CHALMERS

GÖTEBORGS UNIVERSITET

Bortom normalfördelningen

Tunga svansar hos fördelningen av totala antalet individer i Galton-Watson-processen

Beyond the Normal Distribution

Heavy tails of the distribution of total progeny from the Galton-Watson process

Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Esmée Berger
Christoffer Ekgrim
Julia Jansson
David Larsson

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2020

Bortom normalfördelningen

Tunga svansar hos fördelningen av totala antalet individer i Galton-Watson-processen

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk fysik vid Chalmers

Esmée Berger David Larsson

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers

Christoffer Ekgrim Julia Jansson

Handledare: Serik Sagitov

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2020

Förord

Det här arbetet är ett kandidatarbete på institutionen för Matematiska vetenskaper vid Chalmers tekniska högskola under våren 2020. Gruppen består av två studenter på Teknisk Matematik och två på Teknisk Fysik. En loggbok har förts över var medlems insatser under arbetets gång. Alla har varit delaktiga i gruppen och närvarat vid alla möten, och vi har haft en öppen kommunikation. Vi vill tacka vår handledare Serik Sagitov, vars handledning och insikter har varit ovärdeliga. Dessutom vill vi tacka alla som har läst och gett oss återkoppling på vårt arbete, samt varandra för gott samarbete.

Avsnitt	Skrivet av	Övrig information
Pop. pres.	David & Julia	
Notationslista	Esmée & Julia	
Abstract	Christoffer	Inkluderar även sammanfattningen
1 (innan 1.1)	Esmée	Figur 1 & 2: Esmée, Figur 3: Christoffer
1.1 - 1.4	Alla	
2.1	Esmée	Förutom def. 2.3: Christoffer, def. 2.4: David
2.2 (innan 2.2.1)	Julia	
2.2.1	Julia	
2.2.2	Julia	Figur 4: Julia
2.2.3	Christoffer	
2.2.4	Christoffer & Esmée & Julia	
2.3 (innan 2.3.1)	Christoffer & Julia	
2.3.1	Julia	
2.3.2	Julia	
2.3.3	Esmée & Julia	
2.4 (innan 2.4.1)	Esmée	
2.4.1	Esmée & Julia	
2.4.2	Julia	
3.1	Esmée	Figur 5: Julia
3.2	Christoffer	
3.3	Christoffer & Esmée & Julia	
4.1	Esmée & Julia	
4.2.1	Christoffer	Figur 6: Christoffer
4.2.2	Christoffer & Esmée	Figur 7: Christoffer, Figur 8: Esmée
5.1	David	Figur 9: David
5.2	Julia	Figur 10: Julia
6	Esmée & Julia	
A	Esmée	Förutom def. A.13, A.16, A.17: Julia, A.15: David
B.1	Esmée	
B.2	Julia	Figur 11: Julia
B.3	Christoffer & Julia	
B.4	Christoffer & Julia	
B.5	Christoffer & Julia	Figur 12: Julia
B.6	Esmée & Julia	
B.7	Julia	
C.1	Christoffer	
C.2	Julia	
C.3	Julia	
D.1	Esmée	Figur 13 & 14: Esmée
D.2	Julia	Figur 15 & 16: Julia
E	Christoffer & Esmée & Julia	Se beskrivning om simuleringar nedan
F	Christoffer & Julia	

Tabell 1: Huvudförfattare av avsnitt i rapporten.

Gällande rapporten har samtliga gruppmedlemmar korrekturläst varandras delar och bidragit med språkliga ändringar, samt jobbat med strukturen på rapporten. Julia och Esmée har dock tagit extra ansvar för rapporten och gått på alla Fackspråkföreläsningar. Vidare har alla varit delaktiga i opponering på andra grupper och har läst de andra gruppernas rapporter, men sammanställning av respons har gjorts av Julia och Esmée. Uppdelningen efter rapportavsnitt finns i tabell 1. Om flera namn står under "skrivet av" är de listade utan hänsyn till inbördes arbetsmängd, det vill säga att dessa personer har bidragit ungefär lika mycket till avsnitten om det inte står annorlunda.

Vår handledare Serik har hjälpt oss med teorin och hållit inledande föreläsningar där Julia, Christoffer och Esmée har tagit ansvar för det matematiska arbetet med teorin och med att skriva in det i rapporten. Litteratursökningen gjordes främst av Julia men även av Esmée. Julia har jobbat med att lägga in källhänvisningar i olika delar av rapporten. I teoriavsnitten har de som skrivit styckena även gjort matematiken bakom, om det inte står annorlunda.

Christoffer har haft huvudansvar för simuleringar men även Julia och Esmée har bidragit med utveckling av konturprocessen samt analys med grafiska metoder. Körningarna är främst gjorda av Christoffer, och Esmée har använt datan för att göra undersökningen i subkritiska fallet. Vad gäller koden så är den samlad i avsnitt E. Gällande simuleringar av träd från Galton-Watson-processen började vi med en förgreningsprocess, vars kod skrevs av Christoffer i `R`. Christoffer gjorde sedan om detta till flertrådig kod i `Java` så att simuleringarna skulle gå mycket snabbare än förut. Sedan gick vi över från förgreningsprocessen till konturprocessen som Julia och Christoffer jobbade med i både `R` och `Java`, men det fungerade inte riktigt. Esmée skrev sedan en helt ny, fungerande implementation av konturprocessen i `MATLAB`. Julia översatte Esmées kod till `Java`. Förklaring av konturprocessen i avsnitt D.2 är skriven av Esmée. Javakoden är främst gjord av Christoffer, förutom konturprocessen som Esmée gjort och Julia översatt till Javakod. Kod till simulering av fördelningen av Z_n/n givet $Z_n \geq 1$ är skriven av Julia. Kod till importerande av data i `R` från simuleringar gjorda i Java är skriven av Christoffer.

David har haft huvudansvar för paretofördelningen och att hitta exempel, som antal barn födda i Sverige. Wikipediasidorna har skrivits av Christoffer och Julia. Med planeringsrapporten skrevs ett första utkast av David, tidsplanen gjordes av Esmée och resten av arbetet med planeringsrapporten gjordes tillsammans.

Avslutningsvis finns saker som har utforskats eller arbetats med men som inte platsade i den slutgiltiga rapporten. Rekonstruktion av grafen i en Jordan Peterson-video (`https://www.youtube.com/watch?v=NusYLzb-Uho`) gjordes av Esmée. Efter inläsning om bland annat econophysics kom sedan Esmée och Julia fram till att det sades fel i videon, och det blev därför oanvändbart för rapporten. Christoffer försökte sig på anpassning av data till potenslag med olinjär optimering, men Esmée kom fram till att det var bättre att testa om den var potensfördelad med andra, grafiska metoder. I samband med detta försökte vi nyttja andra grafiska metoder som Zengaplot (Julia) och Moment ratio plot (Christoffer), men Julia kom sedan fram till att dessa ej kunde användas på vår data eftersom vi har oändligt väntevärde och varians. Christoffer jämförde olika simuleringsmetoder för att se vilken som var mest nogrann, så att vi till slut valde konturprocessen. Christoffer, Julia och Esmée har arbetat en del med felsökning och tolkning av konturprocessen vilket slutligen ledde till Esmées korrekta implementation.

David skrev tidigare en text om paretofördelningens koppling till paretoprincipen, som nu är borttaget då det ej var relevant för arbetet eftersom paretoprincipen i sig kopplade för lite till tunga svansar. Tidigare anpassning av antal barn födda i Sverige till Poissonfördelning gjordes av David.

I början försöktes data kopplas till paretofördelningen till exempel genom en MLE-metod med Lomax-fördelningen av Christoffer, men detta fungerade inte. Vidare gjordes kopplingen till en potenslag med den tauberska satsen till en början med derivatan av $H(s)$ istället för överlevnadsfunktionen av W, där dessa räkningar och teori gjordes av Julia och Esmée. Detta uppfyllde dock inte alla förutsättningar i satsen, och därför föreslog Serik att en annan följd skulle användas. Med olika ansatser och asymptotisk analys på olika sätt lyckades Julia och Esmée till slut med överlevnadsfunktionen istället.

Populärvetenskaplig presentation:
Då normalfördelningen inte räcker till

I dagsläget ägs en majoritet av världens tillgångar av ett fåtal individer. Klyftorna mellan fattiga och rika är stora, och verkar inte följa någon intuitiv regel som är känd för allmänheten. När det istället gäller längden på människor vet vi säkert att man aldrig kommer träffa någon som avviker allt för mycket från genomsnittslängden. Samtidigt finns det individer som Bill Gates vars rikedom avviker avsevärt från normen. Finns det någon modell som kan användas för att förklara exempelvis varför ett fåtal människor kan bli så mycket rikare än resten?

Skillnaden mellan fördelningen av rikedom och längden på människor kan förklaras med att dessa kan beskrivas av olika sannolikhetsfördelningar, vilka anger hur troligt det är för något att inträffa. Det är helt enkelt mer sannolikt för en dollarmiljonär att existera, än en människa som är längre än den 2.16 meter långa basketspelaren Shaquille O'Neal. Fördelningen som beskriver könsfördelad längd på människor är normalfördelningen, där det praktiskt taget är omöjligt att röra sig mycket långt från medelvärdet. Den beskriver även många andra slumpmässiga företeelser i naturen och i samhället, exempelvis årlig nederbörd och födelsevikt på djur. Däremot räcker normalfördelningen inte till för att förklara vissa fenomen som klyftorna mellan fattiga och rika.

Fördelningen av rikedom uppkommer istället från det så kallade "rich-get-richer"-fenomenet, vilket innebär att många har lite och få har mycket. Detta fenomen beskrivs även i Matteusevangeliet där det sägs att "var och en som har, han skall få, och det i överflöd, men den som inte har, från honom skall tas också det han har". Som ett exempel har ett redan stort företag tillgångarna till att bedriva marknadsföring för att nå ännu fler konsumenter än vad de redan har, och kan därför växa sig mycket större och rikare än vad småföretag med begränsad räckvidd och tillgångar kan. Men rich-get-richer-fenomenet beskriver inte bara rikedom; det kan beskriva många olika fenomen. Till exempel tenderar de mest spelade låtarna på Spotify att bli ännu mer spelade i en slags kedjeeffekt, eftersom många lyssnare gärna lyssnar på redan populära låtar.

Sannolikhetsfördelningar som kan ge upphov till extremfall såsom Bill Gates rikedom sägs ha "tunga svansar". För dessa fördelningar är värden långt bortom genomsnittet förhållandevis sannolika jämfört med för normalfördelningen. Ett exempel på en sådan fördelning är paretofördelningen, uppkallad efter den italienske ekonomen Vilfredo Pareto, som under tidigt 1900-tal observerade att 80 % av allt land i Italien ägdes utav 20 % av befolkningen. Just rikedomsfördelning är ett typexempel på en paretofördelning, till exempel har Forbes lista över de 500 rikaste personerna i världen visats följa paretofördelningen.

En enkel matematisk modell som kan ge upphov till rich-get-richer-effekten och tungsvansade fördelningar är den så kallade Galton-Watson-processen. Under 1800-talet var Storbritanniens aristokrati bekymrad över att vissa aristokratefternamn skulle dö ut, och ville säkra deras överlevnad. Statistikerna Francis Galton och Henry William Watson undersökte hur sannolikt detta var genom att modellera hur efternamn förs vidare, i den så kallade Galton-Watson-processen. Med antagandet att efternamn förs vidare till nästa generation genom söner, eftersom döttrar bytte efternamn vid giftermål, kunde de beräkna huruvida efternamnet teoretiskt sett skulle leva kvar långt in i framtiden, eller slutligen skulle dö ut.

Denna förutsägelse påverkas av hur många söner en far förväntas få i genomsnitt. Om fadern i genomsnitt får fler än en son, kommer faderns efternamn sannolikt att överleva. Om fadern däremot inte får några söner, eller om han bara får döttrar, kommer efternamnet istället att dö ut. I gränsfallet får varje far exakt en son i genomsnitt, vilket leder till att efternamnet överlever, men med nöd och näppe. Dessa tre fall kallas för det superkritiska, subkritiska respektive kritiska för Galton-Watson-processen. I det superkritiska fallet kommer många familjeträd att växa sig stora och aldrig dö ut. Här får vi alltså de tidigare nämnda tunga svansarna, och en realisation av rich-get-richer-effekten, där redan stora familjeträd växer sig ännu större. Men ger Galton-Watson-processen upphov till tunga svansar även i det kritiska fallet? Att få i genomsnitt en son kan till exempel betyda att få noll eller två söner med lika sannolikhet. I denna situation kommer hälften av familjeträden att dö ut redan vid den första generationen, och den andra hälften kommer att fortsätta växa. Men hur länge kommer de flesta familjeträden att leva vidare? Kommer långlivade träd att fortsätta växa enligt rich-get-richer-fenomenet, eller kommer de slutligen dö ut?

I vårt arbete undersökte vi Galton-Watson-processens kritiska fall för att se om denna ger upphov till en tungsvansad fördelning av familjeträd, sett till hur stora dessa blir. Vi har betraktat

ett specialfall av Galton-Watson-processen, där antalet barn som varje individ får är sannolikhetsmässigt fördelat efter den så kallade LF-fördelningen (eng. *linear fractional distribution*). Denna fördelning kan intuitivt förklaras med två mynt. Säg att det första myntet vid kast ger krona med en viss sannolikhet, och att det andra myntet vid kast ger krona med en annan sannolikhet. Dessa sannolikheter är parametrar till fördelningen, och kan alltså varieras. Faderns chans att få sin första son avgörs av ett kast av det första myntet. Om myntet visar krona betyder det att fadern får minst en son, och då kastas även det andra myntet. Det andra myntet fortsätter att kastas ändå tills det visar klave. Fadern får då lika många söner till som antalet gånger som det andra myntet visade krona. Denna fördelning är ett användbart specialfall av Galton-Watson-processen eftersom den gör det möjligt att matematiskt beskriva processen långsiktiga beteende på en enkel form.

Den här förenklingen gjorde det möjligt för oss att exakt bestämma Galton-Watson-processens långlivade fördelning i det kritiska fallet. Vi fann att den fördelades efter en potensfunktion, likt Paretofördelningen. Detta innebär att Galton-Watson-processen har tunga svansar i det kritiska fallet! Efter att vi hade konstaterat detta blev det intressant att försöka verifiera resultaten med hjälp av datorsimuleringar. Simuleringarna bestod av att slumpa fram familjeträd genom att upprepade gånger kasta de två mynten som beskrivs tidigare. Det största familjeträdet vi fick bestod av 300 miljarder personer, vars storlek kan jämföras med tillgångarna hos världens rikaste människor. Sedan jämförde vi grafiskt våra resultat med potensfunktionen vi tidigare hade funnit, och vi fann att de stämde väl överens.

Vi ville även undersöka Galton-Watson-processen i det subkritiska fallet, det vill säga när det genomsnittliga antalet söner som varje far får är mindre än ett. Här kunde vi dock inte använda oss av samma metoder som i det kritiska fallet, utan fick förlita oss helt och hållet på datorsimuleringar. Vi såg att de fördelningar som vi erhöll snabbt avvek från potensfördelningen, alltså det som vi hade i det kritiska fallet. Vi kunde därför inte dra slutsatsen att processen ger upphov till tunga svansar även i det subkritiska fallet. I själva verket börjar normalfördelningen så småningom gälla när medelvärdet av antalet barn blir litet nog.

Idag används förgreningsprocesser såsom Galton-Watson-processen för att beskriva saker utöver familjeträd; den kan exempelvis kopplas till fördelningen av rikedom eller antal streams på Spotify. För streams på Spotify antar vi först för en given låt att en lyssnare finns. Denna personen har sedan möjligheten till att fortsätta lyssna på låten och sprida låten vidare till andra personer, eller sluta lyssna på den. I fallet då personen slutar att lyssna och inte delar låten, blir trädet av lyssnare utdött, då antar vi att inga fler hittar låten. Om personen däremot fortsätter att lyssna och delar låten, så har sedan de nya lyssnarna också möjlighet att antingen sluta lyssna eller sprida låten vidare, och detta beslut antas de göra med samma sannolikhetsfördelning. Dessa är förenklade antaganden, men denna modell kan användas för att koppla ihop Galton-Watson-processen med rich-get-richer-effekten.

Verkligheten påverkas dock inte enbart av rich-get-richer-effekten, utan är ett komplext samspel mellan många olika faktorer. Till exempel kan man spekulera om huruvida företag som Google förstärker rich-get-richer-effekten, eller tvärtemot motverkar den. Sökmotorers algoritmer baseras ofta på att visa de mest populära sidorna, men samtidigt så kan riktad reklam kunna få någon att upptäcka guldkorn som den inte annars hade hittat.

Bidrag till allmänheten

Som bidrag till allmänheten har vi skrivit Wikipedia-artiklar på svenska om vårt ämne. Artiklarna kan finnas på

- https://sv.wikipedia.org/wiki/Galton-Watson_processen
- https://sv.wikipedia.org/wiki/Sannolikhetsgenererande_funktion
- https://sv.wikipedia.org/wiki/LF-fördelning

Artiklarna finns även bifogade i appendix F.

Notationslista

$\sim$	Asymptotisk ekvivalens, se definition 2.13
$\overset{D}{\sim}$	En slumpvariabel *har fördelningen*, t.ex. $X \overset{D}{\sim} \mathcal{N}(\mu, \sigma^2)$ i definition 2.1
$\xrightarrow{D}$	Konvergens i fördelning, se definition 2.2
$\overset{D}{=}$	Likhet i fördelning mellan två slumpvariabler
$:=$	Tilldelning
α	Svansindex, se definition 2.4
f	Sannolikhetsfunktion för diskreta och täthetsfunktion för kontinuerliga slumpvariabler (def. A.9 och A.10)
F	Fördelningsfunktion för diskreta och kontinuerliga slumpvariabler, se definition A.8
G	Sannolikhetsgenererande funktion av X fördelad enligt reproduktionsfördelningen
G_n	Genererande funktion för Z_n, alltså $G_n(s) = \mathbb{E}[s^{Z_n}]$
Γ	Gammafunktionen, där $\Gamma(k) = (k-1)!$
H	Sannolikhetsgenererande funktion för W, alltså $H(s) = \mathbb{E}[s^W]$
L	Långsamt varierande funktion, se definiton 2.14
μ	Väntevärde, se definition A.11
n	Generation
N	Antal simuleringar
$\mathbb{N}_0$	De ickenegativa heltalen $0, 1, 2, \ldots$
$\mathcal{O}$	Ordonotation, storleksordning
p	Sannolikhet i de fall där endast en parameter behövs, som i definition 2.8 eller definition 2.10
p_0	Den andra parametern i LF-fördelningen, där p är den första, se definition 2.11
p_k	Sannolikheten $P(X = k)$ generellt
q	Utrotningssannolikhet för Galton-Watson-processen, se sats 2.3
$\mathbb{R}$	Reella talen
σ^2	Varians, se definition A.12
S	Överlevnadsfunktion för diskreta och kontinuerliga slumpvariabler, se definition 2.3
S_W	Överlevnadsfunktionen för totala antalet individer W
t_n	Taubersk talföljd, som beter sig som en potenslag asymptotiskt
W	Totala antalet individer i Galton-Watson-processen
X	Generell diskret slumpvariabel, se def. A.9, tolkas även som antal barn vid Galton-Watson-processen
Z_n	Antal individer i generation n i Galton-Watson-processen

Sammanfattning

Normalfördelningen är en essentiell sannolikhetsfördelning inom statistiken som kan beskriva många olika sorters fenomen inom naturen och i samhället. Den kan dock inte användas överallt. Exempelvis kan den så kallade *rich-get-richer*-principen, vilket innebär att "många har lite och få har mycket", ge upphov till fördelningar vars svansar inte avtar exponentiellt och som därför inte kan modelleras med hjälp av normalfördelningen. Ett exempel på en process som kan ge upphov till rich-get-richer-fenomenet är Galton-Watson-processen, vilket är en stokastisk förgreningsprocess som ursprungligen användes under 1800-talet för att modellera spridningen och utrotningen av efternamn.

Detta arbete ämnade till att undersöka Galton-Watson-processen, för att ta reda på huruvida den ger upphov till tungsvansade fördelningar i de subkritiska och kritiska fallen. Mer specifikt så betraktades fallet då processens reproduktionsfördelning ges av den så kallade LF-fördelningen (eng. *linear fractional distribution*). Detta gjordes genom en litteraturstudie samt genom simuleringar av processen i `Java` och analys av den erhållna datan i `R`.

Det konstaterades att om Galton-Watson-processens reproduktionsfördelning ges av LF-fördelningen så ges överlevnadsfunktionen S_W av det totala antalet individer W i det kritiska fallet av potenslagen

$$S_W(n) \sim \sqrt{\frac{2}{\pi\sigma^2}} n^{-1/2}, \quad n \to \infty,$$

där σ^2 är reproduktionsfördelningens varians. Approximationer av överlevnadsfunktionen från simuleringar följer formeln väl. Således är överlevnadsfunktionen av W tungsvansad i det kritiska fallet. Vidare så verkar log-log-plots och qq-plots av flera dataset erhållna från simuleringar av processen visa att överlevnadsfunktionen i det subkritiska fallet avtar i en snabbare takt än vad en potenslag beskriver, möjligtvis exponentiellt.

Nyckelord: normalfördelning, rich-get-richer, tung svans, Galton-Watson-process, LF-fördelning.

Abstract

The normal distribution is an essential probability distribution in statistics that can be used to describe many different types of phenomena in nature and society. However, it cannot be used everywhere. For example, the so called *rich get richer* principle, which states that "the rich get richer and the poor get poorer", can give rise to distributions whose tails are not exponentially bounded, and are thus distinctive from the normal distribution. An example of a process that can cause the rich get richer phenomenon is the Galton-Watson process, which is a stochastic branching process that was originally used in the 19th century to model the spread and extinction of surnames.

This paper intended to study the Galton-Watson process to discover whether it gives rise to heavy-tailed distributions in the subcritical and critical cases. More specifically, the case when the offspring distribution is the linear fractional distribution was considered. This was done through a literature review as well as through simulations of the process in `Java` and analysis of the obtained data in `R`.

Using the linear fractional distribution in the critical case, the survival function S_W of the total progeny W was found to be given by the power law

$$S_W(n) \sim \sqrt{\frac{2}{\pi\sigma^2}} n^{-1/2}, \quad n \to \infty,$$

where σ^2 is the variance of the offspring distribution. Approximations of the survival function from simulations of the process match this formula well. Thus, the survival function is heavy-tailed in the critical case. Furthermore, log-log plots and quantile-quantile plots of several datasets obtained from simulations of the process seem to show that the survival function in the subcritical case decreases at a faster rate than what a power law describes, possibly exponentially.

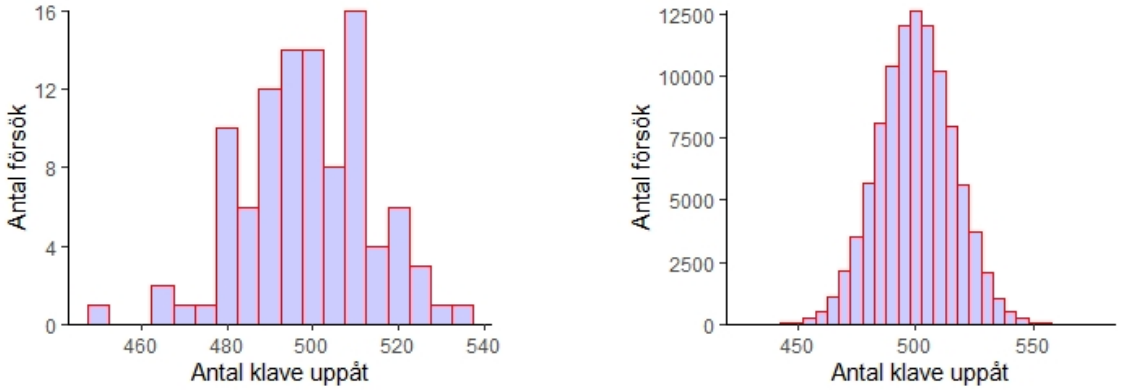
Keywords: normal distribution, rich get richer, heavy tail, Galton-Watson process, linear fractional distribution.

Innehåll

1 Inledning

Vid observationen av många naturliga och samhälleliga fenomen, exempelvis kroppslängd, djurs födelsevikt och årlig nederbörd, uppträder mätvärdena ofta enligt ett välkänt mönster, den så kallade normalfördelningen. Normalfördelningen är en mycket användbar sannolikhetsfördelning, men inte alltid applicerbar. Till exempel följer inte fördelningen av rikedom, akademiska citeringar eller mest streamade låtar normalfördelningen [1]. Det är därför intressant att betrakta fördelningar som kan användas för att beskriva fenomen bortom normalfördelningen, som paretofördelningen, samt modeller som kan ge upphov till sådana fördelningar, som Galton-Watson-processen [2].

Innan vi introducerar modeller bortom normalfördelningen illustrerar vi tanken bakom normalfördelningen genom att betrakta myntkast. Vid myntkast förväntar vi oss att klave i genomsnitt hamnar uppåt hälften av gångerna. Om vi exempelvis låter ett försök bestå av att räkna antalet gånger som klave hamnar uppåt när vi kastar ett mynt 1000 gånger, och om vi upprepar detta försök 100 gånger, skulle vi bli förvånade om resultatet var långt ifrån 500 vid något av försöken. Resultatet kan visualiseras i ett histogram, enligt figur 1a.

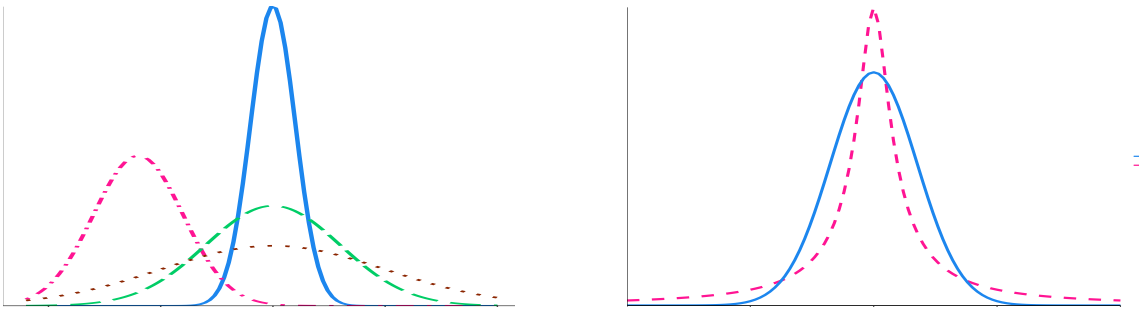


(a) Försöket upprepas 100 gånger.

(b) Försöket upprepas 100000 gånger.

Figur 1: Histogrammen illustrerar resultatet av upprepade myntkast för olika antal försök, där varje försök består av 1000 kast.

Ökar vi antal försök till 100 000 blir resultatet annorlunda, enligt figur 1b. Detta liknar en normalfördelning. Låter vi antalet försök gå mot oändligheten erhålls en faktisk normalfördelning, vars utseende för några olika parameterval åskådliggörs i figur 2a.



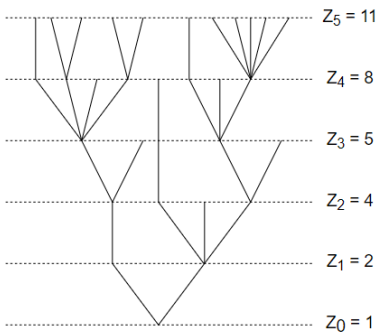
(a) Normalfördelningens utseende för olika väntevärden μ och varianser σ^2.

(b) Heldragen kurva är en normalfördelning och streckad kurva är en tungsvansad fördelning.

Figur 2: Normalfördelningens utseende för olika parametrar samt en normalfördelning jämfört med en tungsvansad fördelning. Observera att de normalfördelningar som har samma väntevärde i figur 2a korsar varandra två gånger medan den tungsvansade fördelningen i figur 2b korsar normalfördelningen fyra gånger trots samma väntevärde.

Exemplet med myntkasten illustrerar att en storhet som kan betraktas som en summa av många små, oberoende och slumpmässiga fenomen kommer att vara normalfördelad i gräns. Detta är en anledning till att normalfördelningen kan användas för att approximera så många naturliga och samhälleliga fenomen, vilket beskrivs formellt av centrala gränsvärdessatsen [3]. Bortom normalfördelningen går det att erhålla något som istället ser ut som den streckade kurvan i figur 2b. Denna fördelning har så kallade tunga svansar, vilket innebär att den har en förhållandevis stor area under grafen även långt bort från väntevärdet. Kurvan går alltså inte mot noll i samma takt som den gör för normalfördelningen, vilket innebär att förutsättningarna för centrala gränsvärdessatsen inte är uppfyllda och därmed att normalfördelningen i detta fall inte är en bra approximation. För att beskriva tunga svansars beteende behövs alltså andra fördelningar. Paretofördelningen är ett exempel på en fördelning som kan användas i fall bortom normalfördelningen, eftersom den fångar beteendet hos tunga svansar. Paretofördelningens svansar följer en potenslag och avtar således långsammare än normalfördelningens exponentiellt avtagande svansar.

Ett exempel på en modell som kan ge upphov till fördelningar med tunga svansar är, som tidigare nämnt, Galton-Watson-processen, vilket är en så kallad förgreningsprocess. Dess namn kommer från de brittiska matematikerna Francis Galton och Henry W. Watson som under 1800-talet modellerade sannolikheten för att efternamn skulle försvinna, eftersom den dåvarande aristokratin oroade sig för det [4]. Galton-Watson-processen går ut på att det finns en startindivid som ger upphov till barn efter en given sannolikhetsfördelning, och dessa barn ger i sin tur upphov till egna barn efter samma fördelning. Figur 3 illustrerar hur denna process skulle kunna utvecklas. Idag används denna typ av modell exempelvis inom populationsgenetik för att beskriva förändringar i populationsstorlek över generationsskiften.



Figur 3: En Galton-Watson-process, där Z_k är antalet individer i generation k. Den kan exempelvis illustrera hur efternamn sprids från far till son, utgående från en man som bär på efternamnet, då startgenerationen Z_0 består av en individ.

Galton-Watson-processen har tre fall: det subkritiska, där väntevärdet av antal barn per individ är mindre än 1, det superkritiska, där väntevärdet är större än 1 och det kritiska, där väntevärdet är exakt lika med 1. Långsiktigt leder det subkritiska och kritiska fallet till att processen slutligen utrotas medan det superkritiska fallet innebär en långlivad process som med en viss sannolikhet aldrig utrotas. I det superkritiska fallet ger Galton-Watson-processen upphov till tungsvansade fördelningar och därmed kan den användas för att modellera *rich-get-richer*-fenomenet, vilket innebär att "många har lite och få har mycket". Ett exempel på ett sådant fenomen är Paretoprincipen som har sitt ursprung i att ekonomen Vilfredo Pareto insåg att 80 % av all mark i Italien ägdes av 20 % av befolkningen och som matematiskt kan beskrivas genom ett särskilt parameterval i paretofördelningen [5].

Med hjälp av sannolikhetsgenererande funktioner kan egenskaperna hos fördelningarna som Galton-Watson-processen ger upphov till i de olika fallen studeras. Sannolikhetsgenererande funktioner är potensserierepresentationer av sannolikhetsfunktionerna för diskreta slumpvariabler och kan användas för att beräkna utrotningssannolikheten för förgreningsprocesser [6]. Ett specialfall av dessa funktioner är så kallade LF-fördelningar (eng. *linear fractional distributions*), vars genererande funktioner kan skrivas som kvoten mellan två linjära funktioner. LF-fördelningar är användbara för att få den genererande funktionen på en enklare form, särskilt efter upprepad applikation av funktionen.

1.1 Problem och syfte

Med utgångspunkt i problemet att många fenomen ej kan beskrivas med hjälp av normalfördelningen och centrala gränsvärdessatsen är syftet med detta arbete att undersöka fördelningar bortom normalfördelningen. I synnerhet betraktas huruvida Galton-Watson-processen ger upphov till fördelningar med tunga svansar i det kritiska och subkritiska fallet. Arbetet ämnar också till att genom denna undersökning ge den tänkte läsaren en inblick i metoder som kan användas bortom normalfördelningen.

1.2 Metod

Arbetet består av en litteraturstudie och av simuleringar. Teori för LF-fördelningen appliceras på det kritiska fallet i Galton-Watson-processen. Genom asymptotisk analys och teori om konvergens för potensserier kan det undersökas hur LF-fördelningen beter sig i asymptotiskt, och vad detta innebär för Galton-Watson-processens egenskaper i det kritiska fallet. Särskilt visas en asymptotisk potenslag för fördelningen av totala antalet individer för Galton-Watson-processen i avsnitt 2.4.1. Vidare simuleras det subkritiska och kritiska fallet för att undersöka kopplingen till tunga svansar. Simuleringarna görs i `Java` med hjälp av konturprocessen som relaterar ett träd från processen till en slumpvandring, och beskrivs i avsnitt 3.1. För att undersöka om den simulerade datan asymptotiskt fördelas enligt en potenslag analyseras den i `R` med grafiska metoder såsom log-log-plot, qq-plot och Hill-estimering, vilka förklaras i 3.3. Därutöver behandlas konkreta exempel på Galton-Watson-processer, till exempel antal barn födda i Sverige, antal streams på topplåtar och antal citeringar på Chalmers, i avsnitt 5.

1.3 Avgränsningar

För den asymptotiska potenslagen betraktas endast fallet då Galton-Watson-processens reproduktionsfördelning ges av LF-fördelningen i det kritiska fallet. I simuleringsstudien jämförs datan i det kritiska fallet med potenslagen, samt undersöks med hjälp av grafiska metoder. Det subkritiska fallet undersöks enbart med grafiska metoder, och undersökningen begränsas till jämförelser mellan potenslagar och exponentialfördelningen. Mellanliggande fördelningar som är lättare än en potenslag men tyngre än exponentialfördelningen behandlas alltså inte. Detta eftersom det är egenskapen av tunga svansar som främst är intressant, och inte vilken exakt fördelning som det subkritiska fallet ger upphov till. Det superkritiska fallet undersöks inte eftersom Galton-Watson-processen är svårsimulerad i detta fall då långlivade träd aldrig dör ut, och därmed är det även trivialt att den ger upphov till tungsvansade fördelningar. Därutöver undersöks endast enkönade Galton-Watson-processer, alltså reproduktion från far till son eller mor till dotter. Vidare är Galton-Watson-processerna som betraktas storleksoberoende, alltså beror reproduktionsfördelningen inte av generationen. Dessa två utvidgningar skulle ej bidra ytterligare till att besvara syftet än vad undersökningen av den enkla Galton-Watson-processen gör, eftersom de skulle ha komplicerat teorin anmärkningsvärt.

1.4 Samhälleliga och etiska aspekter

En samhällsetisk risk med detta arbete är att läsaren får en förenklad bild av verkligheten. Verkligheten påverkas inte enbart av rich-get-richer-effekten, utan är ett komplext samspel mellan många olika faktorer. Nyttan med arbetet är att det bidrar till att läsaren blir mer upplyst om modeller, fördelningar och metoder bortom normalfördelningen och får därmed en inblick i kopplingen mellan potenslagar och verkliga fenomen, så att det skapas förståelse för rich-get-richer-effekten. Dessa aspekter diskuteras vidare i avsnitt 6.

2 Teori

I detta avsnitt presenteras relevant teoretisk bakgrund såväl som teorin som erhållits genom litteraturstudien. Efter att sannolikhetsteoretisk bakgrund har presenterats utforskas långsiktiga egenskaper av Galton-Watson-processer i dess olika fall. Sedan introduceras LF-fördelningen (eng. *li-*

near fractional distribution) samt några av Galton-Watson-processens egenskaper när dess reproduktionsfördelning följer LF-fördelningen. Från dessa resultat samt asymptotisk teori härleds en potenslag för den kritiska Galton-Watson-processen i LF-fallet.

2.1 Bakgrund

I detta avsnitt redogörs för relevant sannolikhetsteoretisk bakgrund. Först ges en definition av normalfördelningen samt en formulering av centrala gränsvärdessatsen. Sedan introduceras överlevnadsfunktioner och paretofördelningen, som är ett exempel på en potenslag. Slutligen introduceras momentgenererande funktioner samt begreppet tungsvansad fördelning, som är centralt för arbetet. Ytterligare sannolikhetsteoretisk bakgrund finns i appendix A, där bland annat begreppen som detta avsnitt bygger på definieras. Vidare sammanfattas alla kontinuerliga och diskreta fördelningar som används i arbetet i tabell 2. Vi börjar med den formella definitionen av normalfördelningen och konvergens i fördelning, samt ger en formulering av centrala gränsvärdessatsen [3].

Definition 2.1 (Normalfördelning)**.** En slumpvariabel X är normalfördelad med parametrarna $\mu \in \mathbb{R}$ och $\sigma^2 > 0$ om dess täthetsfunktion ges av

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

Detta skrivs $X \overset{D}{\sim} \mathcal{N}(\mu, \sigma^2)$, där $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$ och D står för engelskans distribution.

En normalfördelning har cirka $99,7\,\%$ av värdena inom tre standardavvikelser från väntevärdet.

Definition 2.2 (Konvergens i fördelning)**.** Om X_1, X_2, $X_3, ...$ är en sekvens av slumpvariabler med respektive fördelningsfunktioner F_1, F_2, $F_3, ...$ och X är en slumpvariabel med fördelningsfunktionen F så konvergerar X_n mot X i fördelning om $\lim_{n\to\infty} F_n(x) = F(x)$ för alla x där $F(x)$ är kontinuerlig. Detta skrivs $F_n \to F$ eller $X_n \xrightarrow{D} X$.

Sats 2.1 (Centrala gränsvärdessatsen)**.** *Låt X_1, X_2, $X_3, ...$ vara en sekvens av oberoende och identiskt fördelade slumpvariabler med ändligt väntevärde μ och ändlig varians σ^2 och låt $\mathbb{S}_n = X_1 + X_2 + ... + X_n$. Då gäller att*

$$\frac{\mathbb{S}_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} \mathcal{N}(0,1) \quad \textit{då} \quad n \to \infty.$$

Beviset av satsen och tillhörande lemmor finns i appendix B.1. Centrala gränsvärdessatsen är fundamental inom statistiken och säger alltså att summan av ett stort antal oberoende slumpvariabler är normalfördelad i gränsen då antal termer i summan går mot oändligheten, även om slumpvariablerna själva inte är normalfördelade. Detta är en orsak till att normalfördelningen dyker upp så ofta. Ett villkor för att centrala gränsvärdessatsen skall hålla är dock att slumpvariablerna har ändlig varians, vilket inte alltid är fallet. Innan vi ger ett exempel på en fördelning som för vissa parameterval har en oändlig varians, definierar vi nu överlevnadsfunktioner, som används i definitionen av paretofördelningen och i avsnitt 2.4 [7].

Definition 2.3 (Överlevnadsfunktion)**.** Låt X vara en icke-negativ slumpvariabel. Då definieras överlevnadsfunktionen av X som $S(x) = P(X > x)$ då $x \geq 0$.

En variabels överlevnadsfunktion i en punkt x specificerar alltså sannolikheten att denna variabel antar ett värde större än x, och kan tolkas som sannolikheten att en variabels livstid sträcker sig bortom tiden x. Med hjälp av dess överlevnadsfunktion definieras nu paretofördelningen, som är ett exempel på en potenslag [8].

Definition 2.4 (Paretofördelning)**.** En slumpvariabel X är paretofördelad med skalparameter $x_m > 0$ och formparameter $\alpha > 0$ om dess överlevnadsfunktion ges av

$$S(x) = \begin{cases} \left(\frac{x_m}{x}\right)^\alpha & \text{om } x \geq x_m, \\ 1 & \text{om } x < x_m. \end{cases}$$

Skalparametern x_m anger den undre gränsen på värdena som slumpvariabeln X kan anta. Formparametern α kallas även för svansindex och anger hur tung svans överlevnadsfunktionen har [9]. För låga α får vi särskilt tunga svansar. Vidare blir variansen oändlig för $\alpha \leq 2$, vilket innebär att en summa av oberoende slumpvariabler som är fördelade efter paretofördelningen med $\alpha \leq 2$ inte är normalfördelad enligt centrala gränsvärdessatsen, sats 2.1. Vi är alltså bortom normalfördelningen. Om vi istället låter α gå mot oändligheten erhålls något som liknar Diracs delta-funktion $\delta(x - x_m)$ kring punkten x_m, och variansen går mot noll. Innan vi definierar tungsvansade fördelningar och funktioner måste vi definiera den momentgenererande funktionen [3].

Definition 2.5 (Momentgenererande funktion)**.** Avbildningen $M_X : \mathbb{R} \to \mathbb{R}$ kallas för den momentgenererande funktionen till slumpvariabeln X och definieras av $M_X(t) = \mathbb{E}[e^{tX}]$.

Den momentgenererande funktionen har egenskapen att $M_X^{(n)}(0) = \mathbb{E}[X^n]$, där n betecknar den n:te derivatan och $\mathbb{E}[X^n]$ kallas för det n:te momentet av X. Låt oss nu definiera tungsvansade fördelningar och funktioner [10].

Definition 2.6 (Tungsvansad fördelning och funktion)**.** Fördelningen av en slumpvariabel X, med fördelningsfunktion F, sägs vara tungsvansad om dess momentgenererande funktion $M_X(\lambda)$ är oändlig för alla $\lambda > 0$. Överlevnadsfunktionen S sägs vara tungsvansad om

$$\lim_{x \to \infty} e^{\lambda x} S(x) = \infty, \quad \text{för alla } \lambda > 0.$$

Att fördelningen av X är tungsvansad är ekvivalent med att dess överlevnadsfunktion S har tunga svansar [10]. Att en fördelning är tungsvansad innebär alltså att dess överlevnadsfunktion asymptotiskt inte kan begränsas av en exponentiellt avtagande funktion. Det finns många exempel på tungsvansade fördelningar där en av de vanligaste är de som följer potenslagar, exempelvis paretofördelningen. Notera att definitionen för tungsvansad fördelning innebär att den momentgenererande funktionen är oändlig för alla $\lambda > 0$, men inte nödvändigtvis att momenten är oändliga. Betrakta exempelvis paretofördelningen vars första moment, det vill säga väntevärdet, är oändligt om $\alpha \leq 1$, men ändligt om $\alpha > 1$, se även tabell 2 i appendix A. I nästa avsnitt undersöks bland annat väntevärdet och variansen hos fördelningarna som Galton-Watson-processen ger upphov till.

2.2 Galton-Watson-processer

Galton och Watsons modell bygger på antagandet att vi börjar med en vuxen man med ett givet efternamn, som antas få $k \in \mathbb{N}_0$ stycken söner med sannolikheten p_k. Under nästa generation antas dessa söner få nya söner med samma oberoende sannolikheter. Om det inte föds några söner under en hel generation har efternamnet dött ut. Modellen kan generaliseras för att beskriva andra sorters förlopp, vilket leder oss in på definitionen av den så kallade Galton-Watson-processen [11].

Definition 2.7 (Galton-Watson-process)**.** En Galton-Watson-process är en diskret och tidshomogen Markovkedja $\{Z_0, Z_1, \ldots\}$ med tillståndsrummet $\mathbb{N}_0$ som karaktäriseras av det rekursiva förhållandet

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_i, \quad Z_0 = 1. \tag{2.1}$$

Här innebär Z_i antalet individer i generation i, och X_i är oberoende och likfördelade slumpvariabler, $X_i \overset{D}{=} X$, där X har fördelningen

$$X = \begin{cases} 0, & \text{med sannolikhet } p_0, \\ 1, & \text{med sannolikhet } p_1, \\ 2, & \text{med sannolikhet } p_2, \\ \vdots & \end{cases}$$

vilket är processens så kallade *reproduktionsfördelning*.

Notera här att Markovkedjan är tidshomogen eftersom samma reproduktionsfördelning gäller oberoende av generation genom det rekursiva förhållandet. Se definition A.16 och A.17 i appendix A för vidare bakgrund. Att reproduktionsfördelningen är densamma oavsett generation är en användbar egenskap av Galton-Watson-processer, eftersom detta förenklar analysen av det långsiktiga beteendet hos processen. Vi definierar den sannolikhetsgenererande funktionen av X fördelad enligt reproduktionsfördelningen som

$$G(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} p_k s^k. \tag{2.2}$$

Den sannolikhetsgenererande funktionen av X har ett antal användbara egenskaper som presenteras i ekvationerna (2.3a)-(2.3d) [4].

$$G(1) = 1, \ \text{(2.3a)} \quad G(0) = p_0, \ \text{(2.3b)} \quad G'(1) = \mu, \ \text{(2.3c)} \quad G''(1) + G'(1) - (G'(1))^2 = \sigma^2. \ \text{(2.3d)}$$

Egenskap (2.3a) följer från att summan av alla sannolikheter $\sum_k p_k$ alltid är 1, och (2.3b) från att sätta in $s = 0$ i ekvation (2.2). Vidare är $G'(s) = p_1 + 2p_2 s + 3p_3 s^2 + \cdots = \mathbb{E}[X \cdot s^{X-1}]$, alltså $G'(1) = \mathbb{E}[X] = \mu$, vilket är egenskap (2.3c). Genom att derivera ytterligare en gång får vi egenskap (2.3d).

2.2.1 Ett exempel på en Galton-Watson-process

Innan vi fortsätter ger vi ett exempel på en enkel Galton-Watson-process. Vi börjar med att definiera en grundläggande diskret sannolikhetsfördelning [7].

Definition 2.8 (Tvåpunktsfördelning)**.** En slumpvariabel X är tvåpunktsfördelad med parametrar $p \in [0, 1]$, x_1 och x_2 om

$$X = \begin{cases} x_1, & \text{med sannolikhet } 1 - p, \\ x_2, & \text{med sannolikhet } p, \end{cases}$$

där $x_1, x_2 \in \mathbb{Z}$.

Väntevärdet för tvåpunktsfördelningen är $\mathbb{E}[X] = (1-p)\, x_1 + p\, x_2$. Ett exempel på en tvåpunktsfördelad Galton-Watson-process är där varje individ får antingen 0 eller 2 barn, så att $x_1 = 0$ och $x_2 = 2$. Utgående från definition 2.7 och definition 2.8 erhålls

$$X = \begin{cases} 0, & \text{med sannolikhet } 1 - p, \\ 2, & \text{med sannolikhet } p, \end{cases} \tag{2.4}$$

där $p \in [0, 1]$ är sannolikheten att en given individ förgrenas. Väntevärdet för tvåpunktsfördelningen i detta specialfall är $\mu = 2p$, så ett enkelt beroende på p avgör hur Galton-Watson-processen kommer att utvecklas, vilket vi kommer att se i nästkommande avsnitt. Antag att vi börjar med en individ i generation 0, alltså $Z_0 = 1$. Då är antalet individer i nästa generation $Z_1 = X$, där X är fördelad enligt (2.4). För Z_n gäller sedan rekursionsformeln (2.1). Den sannolikhetsgenererande funktionen av X i detta fall är $G_{\text{punkt}}(s) = (1 - p) + ps^2$.

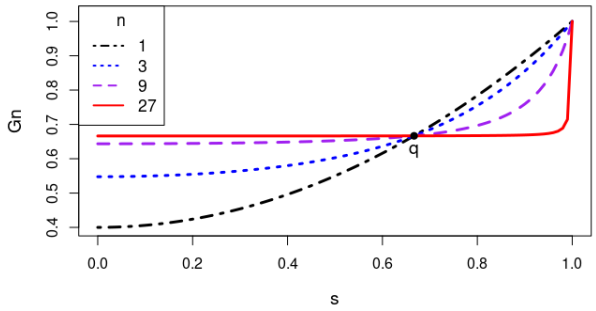
2.2.2 Sannolikhetsgenererande funktion och utrotningssannolikhet

Vi har generellt att den sannolikhetsgenererande funktionen av X är som i ekvation 2.2, där $p_k = P(X = k)$, alltså sannolikheten för en individ att få k barn i en godtycklig generation. Vi vill nu bestämma den sannolikhetsgenererande funktionen för Z_n, alltså $G_n(s) = \mathbb{E}[s^{Z_n} | Z_0 = 1]$, vilket vi kan använda för att bestämma processens utrotningssannolikhet. Från egenskap (2.3b) för genererande funktioner får vi att $G_n(0) = P(Z_n = 0 | Z_0 = 1)$, som tolkas som utrotningssannolikheten i generation n. Vidare har vi från egenskap (2.3c) att $G_n'(1) = E[Z_n | Z_0 = 1]$. För att kunna analysera långsiktig utrotningssannolikhet och väntevärde för Galton-Watson-processen vill vi finna ett enklare uttryck för G_n. Följande lemma formulerar G_n som en upprepad applicering av G – den sannolikhetsgenererande funktionen av X.

Lemma 2.1. *Låt* $G_n(s) = \mathbb{E}[s^{Z_n}|Z_0 = 1]$. *Då gäller*

$$G_n(s) = \underbrace{(G \circ G \circ \cdots \circ G)}_{n \text{ gånger}}(s) = \underbrace{G(G...G(s)...)}_{n \text{ gånger}}.$$

Beviset presenteras i appendix B.2 [4, s. 168]. Vi noterar att $G_n(1) = 1$ eftersom egenskap (2.3a) ger att $G(1) = 1$. I figur 4 kan vi se en illustration av vad som händer med G_n för stora n. Vi noterar att funktionen är monoton och verkar konvergera mot något värde q samt mot 1. Detta formuleras i följande sats, som bevisas i appendix B.2 [12].



Figur 4: Funktionen G_n i det tvåpunktsfördelade fallet för $s \in [0, 1]$, $n = 1, 3, 9, 27$ och $p = 0.6$.

Sats 2.2. *Funktionen* $G_n(s)$ *är monotont växande och konvergerar för alla* $s \in [0, 1]$.

För $s = 0$ är $G_n(0) = P(Z_n = 0)$, så när $n \to \infty$ motsvarar detta värde utrotningssannolikheten för processen. Att $G_n(0) \to q$ då $n \to \infty$ fås genom att använda detta faktum samt sats 2.2. Detta kan även ses i figur 4 och för ytterligare illustration hänvisas till figur 11 i appendix B.2. Vi vill nu bestämma utrotningssannolikheten q, genom en sats som bevisas i appendix B.2 [4, s. 169].

Sats 2.3 (Utrotningssannolikhet)**.** *Låt* G *vara den sannolikhetsgenererande funktionen av* X *där* X *fördelas enligt reproduktionsfördelningen för en Galton-Watson-process. Då är processens utrotningssannolikhet lika med den minsta icke-negativa roten till fixpunktsekvationen* $s = G(s)$.

I det tvåpunktsfördelade fallet med ekvation (2.4) fås fixpunktsekvationen $q = (1 - p) + pq^2$ från sats 2.3, vilken har lösningarna $q_1 = 1$ för alla p och $q_2 = 1/p - 1$ för $p \neq 0$. Om $p < 0.5$ så är $q_2 = 1/p - 1 > 1$ och alltså är då q_1 den minsta icke-negativa roten, varmed utrotningssannolikheten är $q = q_1 = 1$ enligt satsen. I fallet $p = 0.5$ sammanfaller de två rötterna för ekvationen, $q_1 = q_2 = 1$, och även då är utrotningssannolikheten $q = 1$. Att utrotningssannolikheten är 1 innebär att trädet är helt utrotat för tillräckligt stora n, det vill säga det finns n så att $Z_n = 0$. För $p > 0.5$ är istället utrotningssannolikheten $q = q_2 < 1$ som kan ses i figur 4 samt i figur 11 i appendix B.2.

2.2.3 Långsiktig populationsstorlek

För att mer ingående studera egenskaper såsom långsiktig populationsstorlek definierar vi de olika fallen för Galton-Watson-processen [4, s. 161].

Definition 2.9. Låt $\mu = \sum_{k=0}^{\infty} k\, p_k$ vara väntevärdet av reproduktionsfördelningen. En Galton-Watson-process kallas subkritisk då $\mu < 1$, kritisk då $\mu = 1$ och superkritisk då $\mu > 1$.

I det tvåpunktsfördelade fallet är $\mu = 2p$. Då motsvarar det subkritiska fallet att $p < 0.5$, det kritiska att $p = 0.5$ och det superkritiska att $p > 0.5$. Alltså, om sannolikheten p att få 2 barn är mindre än 0.5 kommer trädet utrotas med 100 % sannolikhet, och om sannolikheten p är större än 0.5 kommer det utrotas med sannolikheten $1/p - 1$. I figur 4 ses det superkritiska fallet för $p = 0.6$, där utrotningssannolikheten q således går mot $2/3$. I det kritiska fallet kommer däremot trädet utrotas med 100 % sannolikhet, samtidigt som genomsnittliga antalet barn är 1. Det är därför intressant att betrakta vad som då händer med processen i detta fall. Nu kan den förväntade långsiktiga populationsstorleken delas upp i tre fall, beroende på μ.

Lemma 2.2. *För den långsiktiga populationsstorleken gäller det att*

$$\lim_{n\to\infty} \mathbb{E}[Z_n] = \lim_{n\to\infty} \mu^n = \begin{cases} 0, & \mu < 1, \\ 1, & \mu = 1, \\ \infty, & \mu > 1. \end{cases}$$

Bevis. Beviset görs med induktion. Eftersom Z_1 har fördelningen $\{p_k\}_{k\in\mathbb{N}_0}$ är det klart att $\mathbb{E}[Z_1] = \mu$. Antag att $\mathbb{E}[Z_n] = \mu^n$ för ett godtyckligt n. Eftersom $\mathbb{E}[Z_n] = G_n'(1)$ och $G_{n+1}(s) = G_n(G(s))$ är

$$\mathbb{E}[Z_{n+1}] = G_{n+1}'(1) = G_n'(G(1))G'(1) = G_n'(1)G'(1) = \mathbb{E}[Z_n]\mathbb{E}[Z_1] = \mu^n\mu = \mu^{n+1}.$$

□

Av lemma 2.2 kan vi konstatera att det kritiska fallets beteende är svårtolkat eftersom den genomsnittliga generationsstorleken i gräns är 1, samtidigt som processens utrotningssannolikhet är 100 %. Variansen av Z_n då $n \to \infty$ kan ses i följande lemma.

Lemma 2.3. *För en Galton-Watson-process gäller det i det kritiska fallet att* $\mathrm{Var}(Z_n) \to \infty$ *då* $n \to \infty$.

Beviset presenteras i appendix B.3 [4, p. 163]. Att variansen för Z_n är oändlig i det kritiska fallet antyder att en population skulle kunna växa och bli hur stor som helst. Notera att centrala gränsvärdessatsen, sats 2.1, kräver ändligt väntevärde för att gälla, varmed värden långt bortom väntevärdet är osannolika. Vidare antyder den oändliga variansen att den momentgenererande funktionen i definition 2.6 kan vara oändlig, och alltså finns det möjlighet att Galton-Watson-processen ger upphov till tungsvansade fördelningar i det kritiska fallet.

2.2.4 Totala antalet individer W

Nästa koncept som betraktas är det totala antalet individer W för Galton-Watson-processen, som är besläktat med Z_n. Vi låter W_n beteckna antalet individer upp till och med generation n, alltså $W_n = \sum_{k=0}^{n} Z_k$. Då är W det totala antalet individer för alla generationer, det vill säga att $W_n \to W$ då $n \to \infty$. Följande gäller då för väntevärdet av W.

Lemma 2.4. *För en godtycklig Galton-Watson-process gäller att*

$$\mathbb{E}[W] = \begin{cases} \frac{1}{1-\mu}, & \mu < 1, \\ \infty, & \mu \geq 1, \end{cases}$$

där μ är reproduktionsfördelningens väntevärde.

Bevis. Vi har att

$$\mathbb{E}[W_n] = \mathbb{E}\left[\sum_{k=0}^{n} Z_k\right] = \sum_{k=0}^{n} \mathbb{E}[Z_k] = \sum_{k=0}^{n} \mu^k = \begin{cases} \frac{1-\mu^{n+1}}{1-\mu}, & \mu \neq 1, \\ n+1, & \mu = 1. \end{cases}$$

När $n \to \infty$ följer lemmat. □

Notera att väntevärdet av W är oändligt i det kritiska fallet, samtidigt som processens utrotningssannolikhet är 1. Vidare är variansen av W oändlig i det kritiska fallet likt variansen av Z_n i lemma 2.3 [12, s. 90]. Dessa resultat antyder återigen att Galton-Watson-processen med stor sannolikhet kan fortgå under lång tid och leda till stora populationer i det kritiska fallet. Vidare kan oändligt väntevärde och varians för W innebära att överlevnadsfunktionen av W är tungsvansad i det kritiska fallet enligt definition 2.6. I avsnitt 2.4 visar vi att överlevnadsfunktionen för W i det kritiska fallet är tungsvansad för ett specialfall av Galton-Watson-processen. Nu vill vi dock fortsätta genom att bestämma den sannolikhetsgenererande funktionen för W generellt och en ekvation för den erhålls från följande lemma [12].

Lemma 2.5. *Den genererande funktionen $H(s) = \mathbb{E}[s^W]$ för totala antalet individer W uppfyller ekvationen $H(s) = sG(H(s))$, där G är den genererande funktionen av X fördelad enligt reproduktionsfördelningen.*

Beviset presenteras i appendix B.3 [12, s. 90]. Genom att sätta $x = H(s)$ får vi ekvationen $x = sG(x)$ att lösa för att bestämma $H(s)$. Om vi låter $s = 1$ blir $x = H(1)$ och ekvationen förenklas till $x = G(x)$, vilket är fixpunktsekvationen för funktionen G från sats 2.3. Vi vet därifrån att 1 är en fixpunkt till G och får alltså att $H(1) = 1$. Vi vet också från definitionen av H som en sannolikhetsgenererande funktion att $H(1) = \mathbb{E}[1^W]$. Dessa två uttryck för $H(1)$ måste vara lika, alltså måste $\mathbb{E}[1^W] = 1$, vilket endast är uppfyllt om $W < \infty$. Det ger alltså att $H(1) = P(W < \infty) = 1$, vilket bekräftar att sannolikheten att processen dör ut i det kritiska fallet är 100 %. Vi återkommer till $H(s)$ i avsnitt 2.3.3, efter att vi har introducerat den så kallade LF-fördelningen.

2.3 LF-fördelningen

En specifik fördelning som är användbar som reproduktionsfördelning i Galton-Watson-processen är den så kallade LF-fördelningen (eng. *linear fractional distribution*), som bygger på den geometriska och skiftade geometriska fördelningen. Låt oss därför först definiera dessa [7].

Definition 2.10 (Geometrisk fördelning)**.** En slumpvariabel X är geometriskt fördelad med parameter $p \in [0, 1]$ om dess sannolikhetsfunktion ges av

$$f_X(k) = p(1-p)^k, \quad k = 0, 1, 2, \ldots.$$

Detta skrivs $X \overset{D}{\sim} \text{Geom}(p)$. En slumpvariabel Y är då skiftat geometriskt fördelad med parameter $p \in [0, 1]$ om $Y \overset{D}{=} X + 1$, alltså om dess sannolikhetsfunktion är

$$f_Y(k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots.$$

Tolkningen av den geometriska fördelningen är att den beskriver antal oberoende misslyckade försök tills det första lyckade, där ett lyckat försök sker med sannolikhet p. För den skiftade geometriska fördelningen räknas även det lyckade försöket med. Ett exempel på en situation där den geometriska fördelningen kan användas är när en familj vill få barn ända tills de får en son. Då kan antalet döttrar modelleras med den geometriska fördelningen, som får parametern $p = 0.5$, alltså sannolikheten att de vid varje försök får en son. En fördelning som kan tolkas på ett liknande sätt är LF-fördelningen, men där används istället två parametrar, p_0 och p. LF-fördelningen skulle i det här fallet modellera familjens totala antalet söner, där sannolikheten att få den första sonen är $1 - p_0$ och sannolikheten att få en son efter den första sonen är $1 - p$ per son. I definition 2.11 definieras nu LF-fördelningen genom dess sannolikhetsgenererande funktion och i sats 2.4 presenterar vi den genererande funktionens korresponderande sannolikhetsfunktion [6].

Definition 2.11 (LF-fördelningen)**.** En slumpvariabel X sägs vara LF-fördelad med parametrarna $p_0, p \in [0, 1]$ om dess sannolikhetsgenererande funktion kan skrivas som

$$G(s) = E[s^X] = p_0 + (1 - p_0)\frac{ps}{1 - (1-p)s}. \tag{2.5}$$

Sats 2.4. *Om en slumpvariabel X har sannolikhetsfunktionen*

$$f(k) = \begin{cases} p_0, & \text{om } k = 0, \\ p(1-p_0)(1-p)^{k-1}, & \text{om } k \geq 1, \end{cases} \tag{2.6}$$

har den ekvation (2.5) *som sannolikhetsgenererande funktion för $|s(1-p)| < 1$.*

Beviset finns i appendix B.4. Väntevärdet fås genom egenskap (2.3c) som ger $\mu = G'(1) = (1 - p_0)/p$. Notera att $P(X > 0) = 1 - p_0$ och att X givet $X \geq 1$ är fördelad enligt den skiftade geometriska fördelningen. Vidare, om $p_0 = p$ får vi att $G(s) = p/(1-(1-p)s)$, vilket är den sannolikhetsgenererande funktionen för den geometriska fördelningen. Det innebär att LF-fördelningen med två identiska parametrar är detsamma som den geometriska fördelningen. Detta beror på att sannolikheten att få första sonen då är samma som att få resten av sönerna, vilket är tolkningen av den geometriska fördelningen. Dessutom, om $p_0 = 0$, får vi $G(s) = ps/(1 - (1-p)s)$, vilket är den sannolikhetsgenererande funktionen för den skiftade geometriska fördelningen. I ekvation (2.6) ser vi att $p_0 = 0$ ger $f(0) = P(X = 0) = 0$ vilket kan tolkas som att räkningen börjar från $X = 1$, som i den skiftade geometriska fördelningen.

2.3.1 Egenskaper och parametrar för LF-fördelningen

En användbar egenskap för slumpvariabler med LF-fördelningen är att deras sannolikhetsgenererande funktion kan skrivas som

$$G(s) = \frac{p_0 + (p - p_0)s}{1 + (p-1)s},$$

där alltså både täljare och nämnare är linjära funktioner, vilket är just anledningen till att det kallas för LF-fördelningen. Om vi betraktar ett godtyckligt linjärt bråk är det tydligt att det blir ett linjärt bråk även vid upprepad sammansättning eftersom

$$\frac{a + b\frac{a+bx}{c+dx}}{c + d\frac{a+bx}{c+dx}} = \frac{ac + adx + ab + b^2x}{c^2 + cdx + ad + bdx} = \frac{ac + ab + (ab + b^2)x}{c^2 + ad + (cd + bd)x} = \frac{a' + b'x}{c' + d'x}.$$

Då LF-fördelningen kan skrivas som ett linjärt bråk gäller det att sammansättningen också är en LF-fördelning, alltså är G upprepad n gånger också en LF-fördelning. Vidare ger lemma 2.1 att den sannolikhetsgenererande funktionen G_n för Z_n då reproduktionsfördelningen är LF-fördelad kan skrivas som upprepad applicering av den sannolikhetsgenererande funktionen G för X fördelad efter LF-fördelningen. Därmed är

$$G_n(s) = p_0^{(n)} + (1 - p_0^{(n)})\frac{p^{(n)}s}{1 - (1 - p^{(n)})s}. \tag{2.7}$$

Enligt sats 2.3 får vi utrotningssannolikheten för processen genom att lösa för fixpunkter till den sannolikhetsgenererande funktionen av X fördelad enligt reproduktionsfördelningen. I detta fall är LF-fördelningen reproduktionsfördelningen, således söker vi lösningar s till ekvationen

$$s = p_0 + (1 - p_0)\frac{ps}{1 - (1 - p)s}.$$

Dessa är $s = 1$ och $s = p_0/(1 - p) := s_0$. Med hjälp av ekvation (2.7) och faktumet att s_0 är en fixpunkt kan vi bestämma förhållanden mellan p, p_0 och $p^{(n)}, p_0^{(n)}$, som formuleras i följande sats.

Sats 2.5. *I det sub- och superkritiska fallet gäller* (2.8) *och i det kritiska fallet där* $p = 1 - p_0$ *gäller* (2.9)*. Parametrarna* $p^{(n)}$ *och* $p_0^{(n)}$ *i* G_n *kan uttryckas som*

$$\begin{cases} p^{(n)} = \dfrac{1 - \frac{p_0}{1-p}}{\left(\frac{1-p_0}{p}\right)^n - \frac{p_0}{1-p}}, \\ p_0^{(n)} = \dfrac{1 - \left(\frac{p}{1-p_0}\right)^n}{\frac{1-p}{p_0} - \left(\frac{p}{1-p_0}\right)^n}, \end{cases} \tag{2.8}$$

$$\begin{cases} p^{(n)} = \dfrac{p}{p + (1-p)n}, \\ p_0^{(n)} = \dfrac{(1-p)n}{p + (1-p)n}, \end{cases} \tag{2.9}$$

där p *och* p_0 *är parametrarna i LF-fördelningen med sannolikhetsgenererande funktionen* G*.*

Vi får alltså ett givet uttryck för G applicerat n gånger, till skillnad från i avsnitt 2.2.2, där inget explicit uttryck kunde formuleras. Satsen bevisas i appendix B.4. Då $n \to \infty$ får vi att $p^{(n)} = 1 - [(1-p)/p_0]$ och $p_0^{(n)} = 1$ i det subkritiska fallet, $p^{(n)} = 0$ och $p_0^{(n)} = 1$ i det kritiska fallet samt $p^{(n)} = 0$ och $p_0^{(n)} = p_0/(1-p)$ i det superkritiska fallet.

2.3.2 Fördelningen för Z_n givet att processen har överlevt till generation n

När reproduktionsfördelningen för Galton-Watson-processen är LF-fördelad så kan asymptotiska fördelningar för Z_n i det kritiska och subkritiska fallet härledas genom att använda uttrycken för $p^{(n)}$ och $p_0^{(n)}$ från sats 2.5. Först måste dock exponentialfördelningen definieras [7].

Definition 2.12 (Exponentialfördelning)**.** En slumpvariabel X är exponentialfördelad med parameter $\lambda > 0$ om dess täthetsfunktion ges av $f(x) = \lambda e^{-\lambda x}$ för $x \geq 0$.

I lemmorna 2.6 och 2.7 presenteras fördelningen för (Z_n/n) givet $Z_n \geq 1$ i det kritiska fallet och Z_n givet $Z_n \geq 1$ i det subkritiska fallet, alltså fördelningen för Z_n givet att processen ej är utdöd vid generation n.

Lemma 2.6. *I det kritiska fallet är* $P((Z_n/n) \geq x | Z_n \geq 1) \sim \exp(-xp/(1-p))$ *vilket är överlevnadsfunktionen för den exponentiella fördelningen med parameter* $\lambda = p/(1-p)$.

Lemma 2.7. *I det subkritiska fallet är* $P(Z_n \geq x+1 \,|\, Z_n \geq 1) \sim ((1-p)/p_0)^x$ *vilket är överlevnadsfunktionen för den skiftade geometriska fördelningen med parameter* $1-[(1-p)/p_0]$.

Bevisen finns i appendix B.5. Tolkningen av lemma 2.7 är att Z_n givet $Z_n \geq 1$ är skiftad geometriskt fördelad med parameter $1-[(1-p)/p_0]$. Det är rimligt att $(x+1)$ används, då x börjar på 0 och $P(Z_n \geq 0 | Z_n \geq 1) = 0$. I det kritiska fallet i lemma 2.6 betraktas istället Z_n/n eftersom storleksordningen är större än i det subkritiska fallet. Mer om storleksordning kommer i avsnitt 2.4.2. Att den är asymptotiskt exponentialfördelad kan tänkas tyda på lätta svansar enligt definition 2.6 men eftersom fördelningen gäller för Z_n/n kan denna slutsats ej dras, och det långsiktiga beteendet i det kritiska fallet måste undersökas ytterligare. Detta görs genom att undersöka uttrycket för den genererande funktionen för totala antalet individer för LF-fördelningen.

2.3.3 Genererande funktion för totala antalet individer

Lemma 2.5 ger relationen mellan genererande funktionen G av X fördelad enligt reproduktionsfördelningen och genererande funktionen H för totala antalet individer W. Vi har då att ekvationen $x = sG(x)$ ger

$$x = s\frac{p_0 + (p-p_0)x}{1+(p-1)x}.$$

Vi får x genom att lösa andragradsekvationen, vilket ger

$$x = H(s) = \frac{1+(p_0-p)s - \sqrt{1+(p_0-p)^2s^2 - 2s(p_0+p-2p_0p)}}{2(1-p)}.$$

Vi har valt den mindre av de två lösningarna då den större är större än 1, vilket ej är tillåtet då $H(1) = 1$. I det kritiska fallet då $p = 1-p_0$ fås

$$H_k(s) = \frac{1+(1-2p)s - \sqrt{1+(1-2p)^2s^2 - 2s(1-2(1-p)p)}}{2(1-p)}. \tag{2.10}$$

Det här uttrycket är användbart till asymptotisk analys i avsnitt 2.4.1.

2.4 Potenslag i det kritiska fallet

Här presenteras teori för det asymptotiska beteendet av Galton-Watson-processen i det kritiska fallet, när reproduktionsfördelning ges av LF-fördelningen. Genom asymptotisk analys och den tauberska satsen 2.6 ges ett uttryck för överlevnadsfunktionen av det totala antalet individer W som en potenslag. Genom denna potenslag kan slutsatser dras om dess asymptotiska logaritmiska fördelning samt om storleksordningen av maximalt antal individer. Låt oss först definiera asymptotisk ekvivalens för att kunna använda asymptotisk analys [13].

Definition 2.13 (Asymptotisk ekvivalens)**.** Givet funktionerna $f(x)$ och $g(x)$ gäller att $f(x) \sim g(x)$, då $x \to \infty$, om och endast om

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$$

Om $f(x) \sim g(x)$ beskrivs det som att $f(x)$ asymptotiskt beter sig som $g(x)$. Här betecknar $\sim$ asymptotisk ekvivalens, vilket ej ska förväxlas med beteckningen $\overset{D}{\sim}$ som betyder *fördelat som*, se notationslistan. I sats 2.6 som används för att ta fram potenslagen i det kritiska fallet förekommer långsamt varierande funktioner, vilka är ett användbart specialfall av regelbundet varierande funktioner, se definition A.15. Låt oss därför definiera detta [10].

Definition 2.14 (Långsamt varierande funktion)**.** En positiv funktion L, definierad på $[0,\infty)$, är långsamt varierande om $L(ct) \sim L(t)$ för alla $c > 0$.

Den enklaste sortens långsamt varierande funktion är en konstant funktion, $L(t) = K$.

2.4.1 Sats om konvergent potensserie och asymptotisk analys

Följande tauberska sats beskriver asymptotiskt beteende hos konvergenta potensserier [14].

Sats 2.6. *Antag att vi har en monoton följd* $\{t_k\}_{k=0}^{\infty}$ *där* $t_k \geq 0$ *för alla* k. *Definiera polynomet* $T_n(s) = \sum_{k=0}^{n-1} t_k s^k$, *och potensserien* $T(s) = \lim_{n\to\infty} T_n(s)$. *Vi antar vidare att* $T(s)$ *konvergerar för* $0 \leq s < 1$. *Då gäller för* $0 < \rho < \infty$, *och för funktioner* L *som varierar långsamt vid oändligheten att*

$$T(s) \sim \frac{1}{(1-s)^{\rho}} L\left(\frac{1}{1-s}\right), \quad s \to 1^{-}, \tag{2.11}$$

är ekvivalent med

$$t_n \sim \frac{1}{\Gamma(\rho)} n^{\rho-1} L(n), \quad n \to \infty. \tag{2.12}$$

För specialfallet att den långsamt varierande funktionen L är konstant ser vi att ekvation (2.12) ger en potenslag för följden t_n. Vi vill använda denna sats för att visa att överlevnadsfunktionen för totala antalet individer W är asymptotiskt potensfördelad för LF-fördelningen i det kritiska fallet, alltså tungsvansad. Fördelningen för W är nämligen tungsvansad om dess överlevnadsfunktion S_W har tunga svansar, enligt definition 2.6. Vi låter därmed följden t_n ges av överlevnadsfunktionen $S_W(n) = P(W \geq n)$, vilket kan ses som en följd eftersom funktionen är diskret. Enligt definition 2.3 är egentligen $S_W(n) = P(W > n) = P(W \geq n+1)$, men på grund av att det är ett asymptotiskt beteende hos överlevnadsfunktionen som undersöks är den extra ettan försumbar och vi kan skriva $S_W(n) = P(W \geq n)$. Överlevnadsfunktionen kan också skrivas $S_W(n) = \sum_{k=n}^{\infty} P(W = k)$ vilket ger att följden är monotont avtagande eftersom $p_k \geq 0$ för alla k. För att kunna visa att följden uppfyller (2.12) måste vi därmed visa att potensserien $T(s)$ konvergerar för $0 \leq s < 1$, men inte för $s = 1$ och att (2.11) håller med konstant L och $0 < \rho < \infty$. Om vi skriver ut serien får vi

$$T(s) = \sum_{n=0}^{\infty} t_n s^n = \sum_{n=0}^{\infty} \sum_{k=n}^{\infty} P(W = k) s^n.$$

För att kunna undersöka konvergens och få fram en asymptotisk form som uppfyller (2.11) uttrycker vi $T(s)$ i $H_k(s)$, eftersom den har ett slutet uttryck i (2.10). Genom algebraiska omskrivningar som finns i appendix B.6 får vi

$$T(s) = H_k(s) + \frac{1 - H_k(s)}{1-s}.$$

Vi ser här att uttrycket konvergerar för $s \in [0, 1)$ men divergerar för $s = 1$. Vi gör nu asymptotisk analys då $s \to 1^{-}$, som finns i appendix B.6. Analysen ger

$$T(s) \sim \frac{\sqrt{2}}{\sigma\sqrt{1-s}}, \quad s \to 1^{-}. \tag{2.13}$$

Genom användning av satsen ser vi att (2.13) motsvarar (2.11) med koefficienten $\rho = 1/2$ och den långsamt varierande funktionen $L(1/(1-s)) = \sqrt{2}/\sigma$, vilket är en konstant och därmed uppfyller definition 2.14. Eftersom följden är monoton, och $0 < \rho < \infty$ gäller förutsättningarna för (2.12), som då ger

$$t_n = S_W(n) \sim \sqrt{\frac{2}{\pi\sigma^2}} n^{-1/2}, \quad n \to \infty. \tag{2.14}$$

Vidare får vi då

$$\lim_{n\to\infty} e^{\lambda n} S_W(n) = \lim_{n\to\infty} \sqrt{\frac{2}{\pi\sigma^2}} \frac{e^{\lambda n}}{n^{1/2}} = \infty, \quad \text{för alla } \lambda > 0,$$

eftersom exponentialfördelningen växer snabbare än potensfunktioner. Därmed är överlevnadsfunktionen S_W tungsvansad enligt definition 2.6 och således är också fördelningen av W tungsvansad.

2.4.2 Logaritmiskt fall och maximalt antal individer

Att överlevnadsfunktionen är asymptotiskt potensfördelad leder till en del användbara egenskaper, som att dess logaritm är exponentialfördelad och att maximalt antal individer kan bestämmas.

Lemma 2.8. *Om X är en slumpvariabel vars överlevnadsfunktion är asymptotiskt potensfördelad så gäller att slumpvariabeln* $\log(X)$ *är asymptotiskt exponentialfördelad.*

Beviset finns i appendix B.7. I vårt fall har vi att $\log(\pi\sigma^2 W/2)$ är asymptotiskt exponentialfördelad med parameter $1/2$. Antag nu att vi gör N försök av Galton-Watson-processen och får totala antalet individer $W_1, W_2, \ldots, W_N$ för respektive träd. I följande lemma som bevisas i appendix B.7 presenteras hur maximalt antal individer beror på antal försök.

Lemma 2.9. *Låt $W_{max}(N)$ vara stort och beteckna värdet på den maximala individen. Då har vi*

$$W_{max}(N) = \mathcal{O}(N^2). \tag{2.15}$$

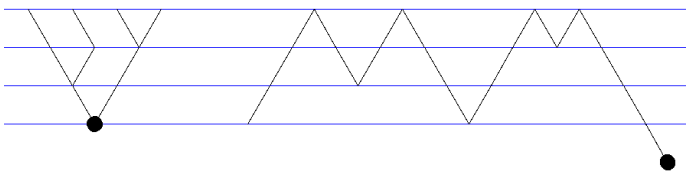
Lemmat ger till exempel en miljon som storleksordningen av antal individer i det största trädet vid tusen försök. Observera dock att detta resultat endast gäller för det kritiska fallet.

3 Simuleringar

I detta avsnitt presenteras metoder för de genomförda simuleringarna. Först förklaras konturprocessen, vilket kan användas för att snabbt simulera Galton-Watson-processen med LF-fördelning. Sedan presenteras teorin bakom approximationen av S_W. Slutligen presenteras de grafiska metoderna log-log-plot, qq-plot och Hill-estimering som använts för att illustrera fördelningen av W.

3.1 Konturprocessen

Vi behöver först en metod för att simulera Galton-Watson-processer. Konturprocessen kan användas som en simuleringsmetod för att effektivt simulera Galton-Watson-processen med en slumpvandring istället för med ett träd. Dess syfte är att snabbt kunna beräkna totala antalet individer för en process. För att förstå principen kan vi utgå från trädet till vänster i figur 5 och betrakta det som hela trädet för en process. Vi startar längst ner och följer grenarna runt trädet, med början på vänster sida. Detta innebär alltså först tre grenar uppåt, vilket motsvaras av tre steg uppåt i tillhörande kontur, som ses till höger i figur 5. Vid grenens ände vänder vi och går nedåt längs två grenar, vilket ger två steg nedåt i konturen, innan vi kan vända och gå uppåt igen. När vi åter är tillbaka vid botten i trädet, efter att ha gått runt hela trädet, har vi i den motsvarande konturen gått lika många steg uppåt som nedåt och alltså tagit med varje gren två gånger. För att modellera en Galton-Watson-process med konturprocessen betraktas således en slumpvandring som börjar på noll och avslutas vid minus ett. Det sista steget från noll till minus ett kan räknas som den ursprungliga individen och det totala antalet individer ges då av denna plus antal steg uppåt i slumpvandringen. Fördelen med denna simuleringsmetod är att antalet individer i varje generation ej behöver sparas, utan det räcker att räkna det totala antalet steg som tagits när processen avslutats, vilket medför att simuleringarna går snabbare än om hela trädet simulerades.



Figur 5: Ett träd över en Galton-Watson-process med 4 generationer och 10 individer (till vänster) samt dess tillhörande kontur som ges av konturprocessen (till höger). Konturen erhålls från trädet genom att starta längst ner och följa grenarna runt hela trädet, med början på vänster sida.

I praktiken är det smidigt att använda LF-fördelningen för att simulera Galton-Watson-processer. I LF-fallet är sannolikheten för en individs första barn $1 - p_0$ och sannolikheten att den får barn

efter det första barnet $1-p$ för varje barn. Således ges, oberoende av varandra, fördelningen för antal grenar uppåt av en geometrisk fördelning med parametern p_0 medan fördelningen för antal grenar nedåt ges av geometrisk fördelning med parametern $1-p$. För en mer ingående beskrivning av hur detta simuleras i praktiken, se appendix E. I det kritiska fallet är $p = 1 - p_0$, varmed fördelningarna uppåt och nedåt är symmetriska. I det subkritiska fallet leder de två oberoende fördelningarna dock till att slumpvandringen nedåt sker snabbare än den uppåt.

3.2 `Java`-simulering och approximering av överlevnadsfunktionen av W

Simuleringar av Galton-Watson-processen genom konturprocessen görs i `Java`, där varje simulering får fortgå ända tills nivån vid minus ett i konturprocessen nås. Flera simuleringar körs parallellt så att stora dataset kan erhållas. Fullständig `Java`-kod för detta finns i appendix E. Eftersom konturprocessen kräver generering av slumpvariabler som följer den skiftade geometriska fördelningen så har även detta implementerats i `Java`. Den matematiska bakgrunden finns i appendix C.1.

Antag att N stycken simuleringar har gjorts. Givet detta dataset är målet att approximera överlevnadsfunktionen $S_W(n) = P(W \geq n), n \in \mathbb{N}$. Detta görs genom att låta

$$P(W = k) := \frac{f_k}{N}, \tag{3.1}$$

där f_k är antalet simuleringar som gav träd av storleken k, och därmed

$$S_W(n) := \sum_{k=n}^{W_{\max}} \frac{f_k}{N}, \tag{3.2}$$

där $W_{\max}$ är antalet individer i det största trädet som simuleringarna gav. Eftersom träd av storlekar nära $W_{\max}$ är relativt sällsynta så approximeras $S_W(n)$ för $n \approx W_{\max}$ relativt dåligt. Summans övre gräns i ekvation (3.2) är $W_{\max}$ och inte ∞ som den borde vara enligt definitionen av överlevnadsfunktion av W. I fallet av exempelvis $n = W_{\max}$ så approximeras $S_W(W_{max})$ då som $P(W = W_{\max})$, och värdena $W_{\max} + 1, \ldots, \infty$ förloras. När den simulerade $S_W(n)$ betraktas för stora n är det nödvändigt att ta hänsyn till detta.

3.3 Grafiska metoder för att avgöra om data är tungsvansad

Det finns flera grafiska metoder för att avgöra vilken fördelning data har, som log-log-plot och qq-plot, och det är bäst att kombinera dessa metoder för att få ett rättvisande svar [15]. Dessa metoder används i resultat- och diskussionsdelen, men främst i avsnitt 4.2. Även en metod som ej kräver parameteruppskattning används, nämligen Hill-estimatorn. Det finns en mängd ytterligare metoder för att avgöra om en fördelning är tungsvansad, men de riktar sig främst mot fördelningar som har ändlig varians eller ändligt väntevärde, vilket inte gäller för W i det kritiska fallet [15].

Det är möjligt att visa att ett datasets överlevnadsfunktion inte följer en potenslag genom att rita en graf av överlevnadsfunktionen med logaritmiska skalor på båda axlar (eng. *log-log plot*). Antag att $t_n = cn^\alpha, c > 0, \alpha < 0$. Då är $\log t_n = \log(cn^\alpha) = \log c + \alpha \log n$. Om överlevnadsfunktionen följer en potenslag skall grafen visa en linjär funktion med negativ lutning, $T(N) = C + AN$, $A < 0$. Ett exempel på en log-log-plot ses i figur 7 i avsnitt 4.2 som visar t_n för olika μ. Log-log-plots är användbara för att avvisa hypoteser, till exempel följer datan ej en potenslag om dess kurva i en log-log-plot är olinjär [15]. Däremot kan log-log-plots inte lika säkert användas för att bestämma vilken fördelning data har eftersom många fördelningar, exempelvis exponentialfördelningen och normalfördelningen, är svåra att skilja från varandra. Därmed bör log-log-plots främst användas för att avgöra om data är potensfördelad eller inte, och för mer information behövs en kombination av grafiska metoder.

En qq-plot (förkortning för engelskans *quantile-quantile-plot*) är ett grafiskt verktyg för att jämföra data med en given fördelning. Detta avsnitt fokuserar på den grafiska tolkningen men mer matematisk bakgrund finns i appendix C.2. Om ett linjärt samband erhålls kan slutsatsen dras att datan och fördelningen är lika. Om sambandet däremot inte är linjärt finns olika fall, kurvan kan vara linjär men ha en annan lutning, eller så kan kurvan vara böjd uppåt eller neråt. I fallet där kurvan är linjär men med annan lutning så skiljer sig varianserna mellan fördelningarna åt. Då

kurvan är böjd uppåt tolkas det som att datan har lättare svansar än fördelningen, och i fallet då den är böjd neråt har datan tyngre svansar.

I detta arbete jämförs datan med exponentialfördelningen i qq-plots för att det finns paket i `R` för detta ändamål, där parametrar ej behöver ansättas utan jämförelsen görs med en generell exponentialfördelning. Om datans svans följer en exponentialfördelning är datan inte tungsvansad, enligt definition 2.6, vilket utnyttjas för att avgöra för vilket värde på μ gränsen till tunga svansar går i det subkritiska fallet. Ett exempel på en qq-plot i det subkritiska fallet ses i figur 8b. Utöver detta jämförs logaritmen av datan med exponentialfördelningen, eftersom logaritmen av en asymptotiskt potensfördelad slumpvariabel är asymptotiskt exponentialfördelad enligt lemma 2.8. Detta används för att avgöra för vilket värde på μ som svansen i det subkritiska fallet inte längre följer en potenslag, och ett exempel på en sådan qq-plot ses i figur 8a.

Hill-estimatorn har en rik bakomliggande matematisk teori och används för att uppskatta så kallade svansindex, som exempelvis svansindexet α, se definition 2.4. Denna teori presenteras inte här eftersom det finns paket i `R` som gör denna uppskattning och presenterar resultatet grafiskt, så fokus ligger på tolkningen av dessa grafiska resultat. Tolkningen är att om grafen stabiliserar sig för något svansindex α så tyder det på att fördelningen följer en potenslag med denna exponent. Stabiliseras däremot inte grafen tyder det på att fördelningen ej följer en potenslag.

4 Resultat och diskussion

I detta avsnitt sammanfattas arbetets teoretiska resultat och simuleringsresultaten presenteras.

4.1 Teoretiska resultat

Galton-Watson-processers långsiktiga beteende har undersökts generellt, till exempel processens utrotningssannolikhet q i de olika fallen, vars värde erhölls från upprepad användning av sannolikhetsgenererande funktioner i avsnitt 2.2.2. Det långsiktiga beteendet av väntevärdet och variansen för Z_n och W har också beskrivits i de olika fallen, i avsnitt 2.2.3 respektive 2.2.4. Det noterades att både väntevärdet och variansen för det totala antalet individer W är oändligt i det kritiska fallet samtidigt som processens utrotningssannolikhet är 100 %. Detta motsäger en exponentiellt avtagande svans hos fördelnings- eller överlevnadsfunktionen av W och antyder att en potenslag med svansindex $\alpha < 1$ istället kan vara en relevant modell.

Vidare har Galton-Watson-processen undersökts i fallet där reproduktionsfördelningen ges av LF-fördelningen. Från LF-fördelningens egenskaper, erhölls ett slutet uttryck för den genererande funktionen för Z_n, eftersom parametrarna i n-fallet kan uttryckas i originalparametrarna, vilket formulerades i sats 2.5. Specifikt gäller i det subkritiska fallet att Z_n givet $Z_n \geq 1$ är asymptotiskt skiftad geometriskt fördelad med parameter $1 - [(1-p)/p_0]$ enligt lemma 2.7. Vidare gäller i det kritiska fallet att (Z_n/n) givet $Z_n \geq 1$ är asymptotiskt exponentialfördelad med parameter $\lambda = p/(1-p)$, vilket presenterades i lemma 2.6.

Genom att härleda en ekvation mellan den genererande funktionen av X fördelad enligt reproduktionsfördelningen och den genererande funktionen för W, i lemma 2.5, visades i avsnitt 2.3.3 även hur ett slutet uttryck för den genererande funktionen av W i det kritiska fallet kan erhållas. Fortsättningsvis användes sats 2.6 i avsnitt 2.4.1 för att visa att överlevnadsfunktionen för W asymptotiskt följer en potenslag. Härledningen krävde uttrycket för den genererande funktionen av W samt asymptotisk analys. Slutligen erhölls

$$S_W(n) \sim \sqrt{\frac{2}{\pi\sigma^2}} n^{-1/2}, \quad n \to \infty,$$

och överlevnadsfunktionen S_W av W visade sig vara tungsvansad enligt definition 2.6. Detta innebär att fördelningen för totala antalet individer i Galton-Watson-processen är tungsvansad i det kritiska fallet med LF-fördelningen som reproduktionsfördelning. Målet i kommande avsnitt är att verifiera detta resultat med simuleringar, samt undersöka det subkritiska fallet.

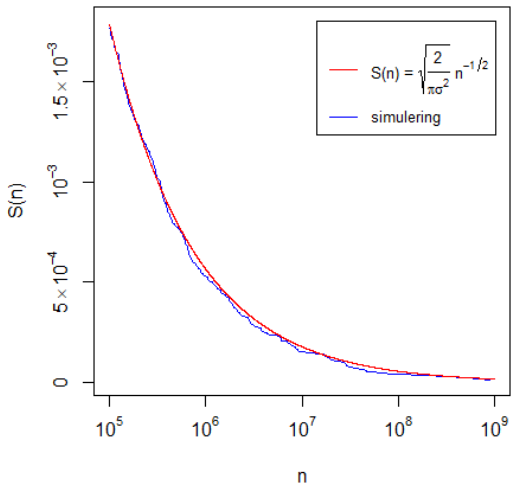
I avsnitt 2.4.2 visades även att logaritmen av en asymptotiskt potensfördelad variabel är asymptotiskt exponentialfördelad, se lemma 2.8, vilket används i de grafiska metoderna som beskrivs i avsnitt 3.3. Från potenslagen för S_W härleddes även regeln i lemma 2.9 att tusen försök ger ett tak på en miljon som största träd, vilket kan vara användbart i simuleringar.

4.2 Simuleringsresultat

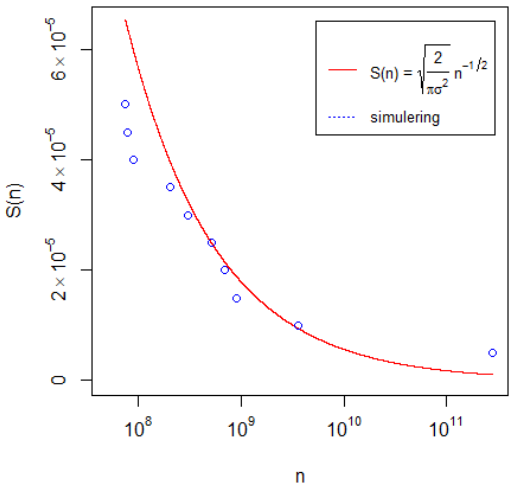
Härmed presenteras data som erhållits genom att simulera Galton-Watson-processen. Det undersöks huruvida datan har tunga svansar genom användningen av grafiska metoder som log-log-plot och qq-plot. I det kritiska fallet verifieras den analytiska potenslagen genom att jämföra dess graf med datans approximerade överlevnadsfunktion. I det subkritiska fallet undersöks huruvida datan följer en potenslag för olika μ och om det finns en gräns med avseende på μ där fördelningen inte längre följer en potenslag, och inte längre har tunga svansar. Fler figurer som styrker de erhållna resultaten presenteras i appendix D.1.

4.2.1 Verifiering av tunga svansar i det kritiska fallet

Figur 6 visar två jämförelser mellan den simulerade samt teoretiska överlevnadsfunktion av W, som ges av $S_W(n) \sim \sqrt{(2/\pi\sigma^2)}n^{-1/2}$, i det kritiska fallet för olika intervall på n. Det dataset som användes bestod av 200 000 simuleringar, och används alltså i både figur 6a och 6b.



(a) En jämförelse mellan den simulerade och teoretiska överlevnadsfunktion av W i det kritiska fallet för intervallet $n \in [10^5, 10^9]$. Skillnaden mellan de två graferna är liten över hela intervallet, vilket tyder på att den analytiska formeln är korrekt.



(b) En jämförelse mellan den simulerade och teoretiska överlevnadsfunktion av W för de tio största träd som simuleringarna i det kritiska fallet gav, i intervallet $n \in [5 \cdot 10^7, 3 \cdot 10^{11}]$. Det största trädet är av storleken 300 miljarder.

Figur 6: Jämförelser mellan den simulerade och teoretiska överlevnadsfunktion av W. Båda figurerna visar samma dataset, men i olika intervall för n.

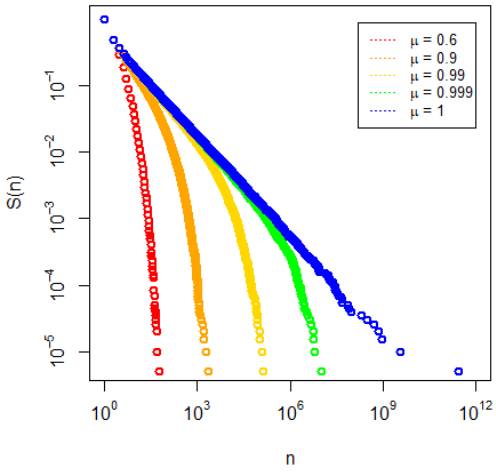
Figur 6a visar att den simulerade fördelningen följer den egentliga väl i intervallet $n \in [10^5, 10^9]$. Att betrakta fördelningen för $S_W(n)$ för $n < 10^5$ kan generellt vara intressant men är ej meningsfullt då den teoretiska formeln för $S_W(n)$ endast gäller asymptotiskt, samt då detta arbete fokuserar på svansar hos fördelningar. Det genomsnittliga felet genom intervallet $[10^5, 10^9]$ är ungefär $2.1 \cdot 10^{-5}$, och det största felet är ungefär $5.8 \cdot 10^{-5}$, som fås vid $n \approx 725\,000$. Detta kan jämföras med det totala spannet för $S_W(n)$ som figuren visar, vilket är $S_W(n) \in [0, 1.8 \cdot 10^{-3}]$. Avvikelserna är alltså små, vilket tyder på att den erhållna formeln är korrekt.

Vidare visar figur 6b överlevnadsfunktionen av W i det kritiska fallet i intervallet $n \in [5 \cdot 10^7, 3 \cdot 10^{11}]$, där de tio största trädstorlekarna som simuleringen gav har märkts ut med ringar. Avvikelserna är här större än i figur 6a, vilket beror på att y-axeln visar ett mindre intervall än förut, samt för att approximationen av överlevnadsfunktionen blir relativt dålig för mycket stora n, eftersom så få simuleringar ger träd av dessa storlekar. Denna effekt förklaras mer i avsnitt 3.2. Av denna anledning kan det vara av intresse att simulera fler träd än 200 000, men detta kan ta lång tid, och resultatet i figur 6a bör inte påverkas märkbart av fler simuleringar. Trenden av ringar observeras ändå följa den analytiska formeln, och det största träd som simuleringarna gav hade den ungefärliga storleken 300 miljarder. Detta kan jämföras med lemma 2.9, som förutspår att den största trädstorleken bland 200 000 simuleringar blir 40 miljarder.

4.2.2 Undersökning av tunga svansar i det subkritiska fallet

I det subkritiska fallet kan en teoretisk potenslag för överlevnadsfunktion av W inte bestämmas, så för att avgöra om Galton-Watson-processen är tungsvansad i det subkritiska fallet behövs data från simuleringar. Figur 7 visar den simulerade överlevnadsfunktion av W för olika värden på μ, med logaritmiska axlar. Graferna är ordnade så att de går från vänster till höger med ökande μ. I samtliga fall gjordes 200 000 simuleringar. Notera att de maximala träden blir mycket mindre i det subkritiska fallet, med så lite som $W_{\max} \approx 60$ för $\mu = 0.6$. Ökningen i storleksordning går snabbt med $W_{\max} = \mathcal{O}(10^3)$ för $\mu = 0.9$, $W_{\max} = \mathcal{O}(10^5)$ för $\mu = 0.99$ och $W_{\max} = \mathcal{O}(10^7)$ för $\mu = 0.999$. Men även nära det kritiska fallet, med $\mu = 0.999$ är beteendet långt ifrån regeln $W_{\max}(N) = \mathcal{O}(N^2)$ från lemma 2.9 som gäller i det kritiska fallet. Det här tyder på att svansarna ej är potensfördelade i det subkritiska fallet.

Överlevnadsfunktionen verkar inte följa en potenslag i något av de subkritiska fallen, inte ens för $\mu = 0.999$, eftersom kurvorna inte påvisar linearitet, vilket en potenslag ska göra när båda axlarna är logaritmiska. Faktumet att kurvorna buktar sig neråt antyder att överlevnadsfunktionen av W i det subkritiska fallet avtar snabbare än en potenslag, men detta behöver undersökas vidare med andra grafiska metoder. Avvikelsen som kan ses i fallet $\mu = 1$ efter $n = 10^9$ beror sannolikt på att överlevnadsfunktionen approximeras dåligt för stora n, se avsnitt 3.2.

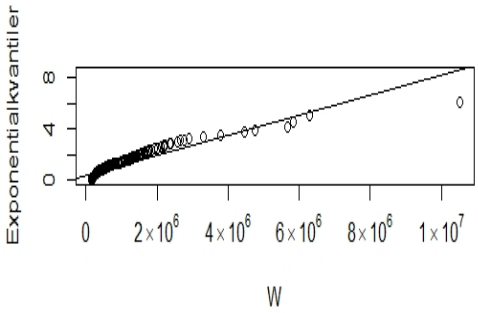


Figur 7: Grafer över den simulerade överlevnadsfunktion av W för olika μ, med logaritmiska axlar. Värdet av μ ökar när graferna går från vänster till höger. Endast fallet $\mu = 1$ verkar följa en linjär funktion, dock med en avvikelse efter $n = 10^9$.

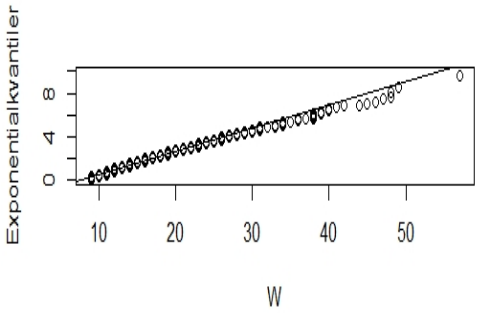
Även om svansarna i det subkritiska fallet inte följer en potenslag kan de fortfarande vara bortom normalfördelningen, då det finns andra fall än potenslagar som ger upphov till tungsvansade fördelningar. Det är således även intressant att se när det subkritiska fallet inte längre är bortom normalfördelningen, eftersom många vanligt förekommande modeller fungerar under den gränsen. I figur 8a kan resultatet av datans svans (bortom tre standardavvikelser) jämfört med exponentialfördelningen ses för $\mu = 0.999$. Att datapunkterna i slutet avviker nedåt från den räta linjen tyder på att svansarna i det fallet är tyngre än exponentialfördelningen, men det är svårt att avgöra eftersom olika tolkningar av kurvans form kan ge olika slutsatser, se avsnitt 3.3. Ytterligare resultat som styrker fördelningen av svansarna för fallet $\mu = 0.999$ finns i appendix C.2. Kring ungefär $\mu = 0.6$ börjar dock datapunkterna följa den räta linjen, se figur 8b, vilket tyder på att det inte längre är bortom normalfördelningen, eftersom fördelningen då ej längre är tungsvansad. Resultat som styrker att fördelningen är lättsvansad för $\mu = 0.6$ finns i appendix C.2.

5 Exempel med verkliga data

Här presenteras exempel på Galton-Watson-processen med verkliga data från externa källor, som en LF-simulering för antal födda i Sverige med statistik från SCB. Vidare presenteras exempel på



(a) En qq-plot i det subkritiska fallet med $\mu = 0.999$. Datapunkterna avviker nedåt från den räta linjen, vilket tyder på tyngre svansar än exponentialfördelningen.



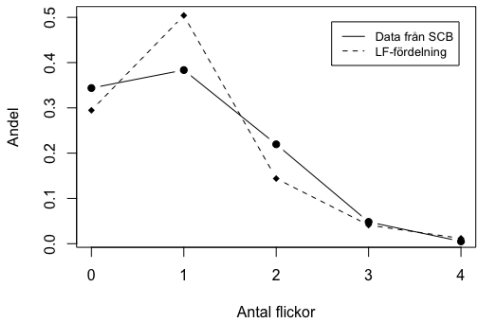
(b) En qq-plot i det subkritiska fallet med $\mu = 0.6$. Datapunkterna följer den räta linjen, vilket tyder på att gränsen till tungsvansade fördelningar går här.

Figur 8: Jämförelse mellan exponentialfördelade svansar och totala antalet individer W i träd som genererats med konturprocessen för olika μ.

verklig potensfördelad data som kan modelleras med en Galton-Watson-process, såsom de mest streamade låtarna på Spotify och de mest citerade forskarna på Chalmers.

5.1 LF-simulering för antal födda i Sverige

Galton-Watson-processens ordinarie syfte var som tidigare nämnt att undersöka hur stor sannolikheten var att olika aristokratnamn skulle leva kvar. En mer modern tagning på detta är att använda modellen för att undersöka om en godtycklig individs gener lever vidare för alltid.



Figur 9: Diskret sannolikhetsfördelning för hur många döttrar en kvinna i Sverige väntas få under sin livstid jämfört med en LF-fördelning med parametrar $p_0 = 0.29$ och $p = 0.71$.

I det här exemplet observeras en genomsnittlig svensk kvinna. Först behövs en uppskattning av hur stor sannolikheten är att en godtycklig kvinna får en dotter. Data från statistiska centralbyrån över hur många barn en genomsnittlig kvinna i Sverige föder innan de fyllt 45 år används, under antagandet att andelen kvinnor som föder barn efter de fyllt 45 är försumbart [16]. Ytterligare antas det vara lika sannolikt att få en son som det är att få en dotter, och därmed kan en uppskattning sammanställas för hur många döttrar en svensk kvinna förväntas föda under en livstid. Anledningen att endast döttrar räknas beror på att den Galton-Watson-process som betraktas i detta arbete är enkönad. Det skall också noteras att mödrar som födde fyra eller fler barn har sammanslagits till samma datapunkt. Sannolikheten att det skulle bli fyra döttrar är dock mindre än 1 %, och sannolikheten att föda fler döttrar än så försummas därför.

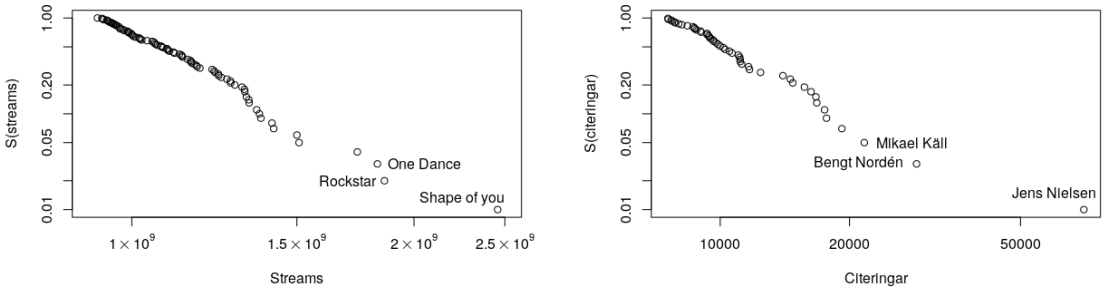
I figur 9 visas att datans fördelning liknar en LF-fördelning, ty även om de inte överensstämmer perfekt så är trenden liknande. LF-fördelningen anpassas att ha ungefär samma väntevärde som den verkliga datan med hjälp av momentmetoden (eng. *method of moments*), se appendix C.3.

Då erhålls parametrarna $p_0 = 0.29$ och $p = 0.71$. Dessa ger ett väntevärde på ungefär $\mu = 0.9871$ för LF-fördelningen, samma som för datan. Skulle dessa parametrar användas för att initiera en Galton-Watson-process landar den i det subkritiska fallet, men med tanke på $\mu = 0.9871$ är den dock bortom normalfördelningen, se avsnitt 4.2.2. Därmed kan slutsatsen dras att trädet för en godtycklig svensk kvinnas gener, enligt Galton-Watson-modellen, slutligen kommer att dö ut.

5.2 Potensfördelade exempel

I inledningen nämndes det att världens rikaste människor enligt Forbes-listan är fördelade enligt en potenslag [2]. Detta kan tänkas modelleras genom en Galton-Watson-process, där generationerna är tidssteg och barnen tolkas som ekonomiska tillgångar. Här spelar alltså inte totala antalet individer roll på samma sätt, utan det är bara antal tillgångar i generationen som spelar roll. I det tvåpunktsfördelade fallet som beskrivs i avsnitt 2.2.1 antas att tillgångarna fördubblas med sannolikhet p och att tillgångarna förloras med sannolikhet $1 - p$. I det kritiska fallet blir det 50 % sannolikhet för fördubblade tillgångar. Det är då tydligt att givet en viss mängd tillgångar kommer processen överleva i framtiden, eftersom det är osannolikt att $Z_{n+1} = 0$ givet $Z_n \gg 0$.

Vidare kan antalet streams av en låt modelleras med en Galton-Watson-process. Då antas först att en lyssnare finns, och att denna sedan har möjligheten att antingen fortsätta lyssna på låten och sprida den vidare till andra, eller sluta lyssna på den och inte sprida den vidare. I fallet då individen slutar att lyssna på låten blir trädet utdött, under ett förenklat antagande att inga fler lyssnare någonsin kommer att hitta låten. Om individen däremot fortsätter att lyssna och delar låten, så har de nya individerna som börjar lyssna på låten sedan också möjligheten till att antingen sluta lyssna eller sprida låten vidare, och de antas för enkelhetens skull göra detta beslut med samma fördelning. Denna enkla modell kan ge upphov till rich-get-richer-effekten, och därmed tunga svansar. Figur 10a visar de mest streamade låtarna på Spotify fram till mars 2020. Notera att figurens axlar är logaritmerade, samt att de mest streamade låtarna verkar följa ett linjärt beroende i grafen. Detta innebär att datan egentligen beter sig approximativt som en avtagande potenslag med en tung svans.



(a) De mest spelade låtarna på Spotify mars 2020 presenterade i en log-log-plot [17].

(b) De mest citerade forskarna på Chalmers presenterade i en log-log-plot [18].

Figur 10: Log-log-plots över mest spelade låtar och mest citerade forskare.

För antalet citeringar på Chalmers antas en liknande förgreningsmodell som för antalet streams på Spotify. Ju mer välciterad en forskare är desto större är chansen att den fortsätter bli citerad. I figur 10b observeras de mest citerade forskarna följa ett linjärt beroende i log-log-grafen. Detta innebär att datan approximativt fördelas som en avtagande potenslag med tung svans.

Notera att exemplet med Forbes-listan samt exemplen i figur 10 beskriver de mest extrema värdena, varmed de karaktäriserar svansarna av fördelningen av tillgångarna, låtarna respektive forskarna. Det linjära beroendet i figur 10 är inte helt tydligt, då de största värdena verkar avvika från resten. Detta känns igen från jämförelsen med överlevnadsfunktionen i avsnitt 4.2.1 och beror på att de mest extrema värdena kan ge avvikelser från det generella beteendet enligt avsnitt 3.2.

6 Avslutning

Syftet med detta arbete var att undersöka fördelningar bortom normalfördelningen och då i synnerhet huruvida Galton-Watson-processer ger upphov till fördelningar med tunga svansar i det kritiska och subkritiska fallet. Vi har analytiskt visat att fördelningen av totala antal individer i en kritisk Galton-Watson-process asymptotiskt följer en potenslag. Detta innebär att den kritiska Galton-Watson-processen ger upphov till tunga svansar, vilket även indikeras av simuleringarna. Vi har med hjälp av simuleringarna också konstaterat att fördelningen av totala antalet individer i det subkritiska fallet inte följer en potenslag för något μ, men att den har tunga svansar som ligger mellan potenslagar och exponentiellt avtagande svansar ner till ungefär $\mu = 0.6$.

Genom undersökningen av tunga svansar har en inblick i grafiska metoder som kan användas bortom normalfördelningen getts. Dessa metoder kan vara en användbar kontroll vid behandling av insamlad data, för att avgöra om något är bortom normalfördelningen eller inte, och således om till exempel centrala gränsvärdessatsen kan användas. Det är från detta arbete tydligt att även vid samma bakomliggande process så kan olika värden på parametrarna ge upphov till vitt skilda fördelningar, och därmed bör försiktighet iakttas innan normalfördelning antas för insamlad data.

Arbetet når också ut till en större publik genom bidrag till Wikipedia om Galton-Watson-processer, sannolikhetsgenererande funktioner och LF-fördelningar på svenska, som finns bifogade i appendix F. Inom det matematiska området som behandlas i detta arbete finns få Wikipedia-artiklar på svenska och för att minska domänförlust behövs fler svenska artiklar inom området.

Som tidigare nämnts kan Forbes-data över de allra rikaste fördelas som en potenslag och även modelleras med hjälp av Galton-Watson-processen, se avsnitt 5. Ofta är rikedom asymmetriskt fördelad vilket ger upphov till klyftor mellan fattiga och rika. Utöver detta leder de flesta fall där popularitet och nätverk är inblandade till rich-get-richer-fenomen, exempelvis genom att människor lyssnar på topplistorna och citerar välciterade artiklar [1]. Som visas i avsnitt 5 fördelas mest streamade låtar på Spotify och mest citerade författare på Chalmers enligt en potenslag och kan modelleras med Galton-Watson-processen. Detta arbete påvisar därmed kopplingen mellan rich-get-richer-fenomen i verkligheten och Galton-Watson-processen, och ger matematisk insikt till varför rich-get-richer-fenomen kan uppstå. Det är dock viktigt att betona att kopplingen inte är självklar, exempelvis är Galton-Watson-modellen av mest streamade låtar på Spotify förenklad. Verkligheten är istället ett komplext samspel av många fler faktorer. Till exempel spekuleras det om huruvida företag med riktad reklam och rekommendationssystem förstärker rich-get-richer-effekten, eller tvärtemot motverkar den [1]. Att diskutera huruvida rich-get-richer-fenomen bör förstärkas eller motverkas är dock utanför ramen för arbetet.

En intressant aspekt av att fördelningen är lättsvansad i det subkritiska fallet för låga μ är att verklighetens rich-get-richer-fenomen kanske kan motverkas på detta sätt. Om till exempel reproduktionsfördelningen för att dela med sig av en låt hade haft väntevärde runt 0.5 så hade vi kanske inte upplevt rich-get-richer-effekten. Eller, som i originalexemplet med efternamn, så skulle aristokratefternamnen dö ut. Arbetet har analytiskt främst fokuserat på det kritiska fallet, så att bestämma en analytisk formel i det subkritiska fallet eller undersöka en tydligare analytisk gräns mellan tungsvansat och icke-tungsvansat hade varit intressant. Här kan kanske de analytiska fördelningarna för Z_n i det kritiska och subkritiska fallet i lemmorna 2.6 och 2.7 komma till användning. Även simuleringar och grafiska metoder kan utökas, med tanke på svårigheterna att tolka de grafiska metoderna som beskrivs i avsnitt 3.3, samt avvikelser som uppstår när de mest extrema värdena betraktas som beskrivs i avsnitt 3.2. Med mer data och fler grafiska metoder kan det undersökas vilken fördelning som passar för olika värden på μ, till exempel exponentialfördelning, normalfördelning, eller någon helt annan fördelning.

En annan aspekt som skulle påverka rich-get-richer-effekten är utökningen av Galton-Watson-processen till en storleksberoende process [19]. Att försöka koppla storleksberoende Galton-Watson-processer i det kritiska fallet till tungsvansade fördelningar är intressant eftersom det visats att antal döda i globala pandemier kan följa en tungsvansad fördelning [20]. Storleksberoende processer skulle kunna modellera spridning av pandemier och virus, genom att att ta hänsyn till antal smittade i reproduktionsfördelningen. Eftersom det är omöjligt att smitta redan smittade individer är detta relevant att ta hänsyn till. För den intresserade läsaren finns en enkel modell av COVID-19 i appendix D.2. Denna modell är en approximation av storleksberoende process, så att modellera virusspridning med en mer teoretiskt korrekt storleksberoende process skulle vara intressant.

Referenser

[1] Easley D, Kleinberg J, et al. Power Laws and Rich-Get-Richer Phenomena. In: Networks, crowds, and markets. vol. 8. Cambridge university press; 2010. p. 543–560.

[2] Klass OS, Biham O, Levy M, Malcai O, Solomon S. The Forbes 400 and the Pareto wealth distribution. Economics Letters. 2006 feb;90(2):290–295.

[3] Grimmett G, Stirzaker D. Probability and Random Processes. 3rd ed. Oxford University Press; 2001.

[4] Dobrow RP. Introduction to Stochastic Processes with R. John Wiley & Sons; 2016.

[5] Lipovetsky S. Pareto 80/20 law: derivation via random partitioning. International Journal of Mathematical Education in Science and Technology. 2009 mar;40(2):271–277.

[6] Athreya KB, Ney PE. The Galton-Watson Process. In: Branching Processes. Springer Berlin Heidelberg; 1972. p. 1–65.

[7] Olofsson P, Andersson M. Probability, Statistics, and Stochastic Processes. 2nd ed. John Wiley & Sons; 2012.

[8] Arnold BC. Pareto distributions. Chapman and Hall/CRC; 2015.

[9] Ishikawa A. Pareto index induced from the scale of companies. Physica A: Statistical Mechanics and its Applications. 2006 may;363(2):367–376.

[10] Foss S, Korshunov D, Zachary S, et al. An introduction to heavy-tailed and subexponential distributions. vol. 6. Springer; 2011.

[11] Sagitov S, Minuesa C. Defective Galton-Watson processes. Stochastic Models. 2017;33(3):451–472.

[12] Stirzaker D. Stochastic Processes & Models. Oxford University Press; 2005.

[13] Bruijn NGD. Asymptotic Methods in Analysis. DOVER PUBN INC; 1981.

[14] Feller W. An Introduction to Probability Theory and Its Applications. vol. 2. 2nd ed. John Wiley & Sons; 1971.

[15] Cirillo P. Are your data really Pareto distributed? Physica A: Statistical Mechanics and its Applications. 2013 Dec;392(23):5947–5962.

[16] Statistiska centralbyrån. Olika generationers barnafödande. Statistiska centralbyrån; 2011. Available from: `http://share.scb.se/ov9993/data/publikationer/statistik/_publikationer/be0701_2011a01_br_be51br1103.pdf`.

[17] Wikipedia contributors. List of most-streamed songs on Spotify. Wikipedia, The Free Encyclopedia; 2020. [Online; accessed 7-April-2020]. Available from: `https://en.wikipedia.org/w/index.php?title=List_of_most-streamed_songs_on_Spotify&oldid=949563511`.

[18] Google Scholar. Profile for Chalmers University of Technology. Google; 2020. [Online; accessed 7-April-2020]. Available from: `https://scholar.google.com/citations?view_op=view_org&org=4746383261807430232&hl=en&oi=io`.

[19] Klebaner FC. On population-size-dependent branching processes. Advances in Applied Probability. 1984;16(1):30–55.

[20] Cirillo P, Taleb NN. Tail Risk of Contagious Diseases; 2020.

[21] Devroye L. Non-Uniform Random Variate Generation. Springer-Verlag; 1986.

[22] Kratz M, Resnick SI. The qq-estimator and heavy tails. Communications in Statistics Stochastic Models. 1996;12(4):699–724.

[23] Prokhorov AV. Moments, method of (in probability theory). Encyclopedia of Mathematics; 2011. Available from: `http://www.encyclopediaofmath.org/index.php?title=Moments,_method_of_(in_probability_theory)&oldid=14059`.

Appendix

A Grundläggande sannolikhetsteoretisk bakgrund

I detta avsnitt i appendix presenterar vi definitioner för en del begrepp som används i rapporten och även senare i bevisen, samt för begrepp som dessa definitioner bygger på. Definitionerna är sorterade så att de kan läsas i ordning. Om inget annat anges följs i presentationen främst [3].

Tabell 2: Sammanfattande tabell över olika fördelningar.

Fördelning	Parametrar	Täthetsfunktion $f(x)$/ Sannolikhetsfunktion $f(k)$	Väntevärde $\mathbb{E}[X]$	Varians $\mathrm{Var}(X)$
Likformig	$-\infty < a < b < \infty$	$\begin{cases} \frac{1}{b-a} & \text{om } x \in [a,b] \\ 0 & \text{annars} \end{cases}$	$\frac{(a+b)}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\mu \in \mathbb{R}$, $\sigma^2 > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$	μ	σ^2
Pareto	$x_m > 0$, $\alpha > 0$	$\begin{cases} \frac{\alpha x_m^\alpha}{x^{1+\alpha}} & \text{om } x \geq x_m \\ 0 & \text{om } x < x_m \end{cases}$	$\begin{cases} \frac{\alpha x_m}{\alpha-1} & \text{om } \alpha > 1 \\ \infty & \text{om } \alpha \leq 1 \end{cases}$	$\begin{cases} \left(\frac{x_m}{\alpha-1}\right)^2 \frac{\alpha}{\alpha-2} & \text{om } \alpha > 2 \\ \infty & \text{om } \alpha \leq 2 \end{cases}$
Exponential	$\lambda > 0$	$\lambda e^{-\lambda x}, \quad x \geq 0$	$1/\lambda$	$1/\lambda^2$
Tvåpunkts	$p \in [0,1]$	$\begin{cases} 1-p & \text{om } k = x_1 \\ p & \text{om } k = x_2 \end{cases}$	$(1-p)\,x_1 + p\,x_2$	$(p-p^2)(x_1^2+x_2^2) - 2(1-p)px_1x_2$
Geometrisk	$p \in (0,1]$	$p(1-p)^k, \quad k = 0,1,2,\ldots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Skiftad geom.	$p \in (0,1)$	$p(1-p)^{k-1}, \quad k = 1,2,3,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

Definition A.1 (Utfallsrum)**.** Mängden Ω av alla möjliga resultat av ett experiment kallas för utfallsrum.

Ett resultat av ett experiment kallas också för ett utfall eller en händelse.

Definition A.2 (σ-algebra)**.** Om $\mathcal{F}$ är en samling av delmängder till Ω så kallas $\mathcal{F}$ för en σ-algebra om den uppfyller:

(i) $\varnothing \in \mathcal{F}$,

(ii) om $A_1, A_2, ... \in \mathcal{F}$ så gäller att $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$,

(iii) om $A \in \mathcal{F}$ så gäller att $A^c \in \mathcal{F}$.

Ofta betraktas ett par $(\Omega, \mathcal{F})$, där Ω alltså är mängden av alla möjliga händelser och $\mathcal{F}$ är en σ-algebra av delmängder till Ω, innehållande alla händelser vars förekomst vi är intresserade av.

Definition A.3 (Sannolikhetsmått)**.** Ett sannolikhetsmått P på $(\Omega, \mathcal{F})$ är en avbildning $P : \mathcal{F} \to [0,1]$ som uppfyller:

(i) $P(\varnothing) = 0$ och $P(\Omega)$=1,

(ii) om $A_1, A_2, ...$ är en samling av disjunkta element i $\mathcal{F}$, det vill säga att $A_i \cap A_j = \varnothing$ för alla par i, j som uppfyller att $i \neq j$, så gäller att

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Definition A.4 (Sannolikhetsrum)**.** En trippel $(\Omega, \mathcal{F}, P)$, bestående av ett utfallsrum Ω, en σ-algebra $\mathcal{F}$ av delmängder till Ω och ett sannolikhetsmått P på $(\Omega, \mathcal{F})$ kallas för ett sannolikhetsrum.

Definition A.5 (Betingad sannolikhet)**.** Låt $B \in \Omega$ så att $P(B) > 0$. Den betingade sannolikheten för att $A \in \Omega$ ska inträffa givet att B inträffat skrivs $P(A \mid B)$ och definieras som

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

där $P(A \cap B)$ betecknar sannolikheten för snittet av utfallen.

Definition A.6 (Oberoende)**.** Händelserna $A, B \in \Omega$ är oberoende om $P(A \cap B) = P(A)P(B)$.

Definition A.7 (Slumpvariabel)**.** En slumpvariabel X är en avbildning $X : \Omega \to E$ med egenskapen att $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ för varje $x \in E$. En sådan funktion sägs vara $\mathcal{F}$-mätbar.

En slumpvariabel kallas också för en stokastisk variabel. Ofta är mängden $E = \mathbb{R}$, vilket vi utgår från i denna rapport, om inget annat anges.

Definition A.8 (Fördelningsfunktion)**.** Fördelningsfunktionen av en slumpvariabel X är en funktion $F : \mathbb{R} \to [0, 1]$, som ges av $F(x) = P(X \leq x)$.

Definition A.9 (Diskret slumpvariabel och sannolikhetsfunktion)**.** En diskret slumpvariabel X tar enbart värden i en uppräkneligt ändlig delmängd $\{x_1, x_2, ...\}$ av $\mathbb{R}$. Dess sannolikhetsfunktion $f : \mathbb{R} \to [0, 1]$ ges av $f(x_j) = P(X = x_j)$, $j \in 1, 2, ...$

Definition A.10 (Kontinuerlig slumpvariabel och täthetsfunktion)**.** En slumpvariabel X, med fördelningsfunktionen F, kallas kontinuerlig om det existerar en funktion $f : \mathbb{R} \to [0, \infty)$ så att

$$F(x) = \int_{-\infty}^{x} f(s)ds, \quad x \in \mathbb{R}.$$

Funktionen f kallas för slumpvariabelns täthetsfunktion.

Definition A.11 (Väntevärde)**.** Om X är en diskret slumpvariabel med sannolikhetsfunktionen f ges dess väntevärde av

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} x_j f(x_j).$$

Om $g : \mathbb{R} \to \mathbb{R}$ ges väntevärdet av $g(X)$ av

$$\mathbb{E}[g(X)] = \sum_{j=1}^{\infty} g(x_j) f(x_j).$$

Om X är en kontinuerlig slumpvariabel med täthetsfunktionen f ges dess väntevärde av

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} s f(s)ds.$$

Om $g : \mathbb{R} \to \mathbb{R}$ ges väntevärdet av $g(X)$ av

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(s) f(s)ds.$$

Definition A.12 (Varians)**.** Variansen av en slumpvariabel X ges av $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Nu kan den likformiga fördelningen definieras [7].

Definition A.13 (Likformig fördelning)**.** En kontinuerlig slumpvariabel X är likformigt fördelad på $[a, b]$ med $a, b \in \mathbb{R}, a \leq b$ om dess täthetsfunktion är given av

$$f(x) = \frac{1}{b - a}, a \leq x \leq b.$$

Detta skrivs $X \overset{D}{\sim} \mathrm{likf}[a, b]$.

Ett vanligt exempel är likf[0, 1], och detta brukar vara standardfördelningen som slumptal dras från inom programmering.

Nu definieras några fler generella funktioner [3].

Definition A.14 (Karakteristisk funktion)**.** Avbildningen $\phi_X : \mathbb{R} \to \mathbb{C}$ kallas för den karakteristiska funktionen till slumpvariablen X och definieras av $\phi_X(t) = \mathbb{E}[e^{itX}]$, där $i^2 = -1$.

Den karakteristiska funktionen till normalfördelningen är $\phi(t) = \exp(i\mu t - \sigma^2 t^2/2)$.

Definition A.15 (Regelbundet varierande funktion)**.** En positiv funktion L, definierad på $[0, \infty)$, är regelbundet varierande om

$$\lim_{t\to\infty} \frac{L(ct)}{L(t)} = \phi(c)$$

för alla $c > 0$ och någon funktion $0 < \phi(c) < \infty$. [14]

Låt oss nu definiera Markovkedja och begreppet tidshomogen [12].

Definition A.16 (Markovkedja)**.** Låt $\{X_0, X_1, \ldots\}$ vara en följd av slumpvariabler med en uppräknelig mängd S som tillståndsrum. Då är följden en Markovkedja om

$$P(X_n = k | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = j) = P(X_n = k | X_{n-1} = j) \tag{A.1}$$

håller för alla $n \geq 1$ och $x_0, x_1, \ldots, j, k \in S$.

Egenskapen (A.1) innebär att varje värde i kedjan endast beror på det föregående värdet i kedjan. Vidare kommer vi främst betrakta tidshomogena Markovkedjor där sannolikheterna i (A.1) ej beror på n.

Definition A.17 (Tidshomogen)**.** En Markovkedja $\{X_0, X_1, \ldots\}$ med tillståndsrummet S är tidshomogen om

$$P(X_n = k | X_{n-1} = j) = P(X_1 = k | X_0 = j),$$

för alla $n \geq 1$ och $k, j \in S$.

B Bevis

För den intresserade läsare följer här bevis till satser och lemman i rapporten, samt långa uträkningar. De presenteras i kronologisk ordning.

B.1 Centrala gränsvärdessatsen

Lemma B.1. *(i) Om $\phi^{(k)}(0)$ existerar så är $\mathbb{E}\left[|X^k|\right] < \infty$ för jämna k och $\mathbb{E}\left[|X^{k-1}|\right] < \infty$ för udda k.*

(ii) Om $\mathbb{E}\left[|X^k|\right] < \infty$ så gäller att

$$\phi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}[X^j]}{j!}(it)^j + o(t^k)$$

och således är $\phi^{(k)}(0) = i^k \mathbb{E}[X^k]$.

Lemma B.2. *Antag att F_1, F_2, F_3, ... är en sekvens av fördelningsfunktioner med respektive karakteristisk funktion ϕ_1, ϕ_2, ϕ_3, ...*

(i) Om $F_n \to F$ för någon fördelningsfunktion F med karakteristisk funktion ϕ så gäller att $\phi_n(t) \to \phi(t)$, för alla t.

(ii) Omvänt, om gränsvärdet $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ existerar och är kontinuerlig vid $t = 0$ så gäller att ϕ är den karakteristiska funktionen till någon fördelningsfunktion F samt att $F_n \to F$.

Bevis av sats 2.1. Börja med att skriva $Y_i = (X_i - \mu)/\sigma$ och kalla den karakteristiska funktionen till Y_i för ϕ_Y. Observera att vi nu kan skriva

$$U_n = \frac{\mathbb{S}_n - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Kalla den karakteristiska funktionen till U_n för ψ_n. Vi har från definitionen av karakteristisk funktion att

$$\psi_n(t) = \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{n}} \sum_{i=1}^{n} Y_i\right)\right]$$

och eftersom alla X_i och därmed alla Y_i är oberoende gäller att

$$\psi_n(t) = \mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\frac{it}{\sqrt{n}} Y_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(i\frac{t}{\sqrt{n}} Y_i\right)\right] = \prod_{i=1}^{n} \phi_Y\left(\frac{t}{\sqrt{n}}\right).$$

Eftersom alla Y_i har samma fördelning och således samma karakteristiska funktion kan detta även skrivas som $\psi_n(t) = (\phi_Y(t/\sqrt{n}))^n$. Enligt Lemma B.1 har vi att $\phi_Y = 1 - t^2/2 + o(t^2)$ vilket ger

$$\psi_n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n = \exp\left[n \ln\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)\right] \to \exp(-t^2/2) \quad \text{då} \quad n \to \infty.$$

Den sista funktionen är den karakteristiska funktionen till $\mathcal{N}(0,1)$ och genom att applicera lemma B.2 följer satsen. □

B.2 Beteende av G_n för stora n och utrotningssannolikhet

Bevis av lemma 2.1. Basfallet är för $n = 1$. Vi får

$$G_1(s) = \mathbb{E}[s^{Z_1} | Z_0 = 1] = \mathbb{E}[s^X] = G(s).$$

Vi antar nu att $G_k(s) = \underbrace{G(G...G(s)...)}_{\text{k gånger}}$, och ska visa att $G_{k+1}(s) = \underbrace{G(G...G(s)...)}_{\text{k+1 gånger}}$. Vi har

$$G_{k+1}(s) = \mathbb{E}[s^{Z_{k+1}} | Z_0 = 1] = \mathbb{E}[\mathbb{E}[s^{Z_{k+1}} | Z_k] | Z_0 = 1].$$

Vidare har vi $Z_{k+1} = \sum_{i=1}^{Z_k} X_i$, alltså

$$s^{Z_{k+1}} = s^{\sum_{i=1}^{Z_k} X_i} = \prod_{i=1}^{Z_k} s^{X_i}.$$

Men då väntevärdet av en produkt av oberoende variabler är produkten av dess väntevärden får vi

$$\mathbb{E}[s^{Z_{k+1}} | Z_k] = \prod_{i=1}^{Z_k} \mathbb{E}[s^{X_i}],$$

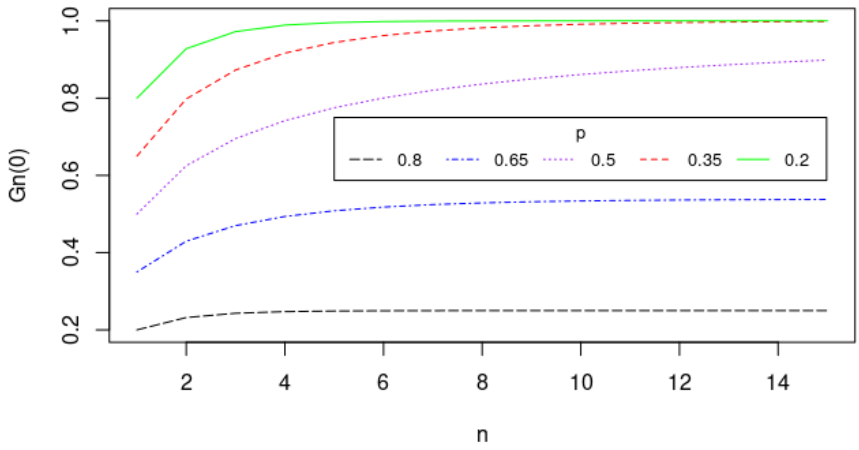
vilket ger

$$\mathbb{E}\left[\prod_{i=1}^{Z_k} \mathbb{E}[s^{X_i}] \middle| Z_0 = 1\right] = \mathbb{E}\left[\prod_{i=1}^{Z_k} G(s)\right] = \mathbb{E}[(G(s))^{Z_k}] \underbrace{=}_{\mathbb{E}[t^{Z_k}] = G_k(t)} G_k(G(s)) = \underbrace{G(G...G(s)...)}_{\text{k+1 gånger}}.$$

Alltså är $G_n(s) = \underbrace{G(G...G(s)...)}_{\text{n gånger}}$ för alla positiva n. □

Bevis av sats 2.2. Först visar vi att funktionen är monoton. Tag s_1 och s_2 där $s_1 \leq s_2$. Då har vi

$$G(s_1) = \sum_{k=0}^{\infty} p_k s_1^k \leq \sum_{k=0}^{\infty} p_k s_2^k = G(s_2)$$



Figur 11: Här illustreras $G_n(0)$ för varierande $n \in [0, 15]$ och $p = 0.2, 0.35, 0.5, 0.65, 0.8$ i det tvåpunktsfördelade fallet.

eftersom $p_k \in [0, 1]$ för alla k. Alltså är funktionen G växande, d.v.s. monoton. Vidare har vi $G_n(s) = G_n(s)$, men då G är monoton så är även G upprepat n gånger en monoton funktion (kan bevisas enkelt genom induktion). Därmed är $G_n(s_1) \leq G_n(s_2)$ och G_n är monoton i intervallet $[0, 1]$.

Vi introducerar nu utrotningssannolikheten för en given generation i som $q_i = P(Z_i = 0)$. Vi har då trivialt $q_0 = 0$ då vi börjar med en individ i generation 0. Vidare har vi en given utrotningssannolikhet i första generationen med $q_1 \geq q_0$, eftersom det nu kan finnas en risk för utrotning. Generellt har vi rekursionsformeln $q_{k+1} = q_k \cdot 1 + (1 - q_k)a = q_k + \Delta q$, där $\Delta q \geq 0$. Här innebär den första termen att om processen var utrotad i generation k kommer den även vara det i generation $k + 1$ med sannolikhet 1. Den andra termen innebär att om processen ej var utrotad finns det en ickenegativ sannolikhet a att den utrotas i generation $k + 1$. Därmed har vi en följd

$$0 = q_0 \leq q_1 \leq q_2 \leq \cdots \leq 1,$$

och eftersom följden är monoton och begränsad konvergerar den till ett värde $q \leq 1$.

I figur 11 illustreras följden av $G_n(0) = q_n$ för ökande n och olika p i det tvåpunktsfördelade fallet. Vi kan se att följden verkar konvergera mot något $q < 1$ för $p < 0.5$ och $q = 1$ för $p > 0.5$. Beteendet av $p = 0.5$ har ännu inte stabiliserat sig då det är i det kritiska fallet, men kommer konvergera mot $q = 1$.

Vi ska nu visa att följden $G_n(s)$ konvergerar för givet $s \in [0, 1]$. Vi vet från ovan att

$$G_n(0) = P(Z_n = 0) = q_n \to q$$

då $n \to \infty$. Vi vet också att $G_n(1) = 1$ för alla n. Alltså konvergerar G_n för $s = 0$ och $s = 1$, och $G_n(s)$ är monoton i intervallet $[0, 1]$. Tag nu $0 < s_0 < 1$. Då är följden $G_n(s_0)$ nedåt begränsad av $G_n(0)$ och uppåt begränsad av $G_n(1)$. Eftersom $G_n(0)$ och $G_n(1)$ har ändliga gränsvärden måste $G_n(s_0)$ också ha det. Alltså konvergerar $G_n(s)$ för $s \in [0, 1]$. □

Bevis av Sats 2.3. Eftersom $G_{n+1}(0) = G(G_n(0))$ och $G_n(0) \to q$ då $n \to \infty$, har vi att q kan finnas genom ekvationen $q = G(q)$. Låt $\hat{s}$ vara en positiv lösning av $s = G(s)$. Vi vill visa att $q \leq \hat{s}$. Eftersom $G_n(s)$ är en monotont växande funktion för $s \in [0, 1]$ så är

$$q_n = G_n(0) \leq G_n(\hat{s}) = \hat{s}$$

och när $n \to \infty$ har vi $q \leq \hat{s}$. □

B.3 Varians för Galton-Watson-process och totala antalet individer

Bevis av lemma 2.3. Genom lagen av total varians är

$$\sigma_{n+1}^2 = \mathrm{Var}[Z_{n+1}] = \mathrm{Var}[\mathbb{E}[Z_{n+1}|Z_n]] + \mathbb{E}[\mathrm{Var}[Z_{n+1}|Z_n]].$$

Eftersom $\mathbb{E}[Z_n] = \mu^n$ är $\mathbb{E}[Z_{n+1}|Z_n] = \mu Z_n$, och eftersom $\mathrm{Var}[Z_{n+1}|Z_n] = \sigma^2 Z_n$ får vi att

$$\sigma_{n+1}^2 = \mathrm{Var}[\mu Z_n] + \mathbb{E}[\sigma^2 Z_n] = \mu^2 \mathrm{Var}[Z_n] + \sigma^2 \mathbb{E}[Z_n],$$

vilket ger differensekvationen

$$\sigma_{n+1}^2 = \mu^2 \sigma_n^2 + \mu^n \sigma^2,\ \sigma_0^2 = 0.$$

Lösningen till en allmän linjär differensekvation av första ordningen, $y_{n+1} = ay_n + d_n,\ n \in \mathbb{N}_0$, kan skrivas

$$y_n = a^n y_0 + \sum_{k=0}^{n-1} a^{n-1-k} d_k,\ n \geq 1,$$

och i vårt fall med $a = \mu^2$ och $d_n = \mu^n \sigma^2$ så blir lösningen

$$\sigma_n^2 = \sum_{k=0}^{n-1} (\mu^2)^{n-k-1} \mu^k \sigma^2 = \sigma^2 \mu^{n-1} \sum_{k=0}^{n-1} \mu^{n-k-1} = \begin{cases} n\sigma^2, & \mu = 1, \\ \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1}, & \mu \neq 1. \end{cases}$$

När $n \to \infty$ följer lemmat. □

Bevis av lemma 2.5. Vi skriver

$$W = 1 + \sum_{k=1}^{Z_1} w_k \overset{D}{\sim} 1 + Z_1 W,$$

där $w_1,\ w_2, \ldots,\ w_{Z_1} \overset{D}{\sim} W$. Här kan $w_k,\ k \in \{1, \ldots, Z_1\}$ ses som det totala antalet individer i oändligheten för nya, oberoende och likfördelade Galton-Watson-processer med en startindivid, ej att blandas ihop med W_k som var totala antalet individer i generation k. Den genererande funktionen för W uppfyller då

$$\begin{aligned} H(s) := \mathbb{E}[s^W] &= \mathbb{E}[\mathbb{E}[s^W|Z_1]] = \mathbb{E}[\mathbb{E}[s^{1+Z_1 W}|Z_1]] = s\mathbb{E}[\mathbb{E}[(s^W)^{Z_1}|Z_1]] \\ &= s\mathbb{E}[H(s)^{Z_1}] = sG(H(s)), \end{aligned}$$

där $G(s) = \mathbb{E}[s^{Z_1}|Z_0 = 1]$ är den genererande funktionen för processens reproduktionsfördelning. □

B.4 LF-fördelning

Bevis av sats 2.4. Den sannolikhetsgenererande funktionen till fördelning (2.6) är

$$G_X(s) = p_0 + \sum_{k=1}^{\infty} p(1-p_0)(1-p)^{k-1} s^k = p_0 + p(1-p_0)(1-p)^{-1} \sum_{k=1}^{\infty} [s(1-p)]^k.$$

Serien konvergerar endast om $|s(1-p)| < 1$ och har då summan $\frac{s(1-p)}{1-s(1-p)}$, vilket ger

$$G_X(s) = p_0 + p(1-p_0)(1-p)^{-1} \frac{s(1-p)}{1-s(1-p)} = p_0 + (1-p_0)\frac{ps}{1-(1-p)s} = G(s).$$

□

Bevis av sats 2.5. Väntevärdet fås genom egenskap (2.3c) som ger

$$\mu = G'(1) = \frac{(1-p_0)p}{(1-(1-p)s)^2}\bigg|_{s=1} = \frac{1-p_0}{p}. \tag{B.1}$$

Vi har

$$s_0 = \frac{p_0}{1-p} = \frac{p_0^{(n)}}{1-p^{(n)}}. \tag{B.2}$$

Från Galton-Watson-processen vet vi också att μ efter n generationer blir μ^n. Genom att använda uttrycket i (B.1) har vi en till ekvation,

$$\mu^n = \frac{1-p_0^{(n)}}{p^{(n)}} = \left(\frac{1-p_0}{p}\right)^n. \tag{B.3}$$

Från (B.2) och (B.3) får vi ekvationssystemet

$$\begin{cases} \dfrac{p_0}{1-p} = \dfrac{p_0^{(n)}}{1-p^{(n)}}, \\ \left(\dfrac{1-p_0}{p}\right)^n = \dfrac{1-p_0^{(n)}}{p^{(n)}}, \end{cases} \iff \begin{cases} p^{(n)} = \dfrac{1-\frac{p_0}{1-p}}{\left(\frac{1-p_0}{p}\right)^n - \frac{p_0}{1-p}}, \\ p_0^{(n)} = \dfrac{1-\left(\frac{p}{1-p_0}\right)^n}{\frac{1-p}{p_0} - \left(\frac{p}{1-p_0}\right)^n}. \end{cases}$$

I kritiska fallet kan vi inte använda ekvationerna (B.2) och (B.3) då de ger samma resultat. Vi har dubbelrot i $s = 1$ då både $s_0 = 1$ och $\mu = 1$, vilket ger $\frac{p_0}{1-p} = 1 \iff p_0 = 1-p$ och motsvarande $p_0^{(n)} = 1 - p^{(n)}$. Därmed måste vi få en till ekvation för att få förhållanden mellan p, p_0 och $p^{(n)}, p_0^{(n)}$. Vi har då från (2.3d) uttrycket för variansen

$$\sigma^2 = G''(1) + G'(1) - (G'(1))^2 = \frac{(1-p_0)(1-p+p_0)}{p^2}.$$

I kritiska fallet har vi $p_0 = 1-p$ så

$$\sigma^2 = \frac{(1-p_0)(1-p+p_0)}{p^2} = \frac{p \cdot 2(1-p)}{p^2} = \frac{2(1-p)}{p}. \tag{B.4}$$

Vidare har vi för Galton-Watson-processen också att variansen efter n generationer är $n\sigma^2$. Alltså får vi genom (B.4) att

$$\begin{cases} n\sigma^2 = \dfrac{2(1-p^{(n)})}{p^{(n)}}, \\ \sigma^2 = \dfrac{2(1-p)}{p}, \end{cases} \iff \frac{2(1-p^{(n)})}{p^{(n)}} = n\frac{2(1-p)}{p}.$$

Tillsammans med ekvationerna (B.2) och (B.3) får vi då slutligen

$$\begin{cases} p^{(n)} = \dfrac{p}{p+(1-p)n}, \\ p_0^{(n)} = \dfrac{(1-p)n}{p+(1-p)n}. \end{cases}$$

□

B.5 Fördelningen av Z_n i det kritiska och subkritiska fallet

Bevis av lemma 2.6. Eftersom vi vill visa att $Z_n/n | Z_n \geq 1$ är asymptotiskt exponentialfördelad introducerar vi $\sim$ för att beteckna asymptotisk ekvivalens (se definition 2.13). Sannolikheten för fler än nx individer givet att generationen ej är utdöd ges av

$$P(Z_n \geq nx \,|\, Z_n \geq 1) = \frac{P(Z_n \geq nx \cap Z_n \geq 1)}{P(Z_n \geq 1)} = \frac{P(Z_n \geq nx)}{P(Z_n \geq 1)}. \tag{B.5}$$

Vi utvecklar

$$P(Z_n \geq nx) = \sum_{k=nx}^{\infty} (1 - p_0^{(n)})(1 - p^{(n)})^{k-1} p^{(n)} \tag{B.6}$$

enligt Sats 2.4. Vi utnyttjar formeln för geometrisk summa och får

$$P(Z_n \geq nx) = (1 - p_0^{(n)})p^{(n)}(1 - p^{(n)})^{nx-1} \frac{1}{1 - (1 - p^{(n)})} = (1 - p_0^{(n)})(1 - p^{(n)})^{nx-1}. \tag{B.7}$$

Sannolikheten att den n:te generationen ej dött ut är

$$P(Z_n \geq 1) = 1 - P(Z_n = 0) = 1 - p_0^{(n)}. \tag{B.8}$$

Vi får då

$$\frac{P(Z_n \geq nx)}{P(Z_n \geq 1)} = (1 - p^{(n)})^{nx-1} \sim (1 - p^{(n)})^{nx} \tag{B.9}$$

eftersom $1 - p^{(n)} = p_0^{(n)} \sim 1$ då n är stort.

Vidare har vi från (2.9) att

$$p^{(n)} = \frac{p}{p + (1-p)n} \sim \frac{p}{(1-p)n}, \tag{B.10}$$

för stora n. Vid insättning av (B.10) i (B.9) får vi

$$\frac{P(Z_n \geq nx)}{P(Z_n \geq 1)} \sim \left(1 - \frac{p}{(1-p)n}\right)^{nx} = \left(\left(1 - \frac{p}{(1-p)n}\right)^{n}\right)^{x}. \tag{B.11}$$

Variabelsubstitution med $m = -\frac{1-p}{p}n$ ger

$$\left(\left(1 + \frac{1}{m}\right)^{m}\right)^{-x\frac{p}{1-p}},$$

vilket då m (och n) är stort ger $\exp(-x\frac{p}{1-p})$. Alltså har $(Z_n/n)|Z_n \geq 1$ exponentiell fördelning med parameter $\frac{p}{1-p}$ då (B.5) och (B.11) asymptotiskt ger

$$P\left(\frac{Z_n}{n} \geq x \,\middle|\, Z_n \geq 1\right) \sim \exp\left(-x\frac{p}{1-p}\right). \tag{B.12}$$

I figur 12 visas simuleringar som bekräftar detta. □

Bevis av lemma 2.7. Liknande som i beviset av lemma 2.6 använder vi samma notation för asymptotisk ekvivalens. Analogt som i (B.5) får vi $P(Z_n \geq x \,|\, Z_n \geq 1) = P(Z_n \geq x)/P(Z_n \geq 1)$. Vi utvecklar likt i ekvation (B.6) och får samma summa fast från x istället för nx. Då $n \to \infty$ får vi $p^{(n)} = 1 - \frac{1-p}{p_0}, p_0^{(n)} = 1$ i subkritiska fallet, eftersom $\frac{1-p_0}{p} = \mu < 1$. Eftersom även $\frac{1-p}{p_0} < 1$ har vi $(1 - p^{(n)})^{k-1} = (\frac{1-p}{p_0})^{k-1} \to 1$ då $k \to \infty$. Vi utnyttjar då formeln för geometrisk summa och får

$$P(Z_n \geq x) = (1 - p_0^{(n)})p^{(n)}(1 - p^{(n)})^{x-1} \frac{1}{1 - (1 - p^{(n)})} = (1 - p_0^{(n)})(1 - p^{(n)})^{x-1},$$

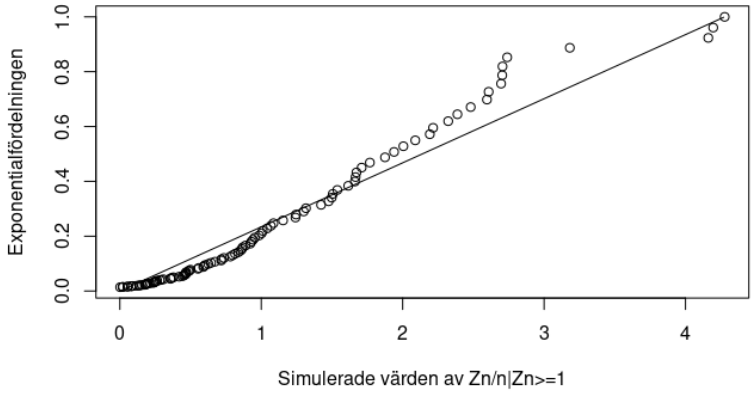
likt resultatet i (B.7). Genom uttrycket för $P(Z_n \geq 1)$ i (B.8) får vi då

$$\frac{P(Z_n \geq x)}{P(Z_n \geq 1)} = (1 - p^{(n)})^{x-1} = \left(\frac{1-p}{p_0}\right)^{x-1}. \tag{B.13}$$

Omskrivning ger nu

$$P(Z_n \geq x + 1 \,|\, Z_n \geq 1) = \left(\frac{1-p}{p_0}\right)^{x}. \tag{B.14}$$

Det är rimligt att $x + 1$ används eftersom $P(Z_n \geq 0|Z_n \geq 1) = 0$. Det här innebär att $Z_n|Z_n \leq 1$ är skiftad geometriskt fördelad med parameter $1 - \frac{1-p}{p_0}$. □



Figur 12: Simulerad data för (Z_n/n) givet $Z_n \geq 1$, jämförd med exponentialfördelningen med parameter $p/(1-p)$ i qqplot.

B.6 Asymptotisk analys

Omskrivning till geometrisk summa: Vi har serien

$$T(s) = \sum_{n=0}^{\infty} t_n s^n = \sum_{n=0}^{\infty} \sum_{k=n}^{\infty} P(W=k) s^n.$$

Vi kan också skriva summan på formen

$$\sum_{n=0}^{\infty} \sum_{k=n}^{\infty} P(W=k) s^n = (p_0 + p_1 + p_2 + \cdots) + s(p_1 + p_2 + \cdots) + s^2(p_2 + \cdots) + \cdots$$

$$= p_0 + p_1(1+s) + p_2(1+s+s^2) + \cdots = \sum_{k=0}^{\infty} P(W=k) \sum_{n=0}^{k} s^k.$$

Genom geometrisk summa får vi då

$$\sum_{k=0}^{\infty} P(W=k) \frac{s^{k+1}-1}{s-1} = \frac{1}{s-1} \left(\sum_{k=0}^{\infty} P(W=k) s^{k+1} - \sum_{k=0}^{\infty} P(W=k) \right) = \frac{sH_k(s)-1}{s-1}.$$

Vi skriver om till

$$\frac{sH_k(s)-1}{s-1} = \frac{1-s+s(1-H_k(s))}{1-s} = 1 + s\frac{1-H_k(s)}{1-s} = H_k(s) + \frac{1-H_k(s)}{1-s}, \quad s \to 1^-. \quad \text{(B.15)}$$

Asymptotisk analys: Vi vill nu göra asymptotisk analys på uttrycket i (B.15), men då måste vi först göra asymptotisk analys på

$$H_k(s) = \frac{1 + (1-2p)s - \sqrt{1 + (1-2p)^2 s^2 - 2s(1-2(1-p)p)}}{2(1-p)}$$

från (2.10). I detta uttryck är rotuttrycket mest komplicerat vilket vi först skriver om genom att subtrahera 1 från s och s^2

$$\sqrt{1 + (1-2p)^2(s^2-1) + (1-2p)^2 - 2(s-1)(1-2(1-p)p) - 2(1-2(1-p)p)}.$$

Sedan faktoriserar vi ut $(s-1)$

$$\sqrt{(s-1)((1-2p)^2(s+1) - 2(1-2(1-p)p)) + (1 + (1-2p)^2 - 2(1-2(1-p)p))}, \quad \text{(B.16)}$$

varpå vi utvecklar resterande termer

$$1 + (1-2p)^2 - 2(1-2(1-p)p) = 1 + 1 - 4p + 4p^2 - 2 + 4p - 4p^2 = 0.$$

Då kan vi bryta ut $(s-1)$ ur uttryck (B.16) vilket asymptotiskt ger

$$\sqrt{s-1}\sqrt{(1-2p)^2(s+1) - 2(1-2(1-p)p)} \sim 2\sqrt{s-1}\sqrt{p(p-1)}, \quad s \to 1^-. \tag{B.17}$$

Om vi nu sätter in (B.17) i ekvation (2.10) får vi

$$H_k(s) \sim \frac{1 + (1-2p)s - 2\sqrt{s-1}\sqrt{p(p-1)}}{2(1-p)} = \frac{1+(1-2p)s}{2(1-p)} - \sqrt{1-s}\sqrt{\frac{p}{1-p}}, \quad s \to 1^-. \tag{B.18}$$

Nu vill vi utveckla resten av uttrycket i (B.15), alltså använder vi (B.18) och får

$$\frac{1-H_k(s)}{1-s} \sim \frac{1 - \frac{1+(1-2p)s}{2(1-p)}}{1-s} + \sqrt{\frac{p}{(1-p)(1-s)}}, \quad s \to 1^-. \tag{B.19}$$

Utveckling av första termen ger

$$\frac{1 - \frac{1+(1-2p)s}{2(1-p)}}{1-s} = \frac{2(1-p) - 1 - (1-2p)s}{2(1-s)(1-p)} = \frac{2(1-p) - 1 + (1-2p)(1-s) - (1-2p)}{2(1-s)(1-p)} = \frac{1-2p}{2(1-p)},$$

vilket insatt i uttrycket (B.19) blir

$$\frac{1-2p}{2(1-p)} + \sqrt{\frac{p}{(1-p)(1-s)}} \sim \frac{\sqrt{2}}{\sigma\sqrt{1-s}}, \quad s \to 1^-$$

och därmed får vi genom användning av (B.15)

$$T(s) = H_k(s) + \frac{1-H_k(s)}{1-s} \sim \frac{\sqrt{2}}{\sigma\sqrt{1-s}}, \quad s \to 1^-. \tag{B.20}$$

B.7 Logaritmisk fördelning och maximalt antal individer

Bevis av lemma 2.8. Vi har

$$P(X \geq x) \sim x^{-\alpha}, x \to \infty,$$

med $\alpha > 0$. Då

$$P(\log(X) \geq x) = P(X \geq e^x) \sim e^{-\alpha x}, x \to \infty.$$

Det ger att $\log(X)$ är exponentialfördelad med parameter α. □

Bevis av lemma 2.9. Beteckna slumpvariablen för det maximala antalet individer $V = \max(W_1, W_2, \ldots, W_N)$. Låt $w = W_{\max}(N)$ för att förenkla notationen i beviset. Vi har

$$\begin{aligned} P(V \leq w) &= P(\max(W_1, W_2, \ldots, W_N) \leq w) = P(\underbrace{W_1 \leq w \cap W_2 \leq w \cap \cdots \cap W_N \leq w}_{W_i \text{ är oberoende } \forall i}) \\ &= \underbrace{P(W_1 \leq W_{\max})P(W_2 \leq w)\cdots P(W_N \leq w)}_{W_i \text{ lika fördelade } \forall i} = (P(W \leq w))^N \end{aligned}$$

Om nu N och w är stora gäller att $P(W \leq w)$ är nära 1 så

$$(P(W \leq w))^N = (1 - P(W > w))^N \sim e^{-NP(W>w)}, \quad N \to \infty, w \to \infty,$$

om $P(W > w) = \mathcal{O}(\frac{1}{N})$. Vi noterar att detta är överlevnadsfunktionen, alltså $P(W > w) = \sum_{k=w+1}^{\infty} P(W = k)$ vilket enligt avsnitt 2.4.1 ger

$$\sum_{k=w+1}^{\infty} P(W = k) \sim \sqrt{\frac{2}{\pi\sigma^2}}(w+1)^{-1/2} \sim \sqrt{\frac{2}{\pi\sigma^2}} w^{-1/2}, \quad w \to \infty.$$

Eftersom vi vill att $P(W > w) = \mathcal{O}(\frac{1}{N})$ så har vi $w^{-1/2} = \mathcal{O}(\frac{1}{N})$ vilket ger $W_{\max}(N) = \mathcal{O}(N^2)$. □

C Teori för simuleringar

I detta avsnitt finns mer ingående teori för matematiska metoder som en del av simuleringarna bygger på.

C.1 Generering av geometriska slumpvariabler

I konturprocessen behöver vi kunna generera slumpvariabler från den skiftade geometriska fördelningen, som definieras i definition 2.10. Detta kan enkelt göras med hjälp av den funktionen `rgeom` i `R`, men för simuleringar i `Java` behöver vi konstruera en egen metod. Detta kan åstadkommas genom att finna den generaliserade inversen till fördelningsfunktionen $F(k)$ [21]:

$$F^{-1}(u) = \inf\{k : F(k) \geq u\}, \quad 0 < u < 1.$$

Fördelningsfunktionen för den skiftade geometriska fördelningen är

$$F(k) = \sum_{i=1}^{k} p(1-p)^{i-1} = 1 - (1-p)^k, \quad k = 1, 2, \ldots \tag{C.1}$$

Vi börjar med att invertera $F(k)$

$$F(k) = 1 - (1-p)^k \implies k = \frac{\log(1 - F(k))}{\log(1-p)}. \tag{C.2}$$

Låt $0 < u < 1$. $F(k)$ är likformigt fördelad på $[0, 1]$ eftersom

$$P(F(k) \leq u) = P(k \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

Därför sätter vi $u := F(k)$ och låter u vara likformigt fördelat på $(0, 1)$. Detta ger att

$$P(F^{-1}(u) \leq k) = P(u \leq F(k)) = F(k),$$

vilket innebär att $F^{-1}(u)$ är fördelad efter $F(k)$. Således får vi ett slumpmässigt k genom att generera ett slumptal $u \sim \text{likf}(0, 1)$ och använda ekvation (C.2). Eftersom vi vill att k ska vara ett heltal så använder vi den generaliserade inversen, och tar därför takfunktionen:

$$k \sim \left\lceil \frac{\log(1-u)}{\log(1-p)} \right\rceil.$$

C.2 Teori bakom qq-plot

Antag att vi har m oberoende observationer från den likformiga fördelningen på $[0, 1]$ och att dessa är ordnade i storleksordning $U_1, \ldots, U_m$ [22]. Då kommer $\mathbb{E}[U_{i+1} - U_i] = 1/(m+1)$ och därmed $\mathbb{E}[U_i] = i/(m+1)$ för alla i. Eftersom U_i inte kommer avvika alltför mycket från sitt medelvärde kommer grafen med $i/(m+1)$ som x-axel och U_i som y-axel vara ungefär linjär.

Antag nu istället att vi har m oberoende och likfördelade observationer $X_1, X_2, \ldots, X_m$ i storleksordning [22]. Om vi vill jämföra dessa med en fördelningsfunktion F så kan vi rita en graf med $i/(m+1)$ som x-axel och $F(X_i)$ som y-axel och undersöka dess linearitet. Detta är ekvivalent med x-axeln $F^{-1}(i/(m+1))$ och y-axeln X_i. Notera att $F^{-1}(i/(m+1))$ är den så kallade teoretiska kvantilen och X_i den empiriska kvantilen, därav namnet quantile-quantile-plot. Om ett linjärt beroende uppfylls är det troligt att datan följer den givna fördelningen.

C.3 Momentmetoden för LF-fördelningen

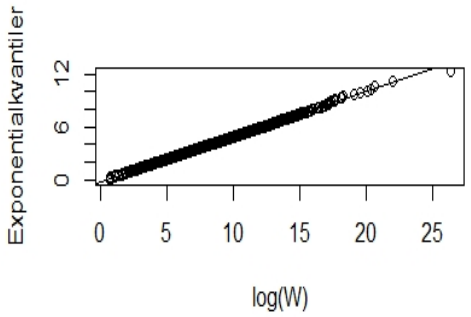
Momentmetoden är en statistisk metod för att uppskatta parametrar från given data genom att beräkna dess respektive moment, såsom väntevärdet och variansen [23]. I vårt fall vi anpassa SCB-datan i avsnitt 5.1 till LF-fördelningen, och därmed bestämma parametrarna p och p_0. Väntevärdet för LF-fördelningen ges av $\mu = (1-p_0)/p$ och variansen av $\sigma^2 = (1-p_0)(1-p+p_0)/p^2$. Genom att beräkna μ och σ^2 för datan får vi då två ekvationer för p och p_0 vilket med SCB-datan ger $p_0 = 0.29$ och $p = 0.71$.

D Övriga resultat och exempel

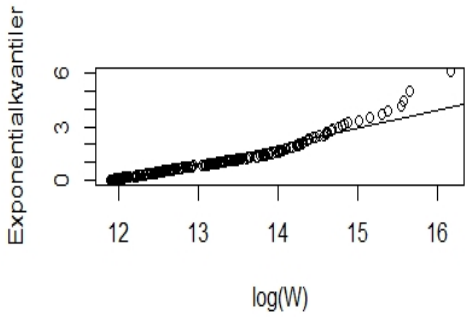
I detta avsnitt finns ytterligare resultat som styrker resultaten i avsnitt 4.2. Dessutom presenteras en enklare modell av COVID-19.

D.1 Resultat

I figur 13 jämförs logaritmen av datans svans (hela svansen för n större än 1 i kritiska och bortom tre standardavvikelser i subkritiska) med en exponentialfördelning i qq-plottar. Att detta går att göra för att avgöra om datan följer en potensfördelning beror på att logaritmen av en potensfördelad slumpvariabel är exponentialfördelad, enligt lemma 2.8. I figur 13a ses att träden som simulerades i det kritiska fallet väl följer den räta linjen, vilket styrker att svansarna i det kritiska fallet följer en potenslag. För träden som simulerades i det subkritiska fallet avviker datapunkterna i slutet uppåt från den räta linjen, vilket tyder på att svansarna är lättare än potensfördelade svansar. I figur 13b visas fallet där $\mu = 0.999$, vilket styrker att det räcker med att vara strax under det kritiska fallet för att inte längre ha svansar som följer en potenslag, precis som figur 7 visar.



(a) En qq-plot i kritiska fallet med $\mu = 1.0$. Datapunkterna följer den räta linjen, vilket tyder på att svansarna i det kritiska fallet följer en potenslag.

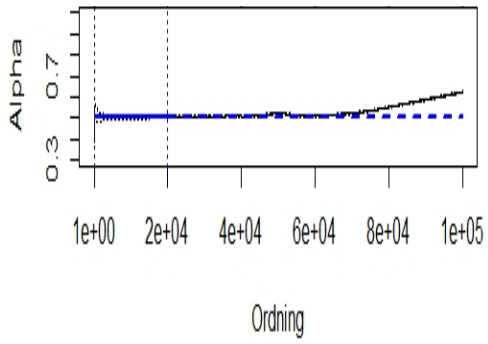


(b) En qq-plot i subkritiska fallet med $\mu = 0.999$. Datapunkterna avviker uppåt från den räta linjen, vilket tyder på lättare svansar än potensfördelade.

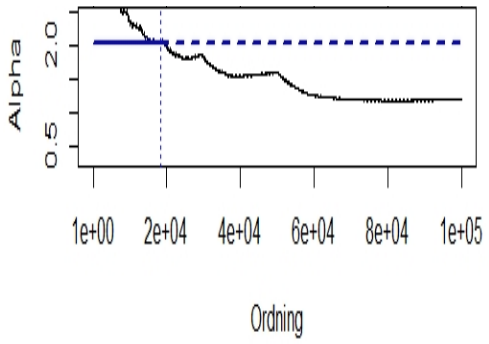
Figur 13: Jämförelse mellan exponentialfördelade svansar och logaritmen av totala antalet individer W i träd som genererats med konturprocessen för olika μ. En jämförelse mellan exponentialfördelade svansar och logaritmerad data är detsamma som en jämförelse mellan svansar som följer en potenslag och datan.

I figur 14 ses en jämförelse mellan resultatet av Hill-estimatorn för det kritiska fallet och det subkritiska fallet med $\mu = 0.6$, vilket är ungefär där gränsen för tungsvansad går enligt tidigare resultat.

Grafen stabiliserar sig på den streckade horisontella linjen för svansindex $\alpha = 0.5$ i det kritiska fallet, se figur 14a, vilket stämmer överens med den teoretiska potenslagen och övriga simulerade resultat. I det subkritiska fallet, se figur 14b, stabiliseras däremot inte grafen, vilket tyder på att denna inte följer en potensfördelning, precis som de andra simulerade resultaten antyder.



(a) Hill-estimatorns resultat i det kritiska fallet med $\mu = 1.0$. Grafen stabiliserar sig för $\alpha = 0.5$.

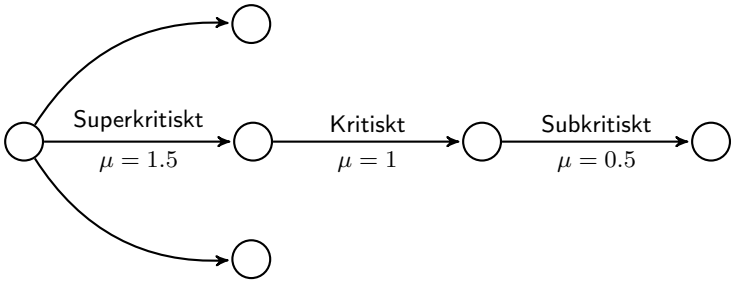


(b) Hill-estimatorns resultat i det subkritiska fallet med $\mu = 0.6$. Grafen stabiliseras inte.

Figur 14: Jämförelse mellan Hill-estimatorns resultat för fördelningen av totala antalet individer W i träd som genererats med konturprocessen för olika μ.

D.2 COVID-19-modell

Här presenteras en enkel modell av COVID-19 med en approximation av en storleksberoende Galton-Watson-process. Storleksberoende Galton-Watson-processer är tidskrävande att simulera och vi försökte istället att approximera storleksberoendet med en enkel modell. Först följer processen det superkritiska fallet, och sedan i varje generation, minskas medelvärdet och efter ett tag blir processen kritisk, och sedan subkritisk. I figur 15 illustreras hur en sådan process kan se ut, där väntevärdet minskar snabbt från $\mu = 1.5$ till $\mu = 0.5$.

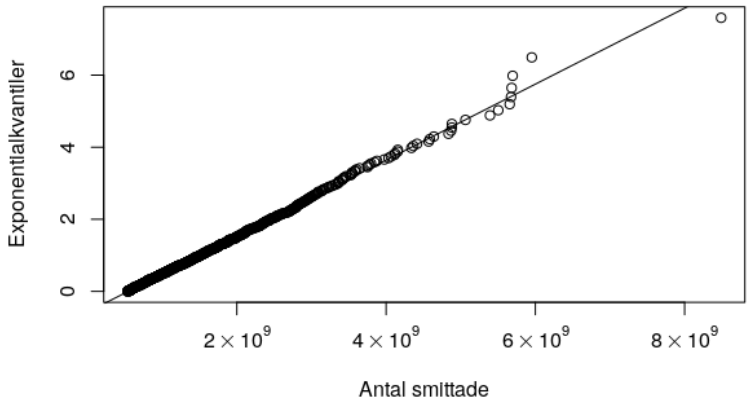


Figur 15: En förenklad version av COVID-19 modellen. I det superkritiska fallet smittar först en person tre, varav två av dessa ej smittar någon. Den tredje personen smittar i det kritiska fallet en ny person och avslutande smittas en sista person i det subkritiska fallet, innan sjukdomen dör ut.

Vi har alltså smittfasen, när viruset sprids snabbt, och sedan efter ett tag är de flesta smittade så att smittan i efterhand dör ut. Denna minskning behöver dock vara långsammare än i figur 15 för att det ska vara realistiskt, eftersom det krävs många smittade för att processen ska bli subkritisk. Här antar vi också att smittade blir immuna och ej kan smittas igen.

Vid simulering av detta fick vi att svansen för viruset ser ut att vara exponentialfördelad, se figur 16. Detta kan bero på att de superkritiska fallen blev väldigt stora men alltid avbröts, de tilläts inte gå för evigt som i våra övre simuleringar. Denna skillnad, ger skillnaden mellan en potensfördelad svans och exponentialfördelad. Det kan liknas med resultatet att $Z_n/n|Z_n \geq 1$ blir exponentialfördelad i kritiska fallet. Det beror på att processen avbrutits vid generation n och det ej är W som vi tittar på. Dock måste denna hypotes undersökas vidare med fler grafiska metoder för att kunna verifieras.

Notera att vår smittspridningsfördelning antyds ha lätta svansar, enligt figur 16, medan antal döda i globala pandemier verkar följa en tungsvansad fördelning [20]. Därav är det av intresse att kontrollera huruvida modellen ger upphov till lätta svansar eller ej. Här får vi dock också notera skillnaden på antal smittade och antal döda, som är särskilt märkbar för COVID-19. Vidare har



Figur 16: QQ-plot för vår virusmodell efter 10000 simuleringar. Endast värden en standardavvikelse från väntevärdet räknas med för att det ska motsvara svansen.

de undersökt alla historiska pandemier med en viss dödlighet och det är alltså fördelningen av pandemier som har tung svans och inte själva pandemiernas svansar.

E Simuleringskod

Här följer koden för simulering av träden för Galton-Watson-processen med LF-fördelningen som reproduktionsfördelning i kritiska och subkritiska fallet med konturprocessen. Om vi i konturprocessen utgår från en individ och tar ett slumpmässigt antal steg uppåt (vilket alltså teoretiskt ges av Geom(p_0)) och sedan ett slumpmässigt antal steg nedåt (vilket teoretiskt ges av Geom($1-p$)) och har en situation där antal steg neråt är färre än antal steg uppåt så har vi kommit till en individ med fler än ett barn. Då tas manuellt ett steg uppåt därifrån, för dess andra barn, och sedan kan på nytt ett slumpmässigt antal steg uppåt och ett slumpmässigt antal steg nedåt tas, samt därefter ett nytt manuellt steg uppåt om processen då fortfarande inte dött ut (d.v.s. nått -1). Detta itereras sedan tills processen dött ut. För högre ordningens barn gäller detsamma som för en individs andra barn. Fördelningen för det slumpmässiga antalet steg uppåt är den vanliga geometriska för annars skulle det innebära att det alltid är minst två barn på grenar i konturprocessens motsvarande träd som går ut för en individs andra (eller högre ordningens) barn, eftersom vi lägger till steget för en individs andra barn manuellt. Fördelningen neråt är den skiftade geometriska för annars skulle det innebära att det efter att ha gått uppåt ett slumpmässigt antal steg skulle kunna vara möjligt att få noll steg neråt och således skulle då ett manuellt steg uppåt tas för det andra barnet för individen längst ut, vilket skulle vara fel eftersom det där inte finns något första barn. Observera dock att eftersom vi manuellt tar ett steg uppåt och låter det slumpmässiga antalet steg uppåt utöver det börja på 0, medan det slumpmässiga antalet steg nedåt börjar på 1, så är det väsentligen samma typ av fördelning, med skillnaden att de beror på olika parametrar.

Programmet nedan körs genom att ange antal simuleringar, värdet på sannolikheterna p, p_0 samt hur många trådar som programmet ska köras på. Konturprocessen beskrivs i egen metod rad 75-90 och simuleringen av slumpvariabler rad 102-109.

```
1 import java.io.*;
2 import java.util.concurrent.*;
3 import java.util.concurrent.atomic.*;
4
5 public class Main {
6
7   private static int no_of_simulations;
8   private static double p0, p;
9   private static volatile AtomicInteger count = new AtomicInteger(0); // counts how
      many simulations have finished
```

```
10  private static volatile ConcurrentLinkedQueue<Long> data = new
      ConcurrentLinkedQueue<Long>(); // contains all simulated data
11
12  public static void main(String[] args) throws InterruptedException, IOException {
13    // parse the command line arguments
14    no_of_simulations = Integer.parseInt(args[0]);
15    p0 = Double.parseDouble(args[1]);
16    p = Double.parseDouble(args[2]);
17    if (p0 < 0.0 || p0 > 1.0 || p < 0.0 || p > 1.0)
18      throw new IllegalArgumentException("p0 and p must be in the interval (0, 1).");
19    double mean = Math.round(1000.0 * (1.0 - p0) / p) / 1000.0; // mean rounded to
      three decimal places
20    int no_of_threads = Integer.parseInt(args[3]);
21
22    // declare and start all threads
23    Thread[] workers = new Thread[no_of_threads];
24    int simulations_per_thread = no_of_simulations / no_of_threads;
25    int remainder = no_of_simulations % no_of_threads;
26    for (int i = 0; i < no_of_threads; i++) {
27      if (i == 0)
28        workers[i] = new GW(simulations_per_thread + remainder);
29      else
30        workers[i] = new GW(simulations_per_thread);
31      workers[i].start();
32    }
33
34    // wait for all threads to finish
35    for (Thread t : workers)
36      t.join();
37
38    // write the data to a .csv file
39    StringBuilder sb = new StringBuilder();
40    for (Long element : data)
41      sb.append(element.toString() + "\n");
42
43    String path = "data " + no_of_simulations + " " + mean + ".csv";
44    BufferedWriter br = new BufferedWriter(new FileWriter(path));
45    try {
46      br.write(sb.toString());
47    } catch (IOException e) {
48      System.out.println(e);
49    } finally {
50      br.close();
51    }
52
53    System.out.println("Done! Data written to " + path);
54  }
55
56
57  private static class GW extends Thread {
58    private int simulations;
59    private static final int progress_step = 100;
60
61    public GW(int simulations) { this.simulations = simulations; }
62
63    // simulate <<this.simulations>> number of GW processes
64    public void run() {
65      int total_progress;
66      for (int i = 1; i <= simulations; i++) {
67        data.add(kontur());
68        if ((i % progress_step == 0 && i > 1) || i == simulations) {
69          total_progress = count.addAndGet(progress_step);
70          System.out.println("Progress: " + total_progress + " of " +
      no_of_simulations);
71        }
72      }
73    }
74
75    // simulates a GW process via the contour process
```

```
76      private long kontur() {
77        long v = 0, s = 1;
78        while (v > -1) {
79          int up = random_geom(p0) - 1;
80          v += up;
81          s += up;
82          int down = random_geom(1-p);
83          v -= down;
84          if (v > -1) {
85            v++;
86            s++;
87          }
88        }
89        return s;
90      }
91
92      // generates a random variable from the shifted geometric distribution
93      private int random_geom(double p) {
94        double u;
95        do {
96          u = Math.random();
97        } while (u == 0.0);
98        return (int)Math.ceil(Math.log(1-u)/Math.log(1-p));
99      }
100
101     // generates a random variable from the linear fractional distribution
102     private int random_LF(double p0, double p) {
103       if (Math.random() <= 1-p0)
104         return random_geom(p);
105       else
106         return 0;
107     }
108   }
109 }
```

Importerande av data i `R` från simuleringar gjorda i `Java`, approximerande av GW-processens överlevnadsfunktion samt plottande av denna funktion mot $\sqrt{2/\pi\sigma^2}n^{-1/2}$.

```
1  # This file is used for
2  # 1. importing data from simulations done in java
3  # 2. approximating the process' survival function (contained within the variable '
       probSum', and can also be computed from the function 'dist(n)')
4  # 3. plotting the approximated survival function against the "real" one.
5
6  ######## Conditions #########
7  p0=0.5
8  p=0.5
9  sigmaSq = (1-p0)*(1-p+p0)/p^2
10 mu = (1-p0)/p
11 tries=200000
12
13 ######## Importing data from java program #########
14 v = as.vector(t(read.csv('data 200000 1.0.csv', FALSE)))
15 tries = length(v)
16
17 ######## Presentation of the data in a matrix #########
18 n_values = c()
19 probabilities = c()
20 unitProb = 1/tries
21 length = 0
22 for (i in 1:tries) {
23   simulation = v[i]
24   index = match(simulation, n_values)
25   if (is.na(index)) {
26     length = length + 1
27     n_values[length] = simulation
28     probabilities[length] = unitProb
29   }
30   else
31     probabilities[index] = probabilities[index] + unitProb
32 }
```

```
prob = cbind(n_values, probabilities)
prob = prob[order(prob[,1]),] # sort according to column 1

# a matrix which lists all values of n that was found through the simulation, and
    the associated cumulative probabilities
probSum = cbind(sort(n_values), rev(cumsum(rev(prob[, 2]))))

# dist(n) = P(W >= n)
dist = function(n) {
  if (n <= 1) return(1)
  if (n > max(probSum[, 1])) return(0)

  index = findInterval(n, probSum[,1])
  if (n %in% probSum[,1] == FALSE)
    index = index + 1
  return(probSum[index, 2])
}
dist_vec = Vectorize(dist)

W_max = max(v) # maximum W that was found through simulation

###### plotting ########
p_n <- function(n) { return(n^(-1/2)*sqrt(2/pi/sigmaSq)) }
n_min = 10^5
n_max = 10^9
n_min_index = findInterval(n_min, probSum[,1])
n_max_index = findInterval(n_max, probSum[,1])
N = n_min_index:n_max_index
M = seq(n_min, n_max, 1000)

xlim = c(10^5, 10^9)
ylim = c(10^-5, 1.8*10^-3)
plot(probSum[N, 1], probSum[N, 2], 'l', col="blue", log="x", xaxt = "n", yaxt = "n"
    , xlim=xlim, ylim=ylim, xlab="n", ylab="t(n)")
axis(1, at=c(10^5, 10^6, 10^7, 10^8, 10^9), labels=c(expression(10^5), expression
    (10^6), expression(10^7), expression(10^8), expression(10^9)))
axis(2, at=c(0, 5*10^-4, 10^-3, 1.5*10^-3), labels=c(expression(0), expression(5 %*
    % 10^{-4}), expression(10^{-3}), expression(1.5 %*% 10^{-3})))
lines(M, p_n(M), 'l', col="red")
legend(x=2.5*10^7, y=0.0018, legend=c(expression(paste("t(n) = ", sqrt(frac(2,
    paste(pi, sigma^2))), " ", "n"^{-1/2}, "     ")), "simulering"), col=c("red", "
    blue"), lty=1, cex=0.8)
```

Simulering av $Z_n/n|Z_n \geq 1$ i kritiska fallet med LF-fördelning och jämförelse med exponentialfördelningen i en qq-plot, se figur 12.

```
library(svMisc)
library(evir)
library(EnvStats)

######## Conditions #########
p0=0.5
p=0.5
sigmaSq = (1-p0)*(1-p+p0)/p^2
mu = (1-p0)/p
tries=100000

######## Simulation ########
#The n value at which we stop Zn
n = 1000

#Simulates n generations of Z, stops if 0
branching_n <- function(p0, p, n) {
  Z <- 1
  for (i in 1:n) {
    Z = sum(rbinom(Z, 1, 1-p0)*(rgeom(Z, p) + 1))
    if (Z == 0) break;
  }
  return(Z)
}

```

```
v = vector("double")
j = 1
for (i in 1:tries) {
  z = branching_n(p0,p,n)
  #Conditional probability on z>=1 means we are only interested in z!=0
  if (z != 0) {
    v[j] = z
    j = j + 1
  }
  progress(i, tries)
}

v = v/n #We want Zn/n
m = max(v) #Used for scaling the exponential distribution

X = seq(0,m,0.01)

#qq plot comparing the exponential distribution with parameter p/1-p with our
    values
qqplot(v,dexp(X, rate = p/(1-p)), main = "", ylab = "Exponential distribution",
    xlab = "Simulated values of Zn/n|Zn>=1")
lines(X, X/m)
```

F Wikipediaartiklar

I detta avsnitt finns bilder, tagna 2020-05-11, på Wikipedia-artiklarna som skapades som arbetets bidrag till allmänheten. De respektive artiklarna handlar om Galton-Watson-processen, sannolikhetsgenererande funktioner samt LF-fördelningen.

Galton-Watson processen [redigera | redigera wikitext]

Galton-Watson-processen är en stokastisk förgreningsprocesses som ursprungligen användes av Francis Galton och Henry William Watson under den andra halvan av 1800-talet för att undersöka hur familjenamn sprider sig genom generationer, och med vilka sannolikheter olika efternamn eventuellt dör ut[1]. I modellen så sprids efternamn från fader till son, där spridningen beror på hur många söner mannen får under en generation. Om mannen inte får några söner alls så har efternamnet dött ut.

Galton-Watson-processen kan användas för att beskriva många andra slags företeelser där det sker förgreningar mellan individer, exempelvis inom populationsforskning och sjukdomsspridning.

Innehåll [dölj]

Matematisk beskrivning [redigera | redigera wikitext]

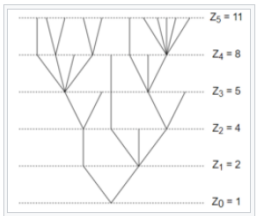
Galton-Watson-processen är en Markovkedja, där storleken av en viss generation bara beror på storleken av den föregående generationen samt hur många barn som föds[2]. Låt Z_n beteckna antalet individer vid generation n, med $Z_0 = 1$, det vill säga, populationen startar med en individ. Processen kan då beskrivas efter rekursionsformeln

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_i,$$

där X_i är en stokastisk variabel som betecknar antalet barn som individ i födde i generation n, och som för alla i är fördelad efter den så kallade reproduktionsfördelningen:

$$X = \begin{cases} 0, & \text{med sannolikhet } p_0, \\ 1, & \text{med sannolikhet } p_1, \\ 2, & \text{med sannolikhet } p_2, \\ \vdots \end{cases}$$

Varje individ ger alltså upphov till k barn med samma sannolikhet p_k.Följden $Z_0, Z_1, Z_2, \ldots$ beskriver då populationens storlek efter varje generation.



Ett exempel på en Galton-Watson-process som börjar med en startindivid.

Förväntad generationsstorlek efter n generationer [redigera | redigera wikitext]

Låt $\mu = \sum_{k=0}^{\infty} k p_k$ beteckna reproduktionsfördelningens väntevärde, det vill säga det genomsnittliga antalet avkommor som varje individ ger upphov till. Väntevärdet av Z_n ges då av $\mathbb{E}[Z_n] = \mu^n$. Långsiktigt, då $n \to \infty$, kan vi se att det finns tre fall för den förväntade generationsstorleken:

$$\lim_{n\to\infty} \mathbb{E}[Z_n] = \lim_{n\to\infty} \mu^n = \begin{cases} 0, & \mu < 1, \\ 1, & \mu = 1, \\ \infty, & \mu > 1. \end{cases}$$

De tre fallen $\mu < 1$, $\mu = 1$ och $\mu > 1$ benämns de subkritiska, kritiska respektive superkritiska fallen för processen[1]. Väntevärdet μ är alltså helt avgörande för populationens tillväxt och långsiktiga utseende. Som ett exempel, antag en Galton-Watson-process vars reproduktionsfördelning är

$$X = \begin{cases} 0, & \text{med sannolikhet } 1-p, \\ 2, & \text{med sannolikhet } p, \end{cases}$$

det vill säga att varje individ dör ut med sannolikhet $1-p$ eller förgrenas till två delar med sannolikhet p. Denna fördelning är en uppskalad Bernoullifördelning. Då blir $\mu = 2p$, och processen är subkritisk om $p < 0.5$, kritisk om $p = 0.5$ samt superkritisk om $p > 0.5$. Vidare ges variansen av generationsstorlekar av

$$\lim_{n\to\infty} \mathrm{Var}[Z_n] = \begin{cases} 0, & \mu < 1, \\ \infty, & \mu \geq 1. \end{cases}$$ [1]

Som kan ses så är variansen oändlig det i kritiska och superkritiska fallet.

Utrotningssannolikhet [redigera | redigera wikitext]

Det är möjligt att beräkna en Galton-Watson-process utrotningssannolikhet, $q = P(\exists n : Z_n = 0)$, genom att använda sig av sannolikhetsgenererande funktioner[1]. Låt $G(s)$ beteckna reproduktionsfördelningens sannolikhetsgenererande funktion, det vill säga

$$G(s) = \sum_{k=0}^{\infty} s^k P(X = k).$$

Då är q lika med den minsta icke-negativa lösningen till ekvationen $s = G(s)$[1]. I det kritiska och subkritiska fallet så gäller det oavsett reproduktionsfördelningen att $q = 1$, det vill säga att populationen är garanterad att eventuellt dö ut. Däremot i det superkritiska fallet så kommer vi generellt att ha $q < 1$, så att processen fortsätter i all evighet. Tag åter reproduktionsfördelningen

$$X = \begin{cases} 0, & \text{med sannolikhet } 1-p, \\ 2, & \text{med sannolikhet } p, \end{cases}$$

som ett exempel. Denna fördelningens sannolikhetgenererande funktion är $G(s) = (1-p) + ps^2$, vilket ger ekvationen $s = (1-p) + ps^2$, som har lösningarna $s = 1$ och $s = \frac{1}{p} - 1$, $p \neq 0$. Mycket riktigt så är $\frac{1}{p} - 1 < 1$ endast när $p > 0.5$, vilket är ekvivalent med att $\mu = 2p > 1$.

För att betrakta långsiktigt beteende av Galton-Watson-processen kan det också vara relevant att betrakta upprepad applicering av sannolikhetsgenererande funktionen G[1]. Här är LF-fördelningen användbar för att få den genererande funktionen på en enklare form, särskilt efter upprepad applikation av funktionen[3].

Totala antalet individer W [redigera | redigera wikitext]

Det totala antalet individer ges av $W = \sum_{k=0}^{\infty} Z_k$, där Z_k alltså betecknar antal individer i generation k. För en godtycklig Galton-Watson-process så är $\mathbb{E}[W] = \begin{cases} \frac{1}{1-\mu}, & \mu < 1, \\ \infty, & \mu \geq 1, \end{cases}$

där μ är reproduktionsfördelningens väntevärde[4]. Om $\mu = 0.9$, till exempel, så är $\mathbb{E}[W] = 10$. Notera att utrotningssannolikheten i kritiska fallet 100% men väntevärdet av totalt antal individer i kritiska fallet är oändligt. För variansen gäller följande:[4]

$$\mathrm{Var}[W] = \begin{cases} \frac{\sigma^2}{(1-\mu)^3}, & \mu < 1, \\ \infty, & \mu \geq 1. \end{cases}$$

I det kritiska fallet är alltså både väntevärdet och variansen oändliga. Fortsättningsvis kan ett förhållande mellan den sannolikhetsgenererande funktionen G_W för W och reproduktionsfördelningens sannolikhetsgenererande funktionen G härledas[4]. Vi får

$$G_W(s) = sG(G_W(s)).$$

Genom att sätta $x = G_W(s)$ får vi ekvationen $x = sG(x)$ att lösa för att bestämma G_W.

Referenser [redigera | redigera wikitext]

1. ^ [a b c d e f] Dobrow, Robert P.,. *Introduction to stochastic processes with R*. ISBN 978-1-118-74072-9. OCLC 922799569. Läst 26 mars 2020
2. ^ Sagitov, Serik; Minuesa, Carmen (2017-07-03). "Defective Galton–Watson processes". *Stochastic Models* 33 (3): sid. 451–472. doi:10.1080/15326349.2017.1349614. ISSN 1532-6349. Läst 26 mars 2020.
3. ^ Athreya, Krishna B. (1972). *Branching Processes*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-65371-1. ISBN 978-3-642-65373-5. Läst 8 april 2020
4. ^ [a b c] Stirzaker, David. (2005). *Stochastic processes and models*. Oxford University Press. ISBN 1-4237-7093-5. OCLC 68623723. Läst 28 mars 2020

Kategorier: Sannolikhetsteori | Stokastiska processer

Sannolikhetsgenererande funktion [redigera | redigera wikitext]

Den **sannolikhetsgenererande funktionen** för en diskret slumpvariabel är en potensserierepresentation av slumpvariabelns sannolikhetsfunktion. Sannolikhetsgenererande funktioner används ofta för deras kortfattade beskrivning av följden Pr (*X* = k) i sannolikhetsfunktionen för en slumpmässig variabel *X*. Vidare, om sannolikhetsfunktionen är reproduktionsfördelningen för en Galton-Watson-process, ger upprepad applicering av den sannolikhetsgenererande funktionen långsiktigt beteende för processen[1].

Innehåll [dölj]

Definition [redigera | redigera wikitext]

Om *X* är en diskret slumpvariabel som har utfallsrummet {0,1, ...}, definieras den *sannolikhetsgenererande funktionen* för *X* som [1]

$$G(s) = \mathrm{E}(s^X) = \sum_{k=0}^{\infty} p(k)s^k,$$

där *p* är sannolikhetsfunktionen för *X*.

Egenskaper [redigera | redigera wikitext]

En del intressanta egenskaper för sannolikhetsgenererande funktioner kan härledas.

1. Sannolikhetsfunktionen för *X* fås genom att derivera *G*[1],
$$p(k) = \Pr(X = k) = \frac{G^{(k)}(0)}{k!}.$$
2. Det följer från egenskap 1 att om två slumpvariabler *X* och *Y* har sannolikhetsgenererande funktioner som är lika, $G_X = G_Y$ så är även $p_X = p_Y$ [1]. Alltså, om *X* och *Y* har identiska sannolikhetsgenererande funktioner, har de identiska sannolikhetsfunktioner.
3. Väntevärdet av X ges av $\mathbb{E}[X] = G'(1^-)$.[1] Vidare ges variansen av *X* av[1] $\mathrm{Var}(X) = G''(1^-) + G'(1^-) - \left[G'(1^-)\right]^2$.
4. $G_X(e^t) = M_X(t)$ där *X* är en slumpvariabel, $G_X(t)$ är den sannolikhetsgenererande funktionen och $M_X(t)$ är den momentgenererande funktionen.

Exempel [redigera | redigera wikitext]

- Den sannolikhetsgenererande funktionen för en konstant slumpvariabel, dvs Pr (*X* = c) = 1, är
$$G(s) = s^c.$$
- Den sannolikhetsgenererande funktionen för en Bernoullifördelad slumpvariabel med parameter *p* ges av
$$G(s) = (1-p) + ps.$$
- Den sannolikhetsgenererande funktionen för en Poissonfördelad slumpvariabel med parametern λ är
$$G(z) = e^{\lambda(z-1)}.$$

Referenser [redigera | redigera wikitext]

1. ^ [a b c d e f] Dobrow, Robert P.,. *Introduction to stochastic processes with R*. ISBN 978-1-118-74072-9. OCLC 922799569. Läst 29 mars 2020

Kategori: Genererande funktioner

LF-fördelning [redigera | redigera wikitext]

En **LF-fördelning** (engelska *linear fractional distributions*) är en fördelning vars sannolikhetsgenererande funktion kan skrivas som kvoten mellan två linjära funktioner. En rationell funktion med linjära funktioner i täljare och nämnare ges av

$$g(x) = \frac{a + bx}{c + dx}$$

där a, b, c och d är godtyckliga konstanter. En slumpvariabel X sägs vara LF-fördelad med parametrarna $p_0, p \in [0, 1]$ om dess sannolikhetsgenererande funktion kan skrivas som[1]

$$f(s) = E[s^X] = p_0 + (1 - p_0)\frac{ps}{1 - (1 - p)s}.$$

Notera att fördelningen kan skrivas på formen av ett linjärt bråk då

$$f(s) = \frac{p_0 + (p - p_0)s}{1 + (p - 1)s}.$$

Det behövs dock endast två fria parametrar p och p_0 eftersom f är en genererande funktion med egenskaperna $f(0) = p_0$ och $f(1) = 1$. Tolkningen av LF-fördelningen är att det ger det totala antalet barn, där sannolikheten för det första barnet är $1 - p_0$ och sannolikheten att få barn efter det första barnet är $1 - p$ för varje barn[1]. LF-fördelningar är användbara för att få den genererande funktionen på en enklare form, särskilt efter upprepad applikation av funktionen.

Innehåll [dölj]

Koppling till geometrisk fördelning [redigera | redigera wikitext]

En slumpvariabel X, fördelad efter den diskreta fördelningen

$$P(X = k) = p_k = \begin{cases} p_0, & \text{om } k = 0 \\ p(1 - p_0)(1 - p)^{k-1}, & \text{om } k \geq 1 \end{cases}$$

har ekvation f som sannolikhetsgenererande funktion för $|s(1 - p)| < 1$[1]. Då kan vi se att LF-fördelningen är nära besläktad med den geometriska fördelningen. Om $p_0 = p$ så får vi att

$$f(s) = \frac{p}{1 - (1 - p)s},$$

vilket är den sannolikhetsgenererande funktionen för den geometriska fördelningen[2]. Så om $X \overset{D}{\sim} \text{Geom}(p)$ så är $f(s)|_{p_0=p} = G_X(s)$. Dessutom, om $p_0 = 0$ får vi

$$f(s)_{p_0=0} = \frac{ps}{1 - (1 - p)s} = G_{X+1}(s)$$

vilket är den sannolikhetsgenererande funktionen för den skiftade geometriska fördelningen (benämnd *för-första-gången-fördelning* i artikeln om geometrisk fördelning). Vidare är $P(X > 0) = 1 - p_0$ och X givet $X > 0$ är fördelad enligt den skiftade geometriska fördelningen .

Användning i Galton-Watson-processen [redigera | redigera wikitext]

LF-fördelningen kan användas som reproduktionsfördelning i Galton-Watson-processen. Vi har från egenskaper av sannolikhetsgenererande funktioner[3] att väntevärdet ges av

$$\mu = E[Z_1] = f'(1) = \frac{(1 - p_0)p}{(1 - (1 - p)s)^2}\Big|_{s=1} = \frac{1 - p_0}{p}.$$

De tre fallen $\mu < 1$, $\mu = 1$ och $\mu > 1$ benämns de subkritiska, kritiska respektive superkritiska fallen för processen[3]. I LF-fallet har vi då att $1 - p_0 < p$ i det subkritiska fallet, $1 - p_0 = p$ i det kritiska fallet och $1 - p_0 > p$ i det superkritiska fallet.

Upprepad applicering [redigera | redigera wikitext]

Om vi betraktar en kvot av två linjära funktioner är det tydligt att den har samma form även vid upprepad sammansättning $g(g(\dots g(x)\dots)$ eftersom

$$\frac{a + b\frac{a+bx}{c+dx}}{c + d\frac{a+bx}{c+dx}} = \frac{ac + adx + ab + b^2x}{c^2 + cdx + ad + bdx} = \frac{ac + ab + (ab + b^2)x}{c^2 + ad + (cd + bd)x} = \frac{a' + b'x}{c' + d'x}.$$

Då LF-fördelningen kan skrivas som en kvot av två linjära funktioner gäller det att sammansättningen också är en LF-fördelning. LF-fördelningen upprepad n gånger kan alltså skrivas som[1]

$$f_n(s) = f(f(\dots f(s)\dots)) = p_0^{(n)} + (1 - p_0^{(n)})\frac{p^{(n)}s}{1 - (1 - p^{(n)})s}.$$

Från egenskaper av sannolikhetsgenererande funktioner och koppling till Galton-Watson-processen kan vi få ett uttryck av f_n i bara p_0, p. Parametrarna $p^{(n)}, p_0^{(n)}$ i f_n kan uttryckas som

$$\begin{cases} p^{(n)} = \dfrac{1 - \frac{p_0}{1-p}}{\left(\frac{1-p_0}{p}\right)^n - \frac{p_0}{1-p}}, \\ p_0^{(n)} = \dfrac{1 - \left(\frac{p}{1-p_0}\right)^n}{\frac{1-p}{p_0} - \left(\frac{p}{1-p_0}\right)^n}, \end{cases}$$

där p och p_0 är parametrarna i LF-fördelningen f. I det kritiska fallet, då $p = 1 - p_0$ gäller specifikt

$$\begin{cases} p^{(n)} = \dfrac{p}{p + (1 - p)n}, \\ p_0^{(n)} = \dfrac{(1 - p)n}{p + (1 - p)n}. \end{cases}$$

Dessa uttryck är användbara för analys av Galton-Watson-processen.

Referenser [redigera | redigera wikitext]

1. ^ [a b c d] Athreya, Krishna B. (1972). *Branching Processes*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-65371-1. ISBN 978-3-642-65373-5. Läst 8 april 2020
2. ^ Olofsson, Peter, 1963- (2012). *Probability, statistics, and stochastic processes* (2nd ed). Wiley. ISBN 978-1-118-23129-6. OCLC 795795373. Läst 8 april 2020
3. ^ [a b] Dobrow, Robert P.,. *Introduction to stochastic processes with R*. ISBN 978-1-118-74072-9. OCLC 922799569. Läst 26 mars 2020

Kategori: Sannolikhetsfördelningar