Day 3

18 May 2024 18:16

Central measures

- Mean
  - AM
  - GM
  - HM

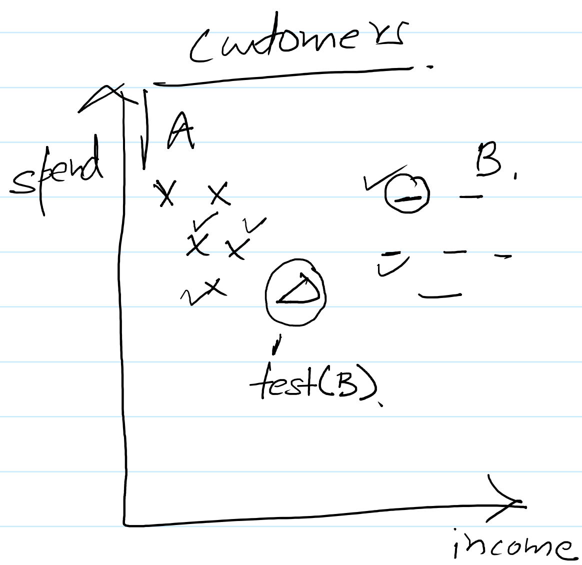
- Median
- Mode

Weighted mean

- Take into account the importance of or weight of each value in the col (series)
- Multiply the weights with the values
- $\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n}{w_1 + w_2 + \ldots + w_n}$

How to determine the weights

- Domain expertise / subjective
- Proportional allocation
  - Based on the value of the data point
- Inverse variance
  - Used in many ML algorithms
    - Example
- Advanced optimization methods
  - Linear programming (OR)
- Data driven methods
  - Empirically



· find out distance of test customer with all the existing customers

income	spend
1K	2k
5k	4k

#	c1	10
	c2	20
	c3	11
	:	15

5 nearest

1K — 12k — c3 — 11
5k — 4k — 15 } 5 heaviest

5
→ 3A ✓ (t = A).
→ 2B.

1.5K — 2.5k. ∴ all 5 hv equal vote/weight.

$$\sqrt{(1.5-1)^2 + (2.5-2)^2} = \text{distance}.$$

euclidean → 8th grade.
powerful.

• preferential weightage.
• distance

$$\frac{1}{\text{distance}} = w$$

Data driven methods
- Example

prediction model ‖ temp. prediction.

test data (prediction) → h, ws, atm, ...

ML1 (NN)	ML2 (SVM)	ML3 Boosting
y ✓	N ✓	y
100 ✓	90	70
(w_1	w_2 ✓	w_3)

(empirically)

33 — 29 — 37.

$$\frac{w_1 * 33 + w_2 * 29 + w_3 * 37}{w_1 + w_2 + w_3}$$

⇒ normalized.

Winsorizing
- Data pre-processing technique
- Capping the extreme values
    - Replace the highest and lowest values
        - With a determined value

Original Income Data:	Winsorized Income Data:
$20,000	$20,000
$30,000	$30,000
$35,000	$35,000
$40,000	$40,000
$1,000,000 (outlier)	$40,000 (capped outlier)
$2,000,000 (outlier)	$40,000 (capped outlier)

What's big with this

Steps
- Identify the % of lower and higher ends (cap)
- Calculate the lower and higher percentile values
- Replace the values
- Calculate the mean

Determining the % of lower and upper ends
- Distribution of data
    - Symmetric, skewed, kurtosis…outliers
    - Sensitivity analysis (outliers)
- Data size
    - Large
        - Higher %

5
15
25
555
777
3

same scale

$\frac{5 - min}{R}$

Range
- MAX - MIN
- Sensitive to outlier
- Limited information u can gather
- Use case
    - Data exploration
    - Cleaning
    - FEATURE SELECTION
    - NORMALIZATION

1 min - max scaler

→ min

→ max

$\left[\frac{x - min}{\underbrace{max - min}_{(R)}}\right]$ [0, 1]

Variance
- Spread or dispersion of data points
- Purpose of calculating
    - Higher variance

- Spread or dispersion of data points
- Purpose of calculating
  - Higher variance
    - Too high - bad
  - Low variance
    - Too low - bad
- Use cases
  - FEATURE SELECTION
  - Model evaluation
  - Making interpretation is hard

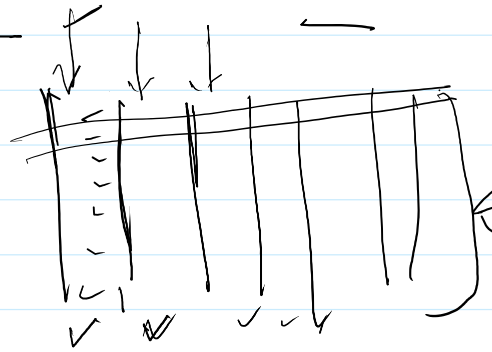
$$\text{Variance} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \text{ for a sample,}$$
$$\text{Variance} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \text{ for a population.}$$

[max - min]

R

X	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
5		-3	4
7		-1	1
9	8	1	1
10		2	4
11,5		3	9

$Var = \cancel{40} = \frac{100}{10} = 10$



Standard deviation
- Square root of var
- Purpose
  - Same unit
    - Customers can easily relate to it

Mean absolute deviation
- Avg abs deviation
- Measure of variability in the data col
- Magnitude of deviation
- Calculate
  - Compute the mean
  - Abs deviation
  - Sum
  - Divide by the number of samples

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \text{mean}|$$

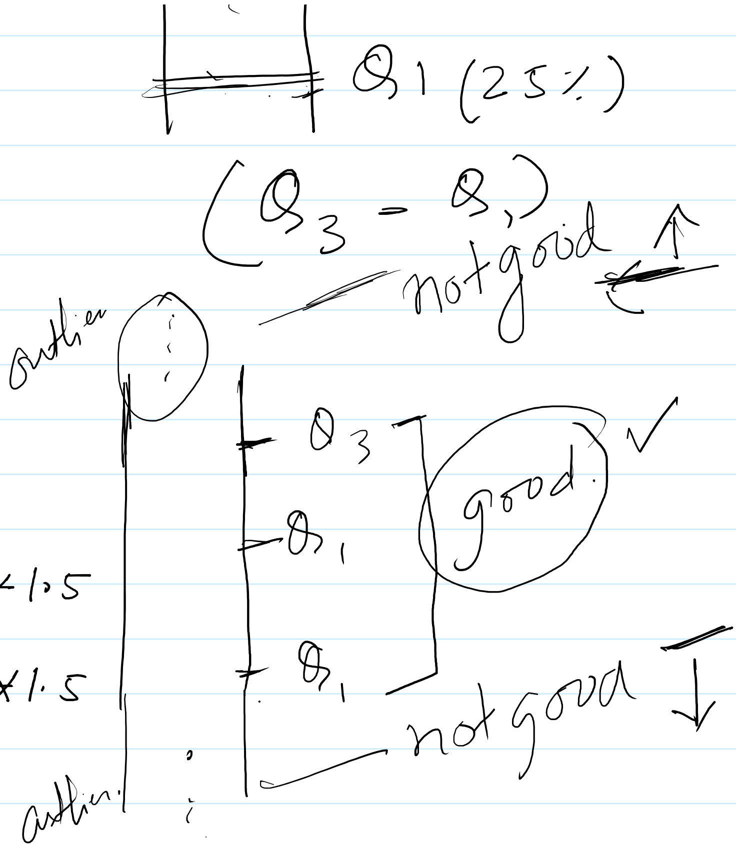
When to use MAD
- Compare this with std dev
- In case of outliers, choose MAD
- In case non normal data, choose MAD

Inter quartile Range (IQR)

- Difference of Q3-Q1
  - Central part of data
  - Excluding the extreme values
- Use case

- Central part of data
- Excluding the extreme values

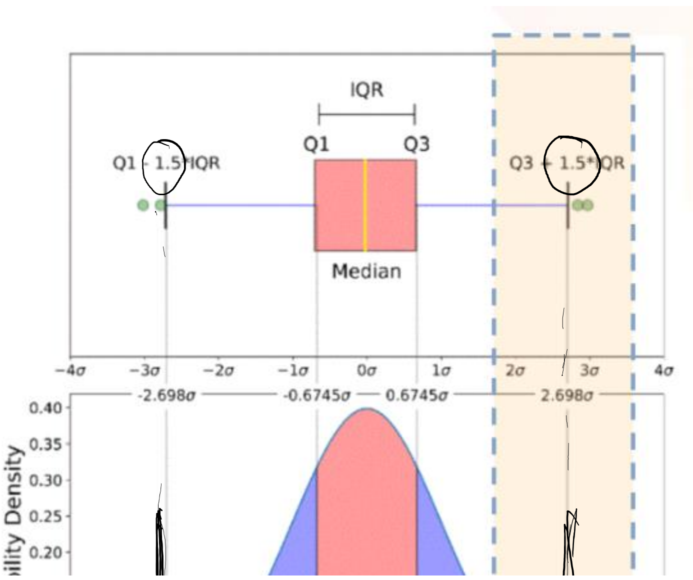
- Use case
  - Less impacted by outliers
  - Easy to explain to customers
  - Box plots - easy to explain
  - Outlier detection



Outlier detection

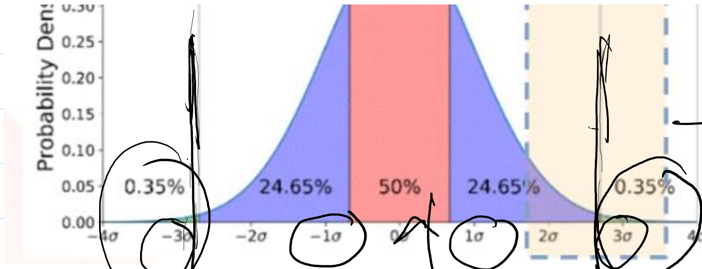
- threshold
  - higher
    - above Q3
      - $= Q_3 + \text{good} * 1.5$
      - $= Q_3 + (IQR) * 1.5$
      - = value
        - ↳ above
          - ↳ possible outliers.
  - lower
    - below Q1
      - $= Q_1 - \text{good} * 1.5$
      - $= Q_1 - IQR * 1.5$
      - = value
        - ↳ below
          - ↳ possible outliers.

Why 1.5?



a typical box plot

imaginary,



imaginary,

extreme

normal distribution.

n.d

σ

: Symmetrical.

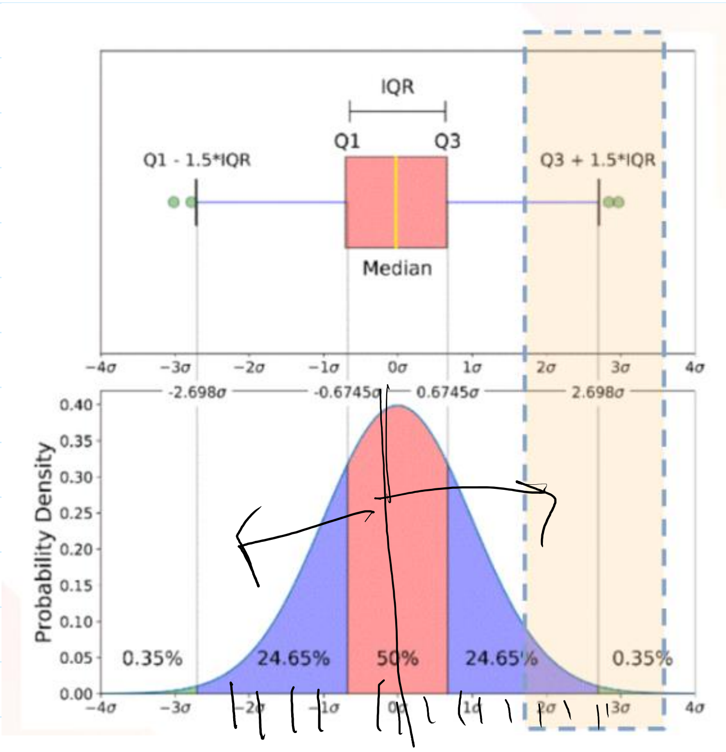
. 68% of data values → −1 to 1 σ.

. 95% - - - - - - → −2 to 2 σ

. 99% - - - - - → −3 to 3σ.

z-score

- Convert the data value
    - In terms of units of std dev



5
7
8
4
8
18

(n.σ).

$\mu = 6$

$\sigma = 2$

$\frac{x - \mu}{\sigma} = \frac{5-6}{2} = -\frac{1}{2}$

$= -.5\sigma$

2σ

2.1σ

3σ

Scaling

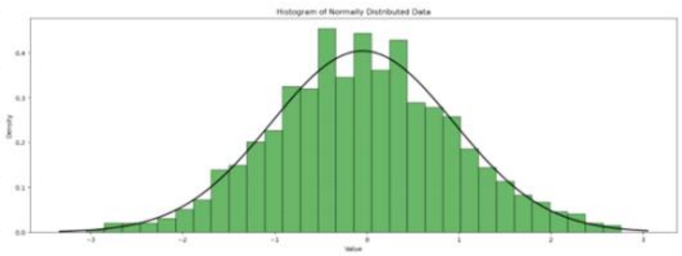
z-score.

Use cases of z-score

- Outlier detection
    - Target +-3.1 sigma
- Scaling purposes (normalization)

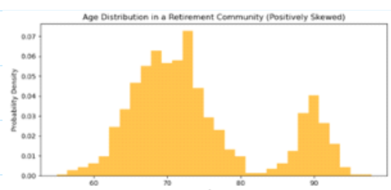
Data symmetry

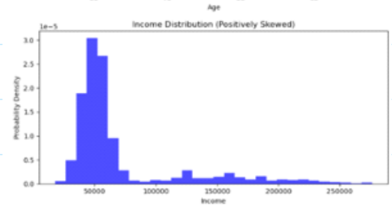
- Why we measure symmetry
    - Symmetrical is GREAT
        - Most analytics/ML - work well
- Methods
    - Viz
        - Histograms
        - Density plots
    - Kurtosis and Skew
- **Kurtosis**
    - Tail of the data distribution
        - Compare this tail with a normal distribution
    - To measure the heaviness
        - More heavy tail
            - More likely the data has outliers
    - Intuition
        - Tailedness
        - Relative amount of extreme values in the col
        - Pos kurtosis
            - More outliers
        - Negative kurtosis
            - Thin tails
            - Fewer outliers
    - Mesokurtic
        - Kurtosis = 3




```
print("Kurtosis (manual calculation):", computed_kurtosis)
Kurtosis (manual calculation): 2.953233675521671
```

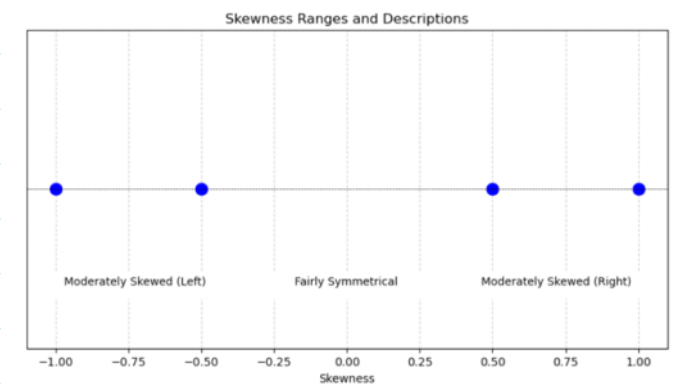
Positive Skew (right skew)
- Income distribution
- Home prices in your area
- Aging pop





Negative skew (left skew)

- Exam score
- Company profits
-

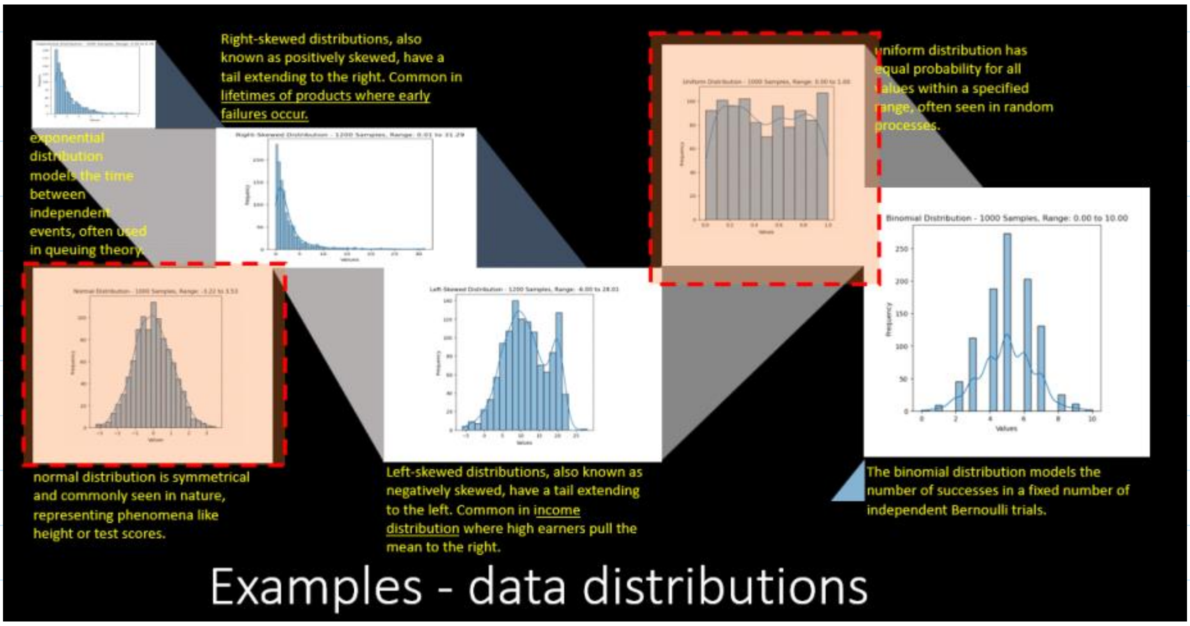


Random Variable

- Some kind of a fn
  - Yields some value
    - Discrete
      - countable
    - Float (continuous)
      - Any value in the range
- Examples
  - Rolling a die {1,2,3,...6}
    - X = {1,2,3,..6}
    - X = sum of 3 dice rolls
  - Avg price of an asset
  - ROI on investment

Data distribution

-



Data distribution vs dispersion

Probability distribution
- Discrete
  - PMF (Probability mass fn)
- Continuous
  - PDF (Probability density fn)

PMF (Probability mass fn)
- Sum of the probs = 1
- PMF will assign prob to each value that X can take

Die

$$P(X=1)=\frac{1}{6}$$
$$P(X=2)=\frac{1}{6}$$
$$P(X=3)=\frac{1}{6}$$
$$P(X=4)=\frac{1}{6}$$
$$P(X=5)=\frac{1}{6}$$
$$P(X=6)=\frac{1}{6}$$

(i) $f(x) \geq 0$ for all $x \in S$,

(ii) $\sum_{x \in S} f(x) = 1.$

Expected value
- Using PMF we can compute the expected value of the variable

$$E(X) = \sum_i x_i \cdot P(X = x_i)$$

where:

- x_i are the possible values of the random variable,
- $P(X = x_i)$ is the probability mass function (PMF) evaluated at x_i,
- and the summation is taken over all possible values of X.

Complaints	Probability

0	0.05
1	0.1
2	0.15
3	0.16
4	0.2
5	0.13
6	0.1
7	0.07
8	0.04

$E[X]$

$= 0\times0.05+1\times0.1+2\times0.15+3\times0.16+4\times0.2+5\times0.13+6\times0.1+7\times0.07+8\times0.04$
$= 0+0.1+0.3+0.48+0.8+0.65+0.6+0.49+0.32$
$= 3.74$

2. What is the probability that the number of complaints will exceed the expected number?

= 0.2 + 0.13 + 0.1 + .07 + .04
= 0.54

$$P(X=x) = \frac{x+2}{38}, \quad x \in S = \{4,\ 5,\ 8,\ 13\}$$

Does the above define a valid probability mass function?

- For all given values of X, the probability is > 0

- Sum of the prob
= P(4) + P(5) +P(8) + P (13)
= 6/38 + 7/38 + 10/38 + 15/38
= 38/38 = 1

Types of PMFs
- Bernoulli distribution
  - Var can take only 2 values
    - Success (1) , prob=p
    - Failure (0), prob = 1-p

- pertains to scenarios featuring a single trial with two potential outcomes
- experiments posing a binary question
  - whether a coin will land on heads,
  - if a die roll will result in a 6,
  - if an ace will be drawn from a deck of cards, or
  - if voter X will opt for “yes” in a referendum.
  - a team will win a championship or not
- Essentially, Bernoulli trials encompass situations where the two potential results can be framed as “success” or “failure,” though these terms aren't strictly literal.
- In this context, “success” simply denotes achieving a “yes” outcome (e.g., rolling a six, drawing an ace, etc.).

- The **expected value** is
  $E(X)$
  $= 0 \times (1-p) + 1 \times p$
  $= p$
- The **variance** is
  $Var(X)$
  $= E(X^2) - E(X)^2$
  $= 1^2 \times p + 0^2 \times (1-p) - p^2$
  $= p - p^2$
  $= p(1-p)$

Binomial distribution

- Prob dist of number of success
  - In a fixed number of independent Bernoulli trials
    - 2 values
      - Prob
- Series of Bernoulli trials
- Binomial actually summarizes the total number of of success
- Calculation

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Where:

- n is the number of trials,
- k is the number of successes,
- p is the probability of success in each trial.

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%55%, and the probability that you lose is 45%45%

What is the probability that you win 15 times if you play the game 20 times?