Introduction to Data Science

Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 1; January 15 - 19, 2024)

Instructor

- Dr. Irfan Yousuf
- irfan.yousuf@uet.edu.pk

Weekly Contents

Course Description

This course introduces students to the basics of Data Science including programming in R or Python, statistical inference, exploratory data analysis, basic machine learning algorithms, feature generation and feature selection. The foundation is laid for big data analytics ranging from social networks to business informatics.

Measurable Student Learning Outcomes

CLOs	Description	PLOs	Domain	Domain Level
CLO1	Explain basic statistical modeling and analysis.	PLO-01	Cognitive	2. Understand
CLO2	Use Exploratory Data Analysis (EDA) on different types of data.	PLO-02	Cognitive	3. Apply
CLO3	Compare tools and techniques to mine big datasets and graphs.	PLO-04	Cognitive	4. Analyze

Weekly Contents

Text Books

1. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. Edition: 1st, Year: 2014, Publisher: O'Reilly Media
2.

Grading Policy

Quizzes	20%
Presentations	10%
Mid Term	30%
Final Term	40%

Weekly Contents

Week	Topics	CLO(s)
1	What is data science? Big Data hype, Skill set needed for a data scientist.	1
2	Basic Statistics, Statistical Inference, Probability Distributions,	1
3	Introduction to Python, Basic concepts	1
4	Model fitting in Python, Data handling in Python	1
5	Exploratory Data Analysis, EDA in data science, Basic tools (graphs, summary).	2
6	Case studies on Exploratory Data Analysis, Social, Medical and Business Data Exploration	2
7	Basic Machine Learning Algorithms (Linear Regression, k-nearest neighbors, k-means)	3
8	Naïve Bayes, Spam Filtering	3

Weekly Contents

9	Feature Generation and Selection, Filters; Wrappers; Decision Trees; Random Forests.	3
10	Recommendation Systems, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis	3
11	Basics of Graph Theory, Big real-world Graphs, Handling graphs in Python	1
12	Social Networks as Graphs, Characteristics of Social Graphs, Social Network Analysis	2
13	Bipartite Graphs, Business Graphs,	2
14	Data Visualization, Different tools to visualize data.	3
15	Data Science and Ethical Issues, Discussions on privacy, security, ethics, A look back at Data Science, Next-generation data scientists	2, 3
16	Revision	1,2,3

Why Data Science?

- One of the topmost professions
- New driving force behind industries is Data.
- Data Science is the Career of Tomorrow.

Skill Set Needed

- Statistics
- Programming skills
- Multivariable Calculus & Linear Algebra

Statistics

- In plural form, it refers to set of **numerical data**.

- In singular form, it is an academic discipline.

Data

- Facts and statistics collected for reference or analysis.

- Data are units of information, often numeric, that are collected through observation.

- Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

What is Statistics

- Statistics is a branch of mathematics that deals with the scientific **collection, organization, presentation, analysis,** and **interpretation** of numerical data in order to obtain **useful** and **meaningful** information.

Descriptive Statistics

- A statistical method concerned with the **collection, organization, presentation** and **description** of sample data.

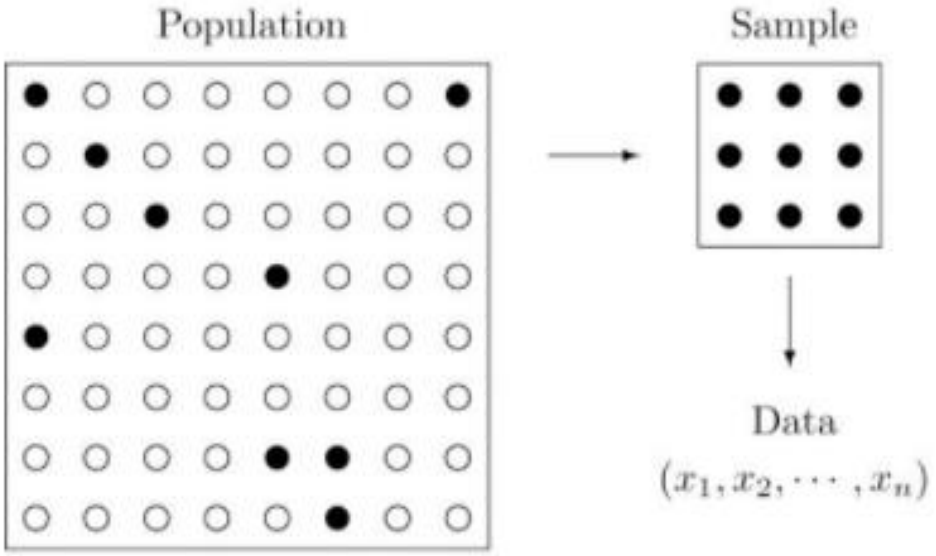
Inferential Statistics

- Inferential Statistics concerned with the analysis of a sample data leading to **prediction, inferences, interpretation, decision** or **conclusion** about the entire population

Population vs. Sample

- Population: The totality of all the elements or persons for which one has an interest at a particular time.
  - Students of 2018 session of CS-KSK

- Sample: It is a subset of a population
  - Students with CGPA > 3.0



Parameter vs. Statistic

- A parameter is a number describing a whole population.

- A statistic is a number describing a sample.

- With inferential statistics, we use sample statistics to make educated guesses about population parameters.

Quantitative vs. Qualitative Data

- Quantitative: These are numerical information obtained from counting or measuring that which can be manipulated by any fundamental operation.
  - Age, Weight, Height

- Qualitative: These are descriptive attributes and characterized by categorical responses.
  - Gender, Weather, Attitude

Variable

- A variable is any characteristics, number, or quantity that can be measured or counted.

- Independent variables: Variables you manipulate in order to affect the outcome of an experiment, e.g., Age

- Dependent variables: Variables that represent the outcome of the experiment, e.g., Salary

Descriptive vs. Inferential Statistics

- Descriptive: concerned with the collection, organization, presentation and description of sample data.

- Inferential: concerned with the analysis of a sample data leading to prediction, inferences, interpretation, decision or conclusion about the entire population

Inferential Statistics

- Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn.
- Because the goal of inferential statistics is to draw conclusions from a sample and generalize them to a population, we need to have confidence that our sample accurately reflects the population.

  - Define the population we are studying.
  - Draw a representative sample from that population.
  - Use analyses that incorporate the sampling error.

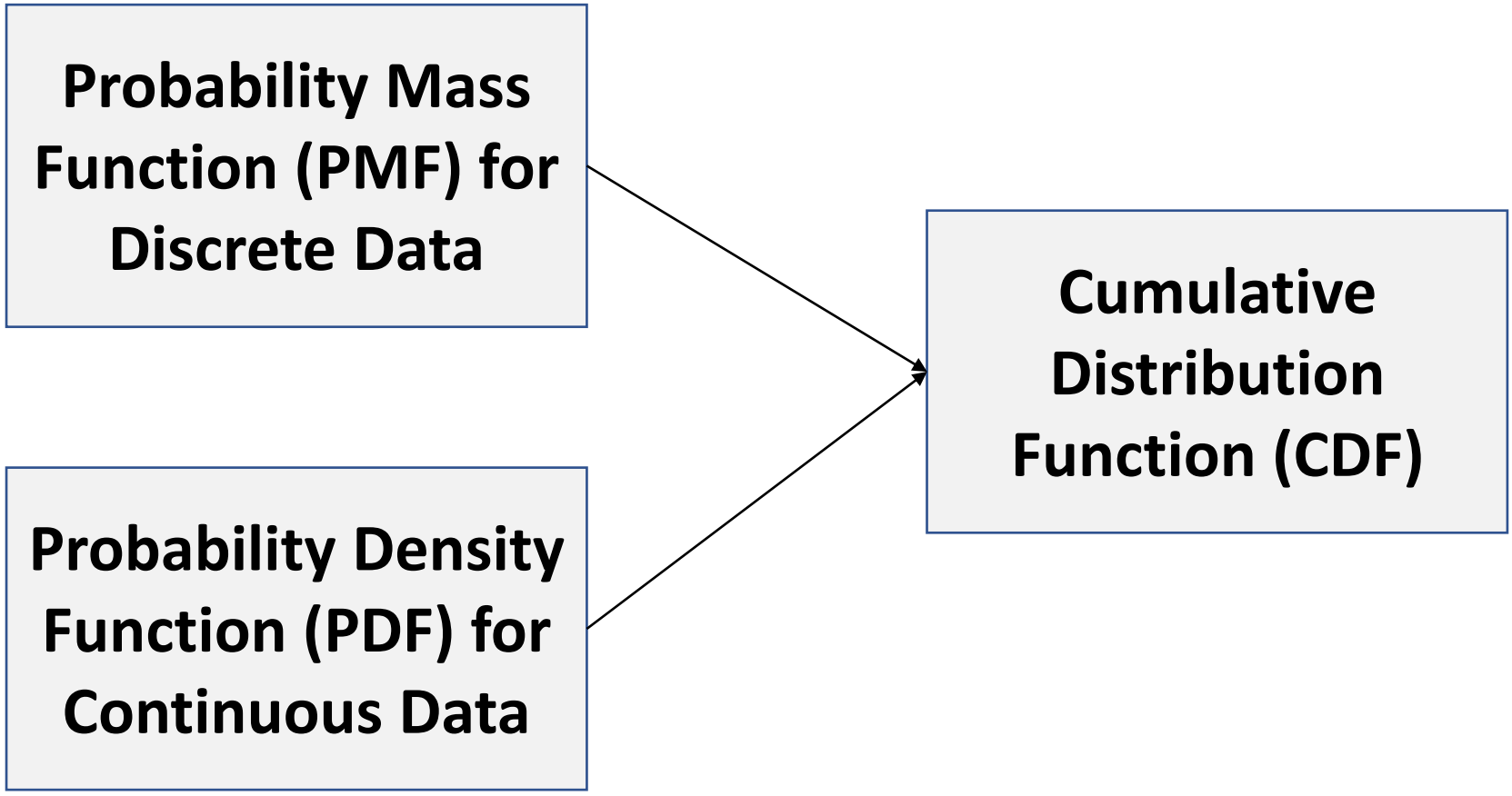
Probability Distributions

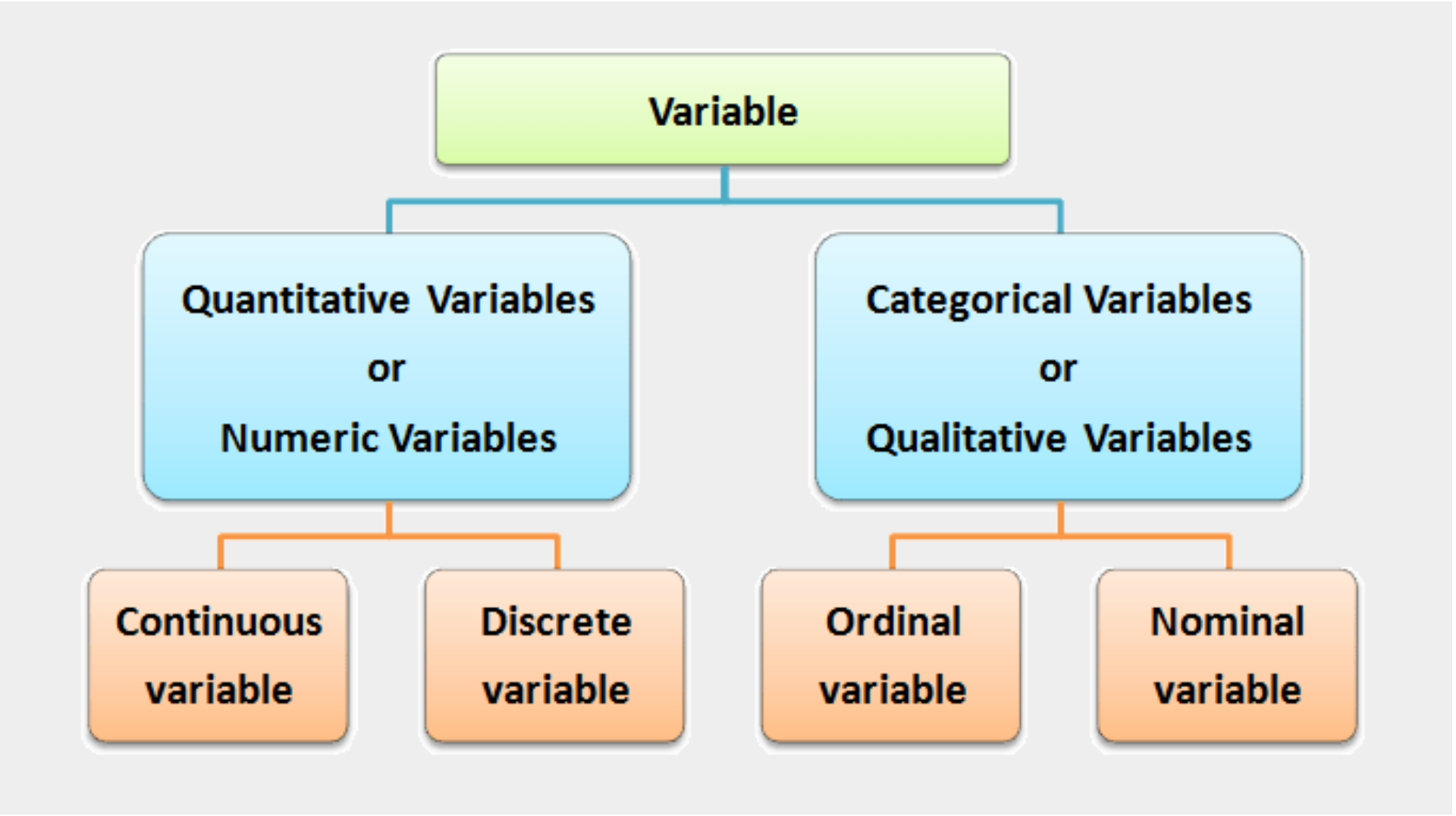
- A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes of an experiment.

  - Tossing a coin
  - throwing a fair die

- Probability distributions are typically defined in terms of the probability distribution functions.

Probability Distribution Functions




Variable
Quantitative Variables
or
Numeric Variables
Categorical Variables
or
Qualitative Variables
Continuous variable
Discrete variable
Ordinal variable
Nominal variable

Discrete vs. Continuous Variable

- A **discrete variable** is a variable that takes on distinct, countable values. In theory, you should always be able to count the values of a discrete variable.

- A **continuous variable** is a variable that can take on any value within a range. Because the possible values for a continuous variable are infinite, we measure continuous variables (rather than count),

Probability Density Functions (PDFs)

- For a discrete random variable X that takes on a finite or countably infinite number of possible values, we determine $P(X=x)$ for all the possible values of X, and call it the **probability mass function** (pmf)

  Example: If you roll a fair six-sided die, the PMF tells you the probability of getting each possible outcome (1, 2, 3, 4, 5, or 6).

- For continuous random variables, the probability that X takes on any particular value x is 0. That is, finding $P(X=x)$ for a continuous random variable is not going to work. Instead, we'll need to find the probability that falls in some interval (a,b) , that is, we'll need to find $P(a < X < b)$. We'll do that using a **probability density function** (pdf).

  Example: If you measure the height of people, the PDF tells you how likely it is to find someone with a height within a certain range (e.g., between 5.5 feet and 6 feet).

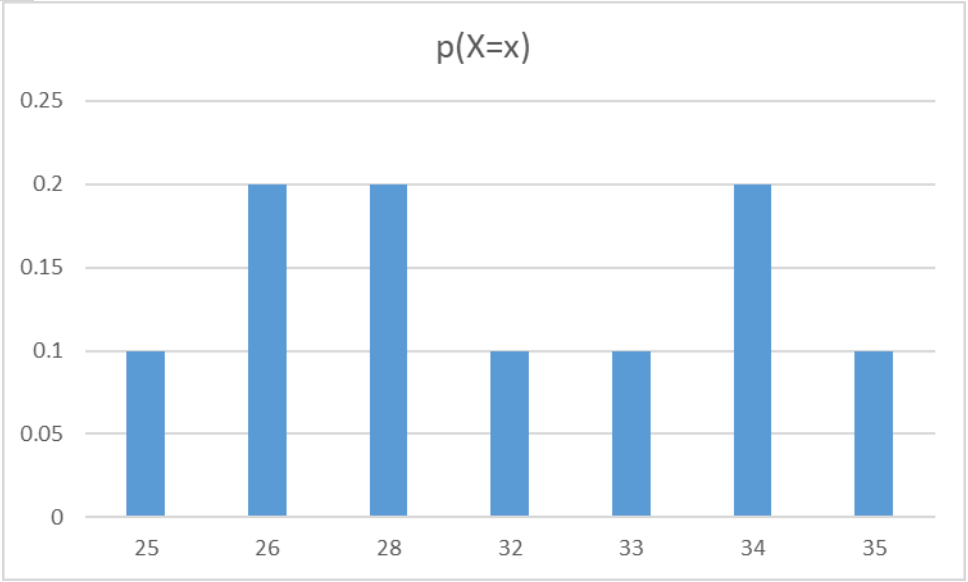
Probability Mass Function

Day	Travel Time (min)	pms
1	25	0.1
2	26	0.2
3	26	0.2
4	28	0.2
5	28	0.2
6	32	0.1
7	33	0.1
8	34	0.2
9	34	0.2
10	35	0.1

X	p(X=x)
25	0.1
26	0.2
28	0.2
32	0.1
33	0.1
34	0.2
35	0.1

$$0 \leq P(X = x) \leq 1 \quad \text{for any } x$$

$$\sum_x P(X = x) = 1$$

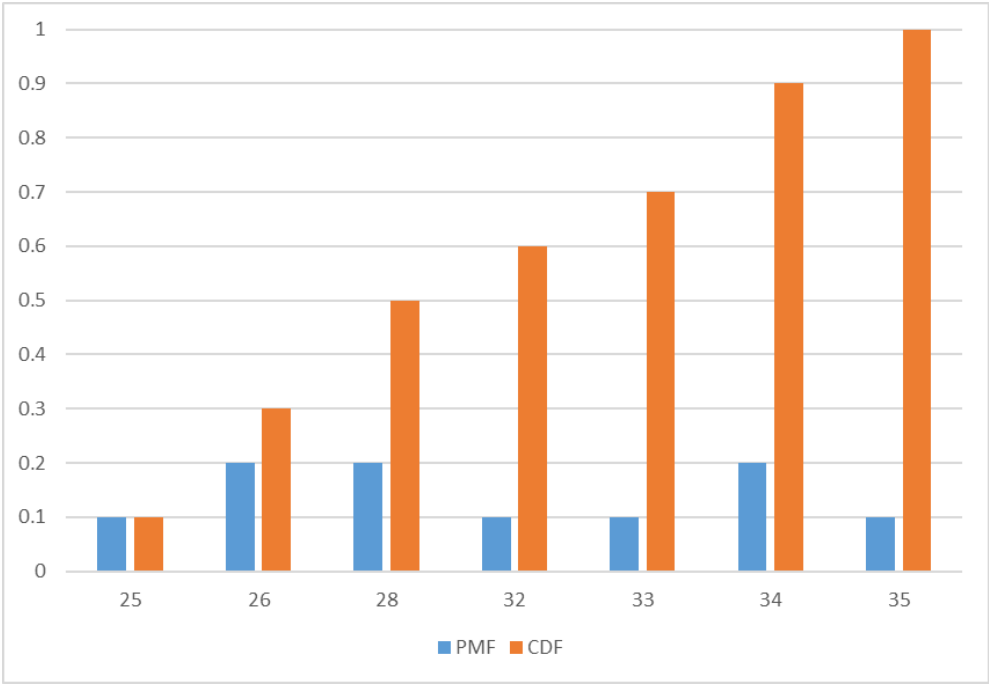


Cumulative Distribution Function of PMF

Day	Travel Time (min)	pms
1	25	0.1
2	26	0.2
3	26	0.2
4	28	0.2
5	28	0.2
6	32	0.1
7	33	0.1
8	34	0.2
9	34	0.2
10	35	0.1

X	PMF	CDF
25	0.1	0.1
26	0.2	0.3
28	0.2	0.5
32	0.1	0.6
33	0.1	0.7
34	0.2	0.9
35	0.1	1

$$F(x) = P(X \leq x)$$



Probability Density Function

The ***probability density function*** **(pdf)**, denoted f, of a continuous random variable X satisfies the following:

1. $f(x) \geq 0$, for all $x \in \mathbb{R}$
2. f is piecewise continuous
3. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
4. $P(a \leq X \leq b) = \int_{b}^{a} f(x)\,dx$

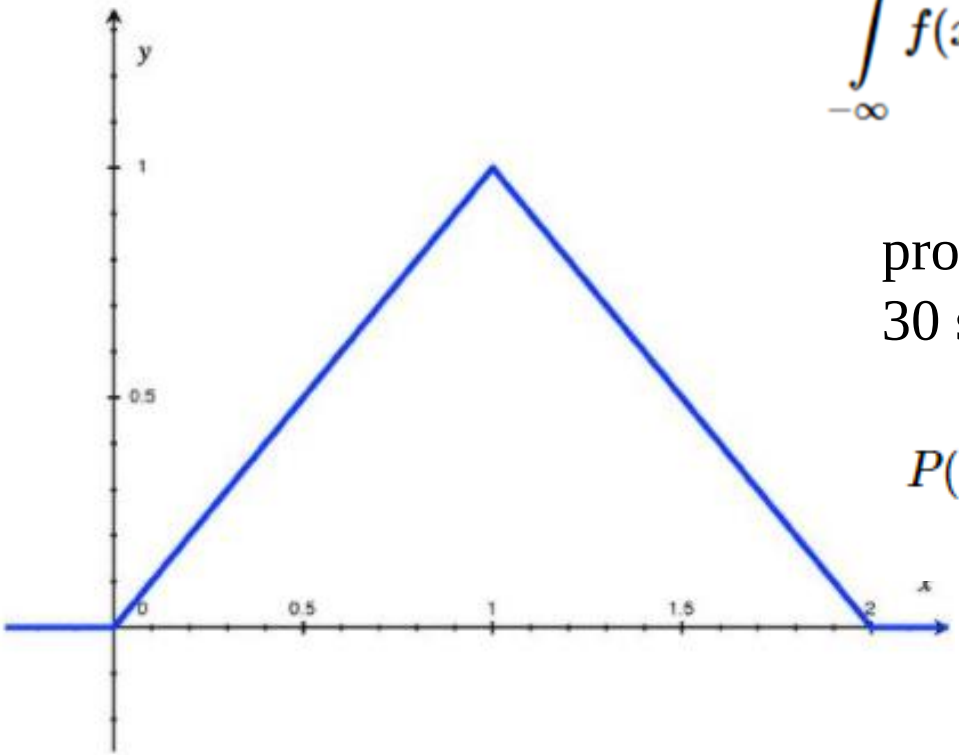
Probability Density Function

Let the random variable X denote the time a person waits for an elevator to arrive. Suppose the longest one would need to wait for the elevator is 2 minutes, so that the possible values of X (in minutes) are given by the interval $[0,2]$.
A possible pdf for X is given by:

$$f(x) = \begin{cases} x, & \text{for } 0 \leq x \leq 1 \\ 2 - x, & \text{for } 1 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Probability Density Function

$$f(x) = \begin{cases} x, & \text{for } 0 \le x \le 1 \\ 2 - x, & \text{for } 1 < x \le 2 \\ 0, & \text{otherwise} \end{cases}$$



$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{2} x\,dx = \int_{0}^{1} x\,dx + \int_{1}^{2} (2-x)\,dx = 1$$

probability that a person waits less than 30 seconds (or 0.5 minutes).

$$P(0 \le X \le 0.5) = \int_{0}^{0.5} f(x)\,dx = \int_{0}^{0.5} x\,dx = 0.125$$

Integral Formula $\int x^n dx = \frac{x^{n+1}}{(n+1)} + c$

Probability Density Function

Continuous random variables have **zero point probabilities**, i.e., the probability that a continuous random variable equals a single value is always given by 0.

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)\,dx = 0.$$

Probability for a continuous random variable is given by areas under pdf's.

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = \int_a^b f(x)\,dx$$

Cumulative Distribution Function of PDF

Let X have pdf f , then the cdf F is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \quad \text{for } x \in \mathbb{R}.$$

Let X be a continuous random variable with pdf f and cdf F.

- By definition, the cdf is found by *integrating* the pdf:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

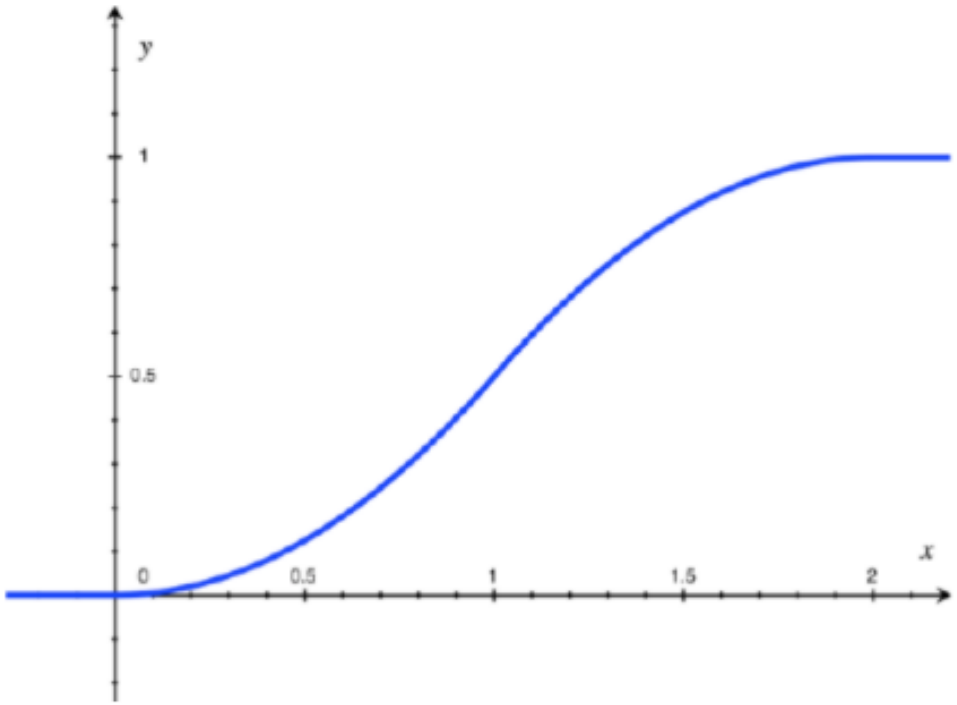
- By the Fundamental Theorem of Calculus, the pdf can be found by *differentiating* the cdf:

$$f(x) = \frac{d}{dx}[F(x)]$$

Cumulative Distribution Function of PDF

PDF to CDF

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{x^2}{2}, & \text{for } 0 \le x \le 1 \\ 2x - \frac{x^2}{2}, & \text{for } 1 < x \le 2 \\ 1, & \text{for } x > 2 \end{cases}$$



Normal Distribution

Normal Distribution Formula

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

$\mu =$ mean of x

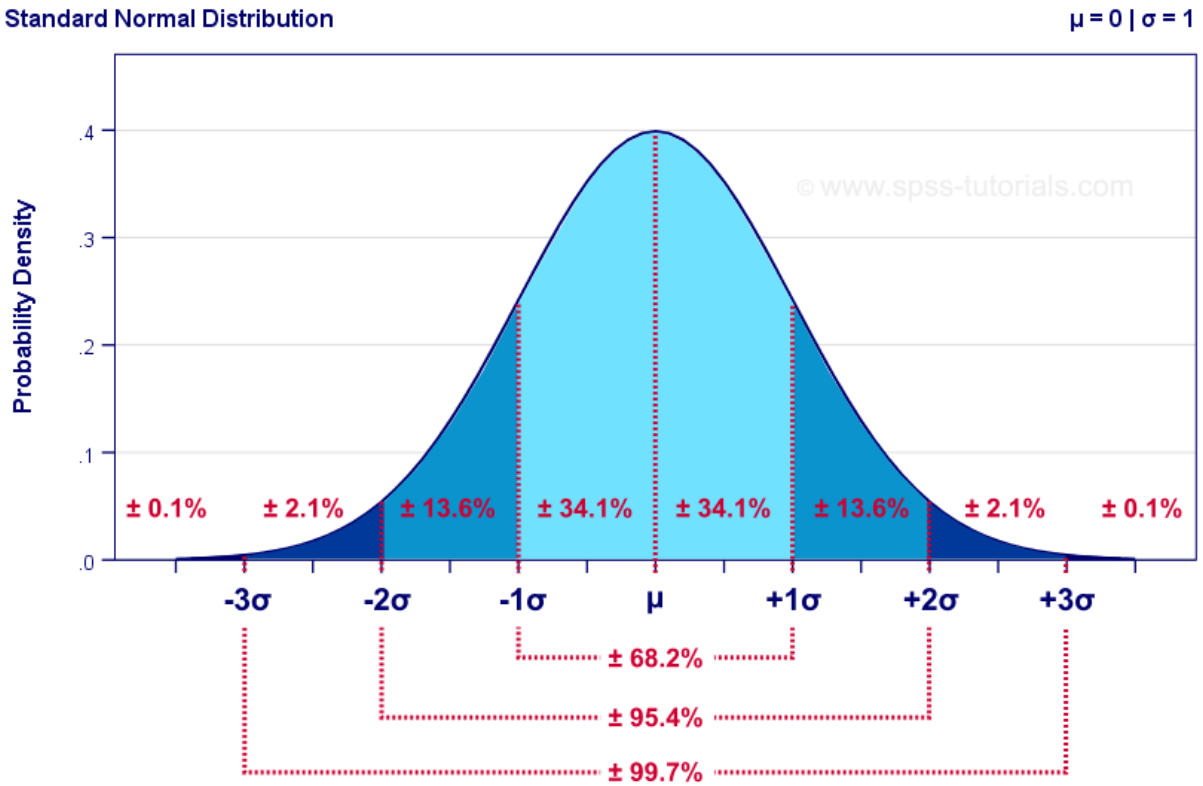
$\sigma =$ standard deviation of x

$\pi \approx 3.14159 \ldots$

$e \approx 2.71828 \ldots$

Normal Distribution

- The mean, median and mode are **exactly the same**.
- The distribution is **symmetric about the mean**—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: **the mean and the standard deviation**.



Normal Distribution

Day	Time
1	32.14
2	31.30
3	29.17
4	28.15
5	30.30
6	30.41
7	32.37
8	33.19
9	31.19
10	30.37
11	28.24
12	29.10
13	28.34
14	28.50
15	29.26
16	28.29
17	25.36
18	27.18
19	30.29
20	27.15

Normal Distribution Formula

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

$\mu =$ mean of x

$\sigma =$ standard deviation of x

$\pi \approx 3.14159 \ldots$

$e \approx 2.71828 \ldots$

Normal Distribution

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

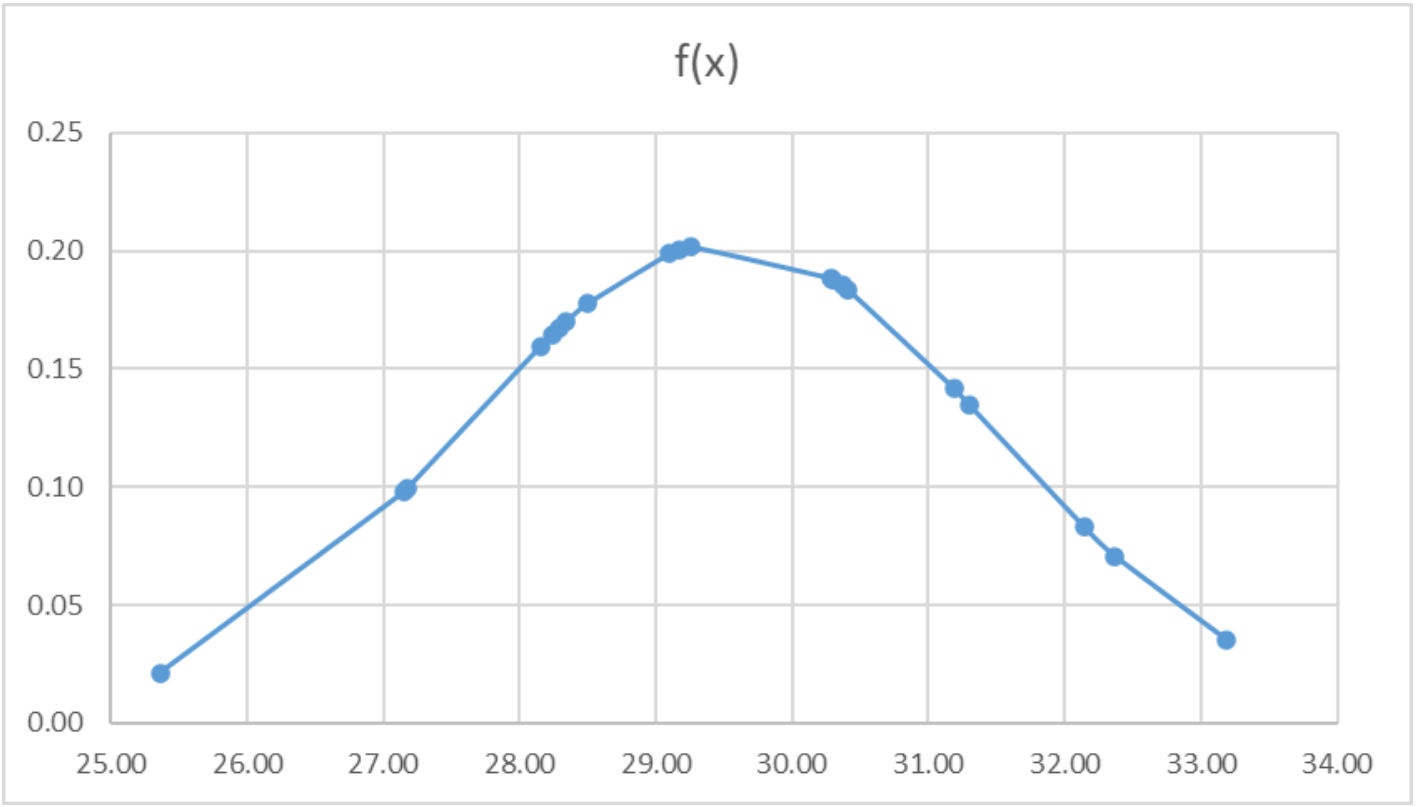
Normal Distribution

Day	Time	f(x)
1	32.14	0.08
2	31.30	0.13
3	29.17	0.20
4	28.15	0.16
5	30.30	0.19
6	30.41	0.18
7	32.37	0.07
8	33.19	0.04
9	31.19	0.14
10	30.37	0.19
11	28.24	0.16
12	29.10	0.20
13	28.34	0.17
14	28.50	0.18
15	29.26	0.20
16	28.29	0.17
17	25.36	0.02
18	27.18	0.10
19	30.29	0.19
20	27.15	0.10

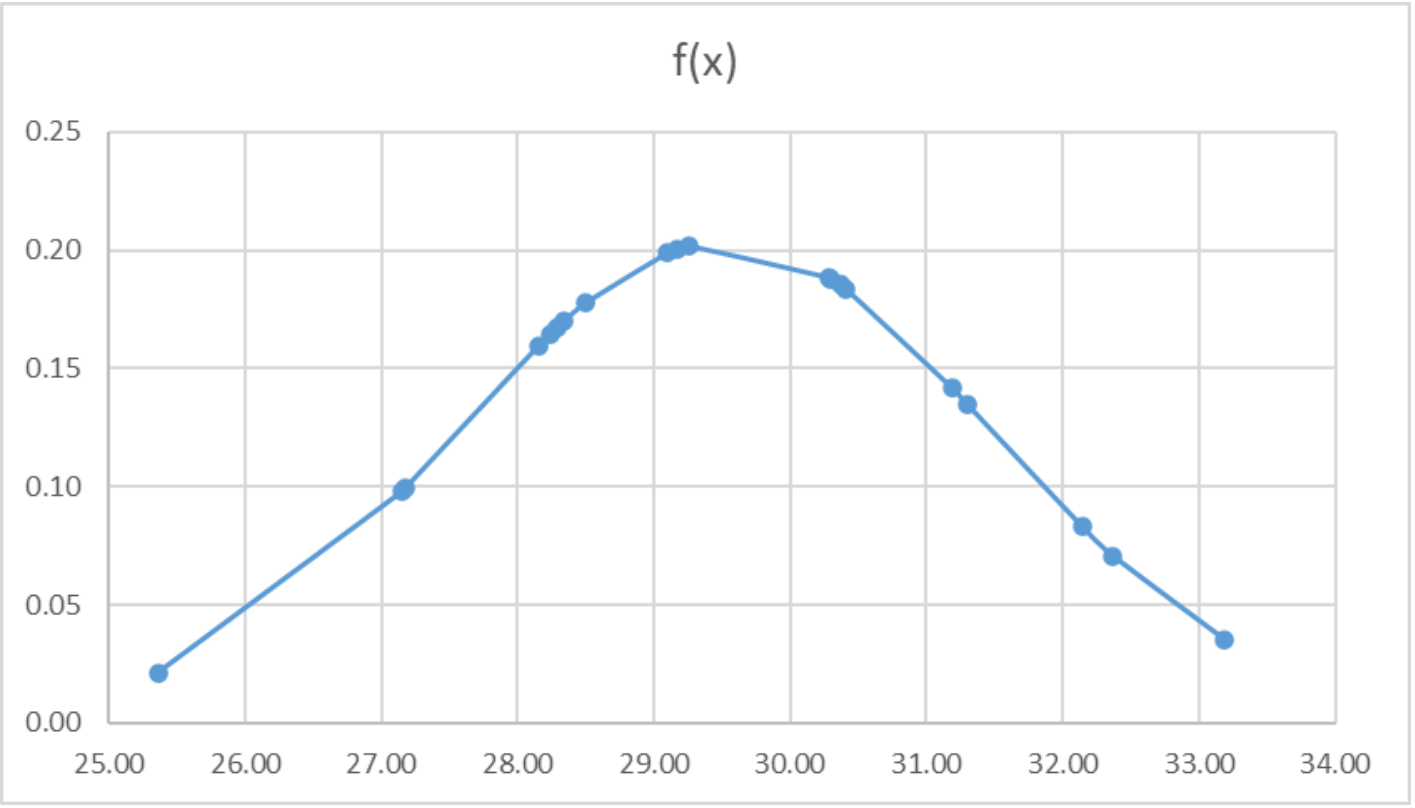
Mean	29.52
St. Dev	1.96

Normal Distribution

Time	f(x)
25.36	0.02
27.15	0.10
27.18	0.10
28.15	0.16
28.24	0.16
28.29	0.17
28.34	0.17
28.50	0.18
29.10	0.20
29.17	0.20
29.26	0.20
30.29	0.19
30.30	0.19
30.37	0.19
30.41	0.18
31.19	0.14
31.30	0.13
32.14	0.08
32.37	0.07
33.19	0.04



Normal Distribution

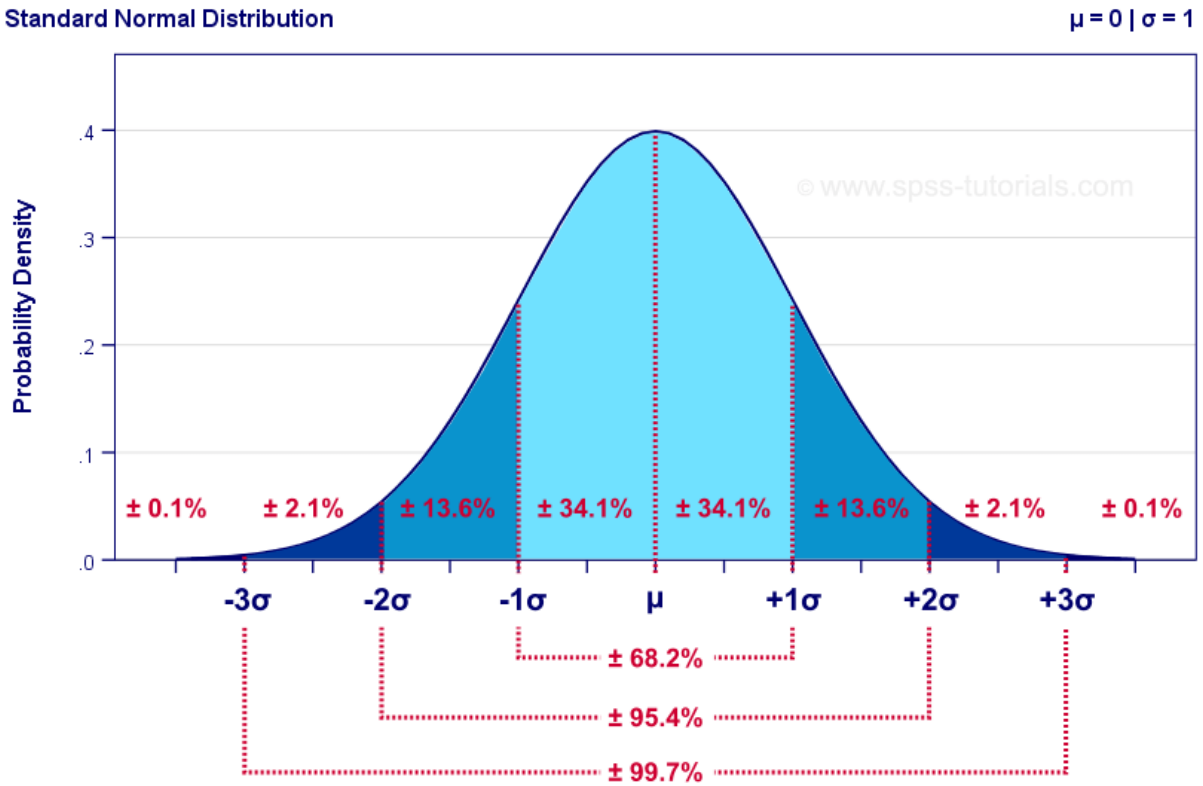


Mean	29.52
St. Dev	1.96

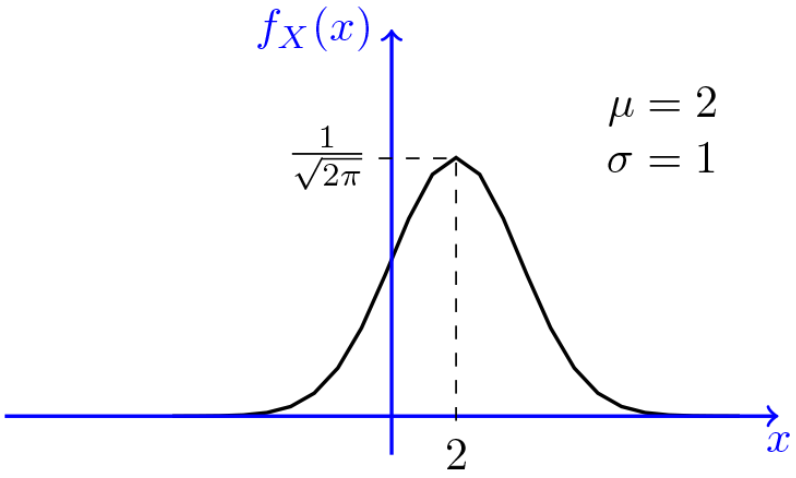
M+SD	31.48
M-SD	27.55

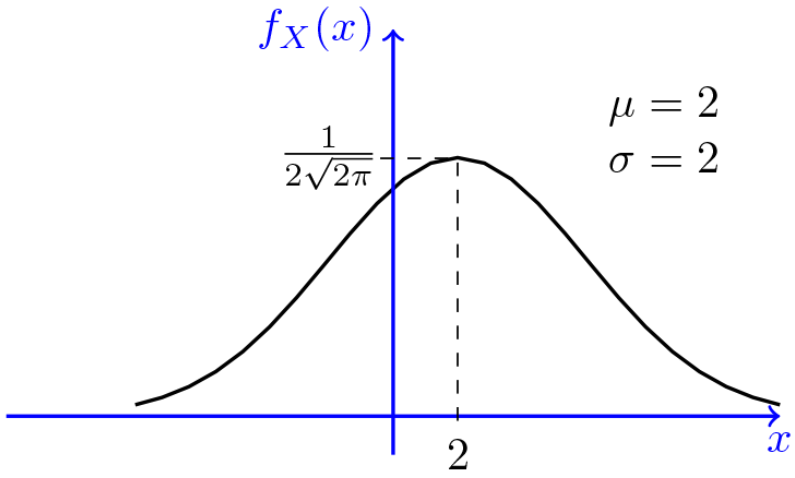
Normal Distribution

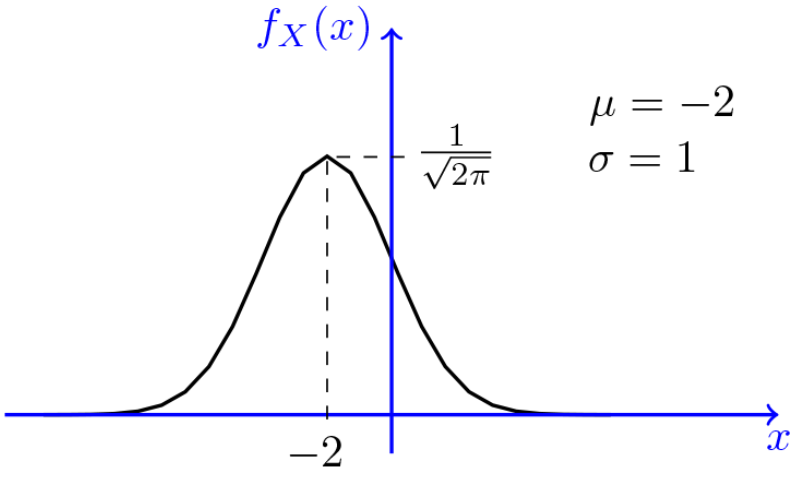
- The mean, median and mode are **exactly the same**.
- The distribution is **symmetric about the mean**—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: **the mean and the standard deviation**.

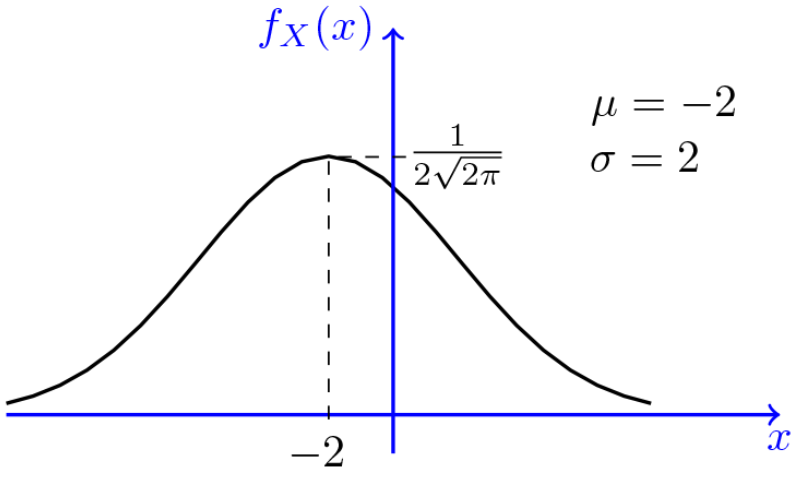


Normal Distribution







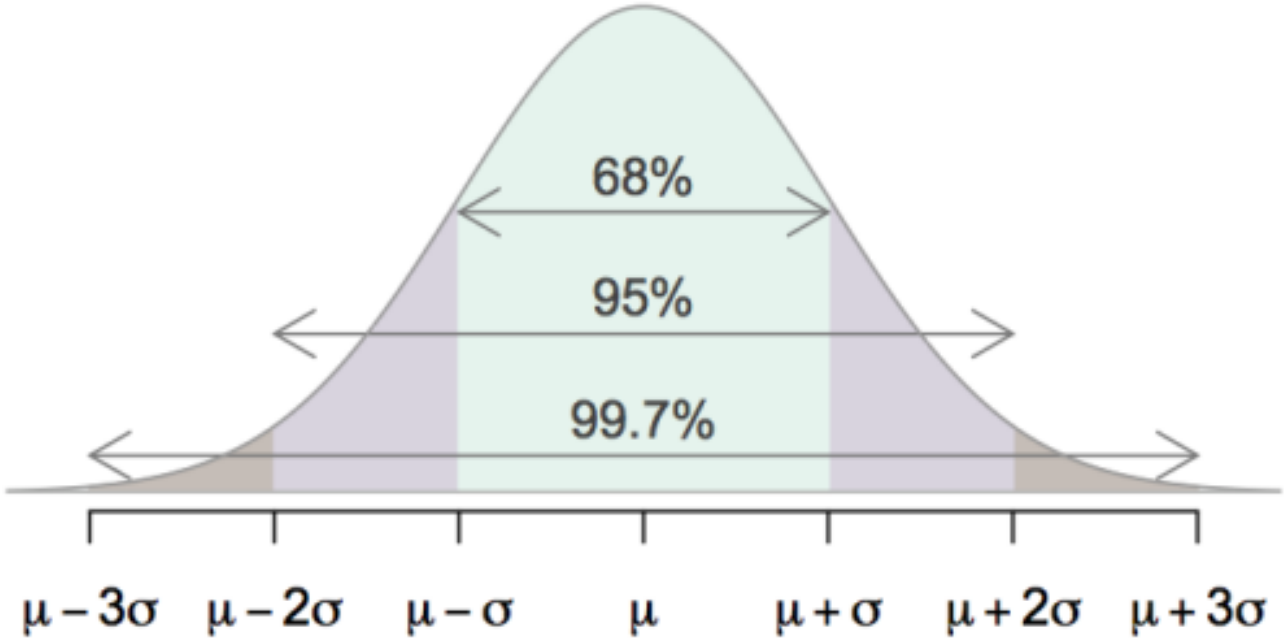


68-95-99.7 Rule

For nearly normally distributed data,

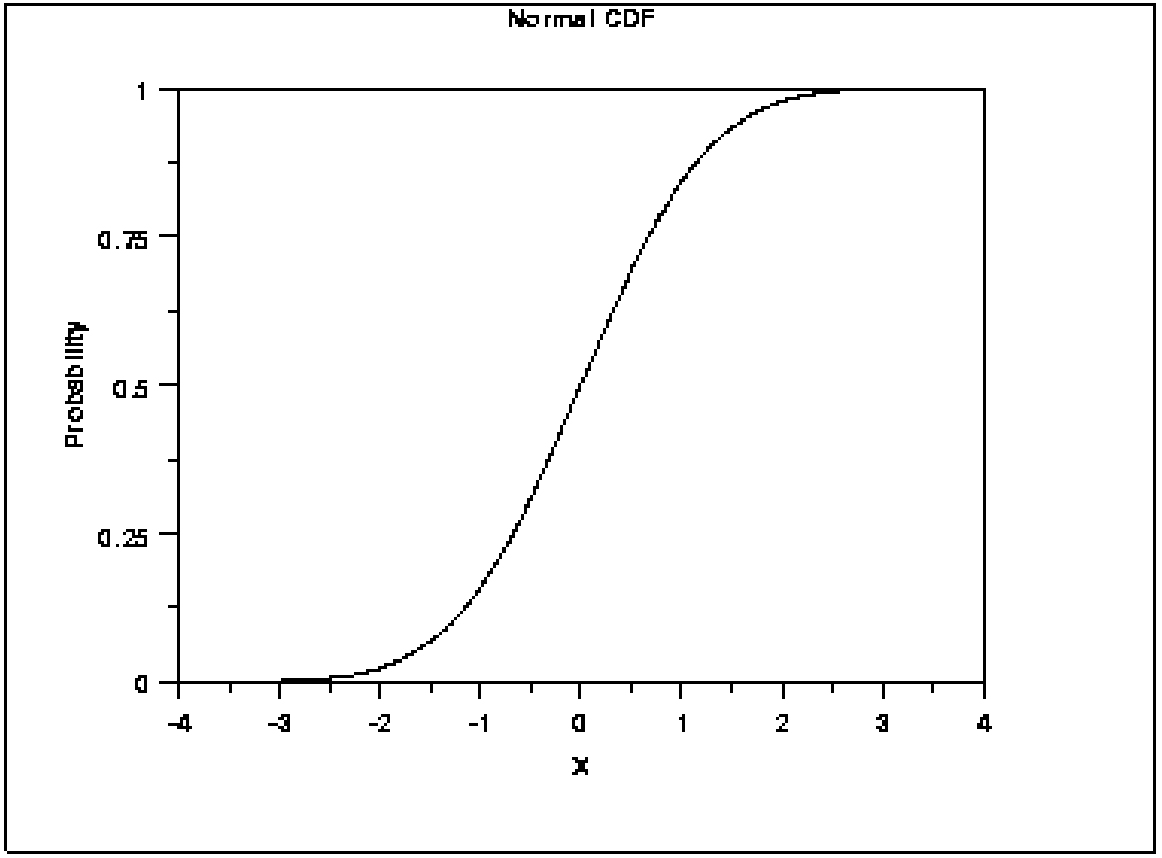
- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



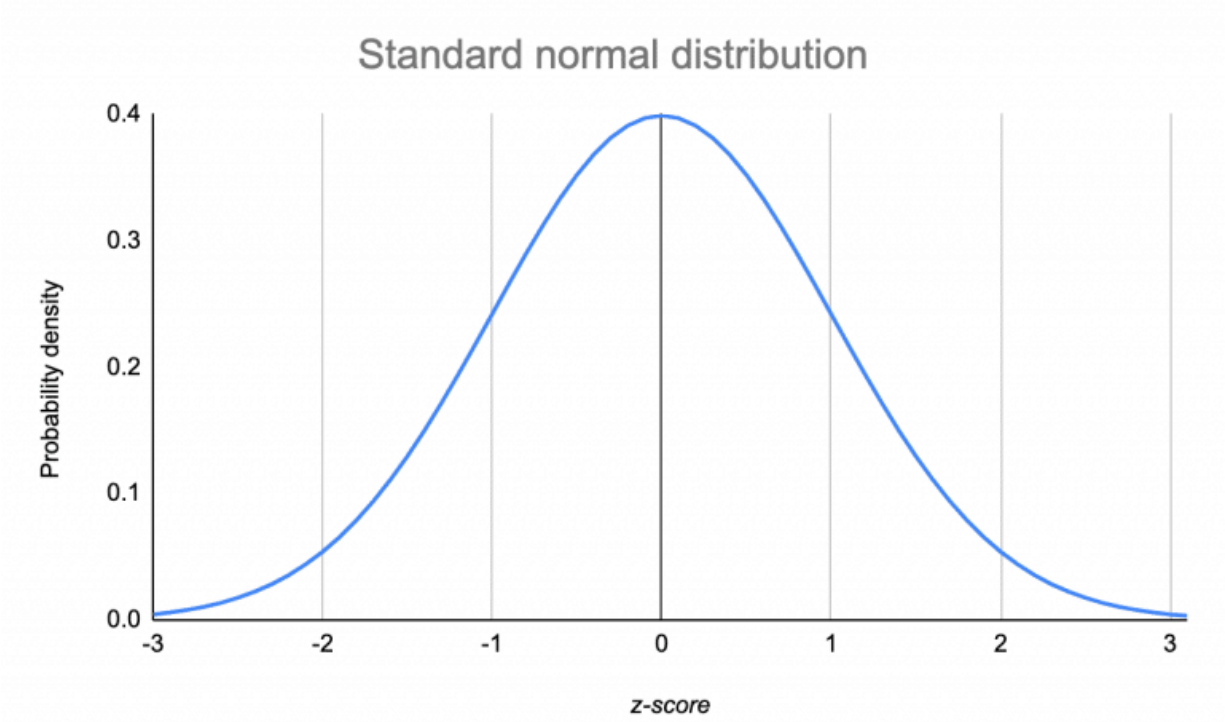
CDF of Normal Distribution

- The cumulative distribution function (cdf) is the probability that the variable X takes a value less than or equal to x.
- (Here in the figure below, Mean=0, SD=1)

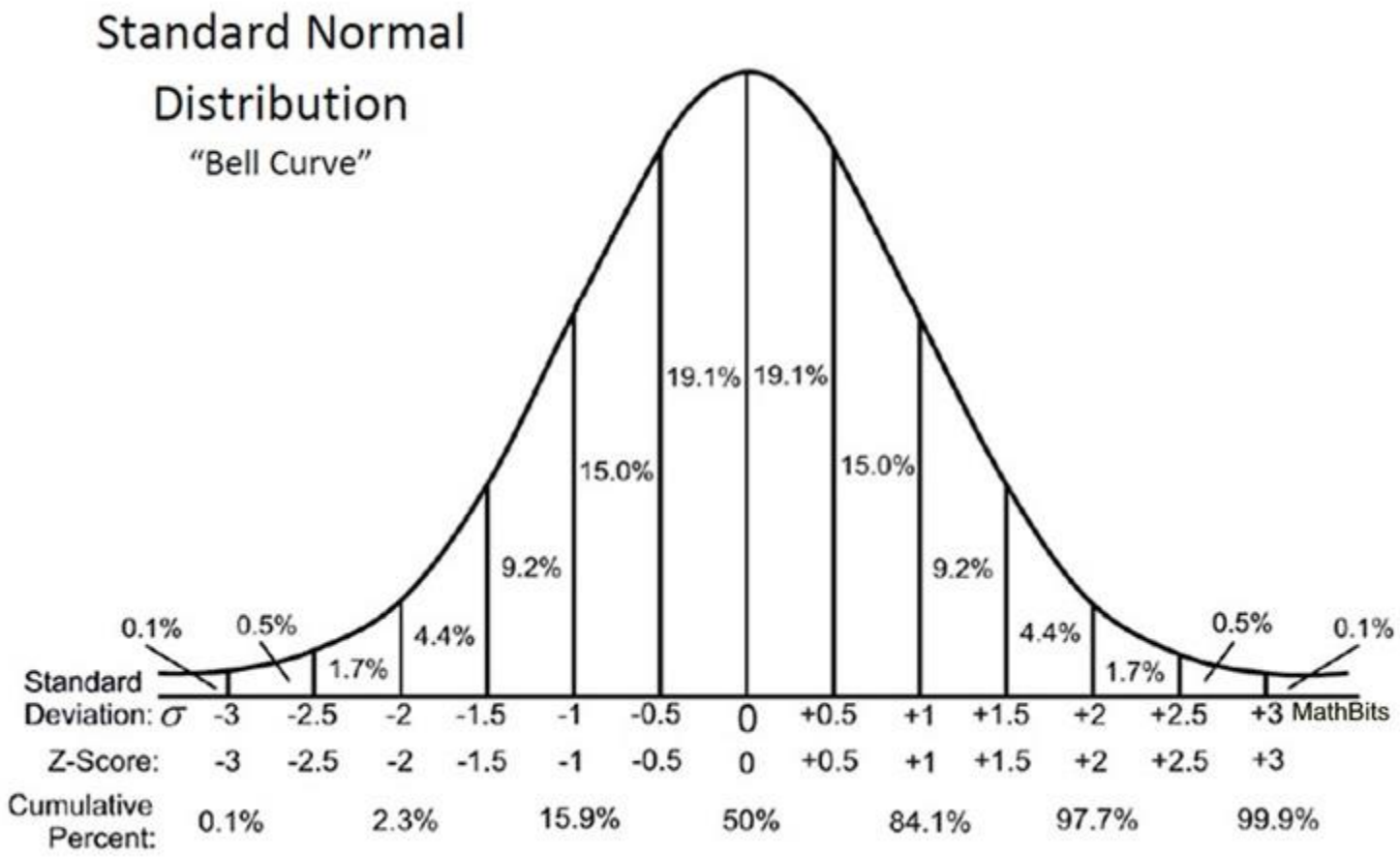


Z-Distribution

- The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.
- Z-scores tell you how many standard deviations away from the mean each value lies.



Z-Distribution



Z-Score

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

x = raw score

μ = mean

σ = standard deviation (std)

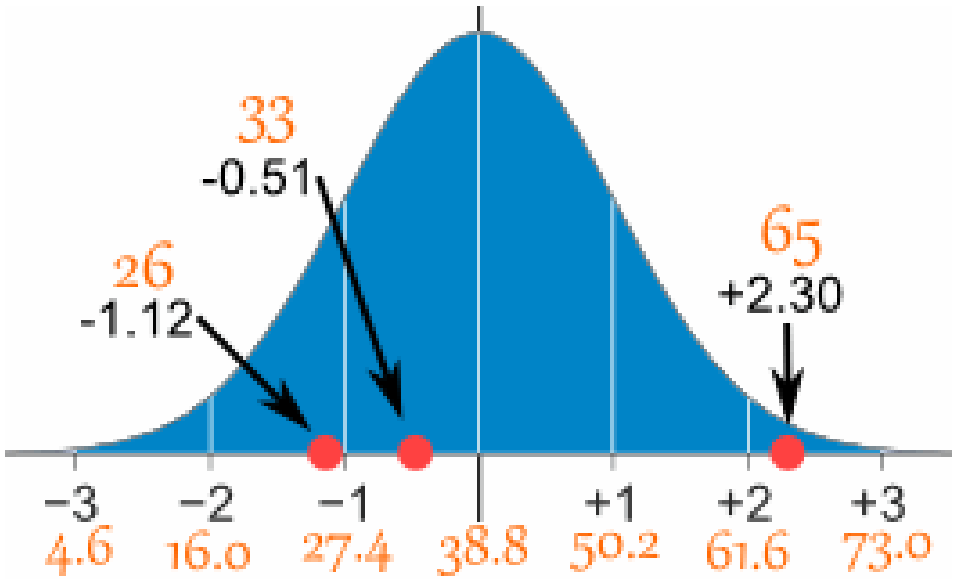
As the formula shows, the **z-score** is simply the raw score minus the population mean, divided by the population standard deviation.

Z-Distribution

Day	Time
1	26
2	33
3	65
4	28
5	34
6	55
7	25
8	44
9	50
10	36
11	26
12	37
13	43
14	62
15	35
16	38
17	45
18	32
19	28
20	34

Mean is 38.8 minutes

Standard Deviation is 11.4 minutes



Summary

- Introduction to Data Science