HCAI5DS02 – Data Analytics and Visualization.
Lecture – 04
Introduction to Statistical Modeling
Turning Uncertainty into Insight. Quantifying Uncertainty.

Siman Giri

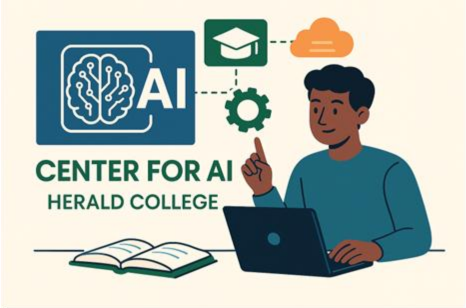


image generated via copilot.

2.3.1 PDF: Discrete vs. Continuous.

For Discrete Random Variables ✓

- We use the **Probability Mass Function (PMF)**
- It gives the probability of each **countable outcome:**
  - $P(X = x_i)$
  - $\sum_i P(X = x_i) = 1$

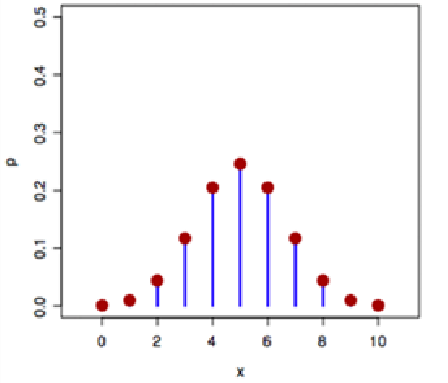


Fig: Discrete Probability Distribution

For Continuous Random Variables

- We use the **Probability Density Function (PDF).**
- Since continuous values are uncountable,
  - the probability of any exact value is
    - $P(X = x) = ?$
- Instead, we calculate the probability over an interval:
  - $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

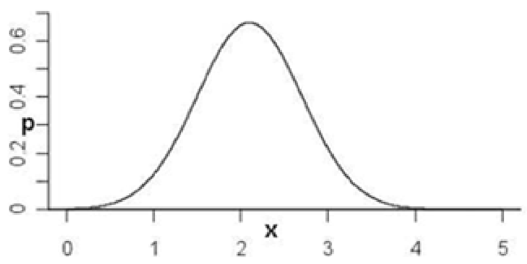


Fig: Continuous Probability Distribution

image from internet : may subjected to copyright.

1. Understanding The Random Variable.

{Extending to Continuous Case.}

1.1 Assign a Probability for Continuous Events

- **Scenario: Customer Arrival Time at E – commerce website.**
    - Your analytics team is studying customer behavior for an online store.
    - Based on server logs, you know that customers typically log in randomly between **1 p.m. and 2 p.m.** on weekdays.
    - You assume there's **no specific peak or bias** during that hour
        - **logins are uniformly distributed** in time between **1 p.m. and 2 p.m**.
    - Let T be the login time of a randomly selected user.
- **Q1:** What is the **Sample Space S**?
- **Q2:** What is the probability $\mathbf{P(T = 1{:}30\ P.m.)}$? Why?

1.1 Assign a Probability for Continuous Events

- **Scenario: Customer Arrival Time at E – commerce website.**
  - Your analytics team is studying customer behavior for an online store.
  - Based on server logs, you know that customers typically log in randomly between **1 p.m. and 2 p.m.** on weekdays.
  - You assume there's **no specific peak or bias** during that hour
    - **logins are uniformly distributed** in time between **1 p.m. and 2 p.m**.
  - Let T be the login time of a randomly selected user.
- **Q1:** What is the **Sample Space S**?
  - $\mathbf{S = [1, 2)\ (from\ 1{:}00\ p.m.\ before\ 2{:}00\ p.m.)}$
- **Q2:** What is the probability $\mathbf{P(T = \ 1{:}30\ P.m.)}$?
  - $\mathbf{P(T = 1{:}30\ p.m.) = 0.}$
- **Why?**
  - Because in **continuous distributions**:
    - The probability at a single point is always zero.
    - Probability is only defined over an interval.
      - You's ask: $\mathbf{P(1.45 \leq T \leq 1.55)\ (10\ \ - minute\ window)}$ instead $\mathbf{P(T = \ 1{:}30\ P.m.)}$.

1.2 The Continuous Probability Paradox.

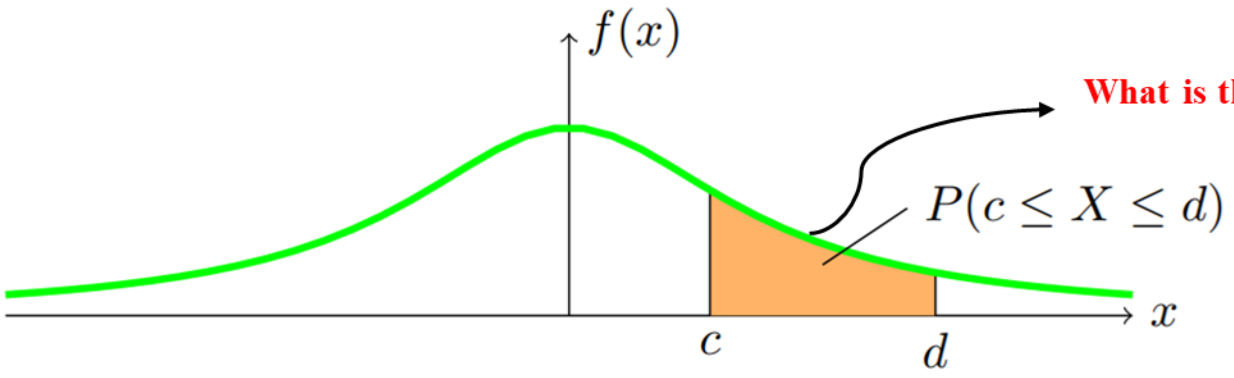
- **Problem:**
  - How can a **variable** take **infinitely many values in an interval**, yet the **total probability** is still **1**?
  - Let $\mathbf{T \sim Uniform}\ [\mathbf{1}, \mathbf{2})$ The contradiction is:
    - If **each point** has **zero probability** i.e. $\mathbf{P(T = t) = 0\ for\ every\ t \in [1, 2)}$ then:
      - **"How does their sum equal 1?"**
    - If **each point** had **non – zero probability** i.e. $\mathbf{P\left(T \in [1, 2)\right) = 1}$ (**By Axiom of Total Probability**) then:
      - **Summing over uncountably many points** $\rightarrow \infty$ **(violates axioms).**
  - In continuous space, we do not assign probabilities to points.
  - We assign a density over intervals and compute area under the **curve**.
- **Where does this curve, or function, come from?**
  - In general, **this function** is **designed to describe how the random variable (continuous) behaves**.
  - It captures **the structure of the experiment** and **how outcomes are likely to occur** across **the sample space**.
    - Such a function is called a **Probability Density Function (PDF)**, and it is defined when we introduce a **Continuous Random Variable**.

1.3 Continuous Sample Space and Probability Function.

- **Continuous Sample Space:**
  - A continuous sample space is a sample space **containing outcomes** defined in the terms of **interval and some interval can have infinite number of points**.
    - Consider {Experiment} observing the distance a ball can be thrown, **say d**. The sample space is now continuous and defined as:
      - $\boldsymbol{\Omega} = \left\{\mathbf{d} \,\middle|\, \mathbf{d} \in \mathbb{R}_{+} \,\&\, \mathbf{d} > \mathbf{0}\right\}$
  - Continuous sample spaces do not **have distinct outcomes** to which **probabilities** can be **assigned.**
  - Instead, probabilities are assigned to ***intervals*** of the **sample space**, and these probabilities are described using a **real-valued** ***probability function***, such as: $\mathbf{f_X(x)}$**.**
- **Why we Need a Function?**
  - To handle infinite possibilities in continuous space, we need:
    - A function to assign **"how dense"** the probability is at each point.
    - A way to compute probability over intervals, not at fixed points.
    - A model to simulate and analyze randomness at scale.

1.4 Continuous Random Variable

- A **continuous random variable** takes a range of values, which may be finite or infinite in extent.
  - Here are a few examples of ranges: $[\mathbf{0}, \mathbf{1}], [\mathbf{0}, \infty), (-\infty, \infty), [\boldsymbol{a}, \boldsymbol{b}]$.
- Formal Definition:
  - A ***random variable*** **X** is ***continuous*** if there is a function $\mathbf{f_X(x)}: \mathbb{R} \to [\mathbf{0}, \infty)$ such that for **any** $\mathbf{a \leq b}$ we have:
    - $\mathbf{P(a \leq X \leq b) = \int_a^b f_X(x)dx}$
    - The ***function*** $\boldsymbol{f_X}$ is called the **probability density function (pdf) of X.**
    - If you graph the **probability density function** of a **continuous random variable X** then:



What is the total area under the pdf $\mathbf{f_X(x)}$?

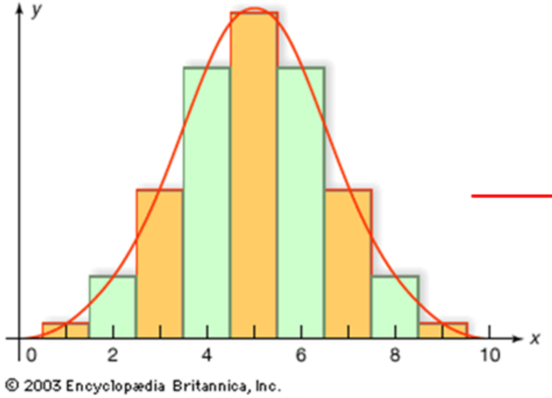
$\mathbf{P(c \leq X \leq d)}$ **= area under the graph between c and d.**

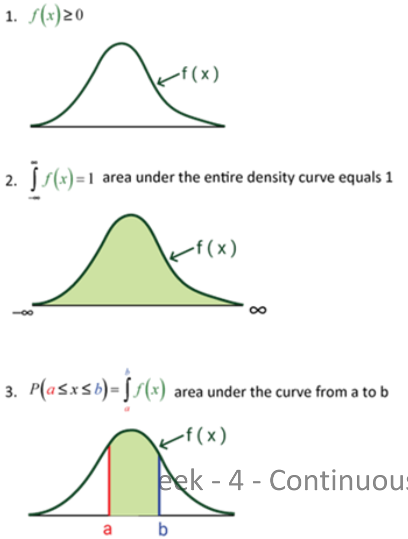
1.4.1 Real world Use Cases of Continuous Variables.

Random Variable	Use Case (Analytics)
Time Spent on website	Estimate likelihood a user stays more than 30 seconds.
Amount spent on purchase	Predict average revenue per customer.
Customer age	Target ads based on age distribution.
Product delivery time	Model reliability of logistic Operations.
Temperature of Cold chain	Ensure food safety in delivery systems.

1.5 What is Probability Density Function (PDF)?

- A **PDF** is a **function $f_X(x)$** that satisfies all the axioms of probability written as:
  - For an **interval** $[a, b]$, the probability is: $P(a \le X \le b) = \int_a^b f_X(x)dx$
  - **Axioms of Probability:**
    - **Non-negativity**: integration over any region of the sample space must never produce a negative value: $f_X(x) \ge 0 \; \forall x \in \mathbb{R}$
    - **Exhaustive:** Over the whole sample space, the probability function must integrate to one i.e. the total area under the curve is 1: $\int_{-\infty}^{\infty} f_X(x)dx = 1$
    - **Additive:** The probability of the union of any non-overlapping regions is the sum of the individual regions.

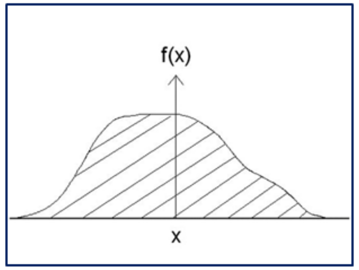




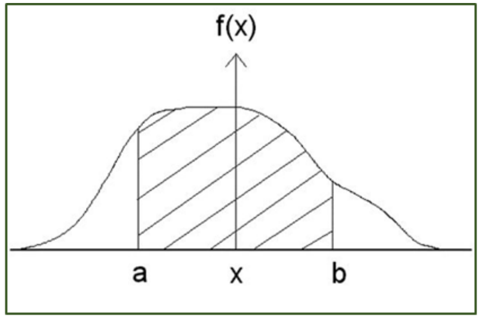
Feature	Discrete PMF	Continuous PDF
Values	Countable	Uncountably infinite
Probability at point	$P(X = x) > 0$	$P(X = x) = 0$
Total probability	$\sum P(X = x) = 1$	$\int f(x)dx = 1$
Plot type	Bar Chart	Smooth Curve
Computation	Exact values	Area under curve over interval.

Table: PMF vs. PDF

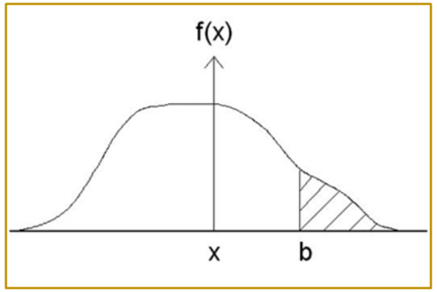
1.5.1 Graphical view of PDF.



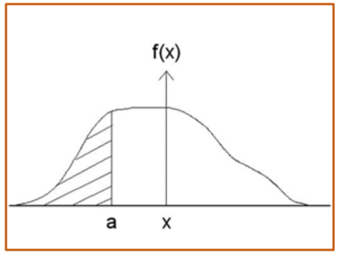
Total area $= \int_{-\infty}^{\infty} f_X(x)\,dx = 1$



Area $P[X \in [a,b]] = \int_a^b f_X(x)\,dx$



Area $P_X(x \geq b) = \int_b^{\infty} f_X(x)\,dx$



Area $P_X(x \leq a) = \int_{-\infty}^{a} f_X(x)\,dx$

1.6 Expectation and Variance of CRV.

- **Expectation of a Continuous Random Variable:**
  - For a *continuous random variable X* with **pdf $f_X(x)$** we define the **expectation** $\mathbb{E}[X]$ as:
    - $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$
  - **Condition for Existence:**
    - The expectation exists (i.e. finite) if and only if:
      - $\int_{-\infty}^{\infty} |x| \cdot f_X(x)\, dx < \infty$
    - If the above **integra diverges** (i.e. equals $+\infty$), then the **expectation is not defined**.
  - **Interpretation:**
    - The expectation is a **weighted average of all possible values of X**, where the weights are given by **the density $f_X(X)$**.
    - It gives us the **center of mass** of the distribution curve.

1.6 Expectation and Variance of CRV.

- **Variance of a Continuous Random Variable:**
  - For a *continuous random variable X* with **pdf $f_X(x)$** and expectation $\mathbb{E}[X]$, **the variance** is defined as:
    - $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f_X(x)\, dx$
      - where $\mu = \mathbb{E}[X]$ is the **expected value** mean of X.
    - If the **expectation** $\mathbb{E}\left[X^2\right]$ **exists**, variance can also be computed as:
      - $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$
  - **Interpretation:**
    - Variance **measures** the **spread or dispersion** of a **distribution around its mean**.
    - A **larger variance** implies the **values of X are more spread out from the mean**.

2. Cumulative Distribution Function.

2.1 Motivation for Cumulative Distribution Function.

- **Scenario:**
  - **You manage a promotional campaign** where you send emails in **batches of 3 users** (micro-segmented delivery). After sending 500 batches, you collect how many users clicked in each batch:

Clicks in Batch	Frequency
0	90
1	180
2	120
3	110

  - Your goal: **Analyze performance and make decisions based on user engagement.**
  - Question you ask: **"What is the probability that a batch gets at most 1 click?"**
  - Can we build a **PMF and answer above question**.

2.1.1 Probability Mass Function.

- **Let X be the number of users who click in a batch,** Thus **the Empirical PMF** can be tabulated as :

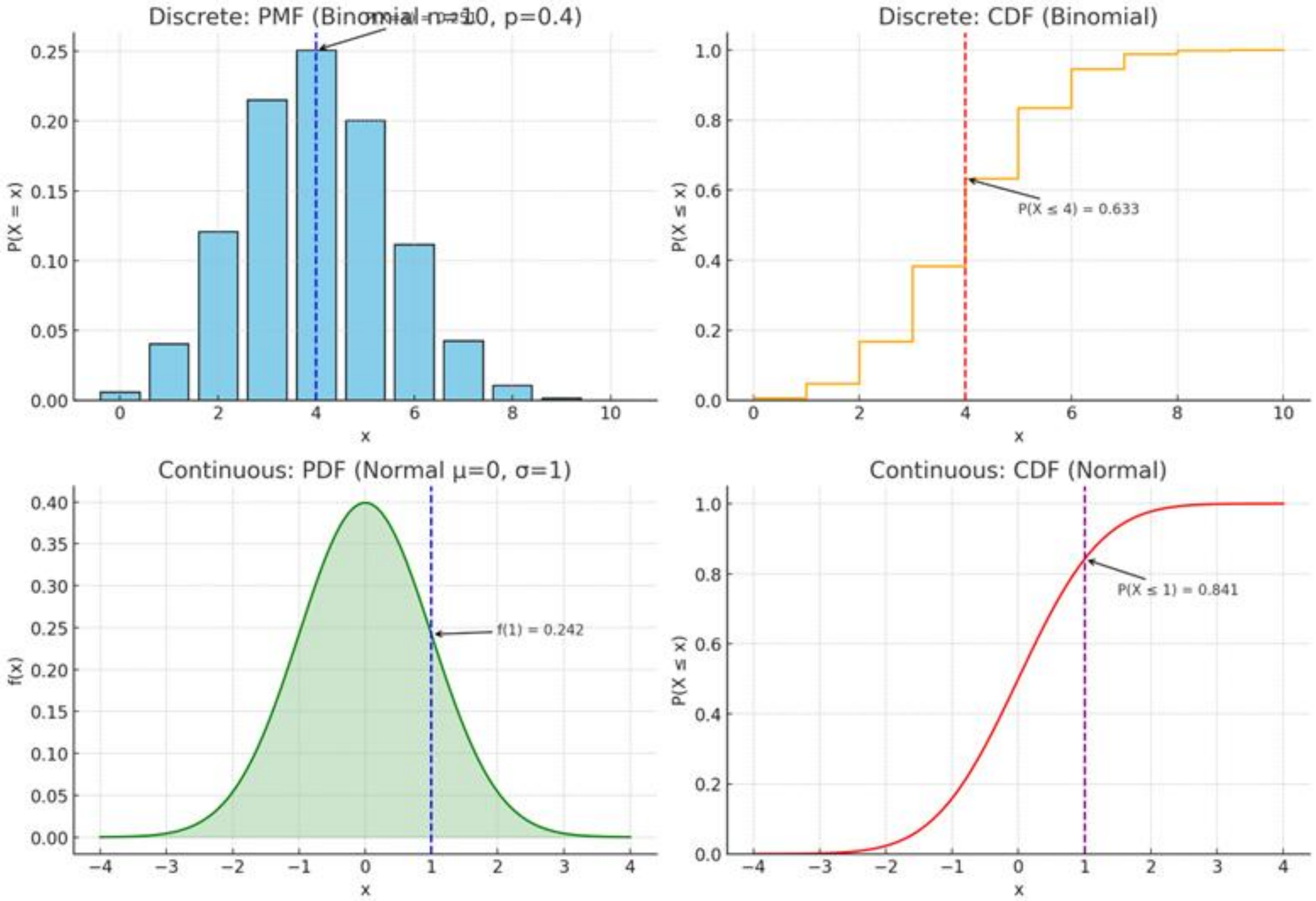
Clicks in Batch	Frequency	$\hat{P}(X = x)$
0	90	$\frac{90}{500} = 0.18$
1	180	$\frac{180}{500} = 0.36$
2	120	$\frac{120}{500} = 0.24$
3	110	$\frac{110}{500} = 0.22$

- **Can we answer this question?**
  - **"What is the probability that a batch gets at most 1 click ?"**
    - $P(X \leq 1) = ?$
- **PMF has limitations:**
  - PMF only gives:
    - $P(X = 0) = 0.18$
    - $P(X = 1) = 0.36$
  - But it does not directly tell us $P(X \leq 1) = ?$
    - Enter the **CDF – Cumulative Distribution Function.**

2.2 Definition: Cumulative Distribution Function.

- For a *random variable X,* the cumulative distribution function (cdf), **denoted by $F_X(x)$**, is defined as:
  - $F_X(x) = P(X \leq x)$
    - which means the probability that the **random variable X** takes on a value less than **or equal to x**.
  - For **discrete random variables**, the CDF is the sum of probabilities of all outcomes less than or equal to x:
    - $F_X(x) = P(X \leq x) \sum_{t \leq x} P(X = t)$
  - For **continuous random variables**, the CDF is the integral of the probability density function (PDF) up to x:
    - $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$
  - the area under the CDF between $-\infty$ **and x**.

2.2.1 Distribution and Cumulative Function.



2.1.4 But What if we Ask …

- "What's **the probability that a batch gets at most 1 click** ?"
  - $\mathbf{P(X \leq 1) = ?}$
  - PMF gives values only at individual points.
  - To answer questions involving ranges or cumulative behavior,
    - PMF becomes tedious and error prone, **especially for large outcome spaces**.
- **CDF to the Rescue:**
  - The **cumulative Distribution Function (CDF)** gives us:
    - $\mathbf{F_X(x) = P(X \leq x) = P(X = 0) + P(X = 1) = 0.54.}$

Table: Possible CDF for our Scenario.

Clicks in Batch	Frequency	$\hat{P}(X = x)$	$F_X(x) = P(X \leq x)$
0	90	$\frac{90}{500} = 0.18$	0.18
1	180	$\frac{180}{500} = 0.36$	$0.18 + 0.36 = 0.54$
2	120	$\frac{120}{500} = 0.24$	$0.54 + 0.24 = 0.78$
3	110	$\frac{110}{500} = 0.22$	$0.78 + 0.22 = 1.00$

2.1.4 Interpretations with CDF.

- Your goal: **Analyze performance and make decisions based on user engagement.**
- **Interpretation 1: Performance Benchmarking:**
  - About 54% of all email batches get at most 1 click.
    - This tells us more than half our email batches are performing below or around average.
    - If our marketing goal is to get **at least 2 clicks per batch**, this insight tells us:
      - Only 46% of batches are hitting that target.
  - **Actionable Insight:** Consider optimizing subject lines, timing, or target segments.

Table: Possible CDF for our Scenario.

Clicks in Batch	Frequency	$\hat{P}(X = x)$	$F_X(x) = P(X \leq x)$
0	90	$\frac{90}{500} = 0.18$	0.18
1	180	$\frac{180}{500} = 0.36$	$0.18 + 0.36 = 0.54$
2	120	$\frac{120}{500} = 0.24$	$0.54 + 0.24 = 0.78$
3	110	$\frac{110}{500} = 0.22$	$0.78 + 0.22 = 1.00$

2.1.4 Interpretations with CDF.

- Your goal: **Analyze performance and make decisions based on user engagement.**
- **Interpretation 2: Audience Segmentation:**
  - Only **22%** of batches have **all 3 users** clicking.
    - These are your **high-performing segments**.
    - Could indicate well-targeted customer profiles or compelling offers.
  - **Actionable Insight:** Analyze what's unique about these batches — location, time, demographics?

Table: Possible CDF for our Scenario.

Clicks in Batch	Frequency	$\widehat{P}(X = x)$	$F_X(x) = P(X \leq x)$
0	90	$\frac{90}{500} = 0.18$	0.18
1	180	$\frac{180}{500} = 0.36$	$0.18 + 0.36 = 0.54$
2	120	$\frac{120}{500} = 0.24$	$0.54 + 0.24 = 0.78$
3	110	$\frac{110}{500} = 0.22$	$0.78 + 0.22 = 1.00$

2.1.4 Interpretations with CDF.

Table: Possible CDF for our Scenario.

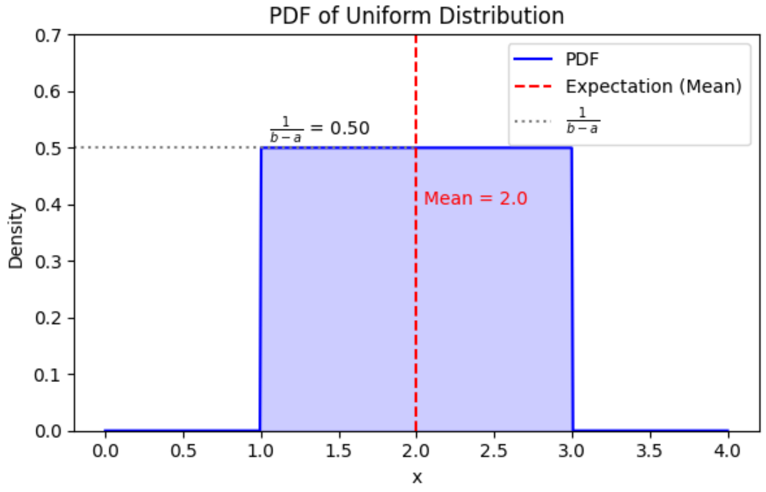
Clicks in Batch	Frequency	$\hat{P}(X = x)$	$F_X(x) = P(X \leq x)$
0	90	$\frac{90}{500} = 0.18$	0.18
1	180	$\frac{180}{500} = 0.36$	$0.18 + 0.36 = 0.54$
2	120	$\frac{120}{500} = 0.24$	$0.54 + 0.24 = 0.78$
3	110	$\frac{110}{500} = 0.22$	$0.78 + 0.22 = 1.00$

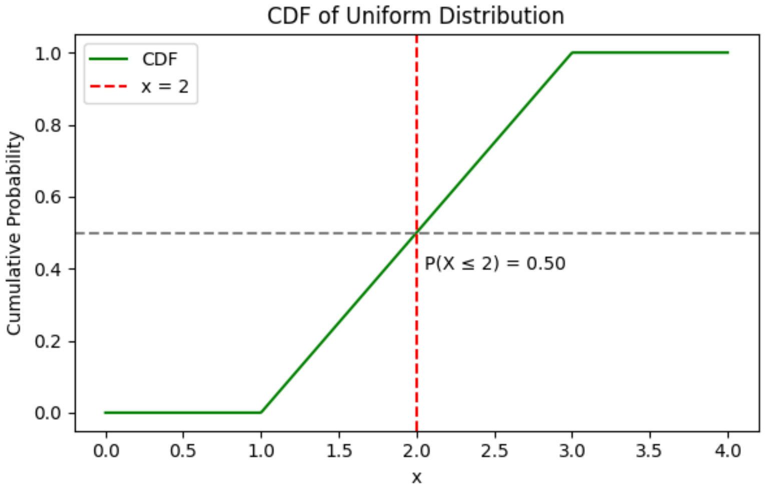
- **Interpretation 3: Risk or Failure Zones:**
  - 18% of batches get **zero clicks**.
  - This signals possible issues:
    - Poorly written content
    - Wrong audience
    - Broken links
  - **Actionable Insight** : Flag these as failure cases and investigate causes.
- **Interpretation 4: Decision-Making Thresholds:**
  - "What is the chance a batch performs **below expectations**?"
  - Let's say your team decides that any batch with **fewer than 2 clicks is a concern**.
    - You ask: $\mathbf{P(X < 2) = F(1) = 0.54}$.
  - Meaning: Over half the batches would **fail to meet your benchmark**.
  - **Actionable Insight** : Rethink email design or customer targeting.

3. Some Common Continuous Probability Distributions.

3.1 Uniform Distribution.

- **Definition:**
  - A continuous random variable $\mathbf{X} \sim \mathbf{U}(\mathbf{a}, \mathbf{b})$ is **uniformly distributed** over $[\mathbf{a}, \mathbf{b}]$ if the probability is equally spread over the interval.
  - The PDF of a Uniform distribution is:
    - $\mathbf{f_X(x)} = \begin{cases} \frac{\mathbf{1}}{\mathbf{b-a}}, \mathbf{a \leq x \leq b} \\ \mathbf{0, otherwise} \end{cases}$
  - The CDF is:
    - $\mathbf{F_X(x; a, b)} = \begin{cases} \mathbf{0 \text{ for } x < a;} \\ \frac{\mathbf{x-a}}{\mathbf{b-a}} \mathbf{\text{ for } a \leq x \leq b;} \\ \mathbf{1 \text{ for } x > b.} \end{cases}$
- **Key properties:**
  - Support: $\mathbf{x \in [a, b]}$
  - Expectation: $\mathbb{E}[\mathbf{X}] = \frac{\mathbf{a+b}}{\mathbf{2}}$
  - Variance: $\mathbf{Var[X]} = \frac{\mathbf{(b-a)^2}}{\mathbf{12}}$



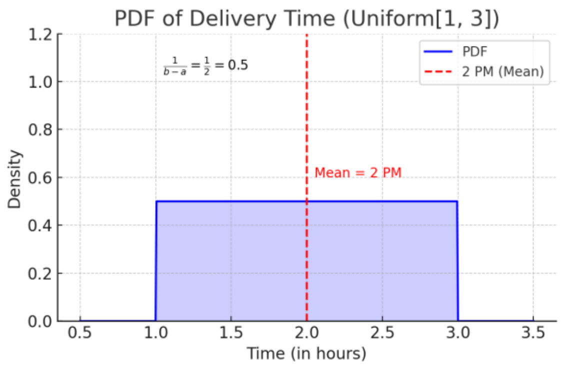


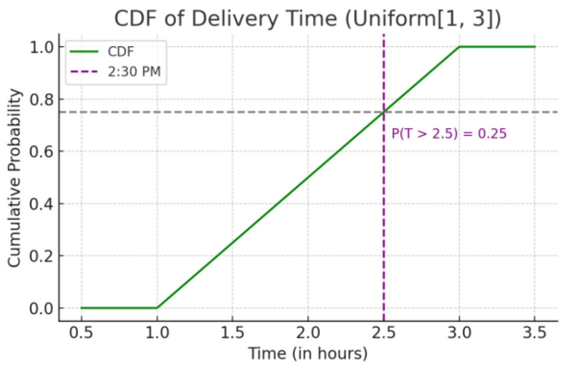
3.1.1 Use Case: Delivery Arrival Time Prediction.

- **Scenario:**
  - A delivery platform (e.g. food or parcel delivery app) promises that an order will arrive sometime between 1PM and 3 PM.
  - **However:**
    - The system does not collect GPS delivery tracking data.
    - The delivery time does not depend on the customer's region.
    - There is no prior historical data on arrival patterns.
    - The only information available is:
      - The delivery will definitely arrive sometime between 1PM and 3PM, all times in that interval are equally likely.
- What is the probability the delivery comes before 2PM?
- What is expected delivery time?
- What is the probability the delivery takes longer than 2:30 PM?

3.1.2 Use Case: Delivery Arrival Time Prediction.

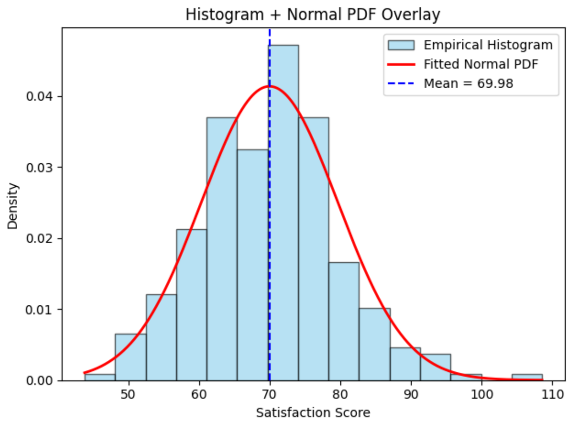
- **Modeling with a Uniform Distribution:**
  - Let $\mathbf{T \sim Uniform(1PM, 3\ PM)}$ represent the **delivery time in hours**.
- **Why Uniform?**
  - You have a **fixed interval** $[1, 3]$and **no reason to favor any sub-interval** over another.
  - This reflects **maximum uncertainty** under a **bounded interval**.
  - The **probability** is **equally spread** over **the interval**.
- **Analytics Tasks:**
  - What is the probability the delivery comes before 2PM?
    - $\mathbf{P(T \leq 2PM) = \frac{x-a}{b-a} = \frac{2-1}{3-1} = 0.5}$
  - What is expected delivery time?
    - $\mathbf{\mathbb{E}[T] = \frac{1+3}{2} = 2PM}$
  - What is the probability the delivery takes longer than 2:30 PM?
    - $\mathbf{P(T > 2.5) = \frac{3-2.5}{3-1} = 0.25}$





3.2 The Normal Distribution.

- **The Most Famous Curve in Statistics:**
  - The normal distribution (**also called the Gaussian distribution**) , named after mathematician Carl Friedrich Gauss is:
    - **The most well – known and widely used continuous probability distribution.**
    - **Often referred to as the bell – shaped distribution because of its symmetrical, bell – like curve.**
  - **Why is it so Important?**
    - Natural Phenomena: Many real – world quantities (e.g. height, test scores, measurement errors) are approximately normally distributed.
    - Mathematical Elegance: Defined completely by two parameters:
      - **mean(μ) and standard deviation (σ).**
    - **Foundational Role:** Central to statistical inference, especially due to the **Central Limit Theorem**.



3.2.1 Definition: Normal Distribution.

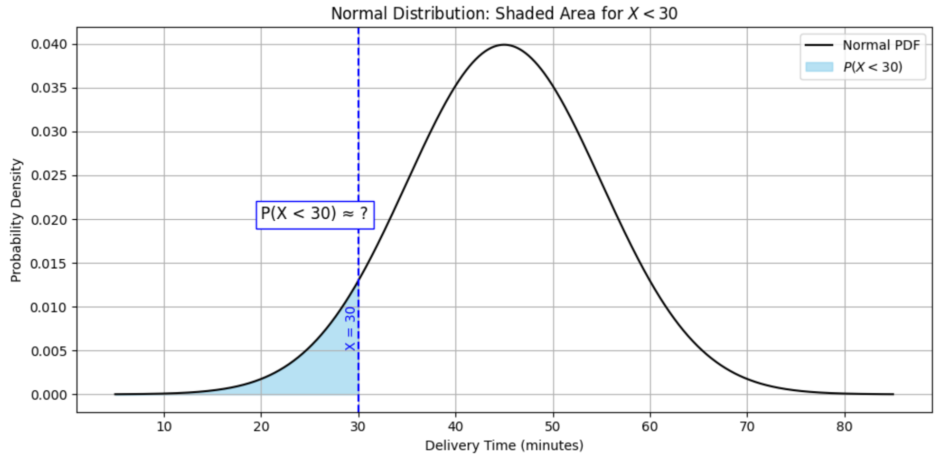
- A *continuous random variable X* is said to follow a **Normal Distribution** with
  - **mean** $\mu \in \mathbb{R}$ and **standard deviation** $\sigma > 0$, written as: $X \sim \mathcal{N}(\mu, \sigma^2)$.
  - **Probability Density Function (PDF):**
    - The **PDF of a normal distribution** is defined as:
      - $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), x \in \mathbb{R}$
    - This function describes **how probability density** is **distributed over the real number line**.
    - The **total area under the curve is 1**.
  - **Cumulative Distribution Function (CDF):**
    - The **CDF** of a **normal distribution** is:
      - $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt.$
    - This CDF does not have a **closed – form expression** in elementary functions,
      - but is computed using the error function or numerical integration.
        - $F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$, **where erf(.) is the error function**.

3.2.2 Use Case: Delivery Time Analysis.

- **Scenario:**
  - A logistics company tracks how long it takes to deliver packages with in a city. Based on thousands of past deliveries, they find:
    - **Most deliveries take around 45 minutes**,
    - **Shorter or longer times are less common, but possible**
    - **The distribution of delivery times resembles a bell curve.**
- **Analytics Task:**
  - What's the probability a delivery is completed in less than 30 minutes?
  - What proportion of deliveries are between 35 and 55 minutes?
  - If a delivery takes more than 60 minutes, should we flag it as late?

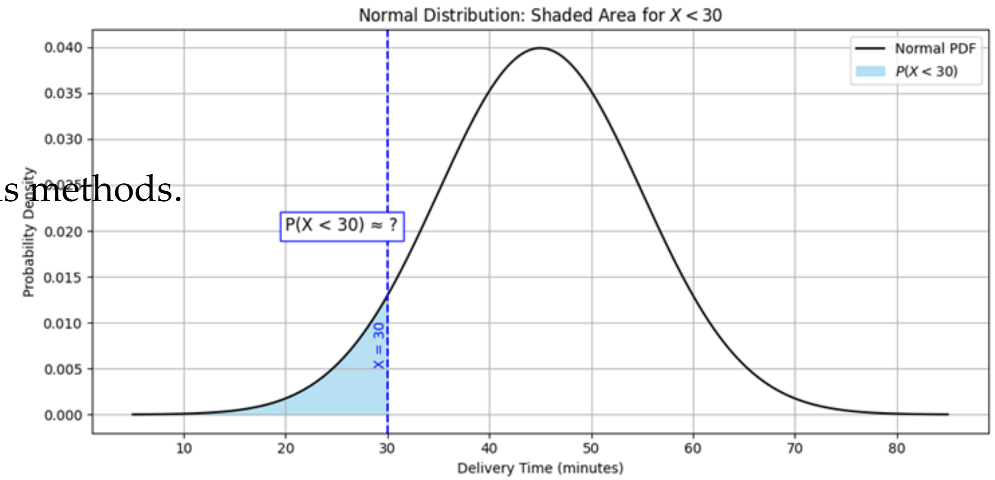
3.2.3 Use Case: Delivery Time Analysis.

- **Modeling Assumption:**
  - Let **X be the delivery time in minutes**, we model:
    - $\mathbf{X \sim \mathcal{N}(\mu = 45, \sigma = 10)}$.
  - **Why Normal Distribution?**
    - Delivery time is **continuous**, **unbounded**, and **symmetrically clustered** around a typical value (mean).
    - Empirical data from sensors or timestamps shows a pattern resembling the bell curve.
  - **Analytics Task:**
    - What's the probability a delivery is completed in less than 30 minutes?
      - $\mathbf{P(X < 30) = F_X(30) = ?}$
    - What proportion of deliveries are between 35 and 55 minutes?
      - $\mathbf{P(35 < X < 55) = F_X(55) = F_X(35) = ?}$
    - If a delivery takes **more than 60 minutes**, should **we flag** it as late?
  - **How do we compute the exact probability?**



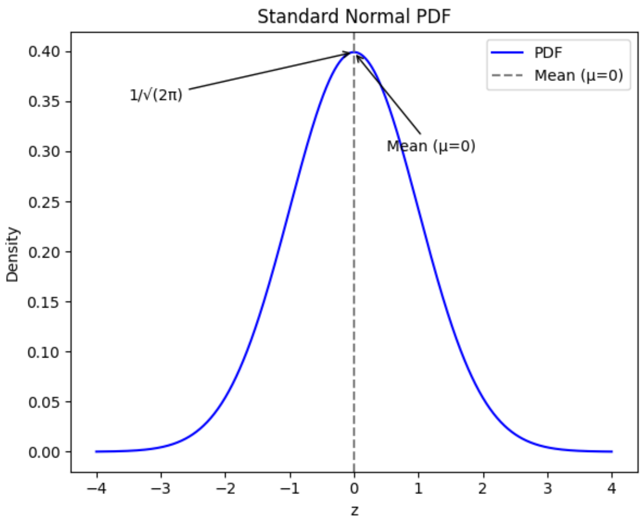
3.2.4 Use Case: Delivery Time Analysis.

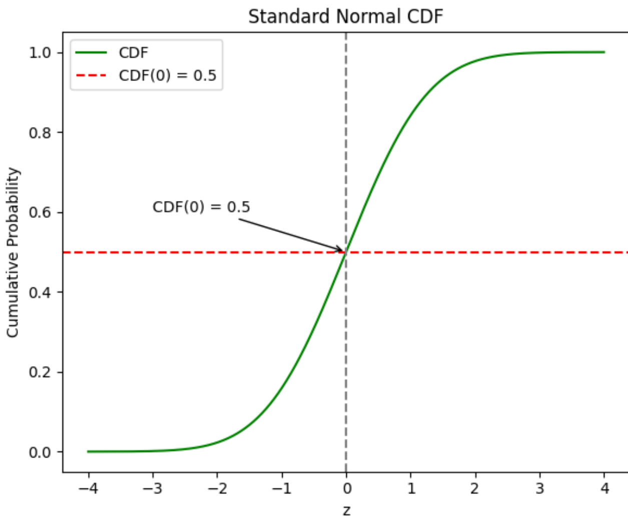
- **Challenges – Area under the Curve:**
  - In continuous distributions, probabilities are computed as **areas under the curve**.
    - For the normal distribution, this means integrating the probability density function:
      - $\mathbf{P(X < x) = \int_{-\infty}^{x} f_X(t)\,dt}$
  - But the Normal PDF:
    - $\mathbf{f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$
      - does not have **closed form integral**
        - i.e. can not be computed exactly using standard calculus methods.
  - The **Normal PDF** is **analytically friendly**,
    - but its integral (i.e., **computing exact probabilities**)
      - requires **numerical methods**, **lookup tables**, or **software**
  - hence the need for **standardization and the CDF**.



3.3 Standard Normal Distribution.

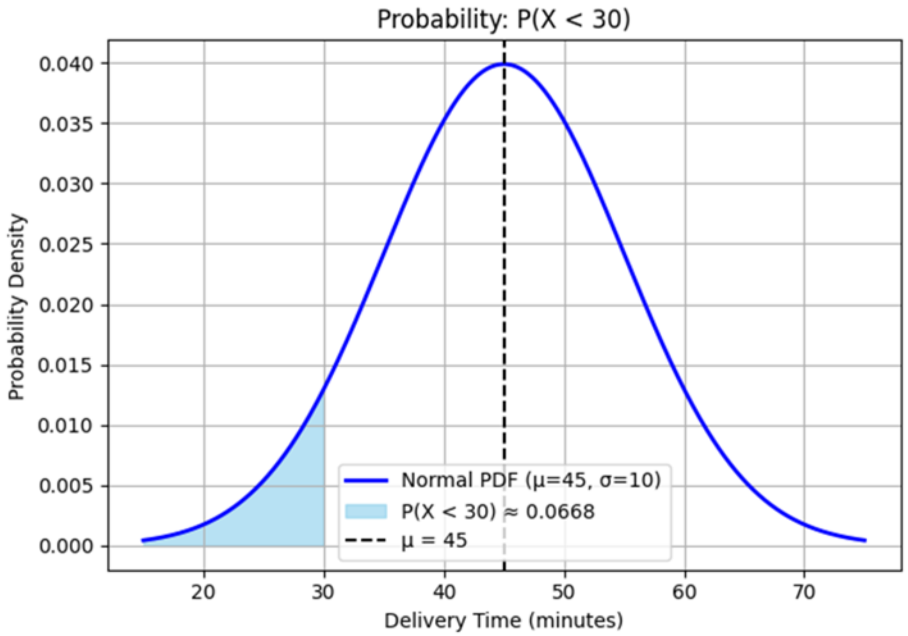
- A **standard normal distribution** is a special case of **the normal distribution** with:
  - **Mean: $\mu = 0$**
  - **Standard Devaition: $\sigma = 1$**
    - We write: $\mathbf{Z \sim \mathcal{N}(0,1)}$
- The probability density function (PDF) of the standard normal distribution is given by:
  - $\mathbf{f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}$ **, for z** $\in \mathbb{R}$
- **Key Properties:**
  - Symmetric around zero.
  - Total area under the curve is 1.
  - Used for standardizing any normal random variable:
    - $\mathbf{Z = \frac{X-\mu}{\sigma}}$
  - The CDF is denoted as $\Phi(z)$, and is defined as:
    - $\mathbf{\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}$





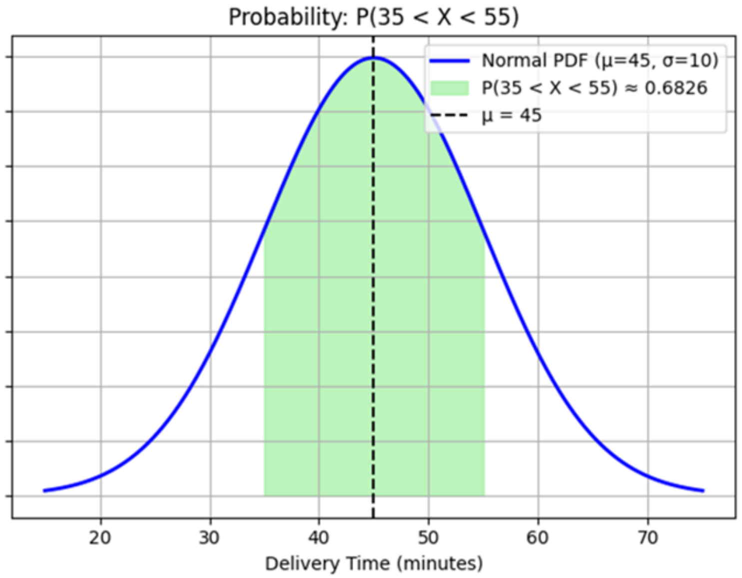
3.3.1 Answering our Question.

- We assume the **delivery time** $\mathbf{X} \sim \mathcal{N}(\mu = 45, \sigma = 10)$:
  - What's the probability a delivery is completed in less than 30 minutes?
    - **You compute:** $\rightarrow \mathbf{P(X < 30) = F_X(30)}$
    - First Standardize:
      - $\mathbf{Z = \frac{X-\mu}{\sigma} = \frac{30 - 45}{10} = -1.5}$
    - Then look up the **CDF of standard normal** $\mathbf{\Phi(-1.5)}$:
      - $\mathbf{P(X < 30) = \Phi(-1.5) \approx 0.0668}$
    - So about **6.68%** of deliveries finish in less than 30 minutes.



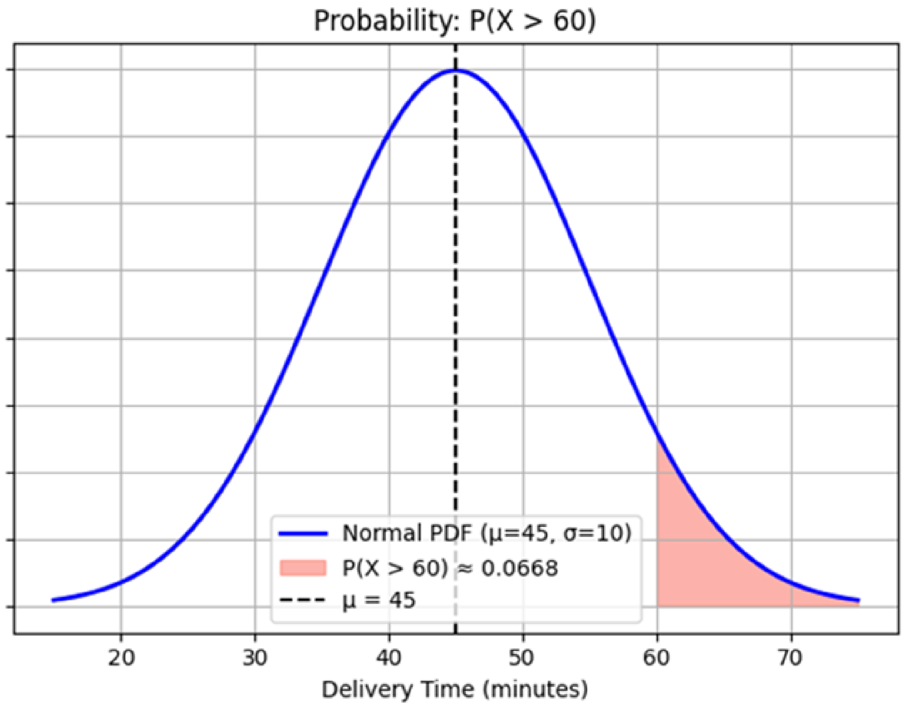
3.3.2 Answering our Question.

- We assume the **delivery time** $\mathbf{X} \sim \mathcal{N}(\mu = 45, \sigma = 10)$:
  - What proportion of deliveries are between 35 and 55 minutes?
    - We compute: $\rightarrow \mathbf{P(35 < X < 55) = F_X(55) - F_X(35)}$
    - First standardize both:
      - $\mathbf{Z_1 = \frac{35 - 45}{10} = -1, \quad Z_2 = \frac{55 - 45}{10} = 1}$
    - Then,
      - $\mathbf{P(35 < X < 55) = \Phi(1) - \Phi(-1) \approx 0.8413 - 0.1587 = 0.6826}$
  - So approximately **68.26%** of **deliveries** fall between **35 and 55 minutes**.



3.3.3 Answering our Question.

- We assume the **delivery time** $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu} = \mathbf{45}, \boldsymbol{\sigma} = \mathbf{10})$:
  - If a delivery takes more than 60 minutes, should we flag it as late?
    - We compute: $\rightarrow \mathbf{P}(\mathbf{X} > \mathbf{60}) = \mathbf{1} - \mathbf{F_X}(\mathbf{60})$
    - First Standardize:
      - $\boldsymbol{Z} = \dfrac{\mathbf{60} - \mathbf{45}}{\mathbf{10}} = \mathbf{1.5}$
      - Then:
        - $\mathbf{P}(\mathbf{X} > \mathbf{60}) = \mathbf{1} - \boldsymbol{\Phi}(\mathbf{1.5}) \approx \mathbf{1} - \mathbf{0.9332} = \mathbf{0.0668}$
- So only **6.68%** of deliveries take **more than 60 minutes**
  - **reasonable to flag as "late."**



3.3.4 Finding the Φ(z):

Using the Z table:

- The table typically gives: $\Phi(z) = P(Z \le z)$
  - That is, the area under the standard normal curve to left of z.
  - This is the cumulative probability up to z.
- **Cautions:**

Type of Probability	What it Means	How to Compute Using Table
$P(Z \le z)$	**Left of z**	**Directly use table: $\Phi(z)$**
$P(Z > z)$	**Right of z**	**$1 - \Phi(z)$**
$P(a < Z < b)$	**Between a and b**	**$\Phi(b) - \Phi(a)$**
$P(Z < -z)$	**Left tail**	**Use symmetry: $\Phi(-z) = 1 - \Phi(z)$**

Standard Normal Probabilities

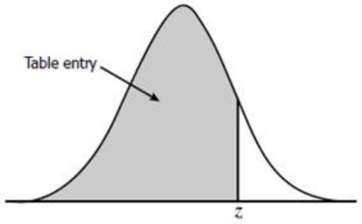


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

3.3.4 Finding the Φ(z):

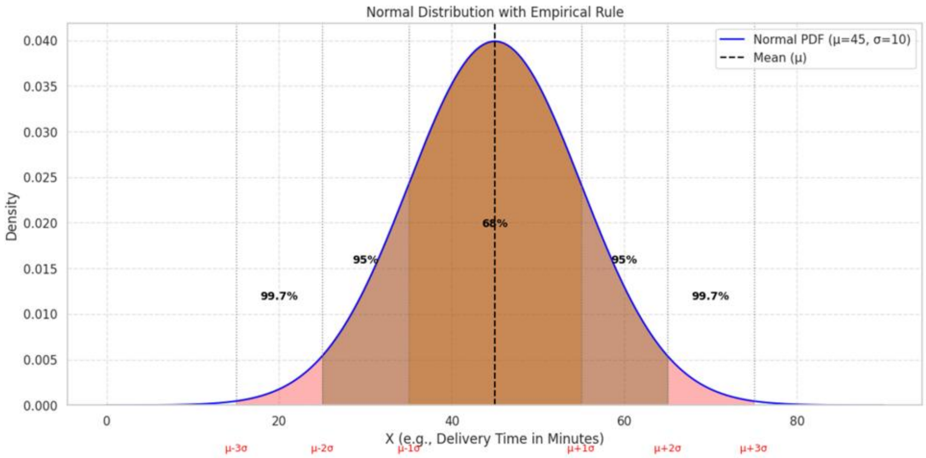
- **Using python matplotlib and scipy.stats :**

```
from scipy.stats import norm
mu, sigma = 45, 10
# 1. P(X < 30)
p1 = norm.cdf(30, loc=mu, scale=sigma)
# 2. P(35 < X < 55)
p2 = norm.cdf(55, mu, sigma) - norm.cdf(35, mu, sigma)
# 3. P(X > 60)
p3 = 1 - norm.cdf(60, mu, sigma)

print(f"P(X < 30): {p1:.4f}")
print(f"P(35 < X < 55): {p2:.4f}")
print(f"P(X > 60): {p3:.4f}")
```

3.4 The 68-95-99.7 Rule.

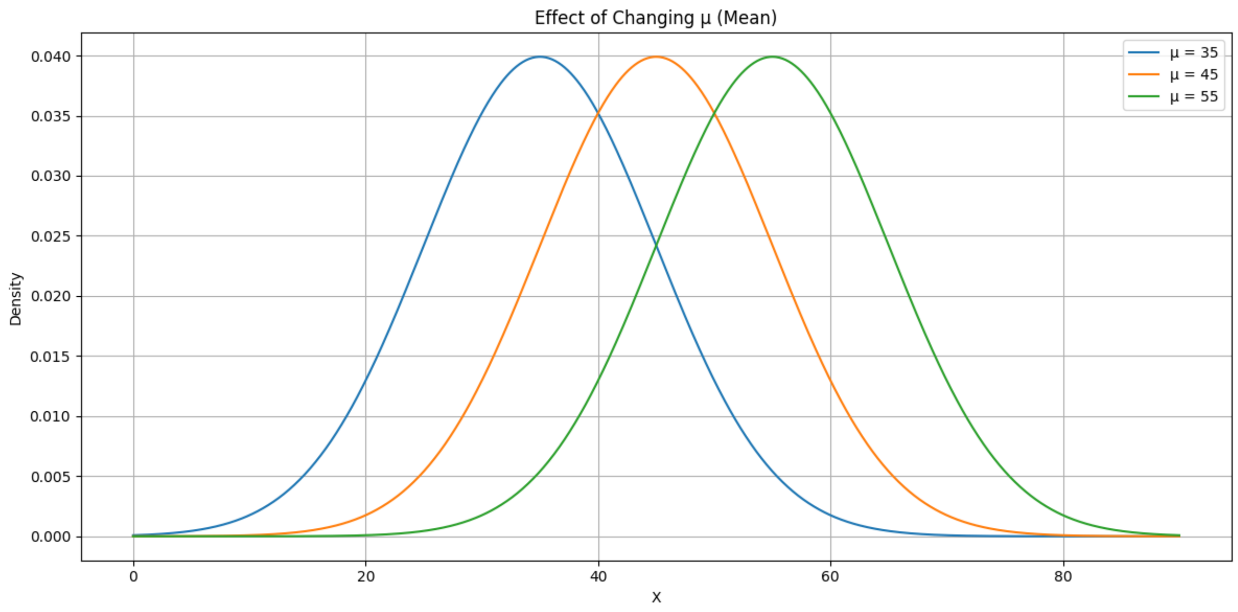
- **aka Empirical Rule:**
  - This rule is a visual and intuitive guide to the spread of data in a Normal Distribution.
  - It states that:
    - For a normal distribution with mean μ and standard deviation σ:
      - 68% of the data falls within 1 standard deviation:
        - $\mathbf{P(\mu - \sigma < X < \mu + \sigma) \approx 0.68}$
      - 95% of the data falls within 2 standard deviations:
        - $\mathbf{P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95}$
      - 99.7% of the data falls within 3 standard deviations:
        - $\mathbf{P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997}$



- **What it means?**
  - If delivery time is normally distributed with a mean of 45 minutes and SD of 10:
    - 68% of deliveries will be between 35 and 55 minutes.
    - 95% between 25 and 65 minutes.
    - 99.7% between 15 and 75 minutes.

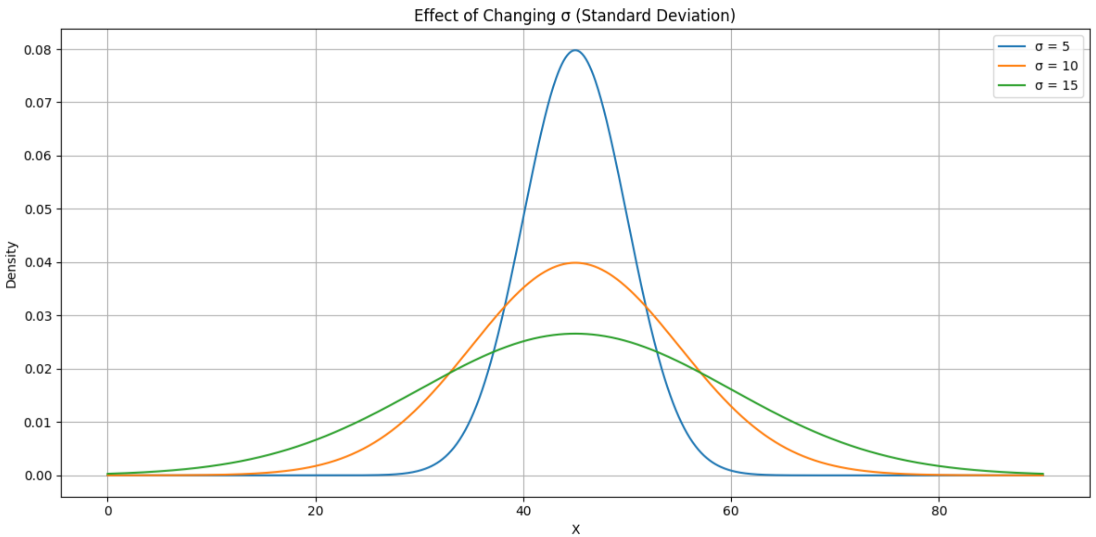
3.5 What Expectation and Variance Control in a Distribution.

- **Expectation (mean, μ):**
  - Controls the **center or location** of **the distribution**.
  - Shift the curve left or right.
  - Does not affect the shape or spread.
- **What it means for business analytics?**
  - Mean delivery time.
  - Average number of clicks per campaign.
  - Mean customer rating.



3.5 What Expectation and Variance Control in a Distribution.

- **Variance (σ^2) and Standard Deviation (σ):**
  - Controls the **spread or dispersion** of the distribution.
  - Affects how **concentrated or spread out** the values are around the mean.
  - Larger $\sigma \rightarrow$ flatter and wider curve.
  - Smaller $\sigma \rightarrow$ taller and narrower curve.
- **In business analytics:**
  - Low variance = consistent performance (e.g., delivery times always ~45 mins)
  - High variance = unpredictable or unstable (e.g., wait time ranges wildly)



Thank You