

Danmarks Statistiks detailomsætningsindeks: Metodedokumentation

Indhold

Indledning	1
Population og stikprøve	1
Specialindberettere.....	2
Stratificering.....	2
Stikprøvefornyelse og paneludskiftning	3
Variable og dataindsamling	4
Dataindsamling	5
Fejlsøgning	5
Opregning og estimater	5
Imputering	6
Måned-til-måned kædning	7
Indeksring og deflatering	7
Deflatering og mængdeindeks.....	8
Sæsonkorrektion.....	8
Offentliggørelser og revisioner	9

Indledning

Denne metodeoversigt giver en indførsel i de anvendte metoder i produktionen detailomsætningsindekset (DOI) i Danmarks Statistik. Fra populationsafgrænsning over opregningen til offentliggørelser og revisioner. Målet er at give et metode-mæssigt overblik til eksterne brugere og skal ses som et supplement til Kvalitetsdeklarationen. Indholdet her vil generelt være mere teknisk i Kvalitetsdeklarationen, men der vil også være tale om en del gentagelser.

Population og stikprøve

DOI er en stikprøvebaseret statistik, der dækker alle detailvirksomheder med undtagelse af detailhandel med motorkøretøjer. Det betyder, at populationen består af virksomheder med en DB07 branche, hvis to første cifre er 47 (hovedgruppe 47). I tillæg hertil kommer branche 10.71.20, Fremstilling af friske bageriprodukter, der udgør en væsentlig del af detailsalget af bageriprodukter i Danmark. Den branche slås sammen med 47.24.00, Detailhandel med brød mv. Derudover slås to andre brancher sammen til én: 47.76.10, Blomsterforretninger og 47.76.20, Planteforhandlere og havecentre.

Specialindberettere

Tællingsenheden er virksomhed, i Danmark identificeret ved et CVR nr., hvilket er jævnfør EU's forordning på kortidsstatistikområdet. Der afviges i nogen grad fra dette, da enkelte enheder består af flere virksomheder, og nogle virksomheder omvendt opdeles i flere. Enkelte udvælges også på trods af, at de har hovedbranche i engroshandel, da de vurderes vigtige for at vise den rigtige udvikling. Alle disse enheder blev tidligere kaldt *fællesindberettere*, men en bedre betegnelse er *specialindberettere*, da de altså ikke kun dækker over enheder, der indberetter fælles for flere virksomheder.

Specialindberetterne tvangsudvælges fra populationen, som naturligvis også renses for alle virksomheder, der indberettes under en specialindberetter, inden der vælges stikprøve. Specialindberetterne holdes i øvrigt ude af den normale opregningsprocedure, og bidrager således altid blot med deres egen omsætning med vægt 1. Ved omlægningen i maj 2012 blev en række almindelige indberettere ændret til specialindberettere. Det var de største indberettere, der ikke allerede var specialindberettere, og det blev gjort netop for, at de ikke skal påvirke opregningen for de øvrige enheder.

Der er pt. godt 100 specialindberettere, hvoraf ca. halvdelen er helt almindelige, men store, virksomheder.

Stratificering

Populationen stratificeres efter deres branche (med de to ovennævnte sammenlægninger) og deres størrelse, målt på deres årlige omsætning fra det Erhvervsstatistiske register (ESR). Den omsætning afspejler det momspligtige salg indberettet til SKAT (ekskl. moms). Størrelsesstrata er som følger:

Størrelsesgruppe	Årlig ESR omsætning
0	<2,5 mio. kr.
1	2,5-5 mio. kr.
2	5-10 mio. kr.
3	10-20 mio. kr.
4	>20 mio. kr.

Størrelsesgruppe 0 er cut-off, hvorfra der ikke trækkes til stikprøven. Nogle brancher har en struktur, der placerer store omsætningsandele i gruppe 0. I 2011 blev der derfor indført en alternativ størrelsesgruppe 1 med intervallet 1-5 mio. kr. Det drejer sig om brancherne:

DB07 Branche
472100
472300
472500
472600
472900
475100
475920
476500
477120

477610
477620
477630
477830

Størrelsesgruppe 4 er take-all, hvorfra alle virksomheder trækkes til stikprøven.

Fra gruppe 1, 2 og 3 trækkes ægte stikprøver med divergerende udvalgssandsynlighed, der dog er stigende indenfor de enkelte brancher.

Stikprøvefornyelse og paneludskiftning

Der skiftes ud i stikprøven en gang om året. Der benyttes en panelstruktur, der både tager hensyn til stabilitet i stikprøven og udskiftning af hensyn til at begrænse indberetningsbyrden for de mindre virksomheder.

Fra 2012 blev metoden til paneludskiftning lagt om. Der benyttes nu en metode, der på simpel vis roterer stikprøven og samtidig bevarer repræsentativiteten. Kort sagt fungerer det ved, at der i hvert stratum gøres følgende:

1. Alle virksomheder tildeles et tilfældigt tal x mellem 0 og 1. Det er det såkaldte samu-tal (SAMU kommer fra Sverige og står for samordnet udvælgelse – systemet kan nemlig bruges på tværs af tællinger).
2. Alle virksomheder tildeles også et tilfældigt samu-gruppe tal, r . Det er et positivt heltal mellem 1 og n , hvor n altså er antallet af samu-grupper (pt. 3 for DOI). Tildelingen sker, så der er lige mange enheder i hver gruppe.
3. Med et givet startpunkt (pt. 0 for DOI) vælges virksomhederne med de laveste x -værdier, der er større end startværdien, til stikprøven indtil der er valgt p procent, hvor p angiver allokeringen for pågældende stratum (hvis stratummet er stort nok, vil enheden med den største x -værdi i stikprøven dermed have en x -værdi omkring $\frac{p}{100}$). På den måde er den første stikprøve udvalgt.
4. Året efter lægges der en fast værdi (pt. 1/3 for DOI) til x -værdien for virksomhederne i samu-gruppe 1. Året efter er det så samu-gruppe 2 og så fremdeles. Hvis en x -værdi kommer over 1, trækkes der 1 fra.
5. Nye virksomheder tildeles tilfældige samu-tal og samu-grupper.
6. Derefter gentages pkt. 3, og på den måde vælges den nye stikprøve.

På den måde udskiftes i DOI eksempelvis hvert år ca. 1/3 af stikprøven, hvis $p < 1/3$, hvorefter de udskiftede enheder normalt vil være ude af stikprøven i minimum 6 år.

Er $\frac{1}{3} < p < \frac{2}{3}$ er det mellem 1/6 og 1/3 der udskiftes hvert år, hvorefter de udskiftede enheder normalt vil være ude af stikprøven i minimum 3 år.

Er $p > 2/3$ kan der gå over 6 år, før en enhed udskiftes – de vil dog stadig som udgangspunkt være ude i minimum 3 år.

Alle disse "tommelfingerregler" skal tages med det forbehold, at ændringer i populationen såvel som ændringer i den enkelte enhed kan have betydning for udskiftningen fra år til år.

Eksempel fra DOI

I 2012 populationen var der i størrelsesgruppe 2 i branchen 47.11.10 (Købmænd) 282 enheder. Det gav 94 i hver gruppe. I det stratum er $p = 24.4$, så der skulle vælges $0.244 \cdot 282 \approx 69$ enheder til stikprøven. De 69 med de laveste samu-numre blev valgt. Tilfældighederne gjorde, at det laveste samu-tal i stikprøven endte på 0.18. Næste år bliver der lagt $1/3$ til samu-tallet for alle enhederne i samu-gruppe 1. Alle fra samu-gruppe 1, der var i stikprøven, vil dermed skiftes ud næste år (da deres samu-tal så vil blive over $1/3$ og under 1 – og der forventeligt ikke ændres på p).

Variable og dataindsamling

Der spørges hver måned til virksomhedernes detailsalg, hvilket er defineret som videresalg af varer til private. Principielt skal der ikke skelnes mellem omsætning fra detailsalg og andre aktiviteter, men det er vurderet, at der for nogle store virksomheder er så væsentlig omsætning fra andre aktiviteter, at det ikke giver mening at inkludere disse i tallene.

Nogle virksomheder har meget vanskeligt ved at udskille deres salg til private fra den samlede omsætning. I så fald accepteres et skøn, og i nogle tilfælde også at hele omsætningen rapporteres.

Respondenterne kan selv bestemme, om de vil indberette ekskl. eller inkl. moms.

For fire brancher spørges til omsætningen fordelt på tre varegrupper:

Fødevarer og andre dagligvarer, beklædning mv. og andre forbrugsvarer.

Det drejer sig om følgende brancher:

47.11.20, Supermarkeder

47.11.30, Discountforretninger

47.19.00, Anden detailhandel fra ikke-specialiserede forretninger (varehuse og stormagasiner)

47.30.00, Servicestationer

For de øvrige brancher spørges kun til ét samlet omsætningstal.

En detaljeret opdeling af supermarkedsvarene på de tre varegrupper kan findes

[LINK](#)

Dataindsamling

DOI er obligatorisk digital, så alle indberetninger skal foretages via et særligt system til indtastning via telefon eller via Virk.dk.

Omsætningen indberettes i hele kroner med angivelse af, om det er ekskl. eller inkl. moms.

Der rykkes for indberetning ved brug af breve, e-mails, telefoner og anbefalede breve.

Fejlsøgning

Der anvendes en forholdsvis avanceret fejlsøgningsmetode, der både kontrollerer de indberettede tal på mikro- og makro-niveau.

Mikrometoderne er klassiske, der kontrolleres for store udsving i forhold til foregående periode, samme periode foregående år, stigningsraten i samme periode sidste år og også for store skift i absolutte værdier.

Makrometoden gør kort sagt det, at den for den enkelte indberetning måler dens effekt på de opregnede estimater, på forskellige niveauer (samlet DOI, varegrupper og brancher). Hvis stigningsraten påvirkes i tilstrækkelig grad af en given observation, markeres den til fejlkontrol. På den måde er det en kombination af en enheds størrelse/vigtighed og afvigelsen af indberetningen i forhold til lignende virksomheder, der gør, om den fejlmarkeres.

Fejlsøgningen foretages ved gennemgang af lister, der kommer som output fra et SAS program. For hver enhed markeret til fejlsøgning noteres om den er rettet og om den er genkontaktet.

Opregning og estimater

Til opregning af den enkelte måneds omsætning anvendes et ratio estimat, hvor opregningspopulationen er et udtræk fra ESR, der normalt er opgjort ultimo det senest afsluttede kvartal, med moms-omsætningen for det seneste år som hjælpevariabel.

Meget kort sagt er formelen for et ratioestimat:

$$X^{ratio} = X^{dir} \cdot Y_{pop} / Y^{dir}$$

Her er X^{ratio} det estimat, vi gerne vil beregne ud fra indsamlet data (i dette tilfælde detailomsætningen) og X^{dir} er det direkte estimat, hvor der altså blot an-

vendes bortfaldskorrigerede udvalgsvægte (var der eksempelvis i et stratum 20 enheder i populationen og 10 svar, er vægten lig 2).

Samtidig er Y_{pop} den summerede hjælpevariabel for hele populationen (f.eks. summen af moms-omsætning), mens Y^{dir} er det direkte estimat beregnet for hjælpevariablen.

Faktoren Y_{pop}/Y^{dir} kaldes også korrektionsfaktoren, idet det er den faktor, hvormed det direkte estimat korrigeres. I estimeringen af DOI korrigeres med denne faktor på brancheniveau (et alternativ er at gøre det helt nede på størrelsesgrupperne for de enkelte brancher, hvilket dog øger sårbarheden overfor størrelsesskift i opregningspopulationen).

En simpel omskrivning af formlen viser, at der gælder

$$\frac{X^{ratio}}{X^{dir}} = \frac{Y_{pop}}{Y^{dir}}$$

hvilket viser at estimeringen bygger på antagelsen om, at forholdet (deraf ratio) mellem det "sande" estimat og det direkte estimat er det samme for tællevariablen og hjælpevariablen. Er der god korrelation mellem tællevariablen og hjælpevariablen, kan ratioestimatet nedbringe stikprøveusikkerheden betragteligt. Bemærk at tællevariablen og hjælpevariablen ikke behøver at være numerisk tæt på hinanden. Således kan et omsætningsestimater udmærket korrigeres med f.eks. beskæftigelsestal – så længe korrelationen er tilstrækkelig.

Der er generelt rigtig god korrelation mellem den indsamlede detailomsætning og moms-omsætningen fra ESR, og usikkerheden på de opregnede totaler reduceres typisk med 20-50 % ved at anvende ratioestimering.

I opregningen indgår specialindberetterne som tidligere nævnt med vægt 1. Både specialindberetterne og alle virksomheder, der indberettes under en specialindberetter fjernes derfor fra opregningspopulationen. Det gør naturligvis, at der skal være data for disse enheder hver måned. Mangler der data for en specialindberetter, når der skal afsluttes, anvendes der imputering.

Imputering

Som hovedregel anvendes kun imputering, når det er absolut nødvendigt. Normalt bortfald håndteres via opregningen, men for de godt 100 specialindberettere anvendes der imputering, ligesom der også anvendes imputering, hvis der helt mangler indberetninger for en (lille) branche.

Imputeringen foregår således, at et SAS program automatisk identificerer de enheder, der skal imputeres værdier for og kommer med et beregnet forslag. Der anvendes tre imputeringsmetoder i prioriteret rækkefølge:

- a) Fremskrivning med enhedens stigning samme måned året før

- b) Fremskrivning med stigningen for enhederne med samme branche og størrelsesgruppe
- c) Gennemsnit for de seneste tre måneder

I månederne omkring påske kan metode a) give problemer, så i marts, april og maj anvendes prioriteringen b-a-c. I de øvrige måneder prioriteres a-b-c.

De foreslåede imputerede værdier gennemgås manuelt og rettes til, hvis der er behov for det. Ved første afslutning af en måned vil der normalt være under 10 imputeringer.

Måned-til-måned kædning

Tidsserierne i DOI er baseret på måned-til-måned kædning. Det betyder, at der hver måned udregnes stigningsrater ved at genberegne den foregående måned og den aktuelle måned samtidig og med samme opregningspopulation. Stigningsraterne beregnes på brancheniveau og for de varegruppeopdelte brancher på varegrupper, og stigningsraterne ganges fortløbende på "kædens start", som pt. er januar 2011, hvor værdierne altså er de faktiske opregnede værdier. For perioderne før januar 2011 regnes således baglæns ved at dividere stigningsraterne på.

Kædningen gør først og fremmest at strukturelle ændringer i populationen eller stikprøven ikke får fuld effekt i de offentliggjorte tal. Benyttede man *ikke* kædning, ville for eksempel et brancheskift fra periode $t - 1$ til periode t resultere i en nedgang i den branche, enheden forlader, og en fremgang i den branche den tilgår. Da der normalt vil være tale om en rettelse af branchen og ikke et egentligt aktivitetsskift er det en u hensigtsmæssig bevægelse i tallene, der ikke afspejler en egentlig trend i nogle af de berørte brancher. På samme vis undgår man også ved at benytte kædning, at det årlige skift i stikprøven i sig selv kan give anledning til et databrud. Der indsamles således i januar for både gammel og ny stikprøve, således, at der kan kædes med overlap i januar.

Et kendt problem ved kædningen er, at der netop kan være strukturelle ændringer man *gerne* vil afspejle i tallene. Nye virksomheder får f.eks. meget lille effekt, så længe de er uden for stikprøven, netop fordi stigningsraterne altid baseres på opregning af to måneder med den samme opregningspopulation. I princippet kan der altså åbne en ny stor og vigtig detailforretning i løbet af året, og den vækst den har, indtil den kommer med i stikprøven, bliver aldrig afspejlet i tallene. Da det omvendt gælder, at lukkede virksomheder repræsenteres ved lukninger i stikprøven, er der altså en risiko for, at kædningen giver et negativt bias. Det skal man naturligvis være opmærksom på.

Indeksering og deflatering

Opregningen og kædningen resulterer i kædede omsætninger på brancher og varegrupper. En oversigt over, hvilke varegrupper de enkelte brancher indgår i, kan findes [LINK](#).

Beregningen af værdiindeks foregår helt simpelt ved, at der indekseres i forhold til det månedlige gennemsnit i basisåret (pt. 2005). Det gøres direkte på omsætningerne på både brancheniveau, varegruppeniveau og på totalen. Der er altså ikke tale om indeks på højere aggregeringsniveauer beregnes ved at vægte indeks fra lavere niveauer. Det kan dog let vises, at der implicit er tale om såkaldte *interne vægte*, og at disse er lig omsætningsandelen i basisåret. Dermed vil eksempelvis en sammenvejning af alle brancheindeksene med disse omsætningsandele som vægte altså give samme resultat som den direkte indeksering af totalomsætningen.

Deflatering og mængdeindeks

Fra Forbrugerprisindekset (FPI) hentes hver måned aktuelle tal for prisudviklingen på forskellige varer. Der hentes fra flere offentliggørelsesniveauer og også enkelte indeks fra basisindeksniveauet, som ikke offentliggøres, anvendes. Indeksene sammenvejes med FPI's egne vægte til prisindeks, der kan anvendes som deflaterer på varegruppeniveau og for EU grupperne. Branchetallene deflateres ikke. Der foregår desuden en omregning fra FPI's basisår (2000) til DOI basisåret (2005).

For varegrupperne deflateres for hver af dem den samlede omsætning, med undtagelse af andre forbrugsvarer i branche 47.30.00, Servicestationer, der deflateres separat med prisindeks for brændstof.

De deflaterede omsætninger giver anledning til beregningen af *mængdeindeks*. Der er tale om et såkaldt *indirekte* mængdeindeks, der beregnes ved formlen

$$I_M = 100 \cdot I_V / I_p,$$

altså er mængdeindekset lig værdiindekset divideret med prisindekset, hvilket igen bygger på den fundamentale identitet, at værdi er lig mængde gange pris.

Også for mængdeindeksene indekseres der direkte på alle niveauer, og der gælder således samme princip vedrørende interne vægte, der her er de deflaterede omsætningsandele i basisåret.

Sæsonkorrektion

Der sæsonkorrigeres værdi- og mængdeindeks for varegrupperne. Til offentliggørelserne af varegrupperne anvendes indirekte sæsonkorrektion til beregning af de sæsonkorrigerede totaler. Til sammenvejning af varegruppeindeksene anvendes omsætningsandelene (hhv. de deflaterede omsætningsandele) i basisåret.

Modelvalg og en række parametre vedr. kalendereffekter holdes så vidt muligt fast for minimum et år ad gangen. Dette sikrer normalt at revisioner, der udelukkende skyldes ændringer i sæsonkorrektionen minimeres. Til gengæld er modelvalg mv. ikke løbende optimeret til at passe til data.

Offentliggørelser og revisioner

Den første offentliggørelse er ca. t+20-t+22, afhængig af kalenderen. Det er foreløbige tal på varegrupper og levering af 5 EU grupper til Eurostat.

Ved næste offentliggørelse (samme tid en måned senere) revideres tallene og der offentliggøres foreløbige tal på brancher. Desuden sendes nu tal for 13 EU grupper til Eurostat.

En måned senere revideres tallene for sidste gang, og herefter revideres tallene kun ved ekstraordinære hændelser.

I praksis betyder dette altså, at der ved hver offentliggørelse af nye tal også offentliggøres reviderede tal for de seneste to måneder. I de nationale offentliggørelser betyder det også, at varegruppetalene revideres to gange, mens branchetallene revideres én gang.