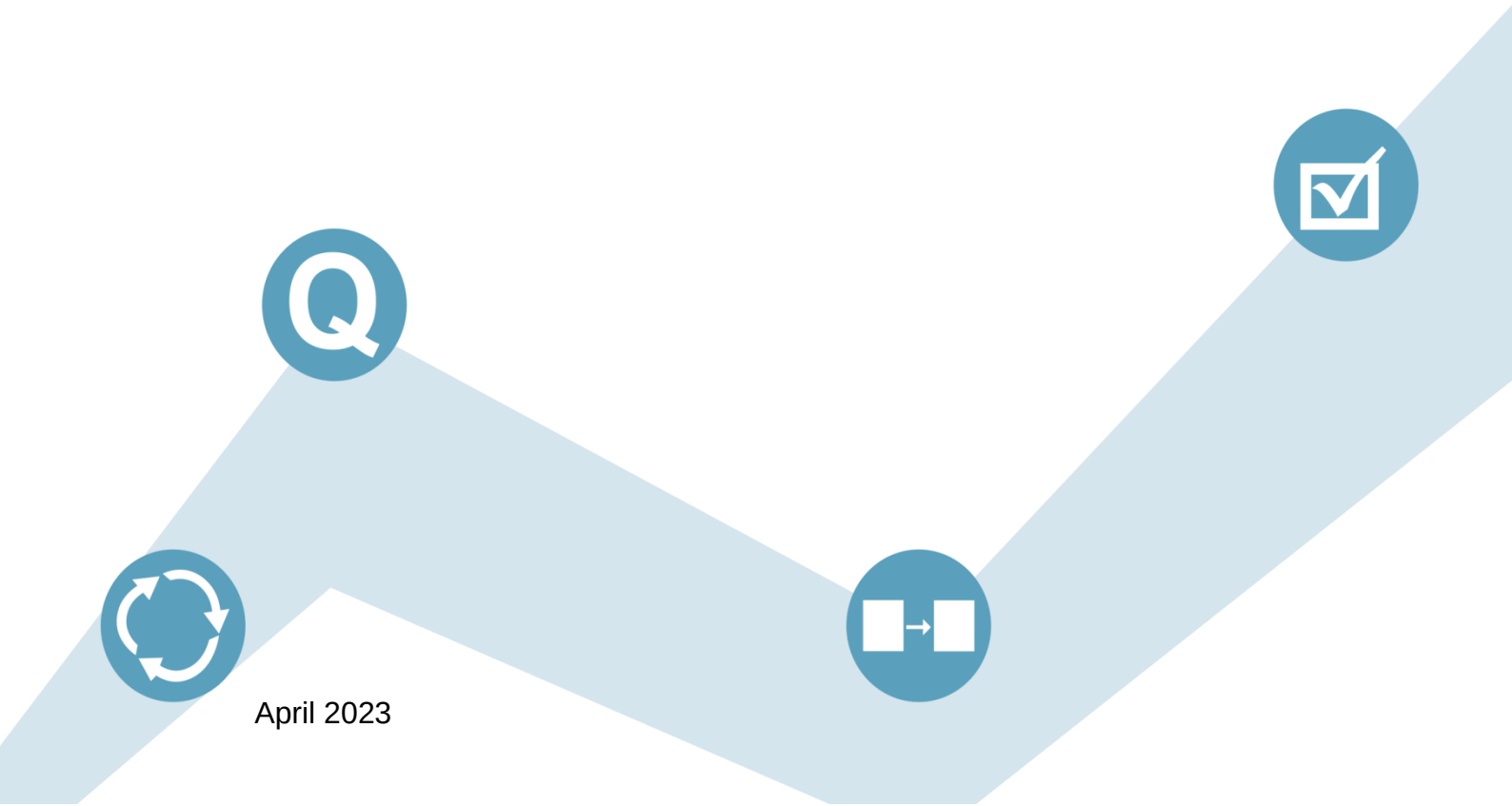
REPUBLIKA SLOVENIJA
STATISTIČNI URAD

SPLOŠNO METODOLOŠKO POJASNILO

Manca Golmajer

STATISTIČNA ZAŠČITA TABEL



April 2023

Kazalo

1 OPIS

Doseganje strateških ciljev in izpolnjevanje nalog državne statistike sta med drugim odvisni od zaupanja opazovanih enot oz. dajalcev podatkov: oseb, gospodinjstev, podjetij, kmetijskih gospodarstev in drugih organizacij. To pomeni, da nam bodo svoje podatke zaupali, če bodo prepričani, da ravnamo tako, da v celoti zagotavljamo njihovo zaupnost (tj. varujemo njihovo identiteto) in zaupnost danih informacij, in da bodo pridobljeni podatki uporabljeni samo za statistični namen. S tem namenom jih tudi nenehno obveščamo o tem, jim razlagamo uporabljene postopke in temu je namenjeno tudi to metodološko pojasnilo.

Del zaščite (varovanja) zaupnosti podatkov temelji na tem, da izkazani podatki ne omogočajo neposredne identifikacije (s pomočjo direktnih identifikatorjev) ali posredne identifikacije (na kateri koli drug način). K temu nas zavezujeta tako slovenska kot tudi evropska zakonodaja. Glede na to, da je poslanstvo državne statistike posredovati in objavljati statistične rezultate v čim širšem obsegu, vendar hkrati tako, da je tveganje razkritja informacij o enotah najmanjše možno, so potrebne metode statistične zaščite, s katerimi dosežemo, da se spoštuje zakonodaja in obenem ohranja zadovoljiva oz. čim večja obvestilna (informacijska) vrednost posredovanih oziroma objavljenih statističnih rezultatov.

Ločimo dve vrsti metod za statistično zaščito tabel:

- **metode, uporabljene na osnovnih mikropodatkih:**
  - *metode, ki spreminjajo podatke* (podatki naj bi se z uporabo take metode spremenili čim manj, vendar v tolikšni meri, da je tveganje razkritja najmanjše možno):
    - zaokroževanje;
    - metoda CTA;
    - metoda celičnih ključev;
  - *metode, ki ne spreminjajo podatkov* (podatki imajo manjšo informacijsko vrednost, vendar so še vedno točni):
    - metoda manjkajočih vrednosti;
    - združevanje v razrede;
- **metode za generiranje sintetičnih podatkov** (podatke dobimo z upoštevanjem statističnega modela).

Metode so implementirane v različnih orodjih (npr. Tau-Argus, SAS-Tool, R (paketi sdcTable, cellKey, ptable, sdcHierarchies)).

2 STATISTIČNA ZAŠČITA TABEL

Tabela (preglednica) je podatkovna struktura, ki realizira matematično strukturo matrike in razvrsti podatke v celice po vrsticah in stolpcih. V njej so prikazani agregirani podatki (različne statistike), ki opisujejo združene podatke več enot. Vsako statistiko, ki jo nameravamo posredovati, si lahko predstavljamo kot tabelo dimenzije 1 x 1. Tabelo definirajo (določajo) pojasnjevalne spremenljivke (spremenljivke, ki napenjajo tabelo), in odvisna spremenljivka (spremenljivka, ki polni tabelo). Število pojasnjevalnih spremenljivk določa število dimenzij tabele.

Pri postopku statistične zaščite tabel določimo pravila za določanje primarno občutljivih celic oz. vrednosti, ki so prikazane v celicah tabele. Od tipa tabele je odvisno, katera pravila bomo uporabili. Ločimo tri tipe tabel:

- **frekvenčne tabele** (v vsaki celici tabele je vpisano število enot, ki pripadajo tej celici);
- **vrednostne tabele** (v vsaki celici je vpisana vsota vrednosti določene spremenljivke enot, ki pripadajo tej celici);
- **druge tabele** (v celicah so vpisane vrednosti drugih tipov statistik, npr. deležev, razmerij, indeksov in drugih), ki so običajno povezane z vrednostnimi in frekvenčnimi tabelami.

Pravila za določanje primarno občutljivih celic so:

- **prag** (če je število enot manjše od praga, potem je celica primarno občutljiva); minimalna vrednost za prag naj bo 3;

- **pravilo dominantnosti** (če n največjih prispevkov prispeva več kot k % k vrednosti celice, potem je celica primarno občutljiva);
- **p%-pravilo** (če lahko enota z drugim največjim prispevkom oceni največji prispevek bolje kot na p % natančno, potem je celica primarno občutljiva).

Pri frekvenčnih tabelah upoštevamo le prag, medtem ko pri vrednostnih in drugih tabelah upoštevamo prag in p%-pravilo ali pa prag in pravilo dominantnosti. Pravila za določanje primarno občutljivih celic so zaupna in se ne smejo posredovati, saj bi z razkritjem parametrov zmanjšali zaščito enot.

Če se določene enote (dajalci podatkov) strinjajo, da se njihovi individualni podatki objavijo (za ta namen pridobimo njihovo pisno dovoljenje), ali če je del podatkov javnih, se lahko uporabi t. i. pravilo zahteve. To pomeni, da tem enotam dodelimo poseben status in tako zmanjšamo izgubo informacije v tabelah, v katerih se pojavljajo.

Z metodo za statistično zaščito tabel primarno občutljive celice vedno zaščitimo. Zaradi povezav znotraj tabele pa najpogosteje zaščitimo še dodatne celice (pravimo jim sekundarno zaščitene celice); če povezav ni, to ni potrebno.

Pri metodi manjkajočih vrednosti zakrijemo primarno in sekundarno občutljive vrednosti tako, da jih ne objavimo, tj. jih ne vpišemo v celico, temveč namesto vrednosti vpišemo črko „z“ (zaupen podatek). Izgubo informacije določa delež zakritih celic.

Pri metodi celičnih ključev slučajno spremenimo vse celice v tabeli ne glede na to, ali so primarno občutljive ali ne. Slučajna sprememba celice je lahko pozitivna, negativna ali nič. Tabela, zaščitena z metodo celičnih ključev, na splošno ni aditivna.

Pri statistični zaščiti tabel je treba paziti na hierarhije pojasnjevalnih spremenljivk, število dimenzij tabel in povezave med tabelami.

3 PRIMER STATISTIČNE ZAŠČITE TABELE

Naj ima frekvenčna tabela pojasnjevalno spremenljivko, za katero velja Total = 1 + 2, 1 = 11 + 12, in pojasnjevalno spremenljivko, za katero velja Total = A + B. Recimo, da je pravilo za določitev primarno občutljivih celic prag 3. Tako ima tabela samo eno primarno občutljivo celico, to je celico [11,A], ki ima vrednost 1. To primarno občutljivo celico lahko zaščitimo na več načinov.

Če uporabimo združevanje v razrede, zmanjšamo število primarno občutljivih celic tako, da tabelo preoblikujemo: odstranimo obe podkategoriji kategorije 1.

Če uporabimo metodo manjkajočih vrednosti, potem poleg primarno občutljive celice [11,A] zakrijemo še sekundarno občutljive celice [11,B], [12,A], [12,B]; s tem preprečimo, da bi uporabnik razkril pravo vrednost primarno občutljive celice s pomočjo enostavnih matematičnih funkcij.

Če uporabimo metodo celičnih ključev, potem slučajno spremenimo vrednost vsake celice v tabeli. Dobljena tabela ni aditivna, čeprav je nezaščitena tabela aditivna.

Nezaščitena tabela

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	1	6
12	14	10	4
2	30	12	18

Primarno zaščitena tabela

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	z	6
12	14	10	4
2	30	12	18

Končna zaščitena tabela

Primer zaščite z združevanjem v razrede:

	Total	A	B
Total	51	23	28
1	21	11	10
2	30	12	18

Primer zaščite z metodo manjkajočih vrednosti:

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	z	z
12	14	z	z
2	30	12	18

		Primer zaščite z metodo celičnih ključev:

	Total	A	B
Total	51	23	27
1	21	12	10
11	7	2	6
12	13	8	3
2	29	12	18