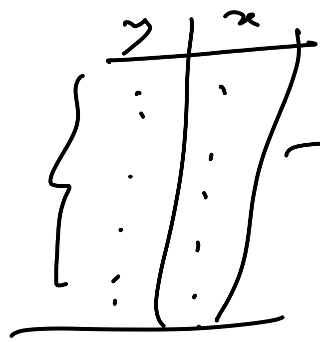
// Regression

1 DV. 1 IV → Simple Reg.

1 DV >1 IV → Multiple "

Ex) SP of the house depending on size of the house.

$$\left.\begin{aligned} SP &= y \\ Size &= x \end{aligned}\right\} \text{Simple Reg.}$$



Actual

$$O = M + E$$

O = observed value (y)

M = Model which based on x.

$= f(x)$.

Sample

Sample.

Error.

$$\text{Error} = O - M. \qquad \text{Let } O = y, \; M = \hat{y}$$

$$= (y - \hat{y}).$$

When there are only one IV (x)

$$\therefore \hat{y} = \hat{a} + \hat{b}x. \qquad \hat{a} = \text{Known}, \; \hat{b} = \text{"}$$

$\hat{a}$ & $\hat{b}$ estimated by OLS.

(ordinary least square method)

$$\sum E^2 = \sum (y - \hat{y})^2 \qquad E \left< \begin{matrix} +ve \\ -ve \end{matrix} \right\}.$$

$$\therefore SSE = \sum (y - \hat{y})^2$$

$y > \hat{y}$ E +ve.

$y < \hat{y}$ E -ve.

E is +ve.

E^2 is +ve.

$$S^2 = \sum (y - \hat{y})^2$$

$$= \sum (y - a - bx)^2.$$

Two parameters a & b. so

which are called indep. Const.

$$\frac{\partial S^2}{\partial a} = 2 \sum (y - a - bx)^{2-1} (-1).$$

$$= (-2) \sum (y - a - bx).$$

$$\frac{\partial S^2}{\partial b} = 2 \sum (y - a - bx)^{2-1} (-x)$$

$$= (-2) \sum (xy - ax - bx^2).$$

N.C. $\quad \frac{\partial S^2}{\partial a} = 0 \quad \frac{\partial S^2}{\partial b} = 0$

$$\left.\begin{aligned} \sum y &= na + b \sum x \quad (1) \\ \sum xy &= a \sum x + b \sum x^2 \quad (2) \end{aligned}\right\} \text{normal equations}$$

From (1) $\quad \frac{\sum y}{n} = a + \frac{b \sum x}{n} \qquad \bar{y} = a + b\bar{x}.$

$$\therefore a = (\bar{y} - b\bar{x}).$$

Putting the value of a in eqn (2), we get

$$\sum xy = (\bar{y} - b\bar{x}) \sum x + b \sum x^2$$

$$\sum xy = \bar{y} \sum x - b \bar{x} \sum x + b \sum x^2.$$

$$b\left(\sum x^2 - \bar{x}\sum x\right) = \sum xy - \bar{y}\sum x$$

$$\therefore b = \frac{\frac{1}{n}\sum xy - \bar{y}\frac{\sum x}{n}}{\frac{1}{n}\sum x^2 - \bar{x}\frac{\sum x}{n}} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}$$

$$\boxed{\mathrm{Cov}(x,y) = r\sigma_x\sigma_y}$$

$$= \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$

$$y = a + bx$$

$$= (\bar{y} - b\bar{x}) + bx$$

$$= \bar{y} + b(x - \bar{x})$$

$$\boxed{y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})} \Rightarrow$$ Reg. eqn of y on x (when y depends on x)

$$\boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}} \longrightarrow (*)$$

Reg. Coeff of y on x

Slly. Reg. eqn of x on y

$$x = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$\boxed{b_{xy} = r\frac{\sigma_x}{\sigma_y}} \rightarrow (**) \Rightarrow$$ Reg. Coeff of x on y

$(x) \quad x\,(xy)$

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}$$

$$r^2 = b_{yx} \times b_{xy}.$$

Coefficient of determination

$$r = \pm\sqrt{b_{yx} \times b_{xy}}.$$

Therefore, Corr. Coeff is the G.M. of two reg. coeff.

{G.M. of two a, b, c ... G.P. $b^2 = ac$, $b = \pm\sqrt{ac}$ }

Case (1) If b_{yx} & b_{xy} are same in sign. (both are +ve) r must be +ve.

Case (2) If b_{yx} & b_{xy} are -ve sign, r is -ve.

Case (3) If b_{yx} & b_{xy} are opp. sign.

x	y

$2x + 3y = 6$ (y on x)

$4x - 9y = 12$ (x on y)

$3y = 6 - 2x$

$y = 2 - \frac{2}{3}x.$ $\quad \bar{y} = 2 - \frac{2}{3}\bar{x}.$

$y - \bar{y} = \left(-\frac{2}{3}\right)(x - \bar{x})$ $\quad b_{yx} = -2/3$ ✓

$4x = 9y + 12$

$$x = \frac{9}{4} y + 3$$

$$\bar{x} = \frac{9}{4} \bar{y} + 3.$$

$$x - \bar{x} = \frac{9}{4}(y - \bar{y}).$$

$$b_{xy} = \frac{9}{4}$$

Models

Exploratory.

NO higher interaction among Ivs.

Predictive

some higher interaction among ~~Bad~~ IVs.

$$CS = f(BS, PE, G).$$

MR

Dv.

$x_1 \quad x_2 \quad x_3$

$$C.S = \beta_0 + \beta_1 \times B.S + \beta_2 \times PE + \beta_3 G$$

Explorations

Math

$$+ \beta_4 \times BS \times PE + \beta_5 BS \times G + \beta_6 PE G + \beta_7 BS \times PE \times G$$

$y = a + bx$

$y =$ (no relation with x) 2

$=$ relation with x.

2^2

$y = a + bx$ $p =$ no of var.

$y = 0 + bx$ Total no of relation

$= 2^p$.

$(a+b)^2$

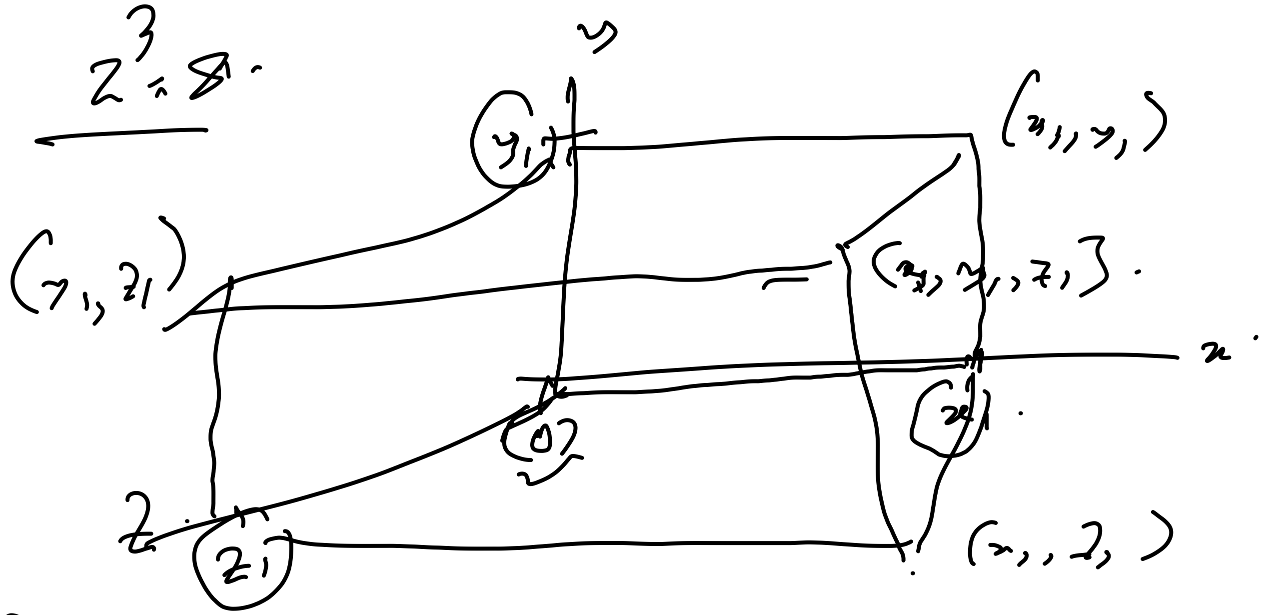
$(\text{no relation} + \text{relation}) = (1+1)^1$

$= 2^1$

$(\text{Su} + \text{Fail})^p$.

$(a+b)^x$.

3 $2^3 = 8$.



Pairwise

$x_1 x_2$, $x_1 x_3$, $x_2 x_3$

$(x_1 x_2 x_3)$

$$\frac{(x_1 + x_2 + x_3)}{\left(({}^3C_0) + ({}^3C_1) + ({}^3C_2) + ({}^3C_3) \right)}$$

$$= (1 + 3 + 3 + 1) = 8$$

$$(1+1)^3 = 2^3 = 8.$$

p. In $\left(\underbrace{{}^pC_0 + ({}^pC_1) + ({}^pC_2)}_{\text{Stat}} + ({}^pC_3) + \dots + ({}^pC_p) \right)$

$p = 20 \Rightarrow 2^p = 2^{20} = 10,48,576$

(Mad)

Min^m no of interaction

$$({}^pC_0) + ({}^pC_1) + ({}^pC_2)$$

$$= 1 + p + \frac{p(p-1)}{2} = \frac{p^2 + p + 2}{2}$$

$p = 20$ $\quad \frac{20^2 + 20 + 2}{2} = 211$ Feature Selection

Flow Chart:

Supervised Model

→ Goal Sec.

Obj

Annotation

SPSS

Unsupervised

PCA.

FA.

Employee data

Reg Lin.

Anova.

M

Reg + T

B.

1. 2 3 4 5.

5. 15 40 20 20 100.

2 4 6 8 10

C_1 C_2 Home

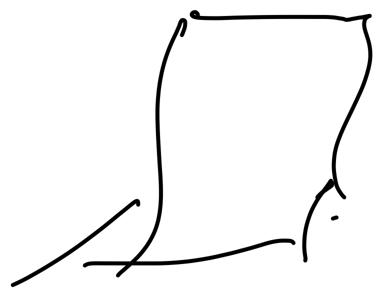
10

HP

$\bar{x}_1 = \bar{x}_2$

$\sigma_1 = \sigma_2$

Skewness



mean = median = Mode

+ve

HD. Mean > median > Mode

-ve

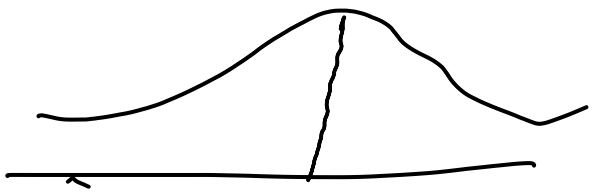
Mean < median < Mode

+ve

$\bar{x} = 100$ $\sigma = 5$

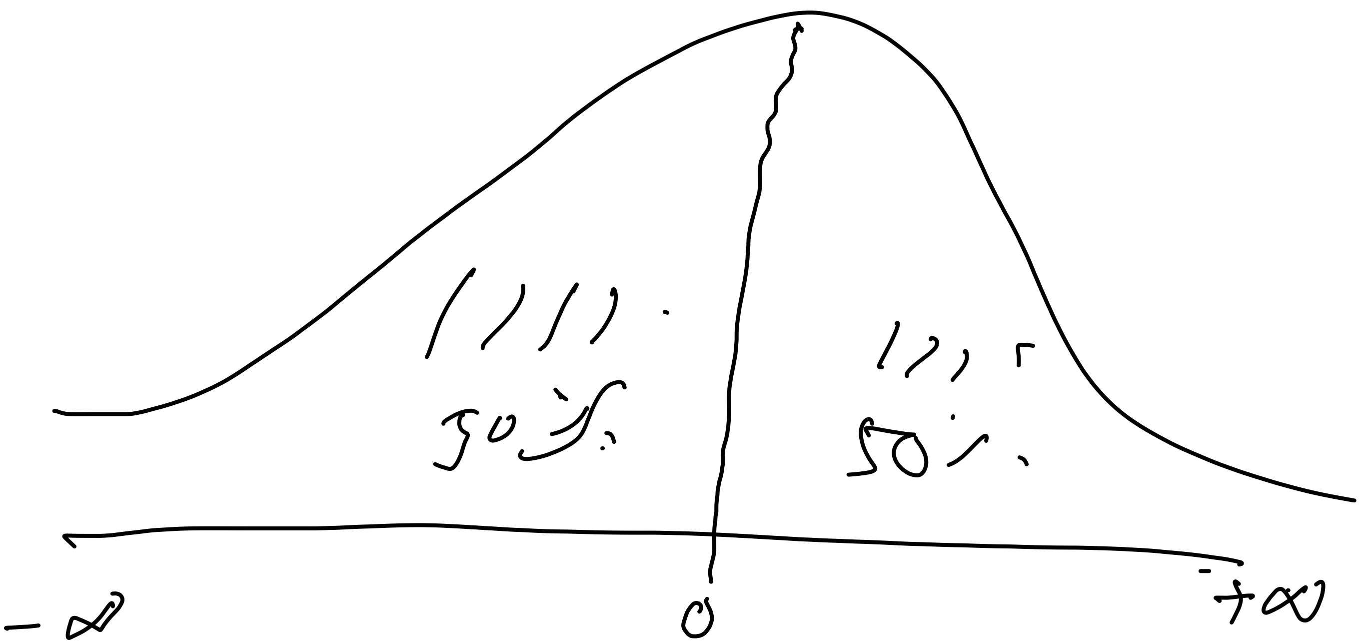
Data

1. Continuous.

2. 

Kurtosis.

3. Symmetric. $SK = 0$



100

$e_{ij} \sim ND(0, \sigma^2)$

Bivariate :

Two variable -

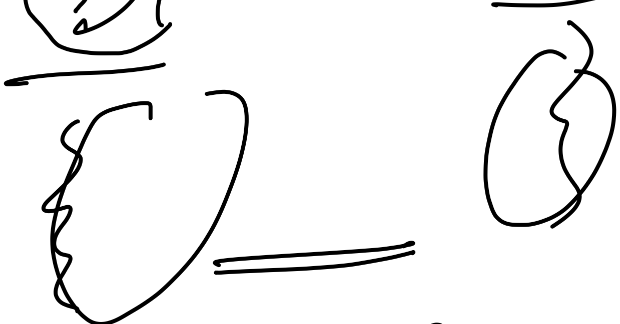
(1) Dr. Ir

1) Selling Price of the house depending on Size of the house (or) density

2] Demand of Temp Avg.

3) Sales dep.

SP. Size on



Q1 Nature of relation (linear). relation / associate

Q2 Degree of

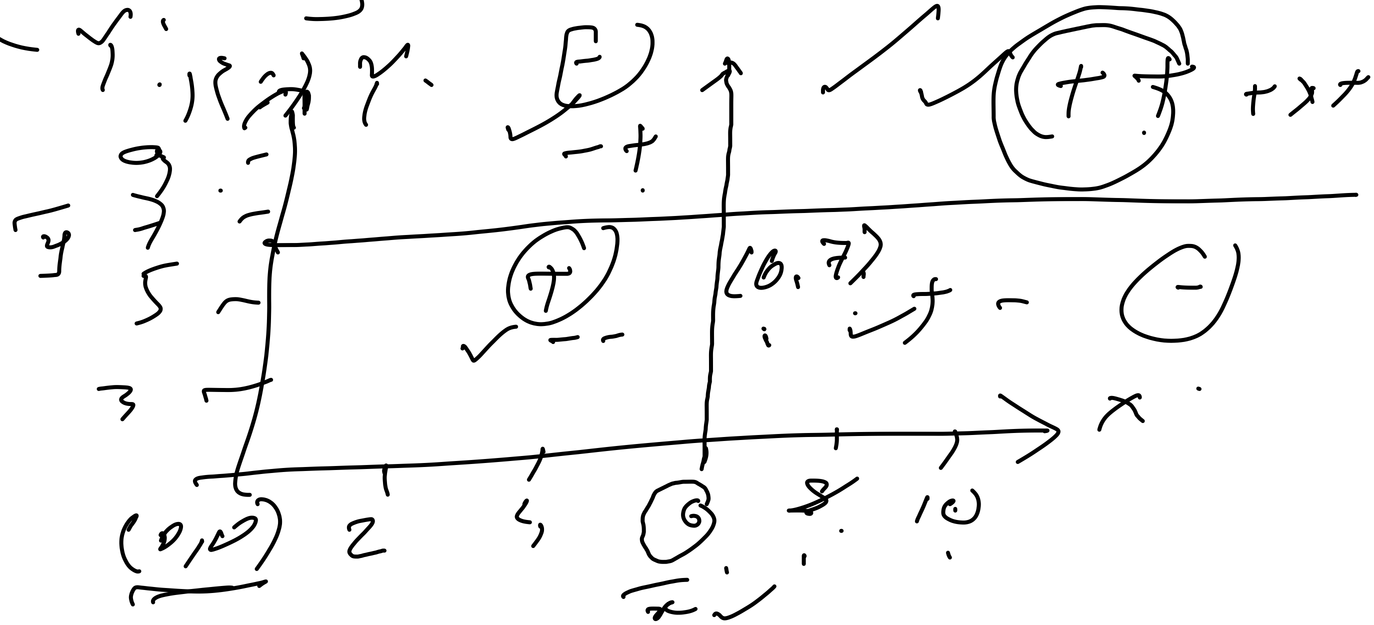
Good. Average. Poor.

$y = x + 1$

x	y
2	3
4	5
6	7
8	9
10	11

x: 2, 4, 6, 8, 10

y: 3, 5, 7, 9, 11



$(x_i - \bar{x}) < 0$ $(y_i - \bar{y}) > 0$

$(x_i - \bar{x}) > 0$ $(y_i - \bar{y}) > 0$

X Y   Place

$$\frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})$$

= Covariance (x, y)

= $\overline{\text{Cov}(x y)}$

$\underline{x}$ $\underline{y}$ $\underline{\text{Cov}(x y)} = 1000$

$= -1000$

$= 0$

$x\uparrow \; y\uparrow \;|\; x\downarrow \; y\downarrow$

$x\uparrow \; y\downarrow \;|\; x\downarrow \; y\uparrow$

$-\infty < \text{Cov}(x y) < \infty$

Height (x)   Weight (y)

Cm   Kg.

Cov(x y)

Cov. Kg.

1. Longer
2. Degree of association
3. Dimension

σ_x

$$= \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

σ_y

$$= \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2}$$

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

Cm   kg

kg

1. $-1 \leq r \leq +1$
2. Unit free.
3. Degree of association

Negative.

PNC.   HNC   MNC.

Positive.

PPC   HPC   MPC.

ZC.

$-1 \quad -0.5 \quad 0 \quad 0.5 \quad +1$

$r = -1 \Rightarrow$ PNC.

$-1 < r < -0.5 \Rightarrow$ HNC.

$-0.5 \le r < 0 \Rightarrow$ MNC.

$r = 0 \Rightarrow$ No linear

$0 < r < 0.5 \Rightarrow$ MPC

$0.5 \le r < 1 \Rightarrow$ HPC.

$r = +1 \Rightarrow$ PPC.

Corr. $r = 0.8$ HPC

y x SP. Size

$r = +1$.

$r = 0.8$

0.2.

Sum of Squares

Error.

Probability

$x_1 \quad x_2 \quad \longrightarrow \quad x_1 \quad x_2 \quad x_3$

$$\mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_2, x_1)$$

$$\begin{array}{c|cc} & x_1 & x_2 \\ \hline x_1 & v(x_1) & c_{12} \\ x_2 & c_{21} & v(x_2) \end{array} \quad 2\times 2$$

$$\begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline x_1 & v(x_1) & c_{12} & c_{13} \\ x_2 & c_{21} & v(x_2) & c_{23} \\ x_3 & c_{31} & c_{32} & v(x_3) \end{array} \quad 3\times 3$$

$$\begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}$$

$$r_{11} = \frac{c_{11}}{\sigma_1 \sigma_1} = \frac{\sigma_1^2}{\sigma_1^2} = 1$$

$$r_{12} = \frac{c_{12}}{\sigma_1 \sigma_2}$$

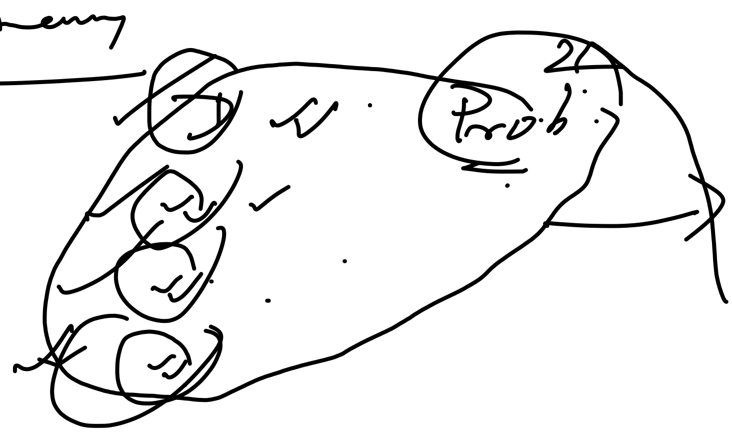
$$r_{22} = \frac{c_{22}}{\sigma_2 \sigma_2} = 1$$

$x_1 \quad x_2$

Covariance & Corr always occurs Pair wise

Covari $= \binom{2}{2} = 1$

How many



Combination → Permutation

A B C D. 2 Cinema tickets

AB AC AD BA CA DA

BC BD CD CB DB DC

Any 2

$$\binom{4}{2} = \frac{4!}{2!\,(4-2)!} = 6$$

3 Q variable

$$\binom{3}{2} = 3$$

3

x_1 x_2 x_3 x_4.

$\binom{4}{2} = 6$ $\binom{5}{2} = 10$ $\binom{p}{2}$

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

$\binom{4}{2}$

x_1 x_2 x_3 x_4.

x_1 x_2, x_3, x_4 } 3 x_2 x_3, x_2 x_4 } 2 x_3 x_4 } 1

$3 + 2 + 1 = 6$

$r = 0.80$

$r^2 = 0.64$

Dv	Iv	
1	1	simple
1	>1	multiple
>1	2	partial
		Canonical

$r = 0.80$

$r = -0.80$

$r^2 = 0.64$

$r^2 \rightarrow 1$

$1 - r^2 = 0.36$

Five-Point Summary

min

Q1

Q2

Q3

max

Box-plot

Prob

P -

$P(G_1)$
$= \frac{5000}{10000}$

Page 6

	No of companies	No of deaths
G_1 < 30	5000	5
G_2 30 - 50	3000	15
G_3 > 50	2000	40
	$n = 10000$	$n = 60$

$P(D/G_1) = 1/2$
$= 5/5000$

$P(D/G_2)$

$P(G_2)$
$= \frac{3000}{10000}$

$$P(D) = \frac{60}{10000} = 0.006 = \frac{15}{3000}$$

$P(G_3)$

$P(D/G_3)$
$= \frac{40}{2000}$

Prob.

Unconditional

Conditional

Forecasting

Past information

Regression

Time Series

Probability

Prior. Posterior.

Known from Past information } Estimate forecast on the basis of Prior

Bayes'n Th

① i	② $P(G_i)$	③ $P(D/G_i)$	④ = ②×③	⑤
1	1/2	Known K_1	$\frac{1}{2} \times K_1 = K_4$	$K_4/0.006$
2	3/10	K_2	$3/10 \times K_2 = K_5$	$K_5/0.006$
3	1/5	K_3	$\frac{1}{5} \times K_3 = K_6$	$K_6/0.006$

Prior

$\Sigma 4 = 0.006 = P(D)$

$\Sigma P = 1$