DATA ANALYSIS TECHNIQUES

DST-SERC School on
Nuclear Matter Under Extreme Conditions

VECC, Kolkata
January 7-25, 2013

Sudhir Raniwala,
University of Rajasthan, Jaipur

PLAN

- Introduction: Empirical Science
  - Logic: Deductive and Inductive
- Formalism: Bayesian Approach
- What are ‘good-estimates’ for a given distribution
- Parameter Determination and Hypothesis Testing
- Straight Line Fit and Outliers
- Error Determination, and Propagation
- Invariant Mass Analysis
- Correlated Variables and Errors
- Introduction to Flow / Neural Networks

- Lectures are based on parts of the following books (including figures, examples, notation!)
  - Data Analysis: a Bayesian approach
    - D S Sivia with J Skillings
  - Statistical for Nuclear and Particle Physicists
    - Louis Lyons
  - Statistical Data Analysis
    - Glen Cowan

  &

  - Wikipedia ☺

- Given a certain set of data
  - How do we verify the validity of an assumed hypothesis
    - Subject to knowing the values of parameters
  - How do we determine the value(s) of unknown parameter(s)
    - Subject to the validity of the hypothesis in question
  - Learn by examples

- Why do we want to do this?
- We believe that
  - the phenomena under study is not arbitrarily random
  - there is an underlying pattern
  - such a pattern is formed in accordance with certain discernible laws
  - these laws can be described in a mathematical form, making them amenable to make prediction and to be tested for subsequent (possible) falsification

- An Example:
  - Tycho Brahe studied the planetary motion
    - Classified the data
  - Kepler looked for patterns
    - The three laws of Kepler describe the pattern
  - Newton gave the law of gravitation, a mathematical form.
    - The law, along with the laws of motion, could make predictions. This was completely deterministic.

- Other Examples: (Innate Randomness)
  - Flipping a coin; Throwing a dice
    - Requires an ‘ability’ to classify results of all flips/throws
  - Radioactive Decay
    - No. of decays in varying time intervals
    - Amount of matter initially
    - Look for patters
    - Obtain the exponential law

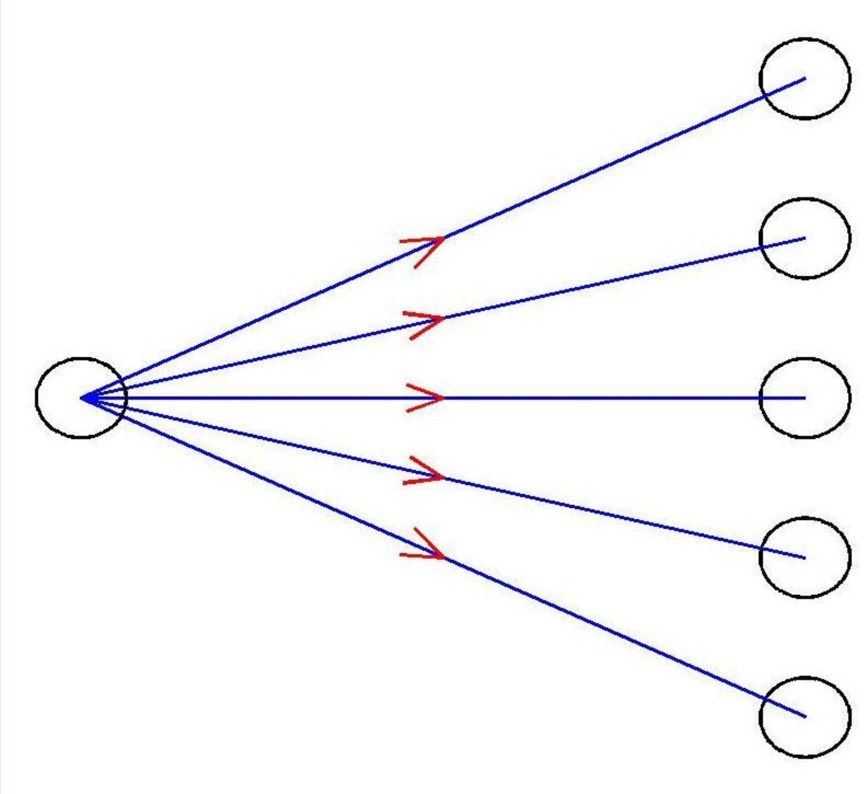
Empirical Science

- Any hypothesis is only (most) probable
- All hypotheses (models/theories) are accepted provisionally, until some data disproves it

- We have learnt to create data in laboratory
  - Enables systematic study
  - Discern Laws of Nature
- Given the data, how do we start? Reverse….

- Deductive Logic
  - Start with a premises
  - Draw definite conclusions

- Fair coin
  5 flips



1H, 4T; p= 0.1562

2H, 3T; p=0.3125

3H, 2T; p=0.3125

4H, 1T; p=0.1562

5H, 0T; p=0.0312

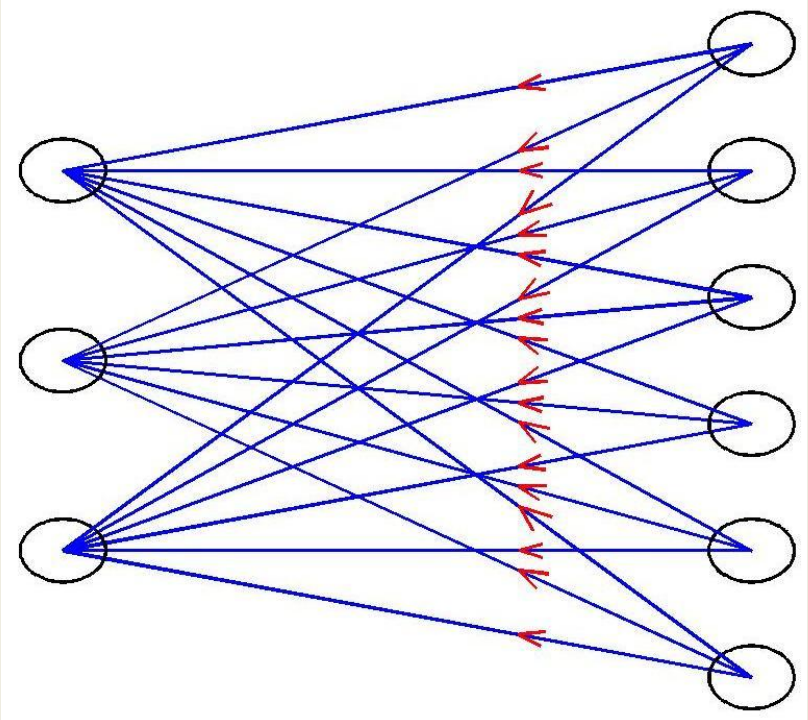
Privilege of a theorist !

- Inductive Logic
  - Experiment flipping 5 coins, 6 (or 6 Xillion) times

P (H) = 0.4

P(H) = 0.5

P(H) = 0.55



0H,5T; p=0.0312

1H, 4T; p=0.1562

2H, 3T; p=0.3125

3H, 2T; p=0.3125

4H, 1T; p=0.1562

5H, 0T; p=0.0312

What can we conclude about the coin? The wonderful and imaginative world of an experimentalist : a data analyst

- Guide inferences , draw objective conclusions
  - Assign Numbers
    - Make rules to assign numbers

Need a formalism

FORMALISM

- Rule 1: Given context 'I'

  $P(X \mid I)$ is probability of obtaining X

  $P(X\text{-}bar \mid I)$ is probability of NOT obtaining X

  $P(X \mid I) + P(X\text{-}bar \mid I) = 1$

- Rule 2: Given context 'I',

  Probability of obtaining X and Y is

  $P(X, Y \mid I) = P(X \mid Y, I) * P(Y \mid I)$

- 'Comma' means AND; '|' means GIVEN

- Useful Result 1: Bayes' Theorem

$$P(X,Y \mid I) = P(Y,X \mid I) \quad \&$$

$$P(Y,X \mid I) = P(Y \mid X,I) * P(X \mid I)$$

$$\therefore P(X \mid Y,I) = \frac{\mathrm{P(Y \mid X,I) * P(X \mid I)}}{\mathrm{P(Y \mid I)}}$$

P(hypo. | data,I) α P(data. | hypo., I)* P(hypo. | I)

(coins from casino)

P(data | hypothesis, I) can be obtained from deductive logic

Bayes' theorem becomes a boon

P(hypothesis | I) is prior probability

P(data |hypothesis, I) is likelihood function

P(hypothesis |data, I) is posterior probability

P(data | I) is evidence

Useful result 2: Marginalisation

$$P(X \mid I) = \int_{-\infty}^{\infty} P(X, Y \mid I)\, dY$$

Normalization

$$\int_{-\infty}^{\infty} P(Y \mid X, I)\, dY = 1$$

Helps to deal with 'nuisance' parameters

- An example:

Given:	Deduce
P(disease | I) = 0.001	P(no disease | I) =0.999
P(+ | disease, I) = 0.98	P(- | disease,I) =0.02
P(+ | no disease, I)=0.03	P(- | no disease,I)=0.97

Need to know

$$P(disease \mid +, I) = \frac{0.98 * 0.001}{(0.98 * 0.001) + (0.03 * 0.99)} = 0.032$$

- Interpretations:
  - In data analysis, probability interpreted as limiting relative frequency

$$P(X) = \lim_{n\to\infty} \frac{M}{N}$$

Here M is No. of occurrences of outcome X in N measurements

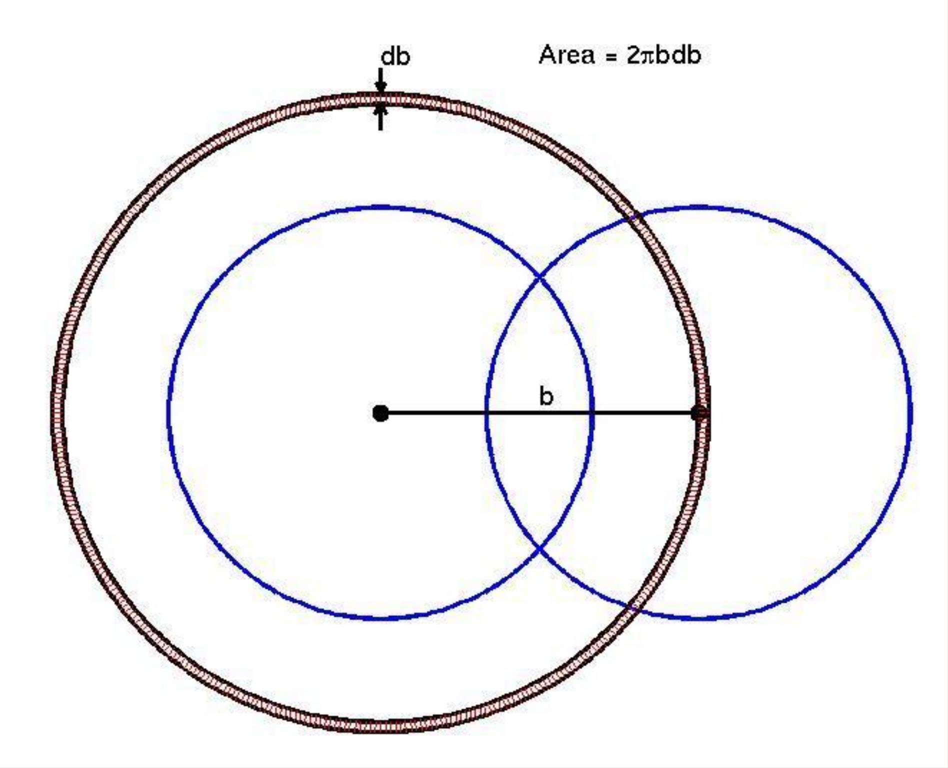
  - N is never infinite
- To estimate the probabilities, given a finite amount of experimental data
- Frequency interpretation may not work:
  - frequency distribution of electron mass ?
  - Probability gives a degree of belief.

- Example from Relativistic Heavy Ion Collisions
- Geometry plays an important role
  - Need to determine impact parameter ‘ b ’

$$\therefore P(b \mid n_{ch}, I) = \frac{P(n_{ch} \mid b, I) * P(b \mid I)}{P(n_{ch} \mid I)}$$

$$P(n_{ch} \mid I) = \sum_{b} P(n_{ch} \mid b, I) P(b \mid I)$$

$$P(b \mid I) \propto b$$



$$P(n_{ch} \mid b, I) \propto \exp\left[-\frac{(n_{ch}-n_0)^2}{2\sigma^2}\right]$$

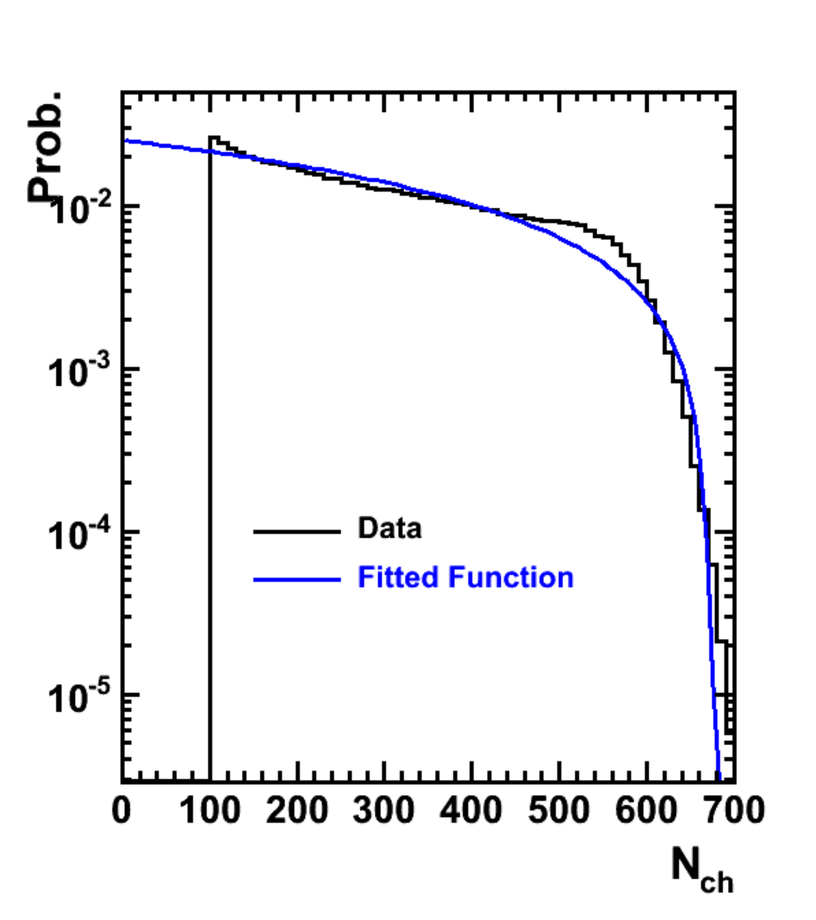
$$n_0 = a_1 + a_2 b$$

Integrate over ‘nuisance’ parameter ‘ b ’, and use

$$b_0 = \frac{n_{ch}-a_1}{a_2} \quad ; \quad \sigma_b = \frac{\sigma}{a_2}$$

$$P(n_{ch} \mid I) \propto \sigma_b^2 \exp\left[\frac{-b_0^2}{2\sigma_b^2}\right] + \sqrt{2\pi} b_0 \sigma_b + \int_0^{\frac{b_0}{\sqrt{2}\sigma_b}} e^{-t^2} dt$$

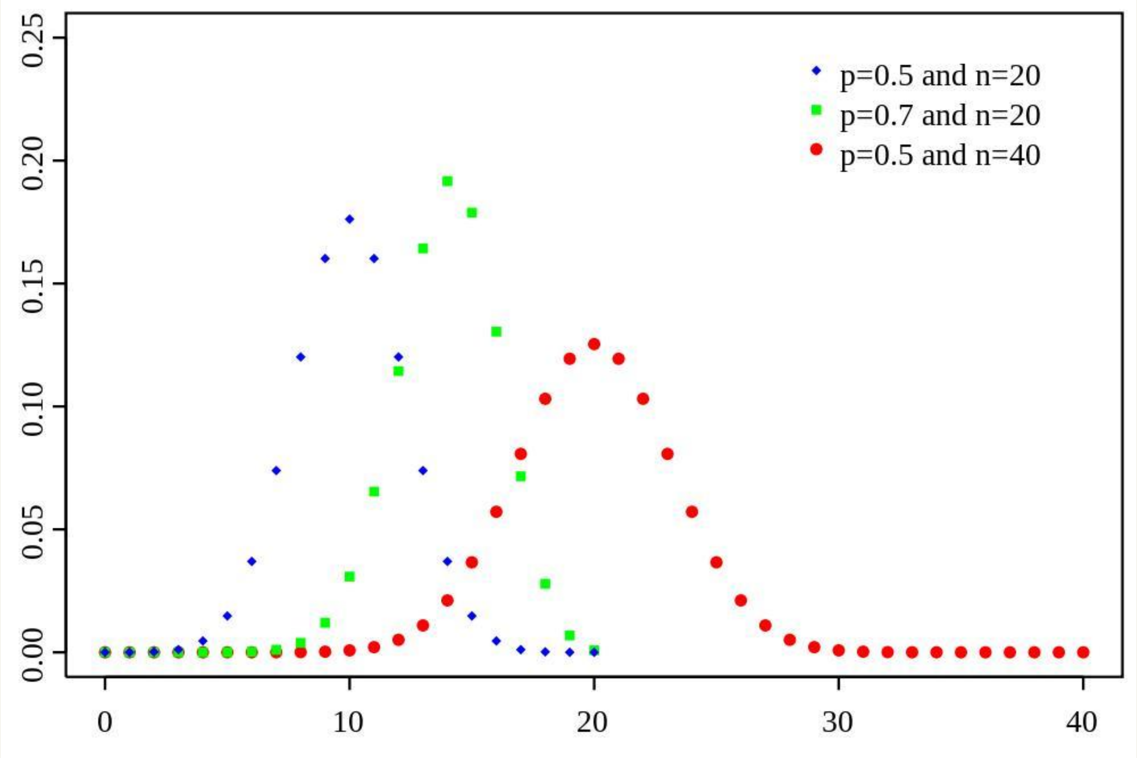
- The result of a certain data



- Gaussian ‘likelihood function’. There are more…

- Binomial
  - Probability of success: p
  - Given n turns, probability of r successes

$$P(r \mid n) = {}^{n}C_{r}\, p^{r} (1-p)^{n-r}$$



Multinomial

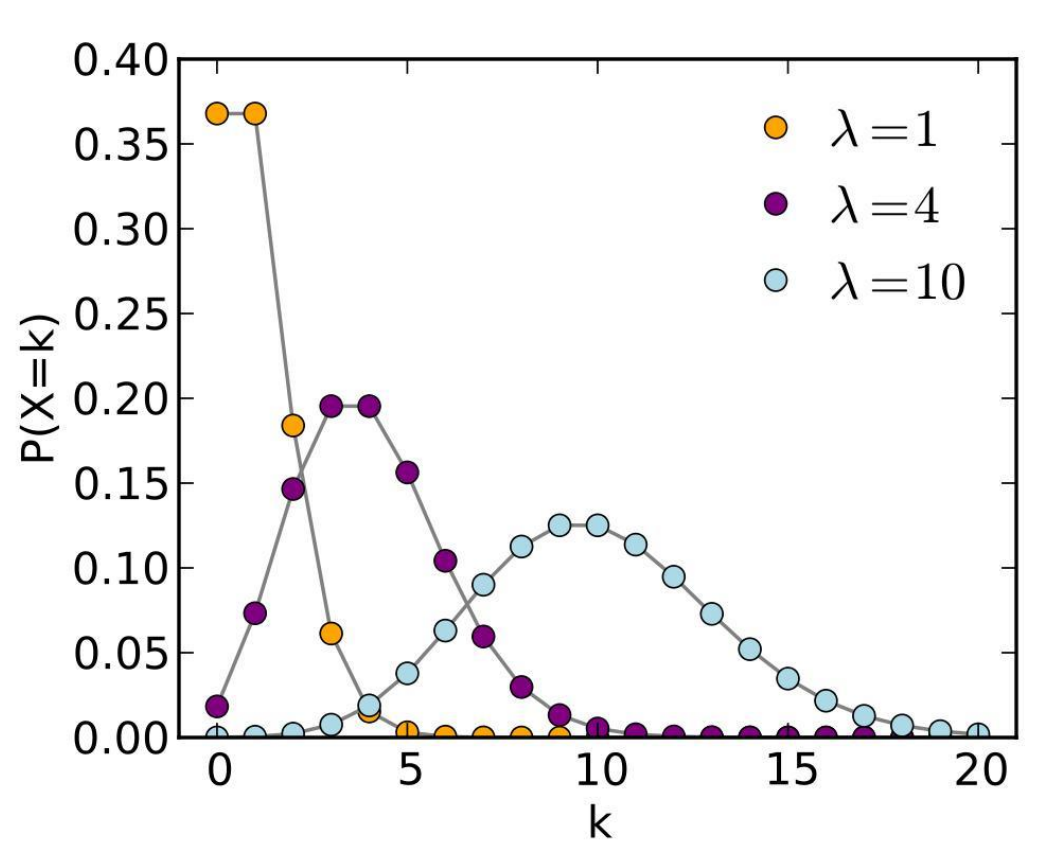
- Poisson

$$P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

$$<n> = \sum_{n=0}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} = \lambda$$

$$\sigma^2 = \sum_{n=0}^{\infty} (n-\lambda)^2 \; P_n = \lambda$$

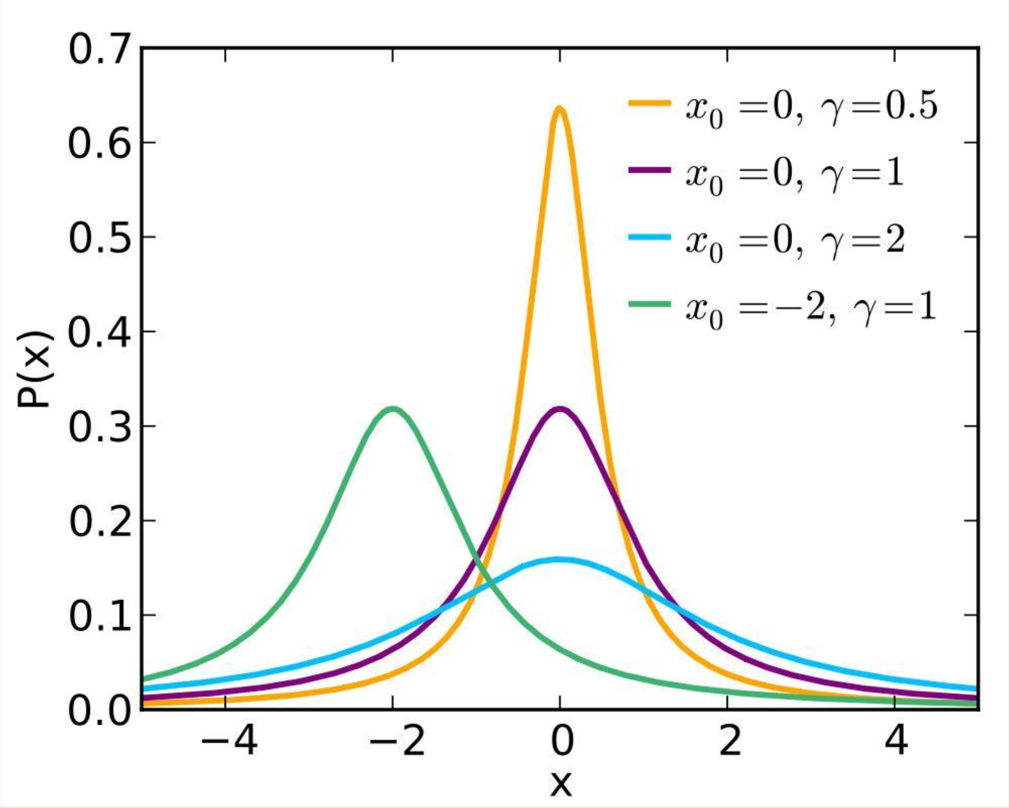
$$\therefore \quad \sigma = \sqrt{\lambda}$$



The forward-backward example with Binomial->Poisson

- Cauchy (Breit-Wigner)

$$P(x|\gamma, x_0, I) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - x_0)^2}$$

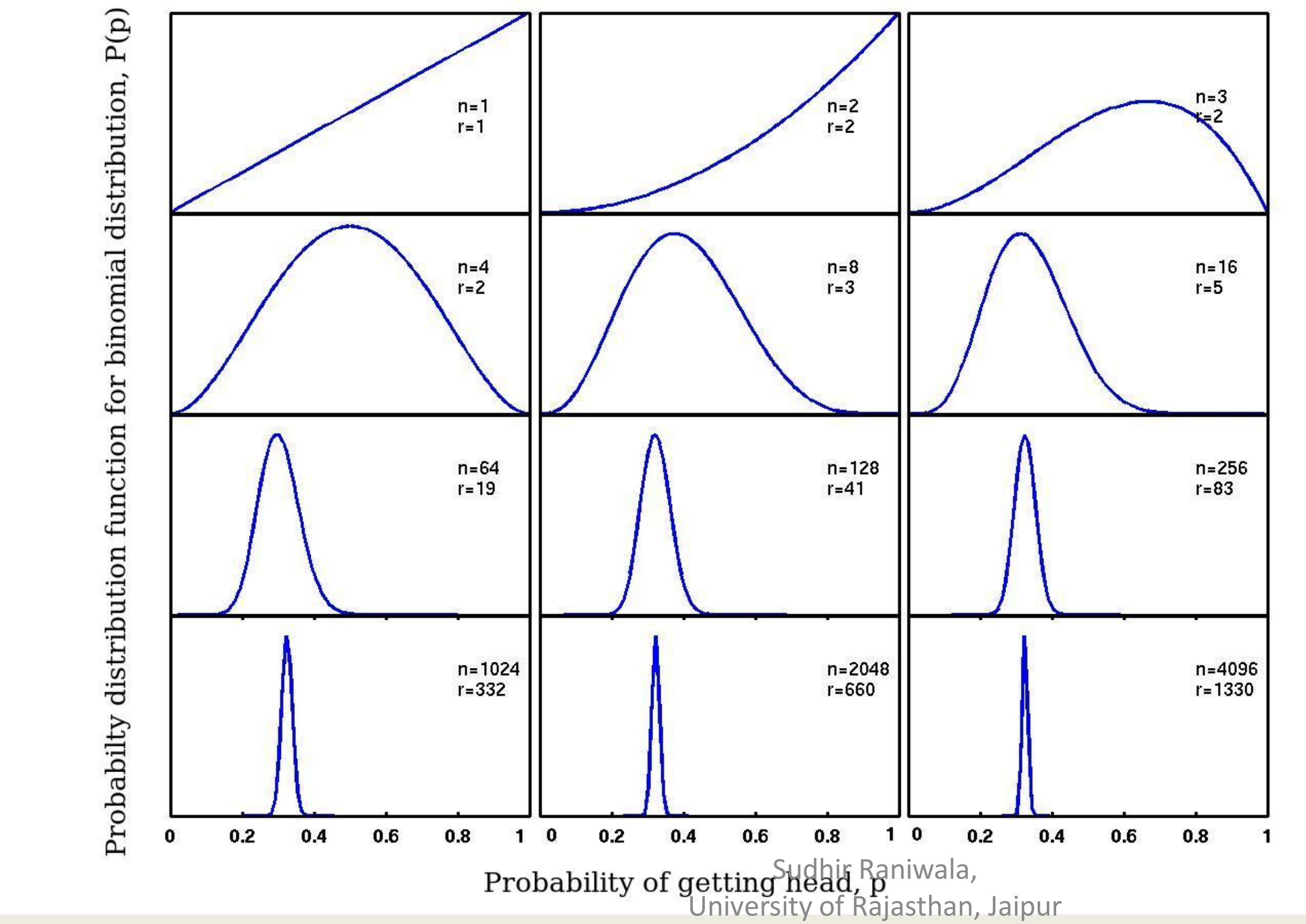


No. of events in a given mass bin…..

- Two fold purpose of data analysis
  - Testing hypothesis:
    - requires knowledge of parameter
  - determining parameter:
    - assumes valid hypothesis
  - deeply inter-related
- Parameter Determination:
  - $x \pm \Delta x$
- Hypothesis testing:
  - XX% probability that the statement is correct

Parameter Determination: Estimate the Bias of a Coin

- Generate data: flip the coin N times
- Need to assume prior probabilities



Purpose: Determine a parameter assuming the likelihood function to be a Binomial distribution. Result independent of prior !