

Dokumentation af CPR-Stikprøve og vægtning

Population og stikprøve

Opgave	CPR-stikprøve til levering
Kunde	Megafon
Delopgaver udført af:	
Population	Danmarks Statistik
Stikprøve	Danmarks Statistik
Population	4.758.579 personer
Stikprøve	78.625 personer.

Særligt for denne undersøgelse

Population Populationen er dannet på baggrund af Befolkningsregistret pr. 30 juni 2023. Populationen består af personer over 18 år pr. 1. januar 2023 og indeholder 4.758.579 personer.

Stikprøve Stikprøven er udtrukket stratificeret på baggrund af adressekoder leveret af Megafon. Adressekoderne inddeles i 198 boligområder, kaldet fokusområder, samt generelle almen sektor. Stratificeringen består således af 200 strata:

- 198 boligområder
- Den resterende almene sektor
- Resten af populationen

Stikprøvestørrelserne i hvert stratum er bestemt af Megafon.

Personer som er flyttet, døde eller udvandrede siden befolkningsregistret blev dannet (30. juni 2023) kan ikke udvælges.

Den samlede stikprøve består af 77.375 personer.

Adressebeskyttede er leveret med i nettostikprøven uden adresseoplysninger. I det leverede dataudtræk findes variabelen *kontakt*, som angiver, om den pågældende person kan kontaktes jf. følgende inddeling:

- 1) Kan deltage
- 2) Adressebeskyttet
- 3) DST-beskyttet (har frabedt sig at være med i Danmarks Statistiks undersøgelser)
- 4) Ugyldig adresse

Det er Danmarks Statistiks anbefaling, at der kun tages kontakt til de personer, der er tildelt kategorien 1) *Kan deltage* i udtrækket.

Variabelliste Der er leveret en kodebog med variabelliste sammen med data.

Populationstabel Der er leveret en populationstabel sammen med data, som viser fordelingen af populationen og stikprøven på udvalgte baggrundsvariable, som fx køn, alder, herkomst og uddannelse.

Dokumentation af vægtning

Efter dataindsamling foretaget af Megafon, foretages en vægtning af swardata hos DST Survey. For at genskabe repræsentativiteten efter et evt. skævt bortfald er der lavet en bortfaldsanalyse hos DST Survey. På baggrund af denne laver vi en vægtning til populationen, så det er muligt at udtale sig om hele populationen generelt.

Megafon har leveret en fil med de respondenter, der skal indgå i vægtningen. I alt 26.088 personer.

Bortfaldsanalysen består af en regressionsanalyse på baggrund af Danmarks Statistiks registre, hvor det undersøges hvilke baggrundsvariable der har haft betydning for bortfaldet. Ved hjælp af disse er der konstrueret vægte via CLAN metoden, som knytter sig til hver respondent og sørger for, at swardata igen bliver repræsentativt for den samlede population på 4.758.579 personer.

Generelt fælles for alle strata var at baggrundsvariable som alder, køn, højeste fuldførte uddannelse, familietype, indkomst og herkomst var signifikante for bortfaldet. Det tilstræbes altid at få så mange af disse baggrundsvariable med i vægtningsmodellen som muligt. I tilfælde af prioritering af baggrundsvariable i modellen, tilstræbes det at inkludere så mange alsidige variable som muligt. Dvs. at det prioriteres at inkludere både geografiske, demografiske og socioøkonomiske baggrundsvariable. I aktuelle undersøgelser indgår den geografiske komponent allerede i stratificeringen.

Vægtene er dannet på baggrund af to forskellige modeller:

- En for fokusområderne samt den generelle almene sektor
- En for den øvrige population

For fokusområderne og den almene sektor indgår strata, alder, køn og højeste fuldførte uddannelse i følgende grupperinger:

- Alder:
 - 18-34 år
 - 35-49 år
 - 50-64 år
 - 65 år og derover
- Højeste fuldførte uddannelse:
 - Grundskole eller uoplyst
 - Gymnasial, erhvervsfaglig (EUD), kort videregående uddannelse
 - Mellemlang videregående uddannelse, bachelor, lang videregående uddannelse, Ph.d

I forbindelse med vægtningen af fokusområderne var det nødvendigt at slå strata 6 (Ågården) sammen med et andet, da få vægte i dette boligområde blev meget høje. I vægtningsmodellen blev strata 6 derfor slået sammen med strata 8 (Aldersrogade). Valget af strata 8 bygger på en sammenligning af strata 6 med de restende fokusområder i Københavns Kommune med hensyn til fordelingen af køn, alder, herkomst, uddannelsesniveau, socioøkonomisk status, familietype samt indkomst.

I vægtningsmodellen for den øvrige population indgår køn, alder, højeste fuldførte uddannelse, herkomst, familietype samt familiens disponible indkomst. Følgende grupperinger er anvendt:

- Alder:
 - 18-29 år
 - 30-39 år
 - 40-49 år
 - 50-59 år
 - 60-69 år
 - 70 år og derover
- Højeste fuldførte uddannelse:
 - Grundskole eller uoplyst
 - Gymnasial, erhvervsfaglig (EUD), kort videregående uddannelse
 - Mellemlang videregående uddannelse, bachelor, lang videregående uddannelse, Ph.d
- Herkomst
 - Dansk oprindelse
 - Indvandre eller efterkommer
- Familietype
 - Enlige uden børn
 - Enlige med børn
 - Par uden børn
 - Par med børn
- Familiens disponible indkomst (i 1.000 kr)
 - Op til 150
 - Fra og med 150 til 250
 - Fra og med 250 til 350
 - Fra og med 350 til 500
 - Fra og med 500

I det returnerede data inkl. vægte er følgende variable inkluderet:

- Personnummer
- Strata
- NpopH (Populationen i strata)
- NrespH (Antal svar i strata)
- Dirpervgt (Direkte vægt, dvs. NpopH/NrespH)
- Pervgt (Personlig/beregnet vægt)

Generelt om population og stikprøve

*Skal have bopæl i
Danmark*

Populationen til CPR-stikprøver dannes udelukkende ud fra oplysninger, der både findes i CPR-registeret og Danmarks Statistiks statusregister for hele befolkningen. Danmarks statistik opdaterer statusregisteret på personniveau hvert kvartal, ca. 40 dage efter kvartalsafslutning. Det vil for eksempel sige, at status for befolkningen pr 1. januar er klar i starten af februar.

For at tælle med i den danske befolkning, skal man have bopæl i Danmark og have tildelt et CPR-nummer.

Kvalitet af cpr

CPR's befolkningsstatus er ret præcis. Usikkerheden på den samlede befolkning er få promille. Kun ved undersøgelser, hvor personer har flyttet ind og ud af Danmark, kan der være en større usikkerhed.

<i>Personer som flytter til udlandet</i>	Personer, som flytter ud af Danmark, er ikke tvunget til at lade det registrere nogen steder. Der er derfor en mindre gruppe personer, som stadig er registret som en del af den danske befolkning, på trods af at de ikke længere har adresse i Danmark.
<i>Udenlandske studerende</i>	Uddannelsessøgende skal have et CPR-nummer for at blive optaget på en dansk uddannelse. De skal dog også være optaget på studiet før de kan få opholdstilladelse, og der kan derfor være et mindre antal personer, som ikke tælles med.
<i>Adressebeskyttelse</i>	Omkring 1,9 pct. af den danske befolkning har adressebeskyttelse, hvilket betyder, at de ikke deltager i interviewundersøgelser fra Danmarks Statistik. Disse personer leveres dog med i nettostikprøven. Desuden er der personer i Danmark som ikke kan kontaktes pga. at der ikke er oplysninger om deres adresse.
<i>Personer under 18 år</i>	For personer under 18 år er forældreoplysninger leveret med i stikprøven, således at det er muligt, at orientere forældre til den udvalgte person, hvis det ønskes.
<i>Sikring af repræsentativitet i undersøgelser</i>	Danmarks Statistik danner populationer med udgangspunkt i Befolkningsregisteret (CPR). I denne proces anvender vi et unikt person-ID for alle personer (en anonym CPR-nøgle). Vi tester og fejlsøger data for at sikre at hver person kun optræder én gang i vores population. Efter at vi har lavet stikprøvedesignet, danner vi stikprøven ved hjælp af SAS Proc SurveySelect. Her anvender vi SRS-metode (Simple Random Sample). Denne metode giver samme tilfældige udvalgschance (uden tilbagelægning) for alle deltagere. Metoden danner en ny tilfældig seedning hver gang og dermed en sikring af, at det ikke er de samme personer, som bliver udvalgt fra populationen gentagene gange. Du kan læse mere om Proc SurveySelect på https://support.sas.com/.
<i>SAS-koden</i>	Kodeeksempel: <pre>proc surveyselect data=pop_select sampsize=10050 method=srs out=stikip OUTSIZE STATS; run;</pre>
<i>Udvalgssandsynligheden</i>	Det er muligt at udvælge en stratificeret stikprøve, hvor udvalgssandsynligheden bliver forskellig i hvert stratum, afhængigt af populations – og stikprøvestørrelse i hver. Vi gemmer altid udvalgssandsynligheden og den direkte vægt for hver udtrukne person.
<i>Test og kvalitetssikring</i>	Efter dannelse af stikprøven tester og fejlsøger vi for dubletter. Vi laver tabeloversigter på hhv. population og stikprøve for at se om sammensætningen er nogenlunde ensartet på en række baggrundsvARIABLE (køn, alder, herkomst, uddannelse, indkomst mv.) Vi validerer herefter stikprøven op mod den aktuelle CPR-befolkning (få dage gammel) for at sikre os, at vi har korrekte oplysninger på alle deltagere, samt markering af beskyttede, døde og udvandrede personer. Ved at anvende denne metode og tilhørende fejlsøgningsproces sikrer vi os at vores stikprøver er repræsentative for den population de er udtrukket fra.

Danmarks Statistiks begreber

<i>Populationen</i>	Vores definition af population er den danske befolkning ifølge seneste kvartals-befolkningsstatus, evt. afgrænset på bestemte oplysninger fra CPR-registeret efter kundens ønske, fx alder på en bestemt dato. Afgrænsningen for den enkelte undersøgelse fremgår af dokumentationen.
<i>Bruttostikprøve</i>	Den udvalgte bruttostikprøve fra populationen indeholder alle, der kan udvælges på det tidspunkt statusregisteret blev dannet.

Nettostikprøve Nettostikprøven indeholder personer, der kan kontaktes efter seneste cpr-opdatering. Dvs. at bruttostikprøven valideres mod den seneste opdatering af cpr-registret for at danne nettostikprøven, som afspejler den aktuelle befolkning.

Adressebeskyttede personer er med i nettostikprøven, men uden adresseoplysninger. Personer, som står med 'Rådhusadresser', dvs. hjemløse og personer uden fast adresse mv., indgår ligeledes uden adresseoplysninger.