DANMARKS STATISTIK

2. maj 2024
Ole Schnor
DST Consulting

Kvalitet af sammenhæng i data, 10 centrale personstatistikregistre 1981-2021

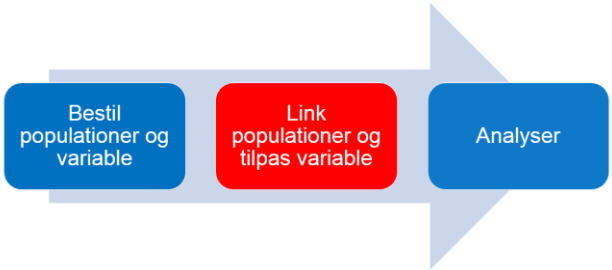
Indledning 2
Statusregistre vs. forløbsregistre 3
Befolkningen (BEF og FAIN) 4
Befolkningens populationer vs. RAS, UDDA og BOL 5
Identiske populationer 1981-2007 5
Forskellige populationer fra 2008 5
Højest fuldførte uddannelse (UDDA) 7
Registerbaseret arbejdsstyrke (RAS) 8
Boligopgørelsen (BOL) 9
Indkomststatistikken (IND) 9
Arbejdsklassifikationer (AKM) 11
Historiske vandringer (VNDS) 12
Indvandrere og Efterkommere (IEPE) 12
Integreret database for arbejdsmarked (IDAP) 13
Elevregistret (KOTRE) 13
Samlet oversigt, kan man ramme Statistikbanken? 14
Nøglerne CPRNR og PERSON_ID 15
Bilag 1. Algoritme i forbindelse med brugen af Elevregistret. ... 16

Indledning

Personstatistikregistrene i Danmarks Statistik består af omfattende datasamlinger, som er op- og udbygget siden begyndelsen af 1980'erne. Data er af høj kvalitet og omfatter hele befolkningen. Dette giver unikke analysemuligheder for brugerne af data - det være sig på et bestemt tidspunkt og over tid.

Dette notat giver indsigt i sammenhængen mellem flere af de mest anvendte forskningsregistre i Danmarks Statistiks Mikrodataordning og deres forbindelse med den offentliggjorte statistik. Notatet henvender sig derfor primært til forskere, analytikere og andre brugere af mikrodata, som ønsker at opnå et bedre indblik i kvaliteten af sammenhængen mellem de forskellige registre.

Der forudsættes et vist kendskab til Grunddata i Forskningsservice, og der henvises til den navngivning af registrene som benyttes af Forskningsservice jf. registeroversigten Hvis man henter sine data fra en projektdatabase eller benytter en myndighedsordning, vil der ofte være en datamanager, man kan spørge til råds.



Inden en bruger kan nå frem til sine analyser skal populationerne linkes og variable skal tilpasses. Her kan der opstå usikkerhed om resultaterne fx: "Hvorfor er der personer, der ikke matcher?" og "Kan man genskabe Statistikbankens tal?

Hvis der ikke er overensstemmelse mellem mikrodata og Statistikbanken, falder tvivlen i første omgang oftest tilbage på brugerens egne programmer, for er der mon en fejl? Men forklaringen kan også være, at man som udgangspunkt slet ikke kan genskabe Statistikbankens tal, og at ikke alle personer nødvendigvis kan finde match. Det er disse spørgsmål, der er omdrejningspunkterne i notatet.

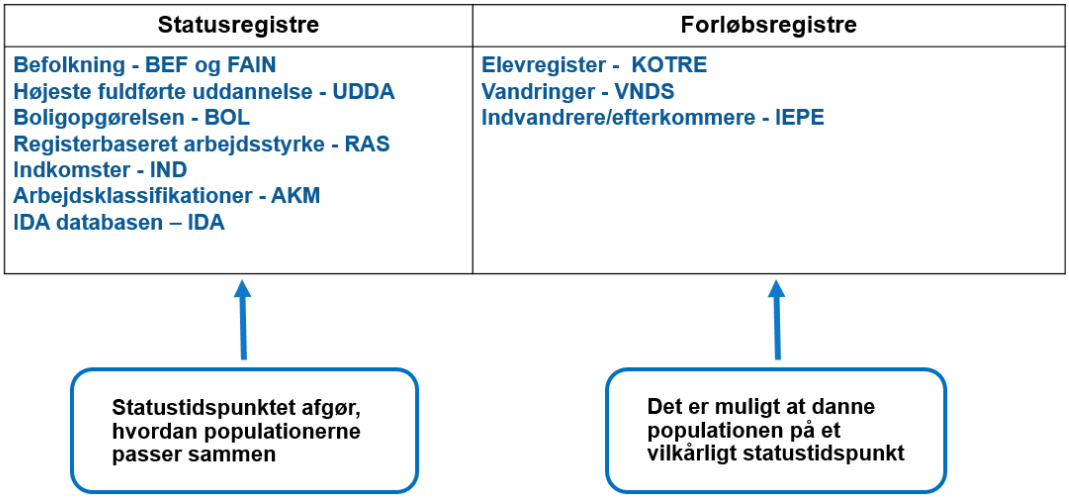
Der ses på kvaliteten i sammenhængen mellem personstatistikregistrene i perioden 1981-2021 i Grunddata hos Forskningsservice (FSE). Der tages udgangspunkt i 10 registre, som er blandt de mest benyttede:

- Befolkningen, Statusbefolkningen, **BEF (1)**
- Uddannelser, Højest fuldførte uddannelse, **UDDA (2)**
- Indkomststatistikken, **IND (3)**
- Befolkningen, Vandringer**, VNDS (6)**
- Registerbaseret Arbejdsstyrkestatistik, **RAS (7)**
- Arbejdsklassifikationer, **AKM (8)**
- Integreret Database for Arbejdsmarkedsforskning, **IDA (10)**
- Befolkningen, Indvandrere og efterkommere**, IEPE (18)**
- Uddannelser, Elevregistret, **KOTRE (21)**
- Boligopgørelsen, **BOL (32)**

Placering på listen over de mest benyttede registre i FSE er angivet i parentes. Befolkningsstatistikkens statusbefolkning er således det mest benyttede register. Kommentarer til notatet modtages, skriv gerne til Ole Schnor OSC@dst.dk

Statusregistre vs. forløbsregistre

Der findes to typer af registre: statusregistre og forløbsregistre. Det er vigtigt at have kendskab til hvilken type, der anvendes, da det kan give nogle muligheder, men også sætte nogle begrænsninger i populationsdannelsen. De fleste personstatistikregistre er statusregistre, dvs. at populationen af personer er opgjort på et på forhånd fastlagt statustidspunkt. Historisk har Befolkningsstatistikkens population pr. 1. januar i året været udgangspunktet for flere af de centrale personstatistikker, hvilket vil fremgå af beskrivelserne i de følgende afsnit. Siden 00'erne har forløbsregistrene vundet indpas, hvilket muliggør populationsdannelse på et vilkårligt statustidspunkt.



I de følgende afsnit er der fokus på statusregistrene, idet det oftest er disse , der benyttes i forbindelse med de publicerede statistikker. Eneste undtagelse er Elevregistret (KOTRE), der er et forløbsregister og benyttes i de publicerede statistikker.

Hvis man ønsker at genskabe den offentliggjorte statistik, skal man derfor næsten altid benytte statusregistrene. Udfordringen kan være, at populationerne ikke nødvendigvis matcher fuldstændigt. Til forskningsformål kan det være underordnet, om man kan genskabe de offentliggjorte tal, men helt afgørende at populationerne matcher hinanden. I det tilfælde må det anbefales, at man benytter forløbsregistrene i det omfang de findes. Bemærk derfor, at ud af de 7 statusregistre er disse 2 baseret på forløbsregistre:

- Højeste fuldførte uddannelse: UDDF – Højeste fuldførte uddannelse (forløb)
- Registerbaseret arbejdsstyrke: AMRUN – Arbejdsmarkedsregnskab uden timenormering. Registret findes fra og med 2008.

Den publicerede statistik, der benytter Elevregistret har sit statustidspunkt d. 1. oktober i året. Da dette kræver en særskilt algoritme af genskabe statistikken er den dokumenteret i bilag 1 i form af SAS-kode.

Befolkningen (BEF og FAIN)

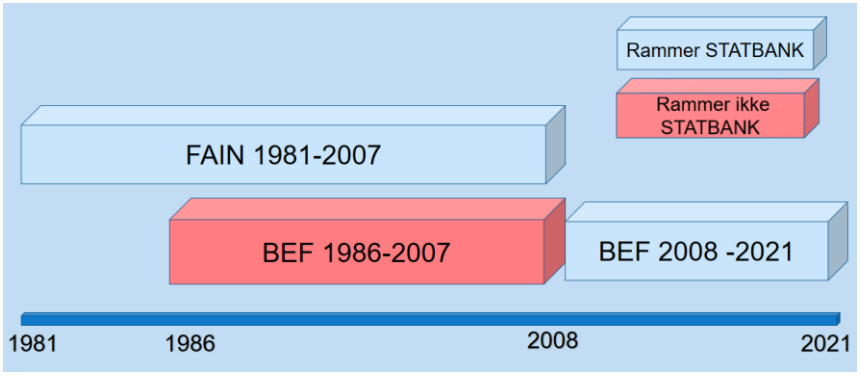
Implementeringen af den Personstatistiske Database (PSD) i 00'erne har stor betydning for at forstå sammenhængene mellem de befolkningsdatasæt, som ligger i FSE's Grunddata. Følgende skal fremhæves:

1. I PSD afløses CPRNR af PERSON_ID som den centrale nøgle. PERSON_ID er den statistiske enhed, hvilket betyder, at personer, der skifter CPRNR, vil fastholde det samme PERSON_ID. I perioden 1981-2020 drejer det sig om ca. 12.000 personer, som i Befolkningsstatistikken har haft et skift i CPRNR. Bemærk, at det er CPRNR, som er nøglen i FSE's Grunddata med variabelnavnet **PNR**. I FSE regi har man dermed ikke mulighed for at følge de personer, der skifter CPRNR.
2. Der indføres et nyt familiebegreb, E-familier. Befolkningsregistret og E-familierne tilbageføres til 1986. På den baggrund dannes en ny statistik i Statistikbanken med familier og husstande fra 1986 og frem.
3. Den personrelaterede statistik i Statistikbanken blev ikke opdateret tilbage til 1986. Først fra og med 2008 bliver PSD benyttet som grundlag for den personrelaterede statistik.

Ovenstående ændringer er baggrunden for, at der findes to udgaver af Befolkningsregistret i FSE's Grunddata:

FAIN – Husstande og familier. Indeholder populationer for perioden 1980-2007. Nøglen er de oprindelige CPRNR, og populationen svarer meget nøjagtigt til det antal personer, som er offentliggjort i Statistikbanken.

BEF – Befolkningen. Indeholder den PSD baserede befolkning fra 1986 og frem. Nøglen er CPRNR, som er konverteret fra PSD's PERSON_ID. Ved en sammenligning med antal personer i Statistikbanken, stemmer populationen ikke i årerne 1986-2007. Det stemmer først fra og med 2008, som skitseret i figuren.



Da **BEF** 1986-2007 er baseret på et nyere udtræk fra CPR registret end **FAIN** i samme periode, vil **BEF** alt andet lige indeholde en mere retvisende population. Til gengæld har **FAIN** en nøje sammenhæng til de øvrige registre **RAS**, **UDDA** og **BOL**, hvilket fremgår af næste afsnit.

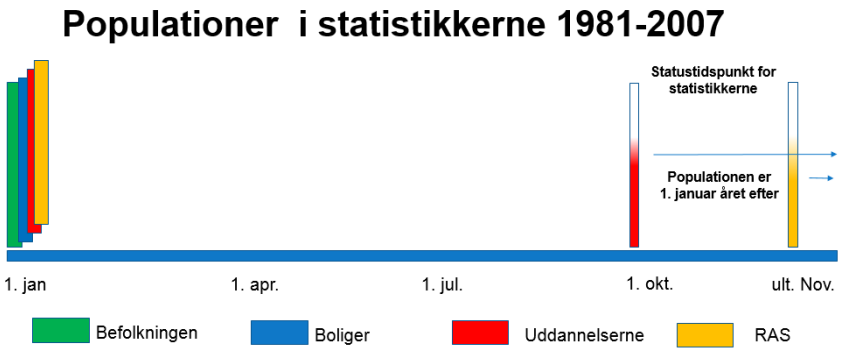
Befolkningens populationer vs. RAS, UDDA og BOL

Her ses på sammenhængen mellem 4 af de grundlæggende statusregistre:

- Befolkningen, Statusbefolkningen, **BEF. (FAIN** 1981-2007)
- Uddannelser, Højest fuldførte uddannelse, **UDDA**
- Registerbaseret Arbejdsstyrkestatistik, **RAS**
- Boligopgørelsen, **BOL**

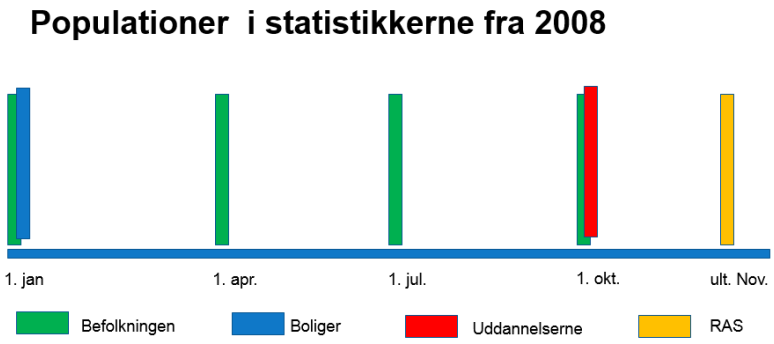
Identiske populationer 1981-2007

Afgrænsning af statistikkernes populationer sker med udgangspunkt i Befolkningsregistret og i perioden 1981-2007 , hvor de 4 registre umiddelbart kan linkes, idet populationerne er helt identiske. Endvidere er de som udgangspunkt direkte sammenlignelige med den offentliggjorte statistik. Det bemærkes, at statustidspunktet for selve statistikken er ultimo november året inden for RAS og 1. oktober året inden for uddannelserne.



Forskellige populationer fra 2008

Fra og med 2008 bliver Befolkningsstatistikken offentliggjort kvartalsvis, mens der fortsat er tale om årlige statistikker for de andre registre. For RAS og Uddannelsesregistret gælder, at statustidspunktet for populationerne skifter i 2008, således at populationerne nu flugter med statustidspunkterne for selve beskæftigelsesoplysningerne (ultimo november) og selve uddannelsesoplysningerne (1. oktober). Det betyder, at man fra og med 2008 ikke har identiske populationer på tværs af de 4 registre. Befolkningen og Uddannelserne kan sammenlignes pr. 1. oktober, Befolkningen og Boligerne kan sammenlignes pr. 1. januar, mens populationen i RAS ikke har et sammenfald med de andre registre.



RAS er fra og med 2008 baseret på Arbejdsmarkedsregnskabet (AMR), som dækker hele kalenderåret. For at få en så retvisende befolkning som mulig, ligger registertidspunktet senere end det, der benyttes i Befolkningsstatistikken. Således er befolkningen for december 2021 trukket i oktober 2022, befolkningen i november 2021 er trukket i september 2022 osv. Derved er der samme tidsmæssige spænd mellem reference- og registertidspunkter måned for måned, og det sikrer, at beskæftigelses- og erhvervsfrekvenserne bliver sammenlignelige hen over året.

Ovennævnte betyder, at når man forsøger at linke efter 2007, vil der derfor være personer, der ikke matcher, og resultaterne kan til tider være svært sammenlignelige med den offentliggjorte statistik. I det følgende ses et konkret eksempel.

Populationen i nedenstående tabel er befolkningen (15-69 år) pr. 1. oktober 2019 med dens højeste fuldførte uddannelse pr. 1. oktober 2019 og socioøkonomisk status fra RAS pr. ultimo november 2019. Populationen danner udgangspunkt for en af tabellerne i forbindelse med regionernes tilskud på udviklingsområdet.[1] De 9.491 personer, der ikke findes i RAS markeres for sig, da de ikke er i befolkningen ultimo november.

tid 2019

	Socioøkonomisk status				
	1. Under uddannelse	2. Beskæftigede	3. Arbejdsløs	4. Uden for arbejdsstyrken	5. Er i befolkningen 1. oktober men ikke i RAS ultimo november
Højeste fuldførte uddannelse					
H10 Grundskole	330.108	336.167	19.656	313.740	2.653
H20 Gymnasiale uddannelser	164.434	199.438	5.788	64.253	2.994
H30 Erhvervsfaglige uddannelser	37.121	878.167	26.477	230.361	3.093
H35 Adgangsgivende uddannelsesforløb	1.116	1.433	63	956	13
H90 Uoplyst	1.258	21.316	1.216	23.480	738
I alt	534.037	1.436.521	53.200	632.790	9.491

Tallene er direkte sammenlignelige med Statistikbanken HFUDD16, hvor man har valgt at kategorisere personer med manglende RAS-oplysninger i gruppen 'Uden for arbejdsstyrken'. Dette betyder, at det ikke er gennemskueligt, hvordan de bagvedliggende populationer er dannet.

Befolkningens højest fuldførte uddannelse (15-69 år) efter bopælsområde, tid, højest fuldførte uddannelse og socioøkonomisk status

	Under uddannelse	Beskæftigede	Arbejdsløs	Udenfor arbejdsstyrken
Hele landet				
2019				
H10 Grundskole	330 108	336 167	19 656	316 393
H20 Gymnasiale uddannelser	164 434	199 438	5 788	67 247
H30 Erhvervsfaglige uddannelser	37 121	878 167	26 477	233 454
H35 Adgangsgivende uddannelsesforløb	1 116	1 433	63	969
H90 Uoplyst mv.	1 258	21 316	1 216	24 218
Subtotal	534 037	1 436 521	53 200	642 281

Eksemplet viser, at man skal være meget påpasselig, når man matcher populationer, hvor statustidstidspunkterne er forskellige.

[1] *Antal personer i arbejdsstyrken uden erhvervsuddannelse.* Indgår i Indenrigs- og Boligministeriets beregning af fordeling af regionernes tilskud på udviklingsområdet. Tabellerne leveres af DST Consulting.

Højest fuldførte uddannelse (UDDA)

Grundpopulationen i registret med befolkningens højeste fuldførte uddannelse (UDDA) lægger sig helt op ad befolkningsstatistikkens populationer jf. det foregående afsnit.

Uddannelseskoden er den centrale variabel i registret. Der er tale om en 4 cifret kode, der benyttes i forbindelse med indberetningerne til Danmarks Statistik. Når statistikken offentliggøres, omsættes den 4 cifrede kode til DISCED-15 koder, som er Danmarks Statistiks hierarkiske klassifikationssystem for uddannelser. Nedenfor ses et eksempel, hvor den 4 cifrede kode '5082 Multimediedesigner' er indplaceret i forhold til DISCED-15 *Hovedområder*. På det mest overordnede niveau er uddannelsen defineret som værende en kort videregående uddannelse.

Se nærmere i Klassifikationsdatabasen:
https://www.dst.dk/da/Statistik/dokumentation/nomenklaturer/disced15-audd

Det betyder, at **hver** gang man skal benytte UDDA's uddannelseskoder, så skal man tage stilling til, hvordan uddannelserne skal grupperes. Ud over *Hovedområde* er det også muligt at opdele på *Uddannelsestyper*, *Uddannelsesniveau* og *Fagområde*. Hertil findes de internationale inddelinger *ISCED Level* og *ISCED Field*. [2] I det omfang, at der sker ændringer i klassifikationerne kan det påvirke mulighederne for at sammenligne med de offentliggjorte tal i Statistikbanken.

- **1981-1990:** Her findes ikke nogen offentliggjort statistik.
- **1991-2005:** En sammenligning med **HFU1** viser, at det ikke er muligt at ramme Statistikbanken. Det er ikke muligt at genskabe den klassifikation (den såkaldte *Forspalte*), der blev benyttet på det tidspunkt. Samtidigt er der tilsyneladende ikke sammenfald mellem de versioner af registret, der findes i UDDA, og dem der er benyttet i forbindelse med de offentliggjorte tal.
- **2006-2007**: Her gælder det, at UDDA's populationer passer med FAIN populationerne. Men den offentliggjorte statistik i **KRHFU1**, benytter populationen pr. 1. januar i BEF. Dette giver en uoverensstemmelse, hvis man forsøger at genskabe Statistikbankens resultater. Endvidere kan ændringer i klassifikationen for uddannelseskoden give forskelle.
- **2008-2021**: Populationerne opgøres nu pr. 1. oktober og stemmer overens med BEF populationerne jf. det tidligere afsnit. Men der har i perioden været en ændring i DISCED-15 klassifikationen, hvilket betyder, at fra 2008-2019 giver den aktuelle klassifikation et andet resultat for *H40 Kort videregående uddannelse* og *H50 Mellemlang videregående uddannelse*. Her er sammenlignet med Statistikbankens **HFUDD16**.

[2] Her kan man med stor fordel benytte de SAS formater og kodelister, som stilles til rådighed på forskermaskinerne.

Registerbaseret arbejdsstyrke (RAS)

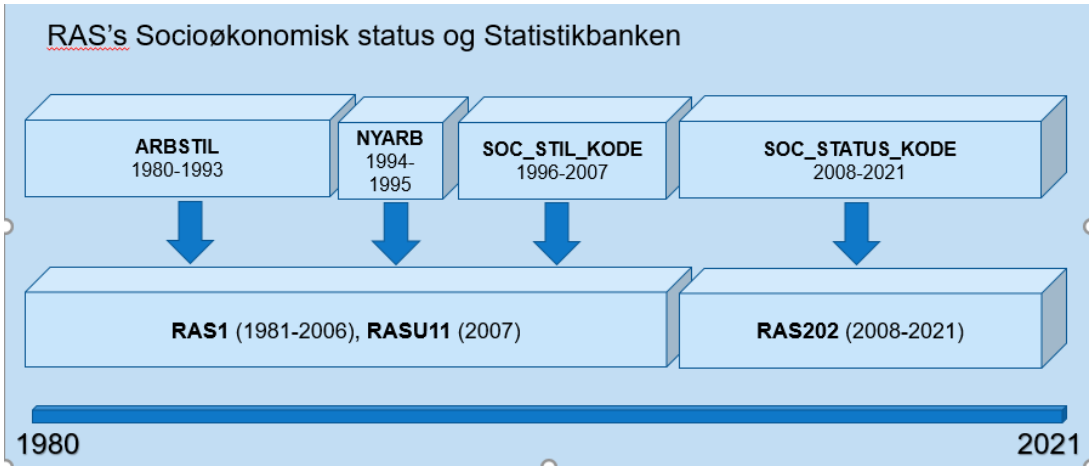
I RAS's grundpopulation indgår alle, som er i befolkningen på statustidspunktet jf. det foregående afsnit om statustidspunkt for populationerne. Derudover er der nogle variationer i grundpopulationerne, når man ser på perioden 1980-2021:

- **1980-1989:** Her er populationen identisk med den population der indgår i befolkningen (FAIN). Indeholder kun befolkningens primære beskæftigelse.
- **1990-2007:** Indeholder ikke kun den primære beskæftigelse men også sekundær og tertiær beskæftigelse samt ikke november beskæftigelse. For selektion af den primære beskæftigelse svarende til den offentliggjorte statistik bruges variablen **NOVPRIO=1**.
- **2008-2021:** For at afgrænse til den offentliggjorte statistik benyttes de to variable **PRIMAER_STATUS_KODE=1** og **I_BEFOLKNINGEN_KODE=1**.

Bemærk, at når man skal matche med FAIN, er årsangivelsen 1. januar året efter, dvs. at RAS1980 fx skal matches med FAIN1981. Bemærk endvidere, at RAS2007 med status pr. november 2007 skal matches med BEF200712 og ikke FAIN. Fra og med RAS2008 (status pr. november 2008) er der ikke længere sammenfald mellem RAS populationen og Befolkningsstatistikken jf. tidligere afsnit.

For RAS gælder, at man som udgangspunkt kan ramme de offentliggjorte tal i Statistikbanken. Der ses dog marginale afvigelser i nogle enkelte år før 2008.

En af de centrale variable i RAS angiver befolkningens tilknytning til arbejdsmarkedet. Variablen har i perioden været genstand for flere databrud og har haft 4 forskellige navne ARBSTIL, NYARB, SOCSTIL_KODE samt SOC_STATUS_KODE. Hvis man følger selektionen af RAS's grundregister (som angivet ovenfor), så ses der stor overensstemmelse i forhold til de offentliggjorte tal i Statistikbanken jf. figuren.



Boligopgørelsen (BOL)

Boligopgørelsen opgør antal boliger/personer med oplysninger om boligerne. Der tages udgangspunkt i den samme population, som findes i Befolkningsstatistikregistret pr. 1. januar i året. Boligopgørelsens fundament er en match på adresserne mellem CPR- og BBR-registret. Der er tilfælde, hvor adresserne er registreret forskelligt i de to registre, hvilket betyder, at der findes personer/boliger uden et match. Der vil dermed aldrig være en fuldstændig overensstemmelse mellem de to registre. Dette forhold er nøje beskrevet i Statistikdokumentationen her: https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/boligopgoerelsen

Når man benytter FSE's Grunddata, er der yderligere manglende match i nogle år:

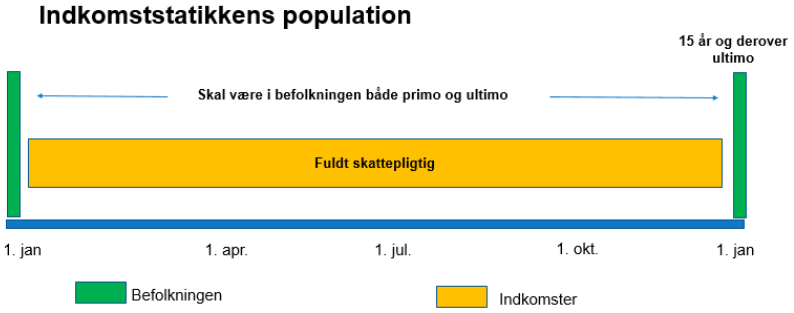
- Det er ikke muligt at genskabe de offentliggjorte tal i Statistikbanken for årene 1986-1987. Årsagen er, at man i 1988 omlagde metoden for match mellem BBR- og CPR-registret med tilbagevirkende kraft for årene 1986 og 1987.

- Det er ikke muligt at genskabe de offentliggjorte tal i Statistikbanken for årene 2005-2009 med brug af de tilgængelige data i FSE, da der mangler boligenheder i BOL.

Indkomststatistikken (IND)

Indkomstregistrets grundpopulation afspejler, at statistikken dækker et helt kalenderår. I registret indgår alle som er i befolkningen ultimo året og alle personer, der er i Skats Slutligningsregister/E-indkomst. For årerne 1981-1986 gælder dog, at personer under 15 år kun ingår, hvis de er fuldt skattepligtige og har bopæl i Danmark både primo og ultimo året. (dvs. OMFANG=1).

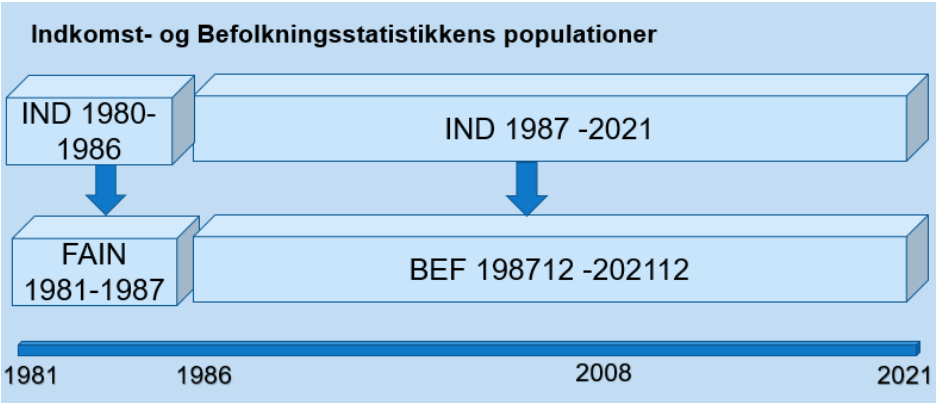
Ved offentliggørelse af den personrelaterede Indkomststatistik indsnævres populationen. Her gælder det, at den kun omfatter alle fuldt skattepligtige, der er mindst 15 år ved årets udgang og har boet i Danmark både den 31. december året før og den 31. december i indkomståret. Dvs. at statistikken ikke inkluderer personer, som er afgået ved døden i løbet i året eller migreret ind og ud af landet.



Den familierelaterede statistik inkluderes alle personer i de familier, hvor mindst én voksen person har boet i Danmark hele året og er fuldt skattepligtig. Opgørelsestidspunktet er 31. december i indkomståret. Statistikken benytter Danmarks Statistiks E-familie-begreb.

For Indkomsterne gælder, at man som udgangspunkt kan ramme de offentliggjorte tal i Statistikbanken 1987-2021. Der ses dog en mindre afvigelse i nogle enkelte år (2008-2009).

Som det fremgår, læner Indkomststatistikken sig i høj grad op af Befolkningsstatistikkens populationer. I forbindelse med en større revision af Indkomststatistikken i 2015 blev populationerne revideret tilbage for hele perioden 1981-2013. Det er disse reviderede indkomstdata, der findes i datasættet **IND –Indkomst.** Her har man benyttet den befolkning, som ligger i **BEF** fra og med 1987. Der er dermed ikke overensstemmelse med den offentliggjorte Befolkningsstatistik i perioden 1981-2007, som er baseret på **FAIN** jf. det tidligere afsnit.

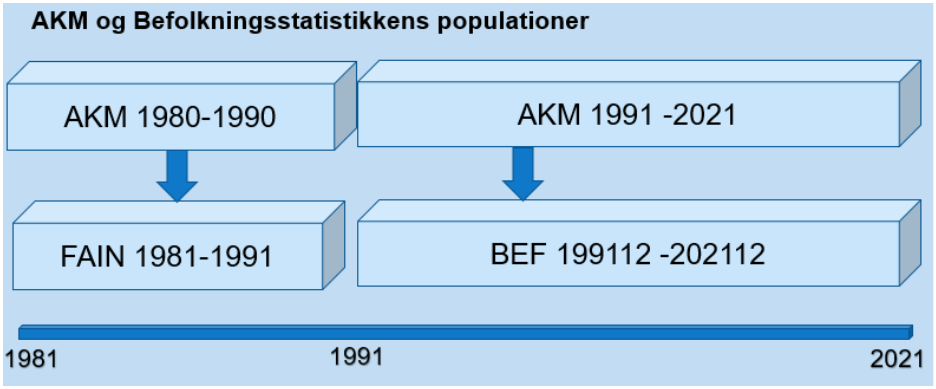


Den reviderede Indkomststatistik er nærmere beskrevet i bilag til Højkvalitetsdokumentation:

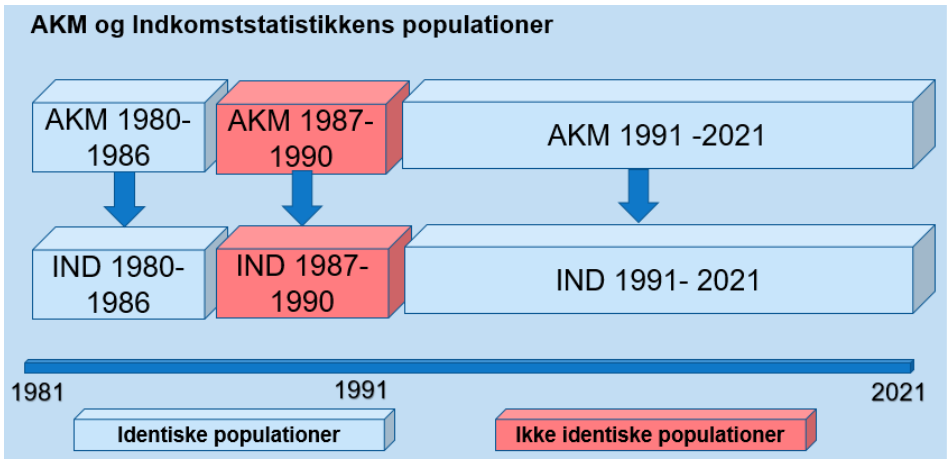
https://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable/personindkomster

Arbejdsklassifikationer (AKM)

Registret med Arbejdsklassifikationer AKM hænger populationsmæssigt sammen med Indkomstregistret. Populationerne er som udgangspunkt identiske, men historisk ses der afvigelser før 1991. Det skyldes, at AKM populationen i perioden 1980-1990 er baseret på FAIN populationen, hvor Indkomsterne (IND) skifter fra FAIN til BEF i 1987.



Derfor vil der være personer, der ikke matcher i årerne 1987-1990, når man sammenligner populationerne fra IND og AKM, jf. figuren.



For alle øvrige år er der stor overensstemmelse mellem populationerne i de to registre bortset fra indkomstårerne 2017 og 2018, hvor der ses en forskel.

For AKM gælder det, at man ikke kan ramme de offentliggjorte tal i Statistikbanken helt præcist, hvilket skyldes, at AKM skal linkes med IND for at foretage den tabellering, der ses i Statistikbanken. Derved påvirkes resultatet af, at de to populationer ikke matcher i alle år (jf. ovenfor), og af det forhold, at nøglen i FSE er CPRNR, hvor det ikke er helt overensstemmelse mellem registrene jf. afsnittet om nøglerne CPRNR og PERSON_ID.

Historiske vandringer (VNDS)

I VNDS ligger oplysninger om befolkningens ind- og udvandringer. Registret kan ikke sammenlignes med de offentliggjorte tal i Statistikbanken. Dette skyldes, at i Statistikbanken benyttes registreringsdatoen i CPR, når den årlige statistik bliver offentliggjort, hvor der er forsinkelser i registreringen af ind- og udvandring. Dette forhold er beskrevet nærmere i Statistikdokumentationen:

Statistikken baserer sig på registreringsdatoen i Danmarks Statistiks befolkningsregister for henholdsvis ind- og udvandringen. Dette skyldes, at der især i forbindelse med udvandringer i en del tilfælde kan være relativt store forsinkelser i registreringerne i CPR- registeret. Ved at opgøre ind- og udvandringer efter registreringstidspunkt opnås, at niveauet for vandringerne bliver bedst muligt, mens enkelte personers vandringer kan indgå i opgørelsen i et senere år, end vandringen egentlig har fundet sted. Danmarks Statistiks opgørelsesmetode prioriterer således, at niveauet for vandringer er så korrekt som muligt, men konsekvensen heraf er, at det faktiske hændelsestidspunkt ikke altid er korrekt på individniveau. For indvandringernes vedkommende er ca. 98 pct. rettidige indberettet til CPR i 2007. For udvandringernes vedkommende er ca. 83 pct. rettidige indberettet til CPR i 2007. I løbet af de seneste år har denne procent for indvandringernes vedkommende været status quo og for udvandringernes vedkommende svinget mellem 83 og 87 pct.
Se: ***https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/befolkningen***

I VNDS benyttes i stedet **det faktiske hændelsestidspunkt,** og oplysningerne tilpasses hvert år med de senest oplysninger. Der kan også ske rettelser flere år bagud i tid. VNDS er dermed mere retvisende i forhold til befolkningens faktiske ind- og udvandringer.

Indvandrere og Efterkommere (IEPE)

I IEPE ligger oplysninger om indvandrere og efterkommere for hele befolkningen. Der er tale om et forløbsregister, der indeholder samtlige personer, som findes i Danmarks Statistiks personstatistiske database. IEPE opdateres én gang årligt med de mest aktuelle oplysninger. Populationen kan ikke sammenlignes med en offentliggjort statistik, men til gengæld kan man inde oplysninger om indvandrerbaggrund for en person uanset hvilken population, der kunne tænkes at være genstand for analyse.

Integreret database for arbejdsmarked (IDAP)

IDA databasen er målrettet mod forskningsverdenen og danner ikke basis for en officiel statistik. I IDA er det muligt at følge personer, arbejdssteder og virksomheder over tid. Det er dermed ikke muligt at sammenligne sine resultater med Statistikbanken. Det gælder derimod, at populationen er helt identisk med RAS's population af personer. Der henvises til bilag i højkvalitetsdokumentationen:

https://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable/ida-arbejdssteder

IDA databasen består af flere datasæt. Ved match mod de øvrige personregistre er det datasættet **IDAP - IDA persondata,** som skal benyttes. En årlig match mellem IDA's og RAS's population giver dette resultat:

- **1980-2007:** Her er de to populationer helt identiske, og der opnås 100 pct. match.
- **2008-2021:** Der er en mindre afvigelse mellem de populationer, som kan tilskrives konverteringen fra PERSON_ID til CPRNR jf. senere afsnit 'Nøglerne CPRNR og PERSON_ID'.

Elevregistret (KOTRE)

Elevregistret (KOTRE) blev etableret helt tilbage i starten af 1970'erne, og er et forløbsregister, hvor man kan følge de enkelte studerende gennem deres uddannelseskarriere. Det gælder, at registret er levende, og der kan forekomme rettelser langt bagud i tid. Det betyder også, at man kun kan genskabe de offentliggjorte tal i Statistikbanken for de seneste to år, men ikke nødvendigvis for de andre år.

- **De seneste to år:** Statistikbanken opdateres med oplysningerne i den seneste version af Elevregistret. Bemærk dog, at **H80 Ph.d. og forskeruddannelser** ikke udstilles i Statikbanken det seneste år, men at disse elever findes i registret.
- **Tidligere end de seneste to år:** Her er ikke nødvendigvis overensstemmelse, idet der kan være rettelser i Elevregistret, som ikke rettes i Statistikbanken.

I forhold til brugen af uddannelseskoden som den centrale variabel i registret henvises til afsnittet med Højest fuldførte uddannelse (UDDA). For at danne de offentliggjorte populationer mht. **tilgang**, **fuldførte** og **status** kræves en algoritme, som fremgår af **bilag 1.**

Samlet oversigt, kan man ramme Statistikbanken?

Grunddata	Revision af registret?	Kan man ramme Statistikbanken?
Befolkningen (**BEF** & **FAIN**)	Revideres aldrig	**Ja** Ret præcist
Højeste fuldførte uddannelse (**UDDA**)	Registret revideres kun et år bagud i tid. Ændringer i klassifikationen kan dog betyde, at man ikke altid kan genskabe statistikbanken bagud i tid.	**Ja/Nej** Kun de seneste år passer
Registerbaseret arbejdsstyrke (**RAS**)	Ved fejl revideres bagud i tid. Statistikbanken ændres samtidigt.	**Ja** Marginale afvigelser i enkelte år
Boligopgørelsen (**BOL**)	Revideres aldrig	**Ja/Nej** Der ses afgivelser i nogle enkelte år
Indkomster (**IND**)	Blev revideret for hele perioden i 2015. Statistikbanken blev samtidigt ændret.	**Ja** Små afvigelser i enkelte år. Man kan dog ikke ramme mht. AKM variablene.
Arbejdsklassifikation (**AKM**)	Blev revideret tilbage til 1991 i 2015. Statistikbanken blev samtidigt ændret.	**Nej** CPRNR nøglen giver problemer i forhold til IND.
Vandringer (**VNDS**)	Revideres hvert år	**Ikke relevant** Der offentliggøres ikke statistik baseret på **VNDS**
Indvandrere og Efterkommere (**IEPE**)	Revideres hvert år	**Ikke relevant** Der offentliggøres ikke statistik baseret på **IEPE**
Integreret database for arbejdsmarked (**IDA**)	Revideres lejlighedsvis fx ved kvalitetsforbedringer af variablene.	**Ikke relevant** Der offentliggøres ikke statistik baseret på IDA.
Elevregistret **(KOTRE**)	Revideres hvert år	**Nej /Ja** Kun de to seneste år kan sammenlignes med Statistikbanken.

Nøglerne CPRNR og PERSON_ID

Der findes to personstatistiske datasamlinger i Danmarks Statistik hhv. *Grunddata* i FSE og *Den Personstatistiske Database*, PSD. PSD blev implementeret i løbet af 00'erne og efter 2007 har FSE hentet mange af deres data i PSD, det gælder fx BEF, UDDA, RAS, IND og KOTRE.

I datasamlingerne arbejdes der med to forskellige nøgler. CPRNR er den nøgle der benyttes i FSE's Grunddata, mens PERSON_ID benyttes som nøglen i PSD. PERSON_ID er Danmarks Statistiks unikke nøgle for en person, idet nøglen fastholdes ved skift i CPRNR.

I det omfang FSE henter deres grunddata i PSD, konverteres nøglen fra PERSON_ID til CPRNR. Her benyttes den aktuelle konverteringsnøgle, og da den aktuelle nøgle ændrer sig fra dag til dag, vil der ikke være 100 pct. match mellem registrene, selvom de har samme population og samme statustidspunkt. Bemærk derfor, at konverteringen påvirker brugernes muligheder for at opnå fuldstændig match registrene imellem. Hvis et projekt/projektdatabase fx får en genleverance af historiske data, kan der også være foretaget en konvertering, som påvirker mulighederne for at genskabe den oprindelige match registrene imellem.

Det påvirker mulighederne for at ramme Statistikbanken helt præcist i de tilfælde, hvor genskabelse af en Statistikbanktabel kræver samkøring af flere registre. Det gælder fx, når man skal samkøre Indkomsterne (IND) og Arbejdsklassifikationerne (AKM), for at sammenligne med den offentliggjorte Indkomststatistik.

Bilag 1. Algoritme i forbindelse med brugen af Elevregistret.

```
*-------------------------------------------------*
  Trækker data fra Elevregistret i en årrække
  og opdeler på status svarende til opdelingen i
  Statistikbanken.
  - Elever pr. 1. okt.
  - Fuldførte
  - Tilgang
*-------------------------------------------------;
*----------------------------------*
  Her ligger data.
*----------------------------------;
libname Data '\\XXXXXXXXXXX\KOTRE';

*----------------------------------*
 Makrovariable
*----------------------------------;
%let aar_start=2019;      /* Det første skoleår der skal være i udtrækket */
%let aar_slut =2022;      /* Det sidste skoleår der skal være i udtrækket */
%let ref_aar=2022;        /* Referenceår for elevregisteret */

*----------------------------------*
 Udvælger alle forløb, der er
 uafsluttede eller afsluttet efter
 starten på det første skoleår
*----------------------------------;
proc sql;
create table kotre as
select    *
from    data.kotre&ref aar
where   ELEV3_VTIL>"30SEP%eval(&aar_start.-1)"d;
quit;

*----------------------------------*
 Dan status og hovedgruppe for
 uddannelsen. Bemærk, at der lægges
 92 dage til datoerne for at ramme
 skoleåret i Statistikbanken.
*----------------------------------;
data kotre2;
set kotre;
length status $50. aar 4. Uddannelse $100.;

/* Et forløb er fuldført, når afgangsarten starter med 1 (ikke bare skift af institu-
tion/trin),
og forløbet er kompetencegivende */
if 0<audd<9999 and AFG_ART in ('1','11') and komp in (1,3,7) then do;
status='Fuldført';
aar=year(ELEV3_VTIL+92);
Uddannelse=put(udd,UDD_HOVED_L1L5_KT.);
if &aar_start.<=aar<=&aar_slut. then output;
end;

/* Der er tale om en tilgang, når tilgangsarten starter med 1 - der er ikke tale om
skift imellem trin
og/eller institution */
if TILG_ART in ('1','11') then do;
status='Tilgang';
aar=year(ELEV3_VFRA+92);
Uddannelse=put(udd,UDD_HOVED_L1L5_KT.);
if &aar_start.<=aar<=&aar_slut. then output;
end;

/* Laver en løkke over hvert år i forløbet, dog tidligst det første år og senest det
sidste år, som kigges på */
/* Tæller forløbet som igangværende, såfremt forløbet inkluderer 1. oktober i det på-
gældende år i - mdy(10,1,i) giver SAS-datoen for den 1. i måned 10 for år i */
do i=max(year(ELEV3_VFRA+92),&aar_start.) to min(year(ELEV3_VTIL+92),&aar_slut.);
if ELEV3_VFRA<=mdy(10,1,i)<=ELEV3_VTIL then do;
status='Elever pr. 1. oktober';
aar=i;
Uddannelse=put(udd,UDD_HOVED_L1L5_KT.);
output;
end;
end;
run;
```

```
*-----------------------------------*
 Tabel
*-----------------------------------;
proc tabulate missing;
class aar status uddannelse;
table status='Status'*aar='Skoleår',
      all='I alt'*n=''*f=commax12. uddannelse='Uddannelse'*n=''*f=commax12. / mis-
stext='0';
title1 'Antal elever fordelt på status, skoleår og uddannelse';
title2 "Opgjort på baggrund af Elevregistret (KOTRE) &ref_aar";
title3 "Metoden svarer til den der benyttes i forbindelse med de offentliggjorte tal i
Statistikbanken fx. UDDAKT10";
run;
```