TK/1172/00.00.00/2021-13 9.6.2021

Luovutettujen aineistojen tietosuoja

Tässä ohjeessa kerrotaan tutkimuskäyttöön fyysisesti luovutetun aineiston ja siitä tuotettujen tulosteiden, esimerkiksi taulukoiden, kuvioiden ja tilastollisten mallien, tietosuojasta. Fyysisesti luovutettua aineistoa käytetään tutkijan tiloissa eikä FIONA-etäkäyttöympäristössä.

Nämä ohjeet on laadittu tukemaan vastuullista toimintaa tietosuoja-asioissa, ja ohjeilla pyritään estämään niin tahattomat kuin tahallisetkin tietosuojarikkomukset.

Jos tietosuoja-asiat herättävät kysymyksiä, ota yhteyttä Tutkijapalveluihin (tutkijapalvelut@stat.fi).

Tilastokeskuksen ja tutkijan vastuu aineistojen ja tulosteiden tietosuojasta

Tutkimusaineistojen tietosuojasta huolehtiminen on sekä Tilastokeskuksen että tutkijoiden tehtävä. Tilastokeskus huolehtii aineistojen tietosuojasta ennen aineistojen luovutusta tutkimuskäyttöön.

Tutkijan on huolehdittava tietosuojasta aineistonsa tutkimuskäytön ja käyttöluvan mukaisen säilyttämisen ajan sekä julkaisuvaiheessa. Kun käyttöluvan voimassaoloaika on päättymässä, tutkijan on huolehdittava luovutetun aineiston ja siitä muodostettujen kopioiden ja välitiedostojen hävittämisestä. Hävittäminen tulee olla tehty käyttöluvan voimassaoloajan päättymiseen mennessä.

Tutkija on vastuussa tietosuojan toteutumisesta julkaisemissaan tutkimustuloksissa.

Miksi tietosuoja?

Tutkimusaineistojen tietosuojasta huolehtiminen on edellytys sille, että Tilastokeskus voi luovuttaa tilastotarkoituksiin keräämiään aineistoja tutkimuskäyttöön. Tilastokeskuksen oikeus kerätä rekisteri- ja kyselytutkimusaineistoja tilastointia varten on taattu tilastolailla, kuten myös oikeus luovuttaa näitä tietoja tutkimuskäyttöön. Toisaalta laki velvoittaa myös huolehtimaan tietojen asiallisesta suojaamisesta. Osa aineistoista sisältää arkaluonteisia tietoja.

Henkilötietojen käsittelyä ohjaa Euroopan unionin yleinen tietosuoja-asetus (EU) 2016/679 ja sitä täydentävä kansallinen tietosuojalaki (1050/2018).

Tutkimushankkeen saamaan käyttölupaan liittyy salassapitovelvoite. Sen mukaan tutkijan on huolehdittava siitä, että tutkimustuloksissa ei ole yksikkötason tietoja, eli yksittäistä henkilöä tai yritystä koskevia tietoja tai mahdollisuutta niiden paljastumiseen. Käsitellessään tutkimuskäyttöön luovutettua aineistoa tutkijan on salassapitovelvoitteen mukaan huolehdittava myös siitä, ettei aineistoa paljasteta tai luovuteta taholle, jolla ei ole siihen käyttöoikeutta.

Eri tulostetyyppien tietosuojaohjeet

Alla on esitetty ohjeita ja sääntöjä erilaisten tulostetyyppien tietosuojasta huolehtimiseen. Tulostetyyppejä ovat esimerkiksi taulukot, kuvat ja tilastolliset mallit. Koska tutkimushankkeita ja erilaisia tulostetyyppejä on paljon erilaisia, ei jokaisen aineiston ja tulosteen tietosuojaa voida arvioida ja ohjeistaa erikseen. Tämän takia yleisten ”nyrkkisääntöjen” antaminen on välttämätöntä.

Otosaineistoista tuotetut tulosteet

Kaikille tulostetyypeille on yhteistä, että tulosten taustalla olevien havaintojen paljastumisriski on yleensä pienempi, jos käytetty aineisto on vain (satunnais-) otos kokonaisaineistosta. Otannasta mahdollisesti johtuva pienempi paljastumisriski ei kuitenkaan tarkoita, että otosaineistoista tuotettuja tulosteita varten voitaisiin automaattisesti lieventää tässä ohjeessa esitettyjä suojaussääntöjä.

On huomioitava, että kokonaisaineisto ei aina tarkoita vain kaikkia Suomessa asuvia henkilöitä tai kaikkia Suomessa toimivia yrityksiä. Tietosuojan näkökulmasta ja havaintojen paljastumisriskiä arvioitaessa esim. tiettyyn ikäluokkaan kuuluvat muodostavat ennemminkin kokonaisaineston kuin otosaineiston, sillä yleensä on pääteltävissä, kuuluuko yksittäinen henkilö ”otokseen”, joka on rajattu kattamaan kaikki ikäluokkaan kuuluvat.

Frekvenssi- ja määrätaulukot

Henkilötietoja sisältävät taulukkomuotoiset tiedot tulee aggregoida ja myös välillinen tunnistaminen estää. Henkilötietoja sisältävien taulukoiden suojauksessa pääsääntönä on kynnysarvo 3. Toisin sanoen taulukon solun tai havaintoryhmän tiedot saa julkaista vain, jos tietojen taustalla on vähintään 3 (painottamatonta) havaintoa. Havaintojen lukumäärätietoakaan ei saa julkaista, jos solun tai ryhmän havaintojen lukumäärätieto on pieni.

Kynnysarvosäännön käyttö pääsääntönä on helpoin tapa pienentää julkaistaviin taulukoihin mahdollisesti kohdistuvaa yksittäisten havaintojen paljastumisriskiä. Havainnon tunnistaminen ryhmänsä ainoaksi tai harvinaiseksi tapaukseksi voi lisätä riskiä tunnistaa havainto ja/tai paljastaa lisää tietoja tästä havainnosta (esim. muiden samaa aineistoa tai aihetta käsittelevien taulukoiden ja tulosteiden avulla). Tämän takia kynnysarvoa tulee soveltaa, vaikka taulukossa julkaistaisiinkin pelkkiä lukumäärätietoja.

Henkilötietojen osalta tietosuojasta on huolehdittava erityisen tarkasti, jos taulukko sisältää arkaluonteisia henkilötietoja[1] (mukaan lukien EU:n yleisessä tietosuoja-asetuksessa[2] luetellut erityiset henkilötietoryhmät). Arkaluonteisten tietojen suojaamisessa saattaa olla tarpeellista käyttää suurempaa kynnysarvoa.

[1] Arkaluonteisia tietoja tässä ovat tiedot, jotka kuvaavat henkilön rotua tai etnistä alkuperää, poliittisia mielipiteitä, uskonnollista tai filosofista vakaumusta, ammattiliiton jäsenyyttä, terveyttä, seksuaalista käyttäytymistä tai suuntautumista, rikostuomioita ja rikkomuksia ja niihin liittyviä turvaamistoimia, kuolinsyytä, kieltä, kansalaisuutta, syntyperää tai synnyinmaata, tuloja, velkoja, varallisuutta, harvinaista ammattia tai muuta sosioekonomista asemaa, sosiaalihuollon tarvetta tai saatuja sosiaalihuollon palveluita, sosiaalihuollon tukitoimia tai muita etuuksia, tai tiedot henkilöön kohdistetuista hoitotoimenpiteistä tai niihin verrattavista toimista.

[2] Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, 9 artikla

Erilaiset jakaumatunnusluvut

Minimi ja maksimi liittyvät yleensä yhteen havaintoon, joten niitä ei useimmiten voi julkaista. Jos minimiä tai maksimia ei voi yhdistää yksittäiseen havaintoon ja kynnysarvo toteutuu, ne voi julkaista.

Jakaumapisteet (pl. minimi ja maksimi), esimerkiksi desiilit, muodostavat erikoistapauksen taulukosta, jossa solufrekvenssejä vastaavat jakaumapisteiden väliin jäävien havaintojen lukumäärät. Jos nämä lukumäärät ylittävät taulukoissa sovellettavan kynnysarvon 3, jakaumapisteet voi julkaista.

Moodin voi julkaista, jos kaikki tai lähes kaikki havainnot eivät saa samaa arvoa.

Keskiarvon, muut suhdeluvut ja jakaumatunnuslukujen korkeammat momentit (esim. varianssi) voi julkaista, mikäli niiden laskennassa on käytetty vähintään kolmea havaintoa.

Osuuksia julkaistaessa kynnysarvon 3 on toteuduttava kaikkien osuuksia muodostavien ryhmien osalta. Jos esimerkiksi halutaan julkaista tieto, että naisten osuus on 58 % koko populaatiosta, tuohon osuuteen täytyy sisältyä vähintään kolme henkilöä. Samoin miesten 42 % osuuden on sisällettävä vähintään kolme henkilöä. Ei siis riitä, että naisia ja miehiä on yhteensä koko populaatiossa vähintään 3.

Muut numeeriset tulostetyypit

Indeksipisteluvut, korrelaatiokertoimet ja testisuureet (t, F, X^2, yms.) voi yleensä julkaista, mikäli niiden laskennassa on käytetty vähintään 10 havaintoa.

Regressiomallin voi kokonaisuudessaan julkaista, mikäli mallin taustalla on riittävästi havaintoja (vähintään 10) ja malli ei kuvaa aikasarjaa yhteen yritykseen tai henkilöön perustuvista havainnoista. Mallin yksittäisiä kertoimia voi yleensä aina julkaista.

Monimutkaisempien tilastollisten mallien numeeriset tulokset voi yleensä aina julkaista, mikäli mallin taustalla on riittävästi havaintoja ja malli ei kuvaa aikasarjaa yksittäiseen yritykseen tai henkilöön perustuvista havainnoista.

Kuvat

Kuvatkaan eivät saa paljastaa yksittäisen havainnon tietoja. Aineistoista piirretyt kuvat ovat sallittuja, jos yksittäinen kuvapiste tai kuvan osa ei voi paljastaa sen taustalla olevaa yksittäistä havaintoa.

Pylväsdiagrammeja ja muita luokitellun aineiston esittämiseen käytettyjä kuvia voi yleensä julkaista, kunhan kussakin luokassa on riittävästi havaintoja. Tällaisen kuvan informaation voi yleensä esittää myös taulukkomuodossa, ja siihen voi soveltaa samoja tietosuojasääntöjä kuin muihinkin taulukkoaineistoihin (ks. yllä kohta Frekvenssi- ja määrätaulukot).

Jakaumakuvista tasoitetut tai riittävän karkealla asteikolla esitetyt jakaumat, histogrammit ja kertymäfunktiot ovat sallittuja. Osa jakaumakuvista voi sisältää tietoja poikkeavista havainnoista tai ääriarvoista, ja ne saattavat joskus paljastaa yksittäisen havainnon tietoja. Ohjelmien piirtofunktiot merkitsevät usein

automaattisesti mm. laatikkokuvaajiin (box plot) poikkeavat havainnot. Kuvia, jotka yksilöivät poikkeavat havainnot, ei saa julkaista, ellei tutkija pysty hyvin perustelemaan, etteivät kuvaan merkityt poikkeavat havainnot ole tunnistettavissa.

Hajontakuvia käytetään tyypillisesti kahden jatkuvan muuttujan arvojen esittämiseen. Hajontakuvien pisteet kuvaavat lähtökohtaisesti kukin yksittäisen havainnon tietoja, minkä vuoksi hajontakuva on tietosuojan kannalta edellisiä kuvatyyppejä hankalampi tapaus. Hajontakuvien tietosuojan arvioinnissa tutkijan tulee kiinnittää erityistä huomiota aineiston luonteeseen, kuten otoksen kokoon, tiedon arkaluonteisuuteen ja poikkeavien havaintojen esiintymiseen. Hajontakuva ei täytä tietosuojavaatimuksia, jos siitä on suoraan nähtävissä tai helposti pääteltävissä esimerkiksi harvinaisen sairauden potilaita yksittäisellä paikkakunnalla.

Suojausmenetelmät

Julkaistavista tiedoista tai taulukoista ei saa olla mahdollista tunnistaa yksittäistä henkilöä tai yritystä koskevia tietoja. Paljastumisriskin sisältävät tiedot on suojattava suunnittelemalla tulosteiden sisältö tietosuojan kannalta hyväksyttäväksi esimerkiksi tarpeeksi karkeita luokituksia käyttämällä.

Taulukossa paljastumisriskin sisältävät solut voidaan suojata
- muuttamalla taulukon rakennetta
- peittämällä yksittäisiä soluarvoja tai kokonaisia rivejä
- muuttamalla soluarvoja esimerkiksi pyöristämällä tai
- korvaamalla alkuperäinen soluarvo likimääräisellä satunnaisluvulla.

Taulukon suojausmenetelmän valinnassa on syytä pyrkiä löytämään menetelmä, joka suojaa taulukkoa riittävästi, mutta säilyttää sen käyttötarkoituksen kannalta tärkeät ominaisuudet mahdollisimman hyvin.

Taulukon rakenteen muuttaminen tarkoittaa muuttujien määrän kontrollointia tai luokituksen muuttamista. Luokitusta muuttamalla taulukosta pyritään hävittämään paljastumisriskissä olevat solut yhdistämällä niitä sisältävät luokat muihin taulukon luokkiin. Luokituksen muuttaminen tarkoittaa usein käytännössä koko luokituksen karkeistamista.

Peittämiseen kuuluu ensisijainen, paljastumisriskissä olevien solujen peittäminen ja toissijainen peittäminen. Toissijaisella peittämisellä varmistetaan, ettei taulukon rivi- tai saraketotaalien avulla pystytä paljastamaan ensisijaisesti peitettyjen solujen arvoja. Peittäminen voidaan tehdä myös rivikohtaisesti. Jos taulukon johonkin rivitotaaliin kuuluu vain pieni määrä tilastoyksiköitä (vähemmän kuin käytetty kynnysarvo), peitetään kyseinen rivi kokonaisuudessaan huomioimatta sen eri soluissa olevien tilastoyksiköiden lukumäärää.

Lisätietoja tietosuojamenetelmistä löytyy esimerkiksi Tilastokeskuksen Tutkimusaineistot etäkäytössä -verkko-oppaan materiaaleista.