TK/1172/00.00.00/2021-14 9.6.2021

Dataskydd för datamaterial som överlåtits

Denna anvisning redogör för dataskyddet för material som överlåts fysiskt för forskningsanvändning och för utskrifter som producerats utifrån det (tabeller, figurer, statistiska modeller o.d.). Fysiskt överlåtna datamaterial används i forskarnas utrymmen och inte via Fiona -distansanvändningsmiljön.

Dessa riktlinjer är utformade för att stödja ansvarsfullt agerande i dataskyddsfrågor och syftar till att förhindra både oavsiktliga och avsiktliga dataskyddsföreteelser.

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

Statistikcentralens och forskarnas ansvar över dataskydd av 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. Forskaren ska för egen del sköta dataskyddet av sitt material under tiden för forskningsanvändningen, då datat lagras under den tid tillståndet är i kraft samt vid publicering av forskningsutskrifter. När tillståndet löper ut måste forskaren ta hand om att förstöra det levererade materialet och de kopior och filer som bildats av det. Förstörandet av materialet måste vara slutfört då tillståndsperioden tar slut.

Forskaren är ansvarig för att dataskyddet i de forskningsresultat som hen publicerar.

Varför dataskydd?

En förutsättning för att Statistikcentralen ska kunna lämna ut forskningsmaterial som samlats in för statistiska ändamål för forskningsanvändning 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, 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”.

Utskrifter som producerats frånurvalsmaterial

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.

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 personer till exempel inom en viss åldersgrupp snarare ett totalmaterial än ett urvalsmaterial, eftersom det i allmänhet är möjligt att dra en slutsats om en person hör till det ”urval” som avgränsats till att omfatta alla personer inom en viss åldersgrupp.

Frekvens-och mängdtabeller

Den utskrifter i tabelleformat som innehåller personuppgifter bör aggregeras och även indirekt identifiering bör förhindras. I regel är tröskelvärdet 3 i skyddet av tabeller som innehåller 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 uppgifterom 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.

Dataskyddet av personuppgifter ska skötas särskilt noggrant, om tabellen innehåller känsliga personuppgifter[1] (inkl. de särskilda kategorier av personuppgifter som räknats upp i EU:s allmänna dataskyddsförordning[2]). Vid skydd av känsliga uppgifter kan det vara nödvändigt att använda ett högre tröskelvärde.

[1] 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.
[2] Europaparlamentets och rådets förordning (EU) 2016/679, artikel 9

Olika fördelningsnyckeltal

Minimi och maximi är i allmänhet förknippade med en observation, så de kan oftast inte 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 tröskelvärdet3 överskridas vad gäller alla grupper som bildar andelar. Med andra ord, om man vill publicera till exempel att kvinnornas andel är 66 % av hela populationen, ska dessa 66 %, liksom också männens 33 % 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 förobservationer 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 bild typer 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 utifrån den om till exempel patienter på en specifik plats med en sällsynt sjukdom.

Skyddsmetoder

Filer eller tabeller som skall publiceras 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 exempelgenom 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 skyddartabellen tillräckligt, men bevarardess 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 det första handtäckta 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).