PRIMJENA STATISTIKE U ORACLE BAZI PODATAKA

Zlatko Sirotić, univ.spec.inf.

ISTRA TECH d.o.o.

Pula

Teme

- Primijenjena statistika je danas jako važna za umjetnu inteligenciju (AI), naročito kod strojnog učenja. No i izvan AI, statistika se intenzivno primjenjuje, i to na velikim količinama podataka, smještenim u (različite) baze podataka.
- Često se koriste specijalizirani alati / jezici za statistiku. Neki su komercijalni, kao SAS, SPSS, MATLAB, a neki besplatni, kao jezik (opće namjene) Python i statistički jezik R.
- Oracle SUBP može koristiti jezik R na dva načina. Jedan način je da se podaci iz baze prebace na R klijenta, a rezultati se vraćaju na bazu. Drugi način, bolji i skuplji, je da se direktno koristi R stroj (engine) unutar Oracle baze.
- Postoji još jedna varijanta – korištenje (skoro 300) SQL statističkih funkcija, od kojih je većinu moguće koristiti u svim edicijama Oracle baze, pa i onoj besplatnoj (XE edicija).

Razvoj napredne analitike u Oracle bazi podataka

Oracle Advanced Analytics Database Evolution

1998 — ORACLE 8i INTERNET
- *7 Data Mining "Partners"*

1999 — DATABASE 9i CLUSTER
- *Oracle acquires Thinking Machine Corp's dev. team + "Darwin" data mining software*

2002
- *Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API*

2004 — ORACLE 10g DATABASE
- *Oracle Data Mining 10gR2 SQL - 7 new SQL dm algorithms and new Oracle Data Miner "Classic" wizards driven GUI*
- *SQL statistical functions introduced*

2005
- *ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements*
- *SQLDEV/Oracle Data Miner 3.2 "work flow" GUI launched*
- *Integration with "R" and introduction/addition of Oracle R Enterprise*
- *Product renamed "Oracle Advanced Analytics (ODM + ORE)*

2008 — ORACLE DATABASE 11g

2011 — ORACLE DATABASE 12c
- *New algorithms (EM, PCA, SVD)*
- *Predictive Queries*
- *SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)*
- *OAA/ORE 1.3 + 1.4 adds NN, Stepwise, scalable R algorithms*
- *Oracle Adv. Analytics for Hadoop Connector launched with scalable BDA algorithms*

2014 — hadoop

ORACLE

Oracle Advanced Analytics

- OAA je opcija Oracle baze podataka (Enterprise Edition), što znači da se posebno plaća.
- OAA postoji od baze 11g i sastoji se od dva dijela: **Oracle Data Mining** (koji je postojao i u prethodnim verzijama baze, 9i i 10g) i **Oracle R Enterprise** (koji je uveden u 11g).
- Oracle Data Mining se može ukratko opisati kao skup SQL funkcija i PL/SQL paketa, koji služi za data mining unutar baze podataka (in-database data mining). Naravno, tu su i grafički alati, npr. Oracle Data Miner Workflow GUI (ekstenzija Oracle SQL Developer alata).
- Oracle R Enterprise integrira programski jezik R (koji je open source) unutar Oracle baze podataka.
- Sljedeći slajd pokazuje mogućnosti OOA opcije.

Advanced Analytics Database Option - opcija (dodatno plaćanje)

Oracle Advanced Analytics Database Option

Wide Range of In-Database Data Mining and Statistical Functions

- **Data Understanding & Visualization**
  - Summary & Descriptive Statistics
  - Histograms, scatter plots, box plots, bar charts
  - R graphics: 3-D plots, link plots, special R graph types
  - Cross tabulations
  - Tests for Correlations (t-test, Pearson's, ANOVA)
  - Selected Base SAS equivalents
- **Data Selection, Preparation and Transformations**
  - Joins, Tables, Views, Data Selection, Data Filter, SQL time windows, Multiple schemas
  - Sampling techniques
  - Re-coding, Missing values
  - Aggregations
  - Spatial data
  - SQL Patterns
  - R to SQL transparency and push down
- **Classification Models**
  - Logistic Regression (GLM)
  - Naive Bayes
  - Decision Trees
  - Support Vector Machines (SVM)
  - Neural Networks (NNs)
- **Regression Models**
  - Multiple Regression (GLM)
  - Support Vector Machines
- **Clustering**
  - Hierarchical K-means
  - Orthogonal Partitioning
  - Expectation Maximization
- **Anomaly Detection**
  - Special case Support Vector Machine (1-Class SVM)
- **Associations / Market Basket Analysis**
  - A Priori algorithm
- **Feature Selection and Reduction**
  - Attribute Importance (Minimum Description Length)
  - Principal Components Analysis (PCA)
  - Non-negative Matrix Factorization
  - Singular Vector Decomposition
- **Text Mining**
  - Most OAA algorithms support unstructured data (i.e. customer comments, email, abstracts, etc.)
- **Transactional & Spatial Data**
  - All OAA algorithms support transactional data (i.e. purchase transactions, repeated measures over time, distances from location, time spent in area A, B, C, etc.)
- **R packages—ability to run open source**
  - Broad range of R CRAN packages can be run as part of database process via R to SQL transparency and/or via Embedded R mode

ORACLE

* included free in every Oracle Database

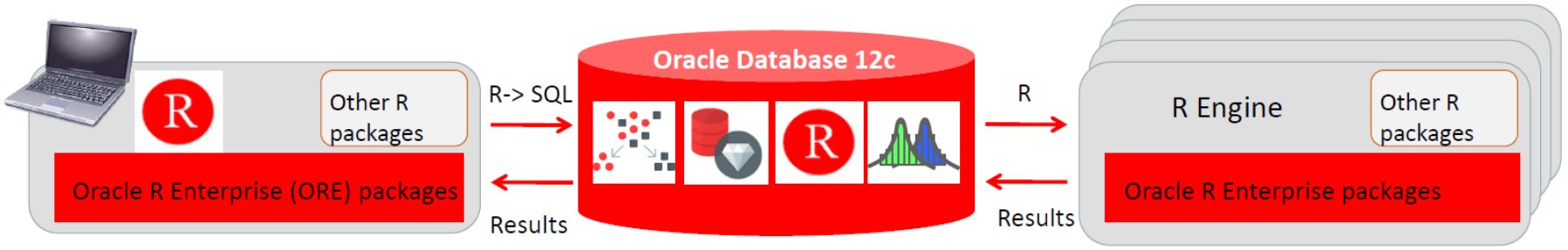
Oracle R i Oracle R Enterprise

- Kako je već rečeno, OAA opciju, uz Oracle Data Mining, čini i Oracle R Enterprise (uveden u 11g).
- R programski jezik se može koristiti nad Oracle bazom i **bez OOA opcije**, ako se R jezik koristi (samo) na klijentskoj strani.
- Postoji i (besplatni) **Oracle R Distribution** (Oracle-supported redistribution of open source R). Omogućuje bolju skalabilnost i bolje performance kada se (kao klijent) koristi sa ORE (koji se izvršava na bazi). Osim toga, omogućuje dinamičko učitavanje biblioteka za linearnu algebru (npr. Intelovu Math Kernel Library) koje omogućuju da se neke R funkcije izvršavaju u višedretvenom modu.
- Oracle R Enterprise integrira programski jezik **R unutar Oracle baze podataka**. Sljedeći slajd pokazuje na koji način radi Oracle R Enterprise engine.

Kako radi Oracle R Enterprise Compute Engine

Oracle Advanced Analytics

How Oracle R Enterprise Compute Engines Work



1 R-> SQL Transparency "Push-Down"

- R language for interaction with the database
- R-SQL Transparency Framework overloads R functions for scalable in-database execution
- Function overload for data selection, manipulation and transforms
- Interactive display of graphical results and flow control as in standard R
- Submit user-defined R functions for execution at database server under control of Oracle Database

2 In-Database Adv Analytical SQL Functions

- 15+ Powerful data mining algorithms (regression, clustering, AR, DT, etc._
- Run Oracle Data Mining SQL data mining functioning (ORE.odmSVM, ORE.odmDT, etc.)
- Speak "R" but executes as proprietary in-database SQL functions—machine learning algorithms and statistical functions
- Leverage database strengths: SQL parallelism, scale to large datasets, security
- Access big data in Database and Hadoop via SQL, R, and Big Data SQL

3 Embedded R Package Callouts

- R Engine(s) spawned by Oracle DB for database-managed parallelism
- ore.groupApply high performance scoring
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables production deployment and automated execution of R scripts

ORACLE

Kako radi Oracle R Enterprise Compute Engine

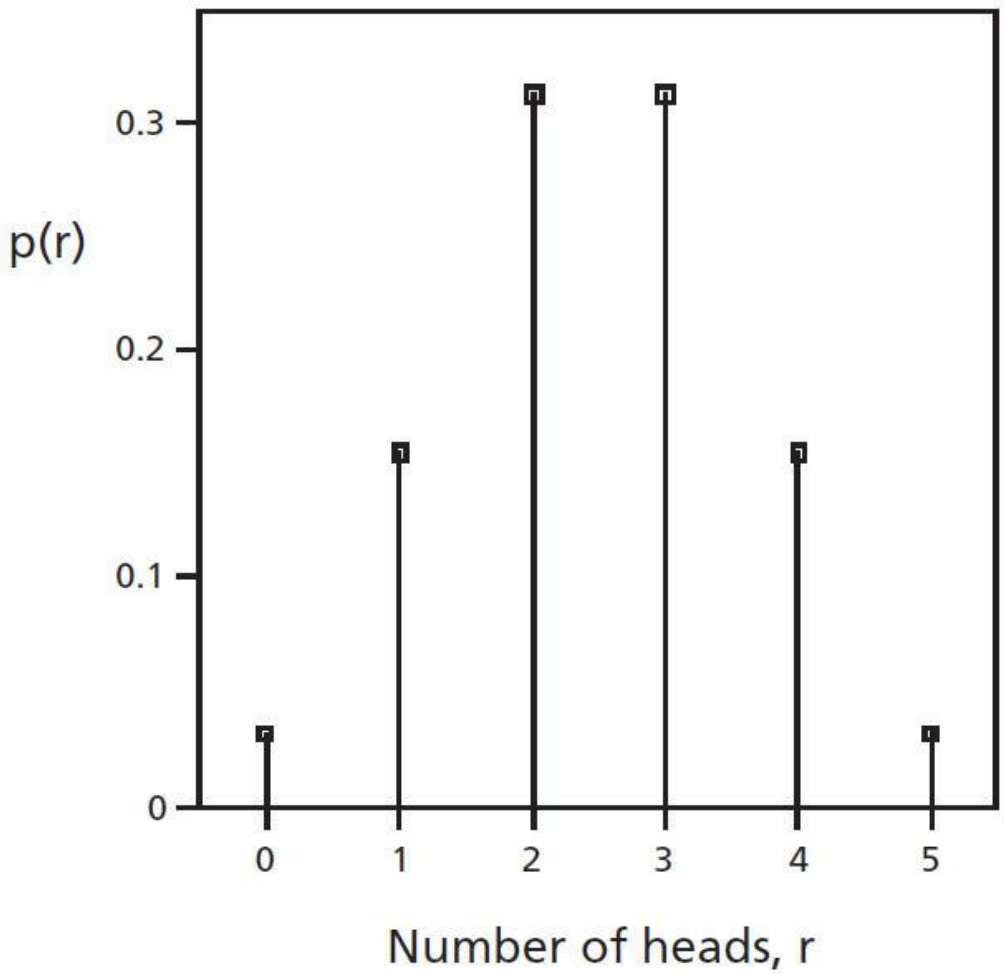
- R na klijentskom računalu presreće R funkcije i prosljeđuje ih serveru. Rezultate dobivene sa baze prikazuje grafički. Može raditi i transformacije podataka i koristiti statističke funkcije i napredne analitičke funkcije.
- R unutar Oracle baze omogućuje obradu velikih količina podataka direktno u bazi, i može koristiti SQL paralelizam (koji ima EE baza). Koristi postojeće statističke i data mining SQL funkcije / PL/SQL pakete koji se nalaze u Oracle bazi.
- R Engine (uključujući Oracle R Enterprise pakete) je R sustav prilagođen od strane Oracle-a. Moguće je istovremeno startati više takvih R sustava, tj. koristiti paralelizam i na R razini.
- Ukratko, ORE eliminira problem ograničenja memorije na klijentskom računalu, omogućuje paralelizam i optimizaciju upita na bazi, ali i paralelizam na R razini.

Matematička statistika (podsjetnik) - diskontinuirane slučajne varijable

- Postoje diskontinuirane (diskretne) i kontinuirane slučajne varijable.
- **Diskontinuirana slučajna varijabla** takva je varijabla *x* koja
  - prima niz vrijednosti x_1, x_2, …
  - ali svaku od njih s određenom vjerojatnošću $p(x_1)$, $p(x_2)$, …
  - pri čemu vjerojatnosti $p(x_i)$ zadovoljavaju jednakost

    $$\sum p(x_i) = 1$$
- Zakon p(x) po kojem svakoj vrijednosti x_i pripada vjerojatnost $p(x_i)$ zovemo **funkcijom (gustoće) vjerojatnosti diskontinuirane slučajne varijable** x (najčešći engl. termin je Probability Mass Function - PMF).

Primjer: Funkcija vjerojatnosti varijable r (broj glava kod bacanja pet novčića)



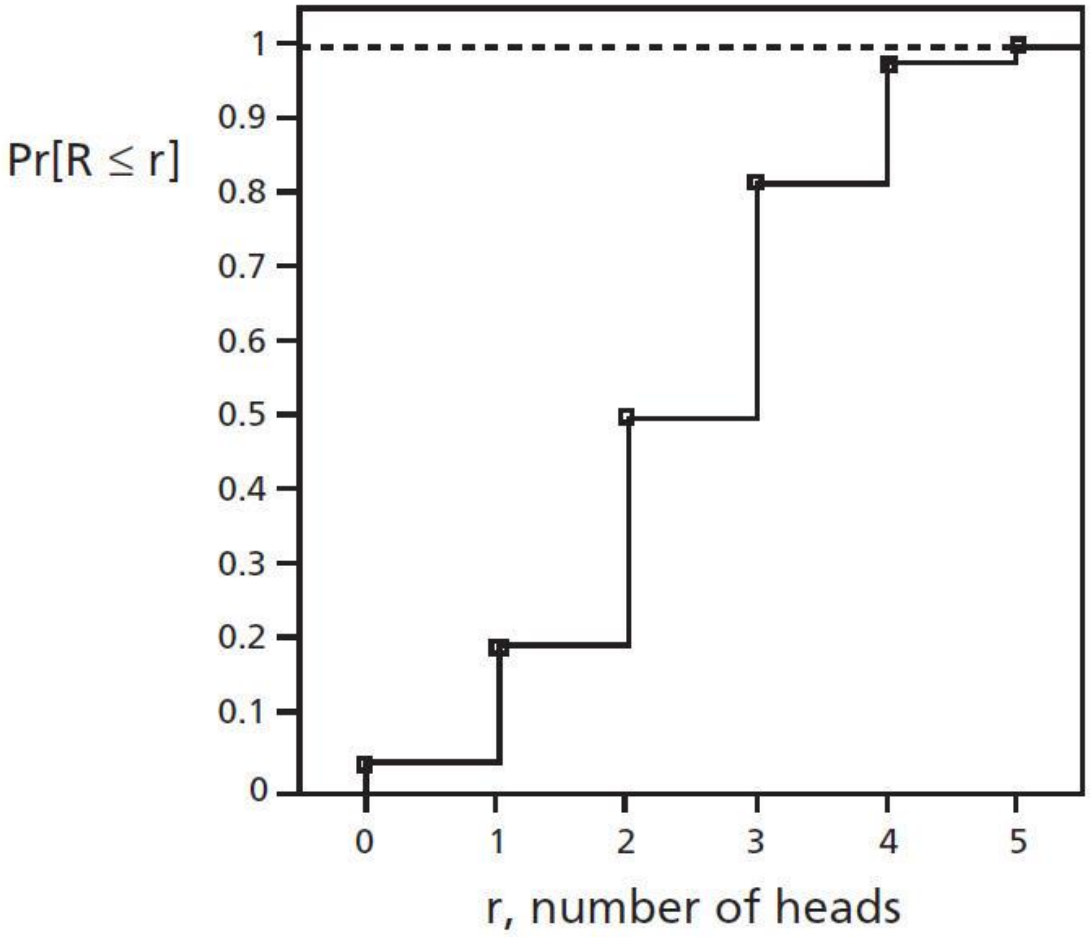
Funkcija distribucije (ili razdiobe) slučajne varijable

- Osim funkcije vjerojatnosti, kod diskretnih slučajnih varijabli važna je **funkcija distribucije (ili razdiobe) vjerojatnosti slučajne varijable** (engl. Cumulative Distribution Functions - CDF).

- Ona pokazuje kolika je vjerojatnost da slučajna varijabla x poprimi bilo koju vrijednost $\leq x_0$:

$$F(x_0) = \sum_{x_i \leq x_0} p(x_i) \qquad \text{tj.} \qquad F(x_0) = P\{x \leq x_0\}$$

Primjer: Funkcija distribucije varijable r (broj glava kod bacanja pet novčića)

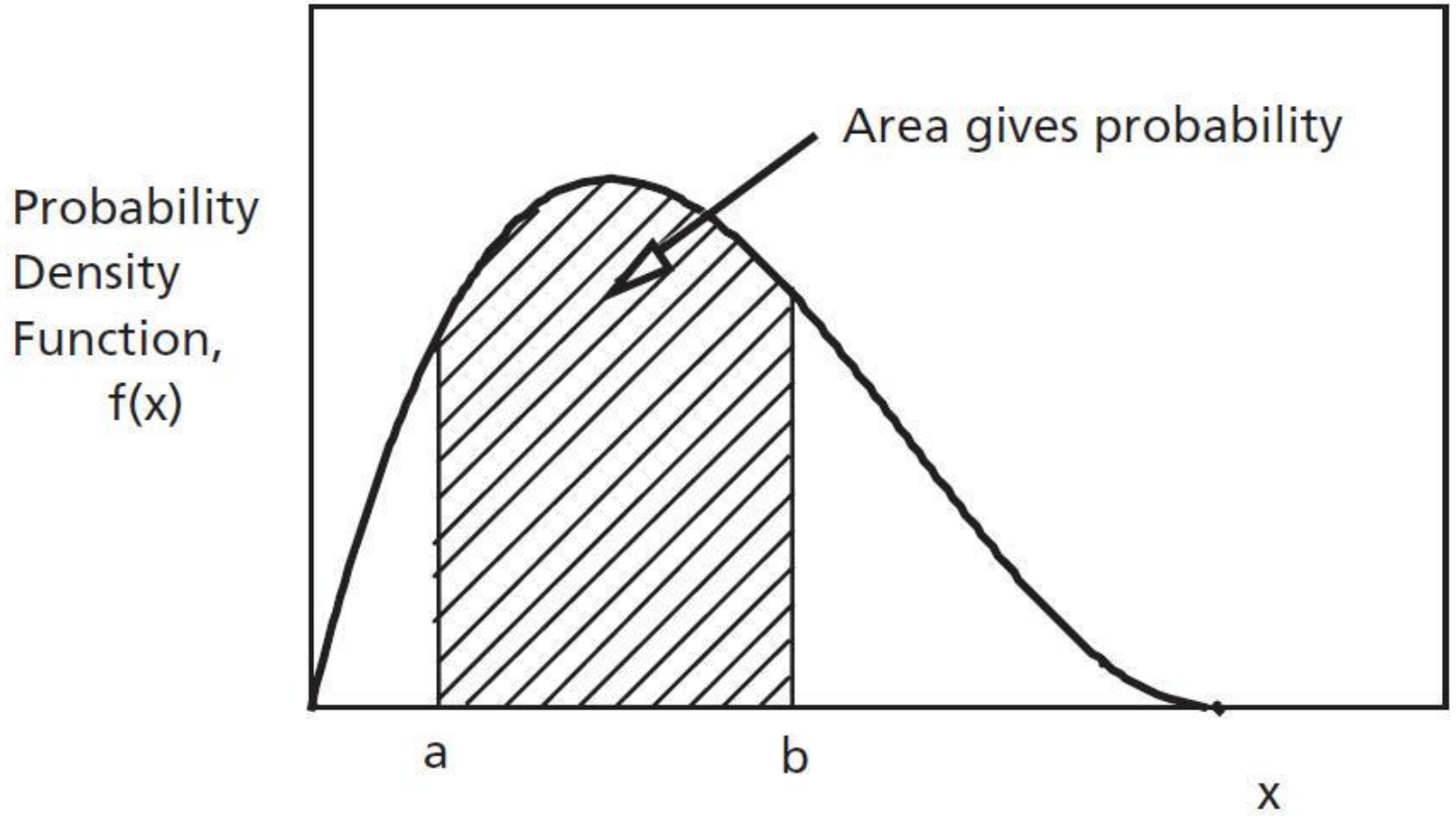


Kontinuirane slučajne varijable

- **Funkcija (gustoće) vjerojatnosti kontinuirane slučajne varijable** x (engl. Probability Density Function - PDF) je takva funkcija f(x) koja ima svojstva:

1. $f(x) \geq 0$ za svaki x iz domene funkcije [a, b]
   (a može biti $-\infty$, b može biti $+\infty$)
2. $\int_a^b f(x)\,dx = 1$
   (površina ispod funkcije unutar domene [a, b] je 1)
3. $\int_{x1}^{x2} f(x)\,dx = P\{x_1 \leq x \leq x_2\}$
   (površina ispod funkcije unutar domene $[x_1, x_2]$ jednaka je vjerojatnosti da varijabla poprimi vrijednost iz te domene).

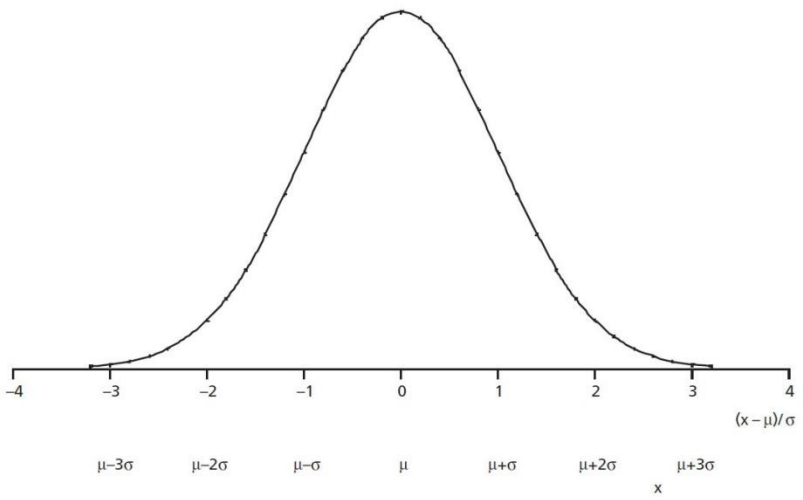
Vjerojatnost kod kontinuirane slučajne varijable

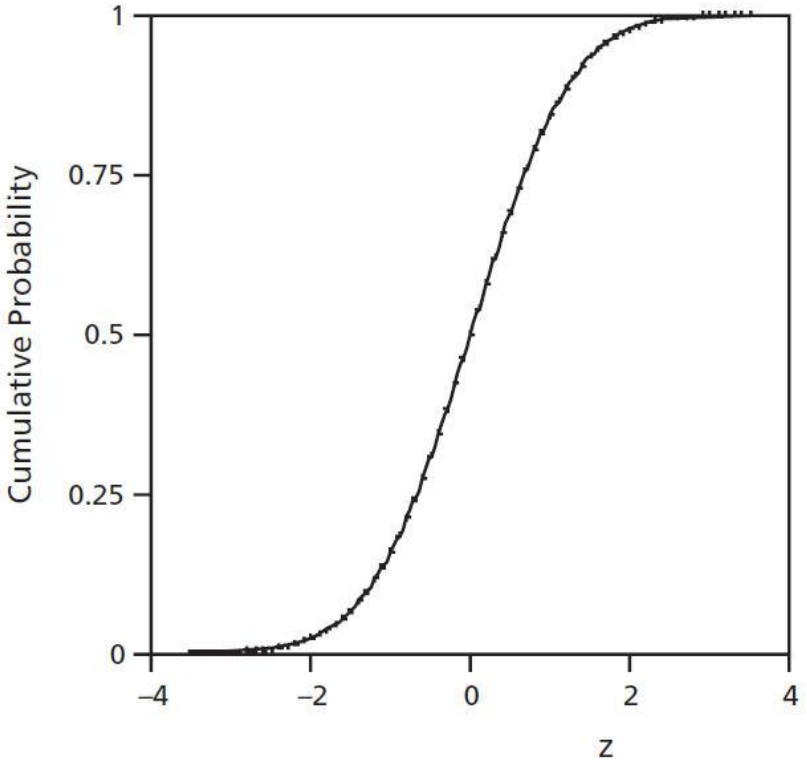


Normalna razdioba (Gaussova krivulja)

- Jedna od najpoznatijih funkcija (gustoće) vjerojatnosti (kontinuirane varijable) je tzv. **normalna razdioba** (poznata i kao Gaussova krivulja, po matematičaru Gaussu).
- Važna je po tome što mnoge druge razdiobe (diskontinuirane i kontinuirane) graniče prema njoj ako neki parametri rastu u beskonačnost.
- Prema njoj graniči i razdioba aritmetičkih sredina uzoraka, bez obzira na razdiobu osnovnog skupa, ako broj elemenata teži ka beskonačnosti (centralni granični teorem o razdiobi aritmetičkih sredina velikih uzoraka).
- Posebno postoji tzv. **jedinična (ili standardna) normalna razdioba**, kod koje je matematičko očekivanje = 0, a standardna devijacija = 1.

Funkcija (gustoće) vjerojatnosti i funkcija distribucije vjerojatnosti kod (jedinične) normalne razdiobe

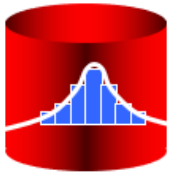




Oracle statističke i analitičke funkcije (još od 10g)

10g Statistics & SQL Analytics

FREE (Included in Oracle SE & EE)



- Ranking functions
  - rank, dense_rank, cume_dist, percent_rank, ntile
- Window Aggregate functions (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value
- LAG/LEAD functions
  - Direct inter-row reference using offsets
- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report
- Statistical Aggregates
  - Correlation, linear regression family, covariance
- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions.

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition

- Descriptive Statistics
  - average, standard deviation, variance, min, max, median (via percentile_count), mode, group-by & roll-up
  - DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, stats_mode, variance, standard deviation, median, quantile values, +/- n sigma values, top/bottom 5 values
- Correlations
  - Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).
- Cross Tabs
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa
- Hypothesis Testing
  - Student t-test , F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA
- Distribution Fitting
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential
- Pareto Analysis (documented)
  - 80:20 rule, cumulative results table

ORACLE

Shema HR u Oracle bazi za testiranje (ima je i XE edicija baze)

- Shema HR ima 7 tablica (međusobno povezanih). Jedna od njih je tablica EMPLOYEES. Zamislimo da ona sadrži samo uzorak iz nekog (velikog) osnovnog skupa. Prikaz nekih podataka iz te tablice:

```
select employee_id, first_name, last_name, salary
  from employees
 order by employee_id;
```

```
EMPLOYEE_ID FIRST_NAME           LAST_NAME                      SALARY
----------- -------------------- ------------------------- ----------
        100 Steven               King                            24000
        101 Neena                Kochhar                         17000
        102 Lex                  De Haan                         17000
...
        204 Hermann              Baer                            10000
        205 Shelley              Higgins                         12008
        206 William              Gietz                            8300
```

Neke jednostavne statističke funkcije

- Minimalna i maksimalna vrijednost, medijan, mod i aritmetička sredina za vrijednosti u stupcu SALARY tablice EMPLOYEES:

```
select min(salary),
       max(salary),
       median(salary),
       stats_mode(salary),
       avg(salary)
  from employees;
```

```
MIN(SALARY) MAX(SALARY) MEDIAN(SALARY) STATS_MODE(SALARY) AVG(SALARY)
----------- ----------- -------------- ------------------ -----------
       2100       24000           6200               2500  6461.83178
```

Neke jednostavne statističke funkcije

- Procjena varijance i standardne devijacije (korijen od varijance) na temelju uzorka (dijeli se sa N – 1; razlika između STDDEV / STDDEV_SAMP je u tome što vraćaju 0 / NULL ako postoji samo jedan redak), i standardna devijacija populacije (ako podaci sadrže cijelu populaciju; dijeli se sa N) za vrijednosti u stupcu SALARY tablice EMPLOYEES:

```
select variance(salary),
       stddev(salary),
       stddev_samp(salary),
       stddev_pop(salary)
  from employees;

VARIANCE(SALARY) STDDEV(SALARY) STDDEV_SAMP(SALARY) STDDEV_POP(SALARY)
---------------- -------------- ------------------ ------------------
      15284813.7     3909.57973         3909.57973         3891.26778
```

PL/SQL paket (u Oracle bazi) DBMS_STAT_FUNCS

```
set serveroutput on
set echo on
declare
  s DBMS_STAT_FUNCS.SummaryType;
begin
  DBMS_STAT_FUNCS.SUMMARY('HR', 'EMPLOYEES', 'SALARY', 3, s);
  dbms_output.put_line('SUMMARY STATISTICS');
  dbms_output.put_line('Count:          '||s.count);
  dbms_output.put_line('Min:            '||s.min);
  dbms_output.put_line('Max:            '||s.max);
  dbms_output.put_line('Range:          '||s.range);
  dbms_output.put_line('Mean:           '||round(s.mean));
  dbms_output.put_line('Mode Count:     '||s.cmode.count);
  dbms_output.put_line('Mode:           '||s.cmode(1));
  dbms_output.put_line('Variance:       '||round(s.variance));
  dbms_output.put_line('Stddev:         '||round(s.stddev));
  dbms_output.put_line('Quantile 5      '||s.quantile_5);
  dbms_output.put_line('Quantile 25     '||s.quantile_25);
  dbms_output.put_line('Median          '||s.median);
  dbms_output.put_line('Quantile 75     '||s.quantile_75);
  dbms_output.put_line('Quantile 95     '||s.quantile_95);
  dbms_output.put_line('Extreme Count:'||s.extreme_values.count);
  dbms_output.put_line('Extremes:       '||s.extreme_values(1));
  dbms_output.put_line('Top 3:
            '||s.top_5_values(1)||','||s.top_5_values(2)||','||s.top_5_values(3));
  dbms_output.put_line('Bottom 3:
            '||s.bottom_5_values(5)||','||s.bottom_5_values(4)||','||s.bottom_5_values(3));
end;
```

PL/SQL paket (u Oracle bazi) DBMS_STAT_FUNCS


```
SUMMARY STATISTICS
Count: 107
Min: 2100
Max: 24000
Range: 21900
Mean: 6462
Mode Count: 1
Mode:  2500
Variance: 15284814
Stddev:  3910
Quantile 5 2500
Quantile 25 3100
Median  6200
Quantile 75 8900
Quantile 95 12702.4
Extreme Count: 1
Extremes: 24000
Top 3:  24000,17000,17000
Bottom 3: 2100,2200,2200
PL/SQL procedure successfully completed.
```

DBMS_STAT_FUNCS
sadrži i distribution-fitting funkcije

- Paket sadrži i funkcije za provjeru "uklapanja" (prilagodbe) vrijednosti iz uzorka u određenu (teorijsku) distribuciju (distribution-fitting functions):

- NORMAL_DIST_FIT function
- UNIFORM_DIST_FIT function
- POISSON_DIST_FIT function
- WEIBULL_DIST_FIT function
- EXPONENTIAL_DIST_FIT function

- Ove funkcije testiraju koliko se dobro vrijednosti iz uzorka prilagođavaju određenoj (teorijskoj) distribuciji.

DBMS_STAT_FUNCS
sadrži i distribution-fitting funkcije

```
DECLARE
  sredina NUMBER := 6000;
  sdevijacija NUMBER := 3000;
  signif NUMBER := 0;
BEGIN
  -- Kod normalne distribucije, 4.parametar (vrsta testa) može biti:
  -- 'CHI_SQUARED', 'KOLMOGOROV_SMIRNOV',
  -- 'ANDERSON_DARLING', 'SHAPIRO_WILKS'
  --
  DBMS_STAT_FUNCS.NORMAL_DIST_FIT ('HR', 'EMPLOYEES', 'SALARY',
    'CHI_SQUARED', sredina, sdevijacija, signif);
  DBMS_OUTPUT.PUT_LINE('Signifikantnost: ' || signif);
END;
/
X-squared value : 84.850467289719626168224299065420560747 65
  Degree of freedom : 15
Signifikantnost: .9999999999103495
```

Testiranje statističkih hipoteza

- Funkcije (gustoće) vjerojatnosti slučajne varijable x ovise o parametrima, npr. parametri ne-jedinične normalne razdiobe su matematičko očekivanje i standardna devijacija.
- Ako jedan **nepoznati parametar promatramo kao varijablu**, a ostale kao konstantu, onda možemo postaviti **hipotezu** H_0 da je vrijednost tog parametra npr. P_0, te alternativnu hipotezu H_1, da je vrijednost parametra P_1 (moguće su i drugačije varijante postavljanja hipoteza).
- Odluku o tome da li prihvaćamo hipotezu H_0 ili H_1 donosimo na temelju **testiranja uzorka**, koji je uvijek konačan. Kod testiranja, uobičajeno su **moguće četiri situacije, dvije u kojima smo donijeli ispravnu odluku i dvije u kojima smo donijeli pogrešnu odluku.**

Testiranje statističkih hipoteza – mogući ishodi

Hipoteza H_0	Istinita	Neistinita
Odbacuje se	**Greška 1.vrste (vjerojatnost je α)**	Pravilan zaključak
Prihvaća se	Pravilan zaključak	**Greška 2.vrste (vjerojatnost je β)**

- Manja greška α rezultira većom greškom β (i obrnuto), ali **ne vrijedi** $\alpha + \beta = 1$.
- Uobičajeno se zadaje (mali) α (5% ili 1%), dok β ne znamo unaprijed. Jakost testa $p = 1 - \beta$.

Statistički testovi (testiranje hipoteza)

- Oracle baza podataka ima nekoliko parametarskih testova i nekoliko neparametarskih testova. Parametarski testovi imaju neke pretpostavke, a tipična je da su podaci u osnovnom skupu distribuirani po normalnoj razdiobi.
- Parametarski testovi:
  - T-test
  - F-test
  - One-Way ANOVA
- Neparametarski testovi:
  - Binomial test
  - Wilcoxon Signed Ranks test
  - Mann-Whitney test
  - Kolmogorov-Smirnov test

T-test

- T-test se najčešće koristi za mjerenje signifikantnosti razlika aritmetičke sredine uzorka i zadane vrijednosti, ili razlika aritmetičkih sredina dvaju uzoraka.
- Postoje 4 funkcije s prefiksom STATS_T_TEST_*

- STATS_T_TEST_ONE:
  test s jednim uzorkom
- STATS_T_TEST_PAIRED:
  test između dva uzorka (paired t-test, crossed t-test)
- STATS_T_TEST_INDEP:
  test dvije nezavisne grupe s jednakom varijancom (pooled variances)
- STATS_T_TEST_INDEPU:
  test dvije nezavisne grupe s nejednakom varijancom (unpooled variances)

STATS_T_TEST_ONE

- T-test hipoteze o signifikantnosti razlike između aritmetičke sredine uzorka i zadane vrijednosti 6500:

```
select avg(salary),
       STATS_T_TEST_ONE(salary, 6500, 'STATISTIC') t_statistic,
       STATS_T_TEST_ONE(salary, 6500, 'TWO_SIDED_SIG') t_sig
  from employees;

AVG(SALARY) T_STATISTIC      T_SIG
----------- ----------- ----------
 6461.83178   -.1009866 .919751863
```

STATS_T_TEST_ PAIRED

- T-test hipoteze o (ne)jednakosti aritmetičkih sredina dvaju uzoraka – uspoređuju se prosječne plaće zaposlenika odjela 50 i odjela 80:

```
select avg(decode(department_id, '50', salary, null)) avg_dept50,
       avg(decode(department_id, '80', salary, null)) avg_dept80,
       STATS_T_TEST_PAIRED(department_id, salary, 'STATISTIC') t_stat,
       STATS_T_TEST_PAIRED(department_id, salary, 'TWO_SIDED_SIG')t_sig
  from employees
 where department_id in ('50', '80');

AVG_DEPT50 AVG_DEPT80     T_STAT      T_SIG
---------- ---------- ---------- ----------
3475.55556 8955.88235 -15.925636 3.1440E-26
```

F_TEST

- F-test hipoteze o (ne)jednakosti između varijanci dvaju uzoraka – uspoređuju se varijance plaća zaposlenika u odjelima 50 i 80:

```
select variance(decode(department_id, '50', salary, null)) var_dept50,
       variance(decode(department_id, '80', salary, null)) var_dept80,
       STATS_F_TEST(department_id, salary, 'STATISTIC', 50) f_stat,
       STATS_F_TEST(department_id, salary, 'TWO_SIDED_SIG') f_sig
  from employees
 where department_id in ('50', '80');

VAR_DEPT50 VAR_DEPT80 F_STATISTIC      F_SIG
---------- ---------- ----------- ----------
2214161.62 4135873.44  .535355264 .053026341
```

One-Way ANOVA

- Analiza varijance s jednim promjenljivim faktorom – analiziraju se plaće po različitim poslovima (job_id), za svaki odjel posebno (group by department_id), ali samo za odjele koji imaju više od 5 zaposlenika:

```
select department_id,
       STATS_ONE_WAY_ANOVA(job_id, salary, 'F_RATIO') f_ratio,
       STATS_ONE_WAY_ANOVA(job_id, salary, 'SIG') p_value,
       avg(salary),
       count(1)
  from employees
 group by department_id
having count(1) > 5
 order by 1;
```

DEPARTMENT_ID	F_RATIO	P_VALUE	AVG(SALARY)	COUNT(1)
30	987.842105	6.1073E-06	4150	6
50	123.493632	2.5691E-18	3475.55556	45
80	26.3968778	.00001331	8955.88235	34
100	23.7247927	.008214884	8601.33333	6

Neparametarska statistika - koeficijent korelacije

- Koeficijent korelacije između trajanja zaposlenosti (u danima) i plaće:

```
select CORR_S(sysdate - hire_date, salary)
         coefficient,
       CORR_S(sysdate - hire_date, salary, 'TWO_SIDED_SIG')
         p_value
  from employees;

COEFFICIENT    P_VALUE
---------- ----------
 .126235719 .195097794
```

Linearna regresija

- Linearna regresija između trajanja zaposlenosti (u danima) i plaće. Rezultati su nagib pravca i odsječak na osi y:

```
select REGR_SLOPE(sysdate - hire_date, salary)
         reg_slope,
       REGR_INTERCEPT(sysdate - hire_date, salary)
         reg_intercept
  from employees;

 REG_SLOPE REG_INTERCEPT
---------- -------------
.063023539    4079.37252
```

Sekvencijalno testiranje hipoteza

- Sekvencijalno testiranje hipoteza (ili sekvencijalna analiza) je posebna statistička metoda testiranja. Smatra se da ju je prvi kreirao Abraham Wald, dok je bio u američkoj vojsci za vrijeme drugog svjetskog rata.
- Njena specifičnost sastoji se u tome da rezultat (jednog koraka) testa može biti trojak:
  - hipoteza H0 se može prihvatiti
  - ili se može se odbaciti
  - **ili (za razliku od uobičajenih statističkih testova) rezultat može biti neodlučan, pa se odabire još jedan element u uzorak**, i tako se nastavlja dok se hipoteza H0 ne prihvati ili odbaci.
- Dalje, kod sekvencijalnog testiranja hipoteza, **moguće je unaprijed zadati (barem približno) vjerojatnosti greške 1. i 2. vrste (α i β)**.

Sekvencijalno testiranje hipoteza

- Oracle baza (za sada) nema sekvencijalno testiranje hipoteza kroz SQL funkcije. No, za tu namjenu postoji više R paketa, koji se mogu koristiti bilo na R klijentu, bilo na Oracle R stroju.
- Opis jednog od tih R paketa – SPRT

- Type Package
- Title **Wald's Sequential Probability Ratio Test**
- Version 1.0
- Date 2015-04-14
- Author Stephane Mikael Bottine
- Maintainer Stephane Mikael Bottine <stephane.bottine@gmail.com>
- Description **Perform Wald's Sequential Probability Ratio Test**
on variables with a Normal, Bernoulli, Exponential and Poisson distribution.
Plot acceptance and continuation regions, or create your own with the help of closures.
- License BSD_2_clause + file LICENSE
- NeedsCompilation no
- Repository CRAN
- Date/Publication 2015-04-15 01:01:11

Zaključak

- Oracle baza podataka još od verzije 9i ima Oracle Data Mining (ODM) opciju, a u verziji 11g je dodan Oracle R Enterprise (ORE), što zajedno čini Oracle Advanced Analytics (OAA) opciju (koja se posebno plaća).
- Statistički jezik R se u Oracle bazi podataka može koristiti i bez ORE, ali se tada podaci iz baze moraju prebaciti na R klijent, a rezultati se (ako je potrebno) vraćaju na bazu. Time se smanjuje mogućnost obrade vrlo velikih količina podataka, i paralelizam (koji omogućuje baza podataka). No ta je varijanta jeftinija, pa i besplatna (može raditi i s XE bazom).
- Postoji još jedna varijanta primjene statistike u Oracle bazi. Od baze 10g na raspolaganju je skoro 300 SQL statističkih funkcija, od kojih je većinu moguće koristiti u svim edicijama Oracle baze, pa i besplatnoj XE ediciji.

Literatura

- DeCoursey, W.J. (2003): Statistics and Probability for Engineering Applications, Newnes, Woburn Massachusetts
- Iličić, K. (2017): Matematičke osnove statistike, Element, Zagreb
- Pavlić, I. (1977): Statistička teorija i primjena (2. izdanje), Tehnička knjiga, Zagreb
- Russell, S.J. & Norvig, P. (2010): Artificial Intelligence – A Modern Approach (3. izdanje), Prentice Hall - Pearson Education, New Jersey
- Oracle priručnici za bazu 12c Release 1 (2013):
  - SQL Language Reference (E17209-14)
  - PL/SQL Packages and Types Reference (E17602-14)
  - Data Mining Concepts (E17692-13)
  - Data Mining User's Guide (E17693-13)
  - R Enterprise User's Guide (E35158-07)