Day 3-3.5

18 May 2024 18:16

Central measures

- Mean
  - AM
  - GM
  - HM
- Median
- Mode

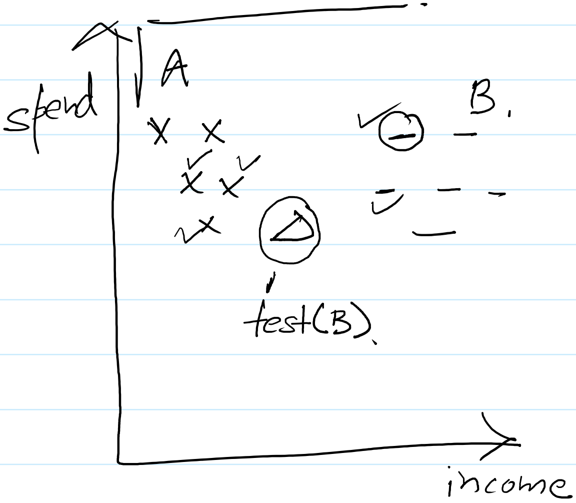
Weighted mean

- 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

customers



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

△ — c1 = 10
— c2 — 20
— c3 — 11
: — 15
5 nearest

income	spend
1K	2K
5K	4K

12k

✓ 1K 4k — c3 — 11] 5 nearest

5k — 15

✓ 1.5K 2.5k

5
→ 3A ✓ (t = A).
→ 2B.

∴ all 5 hv equal vote/weight.

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

euclidean → 8th grade.

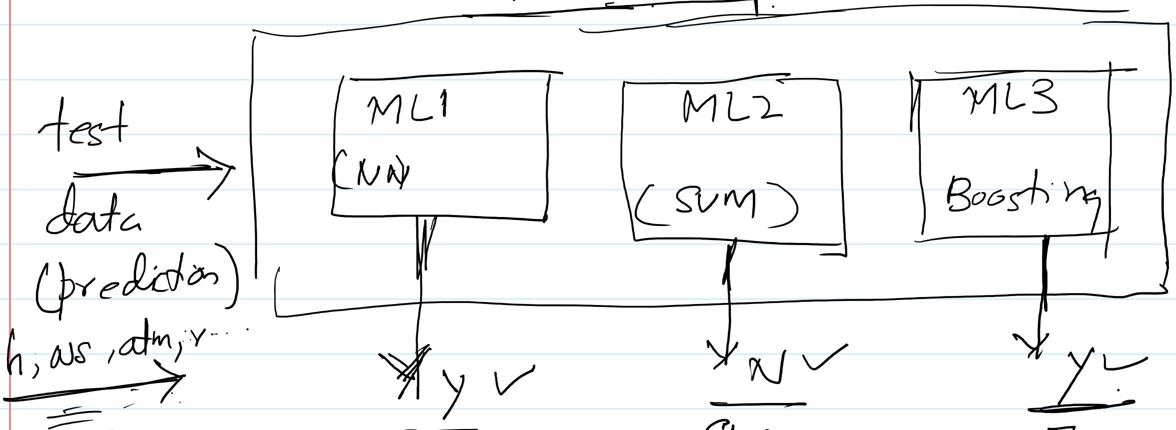
powerful.

- preferential weightage.
- distance

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

Data driven methods
- Example

prediction model || temp. prediction.



100 ✓ 90 70 -✗

(w₁ ✓ w₂ ✓ w₃)

(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:

$20,000
$30,000
$35,000
$40,000
$1,000,000 (outlier)
$2,000,000 (outlier)

What's big with this

Winsorized Income Data:

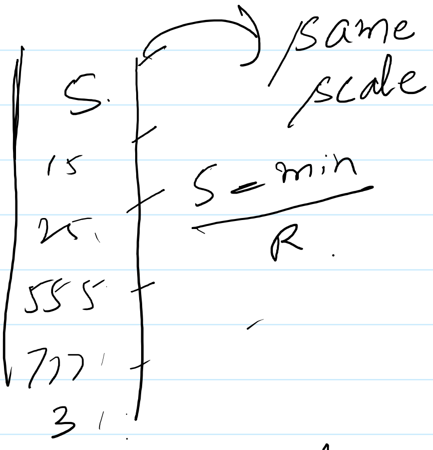
$20,000
$30,000
$35,000
$40,000
$40,000 (capped outlier)
$40,000 (capped outlier)

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 %



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

$$\left[\frac{x - \min}{\max - \min}\right] \quad [0, 1]$$

(R)

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

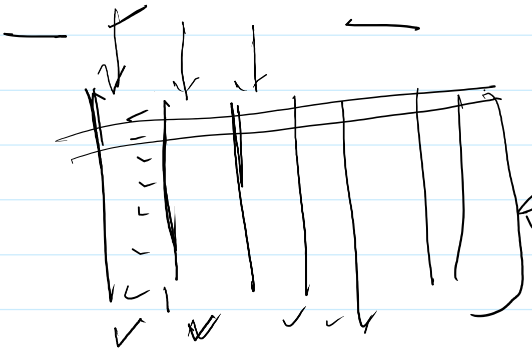
$$\text{Variance} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \text{ for a sample,}$$
$$\text{Variance} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \text{ for a population.}$$

[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

$$\text{Var} = \frac{100}{10} = 10$$

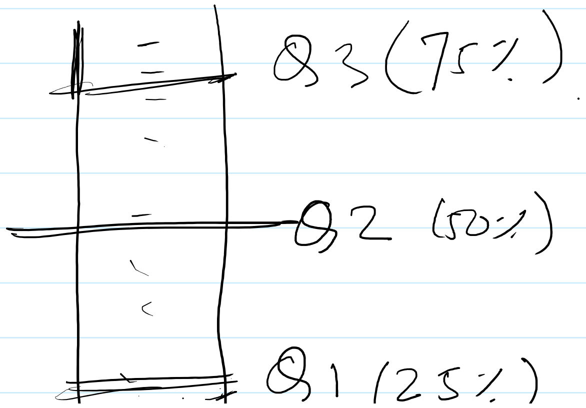


Standard deviation
- Square root of var
- Purpose
  - Same unit
    - Customers can easily relate to it

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

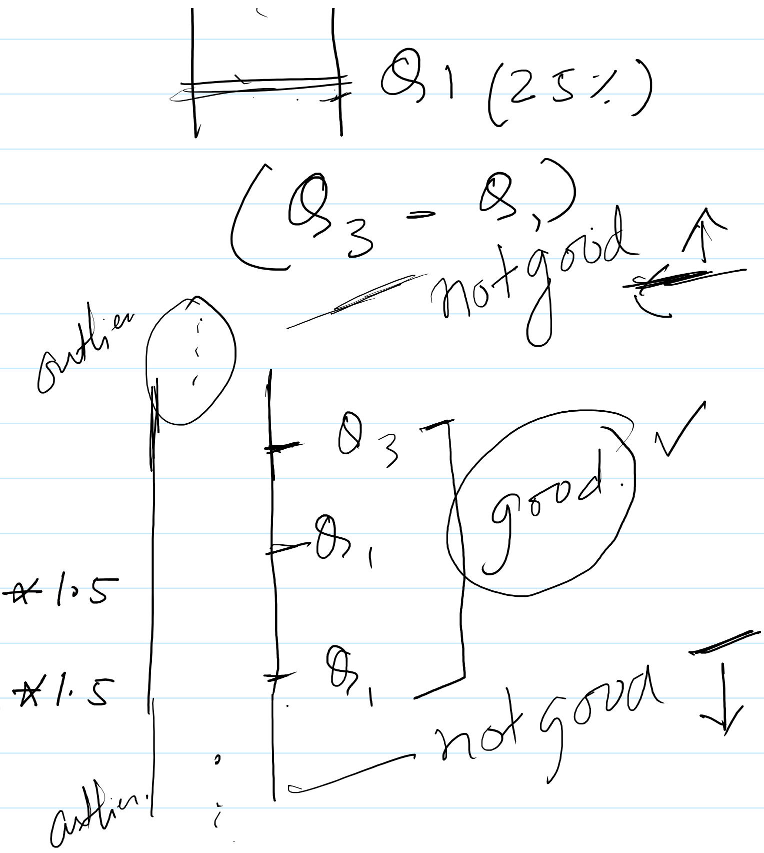


Inter quartile Range (IQR)

- 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

$= Q_3 + \underline{good} * 1.5$

$= Q_3 + (IQR) * 1.5$

$= value$

↳ above

↳ possible outliers.

- lower
    - below Q1

$= Q_1 - good * \underline{1.5}$

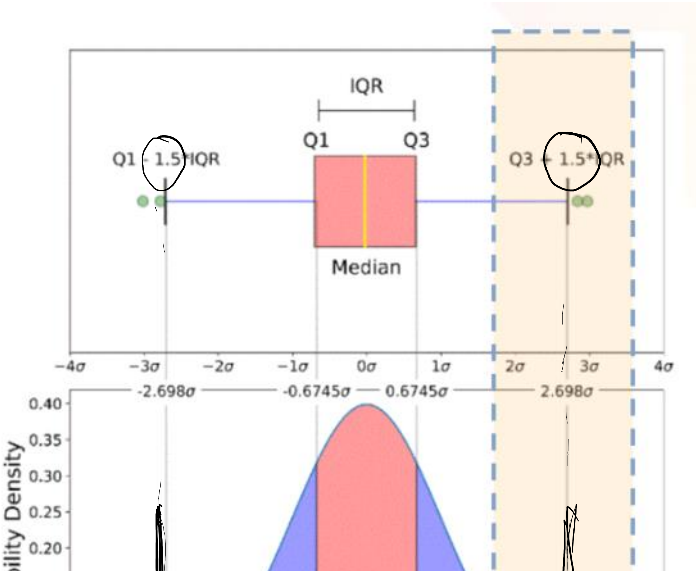
$= Q_1 - IQR * 1.5$

$= value$

↳ below

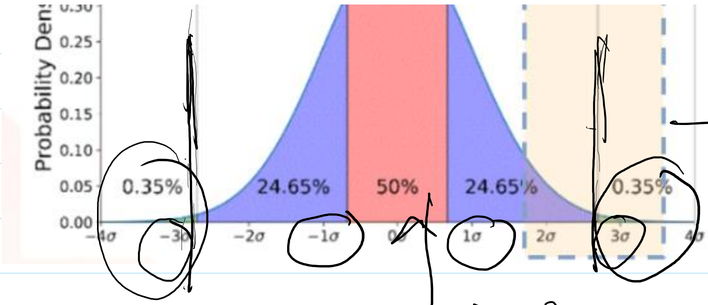
↳ possible outliers.

Why 1.5?



a typical box plot

imaginary,



imaginary,

extreme

→ normal distribution.

σ

n.d

.. Symmetrical ·

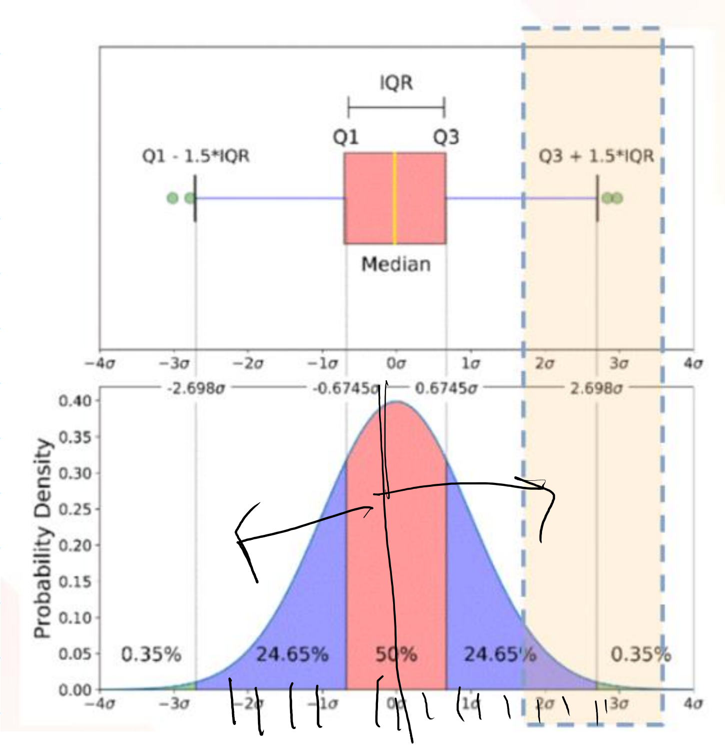
, 68% of data values → -1 to 1 σ

· 95% ------→ -2 to 2σ

· 99% ----→ -3 to 3σ.

z-score

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



5
7
8
4
8
18

f(n.σ).

$\mu = 6$

$\sigma = 2$

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

$= -.5\sigma$

Scaling → z-score.

2σ

2.1σ

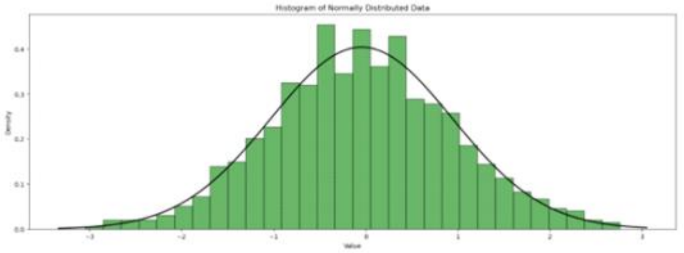
3σ

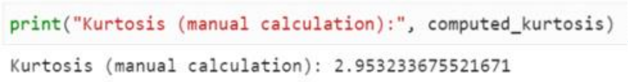
Use cases of z-score

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

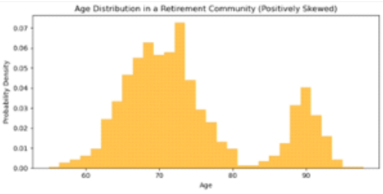
Data symmetry

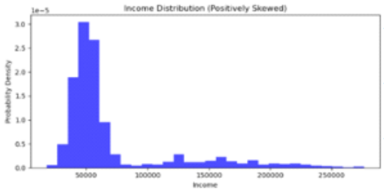
- Why we measure symmetry
    - Symmetrical is GREAT
        - Most analytics/ML - work well
- Methods
    - Viz
        - Histograms
        - Density plots
    - Kurtosis and Skew
- **Kurtosis**
    - Tail of the data distribution
        - Compare this tail with a normal distribution
    - To measure 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
```



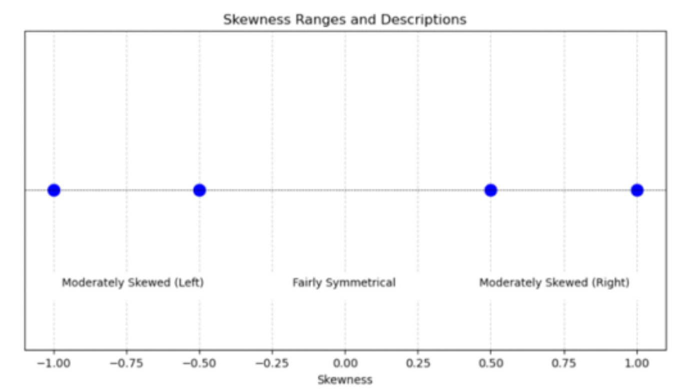


Positive Skew (right skew)

- Income distribution
- Home prices in your area
- Aging pop

Negative skew (left skew)

- Exam score
- Company profits
-

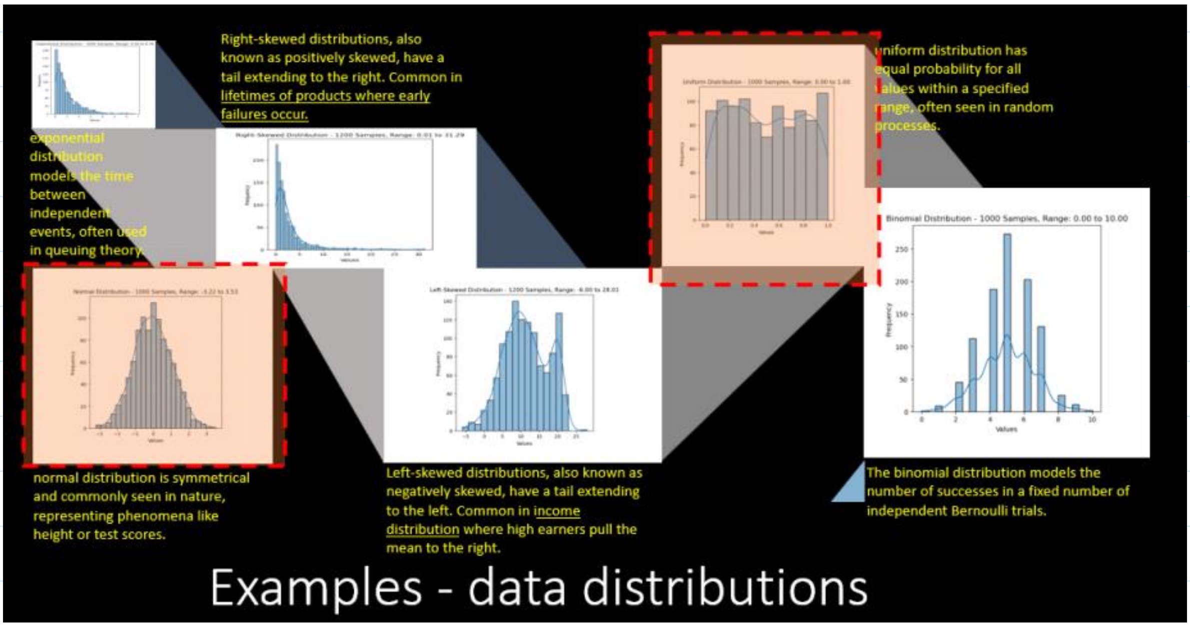


Random Variable

- 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

-


Right-skewed distributions, also known as positively skewed, have a tail extending to the right. Common in lifetimes of products where early failures occur.

exponential distribution models the time between independent events, often used in queuing theory.

uniform distribution has equal probability for all values within a specified range, often seen in random processes.

normal distribution is symmetrical and commonly seen in nature, representing phenomena like height or test scores.

Left-skewed distributions, also known as negatively skewed, have a tail extending to the left. Common in income distribution where high earners pull the mean to the right.

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials.

Examples - data distributions
Right-Skewed Distribution - 1200 Samples, Range: 0.01 to 31.29
Normal Distribution - 1000 Samples, Range: -3.22 to 3.53
Left-Skewed Distribution - 1200 Samples, Range: 4.00 to 28.01
Uniform Distribution - 1000 Samples, Range: 0.00 to 1.00
Binomial Distribution - 1000 Samples, Range: 0.00 to 10.00
Frequency
Values

Data distribution vs dispersion

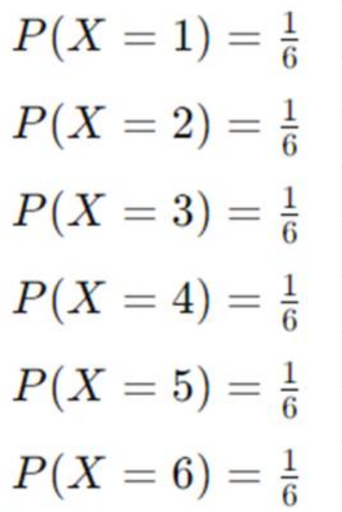
Probability distribution

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

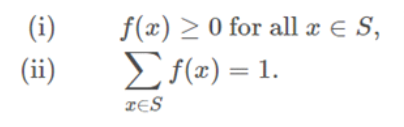
PMF (Probability mass fn)

- Sum of the probs = 1
- PMF will assign prob to each value that X can take

Die



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

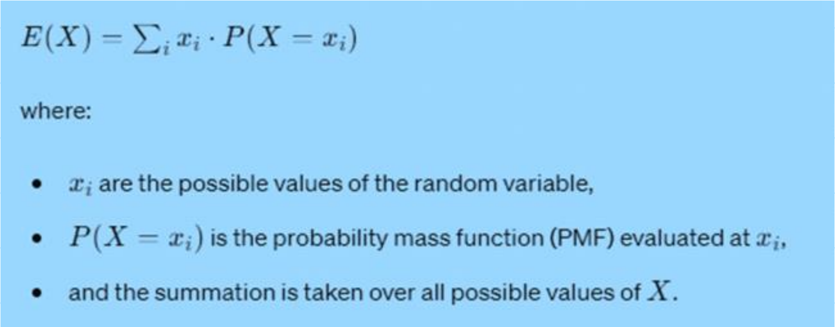


(i) $f(x) \geq 0$ for all $x \in S$,

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

Expected value

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



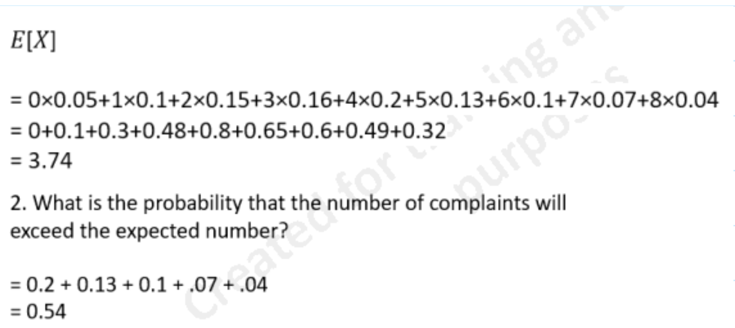
$E(X) = \sum_i x_i \cdot P(X = x_i)$

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

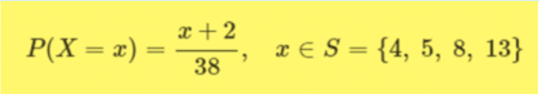


$E[X]$

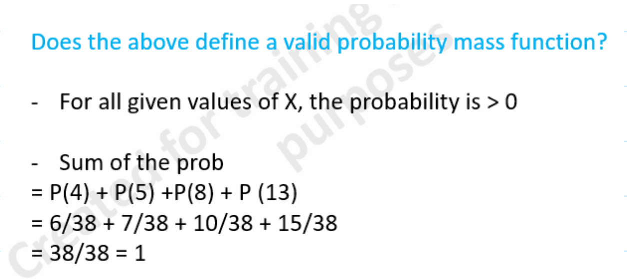
= 0×0.05+1×0.1+2×0.15+3×0.16+4×0.2+5×0.13+6×0.1+7×0.07+8×0.04
= 0+0.1+0.3+0.48+0.8+0.65+0.6+0.49+0.32
= 3.74

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?

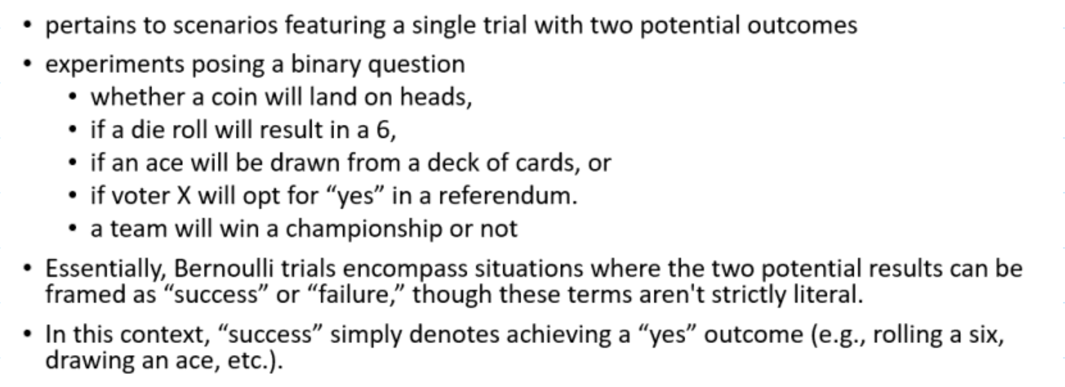
- For all given values of X, the probability is > 0

- Sum of the prob

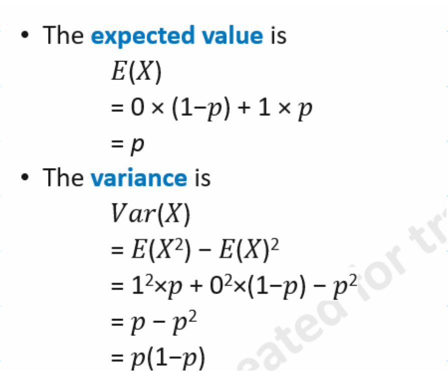
= P(4) + P(5) +P(8) + P (13)
= 6/38 + 7/38 + 10/38 + 15/38
= 38/38 = 1

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

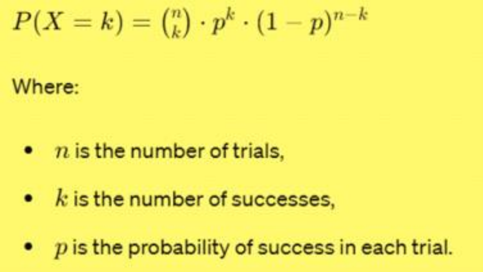
  $E(X)$
  $= 0 \times (1-p) + 1 \times p$
  $= p$

- 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%, and the probability that you lose is 45%

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

= BINOM.DIST(C4,C3,C5, FALSE)

Examples

- Calc the prob that a certain number of emails are spam per day, given p = prob that an email is spam

Mean -

mu = n * p

Variance = n*p(1-p)

Mode = $\lfloor (n+1) * p \rfloor$

$3.7 = 3$

$[2] = 2$

$\lfloor -1.5 \rfloor = -2$

$\lfloor 4.1 \rfloor = \underline{\underline{4}}$

- Hepatitis C (“Hep C”) is a virus affecting the liver, whose symptoms include inflammation, cirrhosis, and all sorts of other nastiness.
- According to the World Health Organization (WHO), 4.1% of people have Hep

C. In humans, it is discovered via a routine blood test.

- In screening for Hep C, some health care providers, to save time and money, combine blood samples from 5 patients to test, and
  - if the combined sample comes back negative, it means that all 5 folks are not infected with Hep C.
  - if the combined sample comes back positive, it means that at least one of the people in the combined sample has Hep C.
    - Then, each of the 5 must be individually tested to see who is infected.
- Suppose 5 randomly selected folks have their blood taken, and their blood is placed into a combined sample. This combined sample is then tested for Hep C.

P(h-c) = 0.041 (positive)
p(no h-c) = 1 - .041 = 0.959 (negative)

Qs 1 : Find the chance of combined sample comes as NEGATIVE

Ans :

- For the combined result to -ve, all 5 patients must be free of hep-C
- p(no h-c) = 0.959 (negative)
- Prob(of 5 being -ve) = .959 ^ 5 = 0.8111 = 0.8111 = 0.8111

Qs 2 : Find the chance of combined sample comes as POSITVE

Ans = 1 - .811 = 0.189

Qs 3 : Suppose a Hep C test (whether done on a single blood sample, or 5) costs around $100. If you sample 100 random people, approximately how much is saved by using combined samples instead of individual testing (assuming you want to individually ID all those infected with Hep C)?

According to a 2010 CDC report, approximately 85% of Americans have health insurance. Suppose we randomly select 10 Americans.

Ans

- Individual costs
  - $100 each and 100 patients
  - $10000
- Combined costs
  - 100 patients, 5 in a group
  - 20 samples
  - Cost = 20 * $100 = $2,000.00

    - Individual testing (after the sample is found positive)
        - 20 * 19% = 3.8 (4)
        - Cost
            - □ 4 * 5 = 20
            - □ 20 * 100=2,000
    - 2000 + 2000 = 4000
- Savings = 10000 - 4000 = 6000

Qs : will the combine test always be beneficial?

- If yes
    - Not true
- If no
    - Reasons
        - High prob for the hep-C

Qs : what type of sampling we used here (dependent / independent)

- Random selection
- Prob is for each individual

Qs : find prob (all of them have insurance - all the 10 patients)

According to a 2010 CDC report, approximately 85% of Americans have health insurance. Suppose we randomly select 10 Americans.

Ans : p = 0.85

P (all) = 0.85 ^ 10=0.1969

Qs : Find prob (5 of them have health insurance)

N = 10

K = 5

P = .85

Qs : Find the most likely number of them that have insurace

Mean -

mu = n * p
Variance = n*p(1-p)
Mode = $\lfloor \underline{(n+1) * p} \rfloor$

Variance = n, p(1-p)

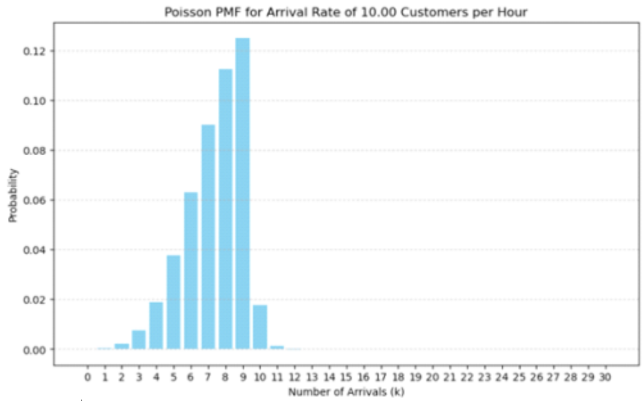
Mode = $\lfloor \underline{(n+1)*p} \rfloor$

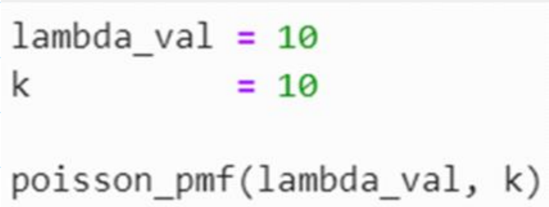
11 * .85=9.35 = 9

Another PMF - Poisson distribution

- Rate - lambda
  - ○ Number of events per unit time
    - ▪ Number of calls
    - ▪ Number of tickets
    - ▪ Number of customers
  - ○ Queueing model (OR)
    - ▪ Arrival rate

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$




```
lambda_val = 10
k          = 10

poisson_pmf(lambda_val, k)
```

PMF = 0.12 = prob associated with observing k = 10

Max PMF : lambda = k

Why Poisson distribution is used in sampling (arrival data)

- ○ Memoryless property
  - ▪ One arrival has nothing to do with prev one
    - □ Independent
  - ▪ Suitable modeling
    - □ Rare event
  - ▪ Empirical
    - □ Studies have shown the results are quite good