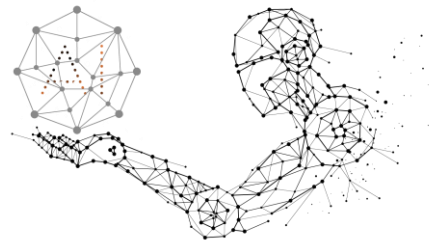
Applied Machine Learning

Chapter 2- A Little Statistics

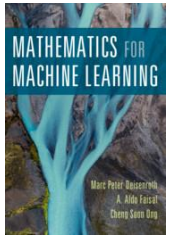
دانشگاه تربیت مدرس
Tarbiat Modares University

Hossein Homaei
Department of Electrical & Computer Engineering

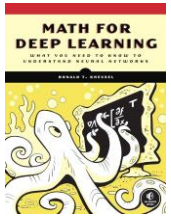


Some resources

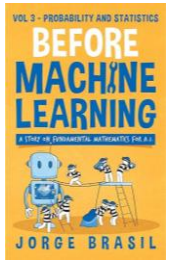
- Books



M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.



R. T. Kneusel, *Math for Deep Learning: What You Need to Know to Understand Neural Networks*. No Starch Press, 2022.



J. Brasil, *Before Machine Learning*, vol. 3, Probability And Statistics. 2024.

… Some resources

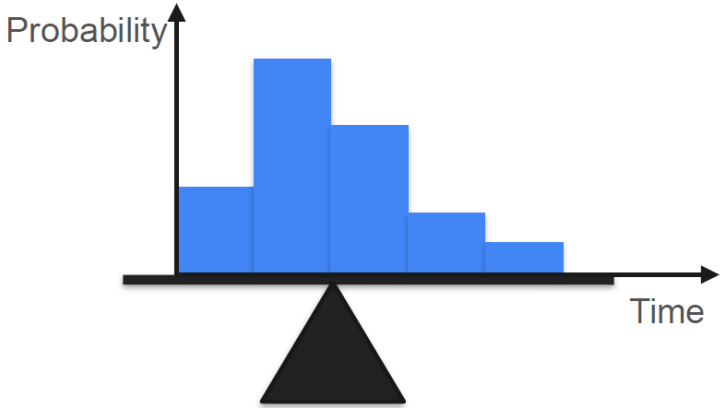
- Online
  - Probability & Statistics for Machine Learning and Data Science

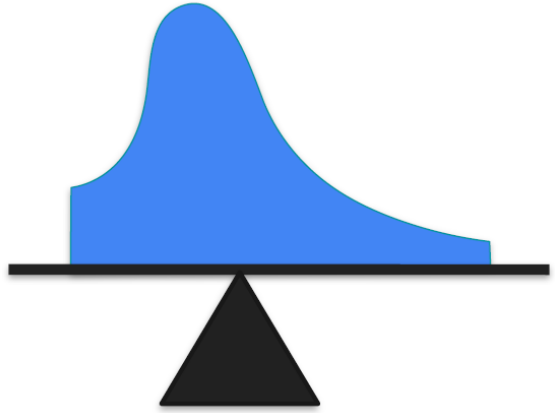
    - Instructor: Luis Serrano
    - DeepLearning.AI
    - **A significant portion of this lecture's content is sourced from this course.**
  - CS229: Machine Learning- The Summer Edition 2019

    - Instructor: Anand Avati
    - Stanford University

Center of the Data

- Mean
  - Sum of the values divided by the number of samples
  - Balance the data





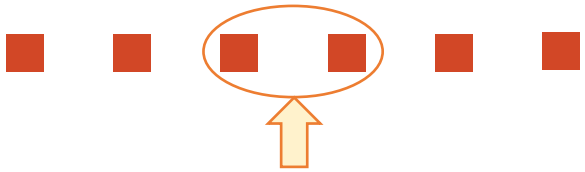
… Center of the Data

- Outlier problem for the mean

… Center of the Data

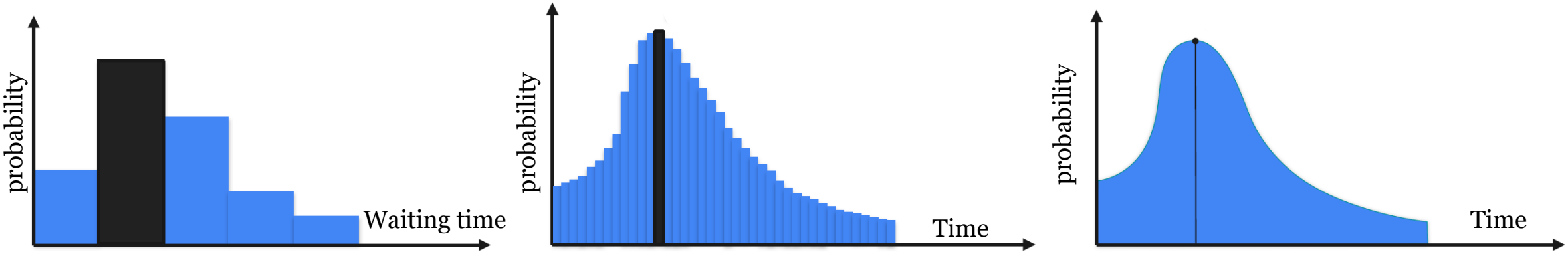
- Median
  - Sort the values and pick the middle

  - For an even number of samples: use the average of the two middle ones

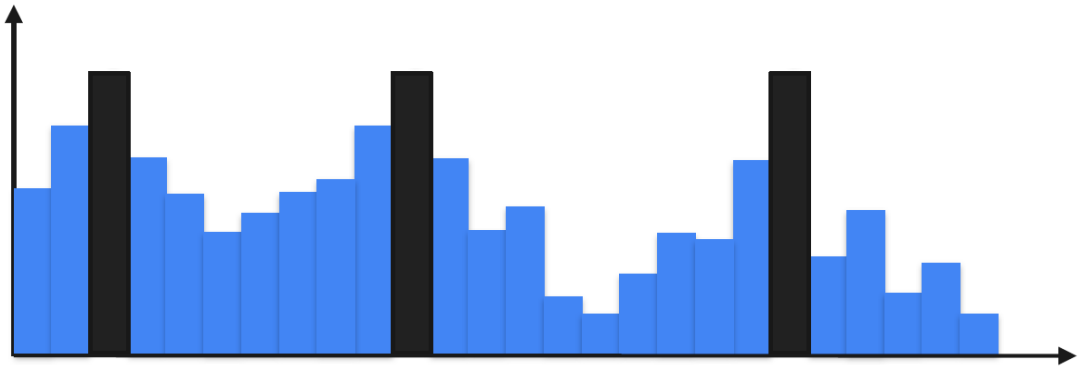


… Center of the Data

- Mode: The value with the highest probability

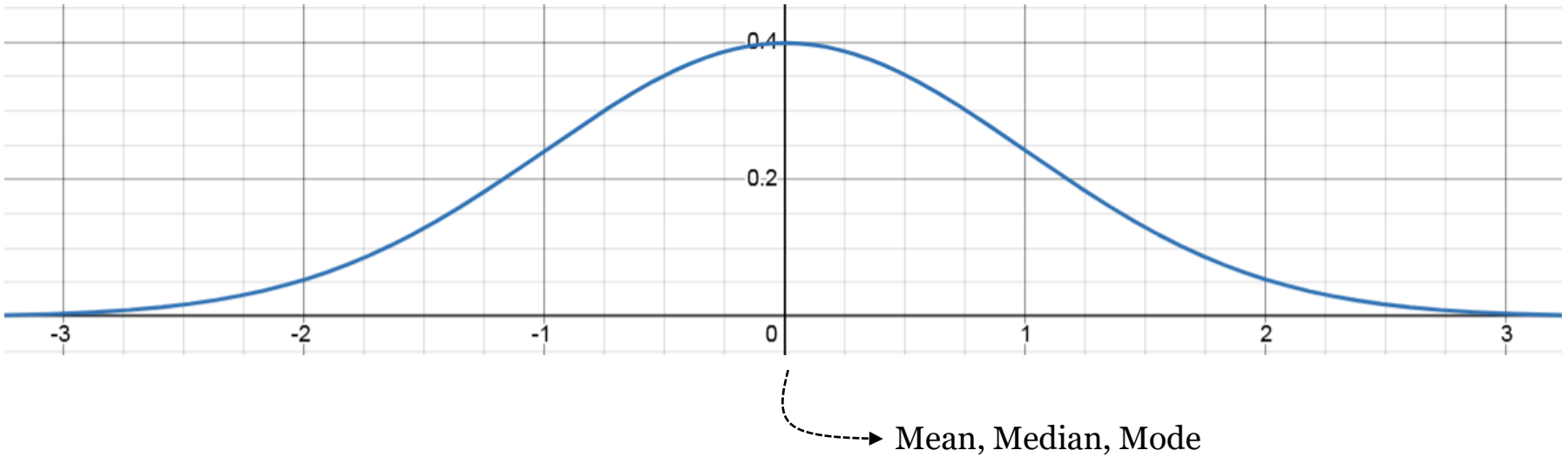


- Multimodal distribution



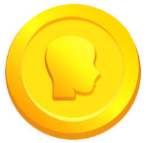
… Center of the Data

- Center of Normal distribution



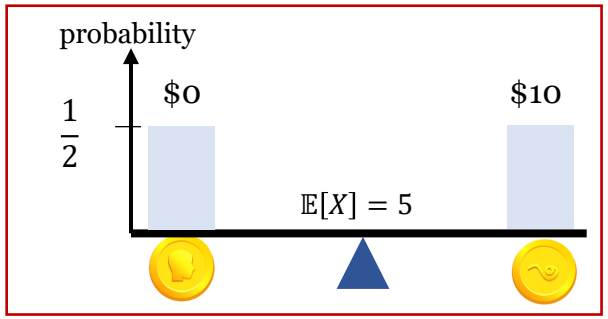
Expected Value

- If you have a probability distribution (any-discrete/continuous), and you are to draw a random sample from it, what do you expect this sample to be on average?
  - Example-1



You win \$10

You win nothing



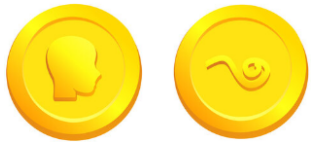
Game cost?

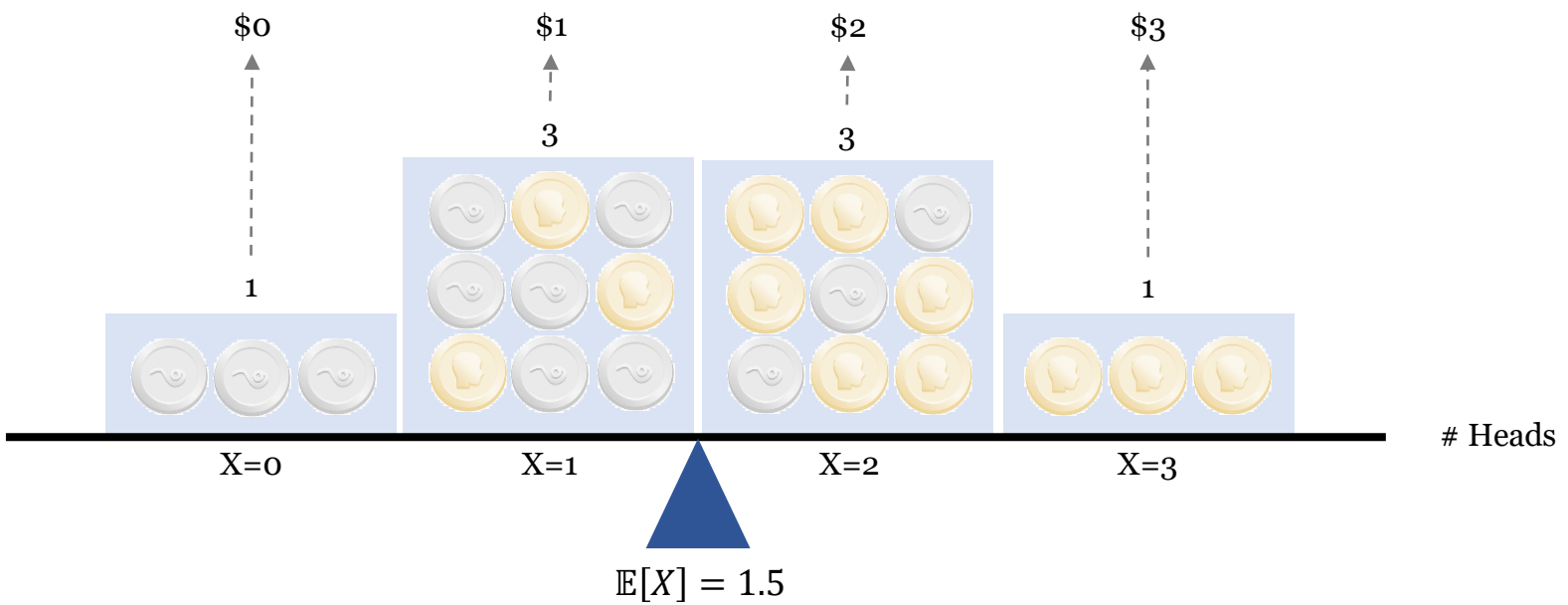
What is the fair price to play for the game?

Long term: $\frac{1}{2} \times \$10 + \frac{1}{2} \times \$0 = \$5 \Rightarrow$ You expect to win \$5 on average $\Rightarrow \mathbb{E}[X] = 5$

… Expected Value

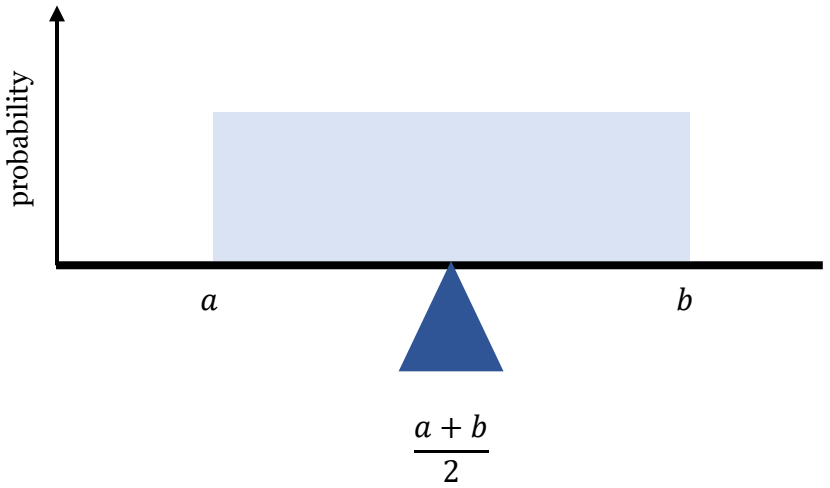
- Example-2
  - Flip 3 coins
  - For each head you win $1
  - Game cost?





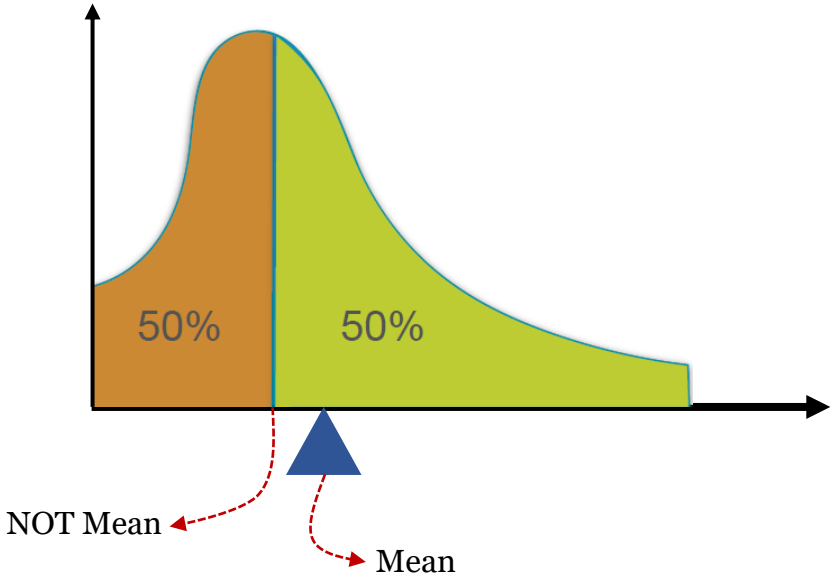
… Expected Value

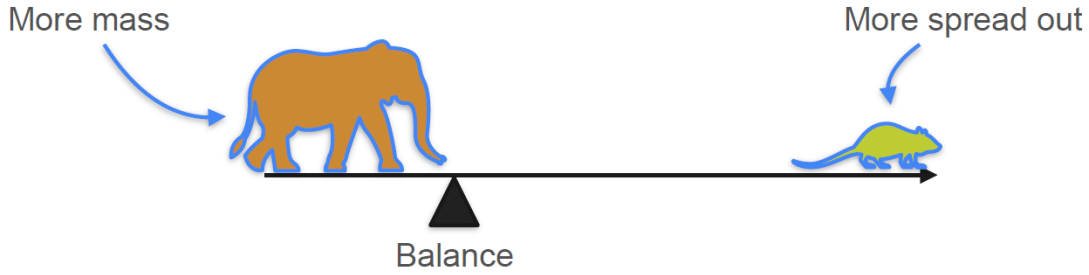
- Example-3
  - Uniform distribution



… Expected Value

- Example-4
  - Unbalanced distribution
  - Misconception: Mean is a place where the data is split in half



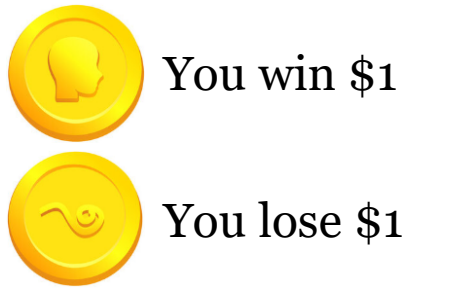


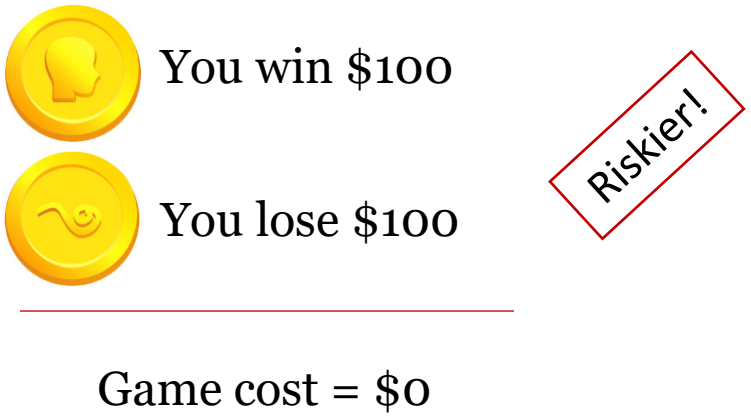
… Expected Value

- General formulation
  - $\mathbb{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \ldots + x_n p(x_n)$
- For a function $f(X)$
  - $\mathbb{E}\big(f(X)\big) = f(x_1)p(x_1) + f(x_2)p(x_2) + \ldots + f(x_n)p(x_n)$
  - Example: Roll a dice. Win $\$x^2$ if get x
    - Expected value of money
      - $\mathbb{E}(X^2) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = \frac{91}{6}$
- Sum of the expectation
  - $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$

Variance

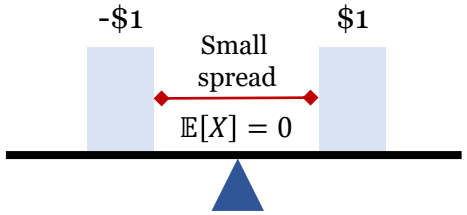
- Specifies how data is spread out
  - Two distributions may have the same expected value, but one of them can be very narrow and the other one can be very wide.
  - Example-1

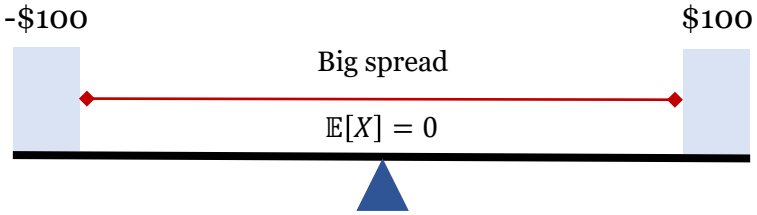




… Variance

- …Example-1





- Why this happens?
  - Positive and negative numbers cancel each others in the expected value
  - Solution: Use the square

$$\frac{1^2 + (-1)^2}{2} = 1$$

$$\frac{100^2 + (-100)^2}{2} = 10000$$

… Variance

- Only Squaring the value do not solve the problem
  - Example-2

- Solution
  - Center the data before squaring

$$\frac{1^2 + (-1)^2}{2} = 1 \qquad \frac{(-1)^2 + 1^2}{2} = 1$$

… Variance

- General formula

$$Var(X) = \mathbb{E}[(X-\mu)^2]$$
$$= \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2]$$
$$= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2$$
$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$
$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

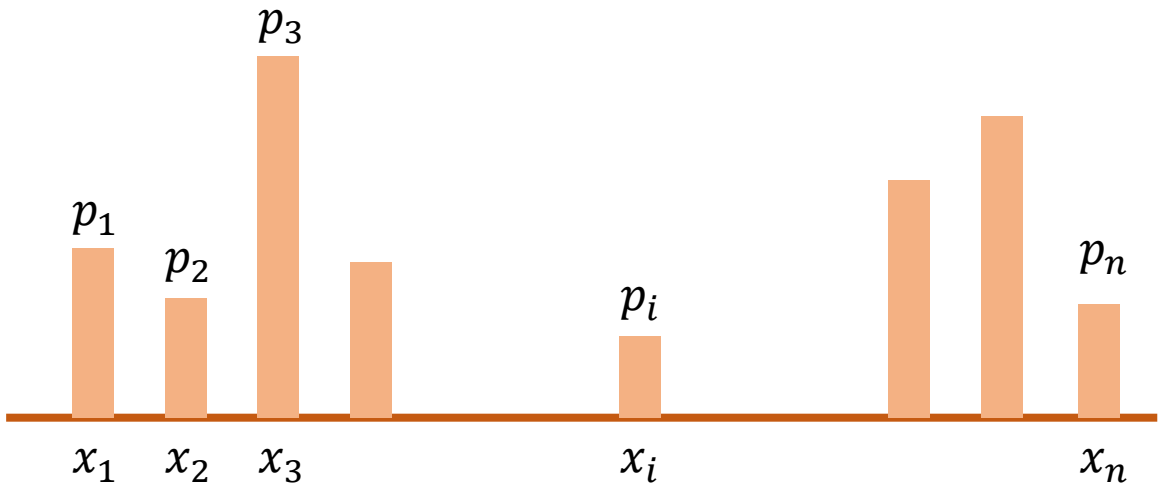
$$Var(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Standard Deviation

- Same as variance but preserves the units
  - $std(X) = \sqrt{Var(X)}$

Moments of a Distribution

- 1st moment
  - $\mathbb{E}[X] = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$
- 2nd moment
  - $\mathbb{E}[X^2] = p_1 x_1^2 + p_2 x_2^2 + \cdots + p_n x_n^2$
- 3rd moment
  - $\mathbb{E}[X^3] = p_1 x_1^3 + p_2 x_2^3 + \cdots + p_n x_n^3$
- …
- Kth moment
  - $\mathbb{E}[X^k] = p_1 x_1^k + p_2 x_2^k + \cdots + p_n x_n^k$



Skewness

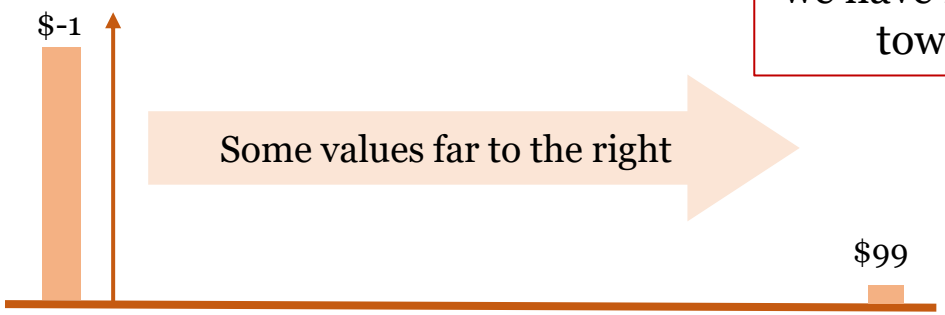
- Game 1
  - Win \$99 with 1% probability
  - Lose \$1 with 99% probability

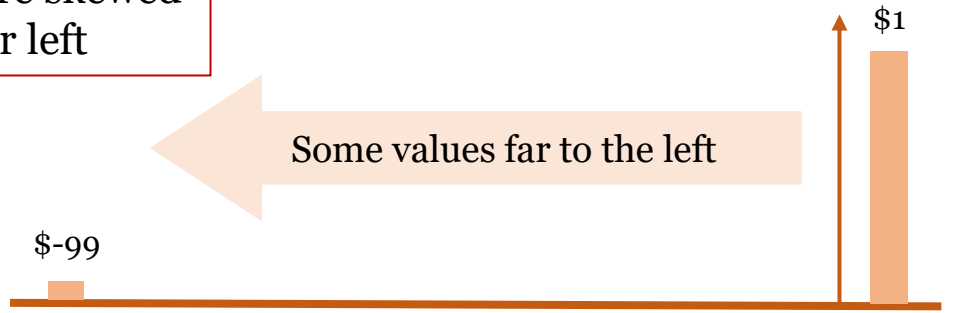
- Game 2
  - Win \$1 with 99% probability
  - Lose \$99 with 1% probability

- $\mathbb{E}[X_1] = -1 \times 0.99 + 99 \times 0.01 = 0$
- $Var(X_1) = (-1)^2 \times 0.99 + (99)^2 \times 0.01 = 99$
- $\mathbb{E}[X_1^3] = (-1)^3 \times 0.99 + (99)^3 \times 0.01 = 9702$

- $\mathbb{E}[X_2] = -99 \times 0.01 + 1 \times 0.99 = 0$
- $Var(X_2) = (-99)^2 \times 0.01 + (1)^2 \times 0.99 = 99$
- $\mathbb{E}[X_2^3] = (-99)^3 \times 0.01 + (1)^3 \times 0.99 = -97.2$

The cube of the variable detects if we have numbers that are skewed towards the right or left



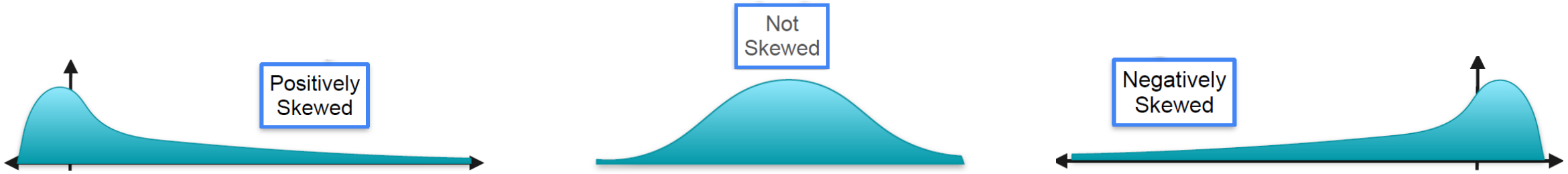


… Skewness

$\mathbb{E}[X_1^3]$ = large positive value

$\mathbb{E}[X_2^3]$ = Near zero value

$\mathbb{E}[X_3^3]$ = large negative value



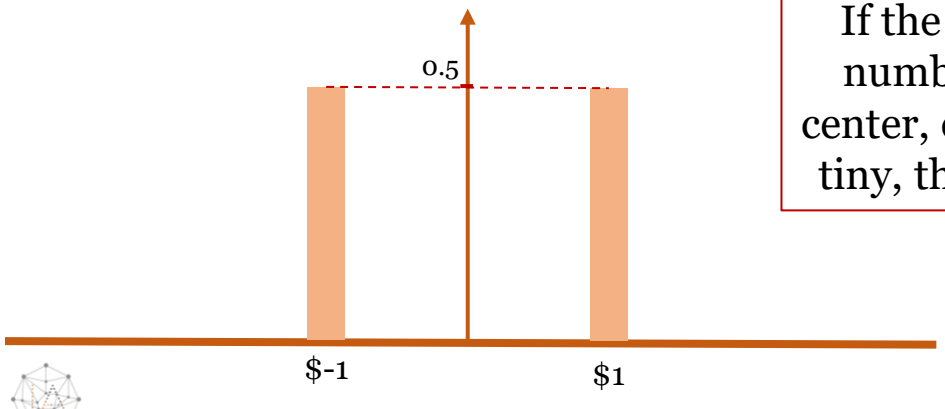
Standard form of the 3rd moment is called skewness

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

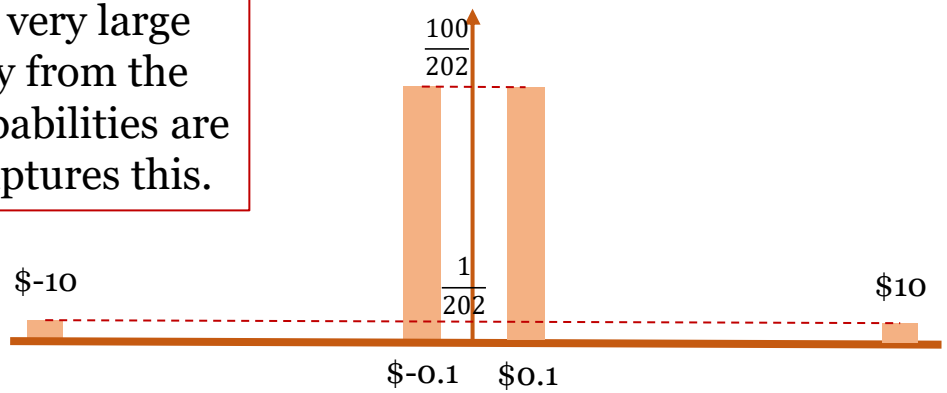
Kurtosis

- $\mathbb{E}[X_1] = \mu = 0$
- $\mathbb{E}[X_1^2] = 1$
  - Already centered- Var and std also equal 1
- $\mathbb{E}[X_1^3] = 0$
  - Already centered- skewness also equals 0
- $\mathbb{E}[X_1^4] = \frac{1}{2}(-1)^4 + \frac{1}{2}(1)^4 = 1$

- $\mathbb{E}[X_2] = \mu = 0$
- $\mathbb{E}[X_2^2] = 1$
  - Already centered- Var and std also equal 1
- $\mathbb{E}[X_2^3] = 0$
  - Already centered- skewness also equals 0
- $\mathbb{E}[X_2^4] = \frac{100}{202}(-0.1)^4 + \frac{100}{202}(0.1)^4 + \frac{1}{202}(-10)^4 + \frac{1}{202}(10)^4 = 99.01$



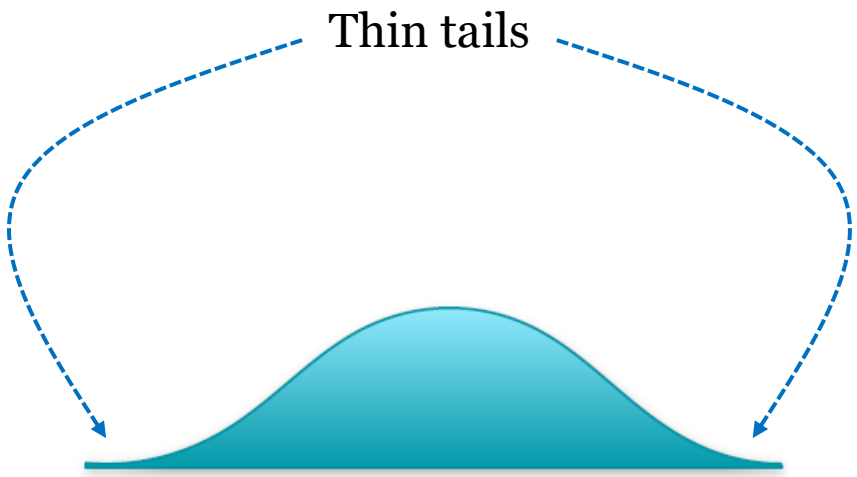
If the distribution has very large numbers very far away from the center, even if their probabilities are tiny, the 4th moment captures this.



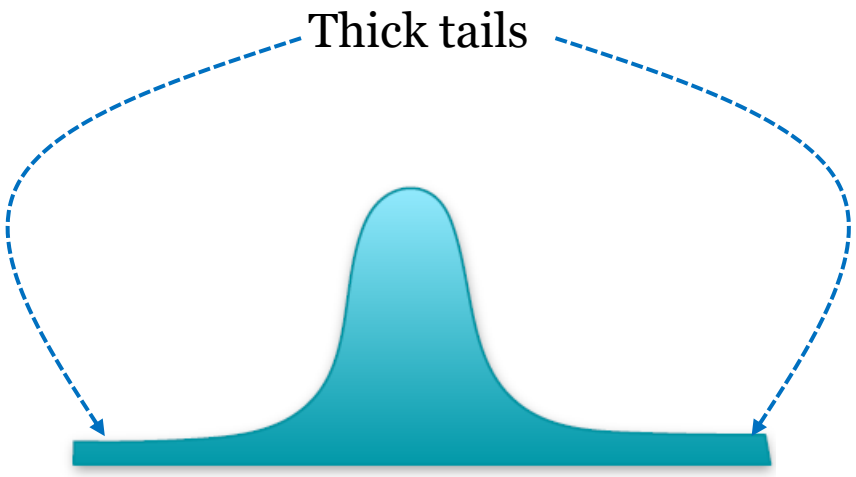
… Kurtosis

- Standard form of the 4^{th} moment is called kurtosis

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$



Small kurtosis



Large kurtosis

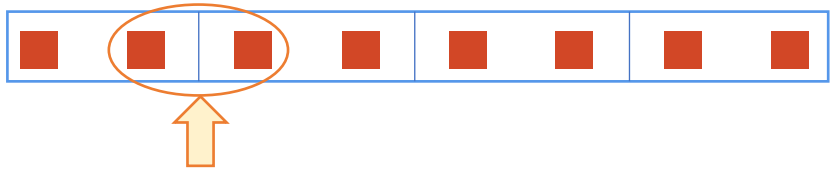
Quantiles

- We previously described the Median
  - Sort the data
  - Median = The middle value or the average of the two values in the middle

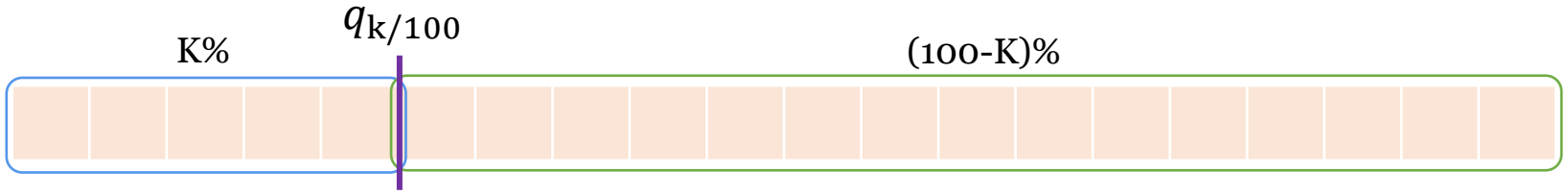
- Median is also called the 50% **quantile** or the 2nd **quartile**
  - The point that leaves 50% of the data to the left and 50% to the right

...Quantiles

- $q_{0.25}$ or $Q1$
  - The 1st quartile
  - 25% quantile
  - The point that leaves ¼ of the data to the left and ¾ to the right



- More general = $q_{k/100}$
  - K% quantile
  - The value that leaves K% of the data to the left and (100-K)% to the right.



…Quantiles

- Common quantiles:
  - 25% quantile = 1st quartile = Q1
  - 50% quantile = median = Q2
  - 75% quantile = 3rd quartile = Q3

- Interquartile range or **IQR**
  - $IQR = Q3 - Q1$

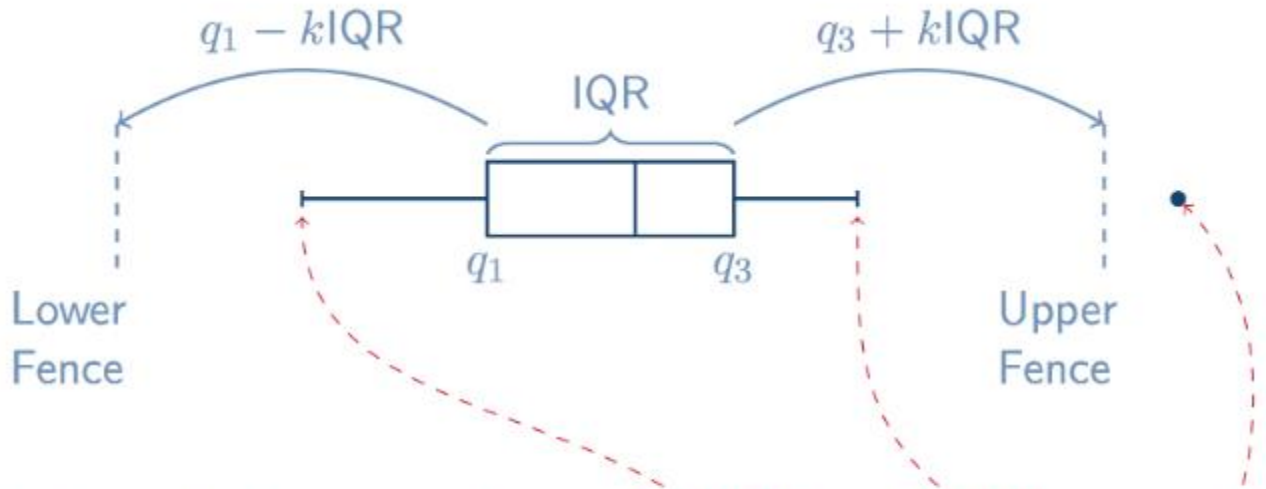
Box-Plots

- Standard way for displaying the distribution of the data based on:
  - The minimum
  - The maximum
  - The median
  - The 1st quartile
  - The 3rd quartile

…Box-Plots

- How?
  - Draw a box from Q1 to Q3
  - Draw a line at the value of the median (Q2)
  - Draw whiskers: Two lines going from the end of the box
    - From Q1 to the maximum of the $Q1 - 1.5 \times IQR$ and the Min point
      - $\max(Q1 - 1.5 \times IRQ, x_{min})$
    - From Q3 to the minimum of the $Q3 + 1.5 \times IQR$ and the Max point
      - $\min(Q3 + 1.5 \times IRQ, x_{max})$
  - Draw outlier data points
    - The points which are too far from the center
      - Bigger than $Q3 + 1.5 \times IQR$
      - Smaller than $Q1 - 1.5 \times IQR$

...Box-Plots



- The "whiskers" extend to the smallest and largest observations that are not outliers.
- Observations that are smaller than the lower fence or larger than the upper fence are identified as dots

...Box-Plots

- Example

8.7	14.2	18.3	18.4	23.2	25.9	29.7	35.2	51.2	54.7	65.9	75

  - $Min = 8.7$
  - $Max = 75$
  - $Q1 = \frac{18.3+18.4}{2} = 18.35$
  - $Q2 = \frac{25.9+29.7}{2} = 27.8$
  - $Q3 = \frac{51.2+54.7}{2} = 52.95$
  - $IQR = Q3 - Q1 = 52.95 - 18.35 = 34.6$
    - $1.5 \times IQR = 51.9$
    - $Q1 - 1.5 \times IRQ = 18.35 - 51.9 = -33.55 < 8.7$
    - $Q3 + 1.5 \times IRQ = 52.95 + 51.9 = 104.85 > 75$

…Box-Plots

- …Example

  - The observations by looking at the plot:
    - Data is skewed
      - $Q_3 - Q_2$ is bigger that $Q_2 - Q_1$
    - There is no outlier
      - Whiskers are cut at the max and min values

Covariance

- In a dataset consists of n samples and 2 columns x and y
  - Mean of samples

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  - Variance: Average squared distance from the mean

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

$$Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_y)^2$$

We will discuss about it later!

… Covariance

- Variance shows how the points spread out along the x or y axes
  - It cannot differentiate between the patterns in the following datasets

- Solution: covariance
  - How two features of a dataset varies with respect to one another

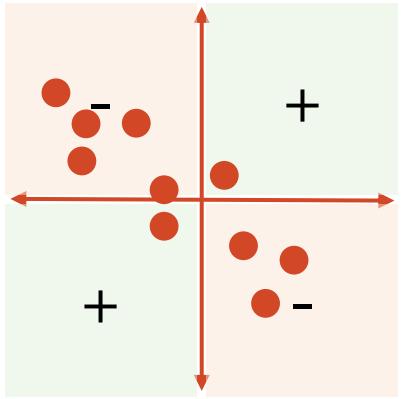
Negative covariance

Positive covariance

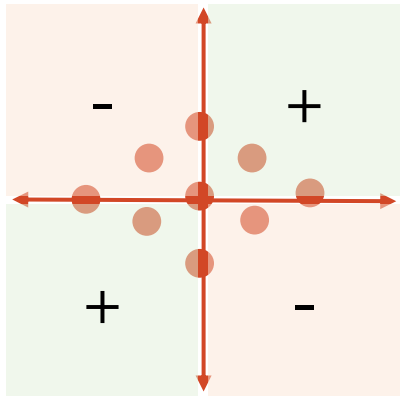
… Covariance

- Covariance formula

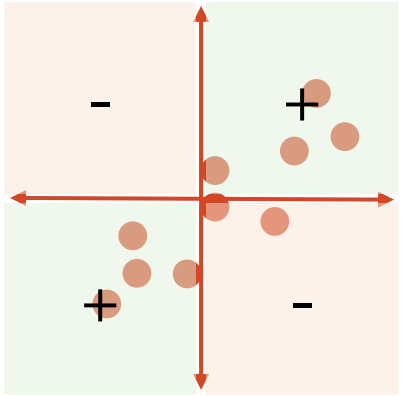
$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$



Negative covariance



Zero or small covariance



Positive covariance

Covariance Matrix

- Covariance matrix for a dataset with two variables (x, y):

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$$

- OR

$$C = \begin{bmatrix} Var(x) & cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix}$$

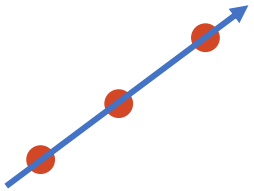
...Covariance Matrix

- For more than two variables:

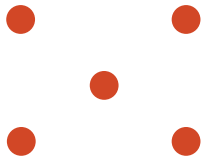
$$\Sigma = \begin{array}{c|ccccc} & V & W & X & Y & Z \\ \hline V & Var(V) & Cov(V,W) & Cov(V,X) & Cov(V,Y) & Cov(V,Z) \\ W & Cov(V,W) & Var(W) & Cov(W,X) & Cov(W,Y) & Cov(W,Z) \\ X & Cov(V,X) & Cov(W,X) & Var(X) & Cov(X,Y) & Cov(X,Z) \\ Y & Cov(V,Y) & Cov(W,Y) & Cov(X,Y) & Var(Y) & Cov(Y,Z) \\ Z & Cov(V,Z) & Cov(W,Z) & Cov(X,Z) & Cov(Y,Z) & Var(Z) \end{array}$$

Correlation Coefficient

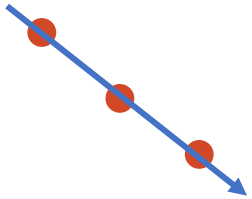
- Covariance shows how the variables are correlated



Positive covariance



zero covariance



Negative covariance

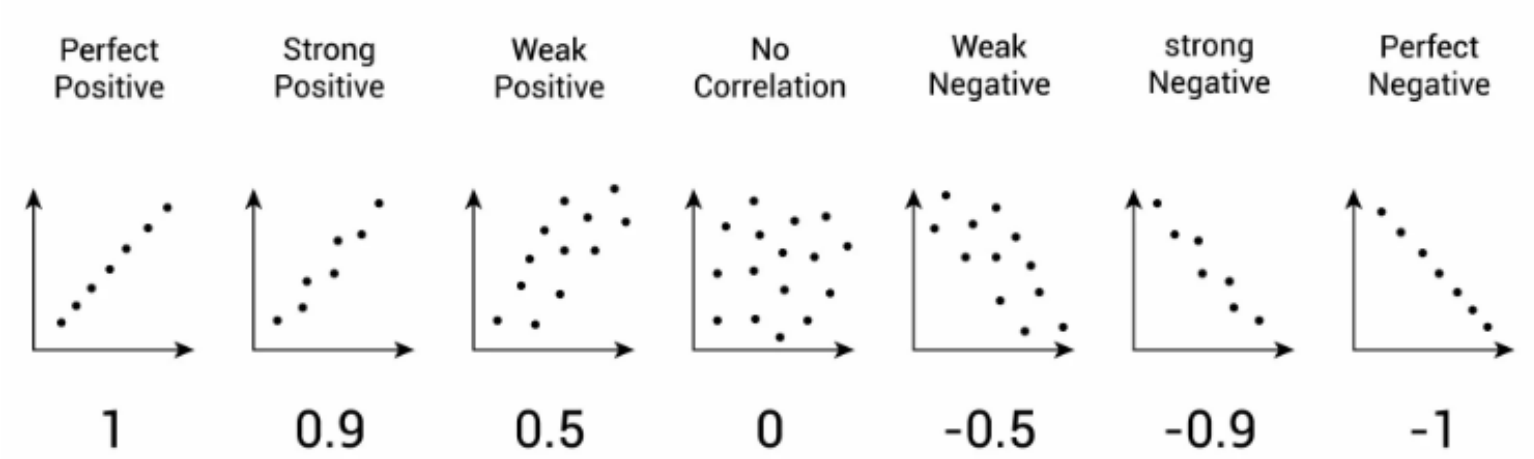
- But we cannot compare covariances magnitude because they use different metrics
  - Example:
    - Cov(Age, Naps-per-day) = -7.45
    - Cov(Age, Height) = 17
    - Which one is stronger?
      - We do not know

… Correlation Coefficient

- Normalizing the covariance make it possible to compare between different covariances

- Formula
  - Correlation Coefficient = $\frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{Cov(X,Y)}{\sqrt{var(x)}\sqrt{var(y)}}$
  - Always between -1 and 1

… Correlation Coefficient



[Ref] https://www.simplypsychology.org/correlation.html

Sampling

Population & Sample

- Population
  - The entire group of individuals or elements you want to study which share a common behavior
  - Size: N
- Sample
  - Subset of the population you use to draw conclusions about the population as a whole
  - Size: n
- Every dataset you work with in machine learning is a sample

Example

- Find the average height of the people living in Tehran
  - Population: all the people living in Tehran
  - Ask everyone for their height, sum them up, and divide by the total number
  - Size: 15,000,000
  - Difficult

- Alternative approach ➔ Only ask a subset (sample) of the people living in Tehran to estimate the average height
  - Sample: the people you select for your study
  - Ask the selected group for their height, sum them up, and divide by the sample size
  - Size: between 1 to 14,999,999
  - Easier

Sampling Requirements

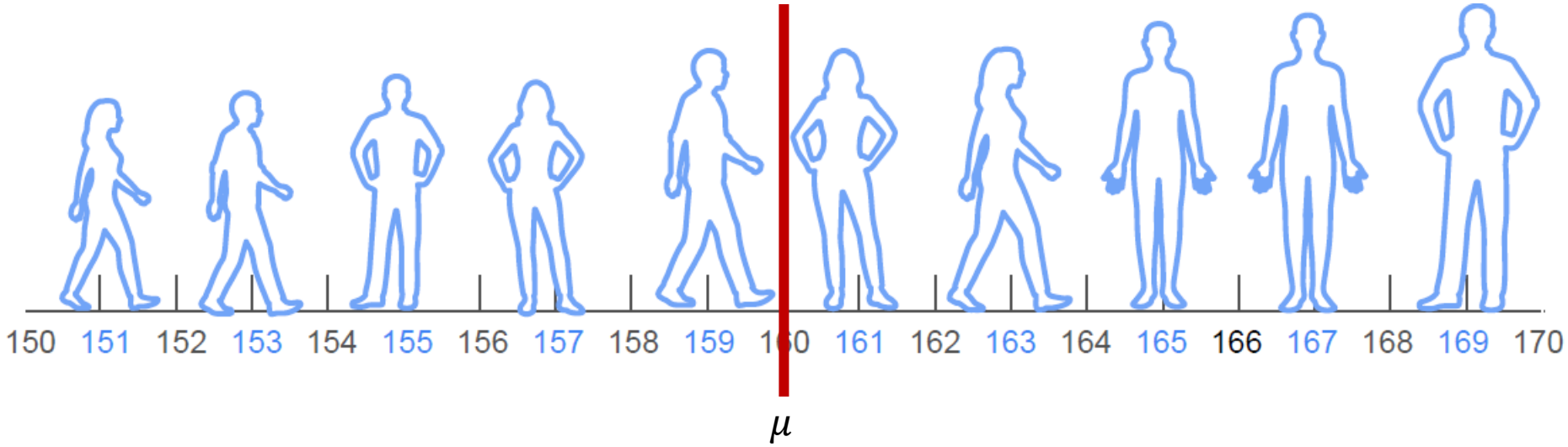
- Random sampling
  - The subset must be selected randomly

- Independent sampling
  - Each sampling must be independent to the other sampling
  - For example, if you select someone in the 1^{st} sampling, it is also possible to select him in the 2^{nd}, 3^{rd}, 4^{th}, and n^{th} sampling too (People are allowed to be repeated). Otherwise, other samples depend on the 1^{st} one.

- Identically distributed sampling
  - Select samples from the entire population
  - For example, do not select the sample from a particular part of the city where people are taller or shorter.

Sample Mean

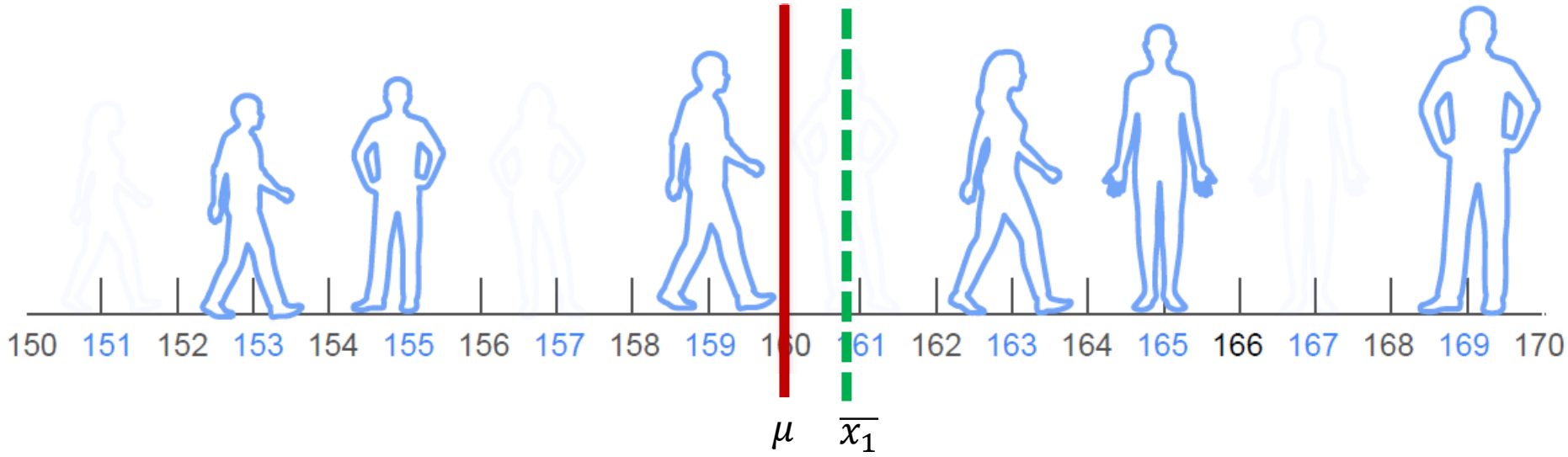
- What is the average height for the following population?



$$\mu = \frac{151 + 153 + 155 + 157 + 159 + 161 + 163 + 165 + 167 + 169}{10} = \frac{1600}{10} = 160$$

… Sample Mean

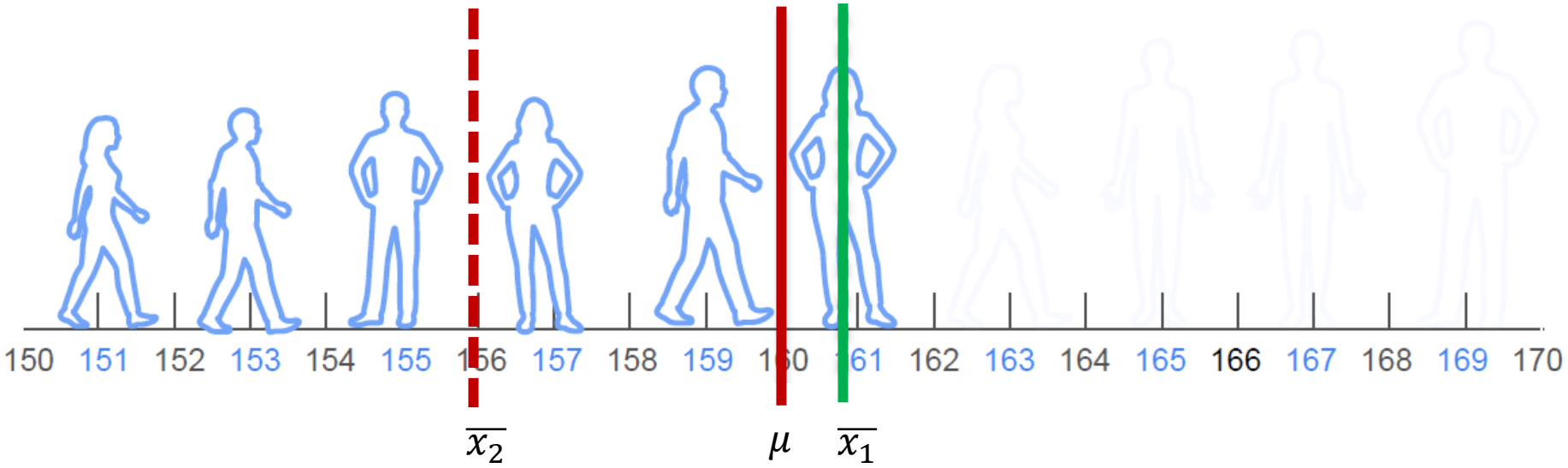
- What is the average height for the following sample?



$$\bar{x} = \frac{153 + 155 + 159 + 163 + 165 + 169}{6} = \frac{964}{6} = 160.97$$

… Sample Mean

- What is the average height for the following sample?

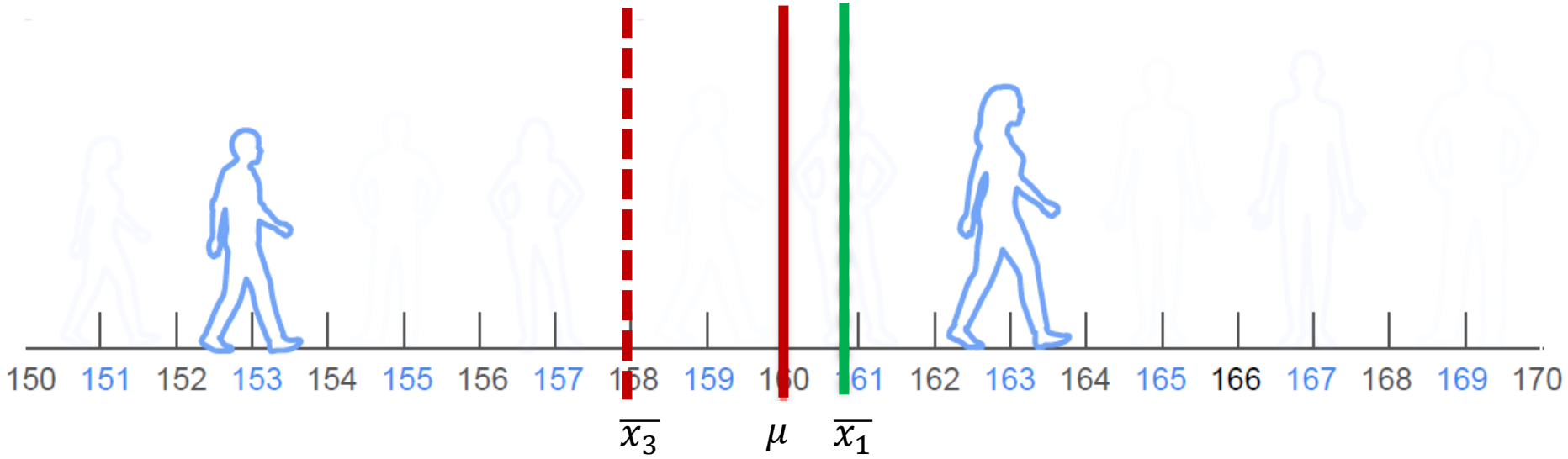


$$\bar{x} = \frac{151 + 153 + 155 + 157 + 159 + 161}{6} = \frac{936}{6} = 156$$

Previous estimation was better! Because of non-random sampling.

… Sample Mean

- What is the average height for the following sample?

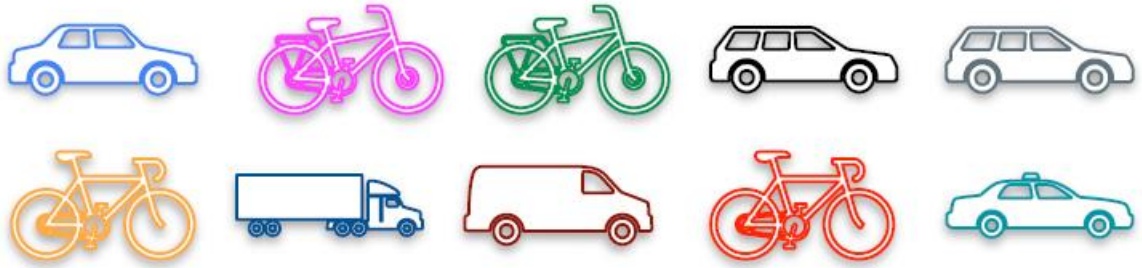


$$\bar{x} = \frac{153 + 163}{2} = \frac{316}{2} = 158 \longrightarrow \overline{x_1} \text{ estimation was better!}$$

Law of large numbers
As the sample size increases, the average of the sample will tend to get closer to the average of the entire population

Sample Proportion

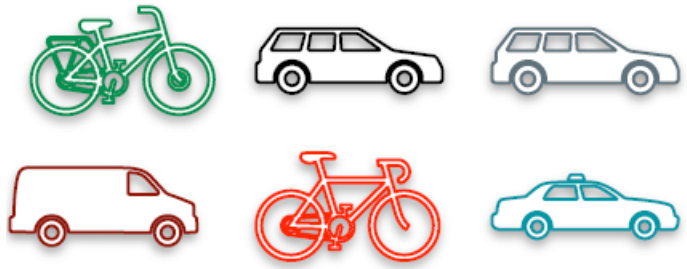
- What proportion of people own a bicycle in the following population?



$$\text{Population proportion } p = \frac{4}{10} = 0.4 = 40\%$$

… Sample Proportion

- What proportion of people own a bicycle in the following sample?



$$\text{Sample proportion } \hat{p} = \frac{2}{6} = 0.333 = 33.3\%$$

- Estimate the population proportion

Sample Variance

- Population variance

$$\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$$

Population size — Population mean

- Sample variance

$$Var(x) = S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

Sample size — Sample mean

… Sample Variance

- Why $n-1$?

- Example- Three-sided dice
  - Population variance
    - All possible outcomes: 1, 2, 3
    - $\mu = \frac{1}{3} \times 1 + \frac{1}{3} \times 2 + \frac{1}{3} \times 3 = \frac{1+2+3}{3} = 2$
    - $\sigma^2 = \frac{1}{N}\sum(x-\mu)^2 = \frac{1}{3}\left((1-2)^2 + (2-2)^2 + (3-2)^2\right) = \frac{2}{3}$

… Sample Variance

- … Example- Three-sided dice
  - Sample variance
    - Flip the dice twice (Sample size = 2) and estimate the variance by averaging on the variance of all possible combinations of the outcomes

Samples		$\bar{x}$	$\frac{1}{n}\sum(x-\bar{x})^2$
1	1	1	0
1	2	1.5	0.25
1	3	2	1
2	1	1.5	0.25
2	2	2	0
2	3	2.5	0.25
3	1	2	1
3	2	2.5	0.25
3	3	3	0

$Average = 2$ ✓

$Average = \frac{1}{3} \neq \frac{2}{3} = \sigma^2$ ✗

… Sample Variance

- … Example- Three-sided dice
  - Sample variance
    - Use $n-1$ instead of n

Samples		$\bar{x}$	$\frac{1}{n-1}\sum(x-\bar{x})^2$
1	1	1	0
1	2	1.5	0.5
1	3	2	2
2	1	1.5	0.5
2	2	2	0
2	3	2.5	0.5
3	1	2	2
3	2	2.5	0.5
3	3	3	0

$$Average = \frac{2}{3} = \sigma^2$$

… Sample Variance

- Mathematical proof
  - We should prove that $\mathbb{E}[S^2] = Var(X)$

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n-1}\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum(x_i - \mu + \mu - \bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum\left((x_i - \mu)^2 + 2(x_i - \mu)(\mu - \bar{x}) + (\mu - \bar{x})^2\right)\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i - \mu)^2 + \sum 2(x_i - \mu)(\mu - \bar{x}) + \sum(\mu - \bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i - \mu)^2 + 2(\mu - \bar{x})\sum(x_i - \mu) + \sum(\mu - \bar{x})^2\right]$$

… Sample Variance

- … Mathematical proof

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 + 2(\mu-\bar{x})\sum(x_i-\mu) + \sum(\mu-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 + 2(\mu-\bar{x})\left[\sum x_i - n\mu\right] + \sum(\mu-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 + 2n(\mu-\bar{x})\left[\frac{1}{n}\sum x_i - \mu\right] + \sum(\mu-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 + 2n(\mu-\bar{x})(\bar{x}-\mu) + \sum(\mu-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 - 2n(\bar{x}-\mu)^2 + \sum(\mu-\bar{x})^2\right]$$

… Sample Variance

- … Mathematical proof

$$= \frac{1}{n-1}\mathbb{E}\left[\sum(x_i-\mu)^2 - 2n(\bar{x}-\mu)^2 + \sum(\mu-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\left(\mathbb{E}\left[\sum(x_i-\mu)^2\right] - 2n\mathbb{E}[(\bar{x}-\mu)^2] + \mathbb{E}\left[\sum(\mu-\bar{x})^2\right]\right)$$

$$= \frac{1}{n-1}\left(\sum\mathbb{E}[(x_i-\mu)^2] - 2n\mathbb{E}[(\bar{x}-\mu)^2] + \sum\mathbb{E}[(\mu-\bar{x})^2]\right)$$

$$Var(X) = \mathbb{E}[(X-\mu)^2]$$

$$= \frac{1}{n-1}\left(nVar(X) - 2nVar(\bar{x}) + nVar(\bar{x})\right)$$

$$= \frac{1}{n-1}\left(nVar(X) - nVar(\bar{x})\right) = \frac{1}{n-1}\left(nVar(X) - n\frac{Var(X)}{n}\right) = Var(X)$$

$$Var(\bar{x}) = \frac{Var(X)}{n} \; ?$$

$$\Longrightarrow \mathbb{E}[S^2] = Var(X), \qquad S^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$$