# Foundations of Data Science

# Probability & Statistics

PG-Level ACP AI&MLOPS Cohort 2

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

# Probability: The Mathematics of Uncertainty

- 80% chance of rain today
- Expected time of arrival is 6 minutes
- Average score of a batsman is 35.3
- Sensor noise is 0.3 units
- Ruling party will win $300 \pm 30$ seats

# Example

- King K is an upcoming batter rising through the U-19 league
- King K has the following scores in 10 matches
  - 24, 43, 124, 22, 156, 98, 76, 51, 102, 89
- King K has the following strike rate in 10 matches
  - 93.2, 52.1, 201.5, 110.2, 90, 124.1, 99.1, 157.2, 165, 178
- Categorical and Numerical Data

| Score | S/R | C or NC | Cat. Var. |
|---|---|---|---|
| 24 | 93.2 | | |
| 43 | 52.1 | | |
| 124 | 201.5 | | |
| 22 | 110.2 | | |
| 156 | 90 | | |
| 98 | 124.1 | | |
| 76 | 99.1 | | |
| 51 | 157.2 | | |
| 102 | 165 | | |
| 89 | 178 | | |

# Probability: Intuitive Frequentist

- When we are not sure of a particular outcome, i.e., we are uncertain, we need a mathematical way to quantify our uncertainty

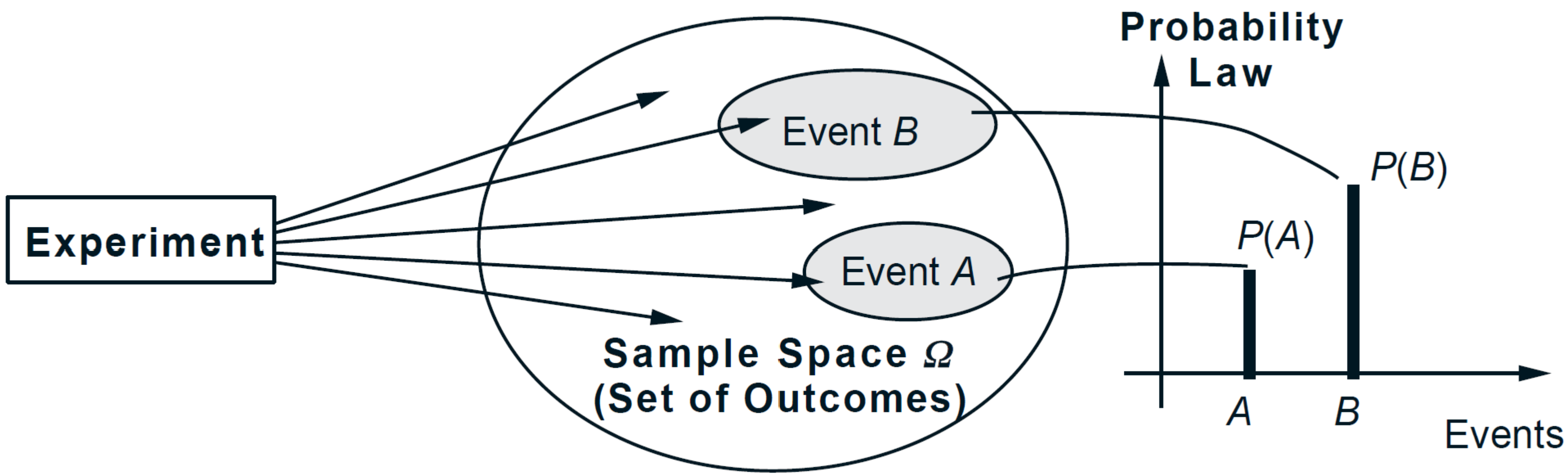
- What is the chance for King K to score a century?

- $P(A) = \dfrac{number\ of\ samples\ with\ scores \geq 100}{total\ number\ of\ samples}$

- We call the above number as the probability of King K to score a century.

# Probability

- **Experiment** - An underlying process of interest
  - A cricket match where King K batted
- It will produce exactly one out of several possible **elementary outcomes**.
  - End of the game King K will have scored runs
- Set of all possible elementary outcomes is called a **Sample Space**
- **Event** of interest – A combination of elementary outcomes
- Probability – A number that quantifies the chance of some event happening



# Throw of a fair dice

- What is the probability of landing a six on throwing a six-sided fair dice?

# Probability: A Fundamental Property

- The uncertain outcome of every experiment has a fundamental property associated with it
- This fundamental property is the **Probability**
- Sample Space – Formal Definition
  - Set of mutually exclusive and collectively exhaustive elementary outcomes
- Probability is defined for events in the sample space and is governed by 3 axioms
  - Non-negativity
  - Normalization
  - Additivity

# Poll 1

- King K has the following S/R in 10 matches

  93.2, 52.1, 201.5, 110.2, 90, 124.1, 99.1, 157.2, 165, 178

1. What is the probability of scoring a S/R >100?
   a. 0.4
   b. 0.6
   c. 0.3
   d. Can't be estimated
2. What is a sample space?
   a. Collection of mutually exclusive elementary outcomes
   b. Collection collectively exhaustive elementary outcomes
   c. Both of the first two options together
   d. Either of first two options

# Random Variables

- Variable defining an uncertain quantity of interest – Random Variable
- Random variable X – Denoted by capital letter
- Random variable assigns a number to an event – Mathematically it is a real valued function from the sample space to the number line
- X = Strike Rate
    - Straightforward
    - The mapping is the numerical value itself
- X = Century
    - How to convert to a number?
    - 1 if yes, 0 if no – Label Encoding
    - Label, One Hot, Weight of Evidence Encoding, Binary etc

# Price of a Pen – A non-cricket Example

- Let us think of Random Variables as variables denoting “items of interest” whose values are not certain
- Let us say the price of a pen is Rs 20. This is a fixed price. If we use X as the variable for the price of this pen, then it is a deterministic variable.
- Now, let us say we don’t know exactly what the price of a pen at the shop is. It can take different values. If we use X as the variable for the price of this pen, then it is a random variable.
- If the price of a pen can be any number from the set 18, 20 or 22.3
  - X is a discrete random variable with three elements in the sample space.
- If the price of a pen can be any number between 18 and 22 including decimals
  - X is a continuous random variable.

# Poll 2

1. X = 24 is a
   a. Discrete RV
   b. Continuous RV
   c. Deterministic Variable
   d. None of the above
2. X = {12,35.5,78.1} is a
   a. Discrete RV
   b. Continuous RV
   c. Deterministic Variable
   d. None of the above
3. The length of left side obtained by breaking a stick is a
   a. Discrete RV
   b. Continuous RV
   c. Deterministic Variable
   d. None of the above

# Frequency Counter: Histogram

- Let us count King K's century scores
- Let us form bins of King K's strike rate and count their frequency
  - 0-50, 51-100, 101-150, 151-200, 201+
  - Data: 93.2, 52.1, 201.5, 110.2, 90, 124.1, 99.1, 157.2, 165, 178

# Probability Mass Function

- $p_X(x) = P(\{X = x\})$
- Probability of the Random Variable $X$, if $X$ were to take the value $x$
- $p_X(x) \geq 0; \Sigma_{x \in \Omega} p_X(x) = 1$

- Frequency plot of century or not – Example of PMF

- Frequency plot of binned strike rate – Example of PMF
  - But wait, didn't we say that strike rate is a continuous variable?
  - Yes, note that we "discretized" the continuous variable by binning it

# Functions

- Mapping from a domain to a range
- It is a relationship that can be parametrized and learnt

# Increase Bins: What Happens?

- Sum of heights of each bin is 1
  - Probability axiom
- We will need uncountably infinite bins
- Each bin’s height will go to zero!
- Not useful!!!
- What do we do? – Move to Continuous Random Variable

# PDF- PMF per unit length

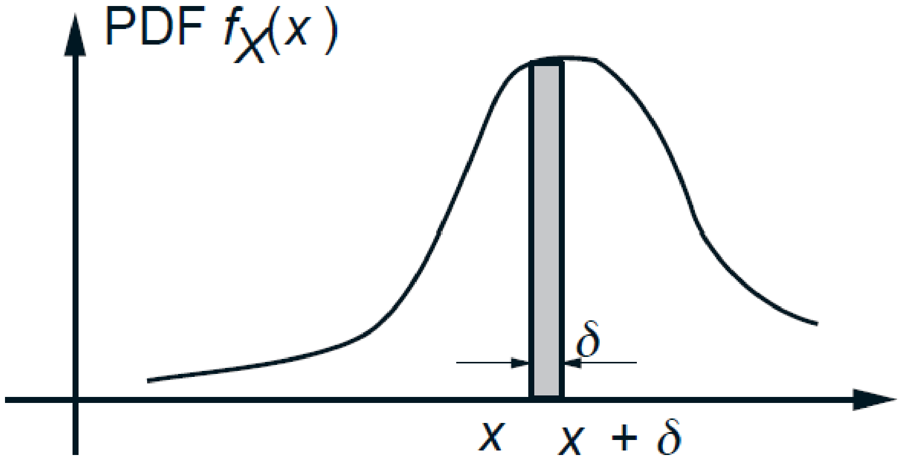



**Figure 3.2:** Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around $x$. If $\delta$ is very small, the probability that $X$ takes value in the interval $[x, x+\delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x) \cdot \delta$.

# The Mean

- The PMF (and PDF) contains the full information
- But we want one (or two numbers)
- Measures of central tendency – The Mean helps us
- The Arithmetic Mean of numerical values is a which can be used to replace all samples, but still have the same number as the sum

$$\mu = \frac{1}{m}\sum_{j=1}^{m} x_j$$

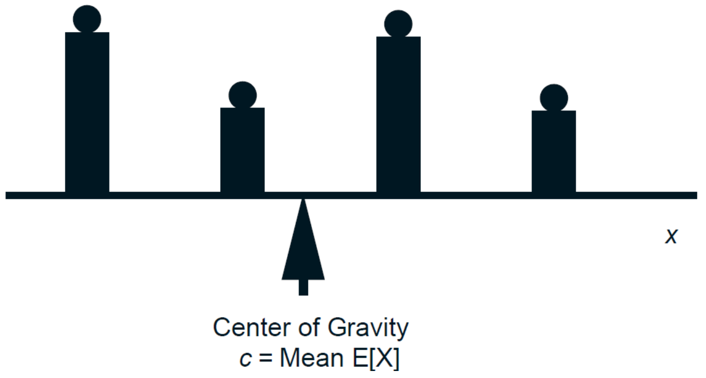
- Let us look at this sum from the frequency plot viewpoint

# Expectation

- Consider a RV X with 5 data samples (1,1,4,4,4)
- What is the mean?
- $\mu = \frac{1}{5}(1 + 1 + 4 + 4 + 4)$
- $\mu = \frac{1}{5}(2 \times 1 + 3 \times 4)$
- $\mu = \frac{2}{5} \times 1 + \frac{3}{5} \times 4$
- *What is this?*
- $\mu = p_X(X = 1) \times 1 + p_X(X = 4) \times 4$
- $\mu = \sum_{x \in \Omega} x p_X(X = x)$

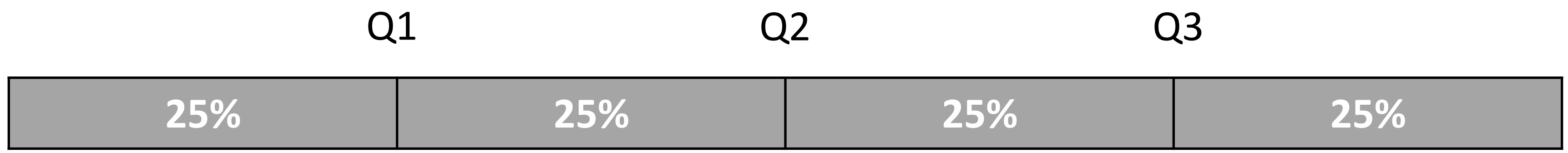
# Expectation of a RV (Mean)

- $E[X] = \sum_{x=X} x p_X(x)$
- $E[X] = \int_{x \in X} x f_X(x) dx$
- Interpretation
  - Center of gravity of the PMF/PDF
  - Average in large number of repetitions of the experiment
- This is one number that we can “expect” on an average for the variable.
- The actual realization of the RV can be different. But in large number of experiments, this is the average.



# Mean, Median, Mode, Quartiles

- Mean is the “average” of a set of numbers
  - Usually we use arithmetic mean
- Median is the middle value of a set of numbers (50%ile)
- Mode is the value that occurs most often in a set of numbers
- Quartiles (25%ile, 50%ile, 75%ile)



# Poll 3

1. Mean and Expectation are different quantities
   - True, False
2. Expectation is a random variable
   - True, False
3. Expectation is a probability
   - True, False
4. PMF is PDF per unit length
   - True, False

# Function of a Random Variable

- If $X$ is a random variable and $g(.)$ is any general nonlinear function
- $Y = g(X)$ is also a Random Variable
- The PMF of $Y$ can be evaluated from the PMF of $X$
- $E[g(X)] = \sum_x g(x) p_X(x)$

# Problem

- Find the PMF of $Y = (X - E[X])^2$ , mean and variance when
  - $p_X(x) = \frac{1}{9}\ if\ x\ is\ an\ integer\ in\ [-4,4]$

# Variance/Std. Dev

- $var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

$$Y = (X - E[X])^2$$

- The square root of variance is standard deviation
- Standard deviation has the same units as the random variable
- Standard deviation is easier to interpret

# Reasoning about one RV when another related RV is known

- What is the likelihood that a person adds "Fried Rice" to their cart in Swiggy?
- A person adds "Gobi Manchurian" to their cart in Swiggy, what is the likelihood that they add "Fried Rice" next?

- How likely is a person Covid+?
- How likely is a person Covid+ if RAT returns –ve?

- How likely are you going to be shouted at by your boss?
- Your boss is in the office shaking his head. How likely are you going to be shouted at?

# The Prediction Problem in Data Science

- What is King K's strike rate when he plays against Sri Lanka?

- King K's strike rate is uncertain and unknown

- We have historical information about X = King K's strike rate
- We also know Y = opposition team
- Now we are asked what is the distribution of X given Y = "Sri Lanka"
  - Actually Y = LabelEncoder("Sri Lanka")

# Condition one RV on another

- Let $X$ and $Y$ be two RV from the same experiment
- The knowledge that $Y = y$ happened may affect our belief about $X$
- $p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$
- $\sum_x p_{X|Y}(x|y) = 1$
- Often easy to calculate
  - $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y)$

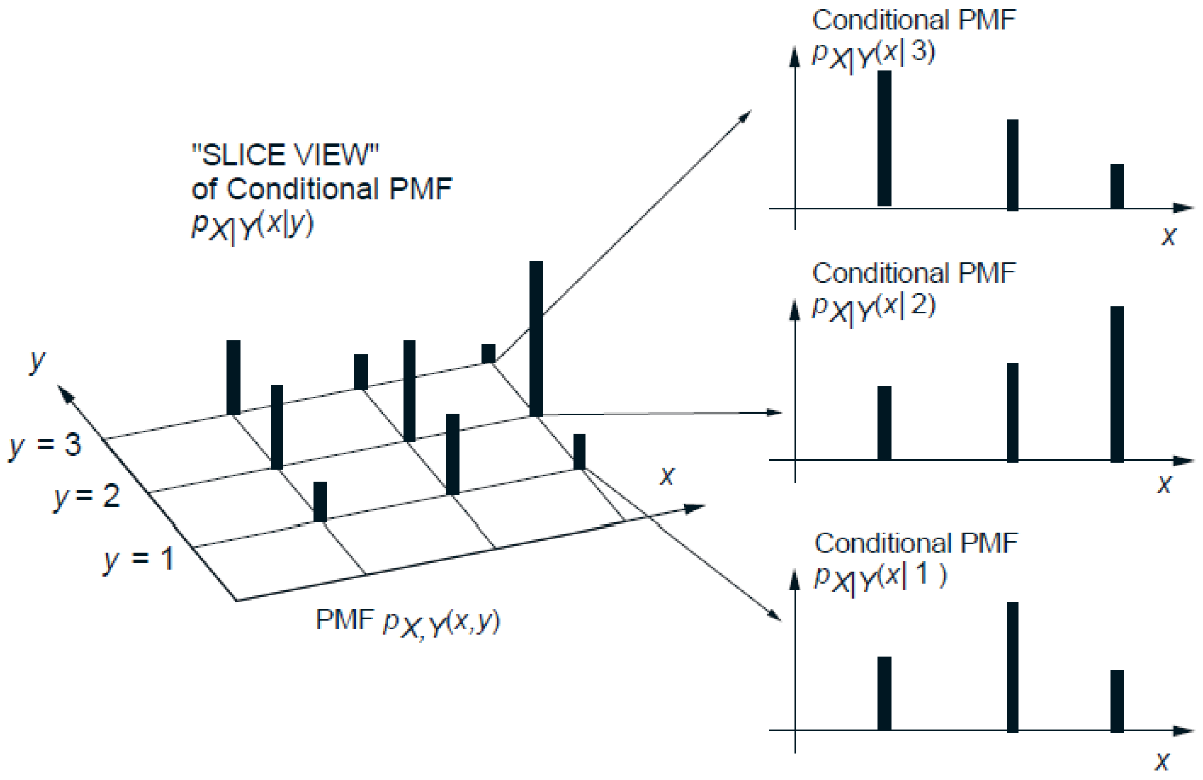
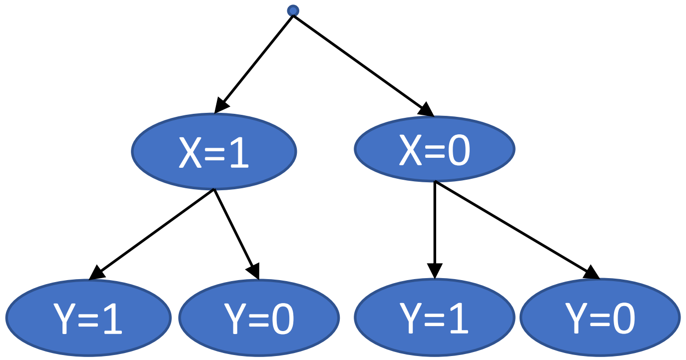



**Figure 2.13:** Visualization of the conditional PMF $p_{X|Y}(x|y)$. For each $y$, we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x|y) = 1.$$

# Joint of 2 RV

- Prob. Models may have several variables of interest
- All variables may be defined on the same sample space
- Their mutual interaction is interesting and useful
- $p_{X,Y}(x, y) = P(X = x, Y = y)$

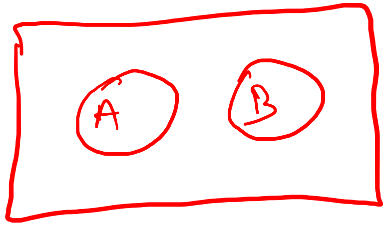


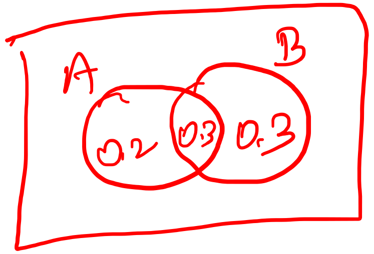
# Independence

- Consider two RVs $X$ and $Y$
- $p_{X|Y}(x|y)$ tells us the improvement in $p_X(x)$ arising out of knowledge of Y=y
- What if $Y$ does not give us any knowledge about $X$?
- $p_{X|Y}(x|y) = p_X(x)$
- By definition of conditional probability
- $p_{X,Y}(x,y) = p_X(x)p_Y(y)$
- This relation is the **DEFINITION** of independence

# Understanding Independence

- If two events are governed by distinct and noninteracting physical processes, such events are usually independent
  - Event A: Prof. Deepak wearing a yellow shirt
  - Event B: ITC Stock trading in upper circuit
- A confusing common thought
  - Two disjoint events are independent
  - Fact: Disjoint events are NEVER independent.
    - The occurrence of one says complete information about the other
  - $P(A \cap B) = 0$ for disjoint, and never equal to $P(A)P(B)$




$P(A) = 0.5$
$P(B) = 0.6$
$P(A \cap B) = 0.3$

# Covariance and Correlation

- We want to see how changes in X are related to changes in Y

- $cov(X,Y) = E[(X - E[X])(Y - E[Y])]$
  - Here the expectation is over the joint $\mathrm{p}_{X,Y}(x,y)$ $f_{X,Y}(x,y)$

- $cov(X,Y) = E[XY] - E[X]E[Y]$

- $cor(X,Y) = cov(X,Y)/std(X)std(Y)$

# Independent and Identically Distributed

- $X_1, X_2$ are called I.I.D if both of them have
  - identical distributions (e.g., both are normal with the same $\mu, \sigma$)
  - Are independent

- Arises in several situations
  - To be seen next week: Binomial is a sum of IID Bernoulli

# Entropy

- Entropy quantifies the randomness in a signal
- Take an example of weather forecast with 4 labels (sunny, sun+cloud, rain, rain+thunder)
- We can encode the above using 2 bits as $2^2 = 4$
  - 00 – Sunny; 01 – Sun+Cloud; 10 – Rainy; 11 – Rain+Thunder
- Now, let us say that with 90% probability, the forecast is sunny, then a more efficient encoding scheme is to reserve one bit for sunny, and then two more bits for the above encoding.
- 90% of the time, only 1 bit needs to be sent, and only 10% needs 3 bits.
- We send on average 0.9*1+0.1*3=1.2 bits, which is lower than 2 bits needed early
- This happened because we have an assumption about the distribution of the information.
- Defined as $H[X] = -\sum_{i=1}^{n} p_X(x_i)\log_2 p_X(x_i)\ \ x_i\ are\ a\ partition$

# Common Distributions

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

# Bernoulli Random Variable

- Each trial has only two possible outcomes (we call success or failure)
- The probability of success is the same in each trial
- Each trial is independent of the previous trials

- X = Outcome is 1 (success, +ve class, H etc)
- $p_X(x) = \begin{cases} p, if\ x = 1 \\ 1 - p, otherwise \end{cases}$
- What is mean and variance?

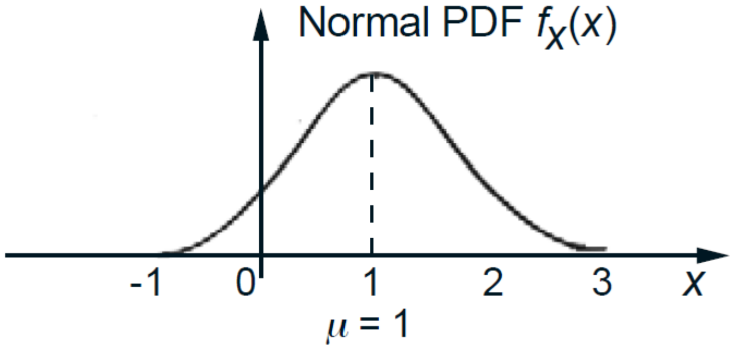
# Binomial Random Variable

- X = Number of success in $n$ trials of a Bernoulli RV
- How many times will I pass $n$ quizzes?
- Example: X = Number of Heads in 4 trials
- What is $p_X(2)$?

- $p_X(X=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, if\ k = 0,1,2,\ldots \\ 0, otherwise \end{cases}$

# Normal Random Variable

- A Continuous RV is Normal (or Gaussian) if it has the PDF of the form
- $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right], if \ -\infty < x < \infty$
- Most used PDF
- Arises in many contexts
  - Score of students in a large class
  - Height of people in a country
- Many default assumptions



# Example

- Weight distribution of college students is normally distributed with mean = 50 kg and standard deviation = 10 kg
- What is the probability of finding a student with weight between 55 to 65?
- The z-transform: $z = \frac{x-\mu}{\sigma}$

$$z_1 = \frac{55-50}{10} = 0.5$$
$$z_2 = \frac{65-50}{10} = 1.5$$
$$P(55 \leq x \leq 65) = P(z_1 \leq z \leq z_2)$$
$$= F(z_2) - F(z_1) = 0.933 - 0.691 = 0.242$$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| z | -2.5 | -2.4 | -2.3 | -2.2 | -2.1 | -2.0 | -1.9 | -1.8 | -1.7 | -1.6 |
| F(z) | 0.006 | 0.008 | 0.011 | 0.014 | 0.018 | 0.023 | 0.029 | 0.036 | 0.045 | 0.055 |
| z | -1.5 | -1.4 | -1.3 | -1.2 | -1.1 | -1.0 | -0.9 | -0.8 | -0.7 | -0.6 |
| F(z) | 0.067 | 0.081 | 0.097 | 0.115 | 0.136 | 0.159 | 0.184 | 0.212 | 0.242 | 0.274 |
| z | -0.5 | -0.4 | -0.3 | -0.2 | - 0.1 | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
| F(z) | 0.309 | 0.345 | 0.382 | 0.421 | 0.460 | 0.500 | 0.540 | 0.579 | 0.618 | 0.655 |
| z | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 |
| F(z) | 0.691 | 0.726 | 0.758 | 0.788 | 0.816 | 0.841 | 0.864 | 0.885 | 0.903 | 0.919 |
| z | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 |
| F(z) | 0.933 | 0.945 | 0.955 | 0.964 | 0.971 | 0.977 | 0.982 | 0.986 | 0.989 | 0.992 |
| z | 2.5 | | | | | | | | | |
| F(z) | 0.994 | | | | | | | | | |

# Intuition for $300 \pm 30$

- Let us say Deepika Kumari hits the 10 cm radius bull’s eye 95% of the time
- Now let us sit behind the target board
  - Bulls eye is not centered on the board
- If the arrow hit at green dot
- Then we can draw a circle of radius 10 cm around it
- This circle will contain the bull’s eye 95% of the time
- In other words:
  - draw a 10 cm circle for every shot of Deepika
  - 95% of those will contain the bull’s eye!

Green – Different shots
Orange – Bulls Eye