ds dnevi slovenske informatike 20 let

"Dvajset let pozneje"

Kongresni center Grand hotel Bernardin
Portorož | Slovenija | 15. - 17. april 2013

20. konferenca
Dnevi slovenske informatike

ALI SO VREDNOSTI VAŠIH PODATKOV V PODATKOVNI ZBIRKI NORMALNO PORAZDELJENE:
PRIMER UPORABE »WINDOW« FUNKCIJ V MS SQL 2012

dr. Uroš Godnov

mag. Tomaž Dular

15. 04. 2013

slovensko društvo informatika

Agenda

- ZAKAJ STATISTIKA
- ZAKAJ NORMALNA PORAZDELITEV
- OBUDITEV „POZABLJENE“ STATISTIKE
- IZVEDBA TESTOV S POMOČJO MS SQL 2012

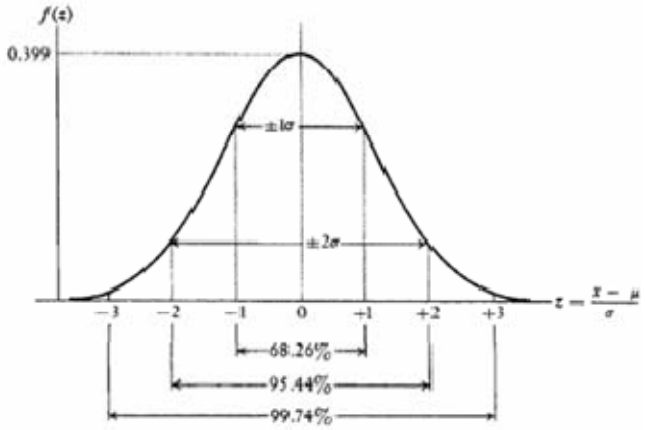
zakaj statistika

- Podatki → Informacije
- Dva načina:
  - ✓ Data driven (podatkovno rudarjenje)
  - ✓ Human driven (statistika)
- Data driven → raziskovalec ribari v kalnem, saj le okvirno pozna vzorce med podatki, skuša pa odkriti še neznane vzorce
- Human driven → raziskovalec mora vedeti, katere vzorce med podatki skuša potrditi ali ovreči
- Uporabnost statistike je zelo raznolika in običajno se vsakdo sreča vsaj enkrat s potrebo po določeni analizi, kjer mu statistika lahko pomaga

Zakaj normalna porazdelitev

- Uporaba statistike ni vedno enostavna
- Obstaja več vrst statistik:
  - ✓ Statistika za majhne vzorce (n<30)
  - ✓ Statistika za velike vzorce (n>200)
  - ✓ Vmesno sivo področje, kjer je potrebno izbrati ustrezne statistične metode glede na porazdelitev podatkov
- Normalna porazdelitev podatkov → parametrični test (t test, ANOVA, regresija,…)
- Nenormalna porazdelitev podatkov → neparametrični testi

Značilnosti normalne porazdelitve



- simetričnost
- zvončasta oblika
- zvezna krivulja od -∞ do +∞
- asimetrija je 0
- sploščenost je 0
- dva parametra, in sicer povprečje (μ) ter standardni odklon (σ)
- približno 2/3 vseh vrednosti se nahaja v intervalu μ ± σ: P(μ - σ ≤ X ≤ μ + σ) = .6826
- približno 95 % vseh vrednosti se nahaja v intervalu μ ± 2σ: P(μ - 2σ ≤ X ≤ μ + 2σ) = .9544

Obuditev pozabljene statistike

- Pearsonov drugI koeficient asimetrije
- Fisherjev koeficient asimetrije
- Kolmogorov – Smirnov test

Pearsonov drugi koeficient asimetrije

- Formula:

$$PI=\frac{3(\bar{x}-median)}{St.odklon}$$

Težava v različicah pred verzijo 2012 je bila odsotnost funkcije za mediano, medtem ko sta funkciji za povprečje in standardni odklon obstajali.

Pred MS SQL 2012

```
DECLARE @Median ASNUMERIC(4,2),
@PI AS NUMERIC(4,2);
-- Izračun mediane
WITH median AS (
SELECT value,rn_up=row_number()OVER
(ORDERBYvalue),
rn_down=row_number()OVER
(ORDERBYvalueDESC)
FROMdbo.DSI2013
)
SELECT @Median=AVG(value)
FROM median
WHERE abs(rn_up-rn_down)<= 1
-- Izračun koeficienta

SELECT @PI=3*(AVG(value)-
@Median)/STDEV(value)
FROMdbo.DSI2013

SELECT@PI
```

MS SQL 2012

```
DECLARE @PI AS NUMERIC(4,2)
SELECT @PI=3*(AVG(1.*Value)-
(SELECT DISTINCT Percentile_cont(0.5)
WITHIN GROUP (ORDER BY Value)OVER
(PARTITION BY 1)
FROM DSI2013))/STDEV(1.*Value) FROM
DSI2013

SELECT@PI
```

Fisherjev koeficient asimetrije

- Formula ni enotna, različni programi uporabljajo rahle modifikacije, ki pa bistveno ne vplivajo na končni rezultat

SPSS, STATA, R	Excel
$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}$	$G_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$

```
DECLARE @Povprecje AS float,
@Stevilo AS float,
@STDEV AS float

SELECT @Povprecje=Avg(1.*value),
@Stevilo= Count(1.*value),
@STDEV=Stdev(1.*value)FROM DSI2013;
----Računanje Skewnessa

WITH InterMediate AS
(SELECT SUM(POWER(1.*value-
@povprecje,3))/@Stevilo as m3,
SQRT(POWER(SUM(POWER(1.*value-
@povprecje,2))/@Stevilo,3)) as m2
FROM DSI2013)

SELECT ROUND(m3/m2,5) AS Skewness
FROM InterMediate
```

```
DECLARE @Povprecje AS DECIMAL(6,2),
@Stevilo AS DECIMAL(6,2),
@STDEV AS DECIMAL(6,2)

SELECT
@Povprecje=Avg(1.*value),@Stevilo=Count(*),
@STDEV=Stdev(1.*value) FROM DSI2013;
----Računanje Skewnessa

WITH InterMediate AS
(SELECT SUM(POWER((Value-
@Povprecje)/@STDEV,3))ASg,
@Stevilo/((@Stevilo-1)*(@Stevilo-2))as n,
SUM(POWER((Value-@Povprecje)/
@STDEV,3))/@Stevilo a sp
FROM DSI2013)

SELECT ROUND(g*n,5) AS Skewness
FROM InterMediate
```

Kolmogorov – Smirnov test

- H0: porazdelitev vrednosti podatkov sledi normalni porazdelitvi
- H1: porazdelitev vrednosti podatkov ne sledi normalni porazdelitvi

Zap. Št.	Gostota populacije	Relativne kumulativne frekvence	Z-vrednost	Z verjetnost	Pričakovane kumulativne frekvence	D_i	D'_i
1	4.13	0.0588	-1.40	0.0808	0.0808	0.0220	0.0808
2	4.53	0.1176	-0.94	0.1736	0.1736	0.0560	0.1148
3	4.69	0.1764	-0.75	0.2266	0.2266	0.0502	0.1090
4	4.76	0.2352	-0.67	0.2514	0.2514	0.0162	0.0750
5	4.77	0.2940	-0.66	0.2546	0.2546	0.0394	0.0194
6	4.96	0.3528	-0.44	0.3300	0.3300	0.0228	0.0360
7	4.97	0.4116	-0.43	0.3336	0.3336	0.0780	0.0192
8	5.00	0.4704	-0.39	0.3483	0.3483	0.1221	0.0633
9	5.04	0.5292	-0.35	0.3632	0.3632	0.1660	0.1072
10	5.10	0.5880	-0.28	0.3897	0.3897	0.1983	0.1395
11	5.25	0.6468	-0.10	0.4602	0.4602	0.1866	0.1278
12	5.36	0.7056	0.02	0.4920	0.5080	0.1976	0.1388
13	5.94	0.7644	0.69	0.2451	0.7549	0.0095	0.0493
14	6.06	0.8232	0.83	0.2033	0.7967	0.0265	0.0323
15	6.19	0.8820	0.98	0.1635	0.8365	0.0455	0.0133
16	6.30	0.9408	1.11	0.1335	0.8665	0.0743	0.0155
17	7.73	0.9996	2.76	0.0029	0.9971	0.0025	0.0563

Kolmogorov – Smirnov test

- Relativne kumulativne frekvence so izračunane kot [Zap. št.]/17, »Z vrednosti« in »Z verjetnosti« so prepisane iz tabele Z vrednosti na podlagi upoštevanih kumulativnih frekvenc
- Pričakovane kumulativne frekvence so izračunane iz »Z verjetnosti«, in sicer, če je Z vrednost < 0, potem je frekvenca enaka »Z verjetnosti«, drugače je frekvenca enaka »1-Z verjetnost«
- Di je absolutna razlika med relativnoter pričakovano frekvenco
- D'i je absolutna razlika med relativno frekvenco ter naslednjo pričakovano frekvenco. Največja razlika med MS SQL 2012 in starejšimi različicami je ravno v tem koraku, saj z novimi »window« funkcijami lahko različne analize opravimo bolj učinkovito

Kolmogorov – Smirnov test

Pred MS SQL 2012	MS SQL 2012
SELECT ABS(A.Upoštevanekumulativnefrekvence-B.Pričakovanekumulativnefrekvence) FROM DSI2013 AS A JOIN DSI2013 AS B ONA.Zap.št=B.Zap.št+1	SELECT ABS(Upoštevanekumulativnefrekvence-LEAD(Pričakovanekumulativnefrekvence) OVER (ORDER BY Vrednosti ASC)) FROM DSI2013

- Poiščemo maksimalno vrednost Di ter D'i in ju primerjamo s kritično vrednostjo za [K-S] testa. Za večje vzorce (>35) pri Sig=0,05 kritično vrednost lahko izračunamo kot $\sqrt{\frac{1,22}{N}}$
- Če je kritična vrednost manjša od Di ter D'i, potem ne moremo zavrniti ničelne domneve, zato obvelja ničelna domneva, ki pravi, da porazdelitev vrednosti podatkov sledi normalni porazdelitvi

Zaključek

- Z uporabo novih Window fukcij v MS SQL 2012 nam je omogočeno, da na lažji način izdelamo statistike, ki nam omogočajo preverjanje hipoteze o normalnosti porazdelitve. Seveda nam te funkcije omogočajo tudi lažjo izdelavo drugih statističnih testov, ki pa niso bili predmet našega članka.

Hvala za vašo pozornost !

Vprašanja?

Pripombe?

Predlogi?