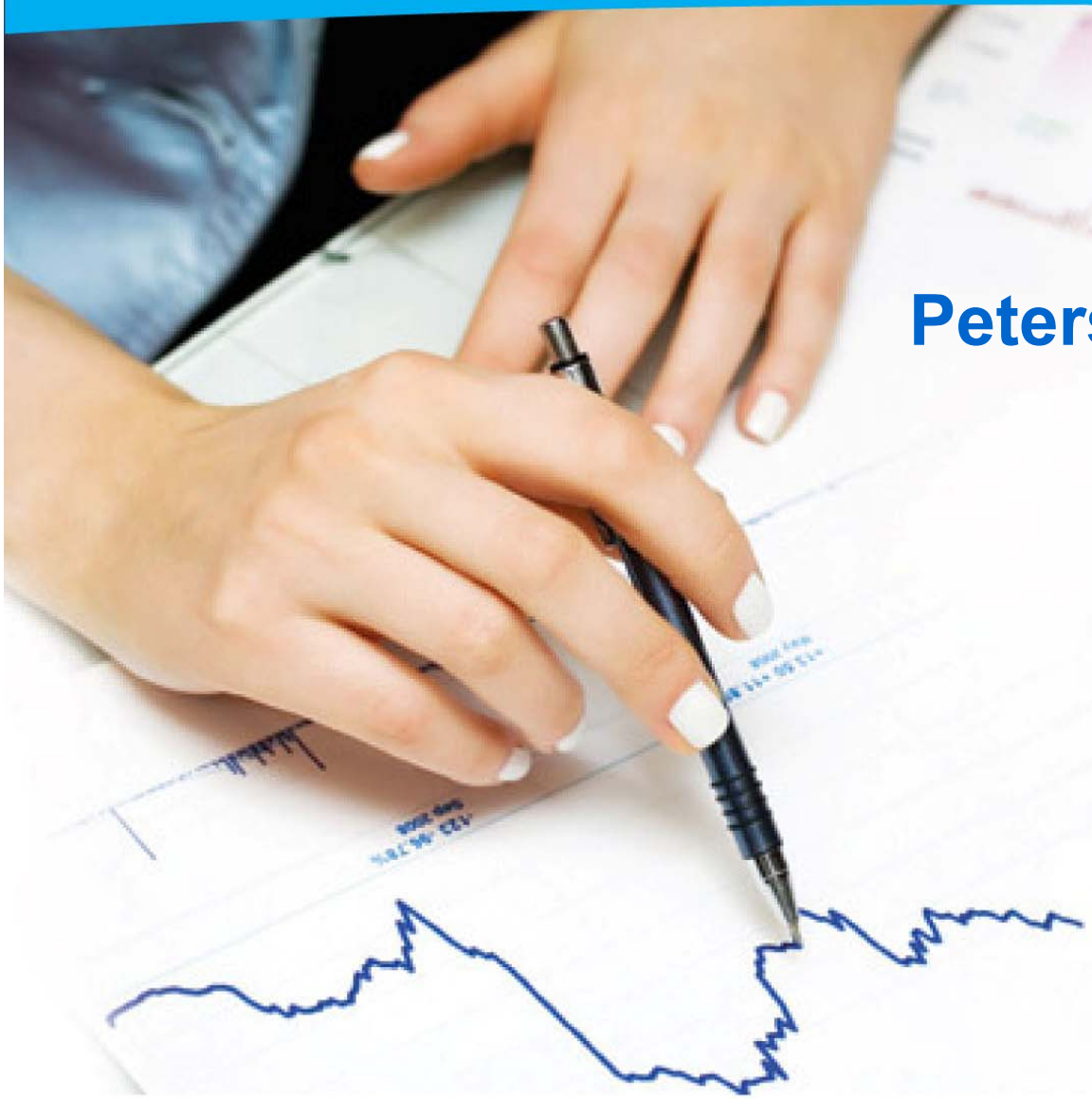
dnevi slovenske informatike

"Nove razmere in priložnosti v informatiki kot posledica družbenih sprememb"

Kongresni center Grand hotel BernardinPortorož
Slovenija | 18. - 20. april 2011



18. konferenca
Dnevi slovenske informatike

Petersonova ter Schnablova metoda za ocenjevanje kakovosti podatkov

dr. Uroš Godnov

mag. Tomaž Dular

18. 04. 2011

slovensko društvo informatika

Petersonova ter Schnablova metoda za ocenjevanje kakovosti podatkov

Agenda

- Dejstva o kakovosti podatkov
- Predstavitev Petersonove ter Schnablove metode
- Potek simulacije
- Rezultati

Dejstva

- 600 milijard dolarjev stroškov letno v ZDA
- Vpliv na informacijske projekte ter poslovne razultate podjetij
- Nezadovoljstvo kupcev (69 %)
- Sklepanje poslov (39 %)
- Izgubo prihodkov (35 %)
- Zvišanje stroškov (67 %)
- Zmanjšanje ugleda oziroma verodostojnosti (77 %)

Dejstva

- Stroški dosegajo 10 % prihodkov
- V storitvenih podjetjih pa celo do 50 % prihodkov
- **Pravilo 1 – 10 – 100 (preprečevanje – popravljanje – posledice)**
- Nekakovostni podatki:
  - izjemne vrednosti,
  - manjkajoče vrednosti,
  - neskladne vrednosti in
  - nepopolne vrednosti

Dejstva

- Sodobna orodja nam olajšajo delo
- Problem so semantično nenatančni podatki, še posebej semantično nenatančne vrstice v bazah podatkov
- Ocenjevanje kakovosti podatkov je odvisno od količine podatkov, ki jo moramo pregledati (milijoni proti nekaj tisoč vrstic)

Metode

- Metode, ki omogočajo sklepanje na velikost celotne populacije s pomočjo vzorcev:
  - Petersenova metoda
  - Schnablova metoda
  - Metoda nemškega tanka

Petersenova metoda

- Capture – recapture
- Ocenjevanje populacije rib v ribniku
- Prvi ulov, drugi ulov, iste ribe v prvem in drugem ulovu
- Če bi torej prvi dan ujeli 300 rib, drugi dan pa 150, od katerih bi bilo 50 takšnih, ki smo jih ulovili že prejšnji dan, bi lahko sklepali, da je v ribniku 900 rib.

$$N = \frac{MC}{R}$$

Schnablova metoda

- Schnablova metoda uporablja isti princip kot Petersenova metoda, le da za sklepanje na celotno populacijo uporablja najmanj tri vzorce.

$$N = \frac{\sum (C_t M_t)}{\sum R_t}$$

Čas (t)	C_t	R_t	U_t	M_t	$CM_t t$
1	20	0	20		0
2	20	5	15	20	400
3	20	7	13	35	700
4	20	10	10	48	960
Skupaj		22			2060

Simulacija

- Uporaba MS SQL R2
- Tabela Population je imela 10000 podatkov
- 100 meritev
- Stored procedura z zunanjo in notranjo zanko (zunanja=št. merjenj, notranja=št. ponavljanj znotraj merjenja)
- Naključna velikost vzorca:

  ABS(CHECKSUM(NEWID())) % @SampleSize + 1,

  kjer je @SampleSize zgornja velikost vzorca

Izsledki simulacije

- Kako so vrednosti porazdeljene in ali kakšen parameter povzroča asimetrično porazdelitev ocenjenih vrednosti
- Izračun povprečja, mediane ter modusa
- Izračun Kurtosisa in Skewnesa

Izsledki simulacije

Velikost posameznega vzorca	2 vzorca	3 vzorci	4 vzorci	5 vzorcev	6 vzorcev
Do 50 %	**10024**	**10024**	**10006**	**10005**	**10005**
Do 10 %	**11166**	**10233**	**10154**	**10121**	**10077**
Do 5 %	**11829**	**10426**	**10066**	**9997**	**9966**

Tabela: Povprečne vrednosti ocenjene populacije po posameznih korakih

Izsledki simulacije

Velikost posameznega vzorca	Vrsta	2 vzorca	3 vzorci	4 vzorci	5 vzorcev	6 vzorcev
Do 50 %	**Skewness**	**0,535**	**-0,11**	**0,734**	**0,627**	**1,160749**
	Kurtosis	**2,083**	**0,493**	**2,022**	**1,276**	**2,922964**
Do 10 %	**Skewness**	**5,601312581**	**1,762**	**0,406**	**0,607**	**-0,21**
	Kurtosis	**41,31149939**	**6,463**	**1,031**	**1,289**	**0,123**
Do 5 %	**Skewness**	**2,221705495**	**0,730201**	**1,593763**	**0,4958882**	**0,409571314**
	Kurtosis	**7,525712242**	**0,050264**	**4,996532**	**0,8738775**	**1,18662677**

Tabela: Vrednost Kurtosisa in Skewnessa ocenjene populacije po posameznih korakih

Sklep

- Manjši vzorci -> več ponavljanj; Večji vzorci -> Petersenova metoda
- Pomembno je upoštevati predpostavke metod, tj. homogenost populacije ter stalnost populacije

Hvala za pozornost!