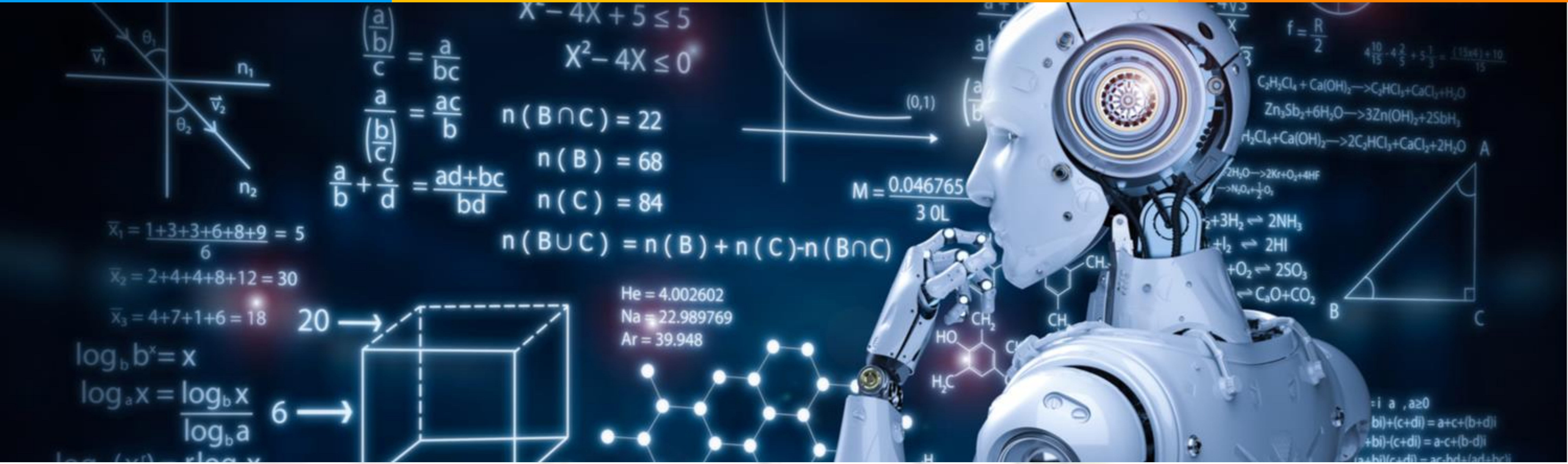
Duration: 60 min | Srikanth Shenoy | AML 5101 | M.E. AI & ML



Lecture 06 – Vector aggregate ops

Recap

- Block Vectors

Semantic meaning of vector magnitude

- E.g. Rainfall in the month of July (31-vector)
- BP of all patients
- Recall magnitude = $\sqrt{{x_1}^2 + {x_2}^2 + {x_3}^2}$

$$x = \begin{bmatrix} 2 \\ 2.25 \\ .. \\ 3.5 \\ .. \\ 3 \end{bmatrix}$$

- Makes no semantic sense
- Does magnitude make sense for a vector with heterogenous data (e.g. Patient vector)?

Units of heterogeneous vector entries

- Calculate nearest/farthest patients data
- Change units of BP to micromm Hg
- Calculate nearest/farthest patients data again
- What do you observe?
- You already know this:
  - Unit-less comparison is ideal
  - Z transformation
  - Standard Scaler in sklearn

$$z = \frac{x - \mu}{\sigma}$$

1. Vector aggregation operations

Aggregation ops on vector: Summary

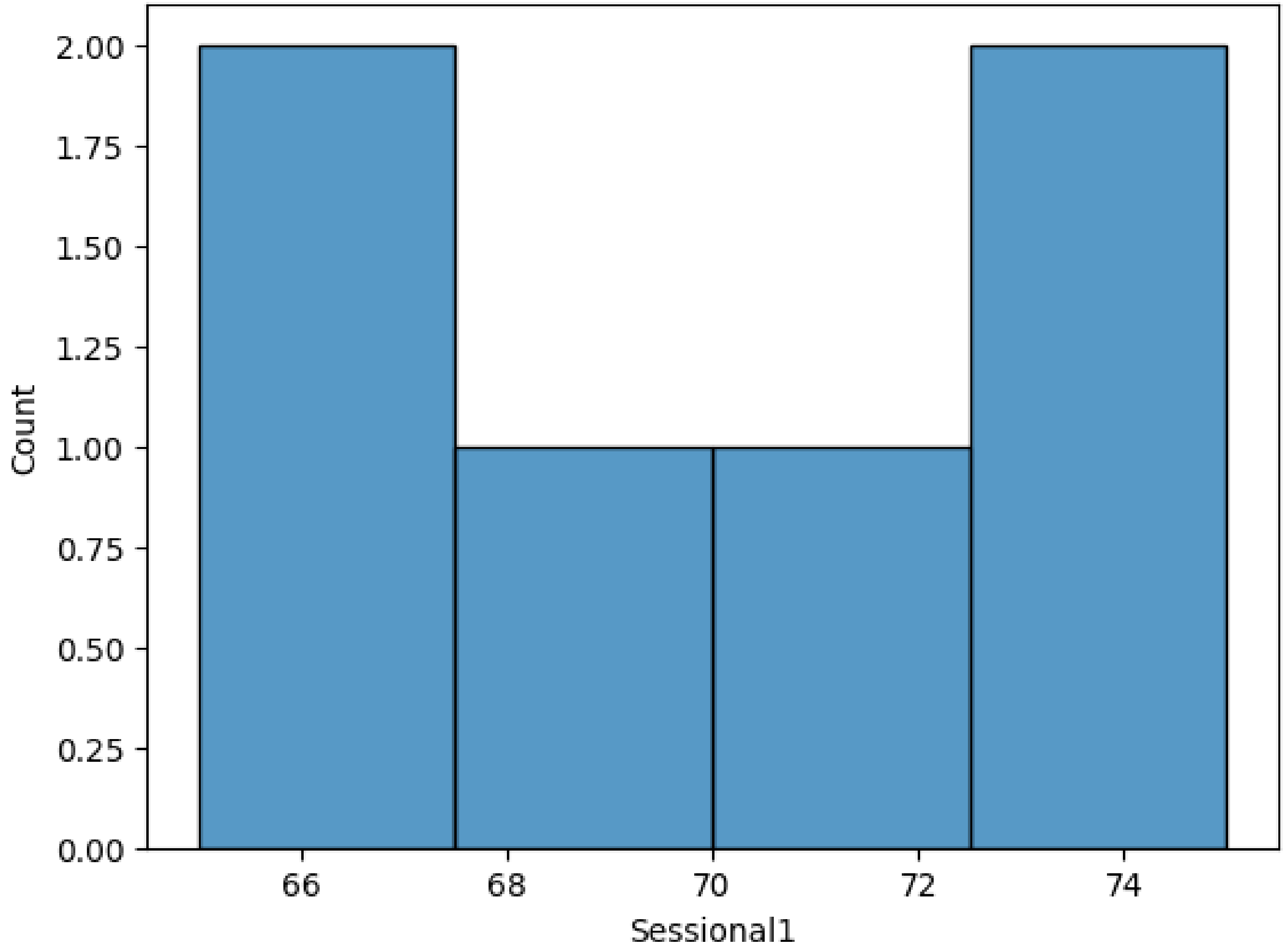
- Should be homogeneous
  - E.g. Rainfall in the month of July (31-vector) $x = \begin{bmatrix} 2 \\ 2.25 \\ .. \\ 3.5 \\ .. \\ 3 \end{bmatrix}$
  - BP of all patients
- Statistical operations on homogeneous data
  - Mean   Absolute mean

    $\frac{1}{n}\Sigma(x_i)$ $\frac{1}{n}\Sigma(|x_i|)$
  - Root Mean Square (RMS) Value
  - Mean Absolute Deviation (MAD)
  - Standard Deviation (SD), Variance

Hint: Read backwards to perform the operation

Calculating RMS value

Student	Sessional1	Sessional1 marks squared
Student1	73	5329
Student2	67	4489
Student3	75	5625
Student4	65	4225
Student5	72	5184
Student6	68	4624



- RMS value = $\sqrt{\frac{1}{n}(x_1^2 + x_2^2 + .. + x_6^2)} = \frac{\|x\|}{\sqrt{n}} = \sqrt{\frac{x^T x}{n}}$

- Meaning: RMS is typical value of vector entry $|x_i|$

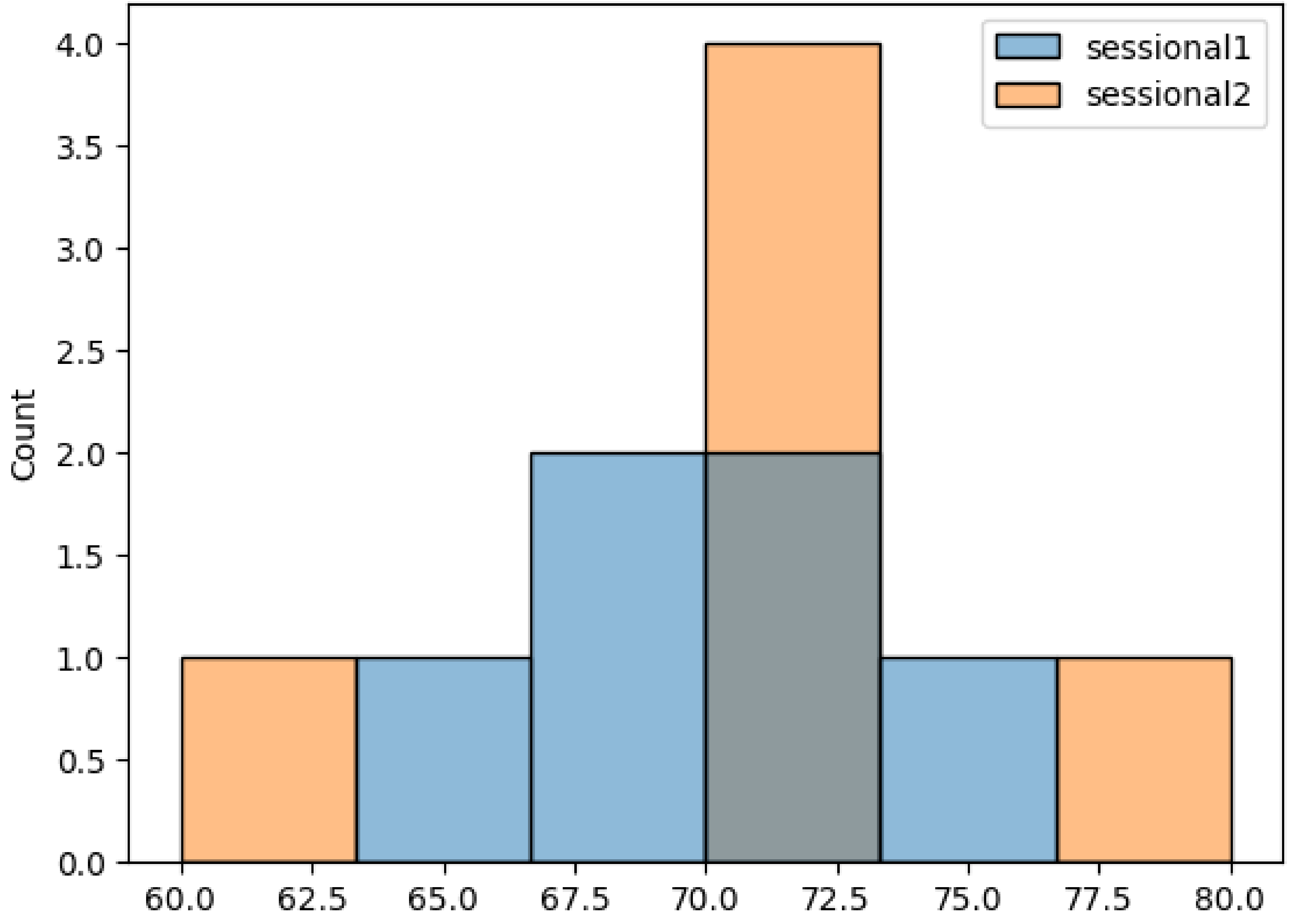
Calculating RMS value (contd.)

- RMS value = $\sqrt{\frac{1}{n}(x_1^2 + x_2^2 + .. + x_6^2)} = \frac{||x||}{\sqrt{n}} = \sqrt{\frac{x^T x}{n}}$
- Meaning: RMS is typical value of vector entry $|x_i|$
- Why not absolute mean? $\frac{1}{n}(|x_1| + |x_2| + .. + |x_6|)$

- Typical value IS NOT expected value
  - Expected Value captures central tendency
  - Typical value takes range (dispersion) into account

Why abs mean is not typical value?

Student	Sessional1	Sessional2
Student1	73	70
Student2	67	80
Student3	75	70
Student4	65	70
Student5	72	60
Student6	68	70



- In both cases abs mean = 70
- But Sessional2 has higher spread.
- Shouldn't it have higher typical value?

Loss functions in machine learning

- Loss functions measure our unhappiness aka error
- Error is deviation. Deviation from what?
  - Deviation of y from y-hat – $y - \hat{y}$
  - y is actual value of target variable
  - y-hat is the predicted value of target variable
- Errors
  - Mean Absolute Error - MAE
  - Mean Square Error - MSE
  - Root Mean Square Error - RMSE

Hint: Read backwards to perform the operation

Errors in machine learning (contd.)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

- MSE for optimization. RMSE for evaluation
  - Because we don't care about actual loss function value. Plus squared function is nice for convex optimization

Which is better – MSE or MAE?

- In MSE
  - Farther points are amplified by squaring
  - ML lingo: Farther points are penalized more
- Is MSE better than MAE?
  - Good & bad based on outliers
  - Good when outliers are removed
  - Bad when outliers should be used (Use MAE)

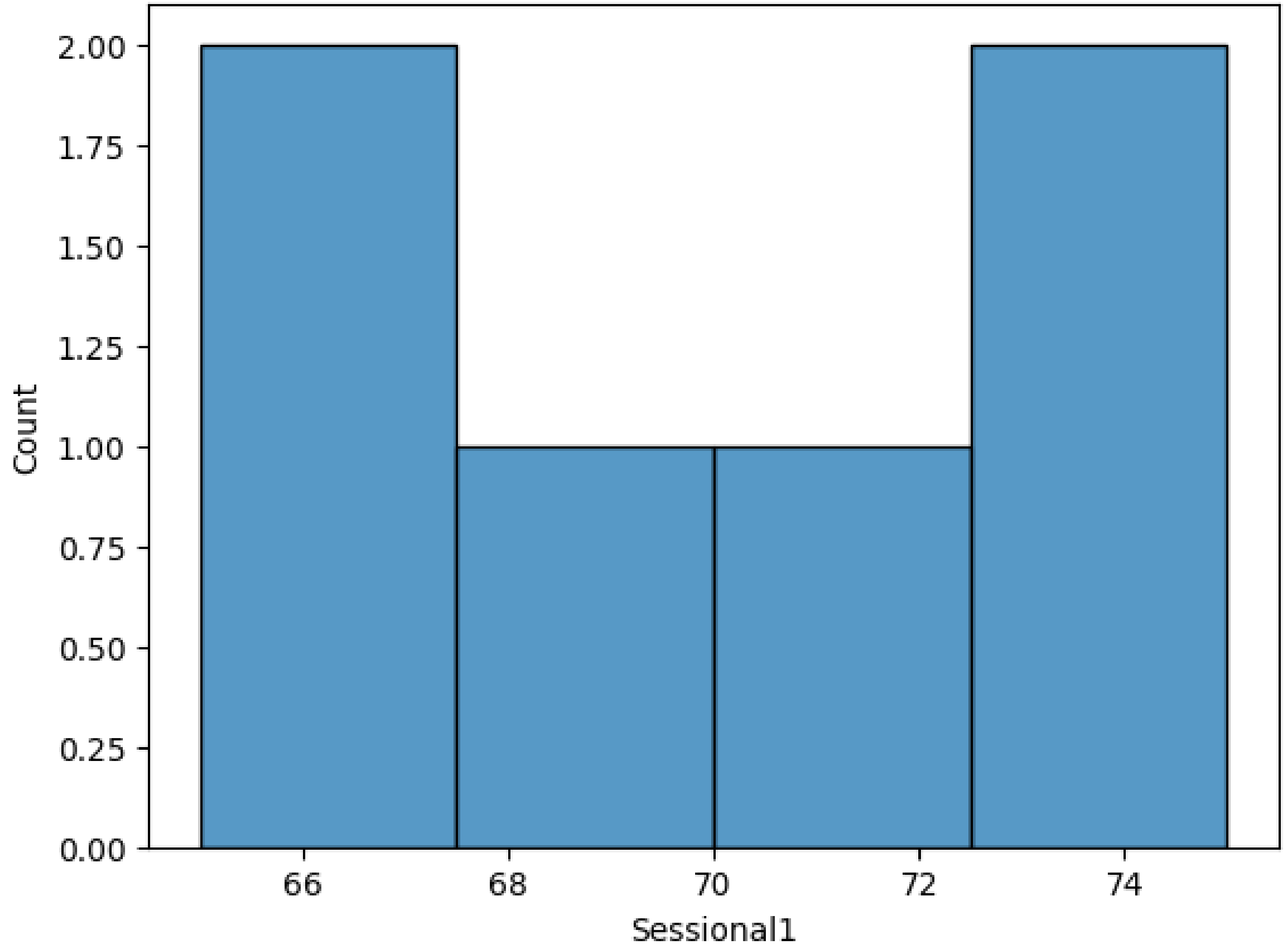
Chebyshev inequality

- Let the vector x contain homogenous data
  - It's entries are probability distribution
- Puts upper bound on fraction of entries > RMS (provides guarantee, regardless of distribution)
- Let k entries of a vector x > a $|x_i| > a \quad (a = \alpha \times RMS)$
- k is limited such that
  - Fraction of x entries (k/n) that can be at most α RMS values away

$$\frac{k}{n} \leq \left(\frac{rms(x)}{a}\right)^2$$

Calculating dispersion with MAD

Student	Sessional1	Sessional1 Abs Deviation
Student1	73	3.0
Student2	67	3.0
Student3	75	5.0
Student4	65	5.0
Student5	72	2.0
Student6	68	2.0

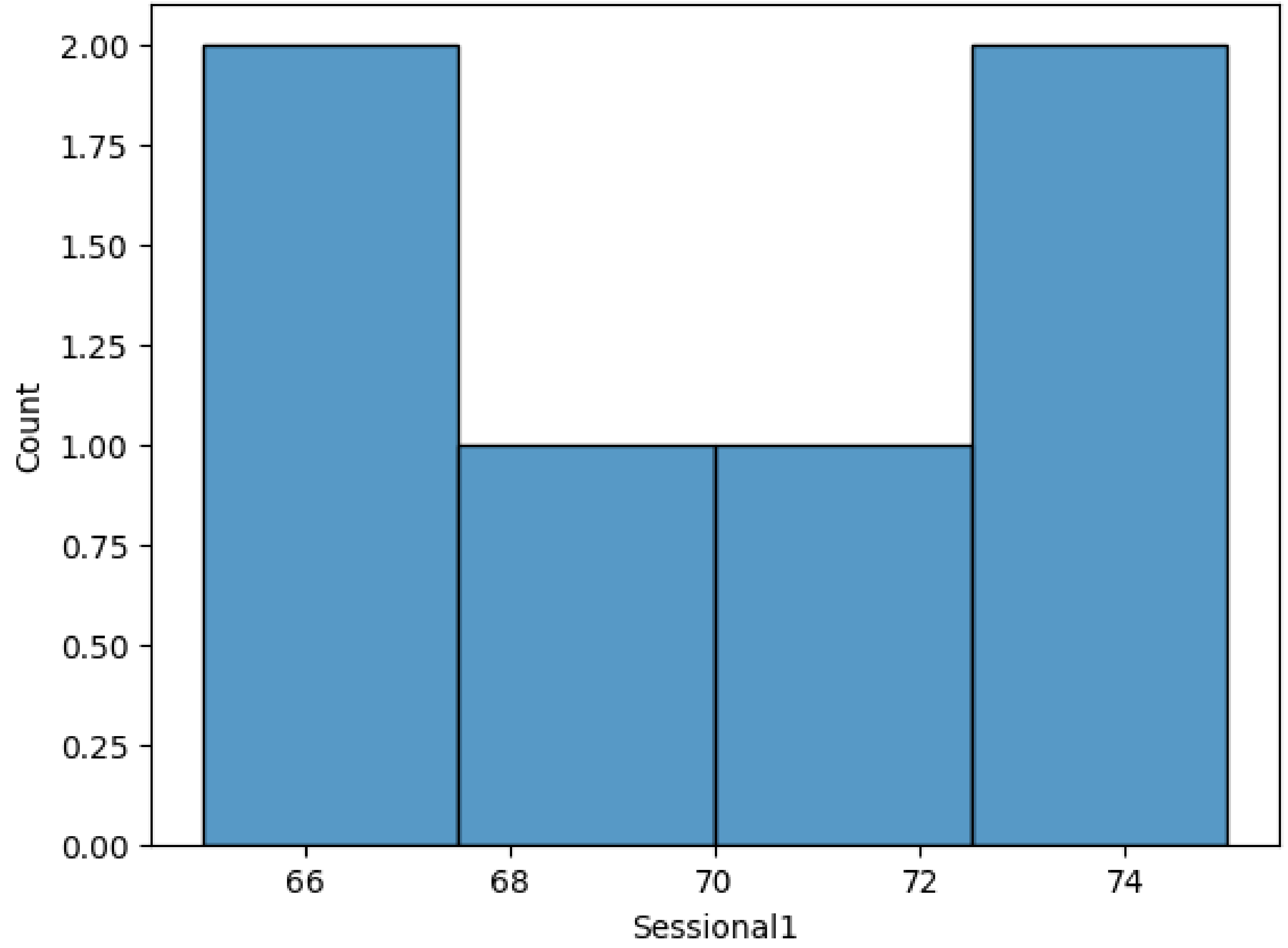


- Absolute Deviation per record = $|x_i - \mu|$
- MAD = $\frac{1}{n}\Sigma_{i=1}^{n}(|x_i - \mu|) = \frac{\|x - \mu\mathbf{1}\|_1}{n}$
- Mean = 70, MAD = 3.33

Use numpy broadcast in code

Calculating dispersion with Standard Deviation

Student	Sessional1	Sessional1 squared deviation
Student1	73	9.0
Student2	67	9.0
Student3	75	25.0
Student4	65	25.0
Student5	72	4.0
Student6	68	4.0



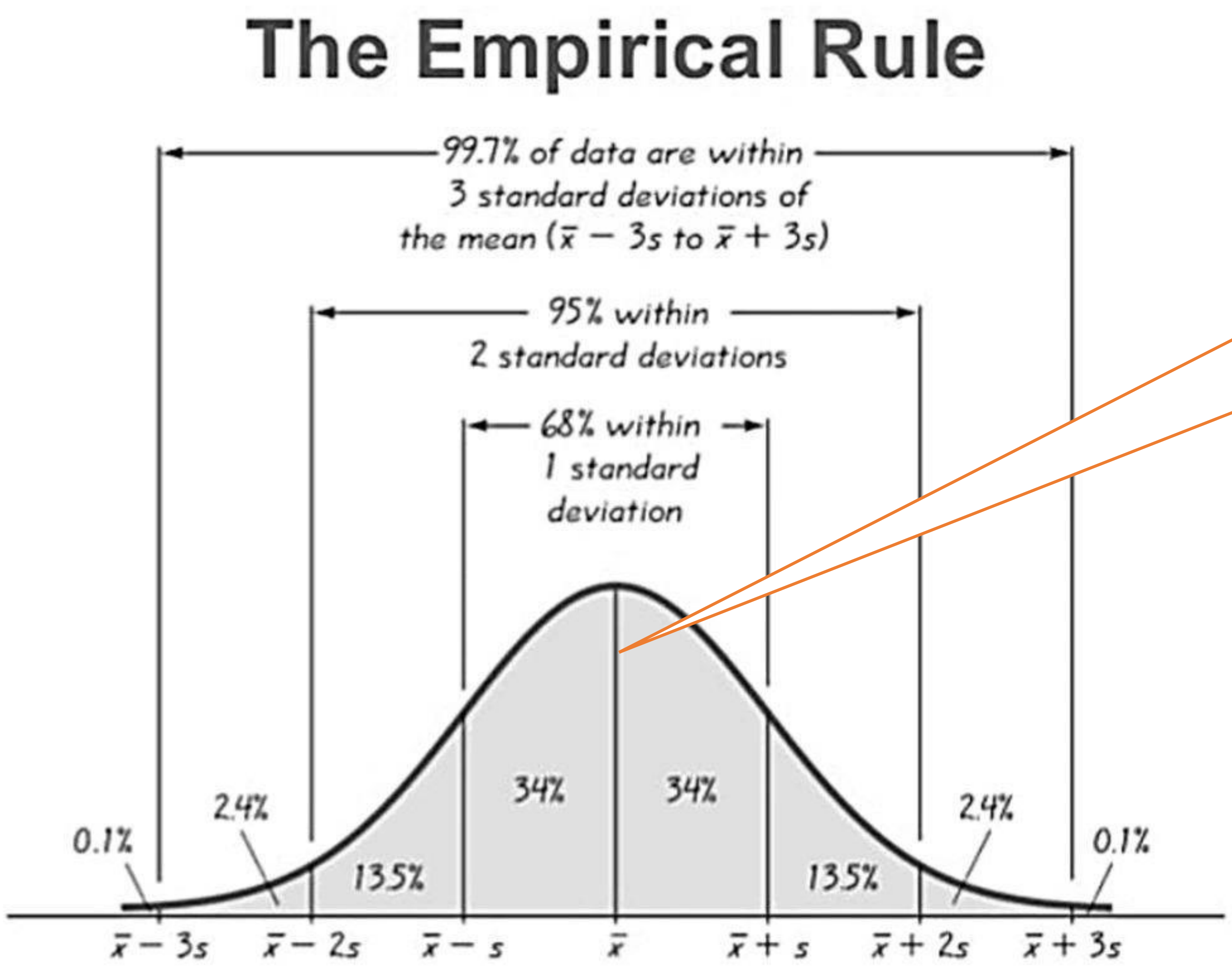
$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \mu)^2} = \frac{\|x - \mu\mathbf{1}\|}{\sqrt{n}} = RMS(x - \mu\mathbf{1}) \qquad \mu = \frac{1}{n}\mathbf{1}^T x$$

- SD is typical value of
  - Mean centered vector
  - deviation of vector entry from mean

SD = RMS value of the mean centered vector

Auditing standard deviation with Gaussian

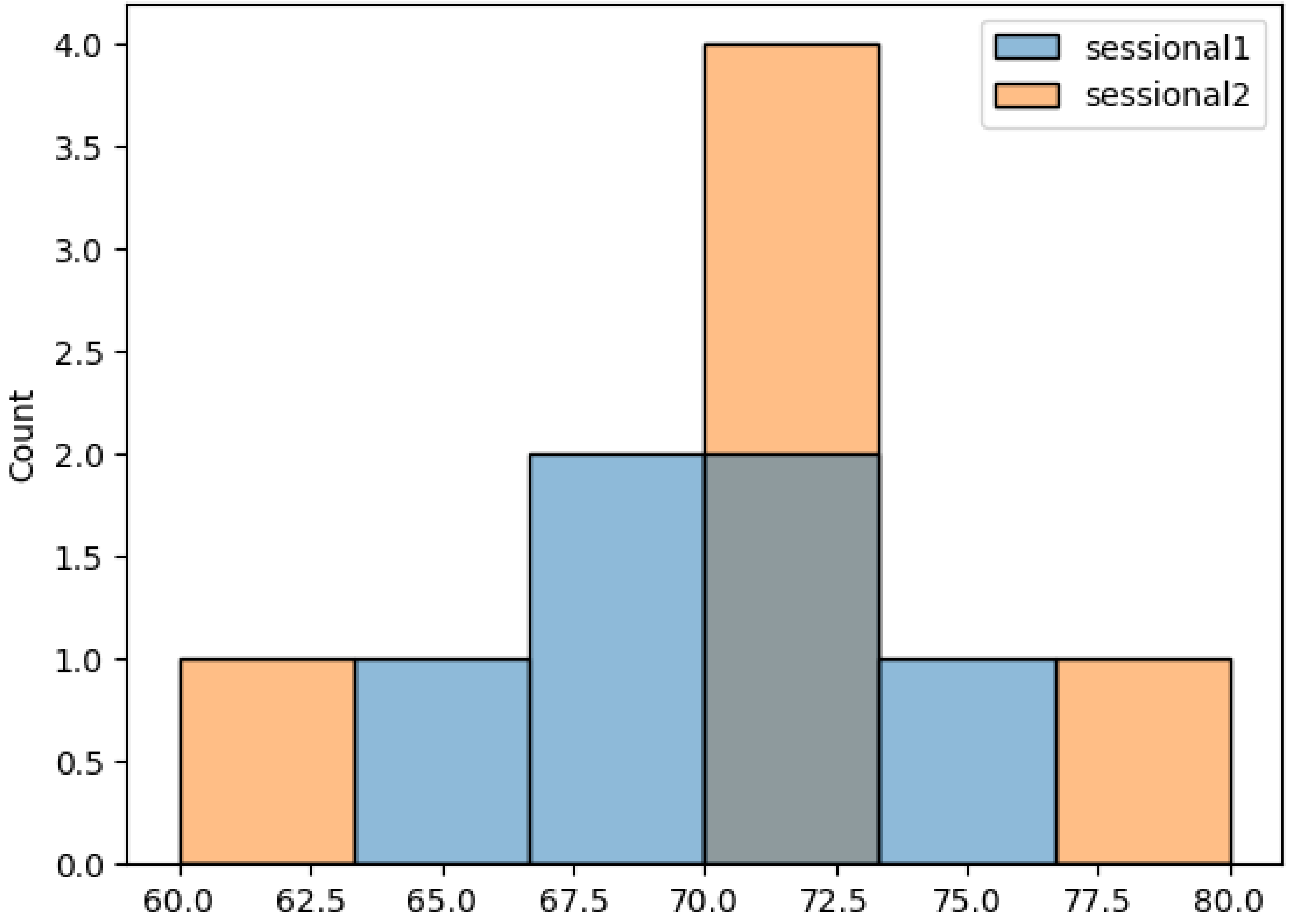
- SD is typical deviation of vector entry from mean implies
  - We can typically find a value within one standard deviation from mean



This is for Gaussian distribution only

Why SD instead of MAD?

Student	Sessional1	Sessional2
Student1	73	70
Student2	67	80
Student3	75	70
Student4	65	70
Student5	72	60
Student6	68	70



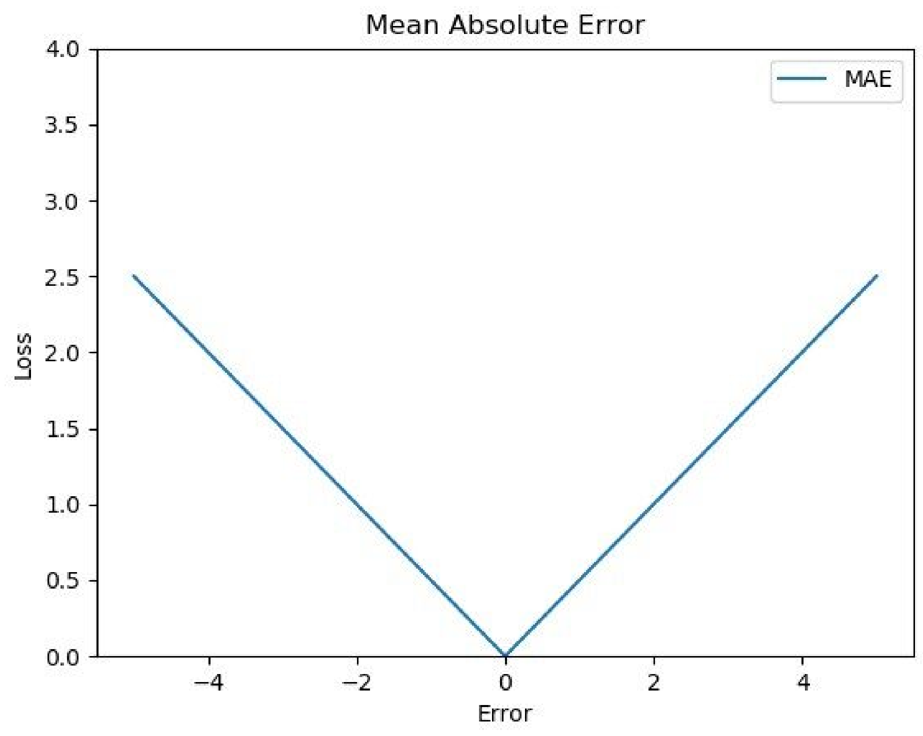
- In both cases mean = 70, MAD = 3.33
- STD for sessional1 = 3.55, sessional2 = 5.77

SD: points to ponder

- When is standard deviation 0 ?
  - Entries of x are constant
- How does SD capture dispersion better?
  - Ans: Farther points are amplified by squaring

SD instead of MAD (optional)

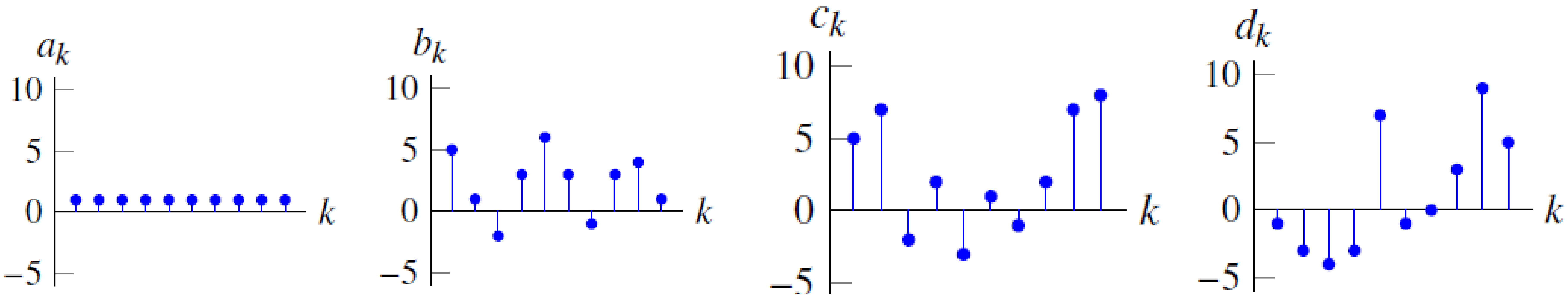
- Mod functions are not differentiable
  - Subgradient work around exists
- Not just a convenient way for easing further calculations

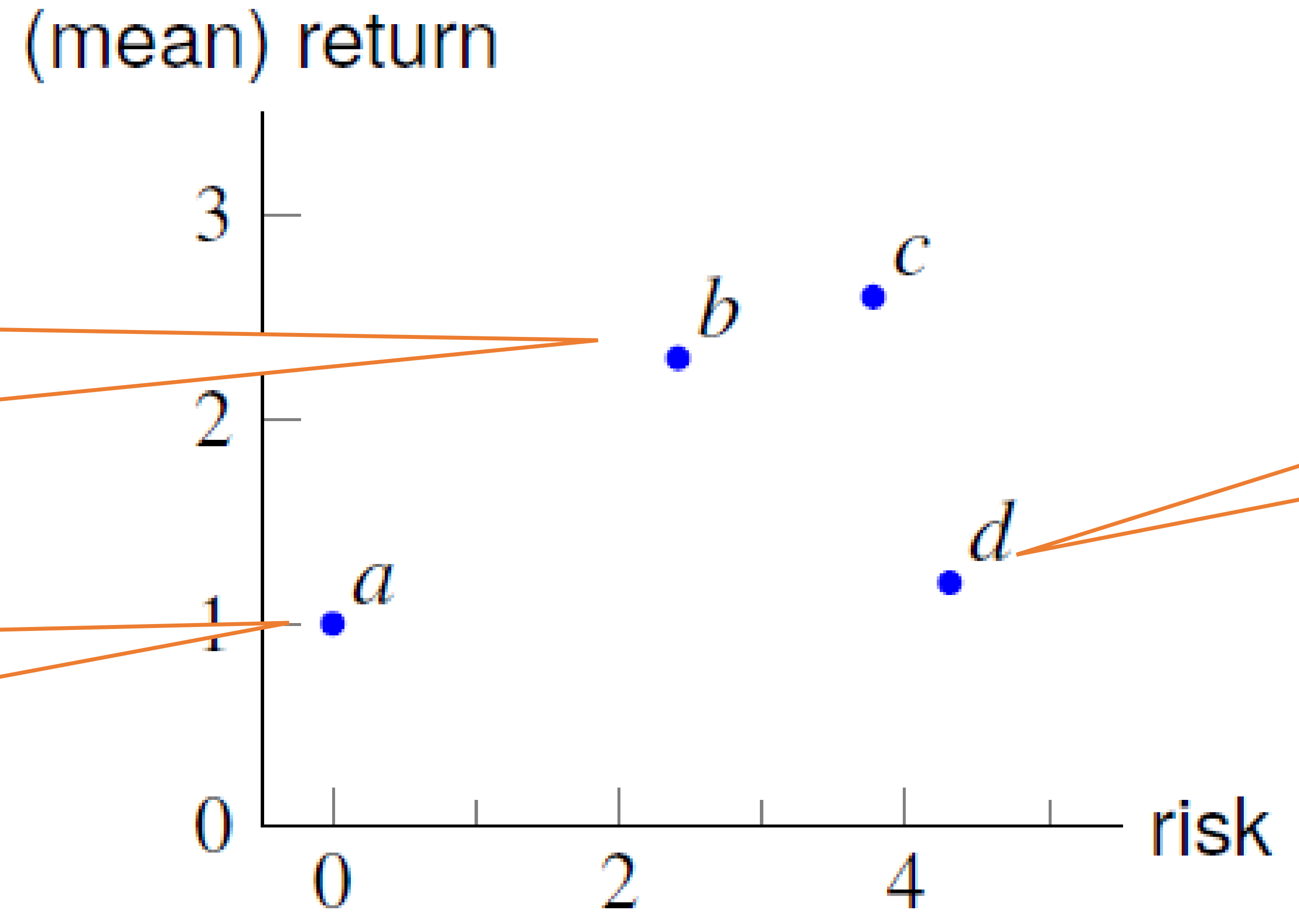


- THE right way to model dispersion for normally distributed phenomena
- Pythagorean analogy - Distance between two points

Standard deviation example: Risk-Return

- Returns over a 10 month period for 4 investments





Each point is mean return for that investment

Investment with 0 SD and 1% return

Is investment d worth the risk given the returns

Risk = fluctuation in return = SD

Chebyshev inequality for standard deviation

- Let vector x contain homogenous data
  - It’s entries are probability distribution
- Puts upper bound on fraction of entries away from mean by certain standard deviation
  - Provides guarantee, regardless of distribution

α

Chebyshev inequality for standard deviation

- Let k entries of a vector x away from mean > a

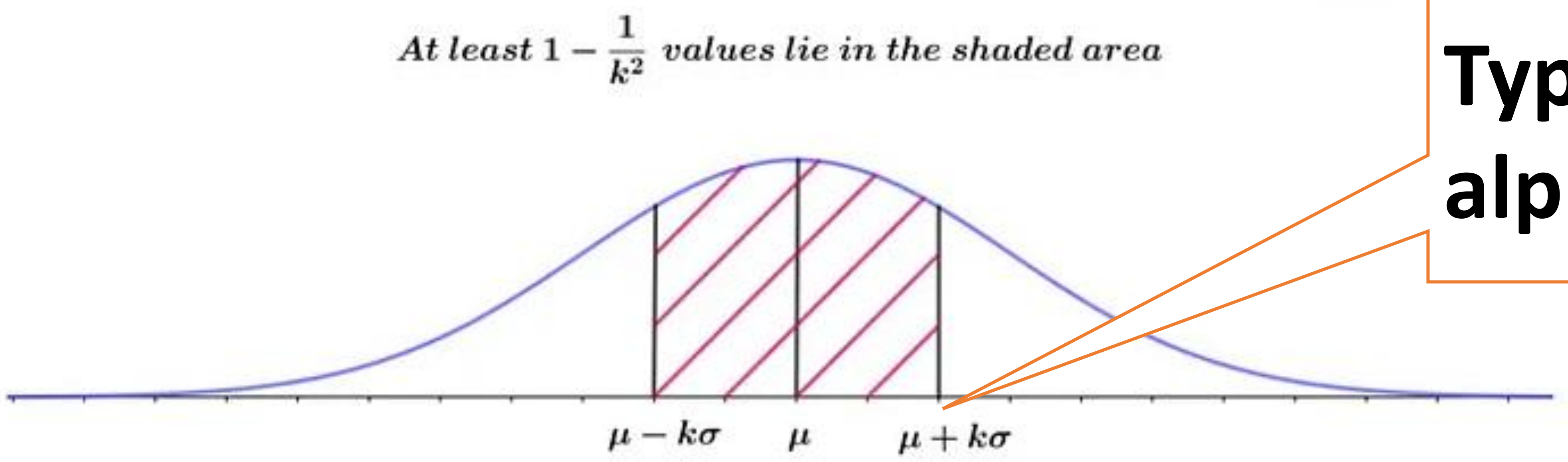
$$|x_i - \mu| > a \quad a = \alpha \times std(x) \quad \alpha > 1$$

- k is limited such that
  - Fraction of x entries (k/n) that can be at most α SD away from mean

$$\frac{k}{n} \leq \left(\frac{std(x)}{a}\right)^2 \implies \frac{k}{n} \leq \left(\frac{1}{\alpha}\right)^2 \implies 1 - \frac{k}{n} > 1 - \frac{1}{\alpha^2}$$

Chebyshev inequality for standard deviation

- Fraction of x entries within α standard deviations from mean is at least $1 - \frac{1}{\alpha^2}$ $\quad \alpha > 1$



Typo: It should be alpha and not k

- Chebyshev inequality is similar to 65-95-99.7 empirical rule of Gaussian distribution
- Chebyshev inequality is applicable to any distribution

Chebyshev inequality applications

- Upper limit of risk without knowing anything about underlying distribution
- Insurance company entering Indian market
  - 90% assured that future claims will be within 3 standard deviations
  - With data over time, company can fit a known distribution
  - If claims was Gaussian distribution, what percent of claims will be within 3 standard deviations?

Properties of Mean vs SD

- Mean / Standard Deviation

$$E[X + \alpha\mathbf{1}] = E[X] + \alpha\mathbf{1} \qquad SD[X + \alpha\mathbf{1}] = SD[X]$$

$$E[\beta X] = \beta E[X] \qquad SD[\beta X] = |\beta| SD[X]$$

SD is never negative (Check this mathematically & logically)

- X is a random vector (vector of random variables)
- x is a realized vector

Topics not covered but included for exam

- Relation between SD, RMS and mean

$$std(x)^2 = rms(x)^2 - avg(x)^2$$

- Textbook contains proof

- Time complexity of statistical vector operations (Read textbook)

Topics not covered but included for exam

- Cauchy Schwarz inequality $|a^T b| \leq \|a\| \|b\|$
- Textbook contains proof.

You already know this

$$a^T b = \|a\| \ \|b\| \ cos\theta$$

$$cos\theta = \left(\frac{a^T b}{\|a\| \ \|b\|} \right) \qquad \theta = arccos\left(\frac{a^T b}{\|a\| \ \|b\|} \right)$$


QUESTIONS

Literature Survey