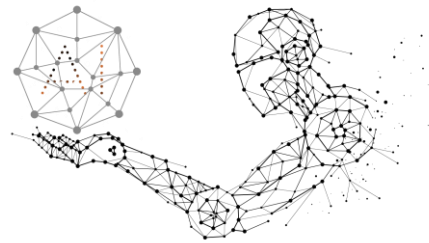
Applied Machine Learning

Lecture 6- A little Probability

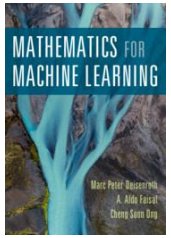
دانشگاه تربیت مدرس
Tarbiat Modares University

Hossein Homaei
Department of Electrical & Computer Engineering

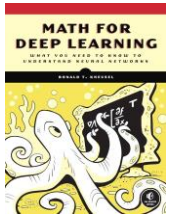


Some resources

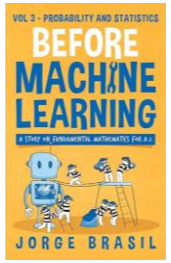
- Books



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

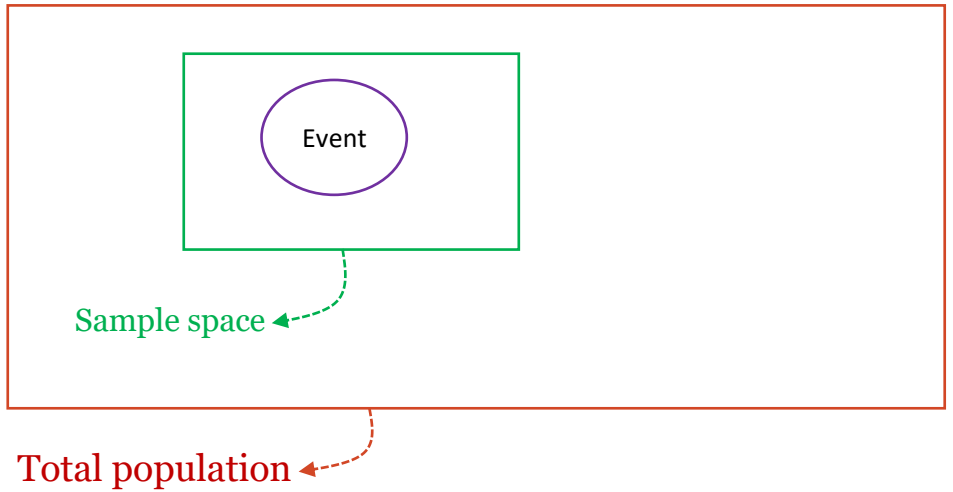
    - Instructor: Luis Serrano
    - DeepLearning.AI
    - **A significant portion of this lecture's content is sourced from this course.**
  - CS229: Machine Learning- The Summer Edition 2019

    - Instructor: Anand Avati
    - Stanford University

Introduction

- Probability
  - A measure of how likely an event is to occur
  - $p = \frac{size\ of\ the\ event}{size\ of\ sample\ space}$

- Example
  - All children = population
  - Children of a school = sample space = 100
  - Event = children play soccer = 40
  - $p = \frac{40}{100} = 0.4$

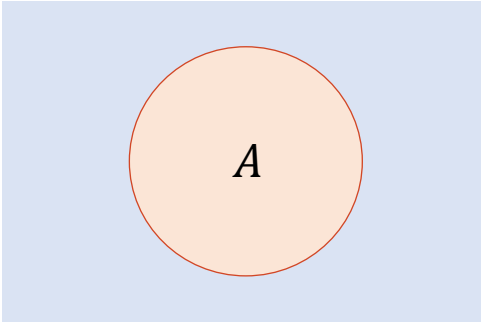


… Introduction

- Experiment
  - Any process that produces an outcome that is uncertain
    - Example: throwing a coin
  - $p = \frac{size\ of\ the\ event}{total\ number\ of\ outcomes}$
    - Examples
      - Roll a fair dice: The probability of obtaining 6= $p(6) = \frac{1}{6}$
      - Flip a fair coin twice: The probability of landing on heads twice= $p(HH) = \frac{1}{4} = 0.25$

Probability Rules

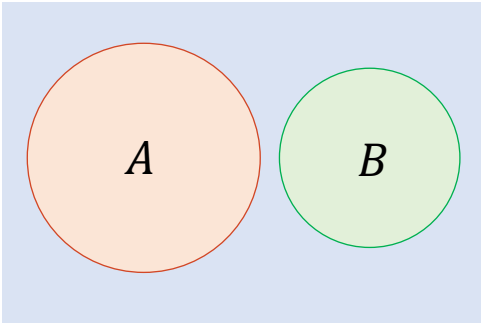
- The complement rule
  - $p(\bar{A}) = 1 - P(A)$
    - Roll a dice. Probability of not obtaining 6?
    - $P(not\ 6) = 1 - p(6) = 1 - \frac{1}{6} = \frac{5}{6}$



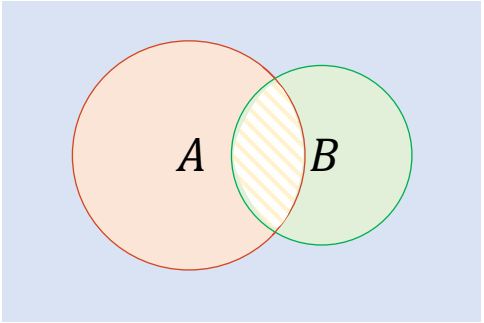
Venn Diagram

… Probability Rules

- Sum of probabilities
  - For disjoint events A and B, $P(A \cup B) = P(A) + P(B)$
    - Roll a dice. Probability of obtaining an even number or 5?
      - $P(even\ OR\ 5) = P(even \cup 5) = \frac{3}{6} + \frac{1}{6} = \frac{2}{3}$



  - For joint events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - Roll 2 dices. Probability of obtaining a sum of 7 or a difference of 1?
      - Sum=7: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)
      - Diff=1: (1, 2), (2, 1), (2, 3), (3, 2), (3, 4), (4, 3), (4, 5), (5, 4), (5, 6), (6, 5)
      - Sum=7 and diff=1: (3, 4), (4, 3)
      - $P(sum7 \cup diff1) = \frac{6}{36} + \frac{10}{36} - \frac{2}{36} = \frac{14}{36} = \frac{7}{18}$



… Probability Rules

- Product rule
  - For independent events A and B, $P(A \cap B) = P(A) \times P(B)$
    - Flip a fair coin 5 times. Probability of landing in heads 5 times?
    - $P(5\ heads) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$
  - Conditional probability: For dependent events A and B, $P(A \cap B) = P(A) \times P(B|A)$

The probability of B given A

  - Roll a dice twice. Probability that the first is 6 and the sum equals 10?
    - $P(first\ 6) = \frac{6}{36} = \frac{1}{6}$
    - $P(sum = 10 | first\ 6) = \frac{1}{6}$
    - $P(first\ 6 \cap sum = 10) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Discrete Distributions

Random Variables

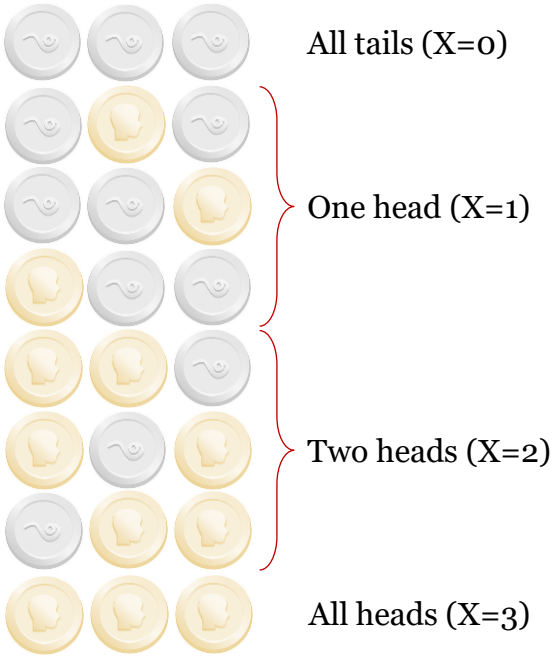
- A quantity that depends on random events (uncertain outcomes)

- Types
  - Discrete random variables: can take only a countable number of values
    - X = number of heads in 10 coin tosses
    - X = number of 1's in 7 dice rolling

  - Continuous random variables: can take values on an interval
    - Wait time until the next bus arrives
    - Quantity of rain (mm)

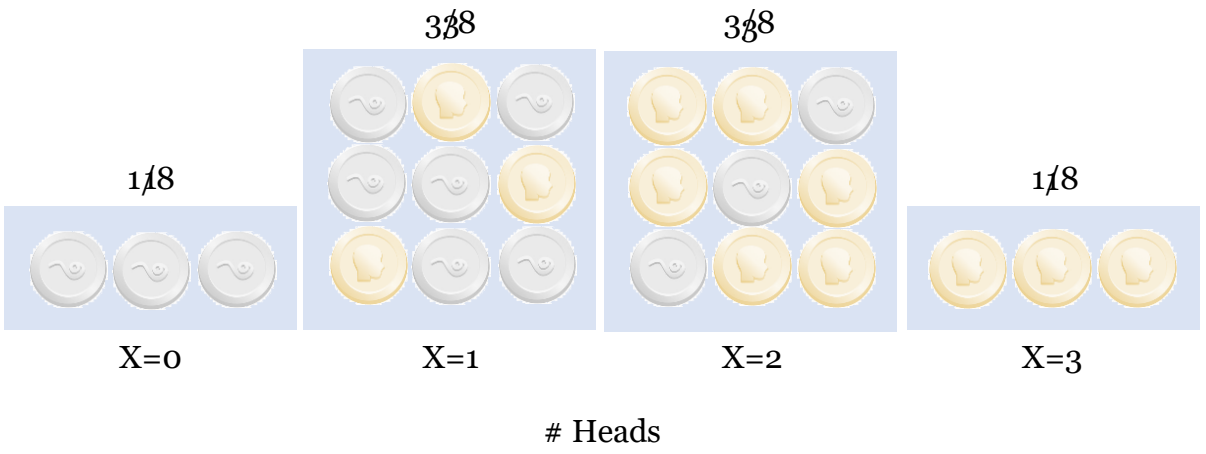
Probability Distribution

- Put all possible scenarios that can happen in the horizontal axis, and for each of them, specify the probability that it happens.
  - Example
    - Flip a fair coin 3 times
    - Random variable (X) = number of heads



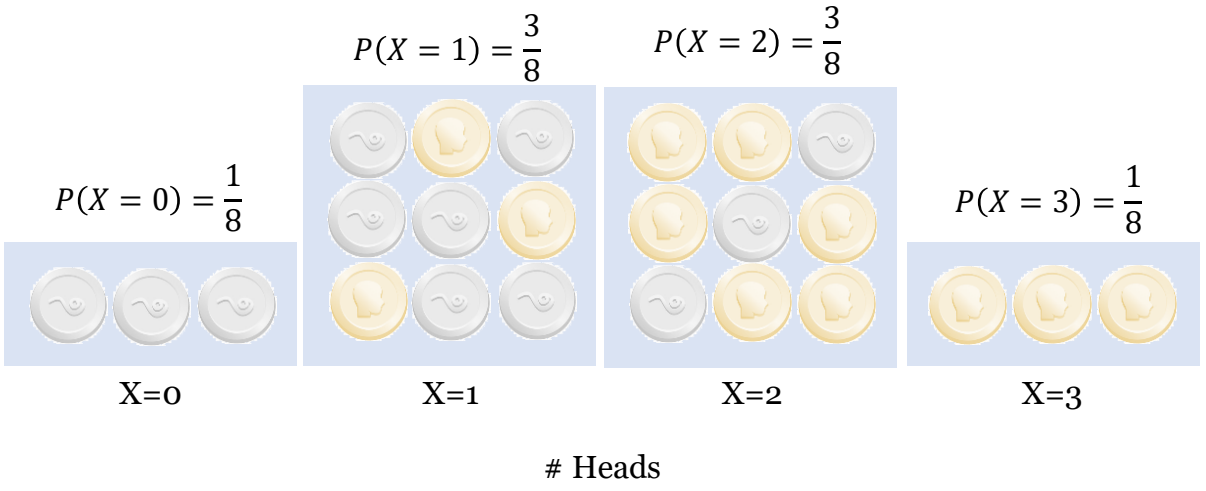
Probability Distribution

- Put all possible scenarios that can happen in the horizontal axis, and for each of them, specify the probability that it happens.
  - Example
    - Flip a fair coin 3 times
    - Random variable (X) = number of heads



… Probability Distribution

- Histogram interpretation
  - For each x from 0 to 3 (all possible values for random variable X), the probability that X is x



Probability Mass Function (PMF)

- What is it?
  - The function that represents how the probability distributes among all the possible values of the discrete random variable
- Notes
  - Use lowercase p to represent it.
    - $p_X(x) = P(X = x)$
  - All discrete random variables can be modeled by their PMF
- Properties
  - $p_X(x) \geq 0$
  - $\sum_x p_X(x) = 1$

Binomial Distribution

- Example: Flip 5 coins
  - What is the probability that 2 of them land in head?
    - For each flip, there is a ½ chance to get heads or tails
      - For the fair coins, each of the outcomes occurs with the same probability 1/32
    - There are 10 possibilities to get two heads out of 5.
    - Is there any general way to find the number of all possible combinations
      - Yes

$$10 = \frac{\text{Number of ways you can order 5 coins}}{\text{Number of ways you can swap 2 heads} \times \text{Number of ways you can swap 3 tails}}$$

$$= \frac{5!}{2!\,(5-2)!} = \binom{5}{2}$$

Binomial coefficient

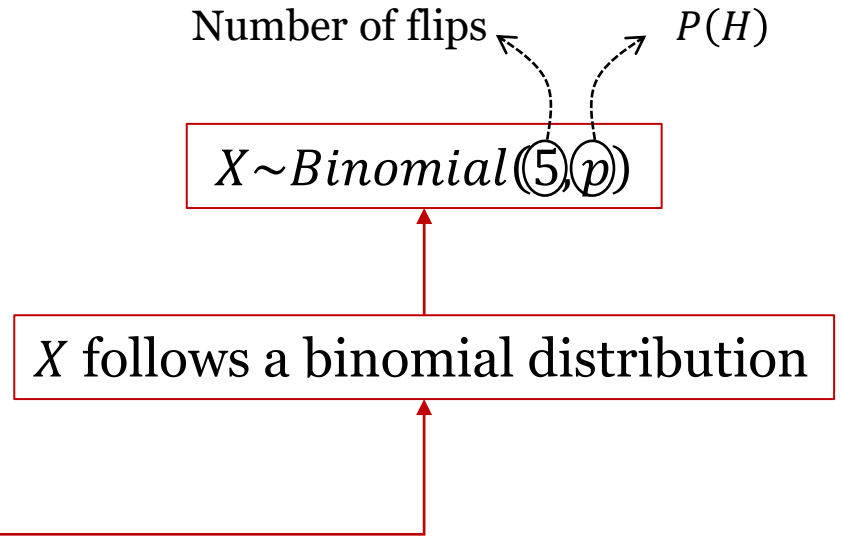
Number of ways you can get 2 heads in 5 coin tosses

… Binomial Distribution

- Binomial coefficient
  - $\binom{n}{k}$
    - read as "n choose k"
    - Counts all the combinations in which you can have k heads in n coin tosses
  - Property
    - $\binom{n}{k} = \binom{n}{n-k}$
      - This is the reason that the PMF of tossing a fair coin has a symmetrical shape

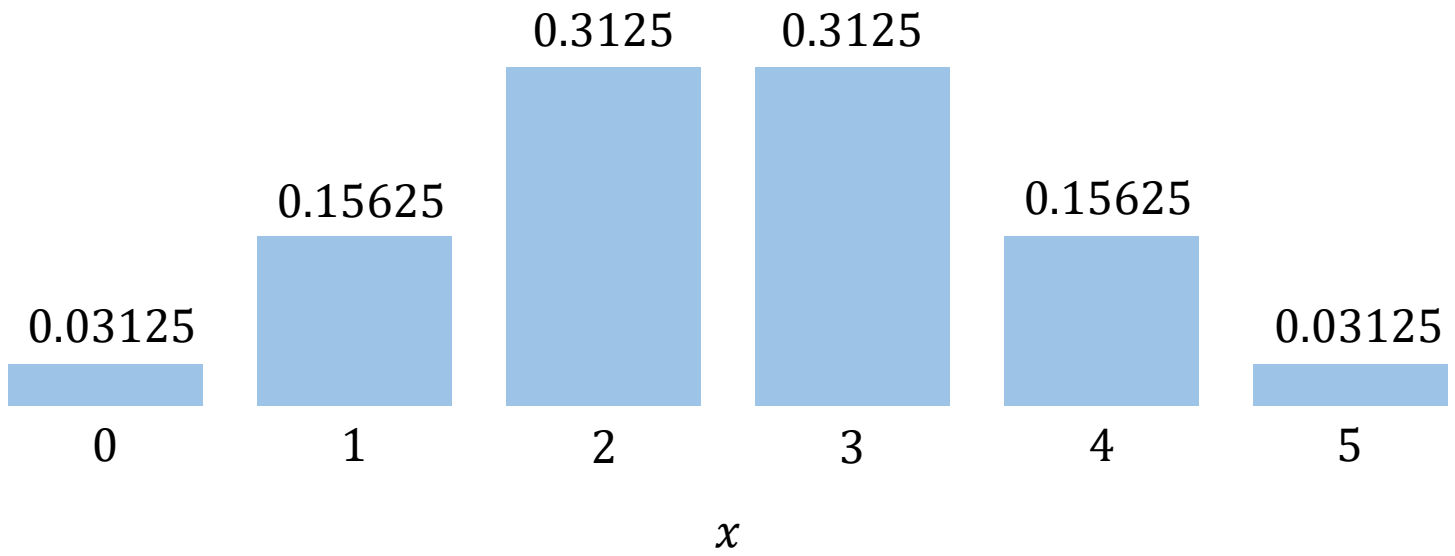
… Binomial Distribution

- Example
  - Looking for a general way to write the PMF for the number of heads in 5 coin tosses
    - X = Number of heads in 5 coin tosses
    - Suppose that the probability of heads is p
      - $P(H) = p$
    - Event $X = x$ means that x heads in 5 coin tosses
      - x can be 0, 1, 2, 3, 4, or 5
    - What is the probability of this event?
      - For one particular order:
        - Probability of seeing x heads = p^x
        - Probability of seeing $5 - x$ tails = $(1-p)^{5-x}$
      - PMF → For all possible orders:
        - $p_X(x) = \binom{5}{x} p^x (1-p)^{5-x}$ for $x = 0, 1, 2, 3, 4, 5$



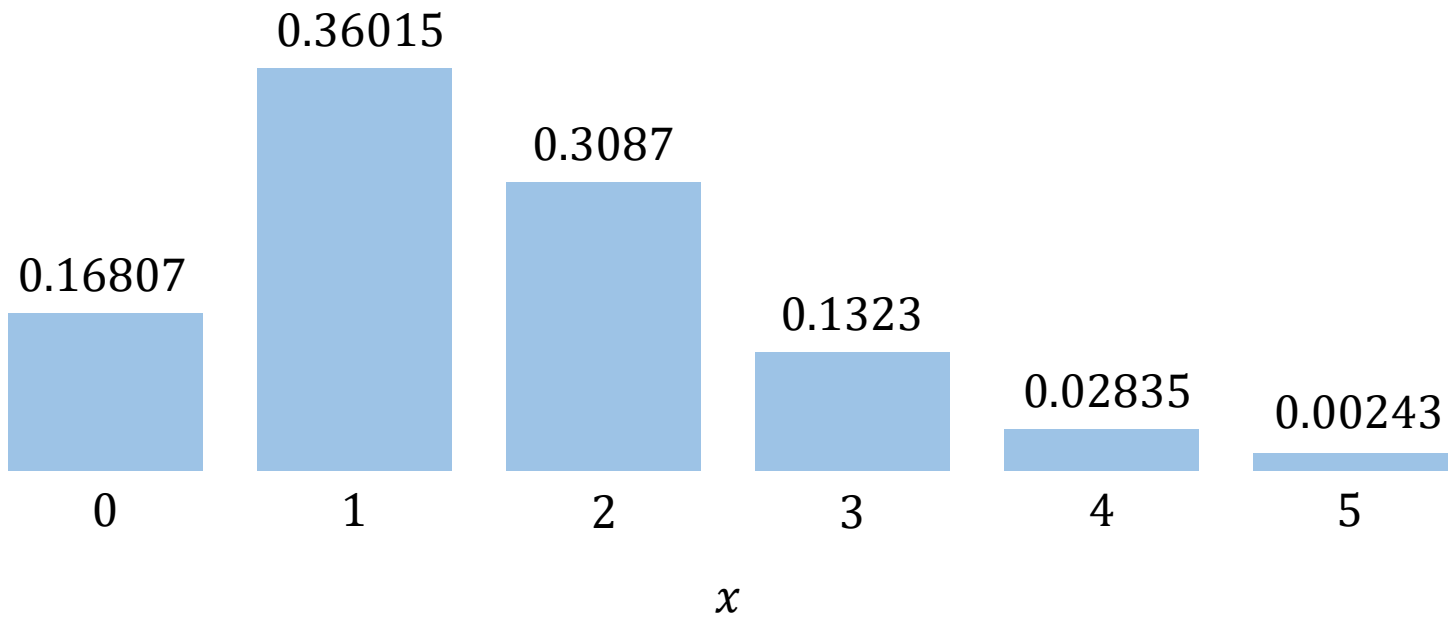
… Binomial Distribution

- Example- Symmetric
  - $X \sim Binomial\left(5, \frac{1}{2}\right)$
    - $p_X(x) = P(X = x) = \binom{5}{x} 0.5^x 0.5^{5-x}$



… Binomial Distribution

- Example- Asymmetric
  - $X \sim Binomial(5,0.3)$
    - $p_X(x) = P(X = x) = \binom{5}{x} 0.3^x 0.7^{5-x}$



… Binomial Distribution

- General PMF formula
  - $X \sim Binomial(n, p)$
    - $p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$
    - n and p are called the parameters of the binomial distribution

    - Number of heads in n coin flips
    - $P(H) = p$
    - Event $X = x$: x heads in n tosses

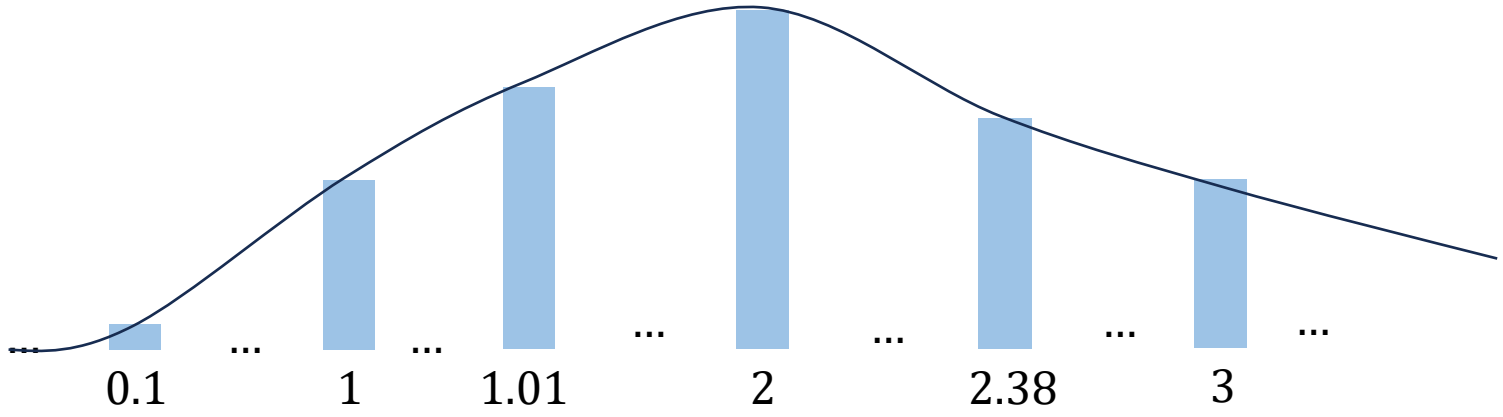
… Binomial Distribution

- Example
  - Rolling a dice 5 times
  - What is the PMF of getting ones?
    - We can assume that the dice is a biased coin
      - Get heads (one) with probability 1/6
      - Get tail (others) with probability 5/6
    - $X \sim Binomial\left(5, \frac{1}{6}\right)$

Continuous Distributions

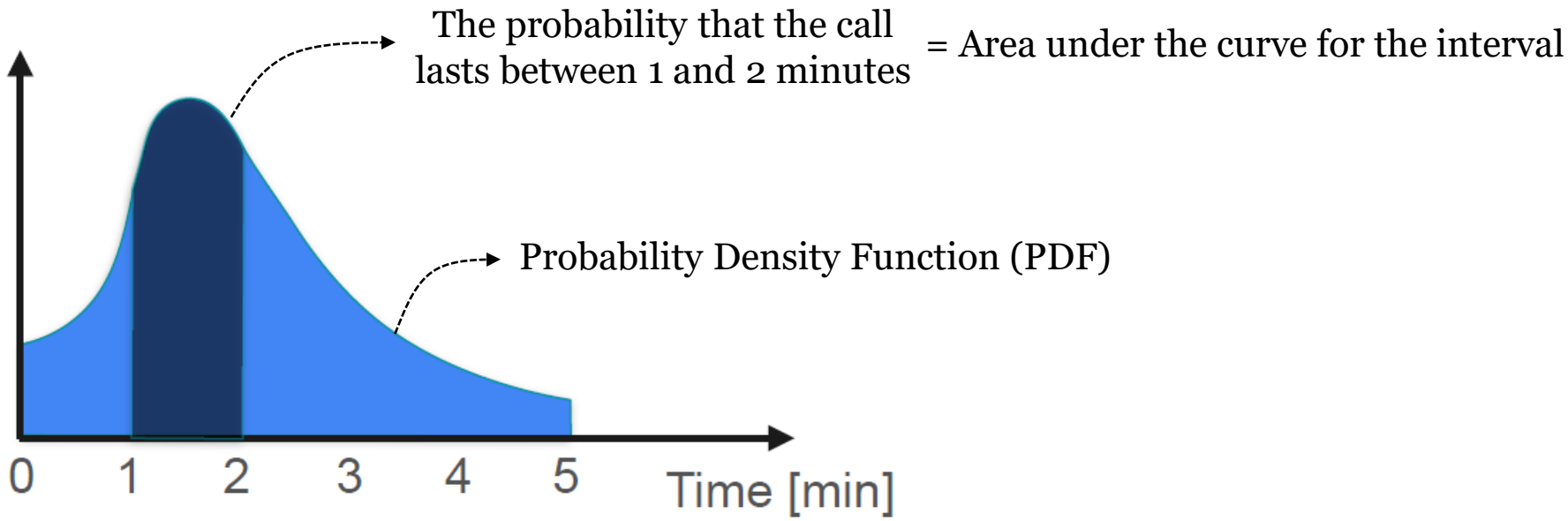
Introduction

- In the discrete distributions, the events always form a list (countable)
- In continuous distributions, the events are an interval
  - Example: waiting time on the phone call
- We cannot draw the probability distribution in the same way we did in PMF
  - Instead, we can draw it for intervals, e.g. 0.01 second intervals
    - Make the intervals more granular to get tiny intervals
      - For infinitely many skinny bars we get a curve



… Introduction

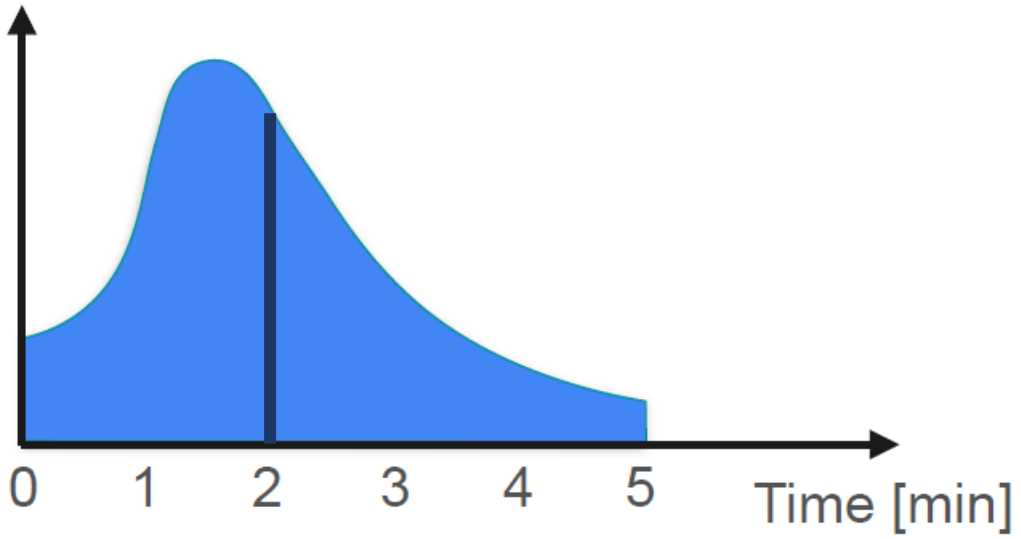
- Discrete variable
  - Sum of heights equals 1
- Continuous variable
  - Area under the curve equals 1



… Introduction

- What is the probability that the call is exactly 2 minutes?
  - Zero!

- In continuous distributions, we can only think of probabilities between window interval, not at a particular single point

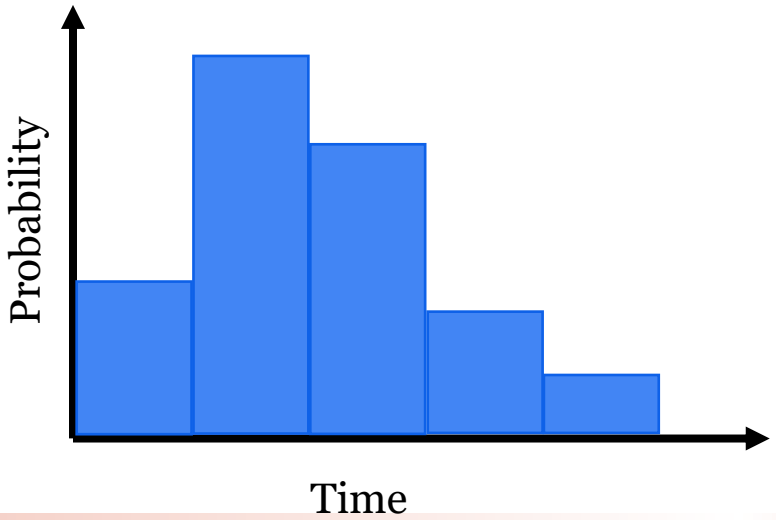


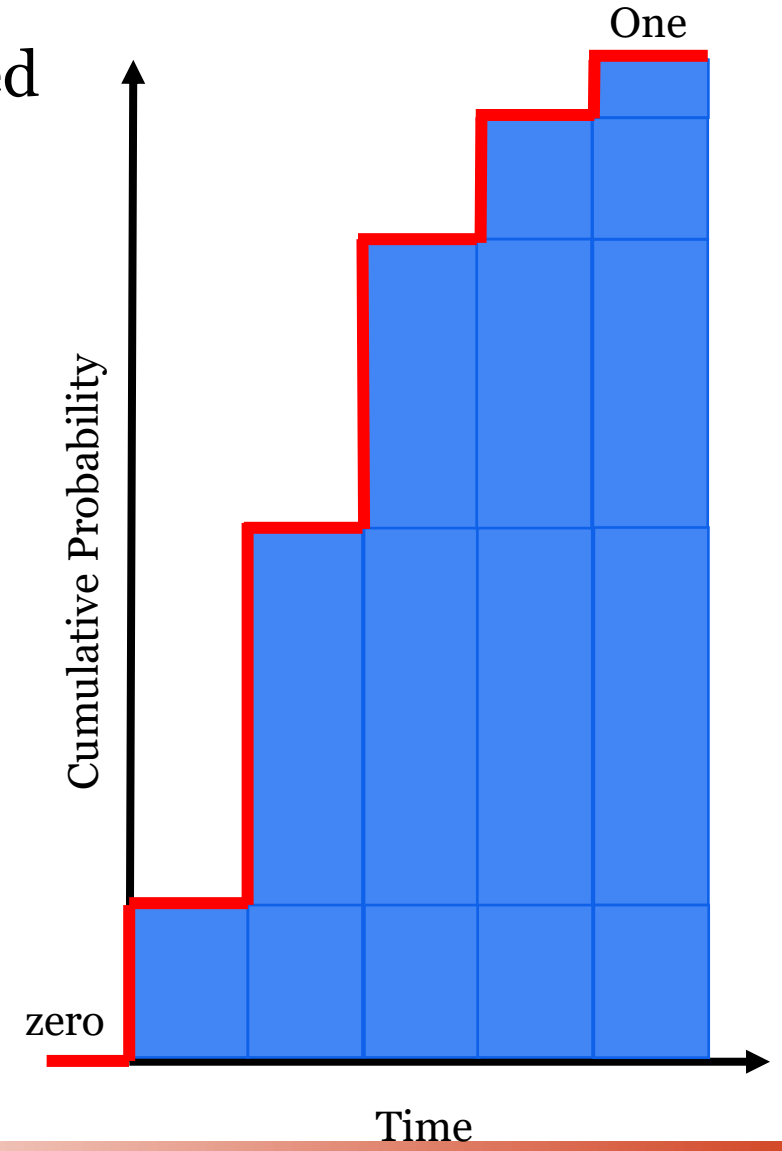
Probability Density Function (PDF)

- What is it?
  - The function that tells you the rate you accumulate probability around each point for continuous random variable
- Notation
  - Use lowercase f to represent it.
    - $f_X(x)$
    - $\mathrm{P}(a < x < b) =$ area under $f_X(x)$
- Properties
  - $f_X(x)$ is defined for all numbers
  - $f_X(x) \geq 0$
  - Area under $f_X(x)$ equals 1 ($\int f_X(x) = 1$)

Cumulative Distribution Function (CDF)

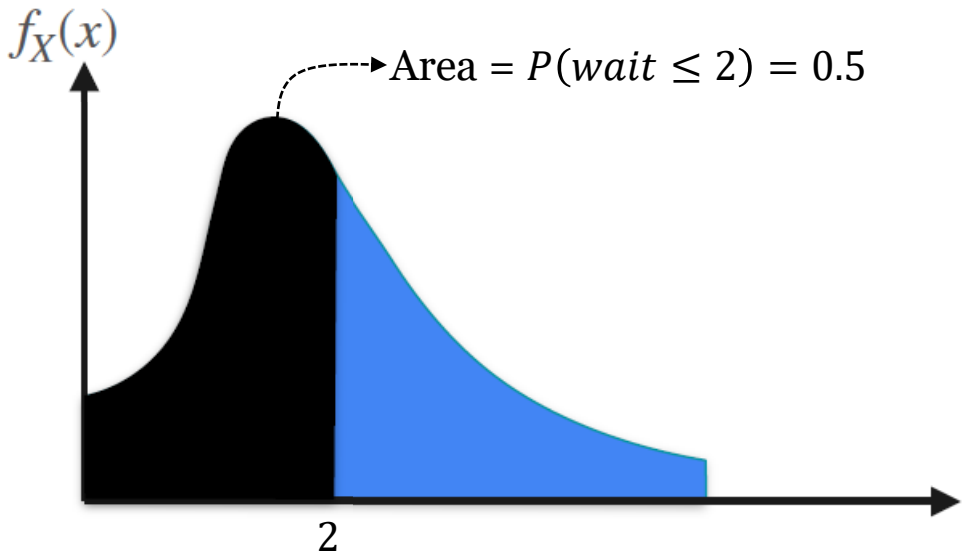
- Specify what is the probability that an event happened before some reference point.
  - Example- call center waiting time
    - The discrete simplification
  - Property
    - Always start at zero and end at one

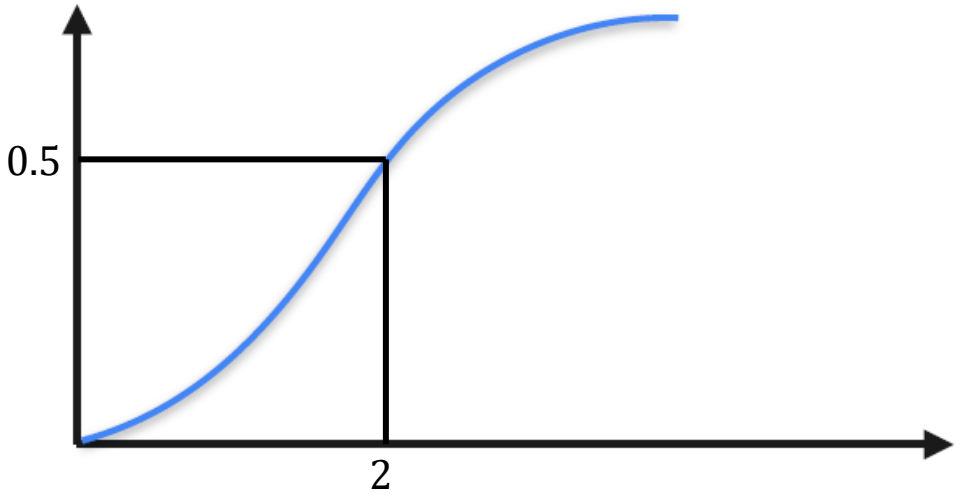




… Cumulative Distribution Function (CDF)

- Continuous distributions
  - Instead of adding the heights of bars, plotting the sum of the area under the curve





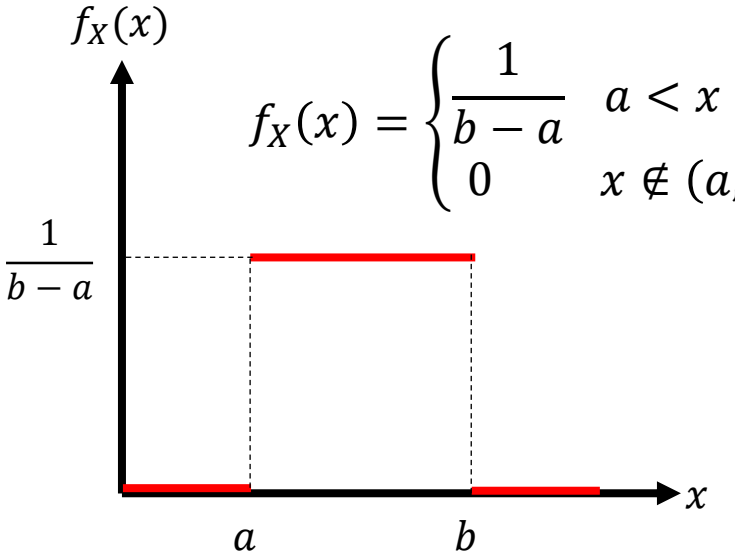
… Cumulative Distribution Function (CDF)

- More formal
  - The CDF shows how much probability the variable has accumulated until a certain value
  - Use uppercase F to represent it
    - $F_X(x) = P(X \leq x)$
  - Properties
    - $0 \leq F_X(x) \leq 1$
    - Left endpoint is 0
    - Right endpoint is 1
      - Can be at infinity
    - Never decreases

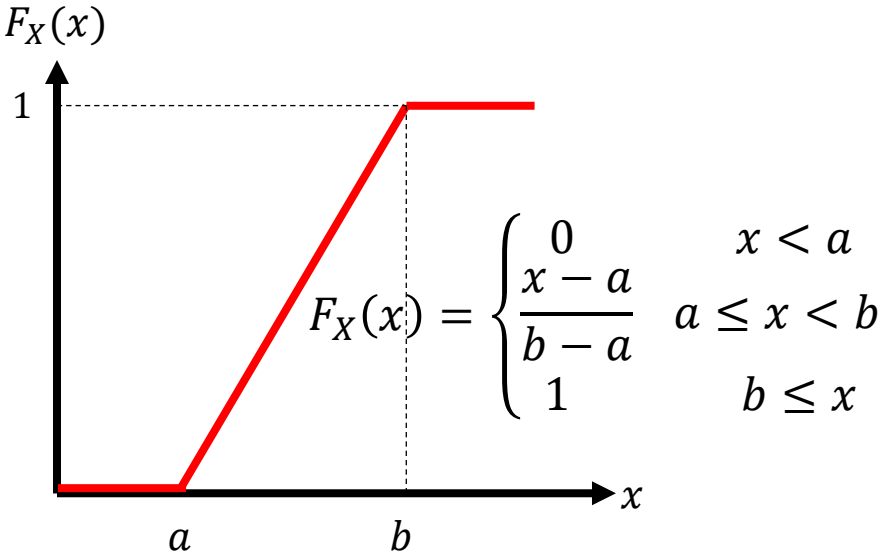
Uniform Distribution

- A continuous random variable can be modeled with a uniform distribution if all possible values lie in an interval and have the same frequency of occurrence
- Parameters
  - Beginning of the interval
  - End of the interval
- Formula

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \notin (a,b) \end{cases}$$

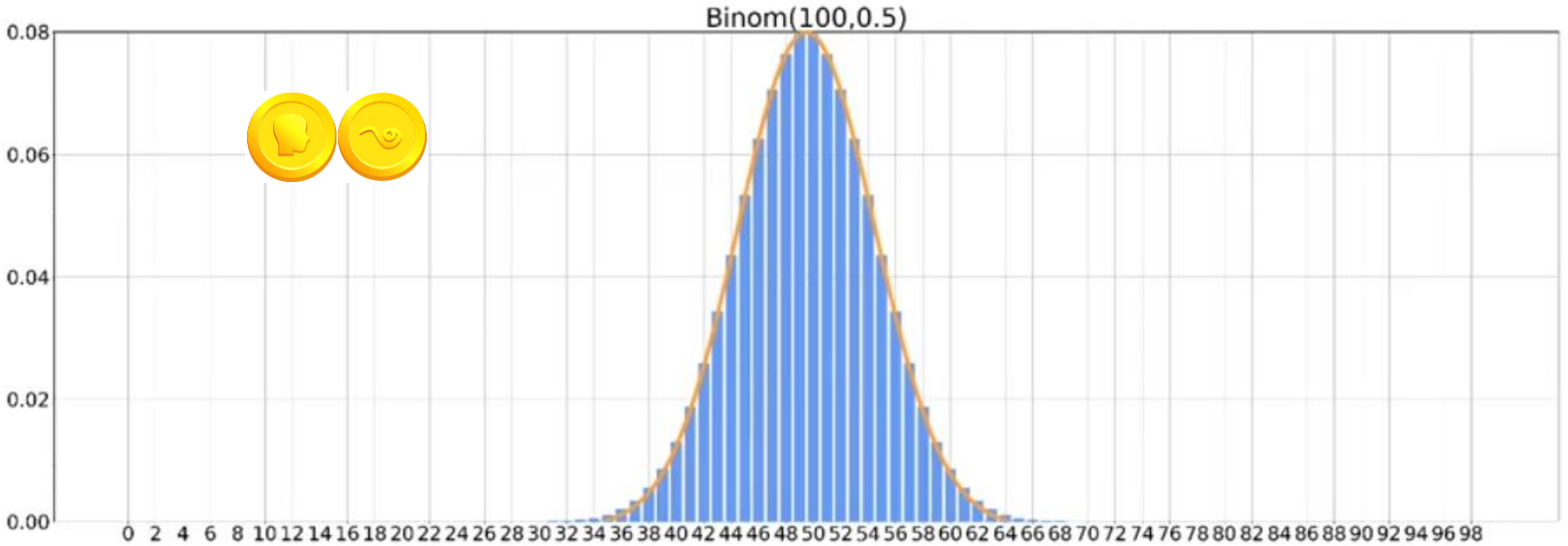


$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & b \le x \end{cases}$$



Normal Distribution

- Bell curve shape
- Also named Gaussian distribution
- The binomial distribution can be approximated by the normal distribution
  - Example- Throw a fair coin for n times, when n is very large

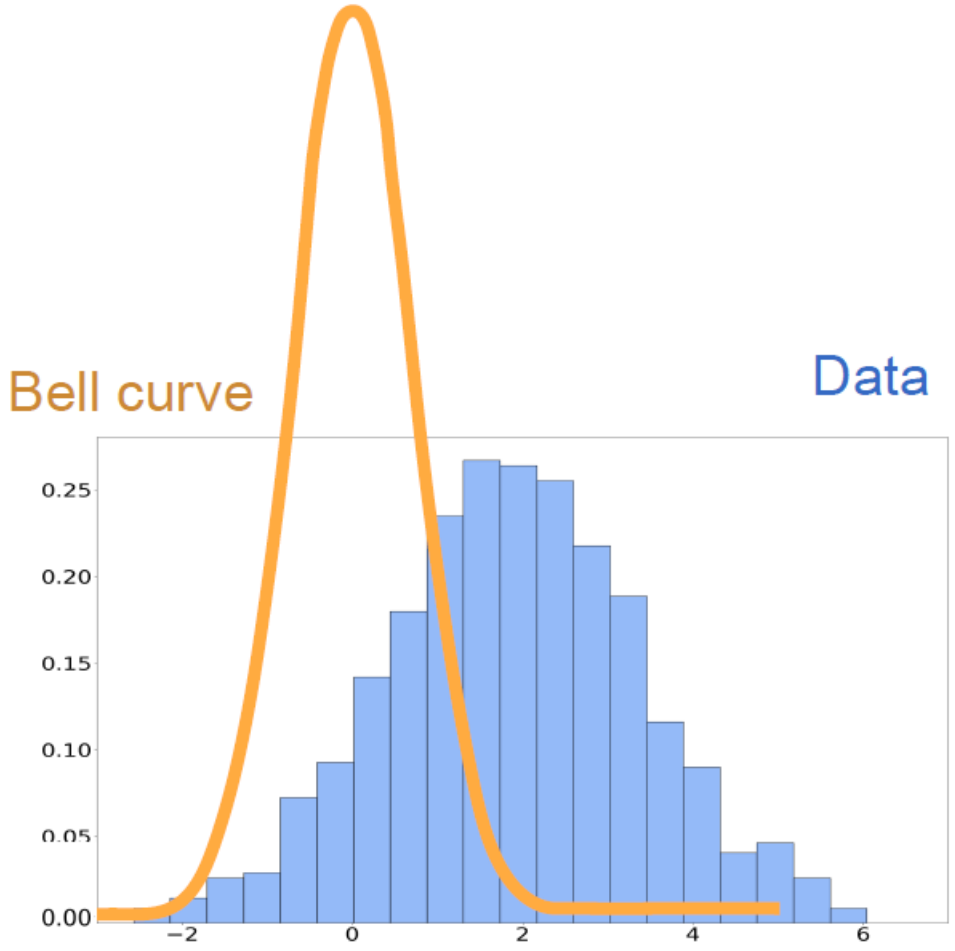


Number of throws (n)

… Normal Distribution

- Formula- example

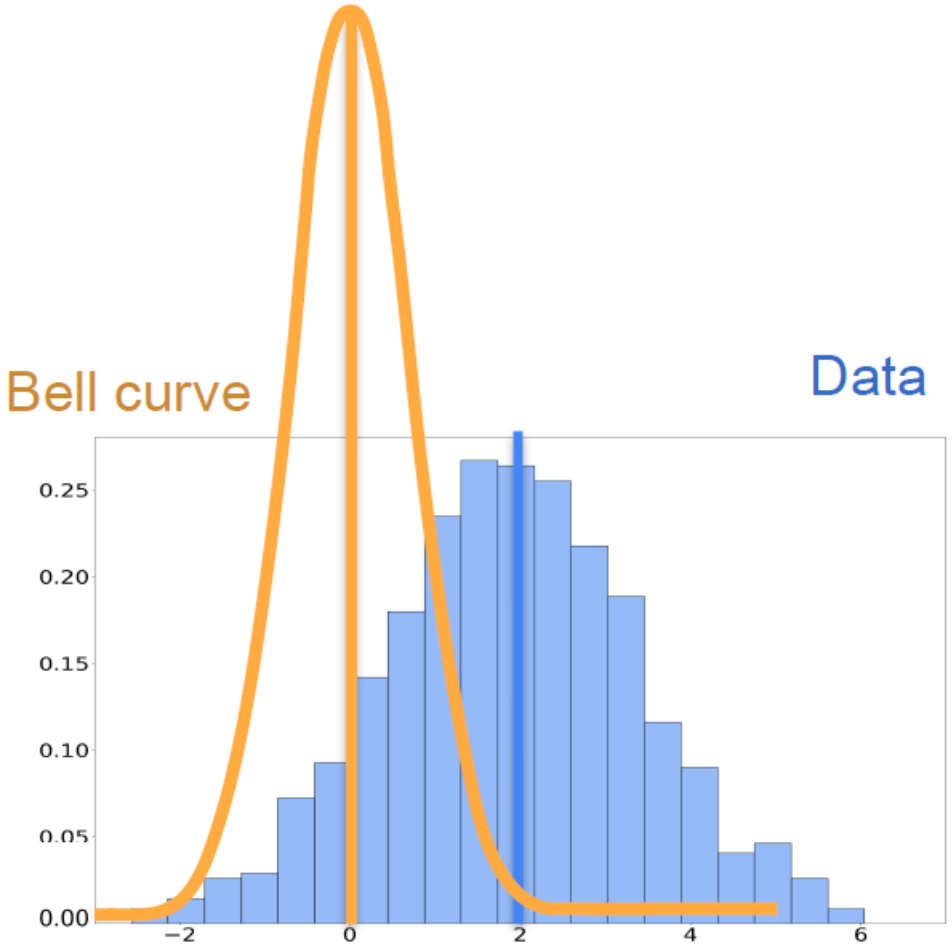
$$e^{-\frac{1}{2}x^2}$$



… Normal Distribution

- … Formula- example
  - Not centered

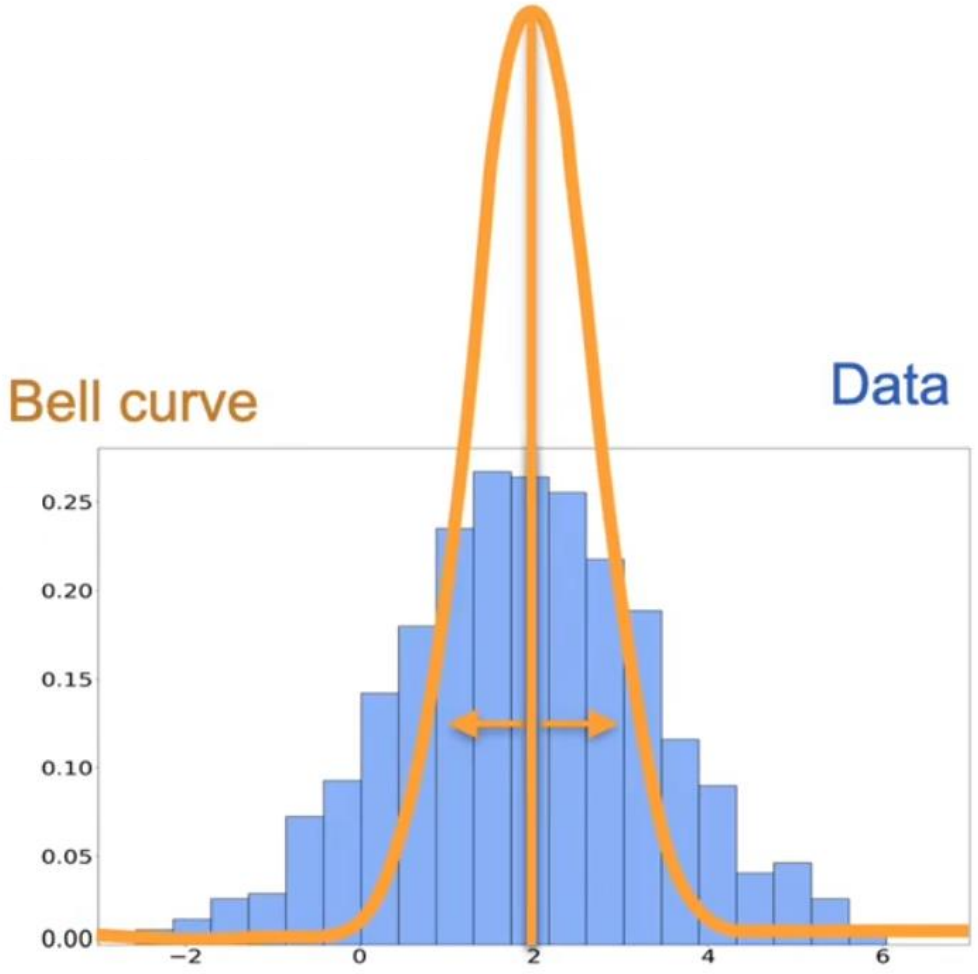
$$e^{-\frac{1}{2}x^2}$$



… Normal Distribution

- … Formula- example
  - Skinny

$$e^{-\frac{1}{2}(x-2)^2}$$



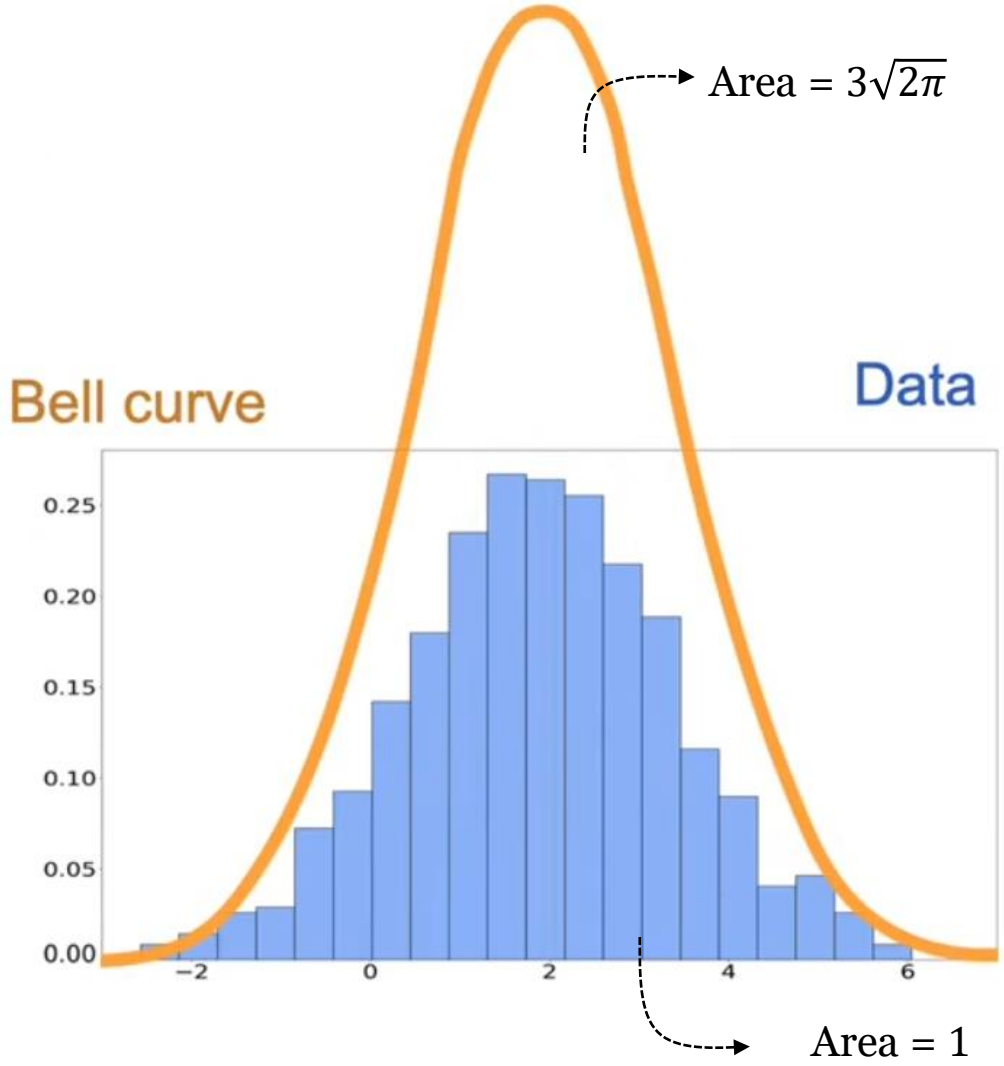
$\sigma = 1$ $\sigma = 3$

Thickness of the curve ← → Thickness of the data

… Normal Distribution

- … Formula- example
  - Different height

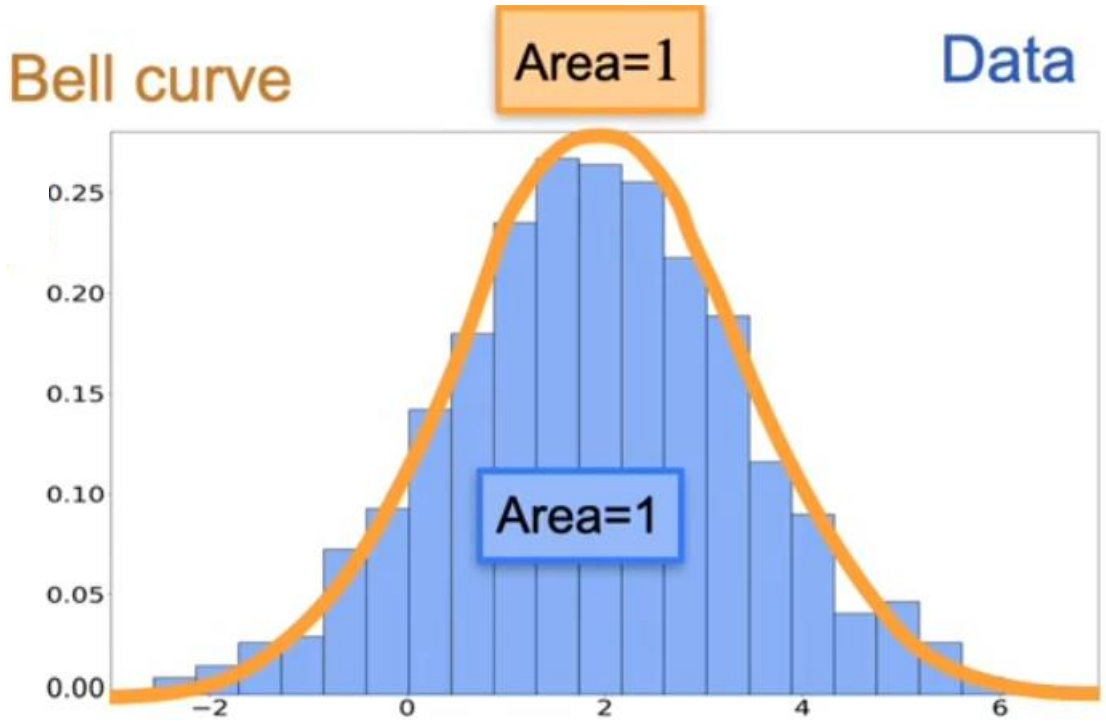
$$e^{-\frac{1}{2}\left(\frac{x-2}{3}\right)^2}$$



… Normal Distribution

- … Formula- example

$$\frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-2}{3}\right)^2}$$



… Normal Distribution

- Formula- general

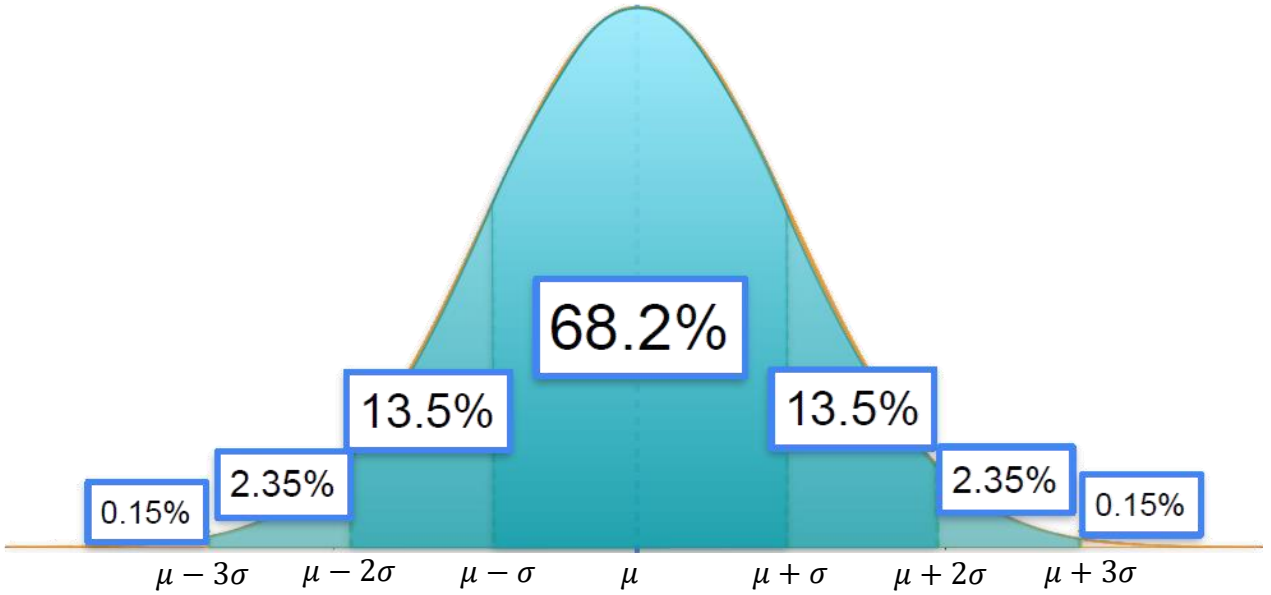
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  - μ is the mean of the data
  - σ is the standard deviation of the data

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Standard Deviation in the Normal Distribution

- 68-95-99.7
  - μ is the mean of the data
  - σ is the standard deviation of the data
  - $X \sim \mathcal{N}(\mu, \sigma^2)$



...Standard Deviation in the Normal Distribution

- Sum of Normal distributions
  - $W = aX + bY$
  - X and Y are independent
  - $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
  - $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

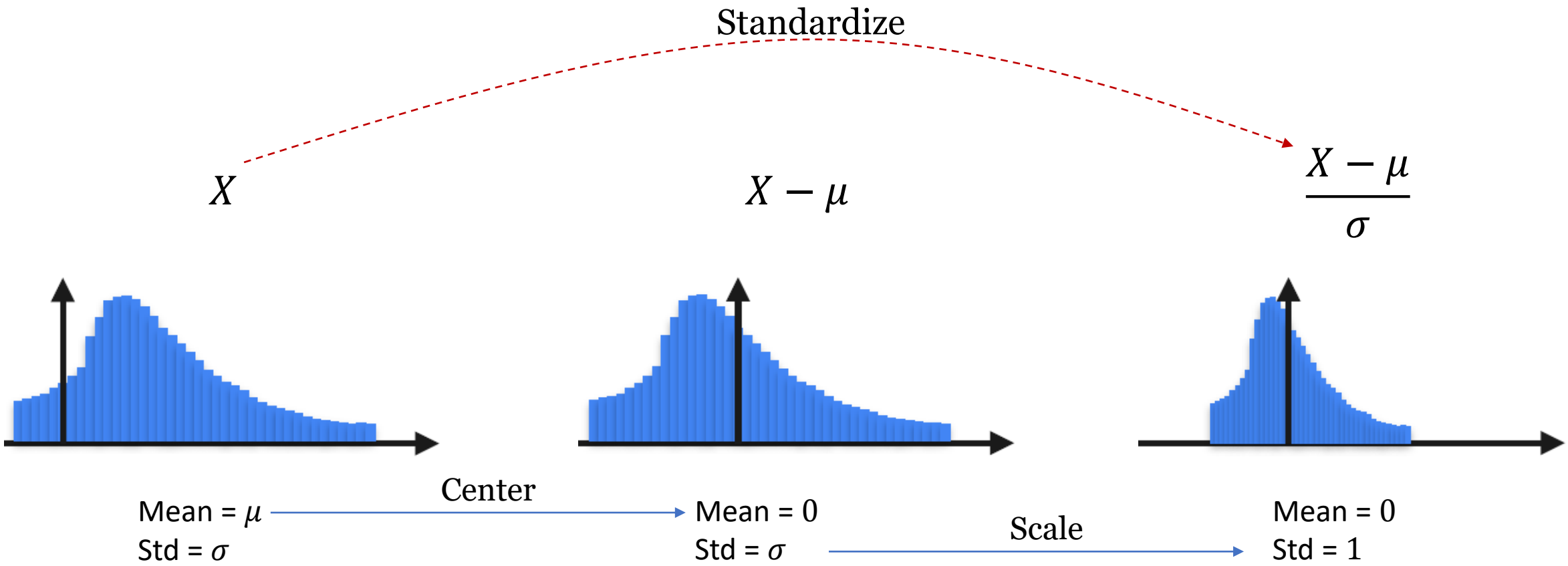
$$W \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

Standardization

- Everything is nicer when the mean is 0 and the standard deviation is 1

- How to standardize data?
  - $X - \mu$ leads to 0 mean
    - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \rightarrow \mathbb{E}[X - \mu] = \mathbb{E}[X] - \mathbb{E}[\mu] = \mathbb{E}[X] - \mu = 0$
  - $\frac{X}{\sigma}$ leads to 1 std
    - $Var(cX) = \mathbb{E}[(cX)^2] - \mathbb{E}[cX]^2 = \mathbb{E}[c^2X^2] - c^2\mathbb{E}[X]^2 = c^2\mathbb{E}[X^2] - c^2\mathbb{E}[X]^2 = c^2(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = c^2Var(X)$
    - $Var\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2}Var(X)$
    - $std\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma}std(X) = \frac{\sigma}{\sigma} = 1$

… Standardizing a Distribution



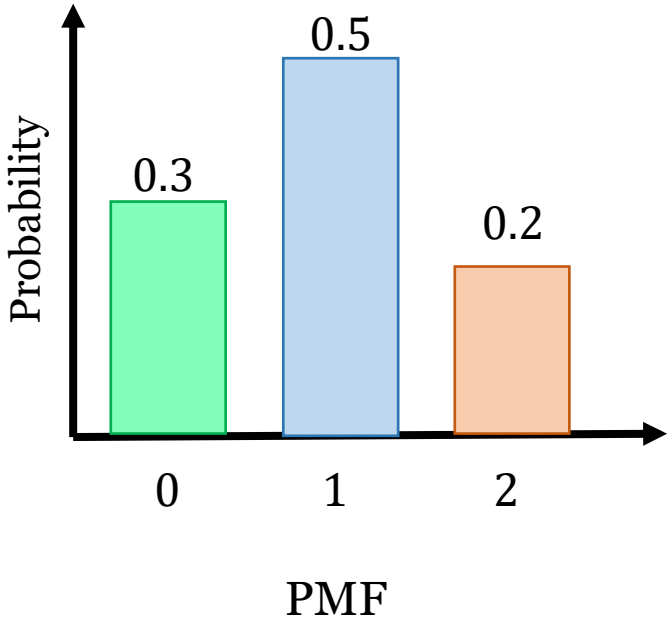
Sampling from a Distribution

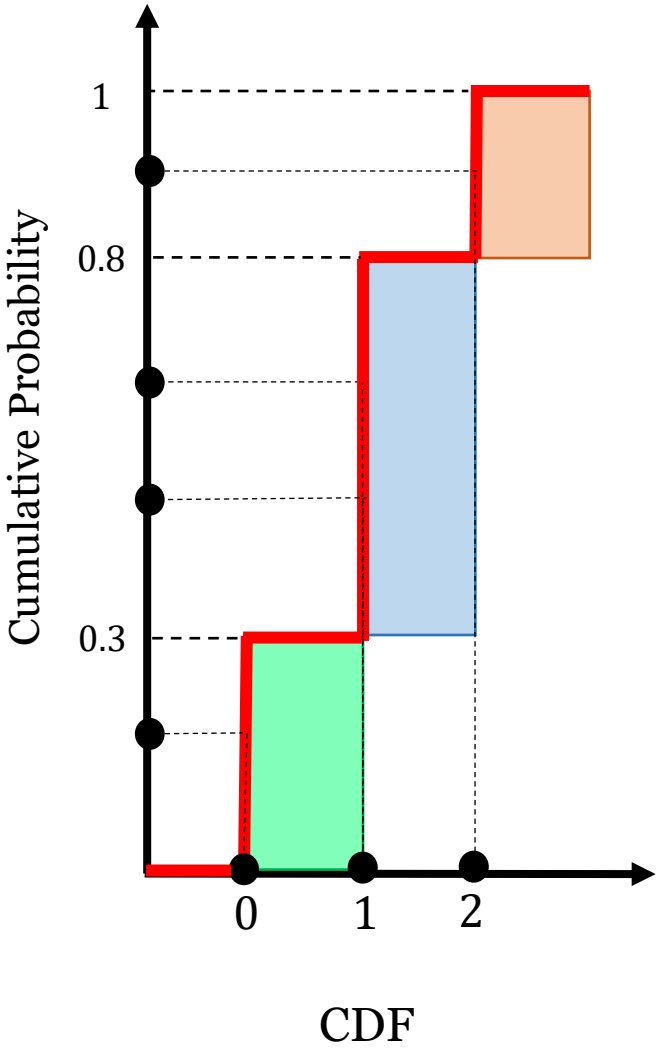
- Create syntactic data that looks like the original one
  - We know the original data distribution
  - We need to pick points that have probabilities given by the original distribution

- How?
  - Draw the cumulative distribution function
  - Generate numbers uniformly between 0 and 1
  - Locate the generated numbers on the vertical axis of CDF
  - Pick related values on the horizontal axis of CDF

… Sampling from a Distribution

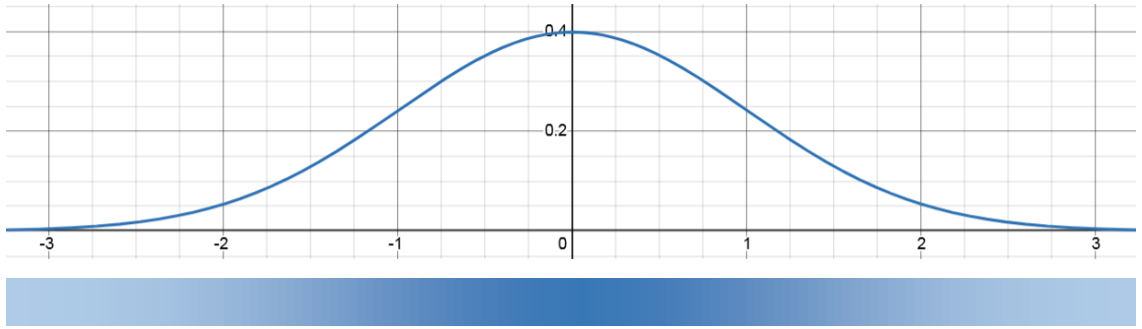
- Discrete distribution



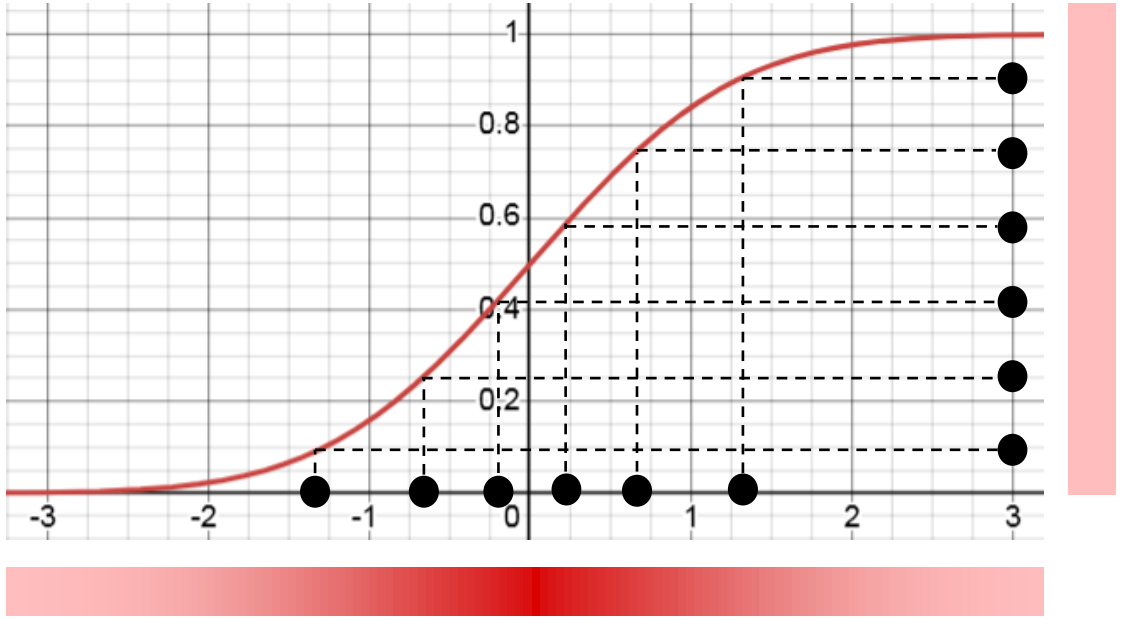


… Sampling from a Distribution

- Continuous distribution



PDF

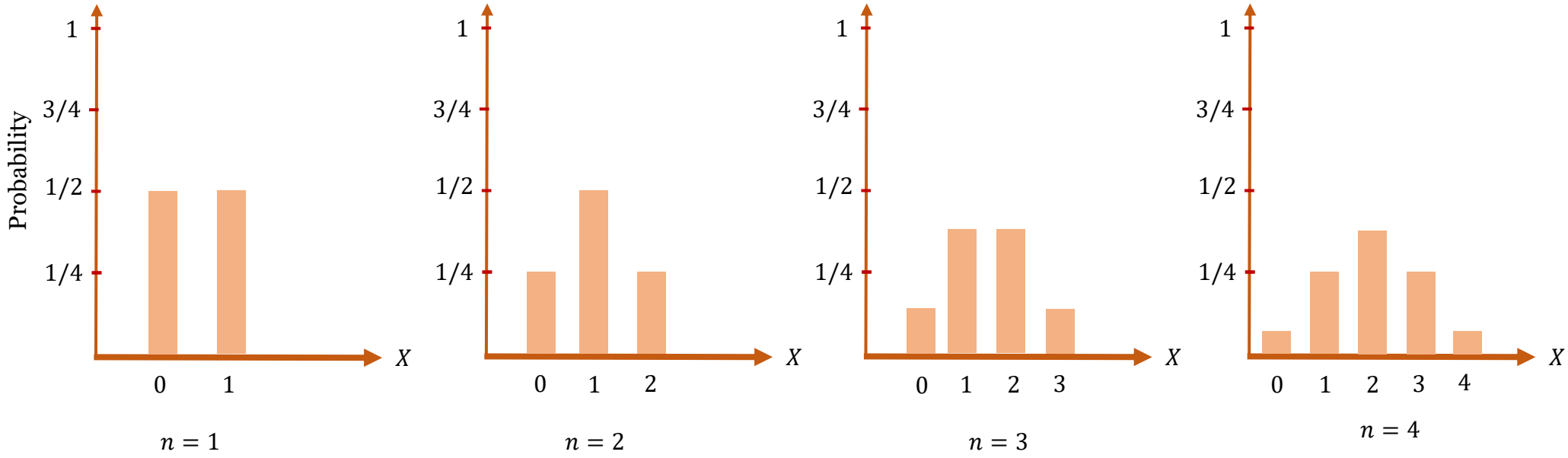


CDF

Central Limit Theorem

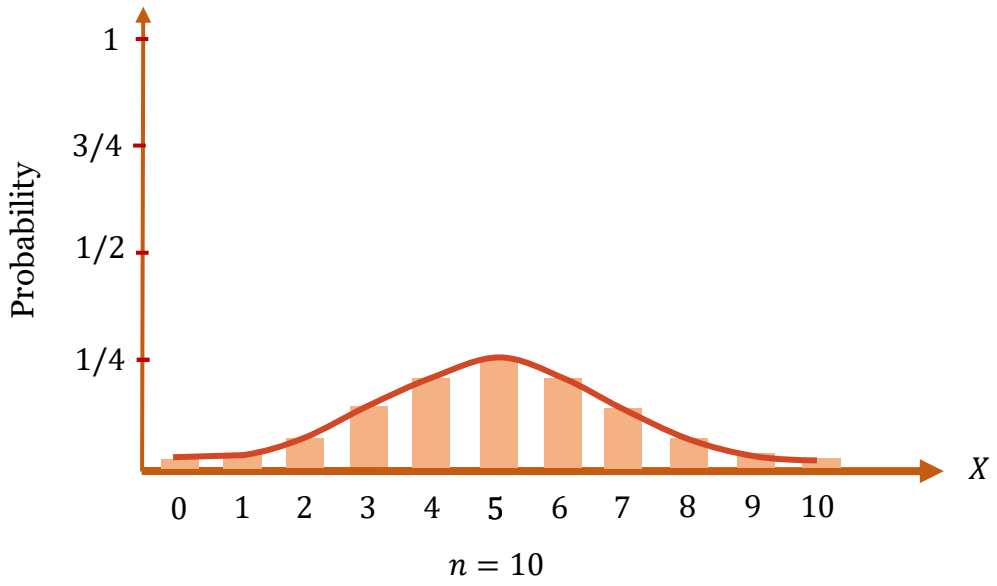
Discrete example

- flip a fair coin n times
  - $P(H) = P(T) = 0.5$
  - Random variable X = number of heads



…Discrete example

- Look like a Gaussian distribution!

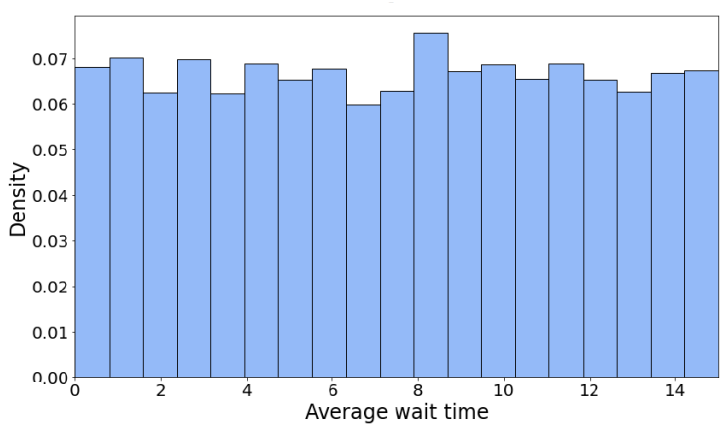


As you increase the number of observations, the probability distribution becomes closer to a Gaussian distribution.

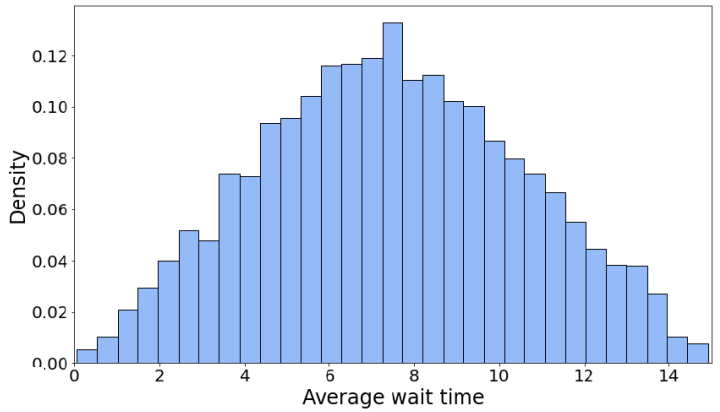
Continuous example

- Random variable X = wait to time for a call to be answered
- Assume that X follows a uniform distribution: $X \sim \mathcal{U}(0,15)$
- Calculate average waiting time (Y) using n samples
  - $n = 1 \rightarrow Y_1 = \frac{X_1}{1}$
  - $n = 2 \rightarrow Y_2 = \frac{X_1 + X_2}{2}$
  - …
  - $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- What is the distribution of Y_n
  - Again, Gaussian!

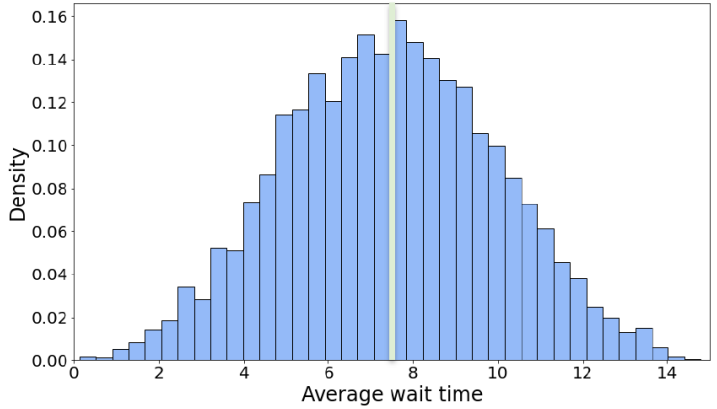
...Continuous example



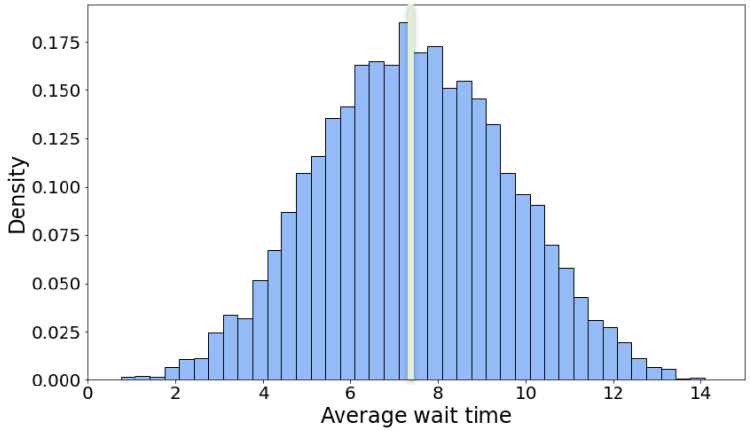
$n = 1$



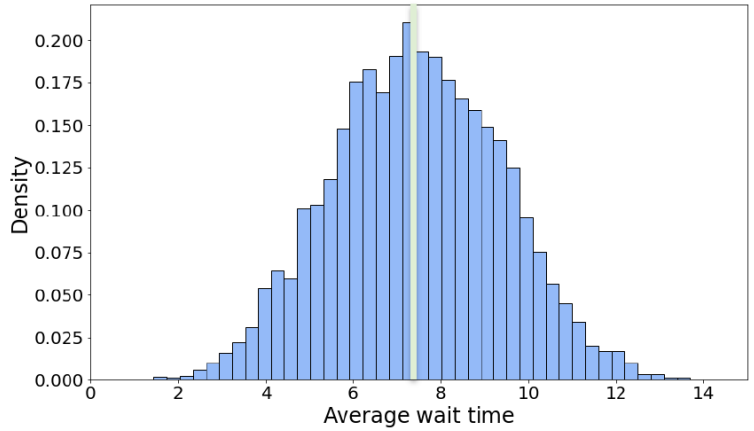
$n = 2$



$n = 3$



$n = 4$



$n = 5$

When you average on a large enough number of samples, the distribution will approximately follow a normal distribution

...Continuous example

- Mean & variance of Y_n
  - $\mathbb{E}[Y_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} n\mathbb{E}[X] = \mathbb{E}[X_i]$

    Assuming each X_i has the same distribution, So $E[X_i] = E[X]$
    - $\mu_{Y_n} = \mu$
    - Mean of $Y_n$equals the population mean
  - $Var(Y_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} nVar(X) = \frac{Var(X)}{n}$
    - $\sigma_{Y_n}^2 = \frac{\sigma^2}{n}$
    - Variance of $Y_n$equals $\frac{1}{n}$ of the population variance

Central Limit Theorem (Definition)

- Formal definition
  - Regardless of the distribution of X, as $n \to \infty$, the variable $\frac{1}{n}\sum_{i=1}^{n} X_i$ follows a normal distribution. We can standardize it and get the standard normal distribution.

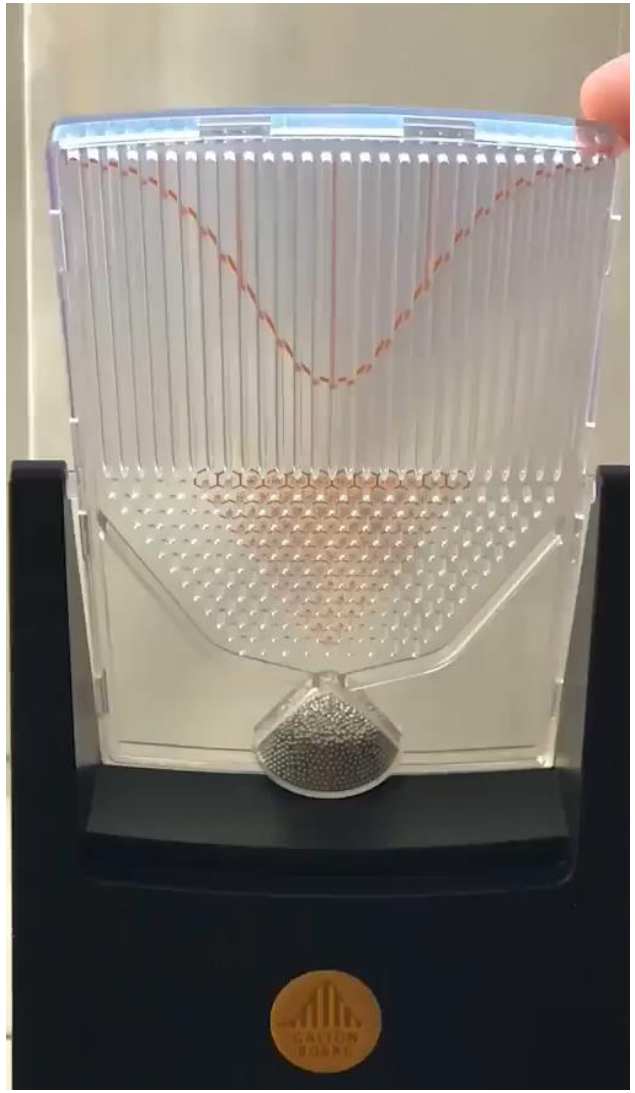
    - $\frac{Y_n - \mu_{Y_n}}{\sigma_Y} \sim \mathcal{N}(0, 1^2)$
    - $\frac{Y_n - \mu_{Y_n}}{\sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]}{\sqrt{\frac{\sigma_X^2}{n}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]}{\sigma_X}\sqrt{n} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}n\mathbb{E}[X]}{\sigma_X}\sqrt{n} = \frac{1}{n}\frac{\sum_{i=1}^{n} X_i - n\mathbb{E}[X]}{\sigma_X}\sqrt{n}$
      $= \frac{\sum_{i=1}^{n} X_i - n\mathbb{E}[X]}{\sqrt{n}\sigma_X}$
    - $\frac{\sum_{i=1}^{n} X_i - n\mathbb{E}[X]}{\sqrt{n}\sigma_X} \sim \mathcal{N}(0, 1^2)$

…Central Limit Theorem (Definition)

- Narrative description
  - Take any arbitrary distribution, take a few samples (always the same number), and look at the average
  - Do this many times and plot all these averages. You get the normal distribution.
    - No matter what distribution you started with in the first place.

- Note
  - In general, a safe rule is that you usually need about 30 samples before the bell-shaped distribution comes in.
    - If the original population of the data is very skewed, you usually need more samples than if you are working with a symmetric distribution.

Galton board



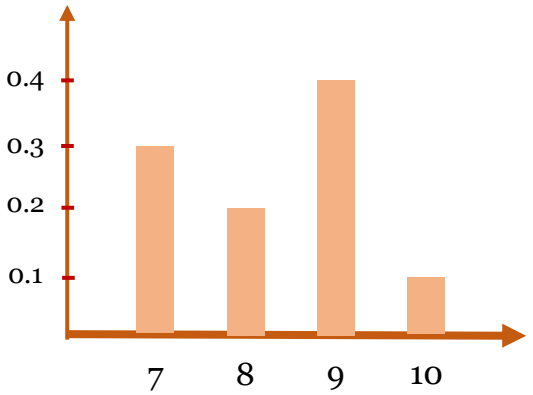
Multiple Variables

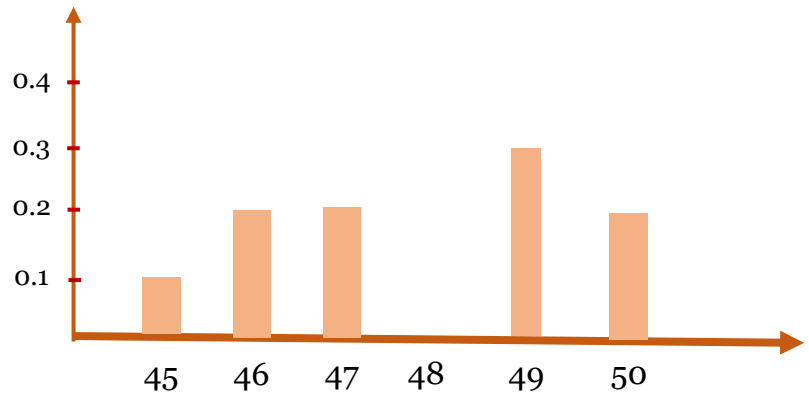
Joint Distributions (Discrete)

- Example-1

Age	Count
7	3
8	2
9	4
10	1

Height	Count
45	1
46	2
47	2
48	0
49	3
50	2





… Joint Distributions (Discrete)

- … Example-1

Age	Height
7	45
7	46
7	46
8	47
8	47
9	49
9	49
9	49
9	50
10	50

Age \ Height	45	46	47	48	49	50
7	1	2	0	0	0	0
8	0	0	2	0	0	0
9	0	0	0	0	3	1
10	0	0	0	0	0	1

… Joint Distributions (Discrete)

- … Example-1

Age \ Height	45	46	47	48	49	50
7	1	2	0	0	0	
8	0	0	2	0	0	
9	0	0	0	0	3	
10	0	0	0	0	0	

Divide by sum

Age (X) \ Height (Y)

	45	46	47	48	49	50
7	1/10	2/10	0	0	0	0
8	0	0	2/10	0	0	0
9	0	0	0	0	3/10	1/10
10	0	0	0	0	0	1/10

Example:

$$p_{XY}(7,46) = P(X = 7, Y = 46) = \frac{2}{10}$$

Probability Mass Function

All probabilities for all possible combinations of X and Y

Joint Distribution

$$\boldsymbol{p_{XY}(x,y) = P(X = x, Y = y)}$$

… Joint Distributions (Discrete)

- Example-2- Toss two dices X = the number rolled on the 1st dice Y = sum of the two dices

Y \ X	1	2	3	4	5	6
12						6, 6
11					5, 6	6, 5
10				4, 6	5, 5	6, 4
9			3, 6	4, 5	5, 4	6, 3
8		2, 6	3, 5	4, 4	5, 3	6, 2
7	1, 6	2, 5	3, 4	4, 3	5, 2	6, 1
6	1, 5	2, 4	3, 3	4, 2	5, 1	
5	1, 4	2, 3	3, 2	4, 1		
4	1, 3	2, 2	3, 1			
3	1, 2	2, 1				
2	1, 1					
1						

Example:
$p_{XY}(3,7)$
$= P(X = 3, Y = 7) = \frac{1}{36}$

Example:
$p_{XY}(1,1)$
$= P(X = 1, Y = 1) = 0$

Y \ X	1	2	3	4	5	6
12						1/36
11					1/36	1/36
10				1/36	1/36	1/36
9			1/36	1/36	1/36	1/36
8		1/36	1/36	1/36	1/36	1/36
7	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	
5	1/36	1/36	1/36	1/36		
4	1/36	1/36	1/36			
3	1/36	1/36				
2	1/36					
1						

… Joint Distributions (Discrete)

- For independent discrete variables
  - $p_{XY}(x, y) = P(X = x, Y = y) = P(x) \times P(y)$

- Example- Toss two dices
  - X = the number rolled on the 1st dice
  - Y = the number rolled on the 2nd dice
  - $p_{XY}(2,3) = P(2) \times P(3) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

Marginal Distribution

- Distribution of one variable while ignoring the other

- To find the marginal distribution for a variable, sum the joint probability distribution over all value of the other variable.

- Formula

$$P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

… Marginal Distribution

- Example
  - Age and height dataset

Age	7	7	7	8	8	9	9	9	9	10
Height	45	46	46	47	47	49	49	49	50	50

  - Joint distribution

		Height (Y)					
		45	46	47	48	49	50
Age (X)	7	1/10	2/10	0	0	0	0
	8	0	0	2/10	0	0	0
	9	0	0	0	0	3/10	1/10
	10	0	0	0	0	0	1/10

… Marginal Distribution

- … Example
  - Marginal distribution of height → Do NOT care about age anymore and only care about height
    - Summarize the behavior of the distribution across only the height variable
    - Add over each height value all the probabilities

		Height (Y)						Marginal distribution of age
		45	46	47	48	49	50	↓
Age (X)	7	1/10	2/10	0	0	0	0	3/10
	8	0	0	2/10	0	0	0	2/10
	9	0	0	0	0	3/10	1/10	4/10
	10	0	0	0	0	0	1/10	1/10
Marginal distribution of height →		1/10	2/10	2/10	0	3/10	2/10	

Conditional Distribution

- Example
  - Age and height dataset
  - Only care about age 9
    - If age = 9, what is the distribution across the height variable?
    - $P_{Y|X=9}(y) = P(Y = y|X = 9)$

Example:

$$p_{Y|X=9}(49) = P(Y = 49|X = 9) = \frac{3}{4}$$

Height (Y)

Age (X)	45	46	47	48	49	50
7	1/10	2/10	0	0	0	0
8	0	0	2/10	0	0	0
9	0	0	0	0	3/10	1/10
10	0	0	0	0	0	1/10

Height (Y)

Age (X)	45	46	47	48	49	50
9	0	0	0	0	3/10	1/10

Sum of the probabilities does not equal 1!

Divide by row sum — Normalize

9	0	0	0	0	3/4	1/4

… Conditional Distribution

- Observe one variable given that the value of the other one is known
- Why divide by the row sum?
  - $P(A,B) = P(A) \times P(B|A) \Rightarrow P(B|A) = \frac{P(A,B)}{P(A)}$

$P(A)$: Row sum

- Example
  - $P_{Y|X=9}(y) = P(Y = y|X = 9) = \frac{P(X=9,Y=y)}{P(X=9)}$
  - $P_{Y|X=9}(49) = P(Y = 49|X = 9) = \frac{P(X=9,Y=49)}{P(X=9)} = \frac{\frac{3}{10}}{\frac{4}{10}} = \frac{3}{4}$
- General formula

$$P_{Y|X=x}(y) = \frac{P_{XY}(x,y)}{P_X(x)}$$

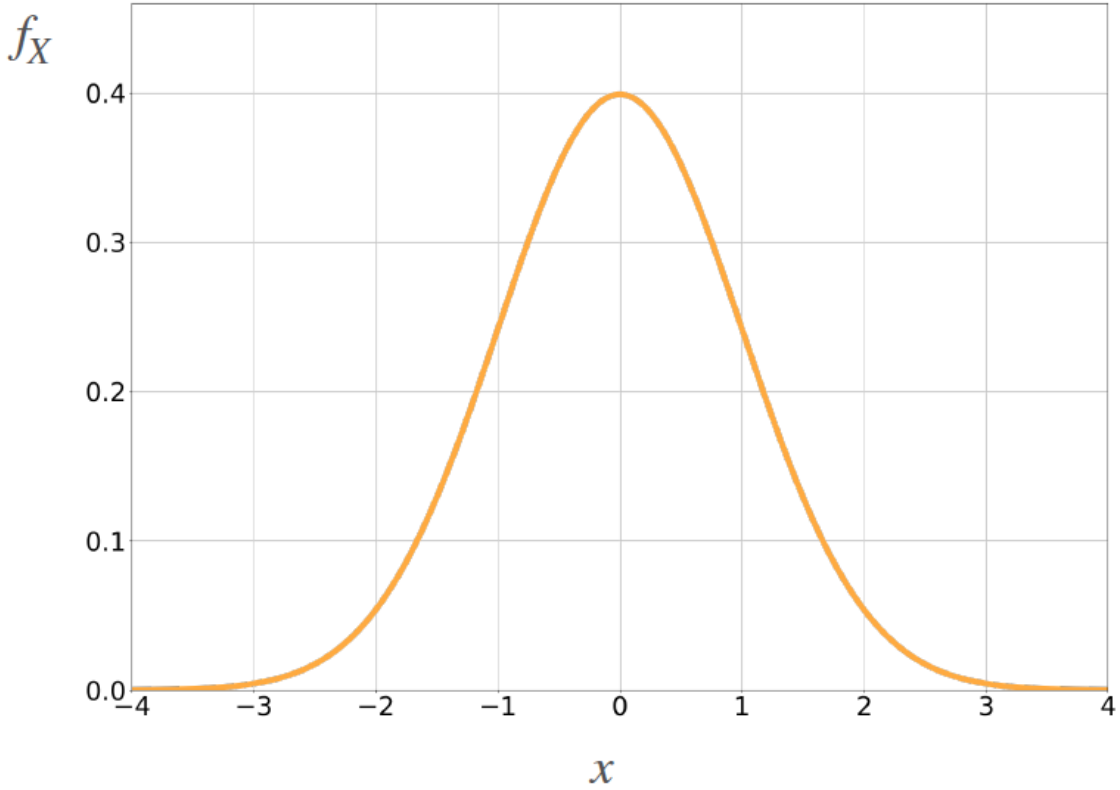
$P_{XY}(x,y)$: Joint PDF of X and Y

$P_{Y|X=x}(y)$: Conditional PDF of Y

$P_X(x)$: Marginal distribution of X

Multivariate Gaussian Distribution

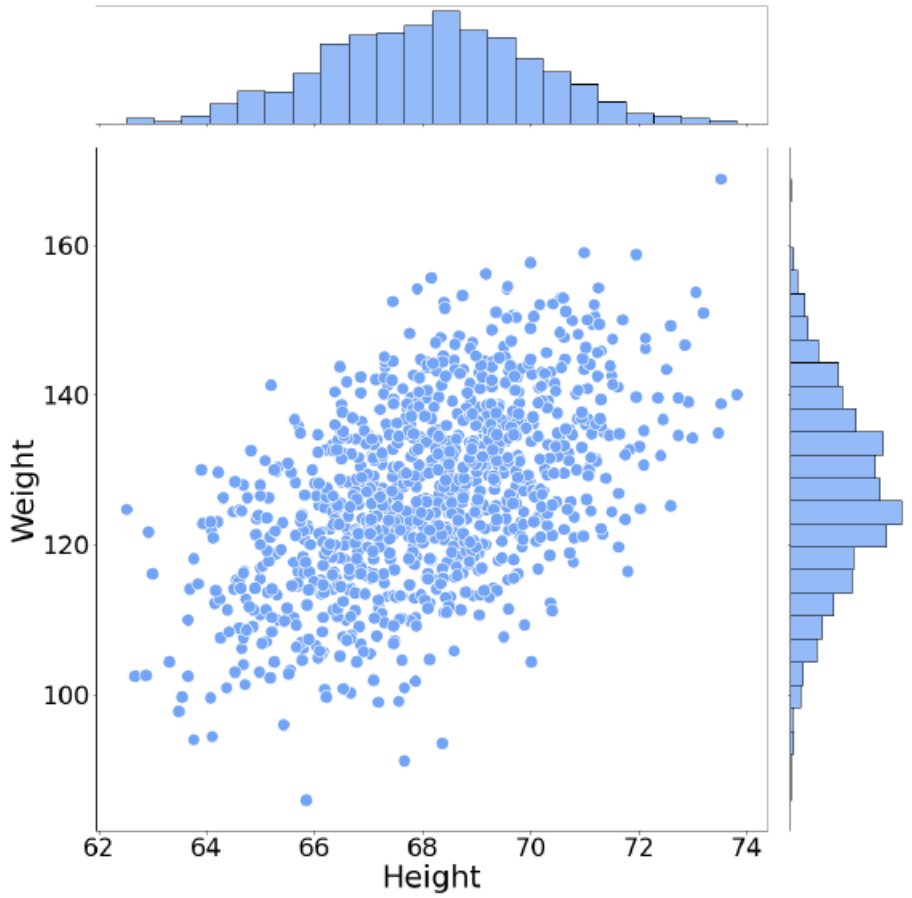
- For a single variable
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
    - μ is the mean of the data
    - σ is the standard deviation of the data
  - $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$



  - What if we have more than one variable?

… Multivariate Gaussian Distribution

- Example
  - H: Height of an adult in inches
    - $H \sim \mathcal{N}(\mu_H, \sigma_H^2)$
  - W: Weight of an adult in pounds
    - $H \sim \mathcal{N}(\mu_W, \sigma_W^2)$
  - If H and W were independent variables we could calculate $f_{HW}(h, w)$ simply by multiplying $f_H(h)$ and $f_W(w)$
    - But it is not the case here
      - Taller adults tend to be fatter
      - There is a correlation between variables



… Multivariate Gaussian Distribution

- General formula is very similar to the formula of the single one ($\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$)
  - Use covariance matrix instead of the variance
  - For vectors instead of scalar values

$$x = [x_1 \quad x_2 \quad \dots \quad x_n]^T \qquad \mu = [\mu_1 \quad \mu_2 \quad \dots \quad \mu_n]^T$$

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Confidence Interval

Introduction

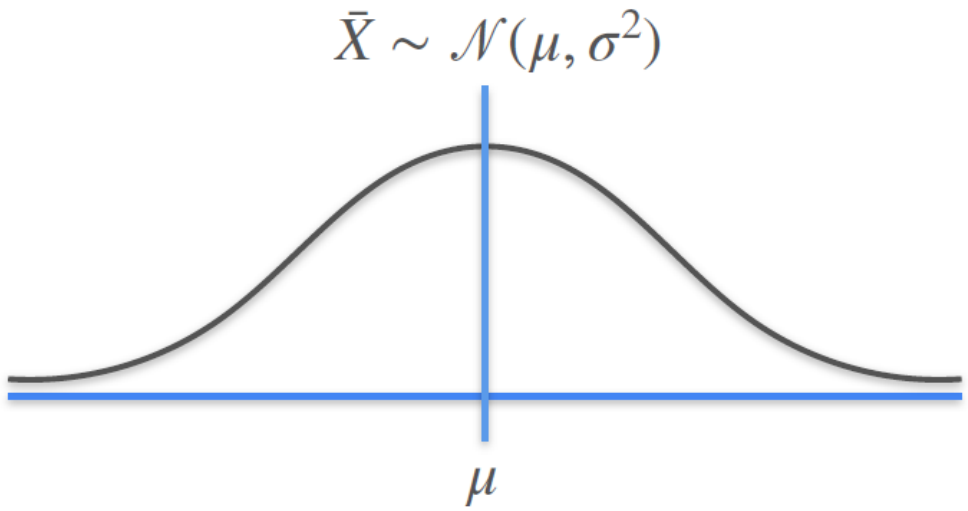
- Example
  - We can use sampling to estimate the mean of population height
  - How good our estimate is?
  - How can we assure the estimation is close to the actual population mean?

- Confidence interval is a technique that is used to establish a degree of certainty to the point estimates from the samples

- Example
  - Instead of $\bar{x}$, we use a lower and upper bounds to report a range to use the sample mean with some degree of certainty
    - $lower\ limit < \bar{x} < upper\ limit$

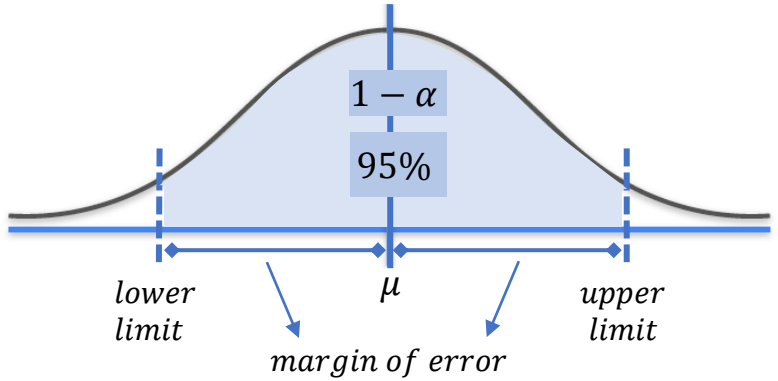
… Introduction

- Example
  - Sample size = 1
  - Find the mean of the sample to use as the estimate for the population mean
  - We take multiple samples of size 1
  - The sampling distribution for the sample means ($\bar{X}$) is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$
    - Based on the Central Limit Theorem
    - μ = population mean
    - σ = population standard deviation



… Introduction

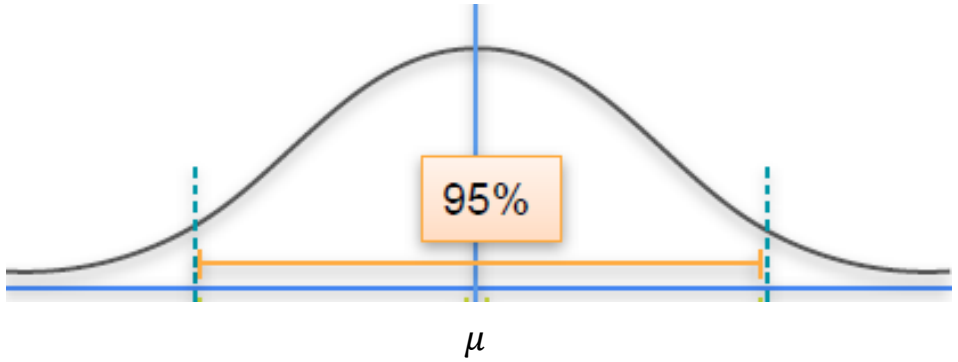
- We want to indicate the frequency with which our sample means lie within an interval (between an upper and a lower limit). How?
  - Decide on a value called the significance level (α)
    - Common value is 0.05
  - The value $1 - \alpha$ is called the confidence level
    - Common value is 0.95 (or 95%)



  - The confidence level indicates the region that contains 95% of the entire sample means in the sampling distribution
  - The lower and upper limits can be calculated based on the standard deviation and the confidence interval we have chosen.
    - This is called the confidence interval
  - The margin error is the distance between the upper/lower limits and the true mean

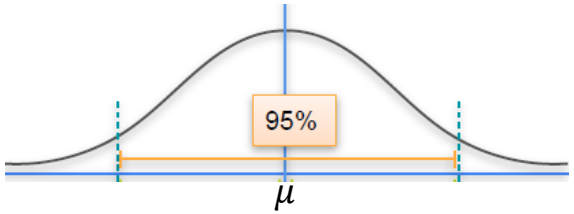
… Introduction

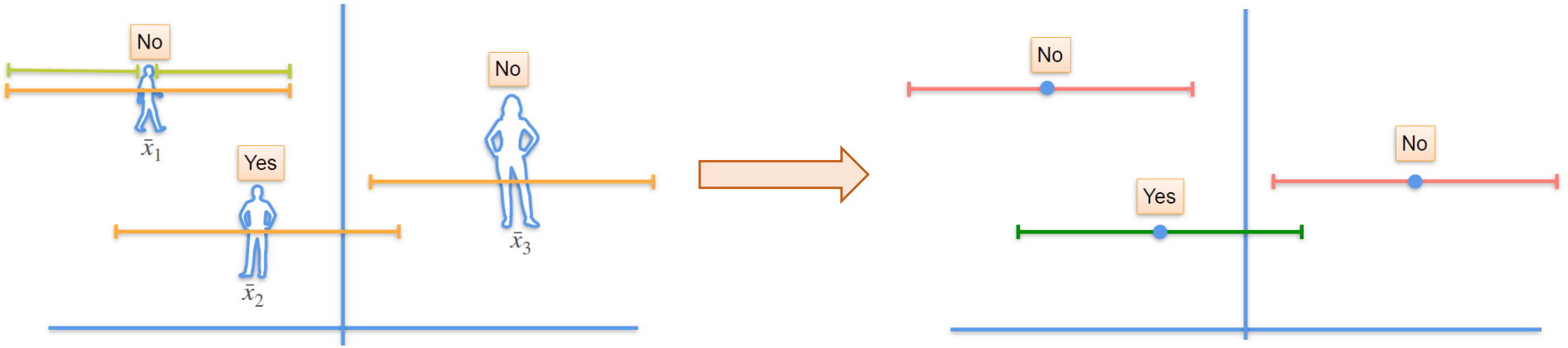
- … Example- sample size = 1
  - Assume that we know σ
    - We calculate the confidence interval for the known confidence level (95%)



… Introduction

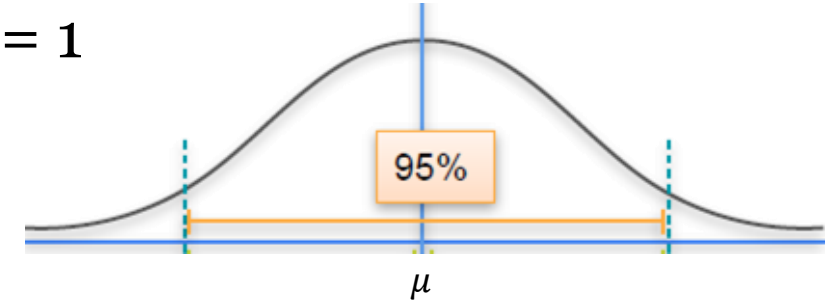
- … Example- sample size = 1



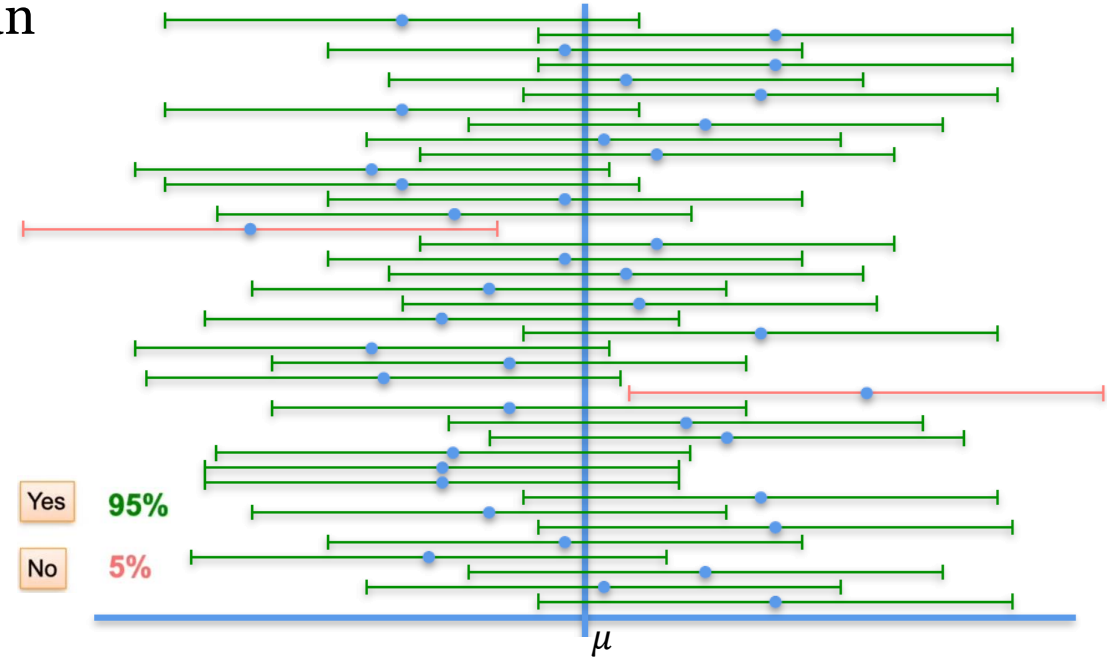


… Introduction

- … Example- sample size = 1



  - Repeat for 100 times
    - 95% of the time, the confidence intervals around the samples (mean) contain the population mean



Mean of the sample means — Population mean

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{1} = \sigma$$

std of the sample means — Population std

… Introduction

- … Example- sample size = 2

Mean of the sample means

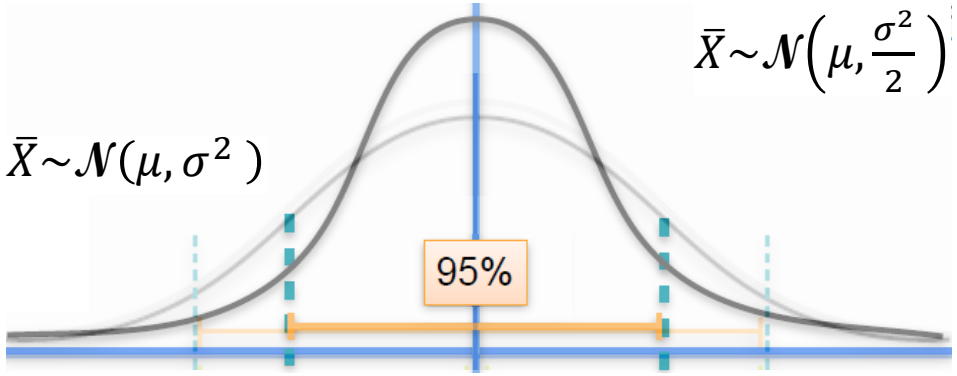
Population mean

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{2}}$$

std of the sample means

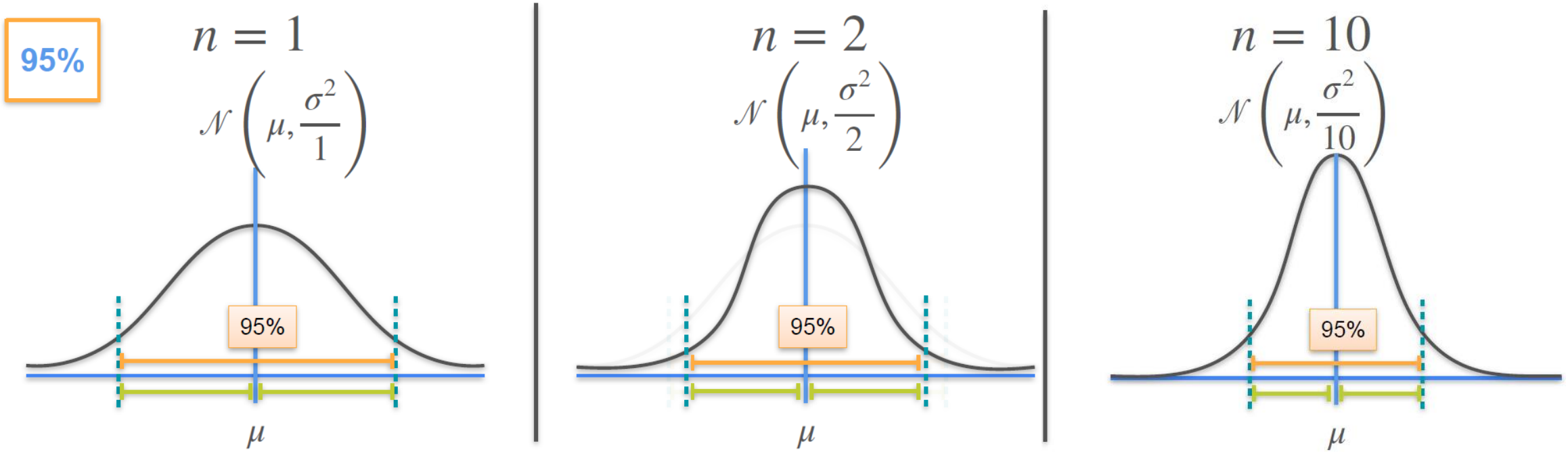
Population std



Margin of error is smaller

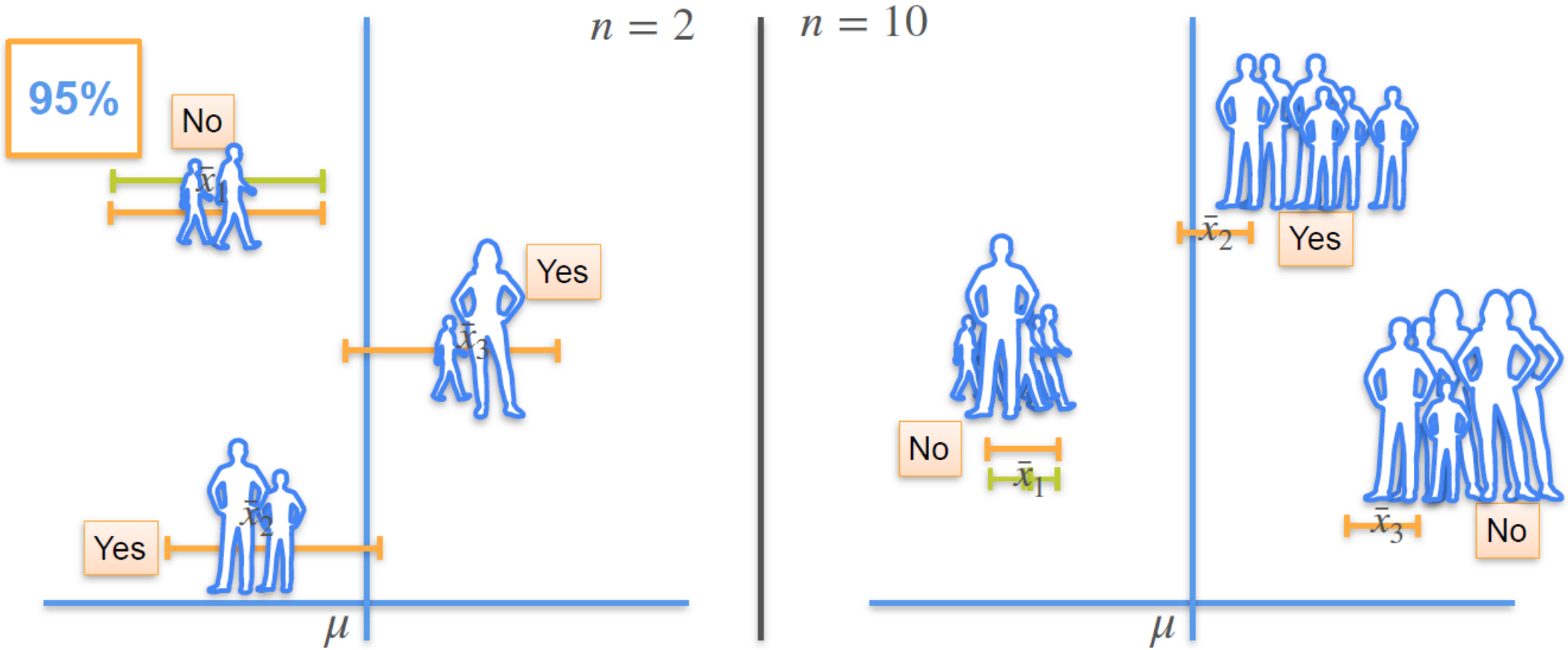
… Introduction

- Effect of the sample size



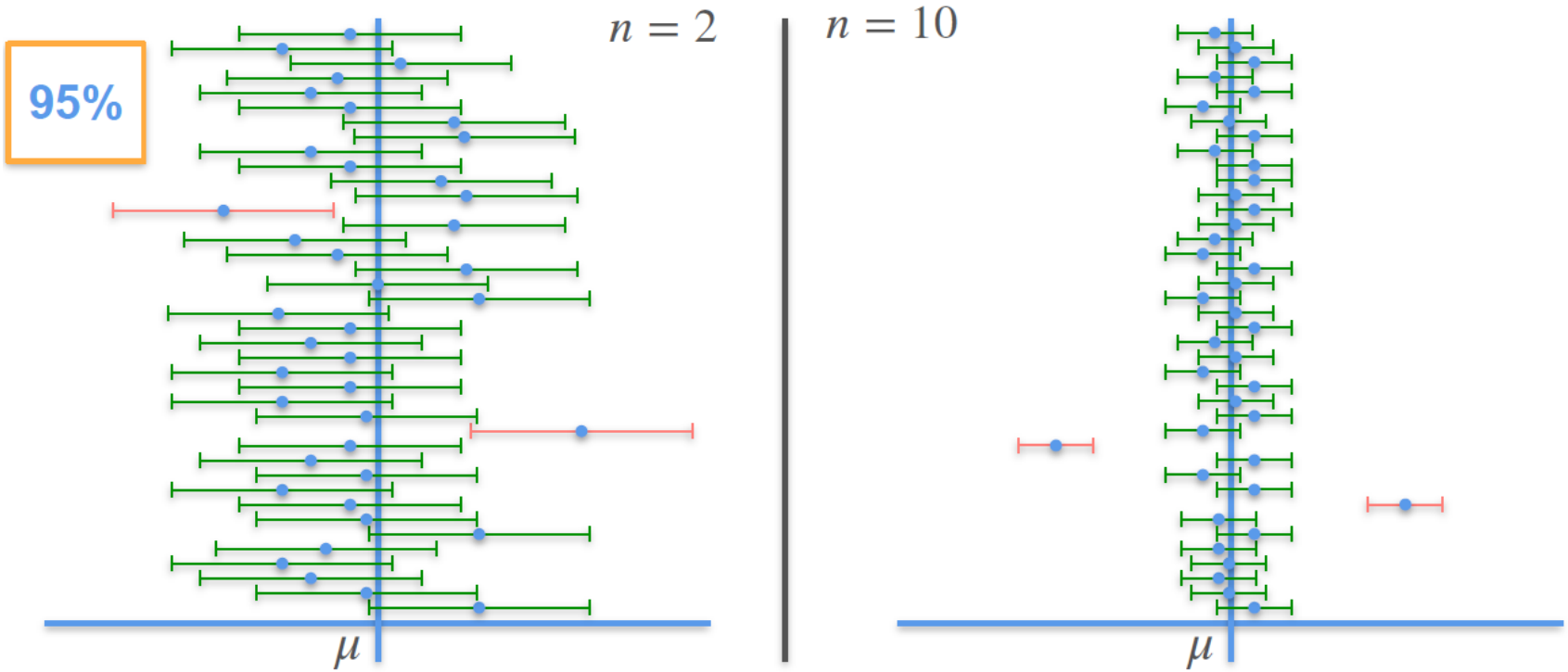
… Introduction

- … Effect of the sample size



… Introduction

- … Effect of the sample size

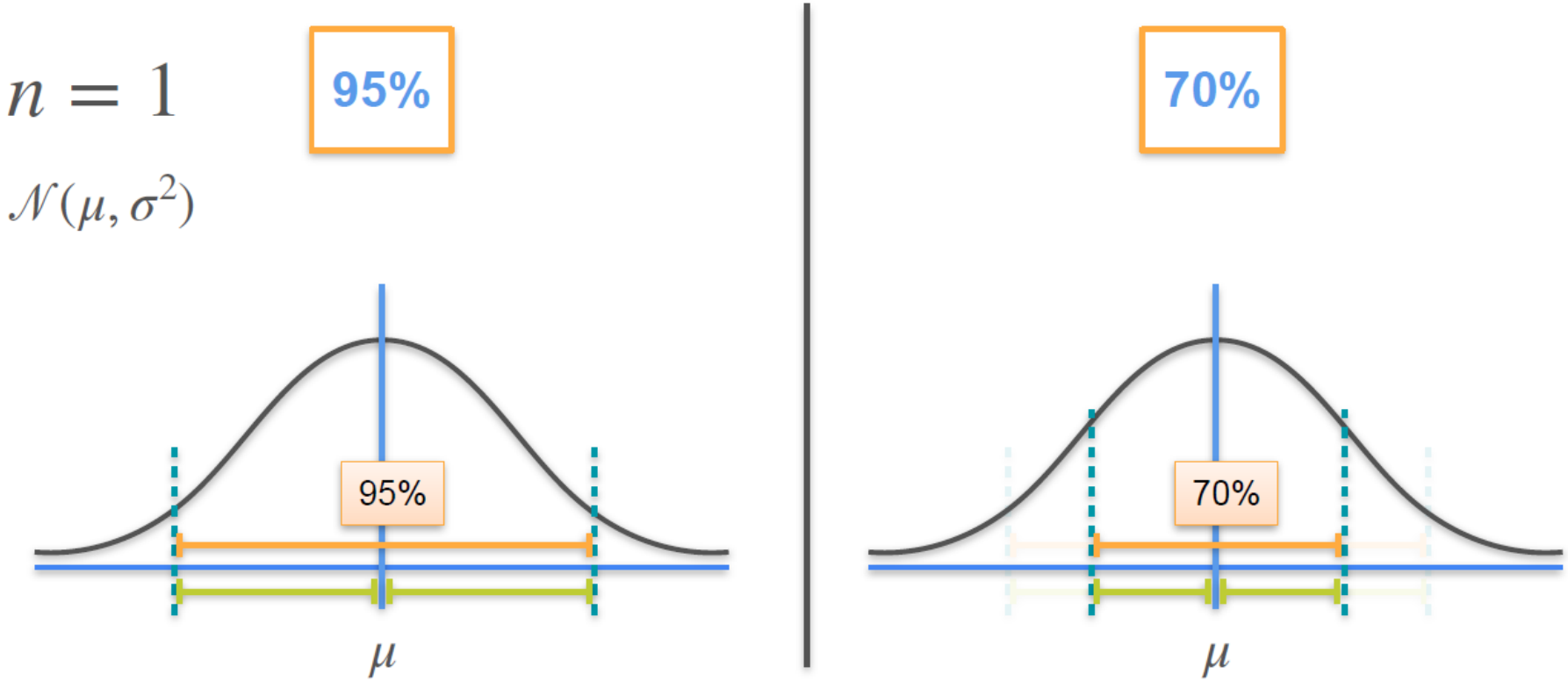


Although we have smaller intervals, we still hit the population mean 95% of the time. Why?

Because the sample means are going to be much closer to the population mean

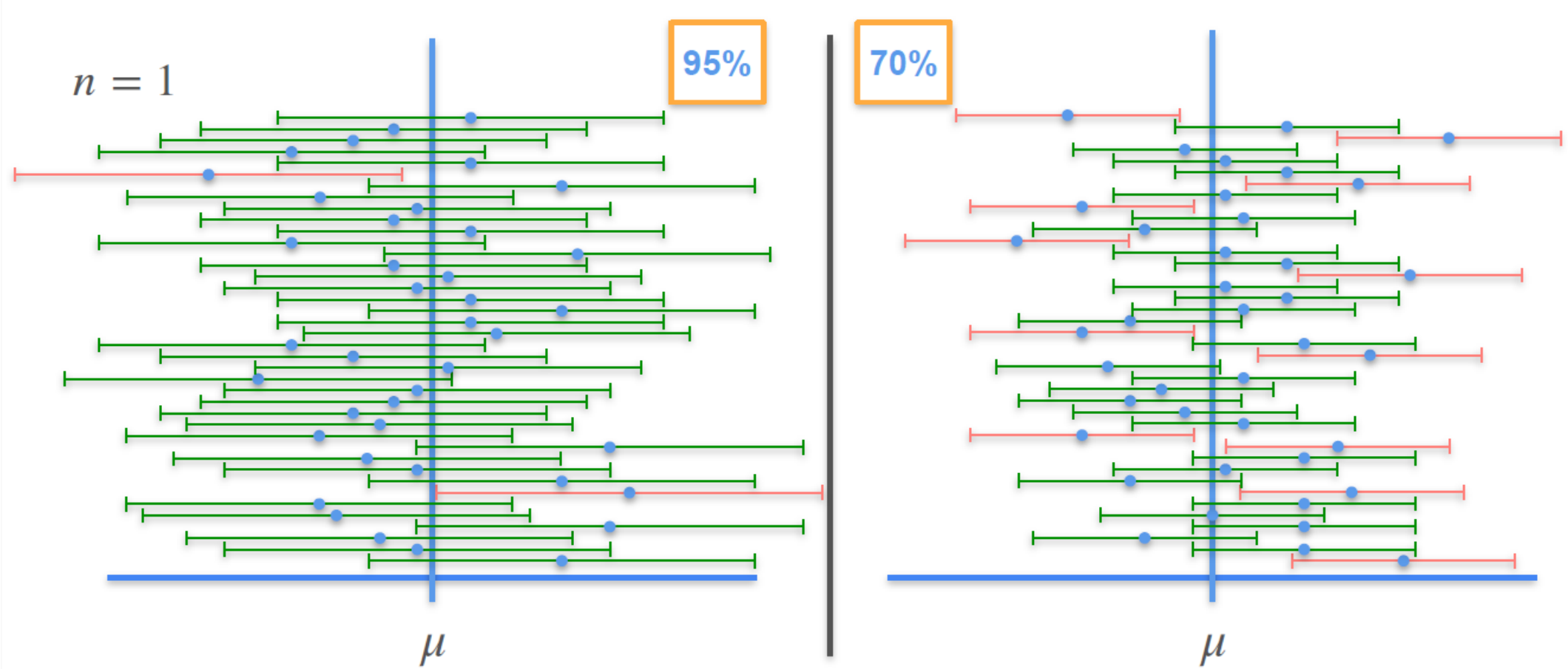
… Introduction

- Effect of the confidence level



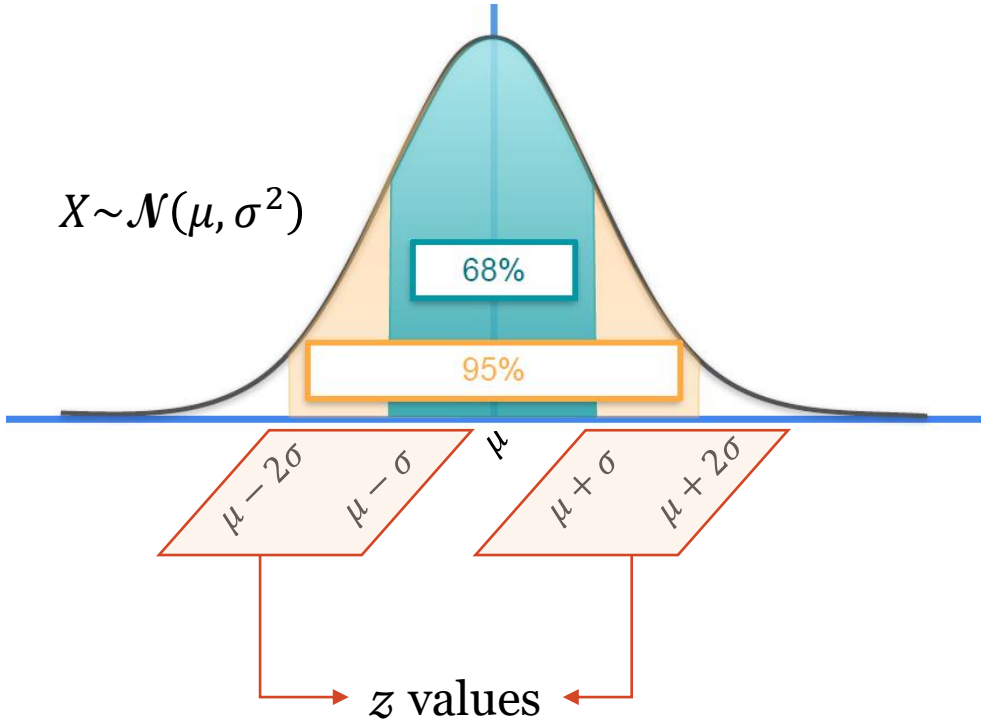
… Introduction

- … Effect of the confidence level



Confidence Interval Calculation

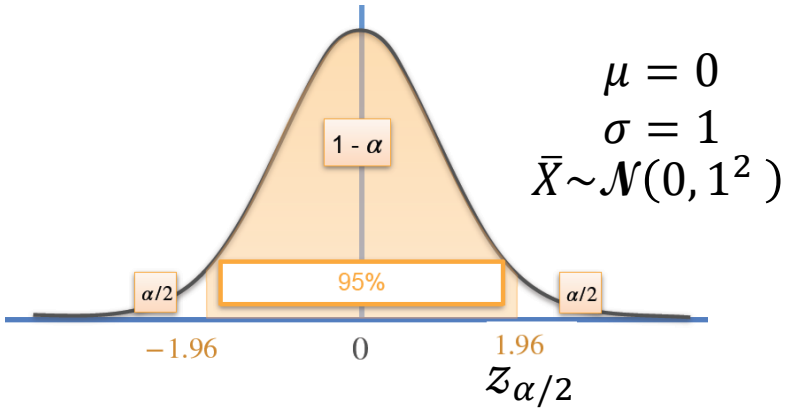
- Margin of error
  - Recall from previous slides
    - 95% of the population falls $2 \times \sigma$ away from the mean



  - Standardize the Normal distribution
    - $\bar{X} \sim \mathcal{N}(0,1)$
    - $\mu \pm 2\sigma \sim \pm 2$
      - The exact points ± 1.96

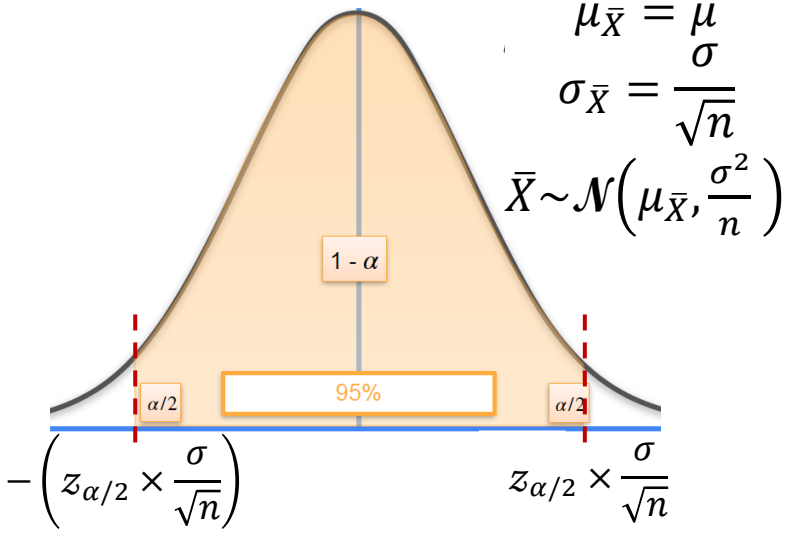
… Confidence Interval Calculation

- … Margin of error
  - For the standardized Normal distribution, $z_{\alpha/2}$ is called the critical value
    - For 95% confidence level $z_{\alpha/2}$ is 1.96



$$\mu = 0$$
$$\sigma = 1$$
$$\bar{X} \sim \mathcal{N}(0, 1^2)$$

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is called the standard error
- Margin of error = $z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$



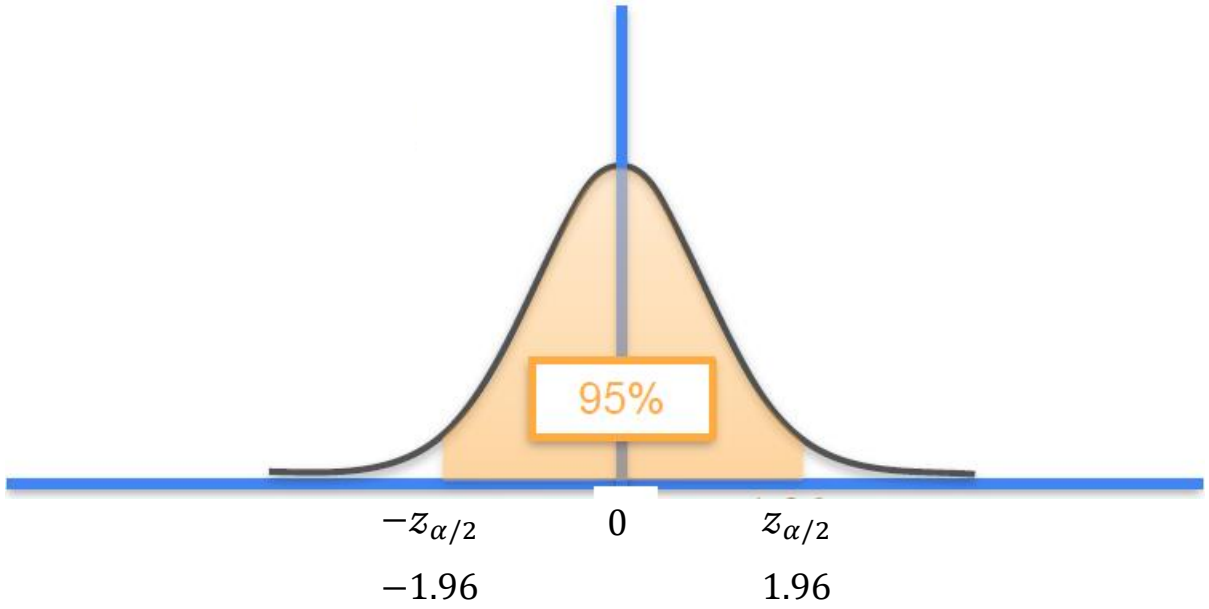
$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$
$$\bar{X} \sim \mathcal{N}\left(\mu_{\bar{X}}, \frac{\sigma^2}{n}\right)$$

… Confidence Interval Calculation

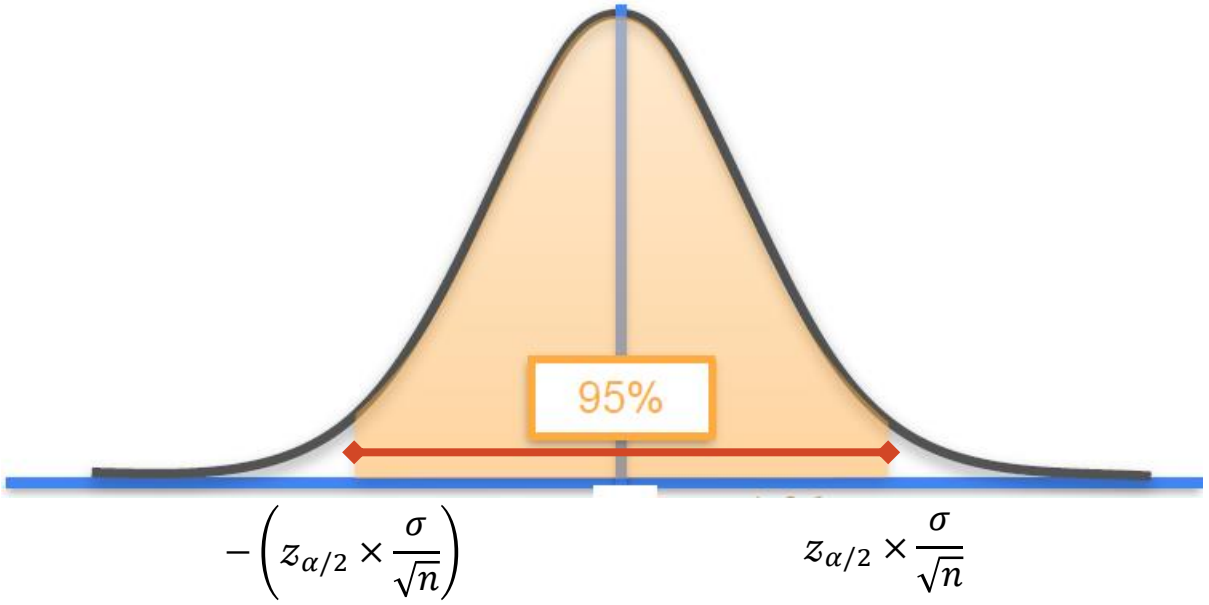
- To find the confidence interval, follow the following steps:

1. Find the sample mean ($\bar{X}$)
2. Define a desired confidence level ($1 - \alpha$), e.g. 95%
3. Get the critical value ($z_{\alpha/2}$), e.g. 1.96
4. Find the standard error ($\frac{\sigma}{\sqrt{n}}$)
5. Calculate the margin of error ($z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$)
6. Add/subtract the margin of error to/from the sample mean
   - Confidence interval = $\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

… Confidence Interval Calculation



… Confidence Interval Calculation



… Confidence Interval Calculation

- Example- Average height of a city with the population 6000
  - Assumptions
    - Random sampling
    - The sample means follow the Normal distribution ($n > 30$)
    - We know that the standard deviation of the population is 25cm ($\sigma = 25$)
    - Sample size (n) is 49
    - The desired confidence level is 95%
  - Calculation
    - $\bar{X} = 170\text{cm}$
    - $z_{\alpha/2} = 1.96$
    - Margin of error = $z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{25}{7} = 7$
    - Confidence interval = $170 \pm 7 = [163, 177]$

We are 95% sure that the average height of the people in the city lies between 163 and 177cm.

Sample Size Calculation

- Example- Continued from the previous slide
  - The 7cm margin of error is a large number.
  - We want to have a smaller margin of error, say 3 cm.
    - This will give us a more exact estimate of the population mean.
  - How can we obtain a smaller confidence interval (be more confident)?
    - Use a bigger sample
  - What is the smallest sample size to obtain the desired margin of error?
    - Use the backward calculation
      - Margin of error = $z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
      - $3 \geq 1.96 \times \frac{25}{\sqrt{n}} \rightarrow \left(\frac{1.96 \times 25}{3}\right)^2 \leq n \rightarrow n \geq 266.78 \approx 267$

… Sample Size Calculation

- Generalized formula
  - The smallest sample size to obtain a particular margin of error, say MOE is calculated by the following formula:

$$n \geq \left(\frac{z_{\alpha/2} \times \sigma}{MOE}\right)^2$$

Unknown Standard Deviation

- Often we do not know the standard deviation of the population (σ)
  - So, we cannot calculate the Confidence interval ($\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$)

- Solution: Use the actual sample standard deviation (s)
  - This change the distribution shape!
    - $\bar{X}$ no longer follows the normal distribution $\mathcal{N}\left(\mu_{\bar{X}}, \frac{\sigma^2}{n}\right)$
    - $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ is no longer a standard Normal distribution $\mathcal{N}(0,1^2)$, it follows a distribution called Student's t-distribution

Normal distribution

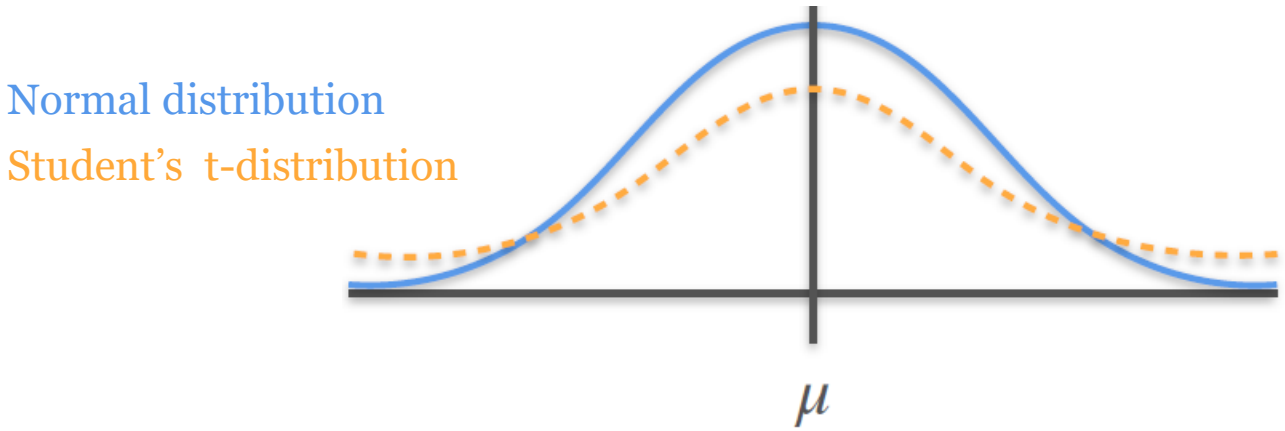
$$\sigma \longrightarrow s$$

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$$

Student's t-distribution

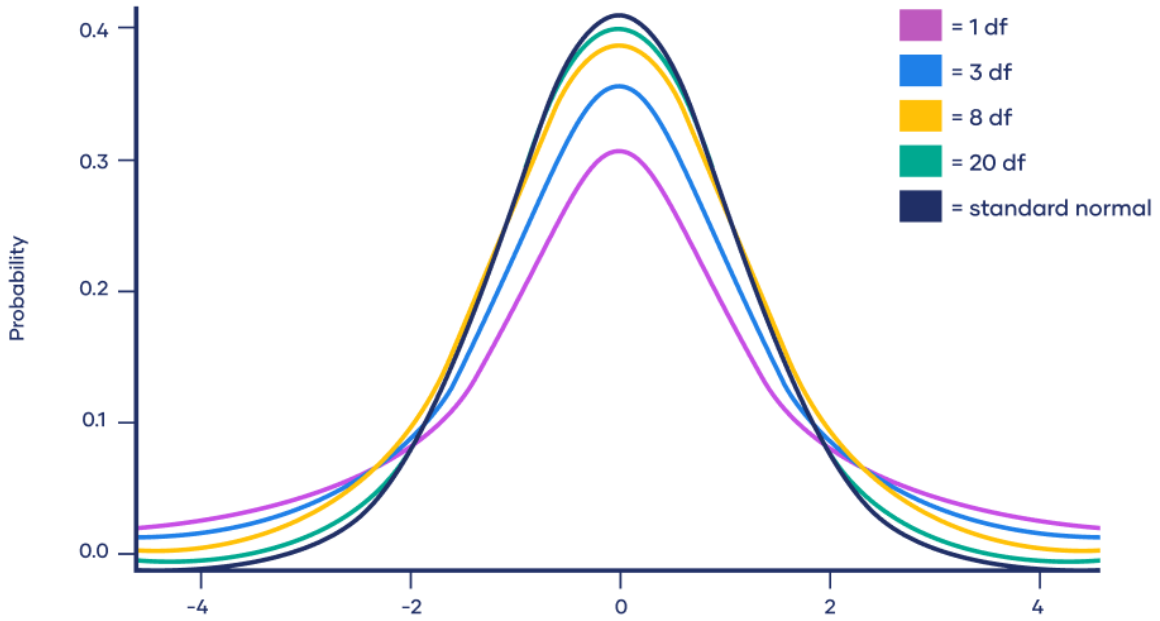
… Unknown Standard Deviation

- Student's t-distribution
  - A continuous probability distribution that generalizes the standard Normal distribution.
  - Like the standard Normal distribution, it is symmetric around zero and bell-shaped
  - In contrast to the Normal distribution, it has thicker tails
    - It means that if you sample a point out of the Student t-distribution, it is more likely to be far from the center than you pick it from the Normal distribution.

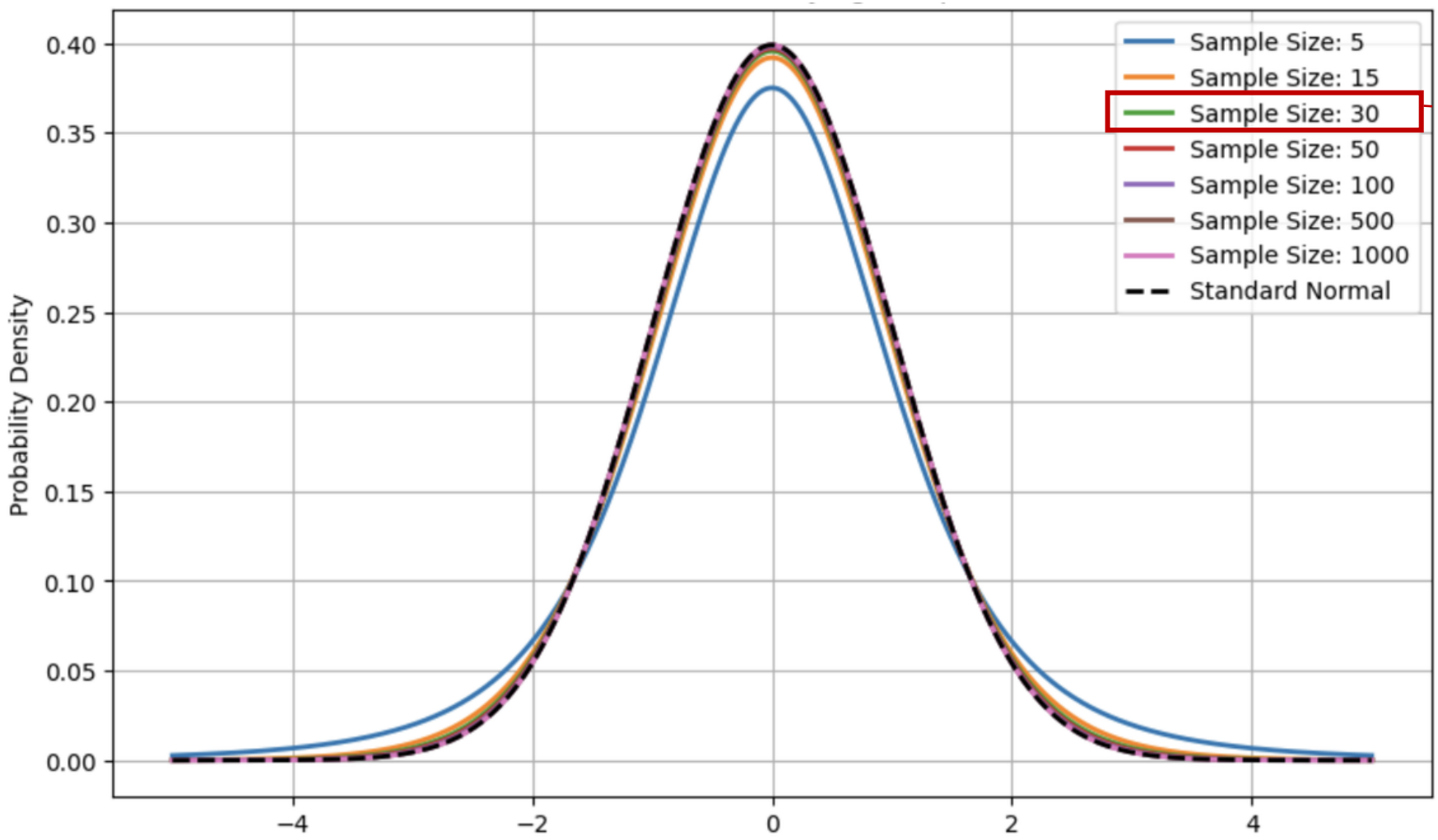


… Unknown Standard Deviation

- … Student's t-distribution
  - It has a parameter called the degrees of freedom and shown by ν
    - $\nu = n - 1$
    - The larger the number of degrees of freedom, the closer you get to a Normal distribution
      - For $\nu = 1$, the Student's t-distribution has heavy tails
      - When $\nu \to \infty$, the Student's t-distribution becomes the standard Normal distribution $\mathcal{N}(0, 1)$



… Unknown Standard Deviation



30 is perfect!

… Unknown Standard Deviation

$1-\alpha$

ν	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
…	…										
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
…	…										
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

… Unknown Standard Deviation

- … Solution
  - Instead of $\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ use the following formula to calculate the confidence interval:

$$\bar{X} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

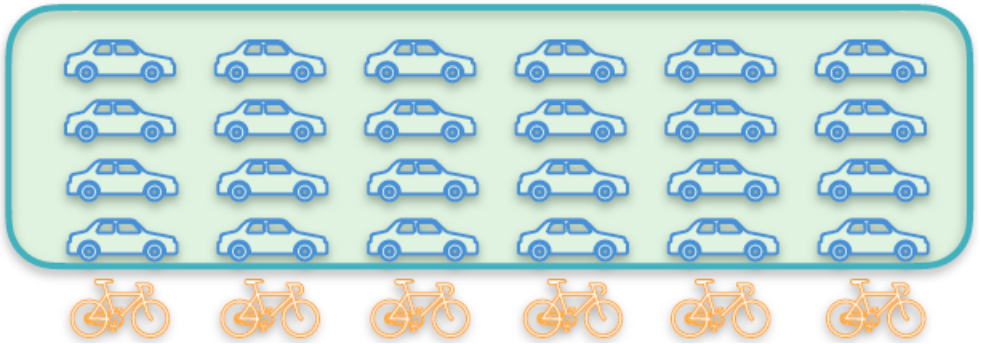
t-score

How to find?

Lookup in the t-table

Confidence Interval for Proportions

- Example- Proportion of people who use car
  - $\hat{p} = \frac{x}{n} = \frac{24}{30} = 80\%$



  - How do you calculate a 95% confidence interval for this sample proportion?

$$\text{Confidence interval} = \hat{p} \pm \text{margin of error}$$

$$\text{margin of error} = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

… Confidence Interval for Proportions

- … Example
  - The desired confidence level is 95%
  - $\hat{p} = 0.8$
  - $z_{\alpha/2} = 1.96$
  - Margin of error = $z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.8(1-0.8)}{30}} = 0.14$
  - Confidence interval = $0.8 \pm 0.14 = [0.66, 0.94]$
    - $66\% < p < 94\%$

… Confidence Interval for Proportions

Confidence Interval for Means

Confidence interval = $\bar{X} \pm$ margin of error

$$\text{margin of error} = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Confidence Interval for Proportions

Confidence interval =$\hat{p} \pm$ margin of error

$$\text{margin of error} = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$