(Descriptive Statistics)

Types of Data

Categorical
(categories or group)

e.g: Car brands
- BMW, AUDI
- Yes and no Question

Numerical

Discrete
(Finite)
- No. of children
- Grades
- Money

Continuous
(infinite)
- weight
- Time
- Distance
- area

Measurement Levels

Qualitative

Nominal
- Car brands
- Weather (spring, summer, winter)

ordinal
- order base data for e.g

Bad food (−)
⋮
good food (+)

Quantitative

Interval
(Doesn't have) True 0
- Temperature (C° or F°)

$\boxed{\begin{array}{l} 0°K = -273.15°C \\ 0°K = -459.67°F \end{array}}$ → common

Ratio
(Has True 0)
- No. of object
- Distance
- Time

Visualization

Types of Data

- Categorical
- Numerical (Discrete and Continuous)

- 1st Categorical Representation
  - Frequency Distribution Tables (row + column)
  - Bar charts ()
  - Pie charts ()
  - Pareto diagram ()

⇓ (80-20 Rule)

* 80% of the effect come from 20% of the causes
* 80% of mistakes can be avoided by fixing 20% of the causes.

(For one Variable)

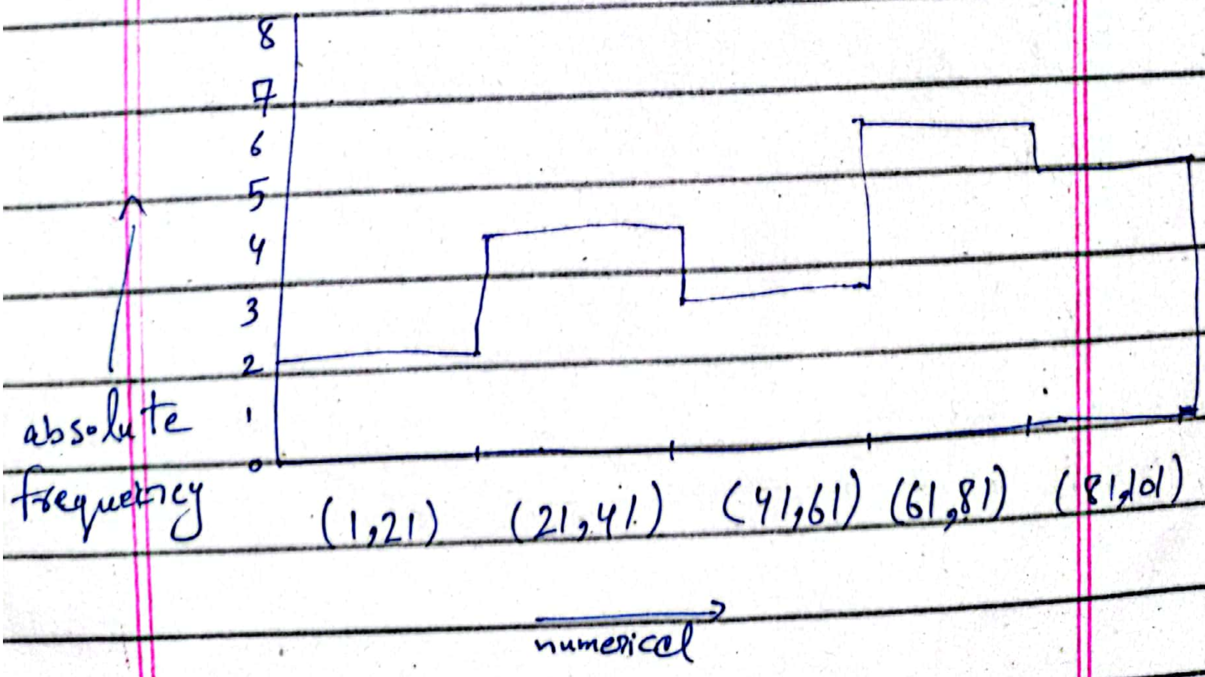
☆ 2nd Numerical Representation

• Frequency distribution Table

Interval start	Interval End	Frequency	(R)
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

$$\text{relative Frequency} = \frac{\text{Frequency}}{\text{Total Frequency}}$$

Histogram

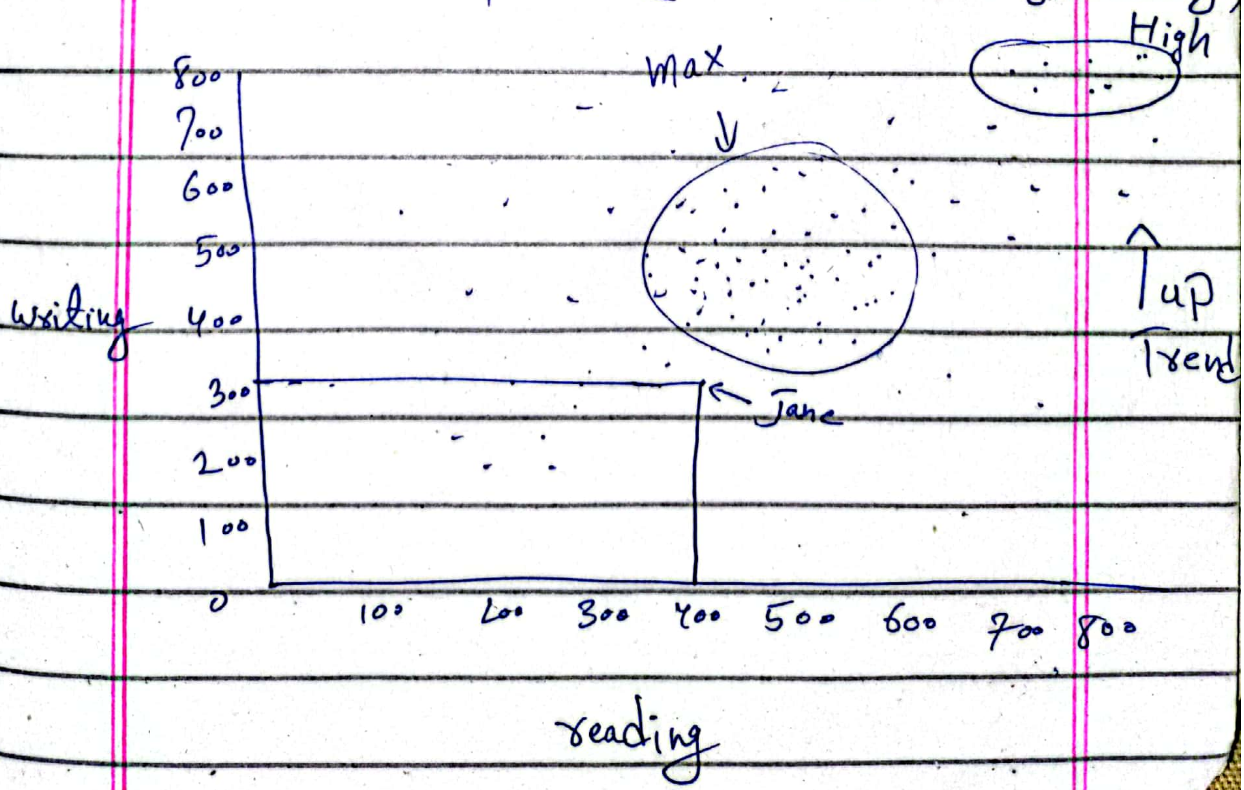


Relationships b/w Two variable

Type of Investor	Investor A	B	C	Total
stocks	46	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
	365	340	210	915

All graphs are very easy to create and read, once you have identified the type of data you are dealing with.

- Scatter Plot (student ID, Reading, Writing)



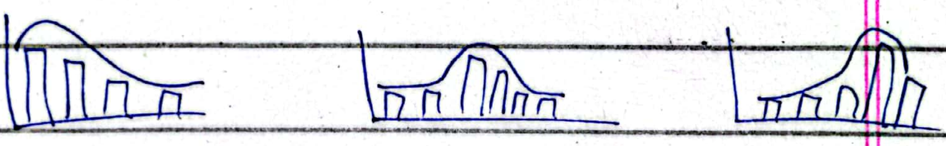
☆ Measures of Central Tendency

- Mean (Avg) (μ) ($x \Rightarrow \bar{x}$) $= \frac{\text{All no.}}{N}$
- Median (center Point of Data)
- Mode (repeated value in Data)

with skewness we calculate which is better measures.

Postive skew, Zero skew, negative skew

Data ⇒



☆ Variance (larg Unit)^2

Population (Mean calculate) (~~[illegible]~~

1
2
3
4
5

$$\sqrt{\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}}$$

For sample variance in dataset → $N-1$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x)^2}{\cancel{N}}}$$

= Standard Dev. ~~Deviation~~

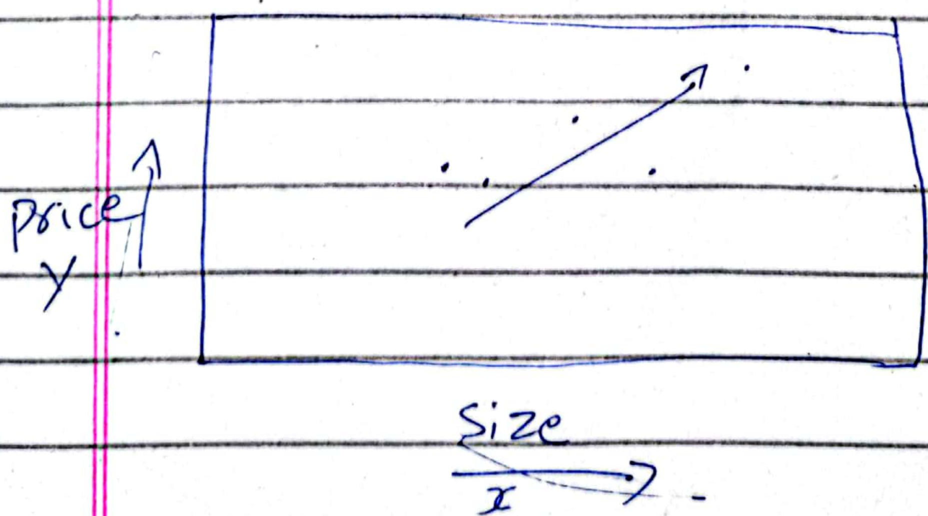
Measures of Relationship b/w variables

Housing Data

	Size(ft)	Price($)
less	650	772,000
price	785	998,000
large	13200	1,200,000
	720	800,000
	975	845,000

Covariance:

The two variables are correlated and the main statistics to measure this correlation is called Covariance. (>0, =0, <0)



Formula Sample:

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

- if $S_{xy} > 0$ (Covariance), the two variable move together
- if $S_{xy} < 0$, the two variable move in opposite directions.

if $S_{xy} = 0$, the two variable are independent.

* Correlation Coefficient

$$= \frac{\text{Covariance}}{std(x) * std(y)}$$

$= 0.87$ (of above housing data)

asymmetric relation

Causality (x causes y is diff from y causes x)
Important to understand the direction of causal relationships.

(Correlation Does not imply Causation)

(Inferential Statistics)

- Probability theory
  - Distributions

☆ · Distribution (Probability Distribution)
- Normal
- ~~B~~ Binomial
- Uniform

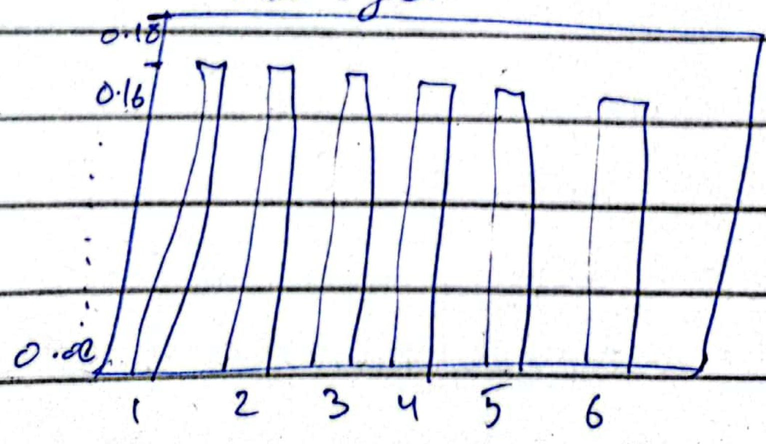
<u>Defination:</u>

A distribution in fn. that shows the possible values for a variable and how often they occur.

e.g:- Rolling a Die (Uniform distr.)

(Discrete) outcome	Probability
1, 2, 3, 4, 5, 6	1/6 = 0.17
impossible → 7	0

visualization



All are same

A Distribution is not a graph itself.
but
the graph is just a visual Representation.

e.g another (Discrete)

Rolling two Dice

6 * 6 = 36 cominotion.

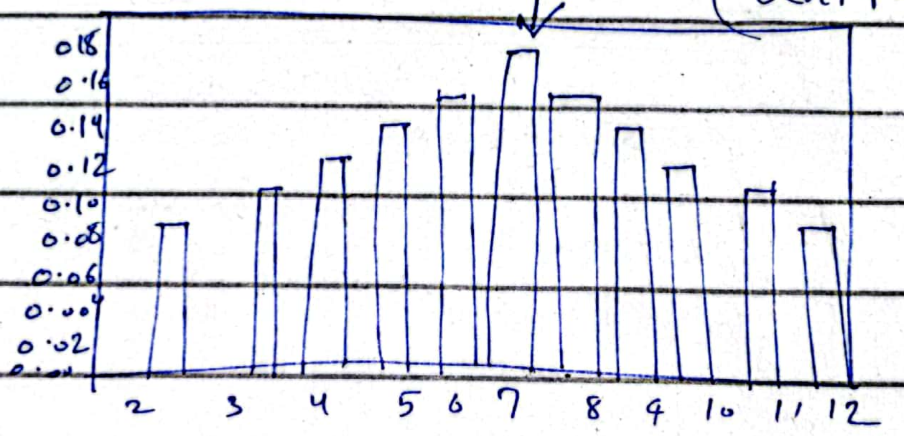
outcome	Probability
2	0.03
3	0.06
4	0.08
5	0.11
6	0.14
7	0.17
8	0.14
9	0.11
10	0.08
11	0.06
12	0.03
All else	0

Rolling 2 Dice Probability of 7 is the highest

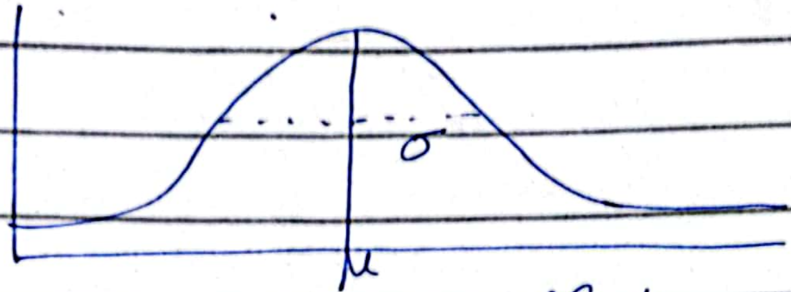
7 = 0.17

visualize

(uniform)



Normal Distribution
(Gaussian distribution)



- Mean = Median = Mode
- It has no skew. all equal

$$N \sim (\mu, \sigma^2)$$

(Normal) (Distribution) (Mean) (variance)

- Standard error decreases when sample size increases

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \downarrow = n \uparrow$$

Scanned with CamScanner

Probability

The likelihood of an event occuring

e.g Flipping a coin

possible two outcom

Heads $\leq$ tails

Events

$\Rightarrow$ High probability

High occurence of event

A $\Rightarrow$ event

P(A) $\Rightarrow$ Probability

$P(A) = \frac{\text{Preferred}}{\text{all}}$ $\Rightarrow$ favorable $\Rightarrow$ sample space

A $\Rightarrow$ Head

$P(A) \Rightarrow \frac{1}{2} = 0.5$

2 $\rightarrow$ possible outcomes

e.g Dice

A = 1

$P(A) = \frac{1}{6} = 0.167$

Scanned with CamScanner

Machine learning

* **Linear Model** :- Data can be classified using linear model is called linear

$$f(x) = x\,w + b$$

input, weight, intercept (bias)

Predicting price of house

Size = 743 ft

$743 \times 336.1 - 3237.51 = \$246,484.29$

1000 " " " " = $\$332,862.4$

Q $y = xw + b$

$x = [2, 3]$

$w = [1.2, -3]$

$b = 7$

$y = x_1 w_1 + x_2 w_2 + b$

$= (2)(1.2) + (3)(-3) + 7$

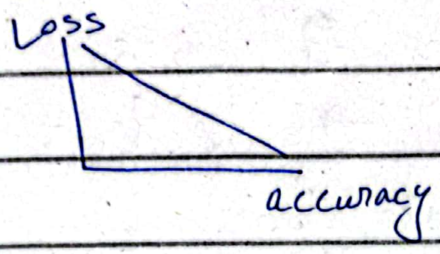
$= 2.4 - 9 + 7$

$= 2.4 - 2$

$= 0.4$

Objective function (Separate block

Loss



if low loss fn. the higher level of accuracy.

Supervised Learning

Reward

In reinforcement learning (AI)

Supervised learning

Regression

l2-norm

Classification

cross-entropy

★ L2-norm = $\sum_i (y_i - t_i)^2$

(Regression) ⇒ (numbers 1.19, 1.21)

The lower the error.

The lower the Loss.

"norm" comes from the fact it is the vector norm, or Euclidean distance of the outputs and their targets.

★ Cross-Entropy

(Classification) ⇒ (categories: cat, Dog)

$$L(y,t) = -\sum_i t_i \ln y_i$$

For eg   0 for no
          1 for Yes

$t = [0, 1, 0] \rightarrow$ category

cat ×, Dog ✓, horse ×

Dog Picture

$y = [0.4, 0.4, 0.2]$

$t = [0, 1, 0]$

→ probably

0.4 means 40% chance to be cat

40% " " Dog

20% " " horse

natural log

$= -(0 \times \ln 0.4) - (1 \times \ln 0.4) - (0 \times \ln 0.2)$

$L(y,t) = 0.92$ ~~Ans~~.

anoter example

Horse Picture

$$
\begin{array}{cccc}
 & \text{Cat} & \text{Dog} & \text{Horse} \\
y = & [0.1, & 0.2, & 0.7] \\
t = & [0, & 0, & 1]
\end{array}
$$

$= -(0 \times \ln 0.1) - (0 \times \ln 0.2) - (1 \times \ln 0.7)$

$L(y,t) = 0.36$ ← better prediction

So the lower the loss fn., more accurate model

=> Better prediction because model $[0.1, 0.2, 0.7]$ 70% probably sure a horse. In other sid $[0.4, 0.4, 0.2]$ 40% probably for cat and dog same.

Any Function that holds the basic property:
Higher for worst results
Lower for Better results
Can be a Loss function.

☆ Optimization Algorithm

- Gradient descent

$f(x) = 5x^2 + 3x - 4$

Goal: Find Min of the Fn.

$f'(x) = 10x + 3$

$x_0 = 4$

$x_1 = ?$

$x_{i+1} = x_i - \alpha f'(x_i)$

$= 4 - \alpha [10 \times 4 + 3]$

$= 4 - \alpha 43$

$X_1 = \cancel{4 - \alpha f'(x_0)}\ 4 - \alpha 43$

$x_2 = x_1 - \alpha f'(x_1)$

$x_3 = x_2 - \alpha f'(x_2)$

Scanned with CamScanner

When minimum is reached

$$x_{i+1} = x_i - \alpha \underbrace{f'(x_i)}_{0}$$

$$x_{i+1} = x_i$$

no longer updates

Question Auto BiKerz

	Predicted: No	Predicted: Yes
Actual: No	TN = 328	FP = 45
Actual: Yes	FN = 65	TP = 562

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{562}{562+45}$$

$= 0.92 = 92\%$

$$\text{recall} = \frac{TP}{TP+FN} = \frac{562}{562+65} = 0.89$$

$= 89\%$

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2$$

$$= \frac{0.92 * 0.89}{0.92 + 0.89} * 2$$

$$= \frac{0.8188}{1.81} \times 2$$

$$= 0.452 * 2 = 0.904$$

$$= 90\%$$

Model: 2

	Predicted: No	Predicted: Yes
Actual: No	TN = 330	FP = 60
Actual: Yes	FN = 50	TP = 560

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{560}{560 + 60} = 0.90$$

$$= 90\%$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{560}{560 + 50} = 0.91$$

$$= 91\%$$

$$F_1 = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} * 2$$

$$= \frac{0.90 * 0.91}{0.90 + 0.91} * 2$$

$$= \frac{0.819}{1.81} * 2 - 0.90\%$$

We should used Model 2 because its Recall higher.
=> In this case of autonomous driving vehicle, it is more important to cature as many of the times a moterbik should "stop" as possible (Recall) than to optimize for accuracy of the times a moterbike does "stop" (Precision).
This is because the consequence of not stopping the bike when it should have are severe (exampl: hitting a human on the street, driving through a stop sign, or hitting another car).