Dataskydd och kontrollförfarandet för resultat

Denna anvisning redogör för dataskyddet för material som är i forskningsanvändning och i synnerhet för utskrifter som producerats utifrån det (tabeller, figurer, statistiska modeller o.d.) och kontrollförfarandet för utskrifter.

Dataskyddsanvisningarna gäller för användning av material såväl i forskningsprojekt som i mikrosimuleringar. Kontrollförfarandet för utskrifter vid mikrosimuleringar avviker dock från förfarandet vid övriga forskningsprojekt.

Vid frågor kring dataskyddsärenden, kan forskaren kontakta forskartjänsterna (tutkijapalvelut@stat.fi / mikrosimulointi@stat.fi).

Dataskydd för forskningsmaterial och utskrifter

Såväl Statistikcentralen som forskarna har till uppgift att sköta dataskyddet för forskningsmaterial. Statistikcentralen sköter för egen del dataskyddet av materialet innan det överlåts för forskningsanvändning och datasäkerheten i distansanvändningsmiljön. Forskaren ska för egen del sköta dataskyddet av sitt material under tiden för forskningsanvändningen och vid publicering av forskningsutskrifter.

Forskaren ansvarar för att dataskyddet genomförs i de forskningsresultat som denne publicerar. Dessa instruktioner har uppgjorts för att främja ansvarsfullt förfarande i dataskyddsärenden och syftet med instruktionerna är att hindra såväl oavsiktliga som avsiktliga dataskyddsförseelser. Forskartjänsterna använder ett kontrollförfarande för forskningsresultat, med vilket också aktsamhet i dataskyddsfrågor främjas. Kontrollförfarandet redogörs närmare i avsnittet Kontrollförfarande för utskrifter.

Varför dataskydd?

En förutsättning för att Statistikcentralen ska kunna lämna ut forskningsmaterial som samlats in för statistiska ändamål för forskningsanvändning och mikrosimulering är att dataskyddet är i ordning. Statistikcentralens rätt att samla in register- och enkätmaterial för statistikföring har tryggats i statistiklagen, såsom också rätten att lämna ut dessa uppgifter för forskningsanvändning. Å andra sidan innehåller lagen också ett åliggande att se till att uppgifterna skyddas på sakligt sätt och en del av materialet innehåller också väldigt känsliga uppgifter.

Behandlingen av personuppgifter styrs av Europeiska unionens allmänna dataskyddsförordning (EU) 2016/679 och den nationella dataskyddslagen (1050/2018), som kompletterar den. Enligt sekretessplikten i anknytning till användningstillståndet för ett forskningsprojekt ska forskaren se till att forskningsresultaten inte innehåller uppgifter på enhetsnivå, det vill säga uppgifter om en enskild person eller ett enskilt företag eller möjlighet att dessa röjs. Vid behandling av material som överlämnats för forskningsanvändning eller vid förfarande i distansanvändningsmiljön, ska en forskare enligt sekretessplikten också se till att materialet inte röjs eller överlåts till någon som saknar rätt att använda det.

Dataskyddsanvisningar för olika utskriftstyper

Nedan presenteras anvisningar och regler för att sköta dataskyddet av olika utskriftstyper. Eftersom det finns många olika forskningsprojekt och utskriftstyper, är det inte möjligt att bedöma och ge anvisningar om dataskyddet för varje material och utskrift separat. Därför är det nödvändigt att ge allmänna ”tumregler”.

I fråga om vissa av Statistikcentralens material och material som utlämnats av andra myndigheter för forskningsanvändning kan dataskyddsanvisningarna avvika från nedan presenterade regler. Dessa avvikande regler finns i materialbeskrivningarna i Taika-katalogen, och vid behov antecknas de också i beslutet om användningstillstånd.

Om forskaren är osäker på dataskyddet för något resultat eller vill begära preciseringar av de givna instruktionerna, ska denne kontakta forskartjänsterna (tutkijapalvelut@stat.fi).

Utskrifter som producerats från urvalsmaterial

Det är gemensamt för alla utskriftstyper att risken för röjande av observationer bakom resultaten i allmänhet är mindre, om det använda materialet är enbart ett (slumpmässigt) urval av totalmaterialet. En mindre risk för röjande vilken eventuellt beror på urvalet innebär dock inte att de dataskyddsregler som presenterats i denna anvisning automatiskt kan lindras för utskrifter som producerats utifrån urvalsmaterial.

Om det dock är nödvändigt att göra en fallspecifik bedömning av risken för att en observation röjs i någon utskrift (till exempel om ett fördelningsnyckeltal är avslöjande eller om det är möjligt att identifiera en observation bakom en enskild bildpunkt i en bild), påverkar användningen av urval i så fall bedömningen.

Det ska observeras att totalmaterial inte alltid betyder enbart alla personer som bor i Finland eller alla företag som verkar i Finland. Med tanke på dataskyddet och bedömningen av risken för röjande av observationer, bildar alla företag till exempel i en viss bransch snarare ett totalmaterial än ett urvalsmaterial, eftersom det i allmänhet är möjligt att dra en slutsats om ett enskilt företag hör till det ”urval” som avgränsats till att omfatta alla företag inom en viss bransch.

Frekvens- och mängdtabeller

De uppgifter som aggregerats i tabellform kan innehålla personuppgifter, företagsuppgifter eller bägge två. Till exempel i arbetstagar-arbetsgivar-material ska såväl person- som företagsnivån skyddas. Också de uppgifter som beskriver yrkesgrupper kan indirekt innehålla företagsuppgifter, om (nästan) alla personer som hör till en viss yrkesgrupp tjänstgör enbart för ett (monopol)företag, då dataskyddet för företaget i fråga ska tryggas.

I regel är tröskelvärdet 3 i skyddet av tabeller som innehåller såväl företags- som personuppgifter. Med andra ord är det tillåtet att publicera tabellceller eller uppgifter i en observationsgrupp enbart om åtminstone 3 (oviktade) observationer

finns bakom uppgifterna. Inte heller uppgifter om antalet observationer får publiceras vad gäller små celler eller grupper.

Användning av regeln om tröskelvärde som huvudregel är det enklaste sättet att räkna en potentiell risk för röjande av enskilda observationer som hänför sig till de tabeller som ska publiceras. Identifiering av en observation som det enda eller ett sällsynt fall i en grupp kan öka risken att identifiera observationen och/eller röja fler uppgifter om denna observation (till exempel med andra tabeller eller utskrifter som behandlar samma material eller ämne). Därför ska tröskelvärdet tillämpas, trots att enbart mängduppgifter publiceras i en tabell.

Utöver ovan nämnda huvudregel ska följande regler beaktas:

- I färska företagsuppgifter (nyare än 15 månader från referenstidpunkten) ska man utöver tröskelvärdet tillämpa dominansregeln[1] (1,75). Det kan förutsättas att dominansregeln används också i äldre företagsuppgifter eller i annat material (till exempel inkomstregistret). Dominansregeln förutsätter att man ska skydda till exempel uppgifter där ett företags omsättning eller lönebelopp eller en annan variabel i materialet står för över 75 procent av företagens omsättning/lönebelopp eller dylikt i en viss bransch (sektor/område etc.) eller om till exempel över 75 procent av löntagarna i en viss yrkesgrupp arbetar hos ett visst företag.
- Vid skydd av uppgifter på arbetsställesnivå ska man också i mån av möjlighet säkerställa skyddet på företagsnivå, det vill säga att det i varje cell ska finnas arbetsställen från minst tre olika företag. Förfarandet ska vara det samma för koncern-företag-förhållanden.
- I uppgifterna om varor (produkter, material och förnödenheter i industrins produktionsstatistik) är antalet företag en konfidentiell uppgift i alla produktionsbenämningar.
- Dataskyddet av personuppgifter ska skötas särskilt noggrant, om tabellen innehåller känsliga personuppgifter[2] (inkl. de särskilda kategorier av personuppgifter som räknats upp i EU:s allmänna dataskyddsförordning[3]). Vid skydd av känsliga uppgifter kan det vara nödvändigt att använda ett högre tröskelvärde.
- Vad gäller yrkesutövare som förekommer i företagsuppgifterna tillämpas samma skyddsregler som för övriga företagsuppgifter, trots att yrkesutövare i princip är personer.
- Publicering av koordinat- eller rutuppgifter är förknippade med särskilda villkor: Rutmaterial i sin helhet får inte publiceras. Uppgifter som baserar sig på rutor får tas ut ur distansanvändningssystemet och publiceras då det finns minst 10 observationsenheter i en ruta. Den information som ska

[1] Enligt dominansregeln (n,k) kan tabellens cellvärde inte publiceras, om dess n största observationer utgör minst k procent av cellens totala värde.
[2] Uppgifter som är känsliga avser här uppgifter som beskriver en persons ras eller etniska ursprung, politiska åsikter, religiösa eller filosofiska övertygelse, medlemskap i fackförbund, hälsa, sexuella beteende eller inriktning, dom i brottmål och förseelser eller säkringsåtgärder i anknytning till dem, dödsorsak, språk, nationalitet, härkomst eller födelseland, inkomster, skulder, förmögenhet, ovanliga yrke eller annan socioekonomisk ställning, behov av socialvård eller erhållna socialvårdstjänster, stödåtgärder eller andra förmåner inom socialvården, eller uppgifter om vårdåtgärder eller därmed jämförbara åtgärder som gäller personen.
[3] Europaparlamentets och rådets förordning (EU) 2016/679, artikel 9

publiceras måste endera sammanställas till en större regional nivå, ges som relationstal eller behandlas på annat sätt så att personer och bostadshushåll inte kan identifieras områdesnivå eller på en karta.

Olika fördelningsnyckeltal

Minimi och maximi är i allmänhet förknippade med en observation, så de kan oftast inte publiceras. Till exempel det största företaget i en bransch kan i allmänhet identifieras, varför relaterade uppgifter om maximiomsättningen inte kan publiceras. Om minimi eller maximi inte kan förenas med en enskild observation eller om tröskelvärdet överskrids, kan de publiceras.

Fördelningspunkterna (exkl. minimi och maximi), såsom deciler, utgör ett specialfall av tabellen där antalet observationer som blir mellan fördelningspunkterna motsvarar cellfrekvenserna. Om dessa antal överskrider det tröskelvärde på 3 som ska tillämpas i tabellerna, kan fördelningspunkterna publiceras.

Typvärdet kan publiceras, om (nästan) alla observationer inte får samma värde.

Medeltalet, andra relationstal och de högre momenten i fördelningsnyckeltalen (till exempel varians) kan publiceras om man vid beräkningen av dem har använt minst tre observationer.

Vid publiceringen av andelar ska tröskelvärdet 3 överskridas vad gäller alla grupper som bildar andelar. Med andra ord, om man vill publicera till exempel att kvinnornas andel är 58 % av hela populationen, ska dessa 58 %, liksom också männens 42 % innehålla åtminstone tre personer. Det är med andra ord inte tillräckligt att det i hela populationen finns åtminstone totalt 3 kvinnor och män.

Andra numeriska utskriftstyper

Indextal, korrelationskoefficienter och teststorheter (t, F, X^2 o.d.) kan i allmänhet publiceras, om åtminstone 10 observationer använts i beräkningen av dessa.

Regressionsmodeller kan publiceras i sin helhet om det ligger tillräckligt med observationer bakom modellen (åtminstone 10) och modellen inte beskriver tidsserien för observationer av ett företag eller en person. Enskilda koefficienter i modellen kan i allmänhet alltid publiceras.

Numeriska resultat i mer komplicerade statistiska modeller kan i allmänhet alltid publiceras, om tillräckligt med observationer finns bakom modellen och modellen inte beskriver tidsserien av observationer om ett enskilt företag eller en enskild person.

Bilder

I likhet med numeriska utskrifter, får inte heller bilder avslöja uppgifter om enskilda observationer. Bilder som ritats av materialet är tillåtna, om en enskild

bildpunkt eller en del av bilden inte kan avslöja en enskild observation bakom den.

Det är vanligtvis möjligt att publicera stapeldiagram och andra figurer för presentation av klassificerat material, så länge det finns en tillräcklig mängd observationer i varje klass. Information i en sådan bild kan i allmänhet också presenteras i tabellform och det är möjligt att på den tillämpa samma dataskyddsregler som på annat tabellmaterial (se ovan i punkten Frekvens- och mängdtabeller).

Fördelningar, histogram och kumulativa fördelningsfunktioner som har utjämnats från fördelningsbilder eller presenterats på en tillräckligt grov nivå är tillåtna. En del av fördelningsbilderna kan innehålla uppgifter om avvikande observationer eller extremvärden, vilka från fall till fall kan röja uppgifter om en enskild observation. Programmens ritfunktioner märker ofta ut avvikande observationer automatiskt bland annat i låddiagram (box plot). Bilder som specificerar avvikande observationer lämpar sig inte för publicering, såvida forskaren inte kan motivera att de avvikande observationer som märkts ut i bilden inte kan identifieras.

Spridningsbilder används vanligen för presentation av värden hos två kontinuerliga variabler. Punkterna i spridningsbilder beskriver i princip uppgifter om varje enskild observation, varför de i jämförelse med föregående bildtyper är ett besvärligare fall med tanke på dataskydd. I bedömningen av dataskyddet för spridningsbilder ska forskaren rikta särskild uppmärksamhet mot materialets natur, med tanke på bland annat urvalets storlek, uppgifternas känsliga natur och förekomsten av avvikande observationer. Spridningsbilden uppfyller inte dataskyddskraven, om det direkt kan ses eller enkelt dras en slutsats till exempel om det största företaget inom en enskild bransch utifrån den.

Skyddsmetoder

Filer eller tabeller som extraheras ur distansanvändningssystemet får inte innehålla möjligheten att identifiera en enskild person eller ett enskilt företag. Uppgifter som omfattas av risk för röjande ska skyddas genom att planera utskrifternas innehåll på ett sätt som är godtagbart med tanke på dataskyddet, till exempel genom att använda tillräckligt grova klassificeringar.

De celler som omfattar en risk för röjande i tabellen kan skyddas genom att ändra tabellstrukturen, täcka enskilda cellvärden eller hela rader eller genom att ändra cellvärden till exempel genom att avrunda eller ersätta det ursprungliga cellvärdet med ett ungefärligt slumpmässigt tal. I valet av skyddsmetod för tabellen finns det skäl att sträva efter att hitta en metod som skyddar tabellen tillräckligt, men bevarar dess egenskaper som är viktiga för användningsändamålet på ett så bra sätt som möjligt.

Ändring av strukturen på tabellen innebär kontroll av antalet variabler eller ändring av klassificeringen. Genom att ändra klassificeringen strävar man efter att ta bort celler förknippade med en risk för röjande genom att kombinera klasser som innehåller dem med andra klasser i tabellen. En ändring av klassificeringen innebär ofta i praktiken att hela klassificeringen förgrovas.

Täckning omfattar primär täckning av celler med risk för röjande och sekundär täckning. Genom sekundär täckning säkerställs att de i första hand täckta cellernas

värden inte kan avslöjas med hjälp av rad- eller kolumntotalerna i tabellen. Täckningen kan också göras radvis. Om någon radtotal i tabellen omfattar enbart en liten mängd statistikenheter (lägre än det använda tröskelvärdet), täcks raden i fråga i sin helhet utan att beakta antalet statistikenheter i dess olika celler.

Närmare information om dataskyddsmetoderna finns till exempel i materialet i webbkursen Forskningsmaterial i distansanvändning (endast på finska), vilken nämns i slutet.

Kontrollförfarandet för utskrifter

Forskartjänsterna använder ett kontrollförfarande för forskningsresultat när det gäller material som är i distansanvändning. Om en forskare är osäker på dataskyddet av en utskrift, lönar det sig att vara i kontakt med forskartjänsterna redan innan överföringen av en utskrift för granskning eller extrahering från systemet.

Genom kontrollförfarandet stöder forskningstjänsten forskaren till ansvarsfull verksamhet i dataskyddsärenden som gäller forskningsresultat. Trots kontrollförfarandet har forskaren ansvar för dataskyddet för sina forskningsutskrifter. Kontrollförfarandet ger forskartjänsterna möjlighet att följa hur dataskyddet genomförs i resultaten i forskningsmaterial och upptäcka behov av att erbjuda ytterligare handledning i dataskyddsärenden.

Kontrollförfarandet fungerar i praktiken på olika sätt i distansanvändning inom forskningsprojekt och i distansanvändning av mikrosimuleringsmodellen. Vad gäller bägge fall ska forskaren dock se till att utskrifter som överförs från distansanvändningssystemet (eller som önskas överföras) inte innehåller material på enhetsnivå eller möjligheten att en uppgift innehållande en enskild observation röjs.

Överföring från distansanvändningssystemet för granskning eller extrahering ska göras med omdöme. Från systemet ska enbart utskrifter som ska publiceras överföras för granskning eller extrahering. Med andra ord ska överföring av så kallade mellanresultat och i synnerhet (stora) filer av log-typ undvikas. Tabeller och figurer som överförs för granskning borde vad gäller innehåll vara i den form som de avses publiceras. Ur systemet är det inte möjligt att få ut utskrifter, som inte kan publiceras på grund av dataskyddet.

Utskrifter ska dokumenteras omsorgsfullt, så att utskriftens datainnehåll står klart för granskaren. Antalet observationer som använts i beräkningen av tabeller, bilder, nyckeltal osv. ska vara synliga i utskrifterna. Om man önskar extrahera uppgifter som avviker från dataskyddsanvisningarna från resultatkontrollen, ska genomförandet av dataskyddet i utskrifterna motiveras väl.

Filformatet för utskrifter, i synnerhet för bilder, ska också vara sådan att det inte orsakar risk för att uppgifter på enhetsnivå röjs. De bildformat som lämpar sig för granskning utgörs av till exempel:

- Bitkarteformat
- PNG (Portable Networks Graphics)
- BMP (Bitmap)

- JPEG (Joint Photographic Experts Group)
- TIFF (Tagged Image File Format)
- Vektorformat
- EPS (Encapsulated PostScript)
- PS (PostScript)
- PDF (Portable Document Format)
- SVG (Scalable Vector Graphics)
- WMF/EMF (Windows Metafile)

I programmet Stata kan ovan nämnda bildformat skapas med kommandot graph export. I programmet SPSS kan bildformatet väljas i funktionen Export output. I R-programmet får man information om ritfunktionen med kommandot help(grDevices). Vissa bildtyper, såsom gph-filer i Stata, sparar i regel det material som använts när figuren ritats och därför lämpar de sig ofta inte för överföring till granskning och extrahering.

Webbkursen Forskningsmaterial i distansanvändning (på finska)

Webbkursen Forskningsmaterial i distansanvändning har publicerats som en del av Statistikskolan på Statistikcentralens webbplats (https://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=ketusivu&course_id=tkoulu_tutki).

Webbkursen ger närmare information om distansanvändning av forskningsmaterial och SISU-mikrosimuleringsmodellen och dataskydd av material. Kursen innehåller också exempel på skydd av tabellmaterial.

I synnerhet för forskare som för första gången använder distansanvändningssystemet rekommenderas det att de, vid sidan om dessa anvisningar, också tar del av materialen i webbkursen Forskningsmaterial i