000 Monte Carlo-simuleringar genomförts, där varje simulering representerar ett möjligt utfall av den totala årliga skadekostnaden för försäkringsbolaget. Glasserman (2004) [16] betonar att antalet simuleringar har stor inverkan på noggrannheten i resultaten. Ju fler simuleringar som genomförs, desto närmare kommer det simulerade medelvärdet det sanna teoretiska väntevärdet, vilket motiveras av stora talens lag. Han framhåller även att ett tillräckligt stort antal simuleringar krävs för att minska slumpmässiga variationer. Mot denna bakgrund har både 10 000 och 100 000 simuleringar genomförts.

Varje simulerat utfall genereras genom att slumpa fram antalet skador för alla skadebelopp som överstiger brytpunkten u och därefter simuleras ett skadebelopp för varje skada per år. I detta fall slumpas antalet stora skador fram från en negativ binomialfördelning, för de stora skadebeloppen (det belopp som överstiger 35 miljoner kronor) från en generaliserad Paretofördelning och de mindre skadebeloppen (de belopp som är mindre eller lika med 35 miljoner kronor) från en lognormalfördelning. Till sist beräknas det sammanlagda skadebeloppet, det vill säga den totala simulerade skadekostnaden. Detta görs genom att summera ihop resultaten från simuleringen av de större respektive mindre skadebeloppen. Syftet med simuleringarna är att skapa en förståelse för både skadekostnaden under ett typiskt skadeår och variationen mellan skadeutfall, för ett försäkringsbolag vars skadedata följer dessa fördelningar.

Resultatet av de totala simulerade skadekostnaderna, det vill säga bruttoskadekostnaden, för respektive simulering kan ses i figur 14a och figur 14b. Det är viktigt för läsaren att ha i åtanke att inom försäkringsbranschen definieras bruttoskadekostnaden som den totala skadekostnaden före tillämpning av återförsäkringsskydd, medan nettoskadekostnaden avser den totala skadekostnaden efter att återförsäkringsskyddet har applicerats. Det är dock viktigt att notera att nettoskadekostnaden inte inkluderar premien som försäkringsbolaget betalar för återförsäkringen. Den avser enbart den del av skadekostnaden som försäkringsbolaget självt behöver stå för efter att återförsäkrarens andel har dragits av.



(a) 10 000 simuleringar (b) 100 000 simuleringar

Figur 14: Histogram över totala simulerade skadekostnaden (bruttoskadekostnaden) för respektive simulering.

Fortsättningsvis begränsas analysen till nettoskadekostnaden, angiven i miljoner kronor, det vill säga skadebeloppen efter avdrag för återförsäkringsersättning. I tabell 19 nedan visas hur nettoskadekostnaden, uttryckt i miljoner kronor, förändras vid olika val av självbehållsnivåer. Den första raden, ”Ingen ÅF” (det vill säga bruttoskadekostnaden), avser scenariot där försäkringsbolaget inte har tecknat återförsäkring, det vill säga där hela den simulerade totala skadekostnaden belastar försäkringsbolaget.

Tabell 19: Nyckeltal för totala nettoskadekostnaden i miljoner kronor vid olika nivåer av självbehåll för 10 000 simuleringar.

Självbehåll	Medel	ΔMedel	ΔMedel_0	Median	SD	Min	Max	95% per	99% per
Ingen ÅF	1 687	–	–	1 677	170	1 152	2 625	1 985	2 134
35	1 637	49,4	49,4	1 631	146	1 152	2 299	1 890	2 008
41	1 651	12,7	36,7	1 643	151	1 152	2 332	1 912	2 029
47	1 659	9,1	27,6	1 651	155	1 152	2 368	1 929	2 050
52	1 665	5,7	21,9	1 657	157	1 152	2 397	1 939	2 064
57	1 669	4,3	17,6	1 661	160	1 152	2 417	1 947	2 075
62	1 673	3,4	14,2	1 664	161	1 152	2 437	1 953	2 083
67	1 675	2,5	11,7	1 667	162	1 152	2 468	1 958	2 092
70	1 677	1,3	10,4	1 669	163	1 152	2 480	1 961	2 096

I tabell 19 visar bland annat hur medel-, median- och percentilvärden för det totala simulerade skadebeloppen påverkas av olika val av självbehåll, där exempelvis medelvärdet ökar från 1 637 till 1 651 miljoner kronor när självbehållet höjs från 35 till 41 miljoner kronor, vilket speglar att en större andel av skadekostnaderna då belastar försäkringsbolaget vid ökat självbehåll.

Observera att självbehållsnivåer endast kan analyseras för belopp som överstiger brytpunkten på 35 miljoner kronor, eftersom information om antalet skador i simuleringen enbart finns tillgänglig för skador över denna nivå. I simuleringarna kan det därmed identifieras hur många av de större skadorna som överstiger olika nivåer av självbehåll. För de

mindre skadebeloppen är detta däremot inte möjligt, eftersom dessa har aggregerats på årsbasis och inte redovisas individuellt. Exempelvis, om en enskild skada på 33 miljoner kronor hade inträffat och självbehållet var satt till 15 miljoner kronor, skulle försäkringsbolaget självt behöva stå för de första 15 miljoner kronorna medan återförsäkringsbolaget täcker resterande 18 miljoner kronorna, men antalet sådana skador i portföljen är okänt. Eftersom information om antalet mindre skador saknas, är det inte möjligt att avgöra hur många av dessa som skulle omfattas av återförsäkringen vid lägre självbehållsnivåer.

I tabell 19 sammanfattas centrala statistiska mått från de 10 000 simuleringarna för att få en förståelse för hur genomsnitt, spridning och storlek av skadebelopp för extrema skador förändras vid varierande självbehåll.

I tabell 19 framgår i kolumnen ”Medel” den genomsnittliga nettoskadekostnaden för olika nivåer av självbehåll. Detta motsvarar medelvärdet av de simulerade nettoskadekostnaderna i miljoner kronor efter återförsäkring och anger det genomsnittliga belopp försäkringsbolaget kan förvänta sig att behöva betala under ett år. I kolumnen ”ΔMedel” redovisas skillnaden i medelvärde jämfört med föregående nivå av självbehåll. Därifrån kan det utläsas hur den genomsnittliga skadekostnaden förändras vid variation av självbehåll. Kolumnen ”ΔMedel_0” visar hur den genomsnittliga nettoskadekostnaden förändras beroende på huruvida försäkringsbolaget väljer att återförsäkra sig eller inte, i relation till vald självbehållsnivå. Den första raden i tabell 19, som avser ett scenario utan återförsäkring, visar att försäkringsbolaget kan förvänta sig en årlig medelkostnad på 1 687 miljoner kronor i genomsnitt, vilket är 50 miljoner mer än vid ett självbehåll på 35 miljoner.

Tillsammans med detta presenteras medianen, som är ett mått på ett typiskt skadeutfall och särskilt intressant eftersom en stor skillnad mellan medelvärde och median tyder på att det finns ett fåtal väldigt stora skador som drar upp snittet, alltså en skev fördelning med tjock svans. Standardavvikelsen visar hur mycket skadeutfallen varierar mellan olika simuleringar. En hög standardavvikelse innebär att skadeutfallen varierar mycket mellan simuleringarna, vilket tyder på en hög osäkerhet i prognosen för framtida skador.

De lägsta och högsta värdena (kolumnerna ”Min” och ”Max”) avser den minsta respektive största simulerade nettoskadekostnaden bland de 10 000 simuleringar som genomförts för varje nivå på självbehållet. Dessa värden ger en indikation på det spann inom vilket skadekostnaderna kan variera, från gynnsamma till ogynnsamma skadeutfall. Kolumnen ”Självbehåll”, andra raden i tabell 19, visar att den lägsta nettoskadekostnaden bland 10 000 simulerade år är 1 152 miljoner kronor, medan den högsta uppgår till 2 299 miljoner, givet ett självbehåll på 35 miljoner kronor. Observera att den lägsta simulerade nettoskadekostnaden är densamma för samtliga nivåer på självbehållet. Detta beror på att det i varje simulering förekommer minst ett utfall utan storskador, vilket alltid resulterar i den lägsta totala kostnaden för storskador. För samtliga självbehåll utgörs den lägsta nettoskadekostnaden därmed enbart av summan av de simulerade mindre skadorna.

För att också förstå hur den simulerade nettoskadekostnaden kan bli under något extremt år, visas 95:e- och 99:e-percentilerna. Dessa anger gränsen för de väldigt ovanliga men fortfarande möjliga skadeutfallen, det vill säga hur stora simulerade nettoskadekostnader som kan behöva hanteras vid särskilt skadedrabbade skadeår som förväntas inträffa en gång på 20 respektive 100 år.

Utifrån tabell 19 framgår att den genomsnittliga nettoskadekostnaden ökar något med högre nivåer av självbehåll. Detta är ett förväntat resultat, eftersom försäkringsbolaget i sådana fall självt bär en större andel av kostnaden för varje enskild skada.

Medelvärdet ligger genomgående något över medianen. Det innebär att det i vissa simuleringar förekommer stora skador som drar upp snittet, även om skillnaden är relativt måttlig. Spridningen, det vill säga standardavvikelsen, ökar något, vilket betyder att det blir större variation i nettoskadekostnaden för högre självbehåll.

De lägsta skadekostnaderna (kolumnen ”Min”) är lika stora oavsett självbehåll, alltså förändras inte skadekostnaderna för de minst skadedrabbade åren. Däremot framgår att de högsta simulerade skadebeloppen (kolumn ”Max”), vilka representerar ovanliga men möjliga skadekostnader per skadeår, tenderar att bli något större när självbehållet höjs. Det betyder att även om den genomsnittliga nettoskadekostnaden inte påverkas särskilt mycket, kan de mest skadedrabbade åren bli dyrare.

I tabell 20 presenteras samma analys som i tabell 19, men baserat på 100 000 simuleringar istället för 10 000 simuleringar. Även i detta fall redovisas nyckeltal såsom medelvärde, median och standardavvikelse.

Tabell 20: Nyckeltal för totala nettoskadekostnaden i miljoner kronor vid olika nivåer av självbehåll för 100 000 simuleringar.

Självbehåll	Medel	ΔMedel	ΔMedel_0	Median	SD	Min	Max	95% per	99% per
Ingen ÅF	1 689	–	–	1 678	171	1 086	2 854	1 988	2 146
35	1 639	49,6	49,6	1 633	146	1 086	2 506	1 889	2 008
41	1 652	12,7	37,0	1 645	150	1 086	2 608	1 911	2 035
47	1 661	9,1	27,8	1 654	154	1 086	2 689	1 927	2 055
52	1 667	5,6	22,2	1 659	157	1 086	2 742	1 937	2 070
57	1 671	4,4	17,8	1 663	159	1 086	2 787	1 946	2 084
62	1 675	3,4	14,5	1 666	161	1 086	2 821	1 953	2 093
67	1 677	2,7	11,8	1 668	162	1 086	2 836	1 958	2 101
70	1 679	1,3	10,5	1 670	163	1 086	2 842	1 961	2 105

Av tabell 20 framgår det att nyckeltalen såsom medelvärde, median och standardavvikelse är relativt stabila för olika självbehållsnivåer med 100 000 simuleringar. En jämförelse mellan resultaten från 10 000 och 100 000 simuleringar visar att nyckeltalen är mycket stabila mellan de två simuleringarna. Medelvärde, median, standardavvikelse samt percentiler uppvisar endast marginella skillnader, vilket tyder på att redan 10 000 simuleringar ger relativt tillförlitliga skattningar.

Detta antyder att valet av självbehåll inom intervallet 35–70 miljoner för försäkringsbolaget har en relativt liten påverkan på den förväntade totala kostnaden. En rimlig förklaring till detta är att skadeutfallen under åren 1999–2024 har varit förhållandevis stabila. För ett annat försäkringsbolag med en mer volatil skadeportfölj hade troligen resultatet sett annorlunda ut, då variationen i skadebelopp kan ge större utslag beroende på val av självbehåll.

5 Diskussion

Det avslutande kapitlet ägnas åt en diskussion av resultaten samt presentation av möjliga inriktningar för framtida studier.

5.1 Diskussion och sammanfattning av resultat

Målet med denna uppsats var att analysera och utvärdera hur försäkringsbolag kan fatta välgrundade beslut vid vad av olika självbehåll i återförsäkringsavtal, samt att undersöka hur olika nivåer på självbehåll påverkar försäkringsbolagets skadekostnader. För att modellera skadeutfallen användes försäkringsbolagets verkliga skadedata mellan åren 1999-2024, där analyserna gjordes separat för större och mindre skadebelopp. För att identifiera en lämplig brytpunkt mellan de större och mindre skadebeloppen har hänsyn tagits till branschpraxis, där olika brytpunkter har testats och utvärderats med hjälp av grafiska och statistiska verktyg, såsom Anderson-Darling- och Kolmogorov-Smirnov-test.

För de större skadebeloppen modelleras antalet skador och de individuella skadebeloppen separat. För att identifiera en lämplig modell för antalet skador undersöks huruvida det föreligger överspridning i data, det vill säga om variansen överstiger väntevärdet. Som framgår av tabell 11 är variansen större än väntevärdet, vilket tyder på att överspridning förekommer i data. Därav modelleras antalet skador med hjälp av en negativ binomialfördelning. Detta är inte ett ovanligt resultat eftersom skadedata ofta varierar mer än vad en Poissonfördelning kan hantera. I vårt fall beror överspridningen på att det har inträffat olika många stora skador under olika år, vilket gör att variationen i antalet skador blir större än väntevärdet. I återförsäkringsbranschen är den klassiska metoden vid modellering av antal skador att använda sig av en Poissonfördelning så länge det inte finns ett starkt mönster eller beroende mellan skadorna. Resultatet från denna studie visar att den negativa binomialfördelningen är tillämplig, så är det fördelaktigt att använda den istället för Poissonfördelningen för att modellera antalet skador. Detta beror på att den negativa binomialfördelningen inte har begränsningen att väntevärde måste vara lika med variansen, vilket kan ge bättre anpassning till data. För att skatta parametrar finns det två klassiska metoder. Den ena är momentmetoden som beskrivs i avsnitt 3.5.2 och den andra är maximum likelihood-metoden som beskrivs i avsnitt 3.5.1. Momentmetoden bygger på att sätta teoretiska moment med empiriska, vilket ofta leder till enklare och snabbare beräkningar. För fördelningar som Poisson- och negativ binomial passar momentmetoden särskilt bra, eftersom den ger enkla beräkningar och tillräckligt noggranna resultat. Den har därför använts för att skatta parametrarna i dessa fall. För de andra fördelningarna har maximum likelihood-metoden använts, eftersom den oftast ger mer tillförlitliga skattningar, särskilt när datamängden är stor.

För modelleringen av Paretofördelningen har två olika metoder använts för att skatta formparametern x_m. Den första metoden innebär att x_m sätts lika med den valda brytpunkten u, medan den andra bygger på att skatta x_m med hjälp av maximum likelihood-metoden. Av figur 3 och figur 5 framgår det att skillnaden i kurvanpassning mellan de två metoderna är marginell. Däremot visar resultaten i tabell 6 och tabell 7 att både Anderson-Darling- och Kolmogorov-Smirnov-testerna ger något bättre resultat när x_m skattas med maximum likelihood-metoden. Skillnaderna är dock små, vilket kan förklaras av att maximum likelihood-skattningen av x_m ofta sammanfaller med det minsta observerade värdet, vilket i praktiken ligger mycket nära den valda brytpunkten. Därmed kan liknande slutsatser dras oavsett metod. Vid modellering av överskott är det därför både enkelt och rimligt att sätta $x_m = u$.

Gällande modellering av de större skadebeloppen visar Anderson-Darling- och Kolmogorov-Smirnov-testen tillsammans med AIC- och BIC-värden att den modell som är mest lämplig att välja är den generaliserade Paretofördelningen. Resultaten för dessa går att hitta i tabell 8 och tabell 9. Fokus i denna uppsats ligger på excess of loss-återförsäkring, vilket behandlas mer ingående i avsnitt 3.1.2. Där beskrivs försäkringskontrakt där återförsäkringsbolaget täcker den del av skadan som överstiger en viss brytpunkt. I sådana fall påverkar inte skadebeloppen under brytpunkten valet av självbehåll. Detta beror på att försäkringsbolaget inte får ersättning från återförsäkringsbolaget för de skador som understiger självbehållet.

Mot denna bakgrund tillämpas en modellering av den totala summan av mindre belopp, utan separat modellering av antalet skador och deras individuella storlek. Resultaten i figur 12, tabell 17 och tabell 18 visar att lognormalfördelningen är en lämplig modell att använda vid modellering av mindre skadebelopp. Detta resultat är rimligt, eftersom lognormalfördelningen har egenskaper som gör den väl lämpad för att modellera skadebelopp. För det första utesluter fördelningen negativa värden, vilket är en naturlig begränsning i en skadeportfölj där skador inte kan vara negativa. För det andra kännetecknas småskador ofta av en högerskev fördelning, där majoriteten av skadorna är små medan ett fåtal är mycket stora, ett mönster som lognormalfördelningen fångar väl.

Simuleringarna i tabell 19 och tabell 20 visar att den genomsnittliga nettoskadekostnaden för försäkringsbolaget inte förändrades nämnvärt beroende på de självbehåll som analyserats i denna studie. Detta kan framstå som något oväntat, men en möjlig förklaring är att skadeutfallen för försäkringsbolaget har varit relativt stabila under perioden 1999–2024. Resultaten visar att ett initialt införande av självbehåll (till exempel 35 MSEK) minskar spridningen i skadeutfallet, vilket syns i lägre percentilvärden och standardavvikelse. Däremot ökar spridningen något igen när självbehållet höjs ytterligare från denna nivå.

Sammanfattningsvis visar analysen i denna uppsats att valet av självbehåll bör grundas på en noggrann analys av hur skadorna har sett ut historiskt, särskilt när det gäller stora och extrema skador.

Även om den simulerade genomsnittliga totala nettoskadekostnaden för försäkringsbolaget inte verkar påverkas särskilt mycket av självbehållets nivå, så förändras riskbilden i de mest skadedrabbade åren, det vill säga hur höga kostnaderna kan bli i extrema utfall, representerat av de 95:e- och 99:e-percentilerna. När ett självbehåll införs minskar dessa extrema utfall initialt, men därefter ökar de åter något i takt med att självbehållet höjs ytterligare. Genom att använda tydliga och väletablerade metoder ger uppsatsen en fördjupad förståelse för hur självbehållet påverkar försäkringsriskerna, vilket kan stödja bättre beslut i praktiken.

5.2 Begränsningar och förslag till framtida studier

För framtida studier skulle denna analys kunna utvidgas genom att bland annat omfatta flera försäkringsbolag, vilket skulle göra det möjligt att jämföra olika skadeportföljer och analyser av hur skillnader i verksamhet påverkar valet av självbehåll och hur utsatt försäkringbolaget är för risk. Dessutom finns det en annan aspekt som skulle kunna behandlas, nämligen återförsäkringspremien, vilken inte har analyserats eller behandlats i denna uppsats. Genom att ta hänsyn till en återförsäkringspremie i analysen erhålls en mer heltäckande bild av den ekonomiska effekten av olika självbehållsnivåer. Det skulle också möjliggöra en mer nyanserad kostnadsoptimering, där valet av självbehåll inte enbart baseras på hur skadeutfallen varierar, utan även på hur mycket återförsäkringsskyddet faktiskt kostar.

Ett annat möjligt spår för framtida utveckling är att utvärdera mer flexibla självbehållsstrategier, där självbehållet kan anpassas efter exempelvis säsong, risknivå eller andra faktorer. Exempelvis kan risken för stormskador vara högre under hösten, medan risken för bränder kan öka under torra sommarmånader.

Det är även värt att understryka att valet av lägesparametern är av stor betydelse vid tillämpning av POT-modellen och vid skattning av de övriga parametrarna i en generaliserad Paretofördelning. Inom återförsäkring är det vanligt att lägesparametern väljs manuellt genom att pröva olika värden, vilket också är den metod som tillämpas i denna studie. Trots dess utbredda användning finns det inga garantier för att den manuella metoden leder till ett optimalt val. Det finns dock en teoretiskt välgrundad metod, Hill-estimator (Hill-skattaren), som används för att estimera lägesparametern. För den intresserade finns mer information gällande denna metod i [11] och delkapitel 7.2.4 i [22].

5.3 Slutsats

Analysen visar att de flesta skadorna är relativt små, men att det då och då inträffar stora skador som påverkar den totala nettoskadekostnaden. Det här mönstret är vanligt inom försäkringar, och tidigare forskning, till exempel av Embrechts med flera (1997) [12, ss. 352- 371], har visat att det passar bra att använda särskilda modeller såsom Pareto- och generaliserad Paretofördelning för att beskriva just sådana extrema skador.

En viktig slutsats är att det är avgörande att identifiera en lämplig brytpunkt för vad som utgör en stor skada. Om brytpunkten sätts för lågt riskerar modellen att inkludera observationer som inte uppfyller GPD-antagandena vilket leder till ökad bias. Sätts den däremot för högt blir parameteruppskattningarna instabila på grund av få datapunkter. Detta innebär att valet av brytpunkt är en kritisk avvägning mellan bias och varians, och valet bör därför styras både av statistiska tester och grafiska analyser. I detta arbete har en brytpunkt på 35 miljoner kronor identifierats som lämplig för att definiera en stor skada för det analyserade försäkringsbolaget, baserat på både grafiska metoder och statistiska tester.

För de mindre skadorna har det fungerat väl att summera skadebeloppen per år istället för att modellera varje enskild skada. Detta antagande bedöms som särskilt rimligt i avsaknad av tydliga tidsmässiga trender eller säsongsmönster i data. En årsvis aggregering av skadorna förenklar modellen, utan att detta medför någon betydande försämring i precision. Denna förenkling möjliggör dessutom användning av väletablerade sannolikhetsfördelningar, såsom lognormal- eller gammafördelningen, för att modellera de totala årliga skadekostnaderna på ett statistiskt tillförlitligt sätt.

Simuleringarna visar att den genomsnittliga nettoskadekostnaden per år inte förändras särskilt mycket beroende på vilken självbehållsnivå som väljs. Däremot påverkas spridningen i resultaten, framför allt när ett självbehåll införs, då variationen minskar tydligt. När självbehållet därefter höjs ytterligare ökar dock variationen och de extrema skadeutfallen något igen. Detta innebär att skadeutfallen först blir något mer förutsägbara, men att detta inte gäller vid mycket höga självbehåll. Därför bör beslut om återförsäkring inte bara baseras på den genomsnittliga skadekostnaden, utan också väga in hur stor variation och risk försäkringsbolaget är villigt att acceptera.

Referenser

[1] Agresti, A. (2013) *Categorical data analysis* (3:e uppl.), Wiley.

[2] Aladejana, J.A., Adesida, A.M. och Akinyemi, L.A. (2023) ”Evaluating the three methods of goodness of fit test for frequency analysis”, *Journal of Research in Applied and Computational Research*, ss.178–187, https://www.jracr.com/index.php/jracr/article/view/151.

[3] Albrecher, H., Beirlant, J. och Teugels, J.L. (2017) *Reinsurance: actuarial and statistical aspects* (1:a uppl.), Wiley.

[4] Alm, S.E. och Britton, T. (2008) *Stokastik: sannolikhetsteori och statistik med tillämpningar*, Liber.

[5] Antal, P. (2003) Quantitative metods in reinsurance. PDF) https://www.actuaries.ch/fr/downloads/aid!b4ae4834-66cd-464b-bd27-1497194efc96/id!96/Reinsurance_Script2009_v2.pdf.

[6] Arshad, M., Rasool, M. och Ahmad, M. (2003) ”Anderson Darling and Modified Anderson-Darling-tests for Generalized Pareto Distribution.”, *Pakistan Journal of Applied Sciences*, volym 3(2), ss. 85-88, https://doi.org/10.3923/jas.2003.85.88.

[7] Bühlmann, H. (1970) *Mathematical Methods in Risk Theory*, Springer.

[8] Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer.

[9] Cummins, J.D., Dionne, G., Gagné, R. et al. (2021) ”The costs and benefits of reinsurance”, *Geneva Pap Risk Insur Issues Pract*, volym 46, ss. 177–199, https://doi.org/10.1057/s41288-021-00216-8.

[10] Danielsson, J. (2011) *Financial risk forecasting: the theory and practice of forecasting market risk, with implementation in R and Matlab* (1:a uppl.), Wiley.

[11] Danielsson, J., Ergun, L.M., de Haan, L. & de Vries, C. (2025) ”Tail index estimation: quantile driven threshold selection”, http://dx.doi.org/10.2139/ssrn.2717478

[12] Embrechts, P., Klüppelberg, C. och Mikosch, T. (1997) *Modelling extremal events: for insurance and finance*, Springer.

[13] Foss, S., Korshunov, D. & Zachary, S. (2013) *An introduction to heavy-tailed and subexponential distributions* (2 uppl), Springer, https://link.springer.com/book/10.1007/978-1-4614-7101-1.

[14] Gary, P.S. (2001) ”Reinsurance (kapitel 7). I: Foundations of casualty actuarial science”, *Casualty Actuarial Society*, volym 4, ss. 343-484, https://www.casact.org/sites/default/files/2021-03/7_Patrik.pdf.

[15] Ghasemi, A. och Zahediasl, S. (2012) ”Normality tests for statistical analysis: A guide for non-statisticians”, *International Journal of Endocrinology and Metabolism*, volym 10(2), ss. 486–489, https://www.researchgate.net/publication/248398138_Normality_Tests_for_Statistical_Analysis_A_Guide_for_Non-Statisticians

[16] Glasserman, P. (2004) *Monte Carlo methods in financial engineering*, Springer, ss. 1-36.

[17] Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (2 uppl.), Springer.

[18] He, Y., Peng, L., Zhang, D. och Zhao, Z. (2022) ”Risk Analysis via Generalized Pareto Distributions”, *Journal of Business & Economic Statistics*, volym 40(2), ss. 852–867, https://doi.org/10.1080/07350015.2021.1874390.

[19] Johansson, B. (2014) *Matematiska modeller inom sakförsäkring.* Kompendium. In: Matematisk statistik, Stockholms universitet.

[20] Marsaglia, G., Tsang, W. W., & Wang, J. (2003) ”Evaluating Kolmogorov's Distribution”, *Journal of Statistical Software*, volym 8(18), 1–4, https://doi.org/10.18637/jss.v008.i18.

[21] Marsaglia, G., & Marsaglia, J. (2004) ”Evaluating the Anderson-Darling Distribution”, *Journal of Statistical Software*, volym 9(2), 1–5, https://doi.org/10.18637/jss.v009.i02

[22] McNeil, A.J., Frey, R. och Embrechts, P. (2005) ”Quantitative risk management: concepts, techniques and tools”, *Princeton University Press*, ss. 283–351, https://www.researchgate.net/publication/235622467_Quantitative_Risk_Management_Concepts_Techniques_and_Tools.

[23] *Negative binomial distribution.* ss. 1-3, https://www.math.ntu.edu.tw/~hchen/teaching/StatInference/notes/lecture16.pdf.

[24] Müller, S., Scealy, J.L. och Welsh, A.H. (2013) ”Model selection in linear mixed models”, *Statistical Science*, volym 28(2), ss. 135–167, https://doi.org/10.48550/arXiv.1306.2427.

[25] Myung, I.J. (2003) ”Tutorial om maximum likelihood estimation”, *Journal of Mathematical Psychology*, volym 47(1), ss. 90–100, https://doi.org/10.1016/S0022-2496(02)00028-7.

[26] Omari, C.O., Nyambura, S.G. och Mwangi, J.M.W. (2018) ”Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions”, *Journal of Mathematical Finance*, volym 8, ss. 137–160, https://doi.org/10.4236/jmf.2018.81012.

[27] Pickands, J. (1975) ”Statistical inference using extreme order statistics”, *The Annals of Statistics*, volym 3(1), ss.119–131, http://dx.doi.org/10.1214/aos/1176343003.

[28] Rosso, G. (2015) ”Extreme Value Theory for Time Series using Peak-Over-Threshold method”, arXiv:1509.01051, https://doi.org/10.48550/arXiv.1509.01051

[29] Rootzén, H. och Tajvidi, N. (1997) ”Extreme value statistics and wind storm losses: A case study”, *Scandinavian Actuarial Journal*, volym 1997(1), ss. 70–94, https://doi.org/10.1080/03461238.1997.10413979

[30] Savani, V. och Zhigljavsky, A.A. (2006) ”Efficient estimation of parameters of the negative binomial distribution”, *Communications in Statistics—Theory and Methods*, volym 35(5), ss. 767–783, https://doi.org/10.1080/03610920500501346.

[31] Simard, M. och L'Ecuyer, P. (2011) ”Computing the Two-Sided Kolmogorov-Smirnov Distribution”, *Journal of Statistical Software*, volym 39(11), ss. 1-18, https://doi.org/10.18637/jss.v039.i11.

[32] Stephens, M.A. (1974) ”EDF Statistics for Goodness of Fit and Some Comparisons”, *Journal of the American Statistical Association*, volym 69(347), ss.730–737, https://www.jstor.org/stable/2286009.

[33] Statistiska centralbyrån (SCB). *Statistikdatabasen*, Hämtad 2025-02-12 från https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/.

[34] Tsun, A. (2020). *Probability & statistics with applications to computing*, http://www.alextsun.com/files/Prob_Stat_for_CS_Book.pdf.

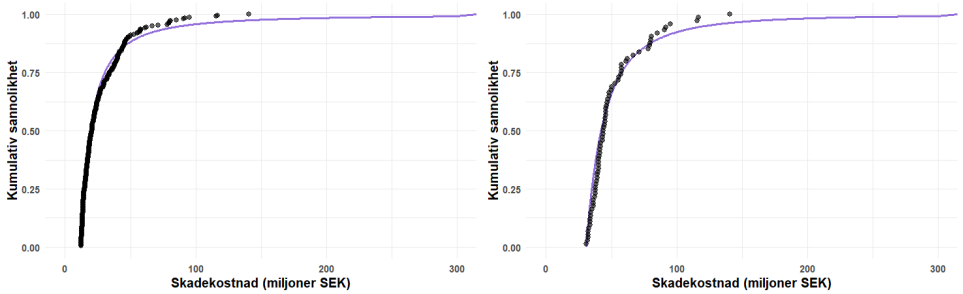
[35] Wang, S. och Gui, W. (2020) ”Corrected Maximum Likelihood Estimations of the Lognormal Distribution Parameters”, *Symmetry*, volym 12(6), ss. 1-20, https://doi.org/10.3390/sym12060968.

[36] Van Zyl, J.M. (2015) ”Estimation of the Shape Parameter of a Generalized Pareto Distribution Based on a Transformation to Pareto Distributed Variables”, *J Stat Theory Pract*, volym 9, ss. 171–183, https://arxiv.org/pdf/1210.7642.pdf.

Appendix

Tabell 21: Parametrarna för Paretofördelningen då $x_m = u$.

Skador	Parameter	Värde
Över 12 MSEK	α	1.5488
	x_m	12 000 000
Över 30 MSEK	α	2.1916
	x_m	30 000 000
Över 52 MSEK	α	2.9186
	x_m	52 000 000

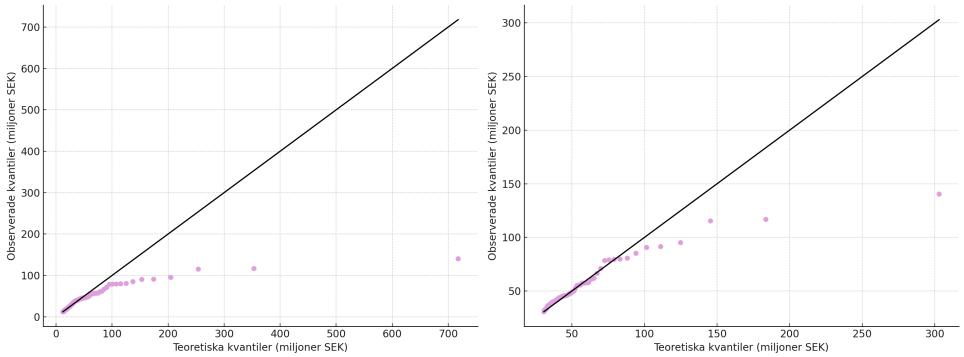


(a) Brytpunkt 12 MSEK

(b) Brytpunkt 30 MSEK

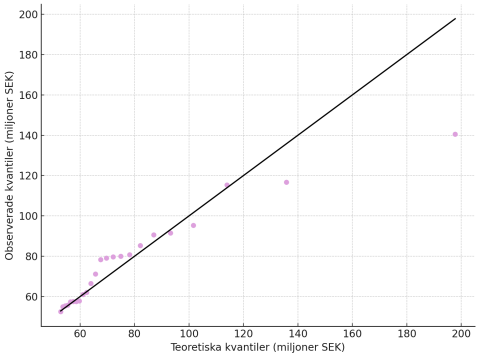
(c) Brytpunkt 52 MSEK

Figur 15: Fördelningsfunktioner för olika brytpunkter för Paretofördelningen då $x_m = u$.



(a) Brytpunkt 12 MSEK

(b) Brytpunkt 30 MSEK

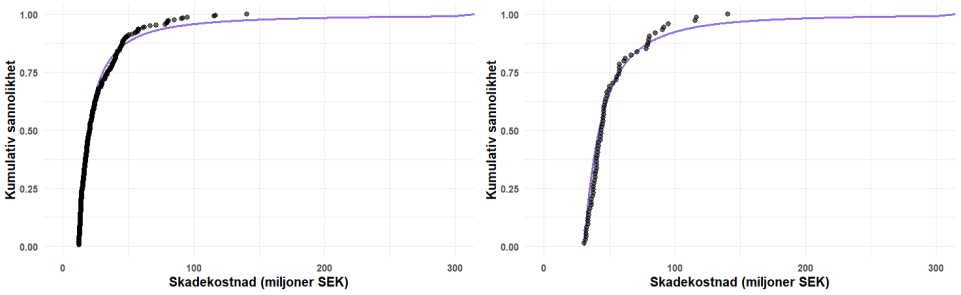


(c) Brytpunkt 52 MSEK

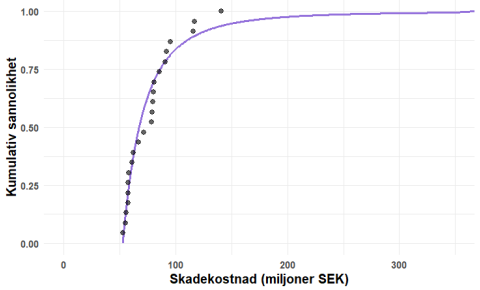
Figur 16: QQ-plot för olika brytpunkter för Paretofördelningen då $x_m = u$.

Tabell 22: Parametrarna för Paretofördelningen då x_m skattas.

Skador	**Parameter**	**Värde**
Över 12 MSEK	α	1.5488
	x_m	12 067 354
Över 30 MSEK	α	2.1916
	x_m	30 588 350
Över 52 MSEK	α	2.9186
	x_m	52 495 723

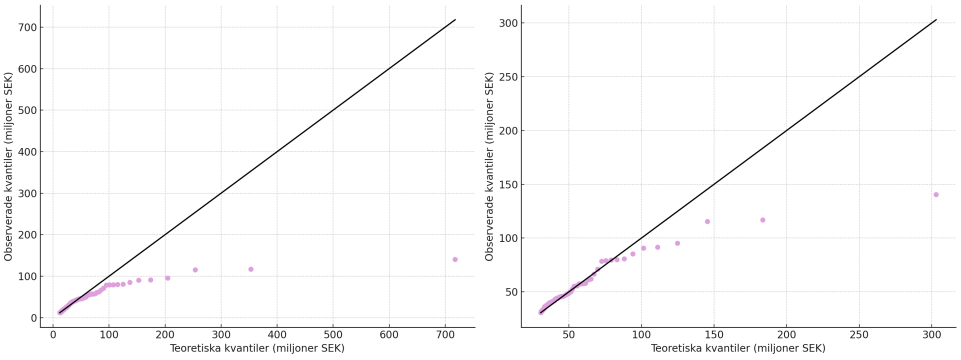


(a) Brytpunkt 12 MSEK (b) Brytpunkt 30 MSEK

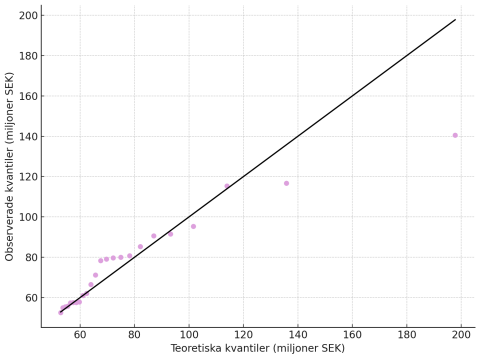


(c) Brytpunkt 52 MSEK

Figur 17: Fördelningsfunktioner för olika brytpunkter för Paretofördelningen då formparametern x_m skattas.



(a) Brytpunkt 12 MSEK (b) Brytpunkt 52 MSEK

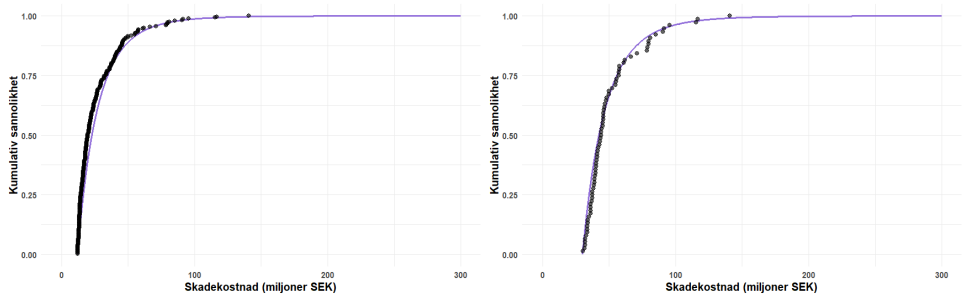


(c) Brytpunkt 52 MSEK

Figur 18: QQ-plot för olika brytpunkter för Paretofördelningen då formparametern x_m skattas.

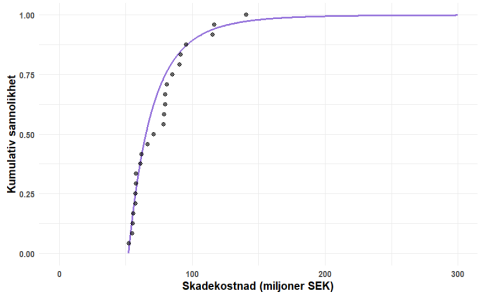
Tabell 23: Parametrar för generaliserad Paretofördelningen.

Parameter	Värde
μ	12 MSEK
ξ	0.1286
β	14 974 572
μ	30 MSEK
ξ	0.1300
β	17 010 830
μ	52 MSEK
ξ	0.1649
β	17 855 315



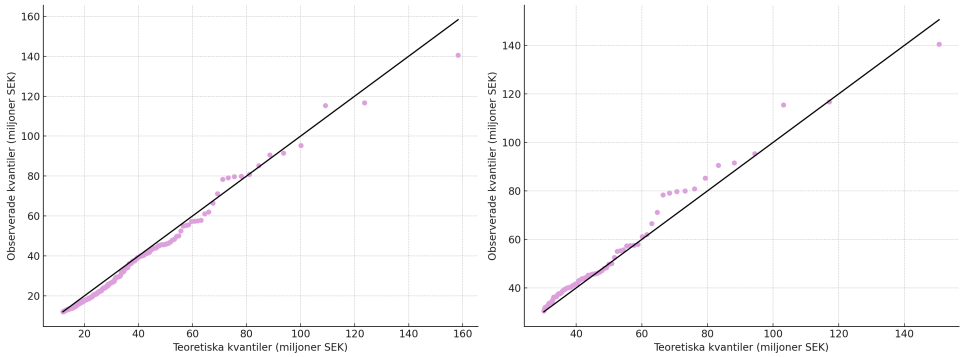
(a) Brytpunkt 12 MSEK

(b) Brytpunkt 30 MSEK

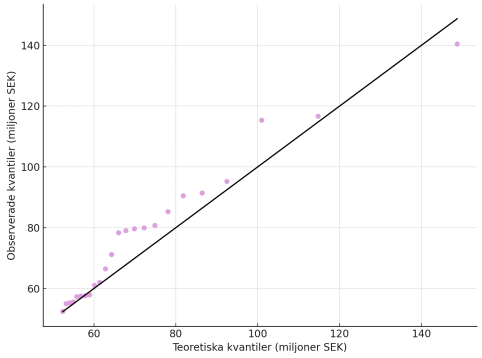


(c) Brytpunkt 52 MSEK

Figur 19: Fördelningsfunktionen för olika brytpunkter för generaliserad Paretofördelningen.



(a) Brytpunkt 12 MSEK (b) Brytpunkt 30 MSEK



(c) Brytpunkt 52 MSEK

Figur 20: QQ-plot för olika brytpunkter för generaliserad Paretofördelningen.