Statistics for DS I

Week 1

Types of ~~data~~

Descriptive statistics: description & summarization of data

Inferential stats: draw conclusions from data
- use possibility of chance (probability)

Data –
1. categorical
2. Numerical
   - Discrete
   - continuous

Data collection requires one of the following scale of measurement: nominal, ordinal, interval, or ratio.

1. Nominal: labels, names

eg: Name, Board, Gender, Blood group etc.
- sometimes numerically coded
- no ordering

2. ordinal scale

- properties of nominal data
- order or rank

eg: customer ratings

- Dataset
  - Cat
    - Nominal
    - Ordinal
  - Num
    - Interval
    - Ratio

3. Interval scale

- all properties of ordinal data
- Interval b/n values (in fixed units)
- Always numeric (no absolute zero)
- eg: temperature

4. Ratio scale

- All prop of Interval data
- Ratio of 2 vals are meaningful

eg: height, weight, age, marks etc.
numerical vals added, subtracted, multiplied, divided

Week 2

Describing categorical data - frequency (count) distributions & relative frequency (ratio)

Graphical display

charts of categorical data
- bar chart, Pie chart - relative frequency

Pie chart
- proportions of category

Bar chart

Pareto charts : categories in bar chart are sorted by frequency good for ordinal ~~data~~ variables

---

Question 2 5 subjects, total 500 Marks

Physics	35%	175
chemistry	25%	125
Biology	10%	50
Maths	18%	90
Hindi	12%	60

$$\frac{125 + 50 + 90}{500}$$

$= 53\%$

Total Players 200

Academy	No of players
A	a
B	b
C	150
D	d
E	75

$a+b+d+125=200$

$a+b+d=75$

$\frac{a+b+d}{200} = \frac{75}{200}$

mode: bimodal multimodal

median: not available unless the data can be put into order

total 80 students

$32.5 \times .80$

Grades	A	25	20
	B	32.5	26
	C	22.5% .	18
	D	20% .	~~16~~

$\frac{18}{20}$ $\frac{22.5}{25.0}$

14 18 20 93 40

Week 3 I measures of central tendency

Sample mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

n – Sample size

N – Population size

Population mean $\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$

Quiz

	nos	freq
(1)	2	$x+6$
	6	$x+2$
	11	$x-3$
	14	x

mean = 5.63

$$\frac{2(x+6)+6(x+2)+11(x-3)+14x}{x+6+x+2+x-3+x} = 5.63$$

$$\frac{2x+12+6x+12+11x-33+14x}{4x+5} = 5.63$$

$$\frac{33x-9}{4x+5} = 5.63$$

$33x-9 = 5.63(4x+5)$

$= 22.52x + 28.15$

$33x - 22.52x = 28.15 + 9$

$10.48x = 37.15 \quad x = 3.54$

$= 4$

Ans 4

(1) Mean adding constant to each value

new mean = mean + c

(2) Mean multiplied by constant

new mean = mean * c

* medium & mode

1. If no of obs odd

medium $\frac{n+1}{2}$ elem

2. If even

~~mean~~ mean of $\frac{n}{2}$ & $\frac{n}{2}+1$ obs.

(1) Medium

- data Inc order

Sample mean is sensitive to outliers whereas sample medium is not

* adding constant

New medium = old medium + c

* Multiplying a Constant

new medium = old medium * c

(2) Mode - most freq

- If no val occurs more than once, no mode

(1) adding const

new mode = old mode + c

(2) multiply

new mode = old mode * c

(6) measure of dispersion

Range, variance, std dev.

Range = Max − Min

Range is sensitive to outliers

Variance

Population variance $\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2}{N}$

Sample variance $s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}$

* (1) adding a constant

new variance = old variance

(2) multiplying a constant

new variance = c^2 * old variance

Standard deviation

$s = \sqrt{\frac{(x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}}$ pop std = $\sqrt{\frac{(x-x_1)^2 + x_n - x}{N}}$

sample std = $\sqrt{\text{sample variance}}$

(1) adding constant: new std = old std

(2) multiply constant: new std = c * old std

Percentile, Quartiles & Interquartile range

Percentiles

1. Arrange the data in inc order
2. If np is not an integer smallest int > np position of 100p.
3. If np is an int. Avg values in positions $\frac{V(np), V(np+1)}{2}$ in the 100p
4. Outliers $< (Q_1 - 1.5 IQR)$ and $> (Q_3 + 1.5 IQR)$

Question

mean = 16 $\quad \frac{\text{sum}}{10} = 16$

Sum = 160 $\quad \frac{166}{10} =$

Sum + 6 = 166

Sample std = 10 $\quad$ mean = 16

variance = 100 $= \frac{(x_1 - \bar{x})^2 + \ldots}{9}$

$(x_1 - \bar{x})^2 + \ldots = 900$

$= 900 - (10-10)^2$

$+ (6+10)^2 = 900 + 16 = \frac{916}{9}$

(4) 97, 64, 62, 71, 47 mean 68.2

$(28.8)^2 + (4.2)^2 + (-6.2)^2 + (2.8)^2 + (21.2)^2$

$$\frac{829.44 + 17.64 + 38.44 + 7.84 + 449.44}{4}$$

(5) 40 38 41 41 96 99 101

38 40 41 41 96 99 101

n = 7 10th - .1 × 7 = 0.7 = 1 - 38

40 .5 × 7 = 3.5 = 4 → 9

(6) IQR = $Q_3 - Q_1$ (75, 25)

26 23 39 25 77 106 92

23 25 26 39 77 92 106

.25 × 7 = 1.75 = 2 Q_1 = 25

Q_3 = .75 × 7 5.25 = 6 Q_3 92

outlier 1.5 IQR above or below Q_1 & Q_3

$Q_1 - 1.5$ IQR $Q_3 + 1.5$ IQR

$Q_1 - 1.5\, IQR$

$= 25 - 1.5 \times 67$

$Q_3 + 1.5\, IQR = 92 + 1.5 \cdot 67$

(9) 44, 52, 53, 56, 68, 7., 80, 81, 83, 89, 89

90

7. 44. 52 53 56 68 80 81 83 89 89

90

= 79-12.30

$n \times .25 = 3$ 4 $= 52.5$

$\frac{}{2}$

$12 \times .75 = 9$ $= 86$

Covariance -   Measure of association

Dev of x   Dev of y   product

$(x - \bar{x})$ $(y - \bar{y})$ $(x - \bar{x})(y - \bar{y})$

Population covariance

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample covariance

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation

Measure of linear correlation b/n 2 numerical variables

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{Cov(x,y)}{S_x \cdot S_y}$$

$$\frac{Cov(x,y)}{S.std\,x \times S.std\,y}$$

Price = 7.77 x size + 130

$R^2 = 0.022$ $r = 0.149$

goodness of the fit

$$R^2 = 1 - \frac{RSS}{TSS}$$

$y = mx + c$

m = slope

c = Intercept

$0 \le R^2 \le 1$

correlated ~~r~~ x

correlation r

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
355	300	-106.9	-74	11427.61	5476	7910.6
487	340	25.1	-34	630.01	1156	-853.6
526	400	64.1	26	4108.81	676	1666.6
590	450	128.1	76	16409.61	5776	9735.6
428	300	-33.9	-74	1149.21	5476	2508.6
398	325	-63.9	-49	4083.21	2401	3131.1
555	450	93.1	76	8649	5776	7075.6
320	400	-141.9	26	20135.61	676	-3689.4
450	375	~~[illegible]~~ -11.9	1	141.61	1	-11.9
510	400	48.1	26	9023.61	676	1250.6

mean = 461.9 374

76358.29 28690

28723.8

$$\frac{28723.8}{\sqrt{76358.29} \times \sqrt{28690}} = \frac{28723.8}{276.33 \times 167.60}$$

$$= \frac{28723.8}{46312.908} = 0.62021$$

Revision

Week 2

		Time
W1	1+1	2
IND	1+1	2
Am	1+1+1+1+1+1	5
JAK	1	1
SL	1	1
CXG	1	1
		12

150
175
175
150
300
1000

150
1000
175
1000

Total no of dollars

$150 + x + y + 250 - 1300$

$= 700 + x + y$

$\frac{250}{700+x+y} = 0.25$

$250 = 0.25(700+x+y)$

$1000 = 700+x+y$

$x+y = 300$

$y = (x - 300)$

$\frac{x}{700+300} = 0.175$

$x = 0.175(1000)$

$= 175$

$\frac{300}{700+x+y} = 0.3$

$300/000 = 700 + x + y$

$x + (x - 300) = 300$

$2x = 300 ... 700$

$y = 300 - 175$

$= 125$