Analys av försäljningsdata från Google Adwords med EM-algoritmen

Erik Bertilsson
David Tobjörk
23-06-2016

Kandidatuppsats, 15 ECTS
Statistiska institutionen, Lunds universitet
Handledare: Peter Gustafsson

Abstract

The purpose of this paper is to find a model that describes the distribution of Google Adwords order data and to find an appropriate method to estimate the expected value for an order. We do this on behalf of the marketing agency Precis Digital. We receive data for one of their customers. To begin with we fit some widely used probability distributions using the Maximum-Likelihood method and investigate how well these distributions describe the empirical data. We assume that our data is not derived from only one probability distribution. Therefore we go further by fitting a mixture of normal distributions with the EM-algorithm and see how well it fits the data. We evaluate the previous single distributions and the mixed distributions using Bayesian information criterion. It turns out that a mixture model fit the data quite well and better than previous single distributions. We evaluate how well mixed models describe data with few observations by using the EM-algorithm on individual campaigns. We come up with the conclusion that mixture distributions are good at describing the distribution of campaigns with few observations. However the estimated expected value of an order does not differ much from an estimation done with the mean value. Another interesting result is that subpopulations of customers probably exist since a mixed model better fits the data. In further research it would be interesting to deeper investigate these subpopulations. These subpopulations probably represent different customers with certain buying behavior. By further investigating these subpopulations campaigns can be optimized.

Sammanfattning

Syftet med uppsatsen är att finna en modell som väl beskriver fördelningen för orderdata från Google Adwords och att finna ett lämpligt sätt att skatta det förväntade värdet för en order. Vi gör detta på uppdrag av marknadsföringsbyrån Precis Digital. Från dem erhåller vi data för en av deras kunder. Till att börja med anpassar vi ett antal olika sannolikhetsfördelningar med Maximum-Likelihood-metoden och undersöker hur väl de beskriver det empiriska datamaterialet. Vi ser tecken på att data inte härstammar från endast en sannolikhetsfördelning. Därmed går vi vidare i att undersöka hur väl en blandning av flera normalfördelningar beskriver data. Vi utvärderar de tidigare enskilda fördelningarna och de blandade fördelningarna med Bayesian information criterion. Det visar sig att blandade fördelningar bättre beskriver data än vad enbart en fördelning gör. Vi prövar hur väl blandade fördelningar beskriver mindre data genom att använda EM-algoritmen på enskilda kampanjer. Vi konstaterar här att blandade fördelningar är ett bra sätt att beskriva kampanjer med färre data. Däremot skiljer sig inte det skattade förväntade värdet för en order särskilt mycket från en skattning gjord med det aritmetiska medelvärdet. Ett annat intressant resultat är att subpopulationer av kunder förmodligen föreligger då en blandfördelning bättre beskriver data. Det vore intressant att i vidare studier djupare undersöka dessa subpopulationer, då de troligtvis representerar olika kundgrupper med olika köpbeteenden. Genom att undersöka dessa subpopulationer kan man bättre anpassa sina kampanjer.

Innehållsförteckning

1. Introduktion 1
1.1 Inledning 1
1.2 Syfte 2
1.3 Avgränsningar 2
1.4 Disposition 2
2. Metod 3
2.1 Datamaterial 3
2.2 Tillvägagångssätt 4
3. Teori 6
3.1 Google Adwords 6
3.2 Teoretiska fördelningar 7
3.3 Maximum-Likelihood 7
3.4 Blandfördelningar 8
3.5 Bayesian Information Criterion 10
4. Resultat 11
4.1 Produktomsättning 11
4.2 Ordervärde 12
4.3 Blandfördelningar 14
4.3.1 Validering av blandfördelningar 16
5. Analys 19
5.1 En fördelning mot empirisk data 19
5.2 Blandfördelningar mot empirisk data 19
5.3 Förväntat ordervärde 21
6. Slutsats 22
Källförteckning 23
Bilaga I

1. Introduktion

Här ges först bakgrunden till det problemområde som behandlas och varför det är intressant. Dessutom presenteras vår uppdragsgivare, syftet med studien, de avgränsningar som görs och disposition för uppsatsen.

1.1 Inledning

Google är idag den särklass största sökmotorn och 2012 gjordes det så många som 1 200 miljarder sökningar med Google (Internet Live Stats, 2016). Dessa sökningar bidrar till en stor mängd intressanta data. Dessa data utnyttjar Google i sitt annonssystem *Google Adwords*. Detta verktyg kan företag använda för att synas när en datoranvändare gör en viss sökning. Bland dessa sökningar kan företag välja vilka sökord som deras annons skall synas för. Annonssystemet är ett sätt att marknadsföra sig. Man får också data på de potentiella kunder som följer annonslänken. Google har alltså skapat en möjlighet att marknadsföra sig digitalt där man direkt kan rikta sig mot kunder som själva visat intresse för varan eller tjänsten.

Tidigare har marknadsföring främst skett via flygblad, tidningar samt radio och tv-sändningar men med internet har nya sätt att marknadsföra sig blivit tillgängliga. Dessa nya sätt att marknadsföra sig på är något som företag inser fördelarna med och utnyttjar i allt större utsträckning. När konsultbolaget Accenture frågade marknadsföringschefer svarade 37 % att de trodde att mer än 75 % av marknadsföringsbudgeten skulle gå till digital marknadsföring inom de närmaste 5 åren (Accenture, 2014).

En fördel med att marknadsföra sig via *Google Adwords* är att man får information om vilka som klickar på annonsen och hur mycket dessa kunder köper för. Denna information kan sedan användas för att utvärdera annonser och förbättra dem. De som använder sig av *Google Adwords* måste bestämma sig för hur annonsen skall se ut och hur mycket de är beredda att betala för den. Som underlag för dessa beslut kan data man har från tidigare *Google Adwords*-annonser användas. Genom att analysera data man får från Google kan slutsatser om kundernas köpbeteende dras och företaget kan maximera avkastningen på den investering de gör i marknadsföringen. För att maximera avkastning på annonserna behöver företaget veta vilka nyckelord de skall investera i och hur mycket de är beredda att betala för dessa. Avgörande vid beslut om hur mycket man är villig att betala för en annons är storleken på den förväntade intäkten. Att bestämma den förväntade intäkten kan dock vara problematisk då intäkterna kan variera mycket. Det kan också finnas olika subpopulationer av kunder vars köpbeteende inom en subpopulation är likt men där det skiljer sig åt mellan dem.

1.2 Syfte

Den digitala marknadsföringsbyrån **Precis Digital** har bett oss att ta fram ett lämpligt sätt att skatta det förväntade ordervärdet för kunderna som klickat på Adwords annonser. Det förväntade ordervärdet är väntevärdet för dessa kunders beställningar i kronor. Syftet med denna uppsats är att besvara frågorna:

- Hur ser en bra modell ut för att beskriva sannolikhetsfördelningen för ordervärde?
- Hur skattas ordervärdet på bästa sätt för det företag vi studerar?

1.3 Avgränsningar

För att besvara frågeställningen används data från en av vår uppdragsgivares kunder. Att undersöka data från fler kunder hade varit intressant, men hade inneburit ett för omfattande arbete.

1.4 Disposition

Kapitel 1 – I det inledande kapitlet introduceras den nya formen av marknadsföring i form av digital marknadsföring. Därefter presenteras syftet med studien och vilka avgränsningar som gjorts.

Kapitel 2 – I det andra kapitlet presenteras den metod som används för att besvara studiens frågeställningar.

Kapitel 3 – Detta kapitel presenterar den statistiska teori som används i studien.

Kapitel 4 – Här presenteras resultatet för studien. Först presenteras resultat från analys av köp av enskilda produkter. Sedan presenteras resultatet för analysen av hela order.

Kapitel 5 – I detta kapitel diskuteras och analyseras de resultat som presenterats i det föregående kapitlet.

Kapitel 6 – Slutligen presenteras slutsatser för studien och förslag på vidare studier ges.

2. Metod

I detta avsnitt beskriver vi först det datamaterial vi har och hur detta bearbetas. Sedan presenteras hur vi går tillväga för att besvara studiens frågeställning.

2.1 Datamaterial

Vår uppdragsgivare Precis Digital har gett oss datamaterial för en av deras kunders Google Adwords-konton. Då denna typ av information är av känslig karaktär kan vi inte avslöja information om kundens identitet. Tabell 2.1 nedan beskriver de variabler den ursprungliga datafilen innehåller.

Tabell 2.1.1: Variabler i det ursprungliga datamaterialet.

Variabel	Definition
KampanjID	Anger kampanjen numeriskt.
GruppID	Anger Annonsgrupp numeriskt.
NyckelordsID	Anger nyckelord numeriskt.
Datum	Anger tidpunkt för en transaktion i dag och timme.
TransaktionsID	Anger transaktion numeriskt.
ProduktID	Anger vilken produkt som köpts numeriskt.
Produktnamn	Namn på den produkt som köpts.
Produktomsättning	Anger intäkten i SEK för en specifik produkt vid en order.
Produktantal	Anger hur många enheter som köpts av en produkt.
Kampanj	Namn på kampanj.
Grupp	Namn på annonsgrupp.
Nyckelord	Nyckelord.
Matchningsnivå	Anger på vilken nivå sökordet matchats med nyckelordet.

Analys och databearbetning har gjorts med programmeringsspråket R. Till R finns en stor mängd så kallade paket tillgängliga, dessa innehåller diverse funktioner och statistiska verktyg. I denna studie har vi använt oss av paketen *dplyr*, *MASS* och *mixtools*. Paketet *dplyr* tillhandahåller verktyg för att manipulera stora datamängder. *MASS* kan bland annat användas för att utföra Maximum-Likelihood-skattningar och *mixtools* är ett paket med diverse verktyg för att analysera statistiska modeller bestående av mer än en fördelning.

Den ursprungliga datafilen består av 321 663 rader, här görs ett antal modifieringar. Till en början konstaterar vi att det finns rader i datafilen som behöver sorteras bort eftersom de saknar värde för en eller flera av variablerna. Detta kan exempelvis vara transaktioner som gjorts utanför Google Adwords och som vi alltså inte har särskilt mycket användning av. Det finns även transaktioner där en kund köpt en produkt för noll kronor som också sorteras bort, dessa kan komma från exempelvis rabattkuponger. Efter att dessa tagits bort återstår 54 177 rader. Vi behöver även göra om vissa variabler i R till numeriska och andra till kategoriska. I studien läggs mycket fokus på hur stor en order är, men även hur mycket en kund köper av en specifik produkt analyseras. För att skilja dessa två åt benämns den första *Ordervärde* (SEK) och den senare *Produktomsättning* (SEK). För att möjliggöra analys av Ordervärde skapar vi en ny datafil som innehåller den nya variabeln *Ordervärde*. Den nya datafilen skapas genom att summera *Produktomsättning* för alla köp inom en order, det vill säga de produktköp som har samma *TransaktionsID*. Den slutliga datafilen består av 24 719 order.

2.2 Tillvägagångssätt

För att besvara vår frågeställning använder vi oss av följande tillvägagångssätt. Vi börjar med att undersöka ursprunglig data, det vill säga omsättningen för respektive produkt, *Produktomsättning*. Detta gör vi först för att försöka förstå hur det i grunden ligger till och om det finns något intressant här att beakta. Här undersöker vi fördelningen för *Produktomsättning*, vi studerar den empiriska fördelningen och ser till att allting har gått rätt till vid datainläsningen i R. Vi bekräftar detta genom att studera den ursprungliga datafilen i Excel. Till den empiriska fördelningen anpassar vi ett antal olika fördelningsfunktioner: Normal-, Lognormal-, Weibull- och Gammafördelning.

Nästa del och den som huvudsakligen besvarar syftet är att undersöka hela order. Det vill säga variabeln *Ordervärde* där produktköpen för en viss order är summerade. Här undersöker vi till en början, på samma sätt som för produktköpen, den empiriska fördelningen. Vi anpassar ovan nämnda fördelningar till data och ser till att allting har gått rätt till vid summeringen av *TransaktionsID:n*. Därefter överväger vi om fördelningen kan beskrivas bättre med en viktad fördelning av flera normalfördelningar. Enligt Aitkin & Rubin (1985) är den underliggande idén med blandfördelningar att det föreligger två eller flera subpopulationer med gemensam fördelningsform men med olika parametrar. Att undersöka blandfördelningar är alltså intressant då datamaterialet potentiellt kommer från flera olika

subpopulationer av kunder. En blandning av fördelningar skulle då kunna beskriva datamaterialet bättre än vad en fördelning gör. För att jämföra de olika fördelningarna som anpassats studeras Bayesian information criterion för respektive modell. Ett högt antal parametrar i modellen ökar BIC-värdet, en enkel modell föredras alltså av BIC (Sheather, 2009). BIC motiveras av att vi inte vill ha en komplicerad modell om det inte är nödvändigt. I praktiken kan det vara så att Adwords kampanjer skiljer sig åt och att det finns lite data för en viss kampanj. Därför valideras och undersöks sedan en blandfördelning för enskilda kampanjer. Först jämför vi olika blandfördelningar med empirisk data för de fyra kampanjer med flest observationer och därefter med empirisk data för fyra slumpade kampanjer med observationer bestående av 30 till 100 observationer. Att vi slumpar kampanjer med 30-100 observationer motiverar vi med att vi inte vill ta kampanjer med alldeles för få observationer då dessa inte är särskilt vanliga och de kan vara opålitliga. Företaget påstår också att de i flesta fall tittar på kampanjer med lite fler köp. För att skatta parametrarna och anpassa teoretiska fördelningar till datamaterialet använder vi oss av Maximum-Likelihood-metoden.

3. Teori

I detta avsnitt beskrivs hur Googles annonssystem Google Adwords fungerar. Därefter beskrivs de fördelningsfunktioner som studeras, skattning av parametrar enligt Maximum-Likelihood-metoden, blandfördelningar och slutligen Bayesian information criterion.

3.1 Google Adwords

Google Adwords är ett annonssystem där en annonsörs annons syns efter att kunderna gjort vissa sökningar. Idag finns över en miljon företag som använder sig av Google Adwords. Annonserna syns på Google men även på andra sidor kopplade till Google. Annonsören betalar när kunderna klickar på annonsen, besöker webbsidan eller ringer till säljaren. Denna tjänst är möjlig att skräddarsy för att annonsören ska nå sina önskemål i sin marknadsföring. Oavsett om man vill få fler besökare till sin hemsida, öka försäljningen eller få tillbaka kunder kan man ta hjälp av denna tjänst. Det är möjligt att bestämma om man vill annonsera globalt eller lokalt, efter ordagranna sökningar eller synonymer och en budget för annonserna kan sättas (Google, 2015a). Det kostar ingenting att registrera sig och budgeten är flexibel och för klicken kan en budget sättas per dag eller månad (Google, 2015b).

Google Adwords kan sägas bestå av tre delar: annonser, annonsgrupper och kampanjer. Annonser består av nyckelord (keyword), ord eller fraser som får annonsen att synas. Annonser som är lika grupperas sedan in i annonsgrupper. Slutligen skapar flera annonsgrupper en kampanj och en budget för kampanjen sätts (Google, 2015e).

När någon söker på ett nyckelord eller liknande termer kan annonsen dyka upp bredvid Googles sökresultat eller på andra hemsidor i Google Network, till exempel partners till Google som är relaterade till annonsens nyckelord. Google bestämmer också vilka annonser som dyker upp i vilka nivåer. Finns det flera som använder sig av samma nyckelord använder Google något de kallar för Ad rank för att bestämma vilka annonser som skall synas och i vilken ordning. Varje annonsör får en ad rank som är baserad på en kombination av: (1) budet, det vill säga hur mycket man är villig att betala, (2) kvalitén på annonsörens annons och hemsida och (3) förväntad effekt av annonsen. Mer i detalj är budet den maxkostnad per klick annonsören är villig att betala och kvalitén är baserad på: förväntad klickfrekvens, annonsrelevans och historik över tidigare besökare. Det hålls alltså en auktion för varje nyckelord där man sätter ett maxbud. Har man samma maxbud som någon annan är nästa kriterium kvalité och förväntad effekt (Google, 2015c).

Det är också möjligt att använda så kallade negativa nyckelord i en annons. Detta innebär att annonsen inte syns vid sökningar som innehåller det negativa nyckelordet. Detta kan till exempel användas då man vill exkludera sökningar på varor som man inte tillhandahåller (Google, 2015d).

3.2 Teoretiska fördelningar

Normalfördelning
En kontinuerlig slumpvariabel X med täthetsfunktionen:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \qquad (1)$$

där parametrarna μ och σ upfyller $-\infty < \mu < \infty$ och $0 < \sigma < \infty$ sägs vara normalfördelad med parametrarna μ och σ^2 . Detta skrivs kort X är $N(\mu, \sigma^2)$.

Lognormal
En kontinuerlig slumpvariabel X där den naturliga logaritmen av denna slumpvariabel är normalfördelad sägs vara lognormalfördelad.

Gammafördelning
En kontinuerlig slumpvariabel X med täthetsfunktionen:

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}, \quad 0 < x < \infty \qquad (2)$$

sägs vara gammafördelad med parametrarna θ samt α och skrivs X är $\Gamma(\alpha, \theta)$. För gammafördelningen kan slumpvariabeln X tolkas som väntetiden till den α:te förändringen skett i en poissonprocess.

Weibullfördelning
En kontinuerlig slumpvariabel X med täthetsfunktionen:

$$f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, \quad 0 < x < \infty \qquad (3)$$

sägs vara weibullfördelad med parametrarna α och β. Weibullfördelningen kan till exempel beskriva livslängden för en tillverkad vara.

3.3 Maximum-Likelihood

Maximum-Likelihood-metoden skattar parametrarna genom att finna de parametervärden som maximerar sannolikheten att erhålla en viss uppsättning av observationer från en given täthetsfunktion. Låt oss säga att vi har slumpvariabeln Y och vektorn av observationer $\boldsymbol{y} = (y_{1,\dots,}y_n)^T$. Sannolikheten att erhålla en viss observation y_i ges då av täthetsfunktionen $f(y_i|\theta)$ med parameter θ. Om observationerna är oberoende ges då sannolikheten för att erhålla hela serien med observationer av:

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \qquad (4)$$

När ovan anges som en funktion av θ kallas denna Likelihood-funktionen $L(\theta)$.

ML-skattningen av parametern θ fås genom att maximera Likelihood-funktionen. Dock är det ofta lättare att maximera den logaritmerade Likelihood-funktionen (5). Detta kan göras då logaritmen är en strängt växande funktion och därmed har den logaritmerade Likelihood-funktionen samma maximum.

$$log\{L(\theta)\} = log\{f(\boldsymbol{y}|\theta)\} = log\{\prod_{i=1}^{n} f(y_i|\theta)\} = \sum_{i=1}^{n} log\{f(y_i|\theta)\}. \quad (5)$$

(Blume, 2002).

3.4 Blandfördelningar

Blandfördelningar kan användas för att ta itu med praktiska problem så som inferens vid förekomst av uteliggare. Idén bakom dessa blandfördelningar är att data i vektorn $X = [x_1, \dots, x_n]$ med stickprovsstorlek n uppstår från två eller fler underliggande subpopulationer med gemensam fördelningsform men olika parametrar. Ett vanligt förekommande fall är att man har en blandning av flera normalfördelningar. Dessa subpopulationer kan vara olika men datapunkter inom samma subpopulation kan vara väl modellerad av en normalfördelning. Då antas alltså att de g olika subpopulationerna (komponenterna) kommer från g olika normalfördelningar (Aitkin & Rubin, 1985).

För blandade fördelningar tänker vi oss att vi har observationerna y_i, där varje observation tillhör en av g fördelningar C. Täthetsfunktionen för den blandade fördelningen för observation y_j är då:

$$p(y_j|\boldsymbol{\theta}) = \sum_{i=1}^{g} \alpha_i p(y_j|C = i, \beta_i) \quad (6)$$

där,

$\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_g, \beta_1^T, \dots, \beta_g^T)^T$.

$\alpha_i =$ är vikten för komponent i.

$\beta_i =$ är parametervektorn för täthetsfunktionen för komponent i. I vårt fall studerar vi blandade normalfördelningar och har alltså väntevärdet μ och variansen σ^2 som parametrar.

Givet att vi har erhållit observationerna $\boldsymbol{y} = (y_1, \dots, y_n)^T$ är målet med Maximum Likelihood att skatta $\boldsymbol{\theta}$ genom att maximera sannolikheten att erhålla dessa observationer. Alltså att maximera:

$$log\{L(\boldsymbol{\theta})\} = \sum_{j=1}^{n} log\left\{\sum_{i=1}^{g} \alpha_i p(y_j|C = i, \beta_i)\right\} \quad (7)$$

Detta är dock svårt då logaritmen är av en summa, som lösning på detta problem kan *Expectation Maximization* (EM)-algoritmen användas. Detta är möjligt genom att introducera okända variabler Z med vektorn av okända data $\boldsymbol{z}$. I fallet med blandade fördelningar kan $\boldsymbol{z}$ med parametrar z_{ij} vara sådan att z_{ij} är ett om observation j kommer från komponent i, annars noll. Fullständig data bestående av både observerad och okänd data är:

$$x = (y^T, z^T)^T \quad (8)$$

Nu studerar vi istället Loglikelihood-funktionen för fullständiga data:

$$log\{L_C(\boldsymbol{\theta})\} = log\{p(x|\boldsymbol{\theta})\} = \sum_{j=1}^{n} log\{\sum_{i=1}^{g} z_{ij}\alpha_i p(y_j|C = i, \beta_i)\} \quad (9)$$

Då z_{ij} är noll för alla utom en term i summan kan (9) nu skrivas:

$$log\{L_C(\boldsymbol{\theta})\} = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij} log\{\alpha_i p(y_j|C = i, \beta_i)\} \quad (10)$$

Istället för att endast maximera Loglikelihood-funktionen, består EM även av det så kallade expectation-steget. Detta steg innebär att väntevärdet $E\left[\log\{L_C(\boldsymbol{\theta})\middle|y, \boldsymbol{\theta}^{(i)}\}\right]$ beräknas. För att kunna utföra detta steg krävs att ett startvärde väljs för parametrarna $\boldsymbol{\theta}^{(i)}$ vilket uttrycket beror på. I det andra steget fås nya parametrar $\boldsymbol{\theta}^{(i+1)}$ genom att maximera hela uttrycket. Denna procedur upprepas till dess att något stop kriterium uppfylls, exempelvis att förändringen mellan nya parametrar blir väldigt liten. EM består alltså av två steg, E-steget och M-steget, vilka beskrivs nedan.

De uppdaterade parametrarna fås genom att lösa:

$$\boldsymbol{\theta}^{(i+1)} = \underset{\boldsymbol{\theta}}{argmax}\, E\left[log\{L_C(\boldsymbol{\theta})\middle|y, \boldsymbol{\theta}^{(i)}\}\right] \quad (11)$$

$$= \underset{\boldsymbol{\theta}}{argmax}\, E\left[\sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij} log\{\alpha_i p(y_j|C = i, \beta_i)\,\middle|y, \boldsymbol{\theta}^{(i)}\}\right]$$

E-steget

På grund av väntevärdets linjära egenskaper kan (11) skrivas:

$$\boldsymbol{\theta}^{(i+1)} = \underset{\boldsymbol{\theta}}{argmax}\left[\sum_{j=1}^{n}\sum_{i=1}^{g} E\left[z_{ij}\middle|y, \boldsymbol{\theta}^{(i)}\right] log\{\alpha_i p(y_j|C = i, \beta_i)\}\right] \quad (12)$$

där vi har att:

$$E\left[z_{ij}\middle|y,\boldsymbol{\theta}^{(i)}\right] = 0\cdot p\left(z_{ij}=0\middle|\boldsymbol{\theta}^{(i)}\right) + 1\cdot p\left(z_{ij}=1\middle|\boldsymbol{\theta}^{(i)}\right) = p\left(z_{ij}=1\middle|\boldsymbol{\theta}^{(i)}\right) \quad (13)$$

I ekvation (13) är sannolikheten att observation j kommer från komponent i.

$$P\left(z_{ij}=1\right) = p(C=i\middle|x_j,\boldsymbol{\theta}^{(i)}) \quad (14)$$

Använder vi nu oss av Bayes sats får vi:

$$P\left(C=i\middle|x_{j,}\boldsymbol{\theta}^{(i)}\right) = \frac{p(x_j\middle|C=i,\boldsymbol{\theta}^{(i)})p(C=i\middle|\boldsymbol{\theta}^{(i)})}{p(x_j\middle|\boldsymbol{\theta}^{(i)})} \quad (15)$$

där $p(x_j\middle|C=i,\boldsymbol{\theta}^{(i)})$ erhålls från täthetsfunktionen för komponent i, $p(C=i\middle|\boldsymbol{\theta}^{(i)})$ är α_i, sannolikheten för komponent i, vilken redan är given av $\boldsymbol{\theta}^{(i)}$. Slutligen har vi $p(x_j\middle|\boldsymbol{\theta}^{(i)})$ som är täthetsfunktionen för hela fördelningen. Med $\boldsymbol{\theta}^{(i)}$ är alla dessa lätta att räkna ut.

M-steget
I detta steg maximerar vi enligt (12). Detta görs genom att sätta de partiella derivatorna för α_i respektive β_i till noll (komponenterna för $\boldsymbol{\theta}$). Vid maximeringen har vi villkoret för vikterna att $\sum_{i=1}^{g}\alpha_i = 1$ (vilket kan göras med hjälp av Lagrangemultiplikator) (Blume, 2002).

3.5 Bayesian Information Criterion

I uppsatsen utvärderar vi olika modeller till data genom att jämföra Bayesian Information Criterion (BIC) för modellerna. Modellen med det lägsta BIC-värdet är att föredra. BIC definieras enligt:

$$BIC = -2\,log(L) + p\cdot log(n) \quad (16)$$

där,

$L =$ likelihooden för modellen,

$p =$ antalet skattade parametrar i modellen,

$n =$ antalet observationer,

(Zucchini, 2000).

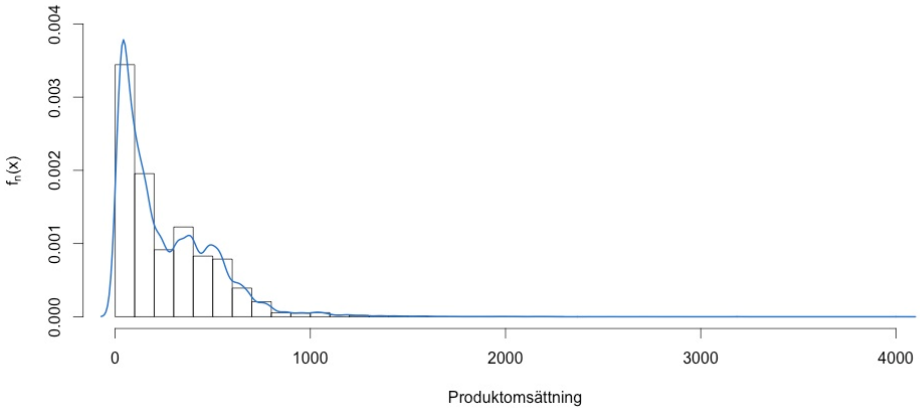
4. Resultat

I detta avsnitt redovisas resultat för studien. Till en början redovisas resultat för ursprunglig data, därefter går vi vidare till att presentera modifierad data, det vill säga för Ordervärde som är centralt för att besvara syftet. Slutligen presenteras hur blandfördelningar kan beskriva data för Ordervärde och hur dessa modeller fungerar på mindre datamängder.

4.1 Produktomsättning

För att få en bättre förståelse för det datamaterial som vi studerar beskrivs till en början ursprungliga data. I den ursprungliga datafilen visas hur mycket en kund köpt av en specifik produkt, detta är vad vi kallar *Produktomsättning*. Nedan presenterar vi deskriptiv statistik för materialet och undersöker hur ett antal olika teoretiska fördelningar skulle kunna beskriva ursprunglig data med hjälp av Maximum-Likelihood-skattningar.

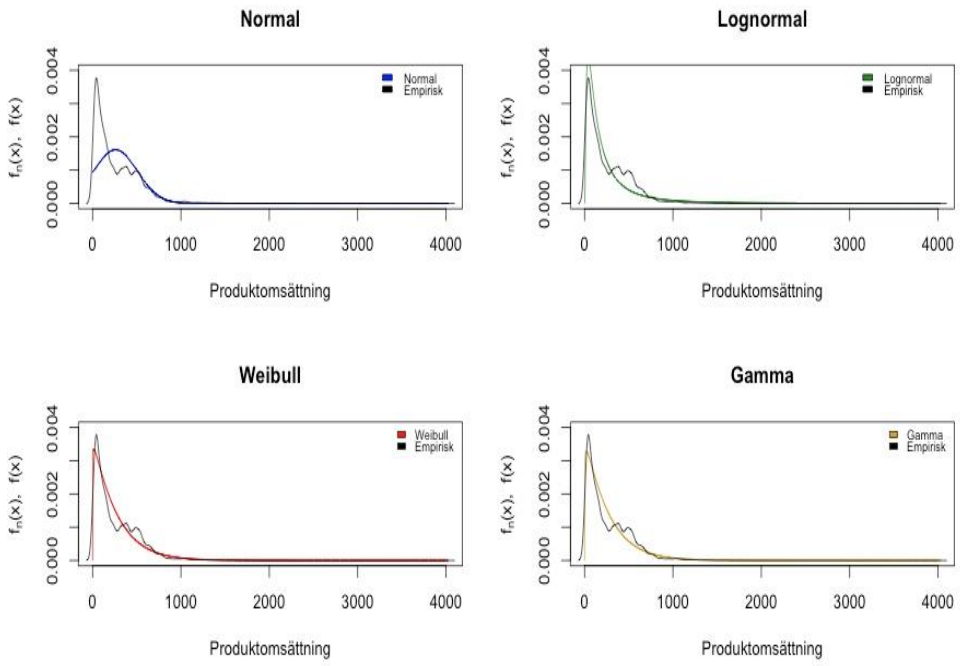
Vi börjar med att beskriva *Produktomsättning* med ett histogram för den empiriska fördelningen $f_n(x)$ (figur 4.1.1). I figuren syns även en kurva för fördelningen. Denna kurva är gjord i R med funktionen *lines* som gör en kärnskattning och bandbredden är enligt standard i R. Materialet består av totalt 54 177 observationer och det aritmetiska medelvärdet är 259,51 kronor och standardavvikelsen är 247,26 kronor. Det minsta värdet är 4 kronor och det högsta är 4 024 kronor.



Figur 4.1.1: Histogram för produktomsättning (SEK).

Vi anpassar ett antal fördelningar (normalfördelning, lognormalfördelning, weibullfördelning och gammafördelning) till empiriska data. I figur 4.1.2 ser vi hur väl de olika teoretiska fördelningarna $f(x)$ är anpassade till den empiriska fördelningen $f_n(x)$ för *Produktomsättning*. Vid skattningen av parametrarna för dessa anpassade täthetsfunktioner har Maximum-Likelihood-metoden använts. Skattningarna av respektive fördelningsparametrar, väntevärde och standardavvikelse redovisas i tabell 4.1.1. 95 % konfidensintervall för fördelningsparametrarna (se bilaga) har beräknats enligt:

$$Parameterskattning \pm 1{,}96 \cdot medelfel. \qquad (17)$$



Figur 4.1.2: Empirisk fördelning med anpassade teoretiska fördelningar (SEK).

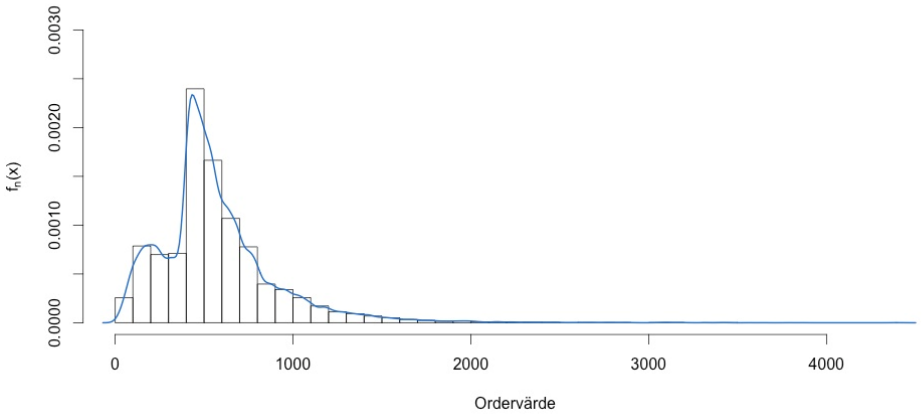
Tabell 4.1.1: Skattade parametrar, väntevärde och standardavvikelse för produktomsättning.

Fördelning	Parameter 1	Parameter 2	$\widehat{\mathbf{E}[\mathbf{X}]}$	$\widehat{\boldsymbol{Std}[\boldsymbol{X}]}$
Normal	$\hat{\mu}$ = 259,5105	$\hat{\sigma}$ = 247,2575	259,5105	247,2575
Lognormal	$log\hat{\mu}$ = 5,0282	$log\hat{\sigma}$ = 1,1438	293,6221	482,4299
Weibull	$\hat{\alpha}$ = 1,0408	$\hat{\beta}$ = 263,7877	259,5825	368,6659
Gamma	$\hat{\alpha}$ = 1,0785	$\hat{\theta}$ = 240,3846	259,2428	249,6356

4.2 Ordervärde

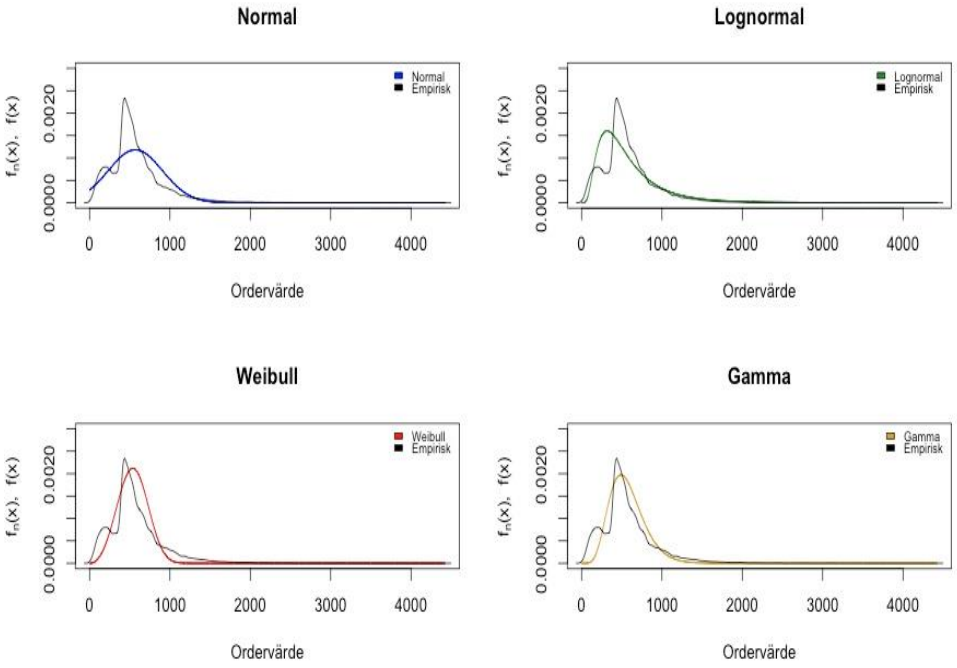
Nedan presenterar vi deskriptiv statistik för *Ordervärde* och undersöker hur ett antal olika teoretiska fördelningar skulle kunna beskriva empirisk data med hjälp av Maximum-Likelihood-skattningar.

Vi börjar med att beskriva *Ordervärde* med ett histogram för den empiriska fördelningen $f_n(x)$, se figur 4.2.1. Materialet består av totalt 24 719 order (observationer) och det aritmetiska medelvärdet är 568,77 kronor, alltså en omsättning på totalt ca 14 mkr. Standardavvikelsen är 337,73 kronor, det minsta värdet är 7,2 kronor och det högsta är 4 425 kronor.



Figur 4.2.1: Histogram för ordervärde (SEK).

Vi anpassar ett antal fördelningar (normalfördelning, lognormalfördelning, weibullfördelning och gammafördelning) till empirisk data. I figur 4.2.2 ser vi hur väl de olika teoretiska fördelningarna $f(x)$ är jämförda med den empiriska fördelningen $f_n(x)$ för *Ordervärde*. Vid skattningen av parametrarna för dessa anpassade täthetsfunktioner har Maximum-Likelihood-metoden (ML) använts. Skattningarna av respektive fördelningsparametrar, väntevärde och standardavvikelse redovisas i tabell 4.2.1. 95 % konfidensintervall för fördelningsparametrarna (se bilaga) har beräknats enligt (17).



Figur 4.2.2: Empirisk fördelning med anpassade teoretiska fördelningar (SEK).

Tabell 4.2.1: Skattade parametrar, väntevärde och standardavvikelse för ordervärde.

Fördelning	Parameter 1	Parameter 2	$\widehat{\mathbf{E}[\mathbf{X}]}$	$\widehat{\boldsymbol{Std}[\boldsymbol{X}]}$
Normal	$\hat{\mu}$ = 568,77	$\hat{\sigma}$ = 337,72	568,77	337,72
Lognormal	$log\hat{\mu}$ = 6,17	$log\hat{\sigma}$ = 0,64	586,87	417,55
Weibull	$\hat{\alpha}$ = 3,28	$\hat{\beta}$ = 600,45	538,42	180,64
Gamma	$\hat{\alpha}$ = 7,00	$\hat{\theta}$ = 83,33	583,31	220,47

4.3 Blandfördelningar

Vi anpassar en blandning av fördelningar till empirisk data för *Ordervärde* med EM-algoritmen (se kapitel 3.4 för beskrivning). Vi testar till att börja med att anpassa två respektive tre gammafördelningar, men då dessa inte passar den empiriska fördelningen väl fortsätter vi med att endast undersöka en blandning av normalfördelningar. Vi anger g=3 (tre normalfördelningar) som startvärde för algoritmen (vi testade även två och fyra normalfördelningar, dessa gav dock ett sämre resultat). Denna metod undersöker först vikterna för respektive normalfördelning (α_i) och därefter dess respektive parametrar (μ_i, σ_i^2). När iterationerna har konvergerat får vi följande värden i tabell 4.3.1 för g=3 (tre normalfördelningar).

Tabell 4.3.1: Vikter och parametrar för blandfördelning.

Fördelning	$\boldsymbol{\alpha_i}$	$\boldsymbol{\hat{\mu}_i}$	$\boldsymbol{\hat{\sigma}_i}$
Normalfördelning 1	0,103	151	58,2
Normalfördelning 2	0,69	511	157,6
Normalfördelning 3	0,207	969	456,2

Konfidensintervall för parametrarna finns att se i bilaga. Konfidensintervall för $\hat{\mu}_i$ har beräknats enligt:

$$\hat{\mu}_i \pm z_{\alpha/2} \frac{\hat{\sigma}_i}{\sqrt{n_i}} \tag{18}$$

Konfidensintervall för $\hat{\sigma}_i$ har beräknats enligt:

$$\left[\sqrt{\frac{(n_i - 1)}{b_i}} \hat{\sigma}_i, \sqrt{\frac{(n_i - 1)}{a_i}} \hat{\sigma}_i \right] \tag{19}$$

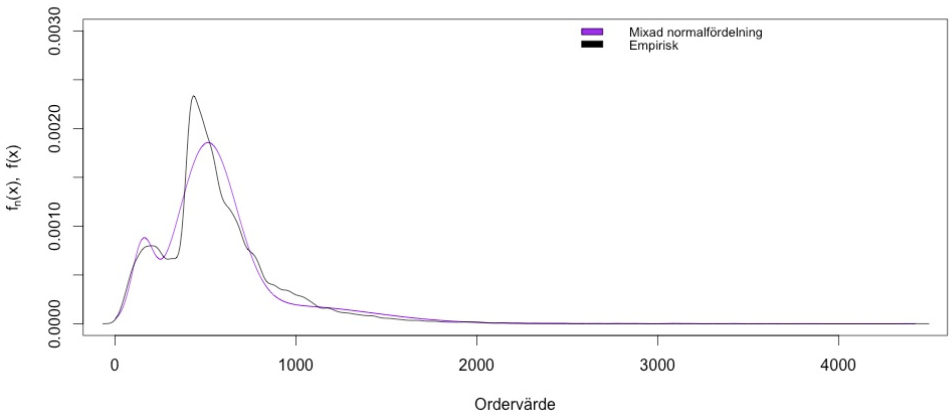
där,

$$a_i = \chi^2{}_{1-\alpha/2}(n_i - 1) \tag{20}$$

$$b_i = \chi^2{}_{\alpha/2}(n_i - 1) \tag{21}$$

där n_i är antalet observationer som tilldelas respektive normalfördelning och α är den valda signifikansnivån.

Därefter plottar vi den blandade fördelningen mot empirisk data för att se hur väl den fungerar som fördelning, se figur 4.3.1.



Figur 4.3.1: Empirisk fördelning och blandning av normalfördelningar $(g = 3)$ *(SEK).*

Skattningen av det förväntade *Ordervärdet*, uträknat från en blandning av g antal normalfördelningar, blir alltså det viktade medelvärdet (22):

$$E[\widehat{Ordervärde}] = \sum_{i=1}^{g} \alpha_i \hat{\mu}_i \qquad (22)$$

För värden från tabell 4.3.1 får vi alltså att skattningen av det förväntade *Ordervärdet*:
$0{,}103 \cdot 151 + 0{,}69 \cdot 511 + 0{,}207 \cdot 969 = 568{,}73$

För att ge en bättre bild av fördelningen ges percentiler nedan i tabell 4.3.2 för den blandade fördelningen. Percentilerna får vi genom att integrera över täthetsfunktionen för den blandade fördelningen. I 90:e percentilen är medelvärdet 1 337,64 kronor och intäkterna i denna percentilen motsvarar ca 21,5% av de totala intäkterna.

Tabell 4.3.2: Percentiler för blandad normalfördelning.

Percentil	**Ordervärde**
10	189,5
25	370
50	521,5
75	680,5
90	994

4.3.1 Validering av blandfördelningar

Vi har anpassat normal-, lognormal, gamma- och weibullfördelning samt en blandning av normalfördelningar till *Ordervärde.* Som jämförelse mellan blandfördelningen och de övriga fördelningarna använder vi oss av BIC. BIC-värde för respektive fördelning presenteras i tabell 4.3.3 nedan.

Tabell 4.3.3: BIC-värden för testade fördelningar.

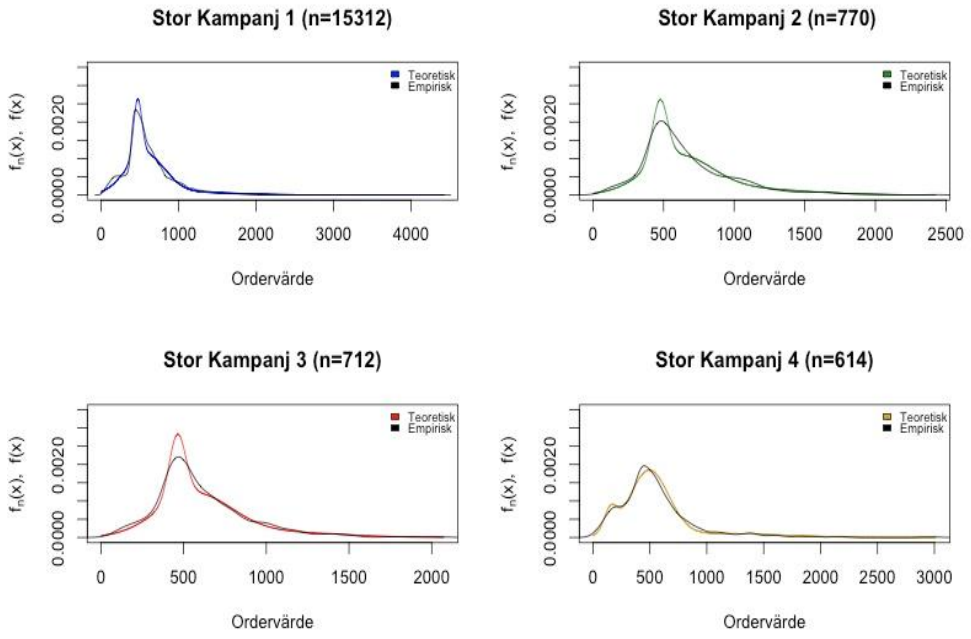
Fördelning	**BIC-värde**
Blandning	349 043,04
Gamma	350 648,83
Weibull	351 735,63
Lognormal	353 057,83
Normal	358 009,03

Vi vill vidare säkerställa hur en blandning av normalfördelningar fungerar i praktiska sammanhang där man inte har ett helt år av data för order som gjorts via Google Adwords, i vårt fall 24 719 observationer. Därmed validerar vi blandade fördelningar på olika kampanjer där observationerna inte är lika många.

Vi börjar med att undersöka hur väl blandfördelningar fungerar på de fyra kampanjer med flest observationer. I tabell 4.3.4 ser vi att en den största kampanjen innehåller betydligt fler observationer än de övriga tre och alltså är en stor del av det totala datamaterialet. I tabell 4.3.4 ges också det aritmetiskt medelvärde ($\bar{x}$) och standardavvikelse (s) för dessa kampanjer. Vi använder även här EM-algoritmen på data från de fyra olika kampanjerna. Algoritmen ger oss vikter och parametrar för respektive fördelning som blandfördelningarna består av. När iterationerna har konvergerat får vi fyra olika blandade fördelningar, alla bestående av tre normalfördelningar (g=3). Dessa blandfördelningar plottas mot empirisk fördelning för respektive kampanj (figur 4.3.2).

Tabell 4.3.4: Antal observationer, medelvärde och standardavvikelse för stora kampanjer.

Kampanj	**Antal**	$\bar{x}$	s
Kampanj 1 (Stor)	15 312	611,9348	340,9935
Kampanj 2 (Stor)	770	658,4561	327,5893
Kampanj 3 (Stor)	712	609,1197	294,0279
Kampanj 4 (Stor)	614	554,0507	358,5783



Figur 4.3.2: Empirisk fördelning och blandning av normalfördelningar för stora kampanjer (SEK).

Vidare skattar vi det förväntade *Ordervärdet* för varje kampanj från de blandade fördelningarna. Detta görs enligt formel 4.3.1. För konfidensintervall för de skattade ordervärden, se bilaga. Därefter jämför vi aritmetiskt medelvärde för varje kampanj (tabell 4.3.4) med respektive skattade förväntade värde för att se hur stor skillnaden är mellan dessa två skattningar. Se nedan i tabell 4.3.5.

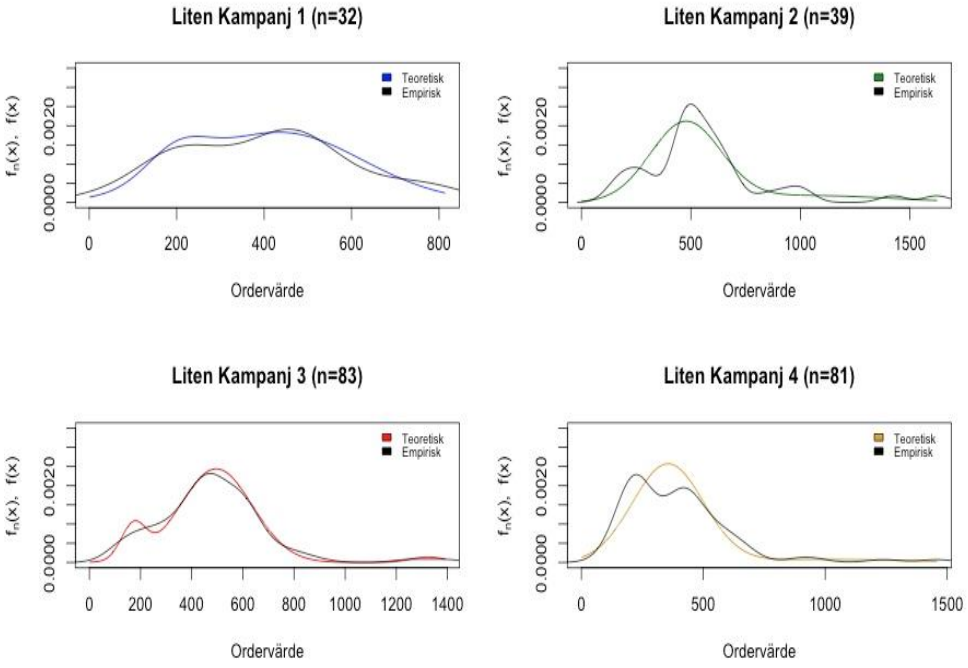
Tabell 4.3.5: Skattning av förväntat ordervärde och differens med medelvärdet.

Kampanj	$\boldsymbol{E[\widehat{Ordervärde}]}$	$\boldsymbol{\left\lvert E[\widehat{Ordervärde}] - \bar{x}\right\rvert}$
Kampanj 1 (Stor)	611,214	0,7208 kronor
Kampanj 2 (Stor)	657,828	0,6281 kronor
Kampanj 3 (Stor)	608,719	0,4007 kronor
Kampanj 4 (Stor)	553,92	0,1307 kronor

Slutligen vill vi också validera hur väl blandfördelningar fungerar på kampanjer med ännu färre observationer. Detta gör vi genom att validera dessa modeller på fyra slumpmässigt valda kampanjer med 30 till 100 observationer. I tabell 4.3.6 nedan ges aritmetiskt medelvärde ($\bar{x}$) och standardavvikelse (s) för dessa kampanjer. Här använder vi EM-algoritmen på data från de fyra olika mindre kampanjerna. Algoritmen ger oss vikter och parametrar för respektive fördelning som modellerna består av. När iterationer har konvergerat får vi fyra blandade fördelningar, där kampanj 1,2 och 4 består av två normalfördelningar (g=2). Kampanj 3 består däremot av tre normalfördelningar (g=3) likt de större kampanjerna i tabell 4.3.4. Dessa modeller plottas mot empirisk fördelning för respektive kampanj (figur 4.3.3).

Tabell 4.3.6: Antal observationer, medelvärde och standardavvikelse för små kampanjer.

Kampanj	**Antal**	$\bar{x}$	s
Kampanj 1 (Liten)	32	403,0312	196,3958
Kampanj 2 (Liten)	39	569,9282	302,7638
Kampanj 3 (Liten)	83	479,9542	214,6122
Kampanj 4 (Liten)	81	394,8414	228,9980



Figur 4.3.3: Empirisk fördelning och blandning av normalfördelningar för små kampanjer (SEK).

Skattningarna för det förväntade *Ordervärdet* från de blandade normalfördelningarna (enligt formel 4.3.1) och differensen mellan denna skattning och medelvärdet ges i tabell 4.3.7 nedan. För konfidensintervall för de skattade ordervärden, se bilaga.

Tabell 4.3.7: Skattning av förväntat ordervärde och differens med medelvärdet.

Kampanj	$\boldsymbol{E[\widehat{Ordervärde}]}$	$\boldsymbol{\left\lvert E[\widehat{Ordervärde}] - \bar{x}\right\rvert}$
Kampanj 1 (Liten)	403,14 kronor	0,1088 kronor
Kampanj 2 (Liten)	569,809 kronor	0,1192 kronor
Kampanj 3 (Liten)	480,1007 kronor	0,1465 kronor
Kampanj 4 (Liten)	395,1435 kronor	0,3021 kronor

5. Analys

I detta avsnitt analyserar vi de resultat som presenterades i föregående avsnitt. Först analyseras produktomsättning som är ursprunglig data. Därefter analyseras hur väl en fördelning samt en blandning av fördelningar beskriver ordervärde. Slutligen diskuteras blandfördelningar för kampanjer med många respektive få observationer och en diskussion förs om det förväntade ordervärdet.

5.1 En fördelning mot empirisk data

När vi tittar på ursprungliga data, det vill säga *Produktomsättningen* i avsnitt 4.1, kan vi konstatera att det här finns en stor spridning. Vi ser i figur 4.1.2 att normalfördelningen inte kan beskriva fördelningen för *Produktomsättning* särskilt väl. De övriga fördelningarna efterliknar data väl och gammafördelningen ser ut att passa ursprungliga data bäst. Detta förklarar vi med att den högra svansen för *Produktomsättning* finns med i denna ML-skattade gammafördelning.

I avsnitt 4.2 presenteras resultat för hela order, *Ordervärde*, som också är centralt för studien. Data är här transformerad från ursprunglig data, där vi har slagit samman alla produktköp med samma *TransaktionsID*. I detta avsnitt har vi som tidigare börjat med att anpassa sannolikhetsfördelningar med hjälp av Maximum-Likelihood-metoden till data för *Ordervärde*. Här ser vi, i figur 4.2.2, att varken en normalfördelning eller en logaritmerad normalfördelning efterliknar den empiriska fördelningen särskilt väl. Weibull- och gammafördelningen passar bättre här, men här är gammafördelningen av samma anledning som i avsnitt 4.1 den som efterliknar data för *Ordervärde* bäst. Av den anledningen kan vi konstatera att gammafördelning är den bästa approximationen för att beskriva *Ordervärde* av de teoretiska fördelningar vi hittills undersökt. Vi utesluter alltså inte att gammafördelningen kan användas för att beskriva data för intäkter från Google Adwords.

5.2 Blandfördelningar mot empirisk data

Däremot ser vi i den empiriska fördelningen för *Ordervärde* att det förekommer "toppar" som endast en fördelning inte kan förklara. Detta kan tyda på att data kommer från en mer komplicerad fördelning och att det föreligger olika subpopulationer i data. För att undersöka dessa "toppar" och troliga subpopulationer närmar vi oss data med en fördelning i form av en blandning av normalfördelningar (figur 4.3.1). Fokus ligger på en blandning av normalfördelningar då en blandning av gammafördelningar inte passade väl. Detta gör vi med hjälp av EM-algoritmen (se avsnitt 3.4) som ger oss värden i form av vikter, väntevärden och standardavvikelser som vi ser i tabell 4.3.1. Själv behöver man ange startvärden för dessa parametrar och antalet normalfördelningar (g) som blandfördelningen skall bestå av. Vi har här valt att endast presentera en blandfördelning bestående av tre normalfördelningar (g=3), då en blandfördelning bestående av två normalfördelningar (g=2) eller fler än tre (g>3) inte gav lika bra resultat i linje med data. Vi konstaterar här att denna fördelning beskriver data för *Ordervärde* bättre än vad de tidigare prövade sannolikhetsfördelningarna

gör då dessa "toppar" samt breda högra svans i empirisk data väl efterliknas av den blandade fördelningen. Vidare kan vi se när vi jämför BIC-värden (tabell 4.3.3) att en fördelning bestående av flera normalfördelningar passar data bäst, då BIC-värdet är lägst för en sådan fördelning. Vi ser också att den, enligt BIC, bästa fördelningen bestående av en fördelning är gamma.

Då den av EM-algoritmen anpassade fördelningen bestående av flera normalfördelningar på ett bättre sätt beskriver hela datamaterialet stödjer detta vår tes om att kunderna kan bestå av olika subpopulationer. Detta stämmer väl överens med Aitkin & Rubin (1985) där den underliggande idén med blandfördelningar är att det föreligger två eller flera subpopulationer med gemensam fördelningsform men olika parametrar. Vi drar slutsatsen att en blandfördelning kan vara lämplig för att beskriva den empiriska fördelningen för *Ordervärde*. Varje subpopulation kan beskrivas väl av en normalfördelning trots att det för empirisk data som helhet inte passade med en normalfördelning.

Då det troligen föreligger olika subpopulationer kan man spekulera kring vad dessa utgörs av. Det kan röra sig om olika typer av kundgrupper. Applebaum (1951) grupperar in köpbeteenden i relation till bland annat inköpsställe och köpta produkter. Författaren menar att inköpsställen varierar från kund till kund där vissa köper det mesta i en viss affär medan andra köper lite här och där. I detta fall kan det vara så att en subpopulation avspeglar kunder som endast köper en eller få typer av varor hos det företag vi studerar medan en annan subpopulation istället köper det mesta hos företaget. Vidare diskuterar han att köpta produkter givetvis varierar där vissa lyxigare och dyrare produkter köps av få medan flera andra istället köper många billigare varor. I de data vi har kan det förslagsvis röra sig om att en subpopulation kan köpa få och billiga produkter, en annan kan istället köpa fler billiga produkter och en tredje liten mindre population kan köpa dyrare produkter. Applebaum diskuterar också impulsköp där man alltså köper något som man inte planerat men som stimulerades av någon form av besök. Detta tror vi absolut kan vara en viktig faktor som sker bland de köp vi undersökt. Då vissa kunder kan tänkas köpa endast den produkt de sökt efter och andra köper produkter utöver sökordet. Med andra ord kan det röra sig om att vissa endast köper någon enstaka produkt som de bestämt sig för sedan innan och andra köper på sig flera andra produkter de inte tänkt köpa från början. Dessa olika köpbeteenden är intressant att undersöka vidare.

För att undersöka hur väl en blandning av normalfördelningar fungerar på data med färre observationer validerar vi fördelningen genom att se hur väl den fungerar för olika kampanjer. Vi börjar med att undersöka dess duglighet på de fyra kampanjer med högst antal order (observationer), se tabell 4.3.4. Vi anpassar EM-algoritmen på dessa fyra kampanjer med olika antal normalfördelningar. Här finner vi att en blandfördelning med två (g=2) samt fler än tre normalfördelningar ($g > 3$) inte är dugliga. Därmed väljer vi att endast visa de blandfördelningar bestående av tre normalfördelningar. Här kommer vi fram till att empirisk data för de fyra stora kampanjerna beskrivs mycket väl av en blandfördelning bestående av tre normalfördelningar i samtliga fyra fall, se figur 4.3.2. Exempelvis har

empirisk data i kampanj 4 två toppar i den empiriska fördelningen, vilka båda efterliknas i hög grad av den blandade fördelningen.

Vidare vill vi också undersöka hur väl en blandad fördelning fungerar på kampanjer med mindre än 100 order (observationer). Här testar vi en sådan fördelning på fyra, slumpmässigt valda, kampanjer med 30-100 observationer, se tabell 4.3.6. För dessa fyra kampanjer passar en blandfördelning bestående av två normalfördelningar (g=2) bäst för tre fall: kampanj 1,2 och 4. Medan för kampanj 3 passar en sådan fördelning med tre normalfördelningar (g=3) bäst. De blandade fördelningarna passar väl för de fyra undersökta kampanjerna. Blandningen passar den empiriska fördelningen bäst för kampanj 1 och kampanj 3. Intressant är att kampanj 3 har störst antal observationer och kampanj 1 har minst antal.

De olika Blandningarna av normalfördelningar beskriver data för *Ordervärde* väl, vare sig datamaterialet består av många eller mindre än 100 order, och vi bestämmer att det är vårt slutgiltiga sätt att beskriva fördelningen för *Ordervärde*.

5.3 Förväntat ordervärde

När det kommer till att bestämma vad för metod som skall användas för att skatta det förväntade Ordervärdet behöver vi klargöra vissa saker. Det aritmetiska medelvärdet, som är en mycket enkel skattningsmetod, och skattningen från en blandning av flera normalfördelningar, ger mycket snarlika skattningar, se tabell 4.3.5 och 4.3.7. Skillnaden är som störst 0,7208 kronor och som minst 0,1088 för de åtta undersökta kampanjerna. Detta följer naturligt av att skattningen från den blandade fördelningen är en vägning av medelvärden från de olika fördelningarna.

En hypotes från vår uppdragsgivare var att medelvärdet inte skulle fungera särskilt bra på grund av uteliggare. Detta stämmer alltså inte för de data som undersöks här, ett skäl kan vara att uteliggarna inte är tillräckligt stora. Ett annat skäl till varför medelvärdet fungerar lika väl kan bero på det stora antalet order vi har i data och att det därmed bör förklaras av satsen om de stora talens lag. Satsen om de stora talens lag säger att medelvärdet av n stokastiska variabler (med samma fördelning och väntevärde), då n är stort, kommer ligga mycket nära väntevärdet (Häggström, 2004).

Slutligen för att skatta det förväntade ordervärdet anser vi att man kan använda sig av både skattningen från en blandning av fördelningar samt det aritmetiska medelvärdet. För att hålla sig till den säkra sidan vid skattningar av *Ordervärde* bör man anpassa en blandfördelning, se över så att denna fördelning följer data väl och därefter räkna ut skattningen. Detta sätt bör förhoppningsvis fungera ännu bättre i jämförelse med det aritmetiska medelvärdet i de fall där order är färre och uteliggare är större. Vill man istället göra det enkelt för sig, och i de fall när man har många observationer, kan man istället använda sig av medelvärdet.

6. Slutsats

I detta avsnitt lyfts centrala resultat fram, begränsningar och förslag på vidare forskning inom området ges.

Den första frågeställningen i syftet vi vill besvara är: ”Hur ser en bra modell ut för att beskriva sannolikhetsfördelningen för ordervärde?”. En slutsats är att blandfördelningar bestående av två eller tre normalfördelningar är ett bra sätt att beskriva fördelningen för ordervärden från en Adwords kampanj. Det faktum att blandfördelningar passar väl även för de små kampanjerna och bäst bland dessa för den med minst respektive störst antal observationer, tyder på att det inte finns ett starkt samband mellan hur väl modellen passar och antalet observationer. Vi konstaterar att blandfördelningar passar väl även vid data med färre än 100 observationer. Har man samlat in data för en kampanj med vissa nyckelord skulle en blandfördelning anpassad till data kunna användas för att beskriva fördelningen av framtida kampanjer med liknande nyckelord.

Ett annat intressant resultat från studien som är värt att lyfta fram är det faktum att blandfördelningar beskriver datamaterialet bra tyder på förekomst av subpopulationer i data. Det kan exempelvis vara subpopulationer av kunder som köper enstaka dyra produkter eller kunder som endast köper billiga förbrukningsvaror. I denna studie har vi inte valt att titta närmare på dessa olika subpopulationer, men att titta närmare på ett företags kunder med hjälp av blandfördelningar är ett intressant område för framtida forskning.

Den andra frågeställningen i syftet vi vill besvara är: ”Hur skattas ordervärdet på bästa sätt för det företag vi studerar?”. Vi kan inte påstå att vår metod för att skatta det förväntade ordervärdet med en blandning av flera normalfördelningar är en bättre metod än det aritmetiska medelvärdet då dessa värden är mycket snarlika. Vår uppdragsgivare kan alltså använda sig av medelvärdet som skattning av det förväntade ordervärdet. Metoden skapar däremot fördelningar som bättre kan beskriva data vilket ger mer fullständig information om ordervärdet.

Blandfördelningar kan även undersökas på data med fler uteliggare. Exempelvis hade det varit intressant att undersöka hur väl dessa fungerar för andra företags ordervärde. Till exempel på ett företag med främst företagskunder, där värdet på en order varierar mer. För ett sådant företag misstänker vi att fördelen med blandfördelningar är större vid skattning av ordervärdet men också för att beskriva fördelningen. För framtida studier är det också intressant att undersöka om det finns något tidsberoende för de olika fördelningarna i blandningen.

Det kan också vara intressant att titta mer noggrant på specifika kampanjer. Exempelvis har vissa kampanjer företagets namn med i många nyckelord och kan då tänkas agera annorlunda. Andra branscher är också intressanta att titta på och jämföra med varandra. Exempelvis att fastställa om och i så fall hur de förväntade ordervärdena skiljer sig mellan förbrukningsvaror och statusvaror.

Källförteckning

Accenture. (2014) *CMOs: Time for digital transformation Or risk being left on the sidelines.* https://www.accenture.com/t20150523T022804__w__/usen/_acnmedia/Accenture/ConversionAssets/DotCom/Documents/Global/PDF/Dualpub_10/Accenture-CMO-Insights-2014-pdf.pdf (Hämtad 2016-02-12).

Aitkin, M. & Rubin, D. B. (1985) *Estimation and Hypothesis Testing in Finite Mixture Models*. Journal of the Royal Statistical Society, Ser. B, 47, 67-75.

Applebaum, W. (1951) *Studying Customer Behavior in Retail Stores*. Journal of marketing. Vol. 16, No. 2, 172-178.

Bilmes, J.A (1998) *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. U.C. Berkeley.

Blume, M. (2002) *Expectation Maximization: A Gentle Introduction*. Technische Universität München.

Dempster, A.P., Laird, N. & Rubin, D.B. (1977) *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Ser. B, 39, 1-38.

Google (2015a) https://www.google.com/Adwords/benefits/?subid=ww-ww-et-awhc-nav (Hämtad 2015-12-10).

Google (2015b) https://www.google.com/Adwords/costs/?subid=ww-ww-et-awhc-nav (Hämtad 2015-12-10).

Google (2015c) https://support.google.com/Adwords/answer/2497976?hl=en&ref_topic=3121763 (Hämtad 2015-12-10).

Google (2015d) https://support.google.com/Adwords/answer/1704371 (Hämtad 2015-12-10).

Google (2015e) https://support.google.com/partners/answer/1704396?hl=en-GB&ref_topic=2795210 (Hämtad 2016-02-01).

Häggström, O. (2004) *Slumpens skördar: strövtåg i sannolikhetsteorin*. Lund: Studentlitteratur.

Internet Live Stats. http://www.internetlivestats.com/google-search-statistics/ (Hämtad 2016-01-30).

Sheather, S.J. (2009) *A Modern Approach to Regression With R.* New York: Springer.

Zucchini, W. (2000) *An Introduction to Model Selection*. Journal of Mathematical Psychology, 44, 41-61.

Bilaga

95 % konfidensintervall för de skattade parametrarna i avsnitt 4.1 Produktomsättning.

Fördelning	**KI parameter 1**	**KI parameter 2**
Normal	$\hat{\mu}$: (257,43;261,59)	$\hat{\sigma}$: (245,79; 248,73)
Lognormal	$log\hat{\mu}$: (5,02;5,04)	$log\hat{\sigma}$ (1,14;1,15)
Weibull	$\hat{\alpha}$: (1,03;1,05)	$\hat{\beta}$: (261,53;266,04)
Gamma	$\hat{\alpha}$: (0,004;0,004)	$\hat{\theta}$: (1,07;1,09)

95 % konfidensintervall för de skattade parametrarna i avsnitt 4.2 Ordervärde.

Fördelning	**KI parameter 1**	**KI parameter 2**
Normal	$\hat{\mu}$: (564,56;572,98)	$\hat{\sigma}$: (334,75; 340,70)
Lognormal	$log\hat{\mu}$: (6,16;6,18)	$log\hat{\sigma}$ (0,63;0,65)
Weibull	$\hat{\alpha}$: (1,77;1,80)	$\hat{\beta}$: (635,72;645,18)
Gamma	$\hat{\alpha}$: (0,005;0,005)	$\hat{\theta}$: (2,95;3,05)

95 % konfidensintervall för de skattade parametrarna i avsnitt 4.3 Blandfördelningar.

Fördelning	**KI $\hat{\mu}_i$**	**KI $\hat{\sigma}_i$**
Normalfördelning 1	(148,74;153,26)	(56,64;59,84)
Normalfördelning 2	(508,63;513,37)	(155,95;159,29)
Normalfördelning 3	(956,50;981,50)	(447,53;465,22)

95 % konfidensintervall för de skattade ordervärdena i avsnitt 4.3.1.

Kampanj	**KI**
Kampanj1 (stor)	(604,67;617,59)
Kampanj2 (stor)	(631,02;684,64)
Kampanj3 (stor)	(584,07;635,43)
Kampanj4 (stor)	(524,62;583,22)
Kampanj1 (liten)	(341,66;464,62)
Kampanj2 (liten)	(475,36;664,26)
Kampanj3 (liten)	(445,98;514,23)
Kampanj4 (liten)	(347,71;442,58)