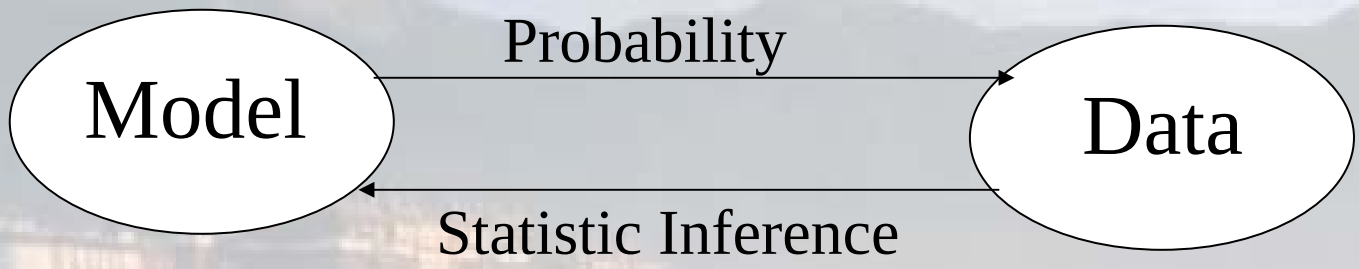
Statistic Analysis
for
Data Mining

Achmad Basuki
EEPIS-ITS
2004

Statistic Inference



Probability: representation of how much data can be took for model

Statistic Inference: How data can be represented the model.

→ Estimation
→ Test of Hypothesis

Probability

We have 4 favorites singer (A,B,C and D) and we pool response from people who's the best. From 20 transaction we has: following data:

A	A	B	D	B	C	A	D	B	C
B	B	C	D	B	C	A	B	A	B
C	D	B	A	B	B	A	C	B	A

Probability favorites for each singer are:

P(X=A) = 8/30

P(X=B) = 12/30

P(X=C) = 6/30

P(X=D) = 4/30

Statistic Inference

We have data of mobile-phone transaction in 10 day:

Day	1	2	3	4	5	6	7	8	9	10
Number of sale	3	1	3	6	3	2	0	6	2	5

We can estimate mean of transaction is 3 (using mean estimation).
Test of Hypothesis using t-student test:

```
>> x=[3 1 3 6 3 2 0 6 2 5];
>> [H,P]=ttest(x,3,0.05,0)
H =
     0
P =
   0.8793
```

Now we now mean-estimator can be represented mobile-phone sale model based on data

T-Test Description

TTEST Hypothesis test: Compares the sample average to a constant.
[H,P,CI,STATS] = TTEST(X,M,ALPHA,TAIL) performs a T-test to determine if a sample from a normal distribution (in X) could have mean M.
M = 0, ALPHA = 0.05 and TAIL = 0 by default.

The Null hypothesis is: "mean is equal to M".
For TAIL=0, alternative: "mean is not M".
For TAIL=1, alternative: "mean is greater than M"
For TAIL=-1, alternative: "mean is less than M"
TAIL = 0 by default.

ALPHA is desired significance level.
P is the p-value, or the probability of observing the given result by chance given that the null hypothesis is true. Small values of P cast doubt on the validity of the null hypothesis.
CI is a confidence interval for the true mean. Its confidence level is 1-ALPHA.
STATS is a structure with two elements named 'tstat' (the value of the test statistic) and 'df' (its degrees of freedom).

H=0 => "Do not reject null hypothesis at significance level of alpha."
H=1 => "Reject null hypothesis at significance level of alpha."

Model of Data

- $X = \{X_1, X_2, \ldots, X_n\}$ is attributes
- $T = \{ (x_{11}, x_{12}, \ldots, x_{1n}), (x_{21}, x_{22}, \ldots, x_{2n}), \ldots, (x_{m1}, x_{m2}, \ldots, x_{mn}) \}$ is tuples
- Y is random variable to estimate the model
- Center of data → mean, median, mode
- Dispersion of data → variance and standard deviation

	Attribut 1	Attribut 2	Attribut 3	…….	Attribut n
tupple 1	xxxxx	xxxxx	xxxxx	…….	xxxxx
tupple 2	xxxxx	xxxxx	xxxxx	…….	xxxxx
tupple 3	xxxxx	xxxxx	xxxxx	…….	xxxxx
…….	…….	…….	…….	…….	…….
tupple 1	xxxxx	xxxxx	xxxxx	…….	xxxxx

Model of Data for Sale-Transaction

Day	Number of Sale	
	Doll	Battery
1	3	7
2	2	0
3	2	3
4	0	0
5	3	4
6	1	2
7	6	5
8	2	3
9	7	5
10	2	1

X1 = Number of Sale for Doll

X2 = Number of Sale for Battery

Center of Data:

	X1	X2
mean	2.8	3
median	2	3
mode	2	3

Dispersion of Data:

	X1	X2
Varians	4.6222	5.3333
St Dev.	2.1499	2.3094

Correlation Two Attributes

Covarians of attribut X1 and X2 have defined:

$$\text{cov}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

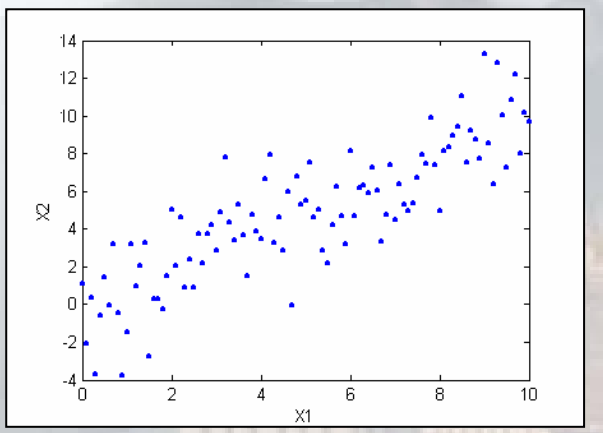
Correlation of attribut X1 and X2 have defined:

$$corr(x_1, x_2) = \frac{1}{n\sigma_1\sigma_2}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

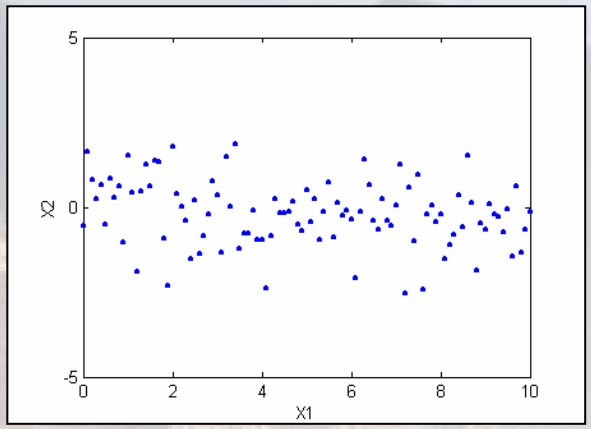
or

$$corr(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sigma_1\sigma_2}$$

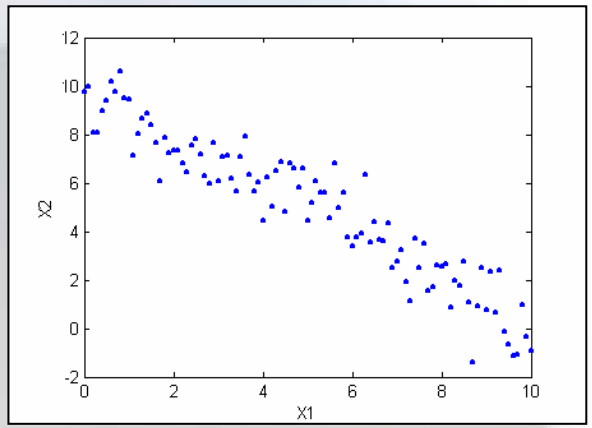
Correlation Description



$Corr(x_1,x_2) > 0$



$Corr(x_1,x_2) = 0$



$Corr(x_1,x_2) < 0$

Statistic independence

Correlation to Sale Attributes

Day	Number of Sale	
	Doll	Battery
1	3	7
2	2	0
3	2	3
4	0	0
5	3	4
6	1	2
7	6	5
8	2	3
9	7	5
10	2	1

$$corr(x_1, x_2) = \frac{\sum_{i=1}^{10} (x_{i1} - 2.8)(x_{i2} - 3)}{(10)(2.15)(2.31)}$$

$$= \frac{38.4}{49.65} = 0.773$$

X_1 and X_2 have positive correlation.

→ If X_1 increased then X_2 increased

→ If X_1 decreased then X_2 decreased

Estimator

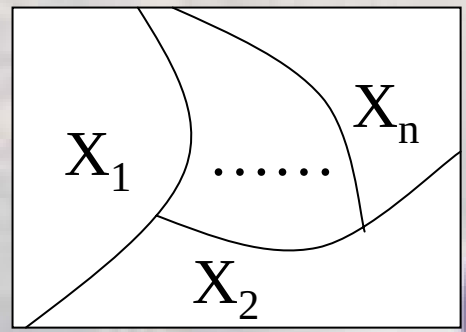
- Y is random variable to estimate the model, Y is called Estimator.
- Y is numeric → estimate process is called Regression
- Y is unordered dataset → estimate process is called classification.

Bayes Theorema

$$p(X_1 \mid X_2) = \frac{P(X_1, X_2)}{P(X_1).P(X_2)}$$

$P(X_1|X_2)$ is probability X_1 with conditional X_2
$P(X_1,X_2)$ is combination probability X_1 and X_2
$P(X_1)$ probability X_1
$P(X_2)$ probability X_2

$$P(X_1 \cup X_2 \cup ... \cup X_n) = 1$$



Bayes Theorem for Who's like coffee [1]

Respondents	Age	Gender	Like Coffee
1	Young	Male	Yes
2	Old	Male	Yes
3	Young	Male	Yes
4	Young	Female	No
5	Old	Female	No
6	Young	Male	Yes
7	Old	Female	Yes
8	Young	Female	No
9	Young	Male	No
10	Old	Male	Yes

Probability people like coffee P(C=Yes) = 6/10

P(C=No) = 4/10

Mr. Bean is old man, Is he like coffee ?

Bayes Theorem for Who's like coffee [2]

Mr. Bean is old man, Is he like coffee ?

Respondents	Age	Gender	Like Coffee
1	Young	Male	Yes
2	Old	Male	Yes
3	Young	Male	Yes
4	Young	Female	No
5	Old	Female	No
6	Young	Male	Yes
7	Old	Female	Yes
8	Young	Female	No
9	Young	Male	No
10	Old	Male	Yes

$P(A=Old \mid C=Yes) = 3/6$

$P(A=Old \mid C=No) = 1/4$

$P(B=Male \mid C=Yes) = 5/6$

$P(B=Male \mid C=No) = 1/4$

X is Old Man :

$$P(X|C=Yes) = P(A=Old|C=Yes).P(B=Male|C=Yes)$$
$$= (3/6).(5/6) = 15/36 = 0.4167$$
$$P(X|C=No) = P(A=Old|C=No).P(B=Male|C=No)$$
$$= (1/4).(3/4) = 3/16 = 0.1875$$

Bayes Theorem for Who's like coffee [3]

Mr. Bean is old man, Is he like coffee ?

X is Old Man : $P(C=Yes|X) = P(X|C=Yes).P(C=Yes)$

$= (0.4167).(0.6) = 0.250$

$P(C=No|X) = P(X|C=No).P(C=No)$

$= (0.1875).(0.4) = 0.075$

The Resume is:

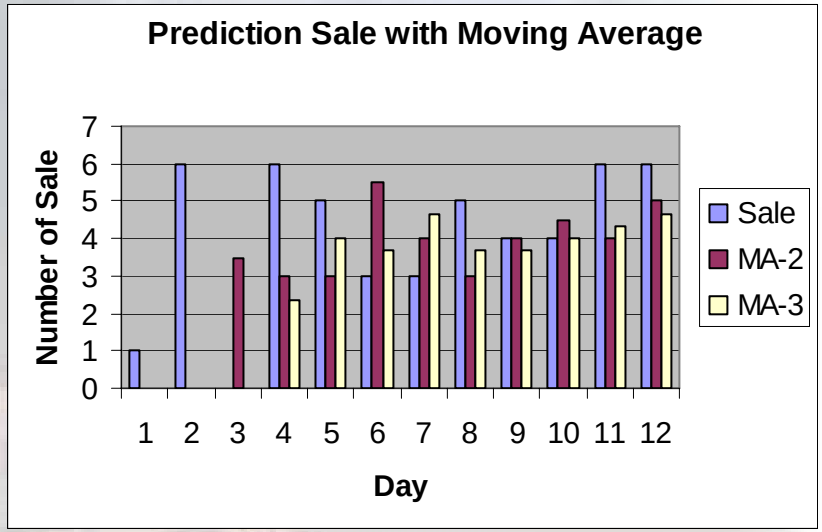
$P(C|X) = Max \{ P(C=Yes|X), P(C=No|X) \}$

$= 0.250$

→ Mr. Bean Like Coffee

Moving Average

Day	Number of Sale	MA-2		MA-3	
1	1				
2	6				
3	0	(1+6)/2	3.5		
4	6	(6+0)/2	3	(1+6+0)/3	2.3
5	5	(0+6)/2	3	(6+0+6)/3	4.0
6	3	(6+5)/2	5.5	(0+6+5)/3	3.7
7	3	(5+3)/2	4	(6+5+3)/3	4.7
8	5	(3+3)/2	3	(5+3+3)/3	3.7
9	4	(3+5)/2	4	(3+3+4)/3	3.7
10	4	(5+4)/2	4.5	(3+5+4)/3	4.0
11	6	(4+4)/2	4	(5+4+4)/3	4.3
12	6	(4+6)/2	5	(4+4+6)/3	4.7



MA with n periodical-times:

$$x_k = \frac{1}{n}\sum_{i=k-n}^{k-1} x_i$$

$$x_k = \frac{1}{n}\left(x_{k-1} + x_{k-2} + \ldots + x_{k-n}\right)$$

Linear Regression

Day	Number of Sale
1	1
2	6
3	0
4	6
5	5
6	3
7	3
8	5
9	4
10	4
11	6
12	6

Linear Regression → $y = ax + b$

Using Least-Square Method we find :

$a = 0.2552$ and $b = 2.4242$

Day	Number of Sale	Regression
1	1	2.6794
2	6	2.9346
3	0	3.1898
4	6	3.445
5	5	3.7002
6	3	3.9554
7	3	4.2106
8	5	4.4658
9	4	4.721
10	4	4.9762
11	6	5.2314
12	6	5.4866