Date Page

Probability & Statistics

Estimate

Estimation — procedure

Estimator — Rule

Data

Primary — First time

Secondary — Existing sources

Statistics

Descriptive — Data into Graphical or

Inferential — Estimation testing or

Date Page

Tabular form. Decision making

Nature of Variable

Qualitative (Attributes labels / names)

Quantitative

Discrete

Continuous

Scales of Measurement:

- Nominal:
  (Qualitative without Ranking) religion, gender etc.
- Ordinal:
  (Qualitative with Ranking) Grades, Military

Date Page

ranks.

- Interval:

(Quantitative and zero presence)

- Ratio:

(Quantitative and zero absence).

True zero : jab kuch bhi na mile.

e.g. 0°C is not absolutely nothing.

Probability & Statistics

Cross Sectional Data:

- Data collected for a specific time of individuals

Time Series Data:

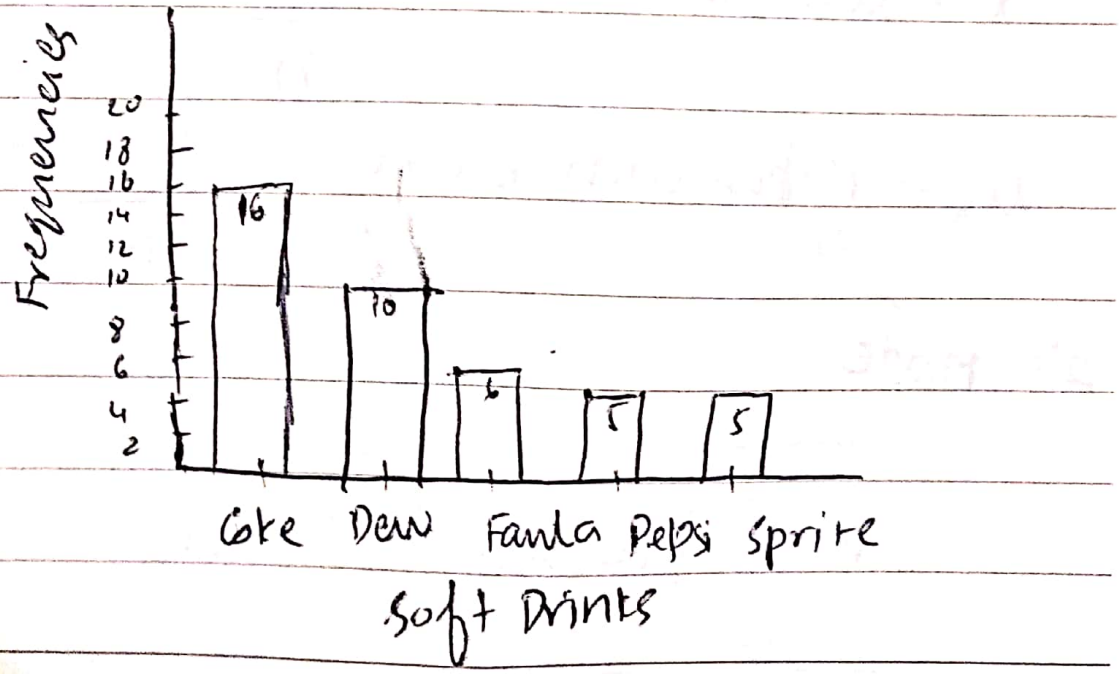
- Data collected over a series of time.

Date Page

$$R.F = \frac{Freq}{Total\ Freq}$$

Soft Drink	Freq (f)	Angle = R.F × 360°
Pepsi	5	5/42 × 360° =
Coke	16	
Sprite	5	
Fanta	6	
Dew	10	
	42	

Bar Chart



Date Page

$\hat{\mu}$ = Population mode

$\hat{x}$ = Sample mode

Measure of Central Tendencies:

1) Mean,

n = sample size

N = Population size

$\bar{x}$ = Sample mean = $\frac{\sum x}{n}$

μ_x = Population mean = $\frac{\sum_{i=1}^{N} x}{N}$

2) Mode

Most frequent value.

Date Page

$\hat{\mu}$ = Pop. mode

$\hat{x}$ = Sample mode.

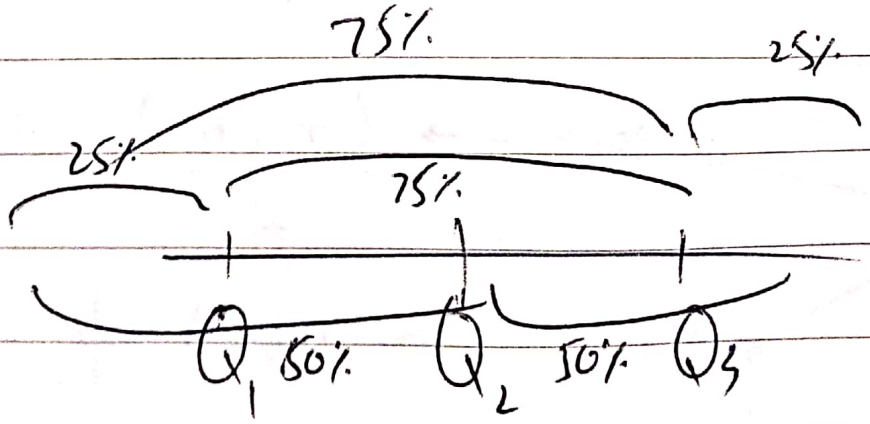
3) Median :

Middle value for sorted data set.

Even : $\dfrac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

Odd : $\left(\dfrac{n+1}{2}\right)^{th}$

4) Quantiles / Percentiles.



lower Quartile | median | upper Quartile

Date Page

P_{10} P_{20} P_{50} P_{60}

D_1 means Decile 10%.

$$i = \frac{P}{100} \times \text{total no. of obs.}$$

$$P_{25} = Q_1 = \frac{25}{100} \times n = 0.25 \times n$$

$$P_{50} = Q_2 = \text{median} = 0.5 \times n$$

$$P_{75} = Q_3 = 0.75 \times n$$

$$P_{80} = 0.80 \times n$$

i :
- integer $\longrightarrow \frac{i + (i+1)}{2}$
- not integer $\longrightarrow$ round-up

Date Page

① ② 3 4 5

3310, 3355, 3450, 3480, 3480,

6 7 8 9 10 11

3490, 3520, 3540, 3550, 3650, 3730

12

3925

$n = 12$

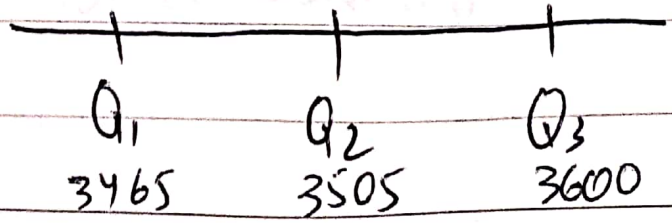
$P_{25} = 0.25 \times 12 = 3 \Rightarrow \frac{3^{rd} + 4^{th}}{2}$

$= 3465$

$P_{50} = 0.50 \times 12 = 6 \Rightarrow \frac{6^{th} + 7^{th}}{2} = 3505$

$P_{75} = 0.75 \times 12 = 9 \Rightarrow \frac{9^{th} + 10^{th}}{2} = 3600$

$P_{80} = 0.80 \times 12 = 9.6 \Rightarrow 10^{th} = 3650$



Date Page

Simple Mean = 3540

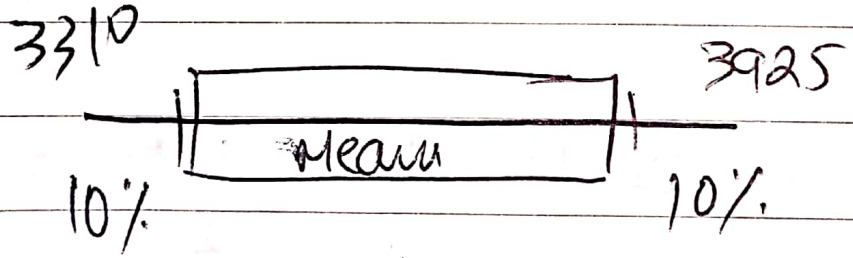
5) Trimmed Mean:

Total trim = 20%

means 10% trim from one side (20% total).

trim ~~means~~ = 10%

means 10% from one side.



trimmed by 10%

Probability & Statistics

Measures of Dispasion

When Average is same we can not analyze which data is better.

1) Range $= R = X_m - X_.$

Date Page

When there are outliars.

2) InterQuartile Range

~~or~~ $IQR = Q_3 - Q_1$

upper - lower quartile

3) Variance ~~or~~ Standard Deviation

Population : σ^2

Sample : s^2 or $\hat{s}^2$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$ (Small population)

$$\hat{s}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$ (close to population)

Date Page

Average of squared deviation from the mean is variation.

Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$$

$$s = \sqrt{s^2}$$

$$S = \sqrt{S^2}$$

:- For small population we divide by n-1 because we have to eliminate biasness from the data set because we didn't take data from the other people

:- Small Sample ($n < 30$)

Date Page

How S.d is related to mean?

For this:

Coefficient of variation:

$$CV = \frac{Sd}{Mean} \times 100$$

↑ lower is better

Prob. & Stats

a) 68, 70, 71, 71, 71, 72, 73, 73, 73
73, 74, 74, 74, 74, 74, 75, 75, 75, 76
77

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

$$= \frac{73+74}{2} = 73.5$$

Date Page

b) Max = 77

Min = 68

Median = 73.5

$Q_1 = \frac{25}{100} \times 20 = 5$

$\Rightarrow \frac{5^{th} + 6^{th}}{2} = \frac{71+72}{2} = 71.5$

$Q_3 = \frac{75}{100} \times 20 = 15 \Rightarrow \frac{15^{th} + 16^{th}}{2}$

$= \frac{74+75}{2} = 74.5$

c) For outliers we calculate IQR

$IQR = Q_3 - Q_1 = 74.5 - 71.5$

$= 3$

Outliers are not there.

Date Page

d) ~~AT&T~~

Range

AT&T : R = 75 - 66 = 9

Sprint : R = 69 - 63 = 6

T-mobile: R = 77 - 68 = 9

Verizon : R = 81 - 75 = 6

Sorting others

AT&T

66, 66, 68, 68, 68, 68, 69, 69, 70, 71, 71, 72, 72, 72, 73, 73, 73, 74, 75, 75

Sprint

63, 64, 64, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 67, 67, 68, 68, 68, 69, 69.

Date Page

Verizon

75, 76, 76, 77, 77, 77, 77, 78, 78, 78,
79, 79, 79, 79, 79, 80, 80, 81, 81, 81

Q_1

$Q_1 = 0.25 \times 20 = 5 = 5^{th} + 6^{th}/2$

AT&T: ~~$Q_1 = 0.25 \times$~~ $68 + 68/2 = 68$

Sprint: $65 + 65/2 = 65$

T-mobile ~~72~~ $+ 71 + 72/2 = 71.5$

Verizon: $77 + 77/2 = 77$

$Q_3 = 0.75 \times 20 = 15 \Rightarrow 15^{th} + 16^{th}/2$

AT&T: $73 + 73/2 = 73$

Sprint: $67 + 68/2 = 67.5$

T-mobile: $74 + 75/2 = 74.5$

Verizon: $79 + 80/2 = 79.5$

$IQR = Q_3 - Q_1$

AT&T: $73 - 68 = 5$

Sprint: $67.5 - 65 = 2.5$

Date Page

T-mobile: 74.5 - 71.5 = 3
Verizon: 79.5 - 77 = 2.5

Mean $\mu = \Sigma X / n$

AT&T: 1413/20 = 70.65
Sprint: 1322/20 = 66.1
T-mobile: 1463/20 = 73.15
Verizon: 1567/20 = 78.35

Variance ~~Standard Deviation~~
$s^2 = \Sigma (X_i - \bar{X})^2 / n-1$

AT&T: 148.55/19 = 7.81
Sprint: 53.8/19 = 2.831
T-mobile: 85.66/19 = 4.508
Verizon: 58.55/19 = 3.081

Date Page

Stand. Deviation

$$S = \sqrt{S^2}$$

ATET: $\sqrt{7.81} = 2.794$

T-mobile: $\sqrt{4.508} = 2.123$

Sprint: $\sqrt{2.831} = 1.682$

Verizon: $\sqrt{3.081} = 1.755$

$$CV = \frac{s.d}{mean} \times 100$$

ATET: $2.794/70.65 \times 100 = 3.95\%$

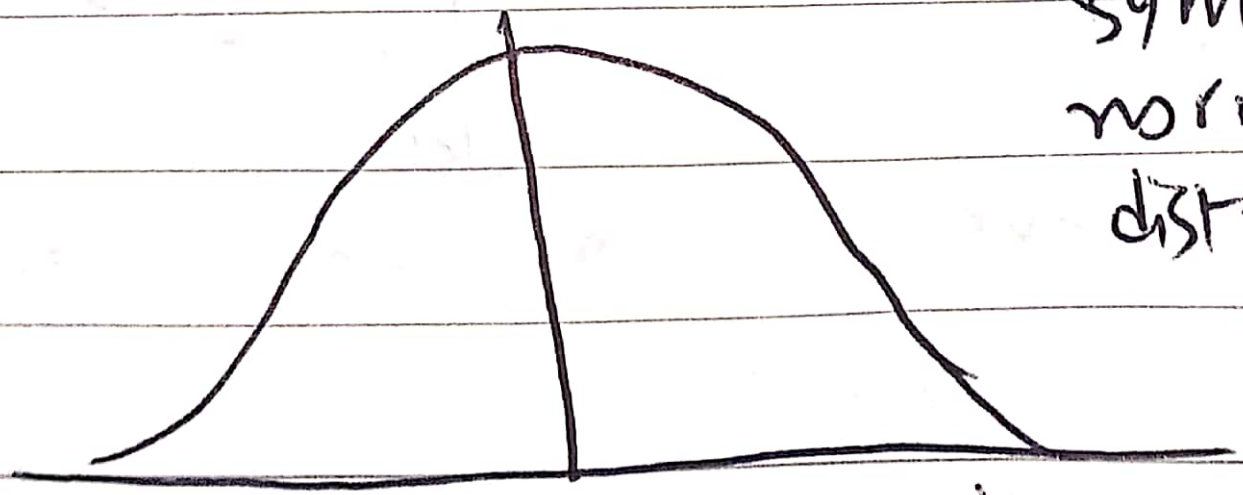
Sprint: $1.682/66.1 \times 100 = 2.54\%$

~~Ver~~ T-mobile: $2.123/73.15 \times 100 = 2.90\%$

Verizon: $1.755/78.35 \times 100 = 2.239\%$

Prob & Stats

Ideal Scenario



Symmetric normal distribution

Mean = mode = median

Date Page

- Everything apart from normality is skewed.
- Skewness can be positive or negative
- Mean > med > mode
  => +vely / right skewed
- Mean < med < mode
  => -vely / Left skewed

- To achieve normality we need transformation

How much the data is skewed?

↓

Pearson's Coeff. of skewness

-3 highly -ve | 0 (1/2) moderate | +3 highly +ve

Date Page

$$S_k = \frac{3(\text{Mean} - \text{mode})}{S.d}$$

↓

amount of skewness

$$\text{T-mobile} = S_k = \frac{3(73.15 - 73.5)}{2.15}$$

$= -0.4884$ (slightly skewed to the left)

Box Whisker Plot

Tmobile : Five point Summary

X_0	Q_1	med	Q_3	X_m
min	lower Quartile		upper Quartile	Max
68	71.5	73.5	74.5	77

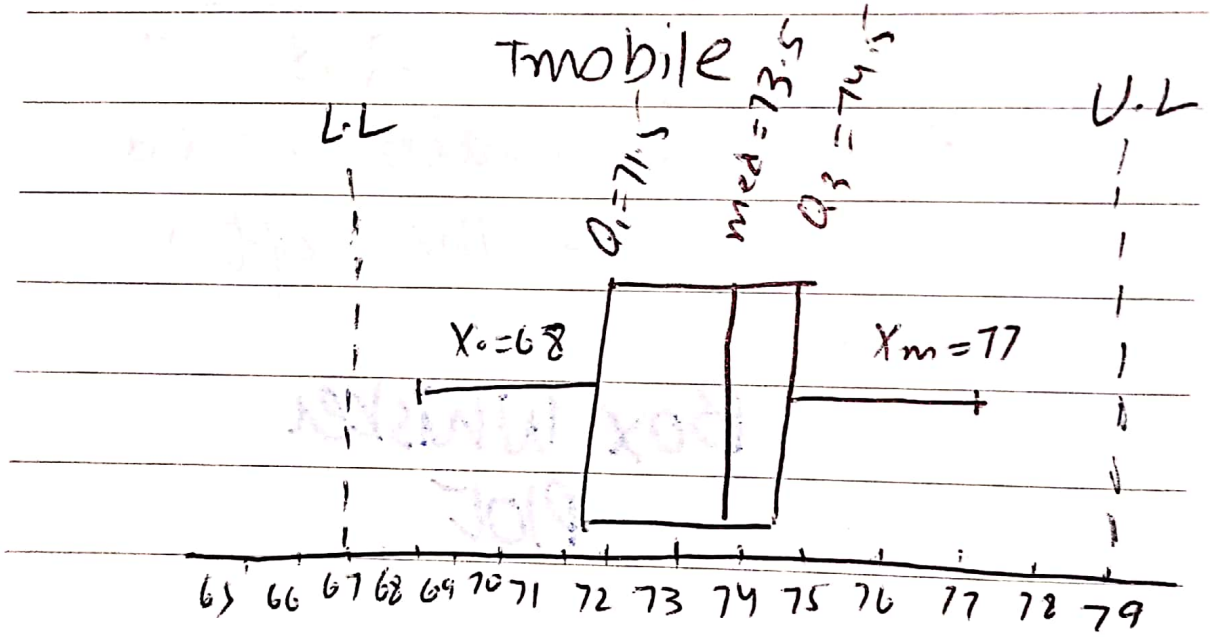
Date Page

$U.L = Q_3 + 1.5(IQR)$

$U.L = 74.5 + 1.5(3) = 79$

$L.L = Q_1 - 1.5(IQR)$

$L.L = 71.5 - 1.5(3) = 67$



(Ch3 41, 63, 20 David)

Date Page

20 Dawson: 11 10 9 10
11 11 10 11
10 10

Clark: 8 10 13 7
10 11 10 7
15 12

Sorting data:

Dawson: 9, 10, 10, 10, 10, 10,
11, 11, 11, 11

Clark: 7, 7, 8, 10, 10, 10, 11, 12, 13
15

Range (Dawson) = 11 - 9 = 2

Range (Clark) = 15 - 7 = 8

Date Page

Five number Summary (Dawson)

X_0	Q_1	Med	Q_3	X_m
9	0.25×10	$\frac{(\frac{n}{2})^{th} + (\frac{n}{2}+1)^{th}}{2}$	0.75×10	
	$= 2.5$		$= 7.5$	11
	$= 3^{th}$	$= \frac{5^{th} + 6^{th}}{2}$	$= 8^{th}$	
	$= 10$		$= 11$	
		$= \frac{10+10}{2}$		
		$= 10$		

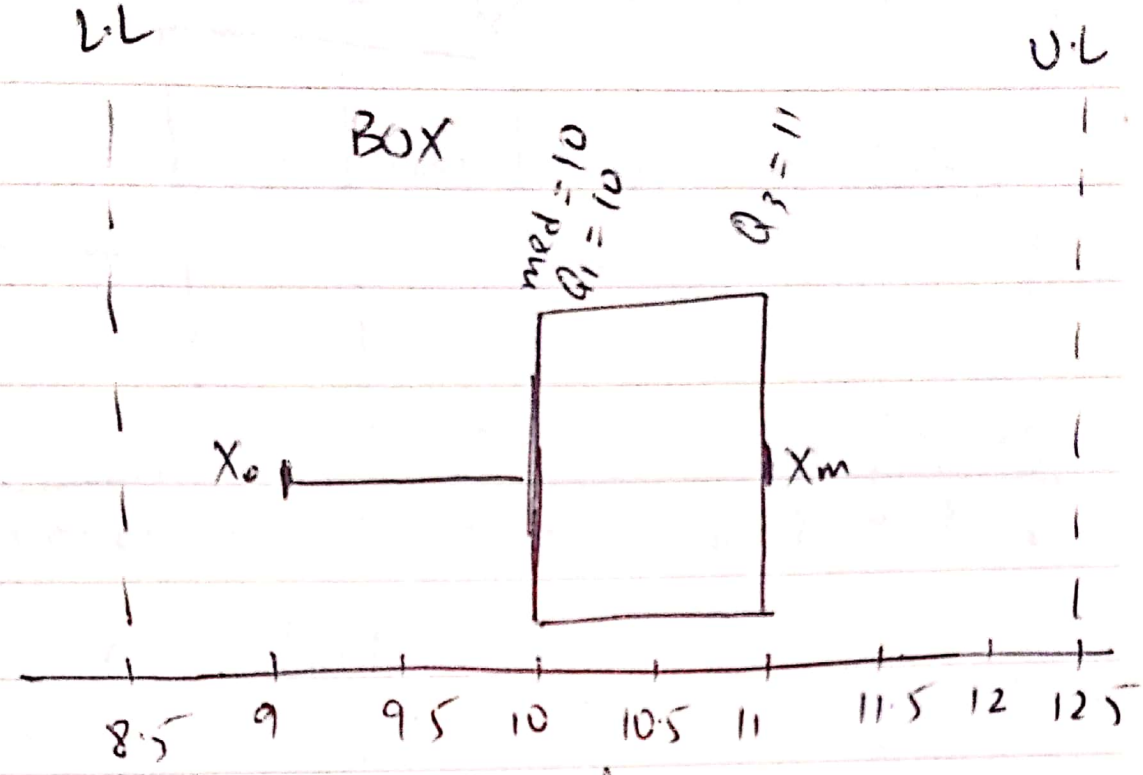
$IQR = Q_3 - Q_1 = 11 - 10 = 1$

$U.L = Q_3 + 1.5(IQR)$

$= 11 + 1.5(1) = 12.5$

$L.L = Q_1 - 1.5(IQR)$

$= 10 - 1.5(1) = 8.5$



Rightly Skewed

(Clark)

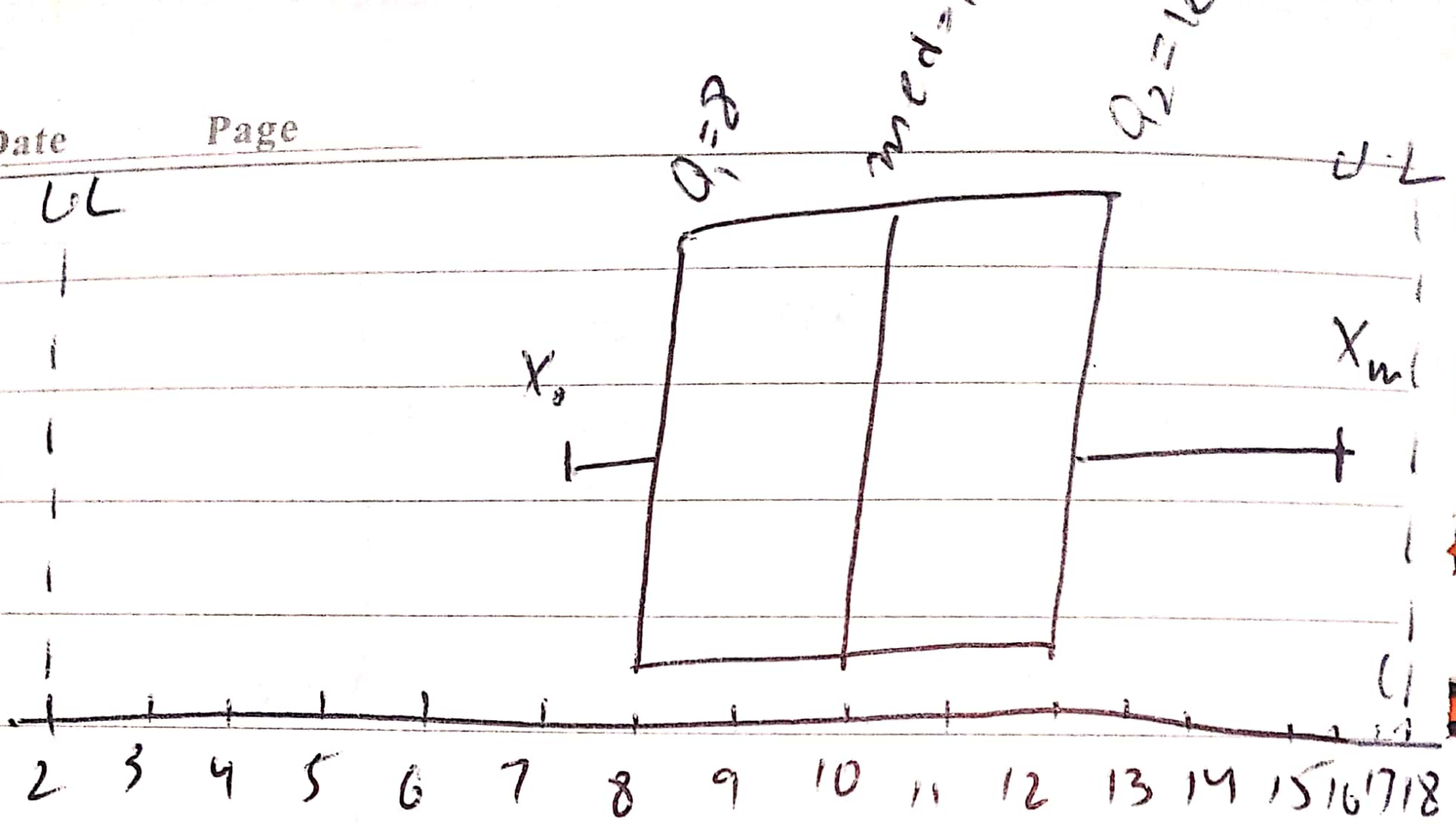
X_o	Q_1	Med	Q_3	X_m
7	0.25×10	$\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$	0.75×10	15
	$= 2.5$		$= 7.5$	
	$= 3^{th}$	$= \frac{5^{th} + 6^{th}}{2}$	$= 8^{th}$	
	$= 8$		$= 12$	
		$= \frac{10+10}{2}$		
		$= 10$		

$IQR = Q_3 - Q_1 = 12 - 8 = 4$

$U.L = Q_3 + 1.5(IQR) = 12 + 1.5(4) = 18$

$L.L = Q_1 - 1.5(IQR) = 8 - 1.5(4) = 2$

Date Page



Normal Distribution

Date Page

6/09/23 (cont.)

Prob & Stats

Year & Audit time in days

12	14	19	18	
15	15	18	17	
20	27	22	23	n=20
22	21	33	28	
14	18	16	13	

	Classes	Tally	Freq(F)	RF	AF / PF	C.F / CRF	CRF	C.P.F	C.B
①	10-14	1111	4	0.2	20	4	4/10=0.2	20	9.5-14.5
②	15-19	1111 1111	8	0.4	40	4+8=12	0.6	60	14.5-19.5
③	20-24	1111	5	0.25	25	17	0.85	85	19.5-24.5
④	25-29	11	2	0.1	10	19	0.95	95	24.5-29.5
⑤	30-35	1	1	0.05	5	20	1	100	29.5-34.5
	Σ	-	ΣF=20 =n	1	100				

-0.5 +0.5

2.2

Date Page

No. of Classes = $1 + 3.3 \log(\text{tot. obv})$

Relative Freq = R.F
$= f_i / \Sigma f$

Percent Freq = P.F
$= R.F \times 100$

Commutative Freq = C.F

$\longleftarrow$ P =

	x_i	$x_i f_i$	$f_i(x_i - \bar{x})^2$
①	12	48	
②	17	136	
③	22	110	
④	27	54	
⑤	32	32	
		Σ	Σ

" R.F = C.R.F
=

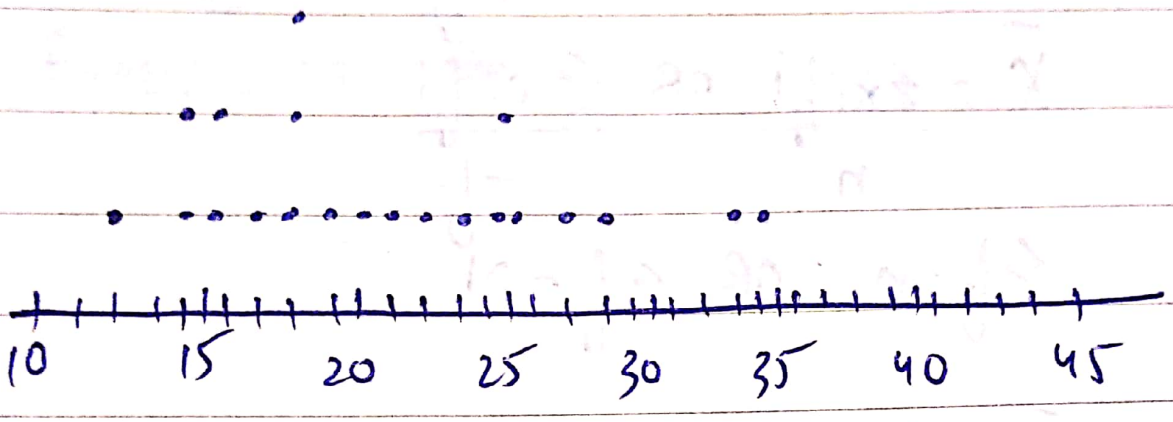
" P.F = C.P.F
=

Class Width = $\frac{\text{Max} - \text{Min}}{\text{No. of Classes}} = \frac{33 - 12}{5} = 4.2$

x_i = Mid points

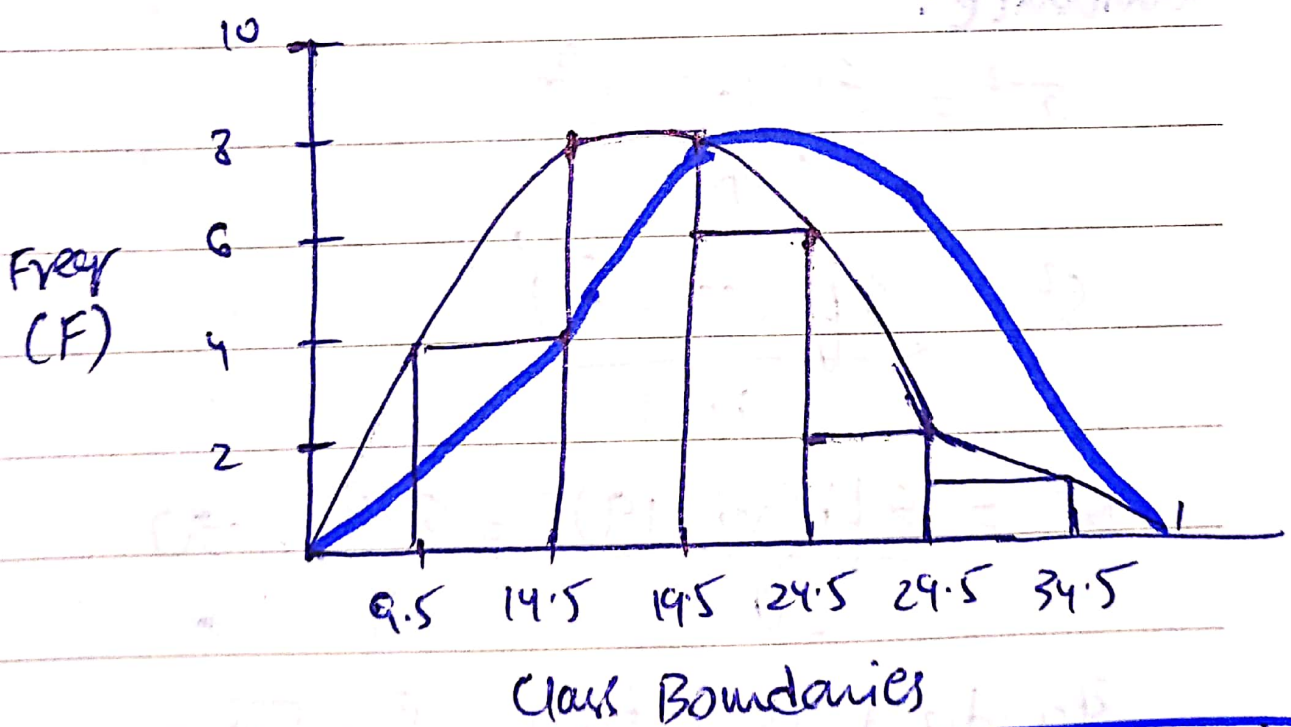
Date Page

Dot Plot



Freq → Histogram

CF →



The data is tvely skewed

- Tea and audit
- Normal curve

Date Page

Mean:

$$\mu = \frac{\Sigma xi fi}{N} \quad OR \quad \frac{\Sigma xi fi}{\Sigma f}$$

$$\bar{x} = \frac{\Sigma xi fi}{n} \quad OR \quad \frac{\Sigma xi fi}{\Sigma f}$$

$$\Sigma f = n \quad OR \quad \Sigma f = N$$

$$\bar{x} = \frac{\Sigma xi fi}{n} = \frac{380}{20} = 19 \text{ days}$$

Variance:

$$\sigma^2 = \frac{\Sigma fi (xi - \mu)^2}{N}$$

$$s^2 = \frac{\Sigma fi (xi - \bar{x})^2}{n-1}$$

$$s^2 = \frac{\Sigma fi (xi - 19)^2}{20-1} = \frac{570}{19} = 30$$

Standard Deviation $= s = \sqrt{30}$

$= 5.48$

Prob & Stats

Probability Basics

Experiment
Trial / Outcome
Render Experiment
Sample Space
Events
Simple / Compound
The Two Events
Mentally Exc / Non-M Exc
Equally likely Events

Counting Principle

→ Rule of Multiplication
→ " " combination
→ " " Permutation

Date Page

Assigning Probabilities:

→ Classical Approach
→ Subjective Approach
→ R.F Approach

→ Repeating Experiments
↓
Monte Carlo Runs.

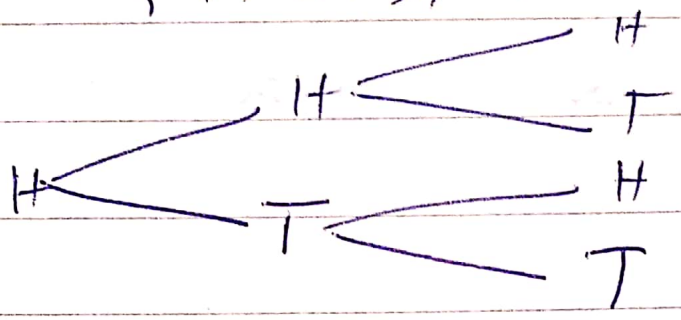
Mutually Exclusive → At once only one category is present and other is negated.

Equally likely → Probability of each event is same.

Date Page

Sample space for a dice

{1, 2, 3, 4, 5, 6}



Rule of Multiplication:

one dice & one coin

sample space = 6 × 2 = 12

26 26 26 10 10 10 10

$26^3 \times 10^4$

Date Page

order imp → Permutation

order not imp → Combination

$n = 6$

A B C D E F

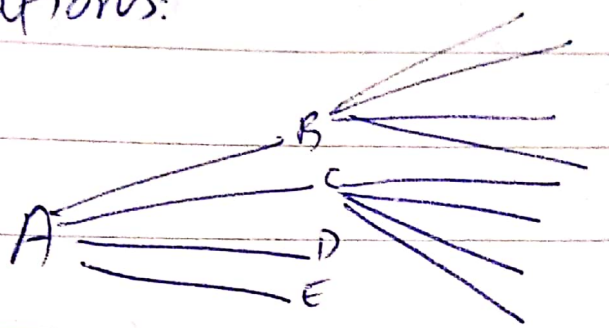
$r = 3$

${}^{6}P_3 = 120$

${}^{6}C_3 = 20$

Date Page

Permutations:



B

C

120

A	B	C	D
A	B	C	D
A	B	C	E
A	B	C	F
A	B	D	E
A	B	D	F
A	B	E	F
A	C	D	E
A	C	D	F
A	C	E	F
A	D	E	F
A	B	C	D	E
---	---	---	---	---
A	B	C	D	F
A	B	C	E	F
A	B	D	E	F
A	C	D	E	F
B	C	D	E	F

Date Page

C D E D E F

C D F

C E F

$n(S) = 6$

$S = \{1, 2, 3, 4, 5, 6\}$

$E = \{2, 4, 6\}$ $n(E) = 3$

$P(E) = 3/6 = 1/2$

Success

$1 - 1/2 = 1/2$ (Failure)

Classical is for → Equally likely Events

(chapter 2 Ward p.01
Henderson Pg 158 $Q_1 \rightarrow Q_6$
Henderson Pg 169 Q_{22})

Prob & Stats

Law of Complementation

$n = 15$

$x = 0, 1, 2, 3, \ldots\ldots, 15$

Success + Failure = Total

$\underbrace{0, 1, 2}_{\text{Failure}} \quad \underbrace{3 \text{----} 15}_{\text{Success}}$

$P(S) + P(F) = 1$

$P(S) = 1 - (P(F))$

$= 1 - [P(x=0) + P(x=1) + P(x=2)]$

$P(F) = 1 - (P(S))$

$A^c = S - A$

Date Page

$S = \{E_1, E_2, E_3, E_4, E_5\}$

$n(S) = 5$

$A = \{E_1, E_2\}$ $B = \{E_3, E_4\}$

$\Rightarrow C = \{E_2, E_3, E_5\}$

$n(A) = 2$ $n(B) = 2$ $n(C) = 3$

$P(A) = \frac{n(A)}{n(S)} = \frac{2}{5}$ $P(B) = \frac{2}{5}$

$P(C) = 3/5$

A and B => Yes, A & B are M.Ex.

-- **Law of Addition**

P (A or B)

Date Page

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

N·M·Ex

$$P(A \cup B) = P(A) + P(B) \quad \text{M·Ex}$$

$$P(A \cup B^c) = P(A) + P(B^c) - P(A \cap B^c)$$

$$P(A \cup B) = P(A) + P(B) = 2/5 + 2/5 = 4/5$$

$A \cup B = \{E_1, E_2, E_3, E_4\}$

$n(A \cup B) = 4$

$P(A \cup B) = 4/5$

$A^c = S - A = \{E_3, E_4, E_5\}$

$C^c = \{E_1, E_4\}$

$P(A^c) = 1 - P(A) = 1 - 2/5 = 3/5$

$P(C^c) = 1 - P(C) = 1 - 3/5 = 2/5$

$$P(B \cup C) = P(B) + P(C) - P(B \cap C)$$

$$= 2/5 + 3/5 - 1/5 = 4/5$$

Date Page

Q24 → ~~tt~~

F ~ fell short of exp
R ~ surpassed exp
M ~ Met exp
D ~ Didn't respond

$P(F) = 0.26$
$P(M) = 0.65$
$P(D) = 4\% = 0.04$

a) $P(R) = 1 - [P(F) + P(M) + P(D)]$
$1 - 0.95 = 0.05$

b) $P(M \text{ or } R) = P(M) + P(R)$
$= 0.65 + 0.05$
$= 0.70$

Date Page

Law of Multiplication

① $P(A \cap B) = P(A)\ P(B|A)$

② or $P(A \cap B) = P(A|B)\ P(B)$

① A and B independent

② A and B dependent

given

$P(\text{Event} \mid \text{cond}) \neq P(\text{Event})$

① $P(A \cap B) = P(A)\ P(B)$

$$P(A \cap B) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Stats 2 \ Stats 1	Y	N	
Y	□	□	$n(Y_2)$
N	□	□	$n(N_2)$
	$n(Y_1)$	$n(N_1)$	$n(S)$

Date Page

Q-35

Pay Rent

		Yes (Y_R)	No (N_R)		
Buy a Car	Yes (Y_B)	56 0.28	52	0.26	108 0.54
	No (N_B)	14 0.07	78 0.39		92 0.46
		70 0.35	130 0.65		200 1

b) $P(Y_B) = 0.54 > P(Y_R) = 0.35$

c)

$$P(Y_R \mid Y_B)$$

$$= \frac{P(Y_R \cap Y_B) = 0.28}{P(Y_B) \quad 0.54}$$

Date Page

d) $P(Y_R | N_B) = \frac{P(Y_R \cap N_B)}{P(N_B)}$

$= \frac{0.07}{0.46}$

$0.52 \neq 0.35$ not ind.

e) $P(Y_R | Y_B) = P(Y_R)$

OR

$P(Y_B | Y_R) = P(Y_B)$

f) $P(Y_B) + P(Y_R) - P(Y_B \cap Y_R)$

$= 0.54 + 0.35 - 0.28$

Q38, 0

Prob & Stats

Q-38 Walpole (4.4)

$M = 200$ $\quad MS = 40$

$W = 200$

$B = 280$

$MS = 80$

Date Page

	B	S	
M	120 (0.3)	80 (0.2)	200 (0.5)
W	160 (0.4)	40 (0.1)	200 (0.5)
	280 (0.7)	120 (0.3)	400 (1)

a) $P(B) = 280/400 = 0.7$

b) $P(S) = 120/400 = 0.3$

c) $P(M|S) = \frac{P(M \cap S)}{P(S)}$

$= \frac{80/400}{0.3}$

$= 0.66$

$P(W|S) = \frac{P(W \cap S)}{P(S)}$

$= \frac{40/400}{0.3}$

$= 0.33$

Date Page

d) $P(M \cap S) = 80/400 = 0.2$

$P(W \cap S) = 40/400 = 0.1$

e) $P(M \cap S) = 80/400) = 0.2$

b) $P(W \cap S) = 40/400 = 0.1$

g) For Independent

$$P(M|S) = P(M)$$

By part (c)

$0.66 \neq 0.5$

$\therefore$ Dependant

$$P(W|S) = P(W)$$

By part (c)

$0.33 \neq 0.5$

$\therefore$ Dependant

Date Page

Prob & Stats

BAYES' THEOREM

$$P(A \cap B) = P(A)\,(P(B|A))$$

$$P(A_i|B) = \frac{P(A_i)P(B/A_i)}{P(A_1)\,P(B/A_1) + \cdots + P(A_n)P(B|A_n)}$$

→ where the email would go spam, inbox, there will be some condition

→ We can classify using Bayes theorem.

$$P(A_i \gamma B) = \frac{P(A_i \cap B)}{\sum_{i=1} P(A_i \cap B)}$$

Date Page

(1) Events	(2) Prior	(3) Cond.	(4) Joint	(5) Posterior

Q40

Events	Prior	Cond	Joint	Posterior
A_i	$P(A_i)$	$P(A_i\|B)$		
A_1	.20	.50	0.1	0.256
A_2	.50	.40	0.2	0.512
A_3	.30	.30	0.09	0.23
			0.39	

Example 2.43

Events	Prior	Cond	Joint	Posterior
P_i	$P(P_i)$	$P(D\|P_i)$		
P1	0.3	0.01	0.003	0.157
P2	0.2	0.03	0.006	0.315
P3	0.5	0.02	0.01	0.526
			0.019	

Date Page

E_1 = Cardholder will default

$P(E_1) = 0.05$

~~E_2 = missing monthly payement~~

~~$P(E_2) = 0.20$~~

E_2 = Card holder not default

$P(E_2) = 1 - 0.05 = 0.95$

M = missing a payement

$P(E_2|M) = 0.2$

$P(E_1|M) = 1$

Events	Prior	Cond	Joint	Posterior
E_1	0.05	1	0.05	0.208
E_2	0.95	0.2	0.19	0.791
			0.24	

A) ~~0.791~~ 0.208

Prob & Stats

W ~ msg contain word

S ~ Spam

H ~ Ham

Let p be success as spam

& failure of spam (& ham)

Date Page

$$P(S|W) = \frac{P(W|S)\,(P(S))}{P(W|S)\,P(S) + P(W|H)\,P(H)}$$

$$= \frac{P(Word)}{P(Word) + q(Word)}$$

$$= \frac{P(Word1)\,P(Word2)}{P(Word1)\,P(Word2) + q(Word1)\,q(Word2)}$$

Without prior knowledge we assum equally likely events $= 0.5$

$$\therefore P(H|D) = \frac{P(H \cap D)}{P(D)}$$

$\because$ it ~~inverse~~ is reverse prob. than $P(D) \longrightarrow$ total prob.

Date Page

→ 1000 emails
700 spam
300 Not Spam

H_1 = Spam = 0.7
H_2 = Ham = 0.3

D = email has word 'cash'

$P(D|H_1) = 350/700 = 0.5$

$P(D|H_2) = 100/300 = 0.33$

$$P(H_1|D) = \frac{P(D|H_1)\,P(H_1)}{P(D|H_1)(P(H_1)+P(D|H_2)P(H_2)}$$

$$= \frac{(0.5)\,0.7}{(0.5)(0.7)+(0.33)(0.3)}$$

$$= \frac{0.35}{0.449} < 0.78$$

$$P(D|H_2) = \frac{0.33 \times 0.3}{0.449} = \frac{0.1}{0.45}$$

$$= 0.23$$

Example

word 'rolex' occurs in 250/2000 messages known to be spam

and 5/1000 messages to be ham which has the word rolex.

incoming message has rolex is spam assuming that it is equally likely that incoming is spam and not spam if our threshold for rejecting a message as spam is 0.9, will we reject?

Date Page

H_1 = spam

$P(H_1) = 0.5$

H_2 = ham

$P(H_2) = 0.5$

D = rolex

$P(D|H_1) = 250/2000 = 0.125$

$P(D|H_2) = 5/1000 = 0.005$

$$P(H_1|D) = \frac{P(D|H_1)\,P(H_1)}{P(D|H_1)\,P(H_1) + P(D|H_2)\,P(H_2)}$$

$$= \frac{(0.125)(0.5)}{(0.125)(0.5) + (0.005)(0.5)}$$

$$= \frac{0.0625}{0.065}$$

$$= 0.961$$

We will reject

Date Page

$H_1 = 2000$

$H_2 = 1000$

$P(H_1) = 2000/3000 = 0.66$ $P(H_2) = 1000/3000 = 0.33$

stalk ↓ undervalued ↓

$P(S|H_1) = \frac{400}{2000} = 0.2$ $P(U|H_1) = \frac{200}{2000} = 0.1$

$P(S|H_2) = \frac{60}{1000} = 0.06$ $P(U|H_2) = \frac{25}{1000} = 0.025$

Both words 'is spam' assuming we have no prior knowledge - 0.4

$$= \frac{P(S|H_1)\,P(U|H_1)}{P(S|H_1)P(U|H_1)}$$

$$= P(H_1|S) =$$

$$= \frac{[P(S|H_1)\,P(H_1)]\,[P(U|H_1)\,P(H_1)]}{P(S|H_1)\,P(H_1)\,(P(U|H_1)\,P(H_1) + P(S|H_2)\,(P(H_2)\,P(U|H_2)}$$

Date Page

$$= \frac{(0.1)(0.66)(0.2)(0.66)}{(0.2)(0.66)(0.1)(0.66) + (0.06)(0.33)(0.025)(0.33)}$$

$$= \frac{0.008712}{0.0105 \quad 0.00887}$$

$$= 0.829$$

$$= 0.98 > 0.9$$

We will reject