# Digitala verktyg för riskbedömning och skyddsåtgärder i tabulära data med personuppgifter

Workshopprogram

Datum: 2024-01-29  
Version: 1



Innehåll

[Digitala verktyg för riskbedömning och skyddsåtgärder i tabulära data med personuppgifter 1](#_Toc159242774)

[Översikt 4](#_Toc159242775)

[Efter genomförd workshop ska deltagarna: 4](#_Toc159242776)

[Praktiskt moment i sdcMicro (sdcApp) 4](#_Toc159242777)

[Vad är sdcMicro? 4](#_Toc159242778)

[Vad är sdcApp? 4](#_Toc159242779)

[Vad används det till? 4](#_Toc159242780)

[Kan jag skaffa det själv? 5](#_Toc159242781)

[Externt material som är bra att gå igenom innan workshopen: 5](#_Toc159242782)

[Guider, manualer och webbinarier: 5](#_Toc159242783)

[Grundläggande koncept som kommer att tas upp: 5](#_Toc159242784)

[Agenda/hålltider 6](#_Toc159242785)

[1. Huvuduppgift labb 7](#_Toc159242786)

[Vid uppgiftens slut förväntas ni: 7](#_Toc159242787)

[Ni förväntas också kunna svara på följande frågor: 7](#_Toc159242788)

[2. Studie, data och kontext 8](#_Toc159242789)

[Studie 8](#_Toc159242790)

[Data 8](#_Toc159242791)

[Kontext 8](#_Toc159242792)

[3. Instruktion steg-för steg 9](#_Toc159242793)

[Microdata 9](#_Toc159242794)

[Anonymize 9](#_Toc159242795)

[Risk/Utilty 9](#_Toc159242796)

[Export data 10](#_Toc159242797)

[Bilaga 1 - Enkätfrågor 11](#_Toc159242798)

[KÖN/ÅLDER 11](#_Toc159242799)

[UTBILDNING 11](#_Toc159242800)

[INKOMST 11](#_Toc159242801)

[KOMMUN 12](#_Toc159242802)

[PARTIVAL RIKSDAGSVALET 2022 12](#_Toc159242803)

[FÖRTROENDE SVENSKA MYNDIGHETER 12](#_Toc159242804)

[POLITISKA FÖRSLAG 12](#_Toc159242805)

[Bilaga 2 - Kodbok 13](#_Toc159242806)

[KÖNSVARIABEL (i dataset) 13](#_Toc159242807)

[ÅLDERSVARIABEL (i dataset) 13](#_Toc159242808)

[UTBILDNINGSVARIABEL (i dataset) 13](#_Toc159242809)

[KOMMUN (i dataset) 14](#_Toc159242810)

[INKOMST (i dataset) 14](#_Toc159242811)

# Översikt

### Efter genomförd workshop ska deltagarna:

* Ha en grundläggande överblick av statistiska koncept och metoder för att pseudonymisera tabulära data.
* Kunna identifiera vilka variabler som bör inkluderas i en riskbedömning och röjandekontroll.
* Vara bekanta med grundläggande tillvägagångssätt och funktioner för pseudonymisering och omkodning i sdcApp.
* Kunna tolka och bedöma resultat av pseudonymisering och omkodning av data samt dess inverkan på datas användbarhet.
* Ha en förståelse för hur statistiska metoder för pseudonymisering och omkodning samspelar och kan kombineras med andra typer av skyddsåtgärder.

### Praktiskt moment i sdcMicro (sdcApp)

Under workshopen används sdcApp för att bearbeta ett syntetiskt tabulärt dataset med ett antal direkt- och indirekt identifierande variabler.

Se den fristående guiden med instruktioner för hur du kommer igång med sdcMicro och sdcApp.

### Vad är sdcMicro?

**sdcMicro** är ett tilläggspaket till programmeringsmiljön R. sdc står för statistical disclosure control (statistisk röjandekontroll). Micro refererar till termen mikrodata vilket avser data om enskilda objekt, till exempel personer. Alla data som laddas in av programmet behandlas endast på datorn som programmet körs på.

### Vad är sdcApp?

**sdcApp** är en del av R-paketet sdcMicro som erbjuder ett interaktivt användargränssnitt där du inte behöver kunna programmera för att använda paketets funktioner. Du kommer åt gränssnittet genom din webbläsare, men webbläsaren arbetar i detta fall endast med lokalt lagrade filer. I användargränssnittet utför du kommandon genom att klicka dig runt i menyerna, i stället för att skriva R-kod.

### Vad används det till?

sdcMicro och sdcApp används i huvudsak för [statistisk röjandekontroll](https://www.scb.se/contentassets/0cd92207266d40eb8829244d51d90b94/handbok-i-statistisk-rojandekontroll.pdf).  Programmet används exempelvis för att identifiera olika variabler, eller kombinationer av variabler som innebär en risk för bakvägsidentifiering. Det hjälper användaren att relativt snabbt få en överblick över en datamängd och huruvida enskilda objekt eller individer löper risk att bli identifierade.

Med hjälp av sdcMicro kan du koda om variabler relativt enkelt och statistiskt utvärdera vilken effekt omkodningar får för risken för bakvägsidentifiering. Du kan också utvärdera hur dina omkodningar påverkar datamängdens kvalitet.

Verktyget är relativt väletablerat och används av både enskilda forskare och internationella organisationer för statistisk röjandekontroll.

### Kan jag skaffa det själv?

sdcMicro med det tillhörande användargränssnittet sdcApp är gratis och [nedladdningsbart](https://sdcappdocs.readthedocs.io/en/latest/installation.html#installing-sdcmicro-package) genom [R eller R-studio](https://sdcappdocs.readthedocs.io/en/latest/installation.html#installing-r-and-rstudio).

Se den fristående guiden med instruktioner för hur du kommer igång med sdcMicro och sdcApp.

Läs mer: <https://sdctools.github.io/sdcMicro/articles/sdcMicro.html>

## Externt material som är bra att gå igenom innan workshopen:

### Guider, manualer och webbinarier:

* [Manual till sdcApp](https://sdcappdocs.readthedocs.io/en/latest/introduction.html) (rekommenderas starkt att du läser introduktionen till manualen)
* [CESSDA webbinarium](https://youtu.be/JeJ6OOxXZwo?si=6WAis0whqYP7-3-R&t=3854) med genomgång av pseudonymisering/anonymisering och sdcApp (rekommenderas starkt)
* [Five safes framework](https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/) – grundläggande ramverk för säkra data
* [Data privacy handbook (Utrecht university)](https://utrechtuniversity.github.io/dataprivacyhandbook/pseudonymisation-anonymisation.html) – Specifikt kring pseudonymisering och anonymisering

### Grundläggande koncept som kommer att tas upp:

* [K-anonymitet](https://youtu.be/Q0DNOIGUzMc) (Youtube)
* [L-diversitet](https://youtu.be/GNhb3PcmjmA) (Youtube)

# 

# Agenda/hålltider

**20 minuter** – Introduktion till grundläggande begrepp.

**20 minuter** – Workshopledaren visa funktioner i sdcApp

**10 minuter** – Gruppindelning, start av sdcApp

**10 minuter** – Paus

**60 minuter** – Deltagarna läser instruktioner och pseudonymiserar ett syntetiskt dataset med hjälp av sdcApp

**30 minuter** – Avslutande genomgång, frågor och diskussion

# 1. Huvuduppgift labb

*Huvuduppgiften är att skapa ett dataset med så liten risk för bakvägsidentifiering som möjligt, samtidigt som datasetets användbarhet bibehålls i så stor utsträckning som möjligt.*

**Steg 1:** Läs Studie, data och kontext.

**Steg 2:** Följ Instruktion steg-för-steg.

**Steg 3:** Redovisa resultat.

## Vid uppgiftens slut förväntas ni:

Kortfattat kunna redovisa resultatet av genomförda skyddsåtgärder i sdcApp och notera de olika överväganden ni gjort under processens gång.

## Ni förväntas också kunna svara på följande frågor:

1. Hur väl lyckades ni med era skyddsåtgärder utifrån den grad av k-anonymitet ni satt som mål?
2. Hur användbara är de modifierade data ni skapat jämfört med originaldata? Går det fortfarande att återanvända data för den sambandsanalys som specificeras i Studie, data och kontext?
3. Finns det andra lämpliga skyddsåtgärder som skulle kunna komplettera eller ersätta de åtgärder som ni genomfört?

Svara på ovan frågor inom gruppen. Diskutera sedan kort vad ni kommit fram till tillsammans med en ytterligare grupp som ni blir hänvisade till av workshopledarna.

# 2. Studie, data och kontext

## Studie

* Studien genomfördes av en forskargrupp vid statsvetenskapliga institutionen vid Göteborgs universitet.
* Syftet med studien är att analysera samband mellan olika bakgrundsvariabler och partisympati, attityder till politiska förslag samt förtroende för svenska myndigheter.

## Data

* Data bygger på en enkätundersökning som skickades ut till ett slumpmässigt urval i åldern 18–85 år i tre av Sveriges folkrikaste län: Stockholms län, Västra Götalands län och Skåne län.
* Totalt skickades enkäten ut till 2000 personer. Svarsfrekvensen var 50%.
* Data innehåller ett antal variabler varav några kan anses utgöra direkta och indirekta identifierare.
* Data innehåller känsliga personuppgifter då respondenterna bland annat har uppgett information om politiska åsikter.

## Kontext

* Fältperiodenpågick mellan oktober och december 2022.
* Urvalet kontaktas via post med en inbjudan till en webbenkät (push-to-web)
* Beställning av urval och adressfil görs av ett privat enkätföretag.
* Samma företag sköter också själva datainsamlingen och leverans av en slutlig datafil till forskargruppen.
* Företaget lovar att radera direkta identifierare och potentiella kodnycklar tre månader efter att fältperioden avslutas. Företaget verkar dock ha missat att ta bort vissa variabler.
* Forskargruppen vill minska risken för bakvägsidentifiering samt att andra ska kunna analysera samband mellan bakgrundsvariabler, partisympati och attityder till politiska förslag samt förtroende för svenska myndigheter

# 3. Instruktion steg-för steg

Obs! Se till att du har läst **2. Studie, data och kontext** innan du börjar följa denna steg-för-steg-guide.

## Microdata

1. Välj fliken **Microdata** och ladda in datasetet *Syntetiskt dataset for sdcApp.csv* med .csv-importfunktionen genom att bläddra fram det på din egna dator i dialogrutan som kommer fram.
2. Gå igenom datasetet under **Display microdata** under huvudmenyn **Microdata** och se till att data har laddats in som det ska.
3. Gå till **Explore variables** i vänstermenyn under huvudfliken **Microdata** och skapa dig en överblick av de variabler som ingår i datasetet. Titta exempelvis på hur fördelningen mellan olika värden inom variabler ser ut. Fundera om det finns något som sticker ut som du bör ta hänsyn till när du senare ska pseudonymisera dina data.
4. Identifiera vilka variabler som kan klassas som direkta respektive indirekta identifierare.

## Anonymize

1. Gå vidare till fliken **Anonymize** i huvudmenyn. Bedöm vilka variabler som eventuellt behöver raderas.
2. Välj sedan de nyckelvariabler som primärt bör ingå i röjandekontrollen. I denna övning kommer vi bara behandla nyckelvariabler som kategoriska/faktorvariabler (“**Cat.**” i sdcApp). När du valt vilka variabler som ska ingå, tryck på knappen “**Setup SDC problem**”.
3. Du hamnar nu på sidan **Show summary** under huvudfliken **Anonymize.** Gå igenom informationen som redovisas på sidan om de variabler du valt ut.
4. Fundera på hur ambitiösa skyddsåtgärder ni bör sikta på (dvs, vilken grad av k-anonymitet som är lämplig för dessa data?). Välj en lämplig nivå som du vill sikta på.
5. Gå till sidan **Recoding** i vänstermenyn under huvudfliken **Anonymize**.
6. Genomför lämpliga omkodningar av dina nyckelvariabler (se Förslag till omkodningar för vägledning).
7. Notera kontinuerligt vilken effekt dina omkodningar får för graden av pseudonymisering enligt k-anonymity, dvs jämför **Modified data** med **Original data**, i tabellen längst ner i högra hörnet på sidan **Recoding**.

## Risk/Utilty

1. Byt huvudflik till **Risk/Utility**. Gå till sidan **Obs. violating k-anon**. Ställ in den grad av k-anonymitet du vill uppnå i slidern längst upp. Gå igenom de observationer som inte uppnår den grad av k-anonymitet som du siktar på. Fundera på om du behöver göra ytterligare omkodningar eller radera specifika värden.
2. Om du bestämmer dig för att radera specifika värden. Gå tillbaka till **Anonymize** och klicka på **k-anonymity** i vänstermenyn. Ställ in den grad av k-anonymitet du vill ha. Klicka på **Establish k-anonymity**. sdcApp raderar då automatiskt värden i nyckelvariabler.
3. Gå igenom sidorna **Barplot/Mosaicplot**, **Tabulations** och **Information loss** i vänstermenyn. Jämför hur dina modifierade data ser ut jämfört med originaldata. Fundera på huruvida dina modifierade data begränsar dess användbarhet givet de samband som forskarna vill kunna analysera (se Studie, data och kontext)?

## Export data

1. Gå till huvudfliken **Export data**. Titta på dina modifierade data. Om du valt att radera specifika värden (se punkt 12), notera vilka värden som har raderats och fundera på vad det kan få för konsekvenser. Fundera på om de skyddsåtgärder du genomfört är rimliga i förhållande till den kontext data hanteras (se Studie, data och kontext).
2. Gå in på sidan **Anonymization report**. Spara rapporten på datorn. Gå igenom rapporten och notera vilka förändringar du gjort i data.

# Bilaga 1 - Enkätfrågor

### KÖN/ÅLDER

En bild som visar text, Teckensnitt, linje, diagram

Automatiskt genererad beskrivningEn bild som visar text, Teckensnitt, skärmbild, diagram

Automatiskt genererad beskrivning

*Kompletteras med data från registerdata.* Kompletteras med data från registerdata. Omkodas.

### UTBILDNING

En bild som visar text, skärmbild, Teckensnitt, nummer

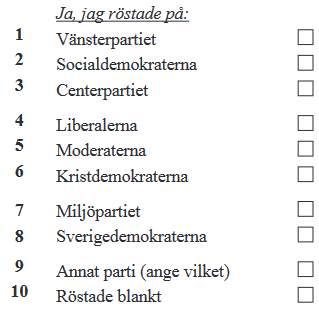
Automatiskt genererad beskrivning

### INKOMSTEn bild som visar text, skärmbild, Teckensnitt, nummer Automatiskt genererad beskrivning

### KOMMUN

*Kompletteras med data från registerdata.*

### PARTIVAL RIKSDAGSVALET 2022



### FÖRTROENDE SVENSKA MYNDIGHETER

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

### POLITISKA FÖRSLAG

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

# Bilaga 2 - Kodbok

### KÖNSVARIABEL (i dataset)

1. Kvinna  
2. Man

### ÅLDERSVARIABEL (i dataset)

Ålder är okodad i data och förekommer som uppgiven ålder.

#### Förslag till omkodning

**ÅLDER 4 -delad**

1. 18–29 år  
2. 30–49 år  
3. 50–64 år  
4. 65–85 år

**ÅLDER 4-delad**

1. 16–29 år  
2. 30–49 år  
3. 50–75 år  
4. 76–90 år

**ÅLDER 8-delad (a)**

1. 16–19 år  
2. 20–29 år  
3. 30–39 år  
4. 40–49 år  
5. 50–59 år  
6. 60–69 år  
7. 70–79 år  
8. 80–90 år

**ÅLDER 8-delab (b)**

1. 16–19 år  
2. 20–24 år  
3. 25–29 år  
4. 30–39 år  
5. 40–49 år  
6. 50–59 år  
7. 60–75 år  
8. 76–90 år

**ÅLDER 9 -delad**

1. 16–19 år  
2. 20–24 år  
3. 25–29 år  
4. 30–39 år  
5. 40–49 år  
6. 50–59 år  
7. 60–65 år  
8. 66–75 år  
9. 76–90 år

**ÅLDER 9-delad (b)**

1. 16–19 år  
2. 20–24 år  
3. 25–29 år  
4. 30–39 år  
5. 40–49 år  
6. 50–59 år  
7. 60–69 år  
8. 70–79 år  
9. 80–90 år

### UTBILDNINGSVARIABEL (i dataset)

1. Grundskola eller motsvarande, kortare än   
9 år  
  
2. Grundskola eller motsvarande, 9 år eller längre

3. Gymnasium eller motsvarande, kortare än 3 år

4. Gymnasium eller motsvarande, 3 år eller längre

5. Eftergymnasial utbildning, ej högskola/universitet, kortare än 3 år

6. Eftergymnasial utbildning, ej högskola/universitet, 3 år eller längre

7. Högskola/universitet, kortare än 3 år

8. Högskola/universitet, 3 år eller längre men kortare än 4 år

9. Högskola/universitet, 4 år eller längre

10. Forskarutbildning

#### Förslag till omkodning

***Utbildning 4-delad***

1. Låg (1, 2 – dvs. max grundskola el. motsv.)  
2. Medellåg (3, 4 – dvs. max gymnasium, folkhögskola el. motsv.)  
3. Medelhög (5, 6 ,7 – dvs. eftergymnasial utb., högsk./univ. Kortare än 3 år)  
4. Hög (8, 9, 10 – dvs. högskola/universitet 3 år eller längre/forskarutbildning)

***Utbildning 2-delad***

1. Låg/medellåg (1–4)  
2. Medelhög/hög (5-10)

### KOMMUN (i dataset)

Okodad med kommunkod enligt registerdata.

1180 Stockholm   
1181 Södertälje   
1182 Nacka  
1184 Solna  
1183 Sundbyberg  
1139 Upplands-Bro  
1280 Malmö  
1281 Lund  
1283 Helsingborg  
1267 Höör  
1270 Tomelilla  
1272 Bromölla  
1480 Göteborg  
1481 Mölndal  
1485 Uddevalla  
1488 Trollhättan  
1490 Borås

#### Förslag till omkodning

***Län***

1. Stockholms län  
2. Västra Götalands län  
3. Skåne län

***Stad/land 4-delad***

1. Stockholm, Göteborg Malmö  
2. Stad eller större tätort  
3. Mindre tätort  
4. Ren landsbygd

***Stad/land 3-delad***

1. Storstad  
2. Stad eller större tätort  
3. Mindre tätort eller landsbygd

***Stad/land 2-delad***

1. Storstad  
2. Ej storstad

[**Se äen SKR:s kommungruppsindelning för vägledning.**](https://skr.se/skr/tjanster/kommunerochregioner/faktakommunerochregioner/kommungruppsindelning.2051.html)

### INKOMST (i dataset)

1. Mindre än 10 000 kr  
2. 10 000 kr – 14 999 kr  
3. 15 000 kr – 19 999 kr  
4. 20 000 kr – 24 999 kr  
5. 25 000 kr – 29 999 kr  
6. 30 000 kr – 34 999 kr  
7. 35 000 kr – 39 999 kr  
8. 40 000 kr – 44 999 kr  
9. 45 000 kr – 49 999 kr  
10. 50 000 kr – 54 999 kr  
11. 55 000 kr – 59 999 kr  
12. 60 000 kr – 64 999 kr  
13. 65 000 kr – 69 999 kr  
14. 70 000 kr – 74 999 kr  
15. Mer än 75 000 kr

#### Förslag omkodning

***Inkomst 4-delad***

1. Max 14 999 kr  
2. 15 000kr - 24 999 kr  
3. 25 000 - 44 999 kr  
4. Minst 45 000 kr

***Inkomst 5-delad***

1. Max 14 999 kr  
2. Mellan 15 000 och 24 999 kr  
3. Mellan 25 000 och 34 999 kr  
4. Mellan 35 000 och 44 999 kr  
5. Minst 45 000 kr