Տվյալների աղավաղման, ուղղման և իմպուտացիայի մեթոդներ

Երևան 8.2 – 12.2 2016

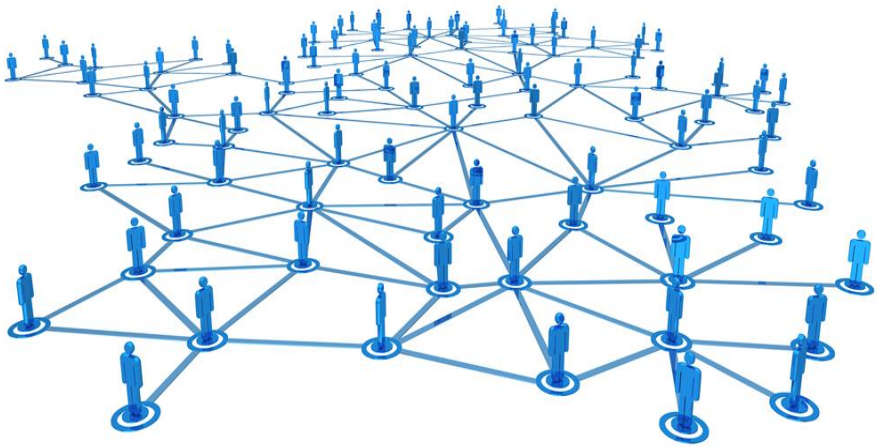
Էնրիկո Տուչի – Իստատ

Istat

Շրջանակ

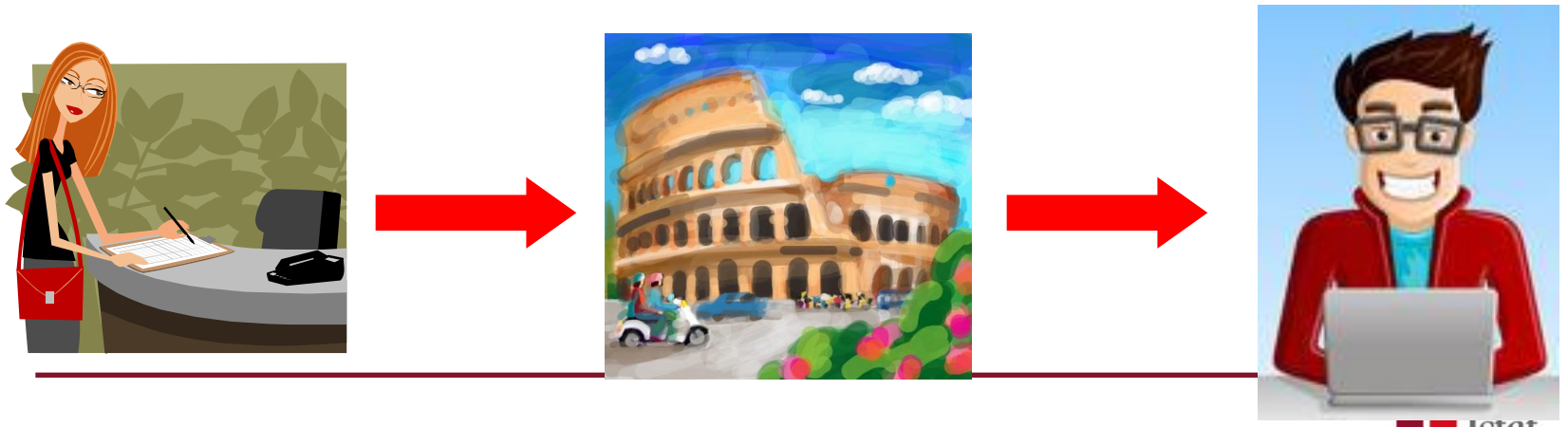
- Տվյալների հավաքագրման մեթոդներ
- Կրկնվող գրառումներ
- Տվյալների աղավաղում
- Տվյալների ուղղում և իմպուտացիա
- Տվյալների վավերացում

Տեղական տվյալների բազայից դեպի կենտրոնական տվյալների բազա



Տվյալների հավաքագրման մեթոդներ

- Թղթային տվյալներից անցում դեպի էլեկտրոնային տվյալների



Տվյալների աղավաղում/1

Տվյալների աղավաղումը վերաբերվում է այն սխալներին, որոնք առաջանում են գրելու, կարդալու, պահպանելու, փոխանցելու կամ մշակելու ժամանակ, ինչի արդյունքում առանց մտադրության փոփոխվում են սկզբնական տվյալները:

Համակարգչային փոխանցման ժամանակ կիրառվում են մի շարք միջոցառումներ տվյալների ամբողջականություն ապահովելու համար

Տվյալների ադավադում/2

Վավերացման կանոններ.

- **Չափի ստուգում.** Տվյալների միավոր արժեքի մեջ նիշերի թվի ստուգում
- **Ձևաչափի ստուգում.** Տվյալները պետք է համապատասխանեն կոնկրետ ձևաչափի
- **Կապակցվածության ստուգում.** Այսպիսով կարող է ստուգվել միմյանց հետ կապ ունեցող տվյալների միավորների կոդերի փոխկապակցվածությունը
- **Շրջանակի ստուգում.** Տվյալները պետք է տեղավորվեն սահմանված առավելագույն և նվազագույն արժեքների սահմաններում

Istat

Ի՞նչ անել հայտնաբերված սխալների հետ

- Կարելի է ճիշտ արժեքը դուրս բերել միևնույն գրառման այլ պատասխաններից
- Անհրաժեշտ է վերադառնալ տվյալի սկզբնաղբյուրին. Եթե սխալը բացահայտվել է, պետք է տեղայնացնել սխալի պատճառը

Եթե մուտքագրվել է ճիշտ արժեք, պետք է հետևել, որպեսզի այն ուղղվի

Ի՞նչ անել բացահայտված սխալների հետ

- Եթե սխալը հնարավոր չէ ուղղել, ապա պետք է նշել այն որպես բացակայող արժեք: Ավելի ուշ բացակայող արժեքները կարելի է փոխարինել լրահաշվված արժեքներով:

- Նշում. Իմպուտացիաները տվյալների շարքը դարձնում են ոչ թե ճիշտ, այլև ավելի կիրառելի

Բացակայող տվյալներ և իմպուտացիա

Բացակայող տվյալը հանդիսանում է սխալի աղբյուր և պահանջում է ուղղում ցանկացած տվյալների շարքում: Սա կարող է հանգեցնել լուրջ խնդիրների՝ վիճակագրական վերլուծության մեջ: Կարելի է օգտագործել իմպուտացիայի մեթոդներ այդ բացերը լրացնելու և տվյալների ամբողջական շարք ապահովելու համար:

Սովորաբար հեշտ է բաժնի պատասխանի բացակայության (ընդհանուր պատասխանի բացակայություն) արդյունքում ստացված բացակայող տվյալները տարբերակել միավորի պատասխանի (պատասխանի մասնակի բացակայություն) բացակայության արդյունքում ստացված բացակայող տվյալներից:

Առաջինն ուղղվում է իմպուտացիայի օգնությամբ, մինչդեռ վերջինը սովորաբար վերակշռվում է կամ վերանայվում գնահատման մեթոդներով:

Ամենաշատ օգտագործվող մեթոդներ/4

Դետերմինիստական իմպուտացիա.

Վերաբերվում է այն իրավիճակին, երբ դաշտի միայն մեկ արժեք կհանգեցնի բավարար արդյունքի (հաշվի առնելով այլ դաշտերի կոնկրետ արժեքները)

Այն փոխարինում է բացակա արժեքը օգտվելով փոփոխականների միջև տրամաբանական կապից և ստանալով արժեք բացակայող տարրի համար

- Օրինակ դա կարող է հանդիպել, երբ տարրերը, որոնք ենթադրվում է, որ պետք է գումարվեն ընդամենը արժեքին, չեն գումարվում: Եթե գումարի մեջ փոխարինենք միայն մեկ տարր, ապա դրա արժեքը բացառապես որոշվում է այլ տարրերի արժեքներով:

Պարզ իմպուտացիոն մեթոդներ /1

Իմպուտացիան մեթոդ է, որի օգնությամբ բացակայող տվյալը լրացվում է հավանական արժեքներով՝ տվյալների ամբողջական շարք ապահովելու համար

Միջին իմպուտացիա

1. Ոչ պայմանական միջին իմպուտացիա. Բացակայող արժեքները փոխարինվում են դիտարկված (այսինքն ռեսպոնդենտ) արժեքների միջինով:
2. Պայմանական միջին իմպուտացիա. Ռեսպոնդենտները և ոչ-ռեսպոնդենտները նախապես դասակարգվում են դասերի (ստրատաների) դիտարկված փոփոխականների հիման վրա և բացակայող արժեքները փոխարինվում են միևնույն դասի ռեսպոնդենտների միջինով:

Մեկուսացված տվյալների ազդեցությունից խուսափելու համար մեդիանան կարող է օգտագործվել միջինի փոխարեն: Կատեգորիալ տվյալների դեպքում մեթոդն օգտագործվում է իմպուտացիայի համար:

Ամենաշատ օգտագործվող մեթոդները/2

Ռեգրեսիոն իմպուտացիա

Այն ներառում է մեկ կամ մի քանի օժանդակ փոփոխականների օգտագործում, որոնց մեջ արժեքները հայտնի են ինչպես ամբողջ բաժինների, այնպես էլ հետաքրքրական փոփոխականի բաժինների համար, որտեղ առկա են բացակայող արժեքներ:

Դետերմինիստական ռեգրեսիոն իմպուտացիա. այս մեթոդը փոխարինում է բացակայող արժեքները կանխատեսված արժեքներով, որոնք ստացվել են բաժնի դիտարկված տարրերի վրա բացակայող տարրի ռեգրեսիայից:

Istat

Ամենաշատ օգտագործված մեթոդներ/3

Hot deck իմպուտացիա. բացակայող տվյալները փոխարինվում են արժեքներով, որոնք ստացվում են նմանատիպ «դոնոր» կոչվող ռեսպոնդենտներից:

Hot Deck իմպուտացիան պարզ տարբերակ է, որը ենթադրում է յուրաքանչյուր բացակայող տարրի փոխարինում պատահականորեն ընտրված տարրով հետաքրքրական փոփոխականի համար: Որպես այլընտրանք, իմպուտացիոն դասերը կարող են կառուցվել՝ դրանց մեջ ընտրելով պատահական դոնոր արժեքներ: Օգտակար կլիներ ստեղծել միատարր ստրատաներ (իմպուտացիոն վանդակներ), որտեղից պետք է լինի և՛ փոխարինվողը և՛ դոնորը:

Nearest-neighbour իմպուտացիա. Կամ հեռավորության գործառույթի համապատասխանեցումը մեթոդ է, որտեղ դոնորն ընտրված է՝ նվազեցնելով որոշակի «հեռավորություն»

Այսպիսով, բնական է փաստելը, որ շերտավորումը և համապատասխանող փոփոխականները պետք է մեծապես բնութագրեն դիտարկումները:

Տվյալների իմպուտացիայի այլ մեթոդներ…

Տվյալների վավերացում/1

Մեկուսացման հայտնաբերում

Մեկուսացումն իրենից ներկայացնում է չափազանց մեծ դիտարկում այն իմաստով, որ այն զարմանալիորեն տարբեր է այլ դիտարկումներից, ինչի արդյունքում կարելի է եզրակացնել, որ դա կարող է առաջացած լինել չափման, հավաքագրման, կոդավորման, գրանցման, մշակման … կամ մոդելի պատճառով առաջացած սխալների հետևանքով:

Տվյալների հետազոտական վերլուծություն

Գրաֆիկական աղյուսակների և թվային չափերի վրա հիմնված տվյալների հիմնական բնութագրիչների ուսումնասիրությունը և գործակիցները ամենայն հավանականությամբ դուրս կթողնեն տվյալների հետ կապված խնդիրների մեծ մասը: