Before beginning:

Books: We follow Feller Vol. 1 with excursions to Alon and Spencer; Mitzenmacher; Hoel-Port-Stone Vol 1. (see moodle page for titles). But you are free (and encouraged) to consult other books.

Some of you asked if we use measure. We shall not. But very important is: you should think about the concepts till you understand. With the advent of internet and social media, no one seems to think that they need to think – except those who want to brainwash you. Life seems to revolve around: click/read/forward! I would like you to be serious about learning.

Grading: Midsem: 40 and semestral: 60.

Our TAs are Ambaye Om and Tejas Oke. They would be glad to help you with problem sets. You should work out problem sets and write down solutions. You should practice writing proofs, especially proper justifications of steps needs practice. The Supreme court, in collaboration with our rulers has invented novel technique: you can be proved guilty based on 'confidential' evidence which you can not see and can not respond to. We shall not follow this technique. Our proofs must be transparent. Remember, I can only evaluate what you write; Not what you have in mind.

Beginning:

The phenomenon of chance is more visible in games and gambling. The origins of probability are indeed in gambling. Let us start with a concrete problem.

Anush and Soham are playing a match which consists of 9 games. The winner, one who wins a majority of games, will receive 800 Rs. Till now 5 games have been played: S won 3 and A won 2. Unfortunately the match has to be stopped for reasons beyond our control. How should the prize money be shared?

1. S says that he is the winner at this stage and so he should receive all. Thus S: 800 and A: 0.

2. A says that the match is stopped for no fault of his. So it should be regarded as draw and they should share equally.
Thus A: 400 and S: 400.

3. Then S says: I won three out of five games, the money should be

shared in the ratio 3:2.
Thus S: 480 and A: 320.

4. Fermat and Pascal then think as follows. Why did this problem arise at all: Because of uncertainty regarding the remaining unplayed 4 games. So the solution should be gotten by trying to understand this uncertainty. Let us see. There are 16 possible scenarios in the remaining four games. Of these 11 scenarios make S winner while 5 scenarios make A winner ($AAAA, AAAS, AASA, ASAA, SAAA$). Therefore, assuming that the players are equally good to win any one game, it is reasonable to award the money in this proportion 11:5.
Thus S: 550 and A: 250.

Thus entered the idea of understanding and analyzing uncertainty. Now-a-days you will reformulate the last argument as: the chances of S winning is 11/16 and the chances of A winning is 5/16 (given the results so far) and the money is shared in accordance to their chances of winning.

uncertainty:

Suppose that an experiment has total number of scenarios n. Of these, you are interested in an event B having k scenarios. Then chances of the event B, (that is our chances of observing an outcome in the set B when we really do the experiment), equals $\frac{k}{n}$ — where we assumed that the scenarios are equally likely. Most of the initial uses of this concept was in gambling.

Later, this concept of chance found interesting applications in several areas. One realized that life is full of uncertainty and thus chance analysis can be applied to understand many practical problems. For instance, will it rain tomorrow? We are not certain, but need to know in order to advise farmers and fishermen. You can try to predict. Agreed, it is possible that sometimes we go wrong in predictions, no problem, good enough to be correct most of the time and miss very very few times. You get data about the necessary parameters — like wind speed, cloud concentration, pressure, temperature etc; — to understand this uncertainty; make a model and predict.

Is it wise to release a particular medicine in the market? Here there is no data, you have to create. You need to conduct experiments; collect data regarding its effectiveness, side effects etc and then make a model and then take a decision.

Understanding chance phenomena was of supreme importance in physics too. It enters in several ways. Imagine a large container with water and also a few pollen particles. Assume that water is in equilibrium. Actually it is never in equilibrium. What we mean is that there are no external forces that make

pollen particle move — like water currents, air bubbles, evaporation etc. In practice you see the pollen particle still continues to perform motion. why does it move and how does it move. This motion is what you call Brownian Motion, because it was first observed by Robert Brown.

Here is a lesson for us. Keep your eyes open and observe things carefully. Brown, not only observed this motion, but also persisted that reasons must be found for this movement! He was even wondering if there is life in the pollen particles.

Many many years later Einstein explained the reason. It is all due to chance phenomenon: the surrounding water molecules keep hitting the pollen particle and displace it 'this way, that way' continuously and this causes motion. More precisely, particles we are talking about being small, one molecular bombardment may not cause observable movement of the pollen particle; but the large number of molecules of water and the consequence large number of bombardments cause an observable motion. Actually you do not consider *one* pollen particle. Throw some pollen particles. Ask: At time $t > 0$, what proportion of particles are at a distance x from their initial position. But let us not complicate life now. You should know that this probabilistic analysis was the basis for the first calculation of Avogadro number. Several years later Norbert Wiener proved that such motions can be mathematically modelled.

When you learn statistical mechanics, or quantum mechanics, you will see probability entering naturally. Now a days we use probability, in an essential way, in computer science too (randomized algorithms, simulation, $\cdots$) ; and modelling of stock market prices and so on. Do you see the similarity: pollen particle moves this way/that way; stock price moves up/down.

Prediction of crop yield is another essential problem. After all, if there is going to be shortage of a produce, Government should initiate purchase orders now for future procurement. Can not wait till you see actual yield to take action, because then it may be too late – you may not get the item or you may have to pay high prices.

Modelling uncertainty:

So how do you model chance experiments. The first thing is the following: you must know what happens when you do the experiment, that is, the possible scenarios. They are called **outcomes**. For example when we toss a coin twice, one after other, the scenarios are HH, HT, TH, TT with obvious understanding of the symbols used. Each one of these is called an outcome of the experiment.

The set of all outcomes is called **sample space** of the experiment. Usually sample space is denoted by Ω. Any subset of the sample space is called an **event**.

Experiment: Toss coin twice
Sample space Ω: $\{HH, HT, TH, TT\}$.
Experiment: Roll a six faced die once.
Sample space Ω: $\{1, 2, 3, 4, 5, 6\}$
Experiment: Toss a coin till you get heads and then stop.
Sample space Ω: $\{H;\ TH;\ TTH;\ TTTH;\ \cdots\cdots\}$.
So on and on.

In the first two examples there are finitely many outcomes. In the third example there are infinitely many outcomes, but countable. It is possible to think of experiments where there are more than countably many outcomes.

Experiment: Observe the bulb glowing above, do not switch off. Note down the life time of this bulb. Sample space Ω: $[0, \infty)$.

You can think of a radio-active particle disintegrating. The time till a beep occurs in the Geiger counter, that is, time till a particle is released.

We want to model the chance phenomenon in such experiments. Any such activity starts with understanding and modelling simple experiments. If they do not fit reality, then use the understanding thus far gained to make a realistic model; and so on. Basic philosophy: understand simple things first.

probability:

We have been using the word experiment and we shall describe several experiments shortly. But you might still wonder what exactly is 'experiment'. What exactly is an outcome. The idea of sample space is made precise as follows. As far us our mathematics is concerned, we simply start with a set (non-empty), and call this sample space and elements of Ω are called outcomes. Of course, the way you get the set Ω depends on the experiment you are considering. But what matters for us in modelling is the set and not who does the experiment, when etc. *Since we want to understand simple experiments, to start with, we assume that our set Ω is a finite or a countably infinite set.*

The idea of chance is made mathematically precise as follows. For each $\omega \in \Omega$ we associate a number, chance of that outcome, $p(\omega)$. This number will tell you the chances of observing the outcome ω, if and when you do the experiment. Since we believe chances of some thing happening should be

non-negative, we should have $p(\omega) \geq 0$ for each $\omega \in \Omega$.

Suppose we have two (different) outcomes ω_1, ω_2. It stands to reason to believe: the chances that one of the outcomes ω_1 or ω_2 appears is $p(\omega_1) + p(\omega_2)$. More generally if A is any set of outcomes then the chances, that one outcome in A occurs, should equal $\sum_{\omega \in A} p(\omega)$.

In particular, the chances that an outcome in Ω is observed, equals $\sum_{\omega \in \Omega} p(\omega)$. But Ω being the set of all possible scenarios, we are sure to observe something from Ω. We somehow feel that if an event never happens, its chances are zero; whereas if an event is sure to happen then its chances are one. In other words, the above sum should equal one, $\sum_{\omega \in \Omega} p(\omega) = 1$.

A probability space is a pair (Ω, p) where Ω is a finite or a countably infinite set; and p a function that associates with each $\omega \in \Omega$ a non-negative number $p(\omega)$ in such a way that all these numbers add to one.

Ω is called sample space; elements $\omega \in \Omega$ are called outcomes. The number $p(\omega)$ is supposed to tell us the chances of observing the outcome ω. Subsets of Ω are called events. We then define Probability of an event A; denoted $P(A)$ by $P(A) = \sum_{\omega \in A} p(\omega)$.

Observe that the definition above does not use the words 'experiment', 'chance' etc. We can refer to (or define) the pair (Ω, p) as experiment!

For example, if the sample space is $\{H, T\}$ then $p(H) = 0.3$ and $p(T) = 0.7$ is a possible assignment. Of course $p(H) = 0.5$ and $p(T) = 0.5$ is also an assignment of probability. In other words these two assignments are two models. At this stage, you might get worried as to how to get these probabilities and who gives them. Instead of getting worried, you should be happy that it is left to your choice and you can start with any assignment of your liking and build models. Proceed to do calculations, evaluate probabilities of complicated events and so on. You can build several models for the same phenomenon.

The natural question that should bother you is the following. I started saying that probability is useful in understanding phenomena and making predictions. So what is the worth of these models and which model should we follow for our predictions. This is where you need to make some observations of the phenomenon and develop techniques that help choosing *one of your models* that fits reality and use it to predict. This body of techniques goes by the name of 'statistics'.

In high school you have learnt ratios (of favourable outcomes to total number of outcomes or whatever your teacher taught you) as probabilities.

So, are we developing something different? Not really, we are developing more general models which include what you learnt, as special cases.

equally likely outcomes:

We first start understanding experiments which have finitely many outcomes. We further assume that the outcomes are equally likely; we have only intuitive idea of what this means and must make a mathematical model. For example in tossing a coin once, there are two outcomes H, T and we assume that they are equally likely, so each must have chance $1/2$ (because total must be 1).

Let Ω be the sample space, finite, of our experiment. If outcomes of Ω are equally likely, there is a number c such that $p(\omega) = c$ for all ω and since they should add to one; we conclude $c|\Omega| = 1 or\ p(\omega) = 1/|\Omega|$ for each outcome. This immediately leads to the conclusion that chances of any event equals the fraction of outcomes which belong to that event. Remember probability of an event is the sum of probabilities of all outcomes in that event.

In an experiment with finitely many equally likely outcomes, the probability of an event A equals the ratio

$$\frac{number\ of\ outcomes\ in\ A}{total\ number\ of\ outcomes}$$

Thus calculating probability reduces to counting number of outcomes in event of interest. This is one reason for considering finite sample spaces and equally likely outcomes. You can use your expertise in counting which you learnt in high school. Also, this formula is exactly what you learnt in high school.

You must understand two things. Most of the time we consider tossing coin or throwing balls into boxes. These are only symbolic and apply to many situations. For example any analysis involving coin tossing can be applied to experiments where there are just two outcomes; H/T or Success/Failure or On/Off or $+1/-1$ or good/defective etc. Balls could be interpreted as particles and boxes could be interpreted as energy levels.

Secondly, 'equally likely' is only a first step. For example if you are manager of a company manufacturing bolts; each bolt could be good or defective. It is disastrous for the company to assume that these two outcomes are equally likely.

We shall now discuss several examples to see how the seemingly meaningless (some of you may have even felt 'trivial to do') calculations you made in high school come to life and also see how calculations are not always that trivial.

Example 1:

I have two usual decks of 52 playing cards each. I arrange one set of cards in a line, one after another. I now shuffle the second deck and place the cards in a line sequentially below each card in the first line. It is interesting to ask: what are the chances of a match; match at place i means the same card (same denomination) appears in the two rows at place i. So the question is: what are the chances that there is match in at least one place.

What is the experiment here? Place the second deck below first one as stated. What is the sample space? All possible arrangements of the 52 cards, of the second deck, in a line. This constitutes Ω. How many outcomes are there? 52!. is it possible to write them down and see how many of them are in our event? It takes a very very long time, so it is not the best way of doing it.

Let A be the set of all outcomes where there is at least one match. It is not easy to directly count $|A|$, number of elements in A. So we split this event into simpler events and use 'inclusion-exclusion' formula. Let, for $1 \leq i \leq 52$, A_i be the set of all outcomes in which the cards at the i-th place are same. Clearly $A = \cup A_i$.

■ Inclusion-Exclusion principle:

$$|\bigcup_1^n A_i| = S_1 - S_2 + S_3 - S_4 + \cdots$$

$S_1 = \sum_1^n |A_i|$; $\quad S_2 = \sum_{i<j} |A_i \cap A_j|$; $\quad S_3 = \sum_{i<j<k} |A_i \cap A_j \cap A_k|$; $\quad \cdots$. ■

Is this formula correct? We shall see later.

Is this formula useful? After all, if we are unable to calculate the numbers S_i then this formula is also useless. We are lucky. I leave for you to check, in our example,

$$|A_i| = 51!; \qquad |A_i \cap A_j| = 50! \quad (i \neq j)$$

$$|A_i \cap A_j \cap A_k| = 49! \quad (i < j < k) \qquad \cdots$$

Thus

$$S_1 = 52 \times 51! = 52!$$

$$S_2 = \binom{52}{2} \times 50! = \frac{52!}{2!}$$

$$S_3 = \binom{52}{3} \times 49! = \frac{52!}{3!}$$

Thus

$$P(A) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots \pm \frac{1}{52!}$$

This is the chances that there is at least one match. I used $\pm$ for the last term, we know the sign is indeed minus (check).

If you want chances of no match, $P(A^c)$ you will see

$$P(A^c) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{52!}$$

Here are some comments on the power of technique/argument used above which is true of many mathematical arguments.

Instead of 52 cards, if we have two similar decks of n cards, then the chances that there is no match is given by

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{n!}$$

This is pretty close to $1/e$ if n is pretty large.

In other words the argument is still applicable and gives an answer. We can also see that the answer is close to a number we already knew.

symbols are unimportant. Consider the following problem: I have 52 letters to different persons and 52 envelopes with their addresses. I place the letters at random in the envelopes. What are the chances that at least one letter goes to its envelope. You will see the answer is exactly same as above. Convince yourself that this is so.

Did you realize that the word 'shuffle' quietly disappeared after statement of the problem. Where did it go? It went into the model! It is difficult to explain what is shuffling – who will do, how many times, by what method, etc. But what is its purpose? Its purpose is that all arrangements are equally likely. That is precisely what we used in our calculation. In other words what is achieved by our mind in using 'shuffling' is precisely achieved by the mathematical modelling.

This is called **matching experiment**. There are several other things you can do with this experiment.

Example 2:

Let $A = \{1, 2, \cdots, N\}$. Experiment consists of picking one after another, twenty times, points from this set A. Remember always selected from A, we are not excluding the already selected elements when selecting a later point. Thus we are making a list of 20 items from A, with repetitions allowed.

Sample space Ω consists of all sequences $(x_1, x_2, \cdots, x_{20})$ where for each i, $x_i \in A$. Check $|\Omega| = N^{20}$. Each outcome has probability $1/N^{20}$.

This experiment is called **sampling with replacement, of size 20**. With replacement because when we take second item the first item is not removed from the set A; the second selection is made as if you noted down the first item and put it back before second selection. **Each outcome of this experiment is called a sample (with replacement) of size 20.**

What are the chances that first observation (first point of the sample) is 1? Answer: there are N^{19} outcomes for which first item is 1. So the required probability is $N^{19}/N^{20} = 1/N$

What are the chances that second item is 1? Again there are N^{19} outcomes where second item is 1. So required probability is again $1/N$.

What are the chances that first and second item are 1? There are N^{18} such outcomes and so the probability is $N^{18}/N^{20} = 1/N^2$.

What are the chances that 1 is included in the sample, that is some x_i in the sample is 1? There are $(N-1)^{19}$ outcomes where $x_i \neq 1$ for all i and hence the required probability equals

$$\frac{N^{20} - (N-1)^{20}}{N^{20}} = 1 - (1 - \frac{1}{N})^{20}$$

Of course you can talk about sample of size n. The set from which you are picking, namely, $\{1, 2, \cdots, N\}$ is called population. It need not be numbers. it could consist of all students of cmi.

Example 3:

Again let $A = \{1, 2, \cdots, N\}$. Experiment consists of picking one after another, twenty times, points from this set A, each time NOT replacing the points selected earlier.

Sample space Ω consists of all sequences $(x_1, x_2, \cdots, x_{20})$ where for each i, $x_i \in A$ and they are distinct. That is $x_i \neq x_j$ for $i \neq j$.

Check $|\Omega| = N(N-1)\cdots(N-19)$.

We assume that $N \geq 20$. Otherwise the sample space is empty set.

Each outcome has probability

$$\frac{1}{N(N-1)\cdots(N-19)}.$$

This experiment is called **sampling without replacement, of size 20**. Without replacement because when we take second item the first item is removed from the set A; similarly at each stage selected item is removed and then next selection is made. **Each outcome of this experiment is called a sample (without replacement) of size 20.**

What are the chances that first observation (first point of the sample) is 1? Answer: there are $(N-1)(N-2)\cdots(N-19)$ outcomes for which first item is 1. So the required probability is $1/N$

What are the chances that second item is 1? Again there are $(N-1)(N-2)\cdots(N-19)$ outcomes where second item is 1. So required probability is again $1/N$.

What are the chances that first and second item are 1? Zero.

What are the chances that 1 is included in the sample, that is some x_i in the sample is 1? There are $(N-1)(N-2)\cdots(N-20)$ outcomes where $x_i \neq 1$ for all i and hence the required probability equals $1 - \frac{N-20}{N} = \frac{20}{N}$

Example 4:

Again let $A = \{1, 2, \cdots, N\}$. Experiment consists of picking a subset of A of size 20.

Remember, we are not picking elements one by one. We grab a subset consisting of 20 elements.

Sample space Ω consists of all subsets of A which have 20 elements. That is all $\omega \subset A$ with $|\omega| = 20$ Check $|\Omega| = \binom{N}{20}$

We assume that $N \geq 20$. Otherwise the sample space is empty set.

Each outcome has probability $\frac{1}{\binom{N}{20}}$.

This experiment is called **selecting a subset of size 20 or sampling a subset of size 20**. Each outcome is called a random subset of size 20.

What are the chances that first observation (first point of the sample) is 1? this question does not make sense. There is no first or second element. We have a subset consisting of 20 points, that is all.

What are the chances that second item is 1? Again meaningless question.

What are the chances that 1 is included in the selected set? that is $1 \in \omega$? There are $\binom{N-1}{19}$ such outcomes and hence the probability equals

$$\frac{\binom{N-1}{19}}{\binom{N}{20}} = \frac{(N-1)!}{19!(N-20)!}\frac{20!(N-20)!}{N!} = \frac{20}{N}$$

same as in the case of sampling without replacement.

For example when you want to make a committee of 3 students, it is the three that matters and not who is selected first etc.

Example 5:

I have 30 boxes numbered: $1, 2, \cdots, 30$. I have 20 balls numbered: $1, 2, \cdots, 20$. Experiment is to throw the balls into the boxes. $|\Omega| = 30^{20}$. all these outcomes are equally likely.

What are the chances that the first box is empty? $\frac{29^{20}}{30^{20}}$

What are the chances that boxes 1,5,7 contain respectively 10,7,3 balls? $\binom{20}{10}\binom{10}{7}/30^{20}$ Remember, any 10 balls could go in box 1.

This experiment is called **Maxwell-Boltzman experiment**. Boxes are energy levels and balls are elementary particles. The interesting point is: apparently, no known particles obey this rule!

Example 6:

I have 30 boxes numbered: $1, 2, \cdots, 30$. I have 20 balls all looking alike. There is no way to distinguish one ball from the other. Experiment is to throw the balls into the boxes. How many different arrangements can our eye perceive? $|\Omega| = \binom{49}{29}$.

This can be seen in several ways. Put 49 star marks in a row; select 29 of these and convert them into vertical lines. You see a picture of 20 balls (the remaining star marks) put in 30 boxes (made by the vertical lines). And every method of putting balls into the boxes is achieved this way. remember, that now it makes no sense to say: where did ball one go? all balls look alike. so two arrangements are different only when the 'occupancy numbers' are different, that is, the vector $(n_1, n_2, \cdots, n_{30})$ where n_i is the number of balls in box i; determines the arrangement. Two different vectors give two different arrangements.

Bose-Einstein Rule: These distinguishable arrangements are equally likely.

This experiment is called **Bose-Einstein experiment**. Again boxes are energy levels and balls are elementary particles.

What are the chances that box one is empty? $\binom{48}{28}/\binom{49}{29}$ You can simplify and see.

What are the chances that boxes 1,5,7 contain respectively 10,7,3 balls? $1/\binom{49}{29}$ This is because there is only one outcome satisfying the given condition.

Particles that obey this rule are called Bosons. Photons are known to obey this rule. To understand how outrageous is this rule consider the following.

I have a box with 1 green ball and 100 red balls numbered 1,2,$\cdots$, 100. I pick a ball at random. What are the chances it is green: 1/101.

Suppose now I tell you that all red balls look alike, there are no numbers on them. How many outcomes your eye can perceive? only two: Red, green; Can I say that these are equally likely and conclude that chance of green ball is 1/2? Ridiculous. The B-E rule *does not* apply here. Thus you should be careful. Without knowing if B-E rule is 'assumed', you can not start calculating just because you are told 'balls (or whatever)' look alike.

Why photons obey B-E rule is unclear, but they do.

Example. 7:

I have 30 boxes numbered: $1, 2, \cdots, 30$. I have 20 balls all looking alike. There is no way to distinguish one ball from the other. Experiment is to throw the balls into the boxes subject to the condition: No more than one ball in a box. $|\Omega| = \binom{30}{20}$ These are equally likely.

What are the chances that box one is empty? $\binom{29}{20}/\binom{30}{20}$. What are the chances that boxes 1,5,7 contain respectively 10,7,3 balls? Zero.

This is called **Fermi-Dirac experiment**. Protons obey this rule.

Example 8:

We want to elect monitor for the class. There are two candidates: Ananya and Dwitimaya. the forty students vote: D got 24 votes and S got 16 votes and thus D is the winner.

What are the chances that through out vote-counting, D is leading? That is through out counting D maintains lead.

We discuss this problem because it teaches a new counting technique and the method has lots of applications. Further the problem itself has a neat answer:

$$\frac{24-16}{24+16} = \frac{8}{40}.$$

What is vote counting? No, do not think of electronic machines; there is no counting you do there, you press a button and answer pops up. Let us count a vote for D as $+1$ and a vote for A as -1. Thus vote counting means all possible sequences of ± 1 of length 40 which have 24 ones and 16 minus ones. This is the sample space. Thus $|\Omega| = \binom{40}{24}$

We assume that all these outcomes are equally likely. Thus the problem boils down to finding $|T|$ where T consists of sequences which have at every stage more $+1$ than -1.

If we think of outcomes as simply sequences of ± 1 then to see if an outcome is in T or not, we need to add and check at every stage. So let us think of Ω in a different and convenient manner.

A path is a sequence $(0, s_0), (1, s_1), \cdots (k, s_k)$ where each s_i is an integer and $s_i - s_{i-1} = \pm 1$ for all $i \geq 1$. The path is said to start at s_0 and end at s_k and is of length k. You can also talk about paths starting at $(1, 5)$ etc.

We claim Ω can be thought of as the set of paths of length 40 starting at 0 and ending at 8. obviously any such path defines a vote counting; simply

$$(s_i - s_{i-1} : i = 1, 2, \cdots, 40)$$

Verify that this has exactly 24 ones and 16 minus ones (because path ends at 8 and starts at zero). Conversely, given a vote counting $(\epsilon_i : i = 1, 2, \cdots, 40)$ you can define a path by taking

$$s_0 = 0, \qquad s_m = \sum_1^m \epsilon_i \quad 1 \leq m \leq 40.$$

Verify that this path starts at zero and ends at 8.

You can draw the axes and join successive points by straight line and visualize a path. such a picturesque visualization suggests new ideas as we see now.

A vote counting corresponds to a path starting at $(0, 0)$ and ending at $(40, 8)$. Number of paths starting at $(0, 0)$ and ending at (n, r) are

$$\binom{n}{\frac{n+r}{2}} = \binom{n}{\frac{n-r}{2}}.$$

This is because, to have such a path the number of ones must be $(n + r)/2$ and number minuses must be $(n - r)/2$. If $k \geq 0$ is not an integer, we take $\binom{n}{k}$ to be zero. In the formula above if $(n+r)/2$ is not an integer (≥ 0) then value is by definition, zero.

You must understand a very important matter here. The formula is stated above as a fact with a proof. However the formula '$\binom{n}{k} = 0$ when k is not integer' is a convention. It is something we adapted. You can not confuse conventions with facts. So the question arises, if $(n + r)/2$ is not an integer, can you show that the number of paths from $(0, 0)$ to (n, r) is indeed zero, so that the convention agrees with the stated formula. Yes.

Thus total number of vote countings are $\binom{40}{24}$. Let us note a simple but powerful fact.

Reflection Principle:

■ Let $k \geq 1$ and $r \geq 1$ be integers.

number of paths from $(0, k)$ to (n, r) which touch or cross the x-axis (after starting)

equals

number of paths from $(0, -k)$ to (n, r). ∎

Note $(0, -k)$ is reflection of $(0, k)$ in the x-axis. The proof of this fact uses reflection in x-axis.

Let A be the first set of paths and B be the second set of paths. Take a path π in A: $(0, s_0), (1, s_1), \cdots\cdots (n, s_n)$; $\quad s_0 = k$; $\quad s_n = r$.

Since it touches the x-axis; let i be the first index with $s_i = 0$. Reflect the path till that point in x-axis. do not reflect the remaining segment of the path. That is, consider the path π^*:

$$(0, -s_0), (1, -s_1), \cdots (i, -s_i), (i+1, s_{i+1}), (i+2, s_{i+2}), \cdots (n, s_n).$$

Remembering that $s_i = 0$, we see (i, s_i) is same as $(i, -s_i)$. We can easily verify that this is also a path, that is, successive s-differences are ± 1. This path starts at $(0, -k)$ and ends at (n, r). Hence $\pi^* \in B$.

Given path $\eta \in B$ it starts below x-axis and ends above; so must hit x-axis, take the first time it hits and reflect the part till then in x-axis, keeping later part as is. This will be a path $\pi \in A$ and $\eta = \pi^*$ and actually the map $\eta \mapsto \pi$ is inverse map of $\pi \mapsto \pi^*$.

This one-one map between A and B proves the result. ∎

We are interested in

number of paths $(0, 0)$ to $(40, 8)$ that do not touch/cross x-axis.

[such a path should pass through $(1, 1)$]

= number of paths from $(1, 1)$ to $(40, 8)$ that do not touch x-axis

[subtract 1 from first coordinate to see]

= paths from $(0, 1)$ to $(39, 8)$ that do not etc

= [Total number of paths $(0, 1)$ to $(39, 8)$] minus [number of paths $(0, 1)$ to $(39, 8)$ that touch etc]

(Use reflection principle)

=[paths $(0, 1)$ to $(39, 8)$] minus [paths $(0, -1)$ to $(39, 8)$]

= [paths $(0, 0)$ to $(39, 7)$] minus [paths $(0, 0)$ to $(39, 9)$]

$$= \binom{39}{23} \text{minus} \binom{39}{24}$$

$$= \binom{40}{24} \frac{24 - 16}{40}.$$

Since the total number of outcomes is $\binom{40}{24}$ we conclude that the required probability is $8/40$ as stated.

This is **path counting example** and there are several interesting consequences of the reflection principle which we see later.

Example 9:

This is **stick breaking example**.

I have 23 sticks numbered $1, 2, \cdots, 23$. Or 23 sticks of different colors. Break each into two parts one small part and one large part. Now I make a bag of 23 sticks again by pairing these 46 pieces. What are the chances of getting back original pairing (original sticks)? What is the probability that long pieces are paired with small parts?

Here

$$|\Omega| = \frac{46!}{2^{23}\ (23)!}$$

This is because, if c is the number of ways of pairing, consider the problem of pairing and arranging the pairs in a line. This can be done using two methods:

(Method 1)

make 23 pairs (c ways) and then arrange these 23 pairs in a row ((23)!ways). so the required number is $c\ (23)!$ ways.

(Method 2)

pick a pair out of the 46 pieces, put in place one;
pick a pair out of the rest, put in place 2 etc.
so can be done in

$$\binom{46}{2}\binom{44}{2}\cdots\binom{2}{2} = \frac{46!}{2^{23}}$$

As a result

$$c\ (23)! = \frac{46!}{2^{23}}; \qquad c = \frac{46!}{2^{23}(23)!}$$

Thus each outcome has probability

$$\frac{1}{c} = \frac{2^{23}\ (23)!}{(46)!}$$

This is also the probability of getting original pairing.

If A is the set of all pairings where large parts are paired with short parts, then $|A| = (23)!$ — you need to put long pieces in a line, no matter how, and then place short pieces one with each long piece. So the required probability equals

$$\frac{2^{23}\ (23)!\ (23)!}{(46)!} = \frac{2^{23}}{\binom{46}{23}}$$

As you could guess, these 23 sticks are the 23 chromosomes in human cell. They break at the centromere, giving you one long part and one small part – centromere is not exactly in the centre. Then they recombine. The first question asks the chances of getting original chromosomes back or original cell back. The second question asks for the chances of the cell dying; if two long pieces join there is not enough space in the cell to fit in. what happens if two short pieces join?

What happens if there are N sticks? For example, Drasofila has seven pairs of chromosomes.

Let us now prove the inclusion-exclusion principle. We do in a little more generality, which will be useful later.

Let Ω be a finite set. Suppose for every point $\omega \in \Omega$, we are given a number $f(\omega)$. Let us define for every subset $A \subset \Omega$,

$$F(A) = \sum_{\omega \in A} f(\omega).$$

We take $F(\emptyset) = 0$.
Here is the formula: Suppose $A_1, \ldots, A_n$ are subsets of Ω, then

$$F(\cup A_i) = S_1 - S_2 + S_3 - \cdots \quad (\spadesuit)$$

$A \cap B$ where

$$S_1 = \sum_i F(A_i);\ S_2 = \sum_{i<j} F(A_i \cap A_j);\ S_3 = \sum_{i<j<k} F(A_i \cap A_j \cap A_k);\ etc$$

When each $f(\omega) = 1$, then $F(A) = |A|$ giving the formula stated in the matching problem.
When we have a probability space (Ω, p) then taking $f(\omega) = p(\omega)$, we get similar formula for $P(A)$, probabilities of events.

Proof 1: Fix $\omega \in \cup A_i$. we show
(i) if $\omega \notin \cup A_i$ then $f(\omega)$ is not added on the right side of ($\spadesuit$).
(ii)if $\omega \in \cup A_i$ then $f(\omega)$ is added exactly once on the right side of ($\spadesuit$).
Statement (i) is clear because then, $f(\omega)$ does not appear in the calculations of $F(A_i)$ or $F(A_i \cap A_j)$ etc.

For proof of (ii), suppose ω occurs in exactly k of the sets; say A_{m_p}, $1 \leq p \leq k$. Then
$f(\omega)$ is added k times in calculation of S_1, it appears in the calculation of $F(A_{m_p})$ for each p.

$f(\omega)$ is added $\binom{k}{2}$ times in calculation of S_2, it appears in the calculation of $F(A_{m_p} \cap A_{m_q})$ for each $p < q$.
and so on.
Thus $f(\omega)$ is added

$$k - \binom{k}{2} + \binom{k}{3} - \cdots = 1 - (1-1)^k = 1$$

This completes the proof.

Proof 2: Use induction on n. Do it for $n = 2$ and proceed. Execute the proof

conditional probability:

One of the most important concepts is that of conditional probability. Most of the time we are not totally ignorant. There is a chance experiment, we do not know the outcome but we know some partial information about the outcome.

For example when you are predicting tomorrow's weather, you are not completely ignorant, you do not close your eye and say there are two possibilities etc. You will see the information about wind speed, temperature, pressure etc. Similarly, if a mother goes to a doctor and asks if her two year old baby is of correct weight, then the doctor does not blindly compare baby's weight with the average weight of two years olds. He will see several parameters like the birth weight etc and compares with the average weight of children with similar parameters.

Example: Consider rolling a fair die.
$\Omega = \{1, 2, 3, 4, 5, 6\}$ and $P(A) = |A|/6$.
Let A be the event $A = \{2, 3, 4\}$. Then $P(A) = 3/6$.
Some one told us that an even number appeared. How shall we define probabilities now? Let us denote $B = \{2, 4, 6\}$. It is meaningless to say that the probabilities of events are as earlier. strictly speaking, it is meaningful but practically useless. If we are now asked chances of the event A then we feel that $P(A)$ should be $2/3$. Because, there are now only three possibilities: $2, 4, 6$, of which the event A has two: $2, 4$. Of course, if we say $P(A) = 2/3$ there will be total confusion because in the original experiment $P(A) = 3/6$. So we say $P(A|B) = 2/3$; read conditional probability of A given B or probability of A given B. Here 'given B' means that an outcome in B occurred.

Similarly, if tomorrow someone tells that an odd number occurred, then you need to rethink and say that $1, 3, 5$, are the possible outcomes with this information $P(A) = 1/3$ etc. It is a nuisance to keep on changing sample space and *then* calculate. One smart fellow found out the quantity $P(A|B)$ is nothing but $P(A \cap B)/P(B)$. You can also check this! This is what was used in high school.

This formula has the advantage that it expresses the conditional probability, that we felt, as ratio of usual probabilities. Thus you can calculate conditional probability without any sample space considerations, using only

probabilities in the original experiment . We take this as the definition.

Definition: Suppose Ω be a probability space with equally likely outcomes (thus $p(\omega) = 1/|\Omega|$ for every outcome). Then we define for events A and B, conditional probability of event A given event B, by

$$P(A \,|\, B) = P(A \cap B)/P(B)$$

of course 'given' means we are told that an event in B occurred.

Warning: $A \,|\, B$ is NOT an event. Thus to say $P(A \,|\, B)$ is probability of the event $A \,|\, B$ is meaningless. It is conditional probability of the event A given the event B.

Before we continue the story further with equally likely outcomes, let us consider general experiments.

not necessarily equally likely outcomes:

Let us now discuss experiments where the outcomes may not be equally likely. Let (Ω, p) be a probability space. Remember, we then defined for any event A; $P(A) = \sum_{\omega \in A} p(\omega)$.

Definition: Let (Ω, p) be a probability space. Then for any two events A and B; we define conditional probability of A given B by

$$P(A \,|\, B) = P(A \cap B)/P(B)$$

We define A, B to be independent if $P(A \,|\, B) = P(A)$, or equivalently if, $P(A \cap B) = P(A)P(B)$.

We talk about conditional probabilities only when $P(B) \neq 0$. You can see that the above definition is imitation of what we arrived at in the equally likely case. Is there any justification to use it in this general case? Yes, shall illustrate by an example; though you can argue more generally.

Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$; $\quad p(k) = k/21$.
Thus the die is loaded/biased, chances of a face are proportional to the number on that face. As earlier let $A = \{2, 3, 4\}$ and $B = \{2, 4, 6\}$ Suppose that we are told B occurred. How should we redefine our probabilities? Clearly then $p^*(1) = p^*(3) = p^*(5) = 0$. Suppose $p^*(2) = a$ then we must have $p^*(4) = 2a$ because for our die chances of 4 are double the chances of 2. Similarly $p^*(6) = 3a$ but since we must have $\sum_{\omega \in \Omega} p^*(\omega) = 1$ we must have: when we are told B occurred

$$p^*(1) = p^*(3) = p^*(5) = 0;\ p^*(2) = 1/6;\ p^*(4) = 2/6;\ p^*(6) = 3/6$$

Hence under the information given, we feel

$$P(A \,|\, B) = p^*(2) + p^*(3) + p^*(4) = \frac{1+2}{6} = \frac{1}{2} \quad \text{(feel)}$$

Under the information, The suggested formula gives

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{6/21}{12/21} = \frac{1}{2} \quad \text{(suggested formula)}$$

Thus the feeling agrees with the suggested formula. As I said, you can justify the suggested formula in all experiments, not only in this example.

The definition of conditional probability can be used to calculate $P(A \cap B)$ if you knew $P(B)$ and $P(A \,|\, B)$; simply using $P(A \cap B) = P(B)P(A \,|\, B)$.

Example: $\Omega = \{HH, HT, TH, TT\}$
probabilities: $p(HH) = 0.1;\ p(HT) = 0.2;\ p(TH) = 0.3;\ p(TT) = 0.4$

This is a legitimate model, these numbers are non-negative and add to one. Is the notation misleading? Does this correspond to tossing two coins? Let us see.

Consider the event that first letter is H; thus $A = \{HH, HT\}$.
Then $P(A) = 0.1 + 0.2 = 3/10$ Similarly $P(A^c) = 7/10$.
Also if B is the event that second letter is H, then

$$P(B \,|\, A) = 1/3. \quad P(B^c \,|\, A) = 2/3$$

$$P(B \,|\, A^c) = 3/7 \quad P(B^c \,|\, A^c) = 4/7$$

Here is an experiment: I have three coins: I, II, III
coin I has chance of Heads 3/10; Hence Tails has chance 7/10
coin II has chance of Heads 1/3; Hence Tails has chance 2/3
coin III has chance of Heads 3/7; Hence Tails has chance 4/7
Toss coin I ; If Heads up, toss coin II; if tails up toss coin III.
Then the outcomes are as above. Interestingly, probabilities for outcomes should also be as prescribed above! For example,

$$p(HH) = P(AB) = P(A)P(B \,|\, A) = (3/10)(1/3) = 0.1$$

Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$
$p(1) = p(2) = p(3) = 0.1 \quad p(4) = p(5) = 0.2 \quad p(6) = 0.3$

This corresponds to rolling an biased die once. What are the chances that even number turns up? Answer: $0.1 + 0.2 + 0.3 = 0.6$.

If A is the event $\{1, 5, 6\}$ then also $P(A) = 0.6$. Similarly, you can calculate probability of any event. If $B = \{2, 4, 6\}$ then you can calculate $P(A|B)$.

Sometimes we would be knowing $P(A|B)$ but we want to know $P(B|A)$. here is an example. Let us say a proportion a of people in Chennai have a certain viral fever. I am not feeling well and go to the doctor. The doctor gets blood test done. Even though the chances of my having the infection are a, we would like to know given the result of the test. No test is foolproof. both kinds of errors occur. It may give false negative: I have the infection but the test says NO with probability α. It may give false positive too: I do not have infection, but the test may say Yes with probability β. The problem now is the following:

A : test said YES $\quad B$: I have infection.

Want $P(B|A)$

What do I know?

$$P(B) = a; \qquad P(B^c) = 1 - a$$
$$P(A|B) = 1 - \alpha; \qquad P(A^c|B) = \alpha$$
$$P(A|B^c) = \beta; \qquad P(A^c|B^c) = 1 - \beta$$

Then our rules tell us

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(AB)}{P(AB) + P(AB^c)}$$

$$= \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)}$$

and all these quantities are known.

Polya urn scheme:

Consider an urn with 30 balls; 20 red and 10 green. here is the game which can be played forever.

Pick a ball at random, note its color and put back and add one ball of that color to the urn. Repeat...

We calculate some probabilities. What are the chances first draw is red?

$P(R_1) = 20/30$

What are the chances second draw is red? We do not know the result of the first draw and hence we do not know the composition of the urn before second draw and hence we can not calculate as a simple ratio. Even if we

knew result of first draw, such a ratio gives only conditional probability given your assumption on the result of the first draw. We proceed as follows

$$P(R_2) = P(R_2G_1) + P(R_2R_1) = P(G_1)P(R_2 \,|\, G_1) + P(R_1)P(R_2 \,|\, R_1)$$

$$= \frac{10}{30}\frac{20}{31} + \frac{20}{30}\frac{21}{31} = \frac{20}{30}$$

See how we split the event into disjoint sub-events by incorporating the lacking information about the first draw.
Thus chance of red ball remains same. In fact chance of red ball at any draw remains 20/30. We can prove by induction. Let the statement be

(S_n) : whatever be $r, g \geq 1$; $P(R_n) = \frac{r}{r+g}$ in Polya scheme starting with r red and g green balls.

S_1 and S_2 are proved above: change 20 and 10 to r and g.
Suppose S_n is proved. We shall prove for S_{n+1}. We need to calculate $P(R_{n+1})$. If you bring in the information about n-th draw, you still do not know the composition of the urn and will not be able to calculate and the best way is to reduce the problem to n-th draw in order to use the induction hypothesis.

$$P(R_{n+1}) = P(R_{n+1}R_1) + P(R_{n+1}G_1)$$

$$= P(R_1)P(R_{n+1} \,|\, R_1) + P(G_1)P(R_{n+1} \,|\, G_1)$$

Note that given R_1; after the first draw you have a Polya scheme starting with $r+1$ red and g green balls and you want the chances of red **now** at n-th draw which by induction hypothesis is $(r+1)/(r+g+1)$. Similarly given G_1 also you can calculate getting

$$P(R_{n+1}) = \frac{r}{r+g}\frac{r+1}{r+g+1} + \frac{g}{r+g}\frac{r}{r+g+1} = \frac{r}{r+g} \qquad \blacksquare$$

Note how we formulated the induction hypothesis, if you simply formulated for (20,10) data you would not have been able to use it after the first draw, for the inductive step.

Though the chances of red at any draw remains same, the number of balls are increasing after each draw, so which probabilities are changing? Answer: conditional probabilities.

$$P(R_1) = \frac{r}{(r+g)} \qquad P(R_2|R_1) = \frac{(r+1)}{(r+g+1)} > \frac{r}{(r+g)}$$

If you see a red ball, then chances of seeing red ball is increased. If the second ball is also red, then chances of red at third are further increased. This is where probability comes to life!

Our friend came from Kolkata. Suppose before starting, some one told him that there is flu epidemic in Chennai. Let us understand our friend's feelings during a short duration after he lands in Chennai. If he right away met a person with flu, he tends to think that there must really be much flu here. The more and more flu people he meets soon after reaching here, he assigns higher and higher chances for flu in Chennai. On the other hand if the people he met had no flu, he tends to think that there is not that much flu here. However, his experience has nothing to do with reality. There is a certain percentage who have flu and the chances of seeing a person with flu is just that fraction. [That is why I said in a short duration, you need not worry about dynamics of spread]. Thus interpreting red as person with flu, the model reflects precisely this phenomenon.

Let us make one more calculation.

$$P(R_2G_1) = P(G_1)P(R_2 \,|\, G_1) = \frac{10}{30}\frac{20}{31}$$

$$P(R_1G_2) = P(R_1)P(G_2 \,|\, R_1) = \frac{20}{30}\frac{10}{31}$$

Thus $P(R_1G_2) = P(R_2G_1)$. Since $P(G_2) = P(G_1)$ we see

$$P(R_1 \,|\, G_2) = P(R_2 \,|\, G_1)$$

In other words there is time symmetry. Given information: G at time 2; ask: R at time 1, or, give the same information G, but at time 1, and ask the same question R, but at time 2. You get the same answer. Only the role of time changed; given information is same and question asked is same. Here again probabilities come alive, because this is what we observe in practice.

Interpret red and green as genes and time in generations. Ask: what are the chances my father has a gene R given I have gene G. Or ask: what are the chances I have gene R given my father has gene G. Answer remains same under certain assumptions on the genetic structure of population. Note that information is same G, question is same R. In one case information is about me (second generation) and question is about my father (first generation). In the other case information is about my father (first generation) and question is about me (second generation) — a time reversal.

There are several interesting calculations that can be done with this urn model, however we shall not do. Before leaving this scheme, let me assure you how mathematics, like music, allows improvisations/embellishments.

Why 20, 10 balls? Yes, you can have r and g initial data.

Why add one ball, why not 7 balls of that color? Yes, you can; the same phenomenon remains. But what does this signify? If you see a red ball, you are now adding 7 red balls; so that the conditional probability of seeing red ball now increased from 20/30 to 27/37 which is higher, more importantly, much higher than the previous increase, namely 21/31. This is also what happens in practice. If our Kolkata friend is a very nervous person, then the moment he runs into one person with flu he might react 'Oh my God, my information is right, there is too much flu here in Chennai'; if he runs into two persons with flu, he might even repent coming here!

Why two colors?why not more colors? Yes, can develop.

Why ball of the color seen? why not add opposite color ball? More generally why not add 3 balls of the color seen and 1 ball of the opposite color? Yes, can be done.

Do you see how things can be improvised?

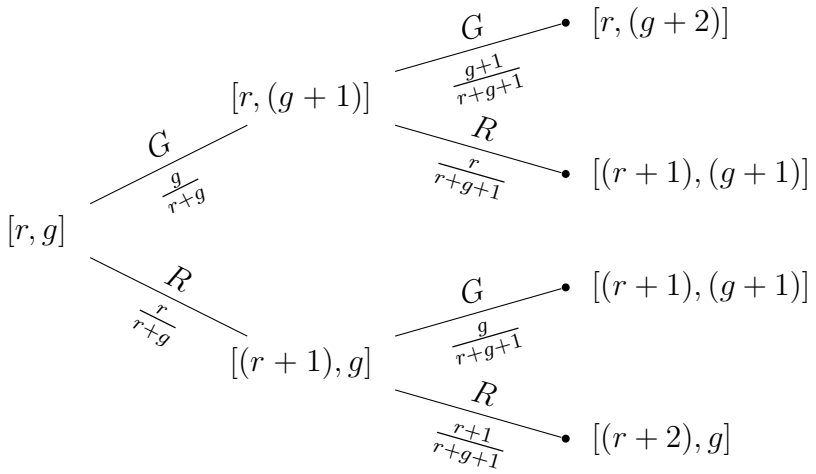
Indeed, the model of adding one ball of opposite color is also important and is called Friedman's urn model. Of course, the above phenomenon is no longer true, chances of red ball at draw n depends on n. Does this model also reflect something that we see? yes, this models 'safety campaign'.

Interpret red as accidents and green as safety measures. In general police are a relaxed lot and do nothing. When accidents increase, they suddenly become active and start implementing safety measures. Once you see more safety measures and accidents decline, then police become slack and accidents start increasing. Once they increase and you see more accidents, they once again wake up and implement safety measures.

Returning to Polya Urn scheme, we can pictorially represent the outcomes. We can draw a tree diagram depicting the composition of the urn at each stage and the outcomes and conditional probabilities. This is what we shall do below starting with general (r, g) urn.

If you read the letters along any particular path, you will get outcomes.

Probability of the event is simply product of the numbers on the branches along the path. On each branch is written conditional probability of that event given the past till then. At the end of each branch is written the composition of the urn to help you calculate probabilities for the next branch. Initially composition is r red and g green. After green draw urn has become $[r, g+1]$ etc.



We shall leave urn models and proceed to the next important idea. But before that let us collect some useful facts which we have seen and used.

Theorem: Let (Ω, p) be any probability space and B an event.

(1) $P(\emptyset) = 0$ and $P(\emptyset \,|\, B) = 0$
$P(\Omega) = 1$ and $P(\Omega \,|\, B) = 1$

(2) $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
$P(A_1 \cup A_2 \,|\, B) = P(A_1 \,|\, B) + P(A_2 \,|\, B) - P(A_1 \cap A_2 \,|\, B)$

(3) if $A_1, \ldots, A_n$ are disjoint events then
$P(\cup A_i) = \sum P(A_i)$
$P(\cup A_i \,|\, B) = \sum P(A_i \,|\, B)$

(4) $P(A^c) = 1 - P(A)$
$P(A^c \,|\, B) = 1 - P(A \,|\, B)$

(5) For any events $A_1, \ldots, A_n$

$$P(\cup A_i) = S_1 - S_2 + S_3 - \ldots$$

where

$$S_1 = \sum P(A_i), \quad S_2 = \sum_{i<j} P(A_i \cap A_j), \quad S_3 = \sum_{i<j<k} P(A_i \cap A_j \cap A_k), \ldots$$

$$P(\cup A_i \,|\, B) = S_1' - S_2' + S_3' - \ldots$$

where

$$S_1' = \sum P(A_i|B), \quad S_2' = \sum_{i<j} P(A_i \cap A_j|B), \quad S_3' = \sum_{i<j<k} P(A_i \cap A_j \cap A_k|B), \ldots$$

(6) (when Ω is infinite) For any sequence of disjoint events $(A_n, n \geq 1)$

$$P(\cup A_n) = \sum P(A_n), \quad P(\cup A_n \,|\, B) = \sum P(A_n \,|\, B)$$

(7) $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \,|\, A_1)P(A_3 \,|\, A_1 \cap A_2)$

Independence:

Example: Consider the usual deck of cards: It has four 'suits':
Clubs ♣; Diamonds ♢; Hearts ♡; and spades ♠
and in each suit 13 cards: $A, 2, 3, \ldots, 10, J, Q, K$.
I pick a card at random. Thus $|\Omega| = 52$. Let A be the event selected card is Ace (A).. Then $P(A) = 4/52$. I tell you that the selected card is spades and ask for probability of A now. Thus if B is the event: selected card is spades, then we want $P(A \,|\, B)$. Clearly it equals $1/13$. Thus

$$P(A \,|\, B) = P(A)$$

The information that the event B occurred did not alter/influence the probability of A. When this happens, that is, $P(A \,|\, B) = P(A)$; equivalently $(A \cap B) = P(A)P(B)$; it appears reasonable to say that the events are independent.

Example: $\Omega = \{HH, HT, TH, TT\}$
Probability of each outcome is $1/|\Omega| = 1/4$. If A is the event 'first letter is H' and B is the event 'second letter is H' then A, B are independent.

Let us consider the experiment: Roll a fair die two times, all outcomes equally likely. Consider the events: A, first throw is even; B, second throw is even; C, sum of the two throws is even. Clearly

$P(A) = P(B) = P(C) = \frac{1}{2}$;

$P(A \cap B) = P(B \cap C) = P(C \cap A) = \frac{1}{4}$.

Thus A, B are independent; B, C are independent; A, C are independent. Should we say that A, B, C are independent? If you know A. and B both occur then you know C must hold. In fact if you know about any two of these, then you know about the third one. For example if both of A, B occur or none of A, B occurs, then C occurs. If exactly one of A, B occurs, then C does not. Thus we should not really say that the three are independent. After all if you say certain things are independent, they have nothing do with each other; if you have any information on any of them it should not reveal about the remaining. So what fails in this example?

$P(A \mid B \cap C) \neq P(A)P(B \cap C)$, equivalently
$P(A \cap B \cap C) \neq P(A)P(B)P(C)$.

If we put this extra condition, we can show the intuitive feeling outlined above is correct and A, B, C are independent. More generally we make the following definition.

Definition: Events $A_1, A_2, \cdots, A_n$ in a probability space are independent if for any number of them probability of their intersection equals product of their probabilities. In symbols, the following is true: for every $k \leq n$ and every $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.
$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$

Does this definition reflect what we had in mind? We can show the following:

■ if for each i, the event B_i is either A_i or A_i^c, then
$P(B_1 \cap B_2 \cap \ldots B_n) = P(B_1)P(B_2) \cdots P(B_n)$ ■
This is done by induction on the number of complements that appear. This can then be used to show any information on some of these does not influence the others. In fact the above statement is equivalent to independence of the events $(A_i : 1 \leq i \leq n)$.

Given a probability space and events, we can use the above equations to check if they are independent events. Or we can use the requirement of independence to assign probabilities.

For example, if we say toss a fair coin twice independently, then it means, assign probabilities so that questions about first and second tosses are independent. For example chances of HH is product of chances that first toss is H multiplied by chances second toss is H. Hence $P(HH) = 1/4$. We are

repeating, independently, the experiment of tossing the coin. This leads to more general concept of repeating experiments independently.

Definition: Suppose we have experiments (Ω_i, p_i) *for* $1 \leq i \leq n$. *We define the product experiment* (Ω^*, p^*) *as follows.*

$\Omega^* = \prod_1^n \Omega_i = \{(\omega_1, \cdots, \omega_n) : \omega_i \in \Omega_i\}$

and

$p^*(\omega_1, \cdots, \omega_n) = \prod_1^n p_i(\omega_i)$

This space represents performing the n experiments, one after the other, independently. Consider $A_i \subset \Omega_i$, and define the events: A_i^* = the set of all outcomes in Ω^* such that i-th coordinate is in A_i. Note that the event A_i^* in the product depends only on i-th experiment, whether an outcome of Ω^* is in A_i^* or not depends only on i-th experiment. Then in the space (Ω^*, p^*) the events $A_1^*, \cdots, A_n^*$ are independent.

You must keep in mind that in defining p^* we multiplied the probabilities because we want events depending on different i to be independent. You must pause and think about matters.

If all the experiments are same, then this is referred to as independent repetitions of the one experiment. For instance, let us consider the experiment of tossing a coin once. $\Omega = \{H, T\}$ and $p(H) = \theta$ and $P(T) = 1 - \theta$ where $0 < \theta < 1$. Thus chance of heads is θ. The experiment of tossing the coin n times independently means the product experiment where the sample space is Ω^* and probabilities are given by

$p(\epsilon_1, \cdots, \epsilon_n) = \theta^i(1-\theta)^{n-i}$ where i is the number of H in the outcome $(\epsilon_1, \cdots, \epsilon_n)$.

Let us calculate probabilities of some interesting events. A is the set of all outcomes having exactly k heads. You see that each such outcome has probability $\theta^k(1-\theta)^{n-k}$ and there are $\binom{n}{k}$ many such outcomes. Thus

$$P(A) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Here we have tossed a coin n-times and counted number of heads. This is a recurring phenomenon, we perform an experiment and make a measurement. For instance, if you want to decide whether to accept a lot of bolts supplied to your company, you need to know whether they confirm to the specifications.

You can not test all items. You take a sample of 100 items one by one, test and count the number of defective found. Base your decision on this number. it is unimportant whether the first is defective or the tenth etc. If you are trying to estimate chance of heads for a given coin, you can toss it, say n times and if you find k heads, then k/n is a good estimate of the chance of heads. It is unimportant where exactly the k heads appeared among the n tosses.
These measurements are called random variables.

Definition: Let (Ω, p) be a probability space. A random variable on the space is a real valued function. The distribution of the random variable is a list or table giving all values of the random variable and the corresponding probabilities; that is against the value a write the probability $P(X = a)$, that is $P\{\omega : X(\omega) = a\}$.

If random variable is simply a function on Ω why do you need a new term, could have called it a function. Well, by using the word random variable we draw your attention to the fact that there is a probability on Ω.

For instance, in the above example of tossing a coin n times (when nothing is said, the tosses are independent), if X is the number of heads then distribution of X is

values	0	$\dots$	k	$\dots$	n
probabilities	$(1-\theta)^n$	$\dots$	$\binom{n}{k}\theta^k(1-\theta)^{n-k}$	$\dots$	θ^n

This distribution is called Binomial distribution; denoted $B(n, \theta)$.

We have a constitutional officer (VP) who 'feels' constitution is unimportant; another constitutional officer (G) who 'feels' no rules apply to him; a Nobel Laureate (Economics) who 'feels' India had no math tradition and we got our math from Greece, $\cdots$. When you do problems do not stop with the feelings like 'easy', 'difficult', 'can do' etc. Feelings can fool you. Take pen and paper, think and work out problems, write solutions. You can do it.

Expectation and Variance:

We play a series of games. I toss a coin. Heads up I give you one Rupee, Tails up you give me one Rupee. Thus if X is your profit, then $X = +1$ if Heads up and $X = -1$ if Tails up.
If chance of Heads is 2/3 and Tails is 1/3; will you play. Yes you will. The argument is that if we play 30 games you are likely to win approximately 20 games and loose 10 so that you gain 10 Rs; or on the average 1/3 Rs per game. Of course instead of 30 games and so on we can calculate $(+1)(2/3) + (-1)(1/3) = 1/3$ as expected profit per game.
If chance of Heads is 1/3 and Tails is 2/3; will you play. No, you will. not. Similar argument applies: You expect to win $(+1)(1/3) + (-1)(2/3) = -1/3$ per game.
What if the coin is fair. You will think, looks fair, neither you win, nor loose on the average.
Let us consider fair coin game but now $X = +1000$ if Heads up and $X = -1000$ if Tails up. Clearly this is also fair, expected gain is zero. Will you play? probability not, because it has 'high risks': you may loose the first three games (though you may win later games). In other words the random variable has X has larger spread. The average value and spread is made precise now.

Definition: If X is a random variable taking values $\{x_i : i \geq 1\}$ with respective probabilities $\{p_i, i \geq 1\}$; then we define the Expected value/Average value/Mean value by the formul;a

$$E(X) = \sum x_i p_i$$

provide $\sum |x_i p_i|$ is finite. If this last sum is not finite, we shall not define $E(X)$. We define the variance of X by

$$Var(X) = \sum x_i^2 p_i - (\sum x_i p_i)^2 = \sum x_i^2 p_i - (E(X))^2$$

when the sums are finite.

Example 1: Toss a coin once, chance of Heads in a toss is p. Let X be the number of Heads.
Sample space: $\{H, T\}$.
Probabilities: $P(H) = p; P(T) = 1 - p = q$.

Values of X are $1, 0$ with probabilities p, q respectively.
Expected value: $E(X) = p$.
Variance: $V(X) = p^2 - p = pq$.
This is called Bernoulli random variable. Sometimes the random variable X taking values 1 and 0 is called Bernoulli random variable. Sometimes *any* random variable that takes two values is called Bernoulli variable.

Example: Same coin above. Toss n times. Let X be the number of Heads.
Sample space: All sequences of H, T of length n.
Probability: $P(\omega) = p^k q^{n-k}$ if k is the number of Heads in ω.
random variable: $X(\omega) =$ number of H in ω.
Distribution: values $\{0, 1, \ldots, k, \ldots, n\}$ and respective probabilities $\{\binom{n}{k} p^k q^{n-k} : 0 \leq k \leq n\}$.
Expected value: $E(X) = np$.
Variance: $V(X) = npq$.
Here is the calculation:

$$E(X) = \sum_0^n k \binom{n}{k} p^k q^{n-k} = np \sum_1^n \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

$$= np(p+q)^{n-1} = np.$$

and

$$\sum_0^n k^2 \binom{n}{k} p^k q^{n-k} = \sum_0^n k(k-1) \binom{n}{k} p^k q^{n-k} + \sum_0^n k \binom{n}{k} p^k q^{n-k}$$

$$= n(n-1)p^2(p+q)^{n-2} + np = n^2p^2 - np^2 + np$$

so that

$$var(X) = n^2p^2 - np^2 + np - n^2p^2 = np(1-p).$$

This distribution is called Binomial distribution. and any random variable X having this distribution is called Binomial random variable, denoted $X \sim B(n, p)$.

Example: Same coin. Toss till you get one Head and X is the number of Tails obtained.
Sample space: $\{H, TH, TTH, TTTH, \ldots\}$
Probabilities: $P(T^k H) = q^k p$.
Random variable: $X(T^k H) = k$.
Distribution: value k with probability $q^k p$ for $k = 0, 1, 2 \ldots$.
Expectation: $E(X) = q/p$.

Variance: $V(X) = ?$

$$E(X) = \sum_0^\infty kq^k p = p\sum_1^\infty kq^k$$

But if $q + 2q^2 + 3q^3 + \cdots = S$ then

$$S - qS = \frac{q}{1-q} = \frac{q}{p}; \qquad S = \frac{q}{p^2}$$

so that

$$E(X) = q/p.$$

This distribution is called Geometric distribution and any random variable with this distribution is called a Geometric random variable: $X \sim G(p)$. Sometimes the random variable Y = number of tosses is called Geometric variable. Clearly $Y = X + 1$. **For us** X **is** $G(p)$.

Example: Example: Same coin. Toss till you get r Heads and X is the number of Tails obtained.
Sample space: All finite sequences of H, T with last letter H and before that exactly $(r-1)$ letters are H. Probabilities: $P(\omega) = q^k p^r$ if ω is of length $(r+k)$.
Random variable: $X(\omega) = k$ if ω is of length $(r+k)$.
Distribution: value k with probability $\binom{r+k-1}{r-1} q^k p^r$ for $k = 0, 1, 2 \ldots$.
Expectation: $E(X) = rq/p$.
Variance: $V(X) = ?$
This distribution is called negative Binomial distribution and any random variable with this distribution is called a negative Binomial random variable: $X \sim NB(r, p)$. Sometimes the random variable Y = number of tosses is called Geometric variable. Clearly $Y = X + r$. **For us** X **is** $NB(r, p)$.
The probabilities are terms appearing in the Binomial expansion with negative index.

Suppose the coin has a very very small chance of heads, say 0.0001 and we toss the coin 10000 times. What are the chances of 5 heads? of course you can give binomial probability which is difficult to compute. When p is very small and n is very large and $np = \lambda > 0$ then chances of value k approximately equals $e^{-\lambda}\lambda^k/k!$. Here is the precise result:

Theorem Let $X_n \sim B(n, p_n)$. Suppose $np_n \to \lambda > 0$ as $n \to \infty$. Then for each k

$$P(X_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}$$

Proof is very simple: Let us denote $\lambda_n = np_n$ so that $\lambda_n \to \lambda$. Also note that hypothesis implies that $p_n \to 0$.

$$P(X_n = 0) = (1 - p_n)^n = (1 - \frac{\lambda_n}{n})^n \to e^{-\lambda}$$

$$P(X_n = 1) = np_n(1 - p_n)^{n-1} = \lambda_n \, (1 - \frac{\lambda_n}{n})^n \, (1 - p_n)^{-1}$$
$$\to \lambda e^{-\lambda}$$

You can prove by induction on k that for ecery k the result holds.

Example: Let $\lambda > 0$. The distribution:
Values: k. for $k = 0, 1, 2, \ldots$
probabilities: $e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.
is called Poisson distribution and if X has this distribution then we say X is a Poisson random variable; $X \sim P(\lambda)$
Unlike the previous examples, we have not given an experiment that produces this random variable. But in practice this is supposed to fit well in several phenomena. The number of accidents during a month at a busy intersection (because, a large number of cars pass through and each car can potentially cause an accident but with a very small chance); the number of particles emitted by a radioactive material during an hour (because there are very large number of atoms and each can potentially fly but has a very small chance — all other atoms are not allowing it) and so on. Thus this is very useful model for several phenomena.

We shall study properties of expectation. why did we define expectation only when the series is absolutely convergent? Because, there is no specific order in which the values of the random variable are to be listed. I may write in one order and you may list the values in a different order. When we multiply values with probabilities and add we both should get the same answer. Otherwise, there is a serious problem. You know that if a series is absolutely convergent then the series can be added in any manner. Because, we put the condition of absolute convergence before making the definition of expectation, the order in which the values of the variable are listed does not matter.
But if the series is convergent but not absolutely convergent, then you can rearrange to get any pre-determined sum!
Note that any sum where all terms are non-negative can be rearranged and added in any manner and we get the same answer. For example, suppose

you are adding

$$\sum_{1}^{\infty} y_n$$

where each $y_n \geq 0$. You can partition natural numbers into disjoint sets

$$A_k : k = 1, 2, 3, \cdots$$

You can add those y_n such that $n \in A_k$ only and obtain the sum. Let us say this subtotal is z_k. Now add these sub totals z_k to get z. No matter how you make the partition $\{A_k\}$ and calculate you will get the same final answer. Thus by one method if you get 25 (or ∞) then you will get 25 (or ∞) by any other method. You must keep in mind that terms (y_n) are non-negative.

If you think that it is your birth right to add 'as you like', then add the following numbers: (1) Do row totals and add them; (2) Do column totals and add them.

$$\begin{array}{cccccccc}
1 & -1 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\!\vdots\!\vdots
\end{array}$$

In our course we come across only mice random variables and you need not worry about absolute convergence.

Theorem: Expectation properties

Let (Ω, p) be a probability space.

1. A random variable X has expectation iff $\sum |X(\omega)|p(\omega) < \infty$. In that case $E(X) = \sum X(\omega)p(\omega)$.

2. If $X_1, \cdots, X_n$ are rvs then so is $\sum X_i$. Further if each X_i has expectation, then so is $\sum X_i$ and $E(\sum X_i) = \sum E(X_i)$

3. if X is a rv, then so is $23X$. If X has expectation, then so has $23X$ and $E(23X) = 23E(X)$.

4. Let X be a rv and $f : R \to R$ be any function. Then $Y = f(X)$ defined by $Y(\omega) = f(X(\omega))$ is a rv. Assuming $E(X)$ and $E(Y)$ are defined we have $E(Y) = \sum x_i f(x_i)$. ∎

Remember that a series of positive numbers can always be added; at the worst the sum may be infinity.

Proof of 1: suppose X takes values $\{x_k : k \geq 1\}$ with probabilities $\{c_k : k \geq 1\}$ respectively. Let us partition the sample space Ω as follows

$$A_k = \{\omega \in \Omega : X(\omega) = x_k\}; \qquad k = 1, 2, 3, \cdots$$

Note that

$$c_k = P(X = x_k) = P\{\omega : X(\omega) = x_k\} = \sum_{\omega \in A_k} p(\omega).$$

Also

$$\sum_{\omega \in A_k} |X(\omega)| p(\omega) = |x_k| \sum_{\omega \in A_k} p(\omega) = |x_k| c_k.$$

From the observation made about absolute convergence,

$$\sum |X(\omega)| p(\omega) = \sum_k |x_k| c_k$$

Hence, one side is finite iff the other side is finite. Thus a necessary and sufficient condition for expectation to be defined is that the left side above be finite.

But if the expectation is defined, that is, left side is finite, then the series $\sum X(\omega)p(\omega)$ is absolutely convergent. Now use the observation about absolutely convergent series to see you can make subtotals over each A_k and then add them. Thus you see

$$\sum X(\omega) p(\omega) = \sum_k x_k c_k = E(X).$$

Proof of 2. Let us do for two rvs.

Let X and Y be random variables and $Z = X + Y$, that is $Z(\omega) = X(\omega) + Y(\omega)$. Note that if

$$\sum |X(\omega)| p(\omega) \quad \text{and} \quad \sum |Y(\omega)| p(\omega)$$

are finite then using $|Z(\omega)| \leq |X(\omega)| + |Y(\omega)|$ you see

$$\sum |Z(\omega)| p(\omega)$$

is also finite. Thus if X and Y have expectations defined then expectation of $Z = X + Y$ is also defined. Clearly then

$$\sum Z(\omega)p(\omega) = \sum X(\omega)p(\omega) + \sum Y(\omega)p(\omega)$$

or by using part 1 above,

$$E(X + Y) = E(X) + E(Y)$$

3. Similarly you can show $E(23X) = 23E(X)$.

4. Note that this is not a tautology. For example the formula is not definition because there is no mention of distribution of Y. It is not from part 1, because there is no ω. It is a formula to calculate $E(Y)$ by using the distribution of X: no need to look at the sample space, no need to calculate the distribution of Y either.
As we go along we shall not make too much fuss about existence etc, we assume that the expectations we are talking about exist and proceed with calculations.

Proof of the formula is simple. By part 1,

$$E(Y) = \sum_{\omega} Y(\omega)p(\omega) = \sum_{i} \sum_{\omega:X(\omega)=x_i} Y(\omega)p(\omega) = \sum_{i} f(x_i)p_i \quad \blacksquare$$

You must appreciate part 1, which allowed you to prove linearity of expectation. After all if you know distribution of X and Y then there is **no** way of calculating distribution of $X + Y$; you need more information. For example, If I toss a fair coin and X is the number of heads obtained and Y is the number of Tails obtained then $X \sim B(2, 1/2)$ and $Y \sim B(2, 1/2)$. Also $P(X + Y = 2) = 1$. On the other hand, consider tossing a fair coin 4 times and X is the number of heads in the first two tosses and Y is the number of Heads in the last two tosses. Even now $X \sim B(2, 1/2)$ and $Y \sim B(2, 1/2)$. However now $X + Y \sim B(2, 1/2)$.
Thus you can not calculate the distribution of $X+Y$ from those of X and Y. S0 if you depended on the definition of expectation via distribution ($\sum x_k p_k$) then there is no way to conclude the above.

Suppose we have a rv X with mean μ. Then

$$V(X) = E[(X - \mu)^2]$$

We simply take $f(a) = (a - \mu)^2$ in the above theorem:

$$E[(X-\mu)^2] = \sum (x_i - \mu)^2 p_i = \sum x_i^2 p_i + \mu^2 \sum p_i - 2\sum \mu x_i p_i$$

$$= \sum x_i^2 p_i + \mu^2 - 2\mu^2 = \sum x_i^2 p_i - \mu^2$$

$$= \sum x_i^2 p_i - (\sum x_i p_i)^2. = .V(X)$$

We can also say by taking the function $f(a) = a^2$ and simplifying that

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

Expected number of matches:

The fact sum of expectations equals expectation of sum is a very very useful result. Let us consider the matching problem with 52 cards and 52 envelopes. There are (52)! outcomes. Let X be the number of matches. that is, if you take an outcome ω then $X(\omega)$ is the number of matches according to the placement ω. This variable can take integer values $\{k : 0 \leq k \leq 52\}$. It is not easy, though we did, to calculate $P(X = k)$. It is not easy to calculate $E(X)$ using the definition. We shall cleverly express this as sum of variables for each of which expectation can be calculated in a painless manner.

For $1 \leq i \leq 52$, let X_i be the following random variable on our space: $X_i(\omega) = 1$ or zero according as there is match at place i or not in the arrangement ω. Thus

$$X(\omega) = \sum_{1}^{52} X_i(\omega)$$

Note that X_i indicates whether there is match at place i or not and is hence called indicator random variable. The beauty is that it takes only two values zero and one. Thus for any i,

$$E(X_i) = P(X_i = 1) = \frac{51!}{52!} = \frac{1}{52}$$

and

$$E(X) = 1.$$

A hopeless situation is rescued by the linearity of expectation. Also note that even if you have 1000 cards instead of 52, the expected number of matches still equals one.

Chebyshev's inequality:
Let $a > 0$.
(i) For any non-negative rv X we have

$$P(X \geq a) \leq \frac{E(X)}{a}$$

(ii) For a random variable X

$$P(|X| \geq a) \leq E(|X|)/a; \qquad P(|X| \geq a) \leq E(|X|^2)/a^2.$$

(iii) X be a rv with mean μ and variance σ^2.

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

It is assumed that the required Expectations are finite.

Proof of Chebyshev is simple.

(i) Let $a_1, a_2, \cdots$ be all the values of X with respective probabilities (p_i).

$$P(X \geq a) = \sum_{i:a_i \geq a} p_i \leq \sum_{i:a_i \geq a} \frac{a_i}{a} p_i \leq \frac{1}{a} \sum a_i p_i = \frac{1}{a} E(X)$$

(ii) follows by applying (i) to $|X|$ and $|X|^2$. Note that the events $(|X| \geq 1)$ and $(|X|^2 \geq a^2)$ are same.
(iii) follows by applying (i) to $(X - \mu)^2$. Note that $E[(X - \mu)^2] = \sigma^2$. ∎

height of bus:

Here is a simple special case. suppose $a = 3\sigma$ then the above inequality says

$$P\{\omega : \mu - 3\sigma < X(\omega) < \mu + 3\sigma\} \geq 1 - \frac{1}{9} = \frac{8}{9}$$

$$\sim 90\%.$$

For example if average height of the adult (?) population of Chennai is five feet with standard deviation equal to 1/6 (two inches), then 90% of adults have height at most five and half feet. In particular, if you are asked to design height of a bus so that ninety percent of adults can travel comfortably, then five and half feet is a good suggestion.

Weak Law of Large Numbers for coin tossing:

We shall now ask whether the mathematics reflects the intuitive ideas with which we started. For example, when we say p $(0 < p < 1)$ is the chance of heads for the coin, our intuitive feeling is that in a large number of tosses a proportion p will be heads. Carefully note, I am not saying expected number of proportion of Heads is p; which is true and we calculated above.

We defined mathematical models, concept of independence and so on. Can we show that actually a proportion p will be heads or proportion of heads is getting close to p? We immediately realize that the question is not well formulated. the proportion of heads is a random variable; there is a non-zero chance that all tosses may result in heads. In fact the proportion of heads takes all values $0/n, 1/n, 2/n, \cdots, n/n$ each with positive probability.

Thus a better question is: can you show that when the number of tosses is large then the proportion of heads is close to p with very very high probability? Yes.

Theorem WLLN:
Let Y_n be the proportion of heads in n tosses of a coin whose chance of heads in a toss equals p $(0 < p < 1)$. Given any $\epsilon > 0$,

$$\lim_n P(|Y_n - p| \geq \epsilon) = 0.$$

More precisely,

$$P(|Y_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \to 0 \qquad \blacksquare$$

Observe that ϵ being fixed, the first conclusion follows from the second.

What does this mean? You give me any error limit for p, say $\epsilon > 0$ and also give me an $\eta > 0$, then I can show you an N such that if you toss n times $(n > N)$ then

$$P(p - \epsilon < Y_n < p + \epsilon) \geq 1 - \eta$$

You give me any error limit, say, 0.01. Demand that observed proportion should be in $(p - 0.01, p + 0.01)$ with very high probability, say with probability at least 0.999. Yes, I can prescribe to you an N so that if you toss at least so many times then the observed proportion is within the limits you prescribed with probability at least 0.999. Think about it.
This is called Weak Law of Large Numbers for coin tossing. This is a satisfactory answer confirming that maths is on right track. Since we are considering

coin tossing now, this theorem is for coin tossing. Since this says something about what happens in large number of tosses, this is law of large numbers. This is 'weak' because there is a better theorem 'strong' law. Strong Law answers the following question: Ok, you are saying that in very very large number of tosses the proportion is close to p; can you show that if I really toss 'infinite number of times' then the proportion *actually equals* p. It may look like a meaningless question since we can never toss infinitely many times; it is not silly.

{such questions are not meaningless. Michelson-Morley experiments essentially showed that no matter how you measure speed of light you get the same answer. Then Einstein thought: Aha, suppose I sit on a photon and try to measure the speed of another photon travelling parallel to me. What do I get. since their speeds are same – whatever it may be – then the other one should look stationary to me, speed zero, right? Where is the catch? These are called 'thought experiments' and are an essential part of any thought process.}

There is another important byproduct of this. In reality, no one tells us the chance of heads. What this says is the following: If you toss it a large number of times and take the observed proportion of heads, then that is a good estimate of the unknown p. The estimate can be made close to the actual value p with very high probability. Carefully understand this.

You might wonder, why would any one play with coin? As I said earlier, the same happens whenever there are two alternatives. For example let p be the proportion of people supporting a political party and $(1-p)$ the proportion that does not support. The same philosophy above helps you to estimate p.

Proof is simple. If X_n is the number of Heads in n tosses, then we know $X_n \sim B(n,p)$ variable, so that its mean value is $\mu = np$ and variance $\sigma^2 = npq$. Since $Y_n = X_n/n$ we see, by Chebyshv

$$P(|Y_n - p| \geq \epsilon) = P(|X_n - np| \geq n\epsilon) \leq \frac{npq}{n^2\epsilon^2} \to 0$$

Weierstrass:

We can use Chebyshev to obtain a theorem of Weierstrass. It says that every continuous function on a closed bounded interval is very close to a polynomial. If you tell how close you want, I can give a polynomial which is

so close. Here is the precise statement.

Weierstrass approximation Theorem:
Let f be a continuous function on $[0,1]$ and $\epsilon > 0$. We can find a polynomial P (in one variable x) such that $|f(x) - P(x)| < \epsilon$ **for every** $x \in [0,1]$.

Given the function f, let us define some polynomials, one for each $n \geq 1$.

$$P_n(x) = \sum_{k=0}^{n} f(\frac{k}{n}) \binom{n}{k} x^k (1-x)^{n-k}.$$

These are called Bernstein polynomials associated with f and this proof is due to Bernstein. Note that even though you see f in the above definition; we only used the values of f at certain points.

Observe whatever x we take

$$f(x) - P(x) = \sum_{k} \left[f(x) - f(\frac{k}{n}) \right] \binom{n}{k} x^k (1-x)^{n-k} \qquad (*)$$

We used binomial probabilities add to one. Note that if k is such that k/n is close to x then the first term $\left[f(x) - f(\frac{k}{n}) \right]$ is small. If x is far from k/n, that is, k is far from nx, then Chebyshev tells the second term $\binom{n}{k} x^k (1-x)^{n-k}$ is small — these are binomial $B(n,x)$ probabilities far from its mean. Here is the execution.

We are given $\epsilon > 0$. Pick, using uniform continuity of f, a number $\delta > 0$ so that

$$|u - v| < \delta \Rightarrow |f(u) - f(v)| < \epsilon/2.$$

Also, using the fact that a continuous function on $[0,1]$ is bounded fix a number $M > 0$ so that $|f(x)| < M$ for all $x \in [0,1]$.
The fix a N such that

$$\frac{1}{\delta^2 N} < \epsilon; \qquad i.e. \qquad N > \frac{1}{\epsilon \delta^2}$$

In what follows $n > N$. How did we know such choices are to be made? Because of the rough calculations as explained in class. From $(*)$

$$|f(x) - P(x)| \leq \sum_{k} \left| f(x) - f(\frac{k}{n}) \right| \binom{n}{k} x^k (1-x)^{n-k}$$

(used $|\sum a_i| \leq \sum |a_i|$)

$$\leq \sum_{k:|\frac{k}{n}-x|<\delta} [***] + \sum_{k:|\frac{k}{n}-x|\geq\delta} [***]$$

$$\leq \frac{\epsilon}{2} \sum_{k:|\frac{k}{n}-x|<\delta} [\text{binomial prob.}] \quad + \quad 2M \sum_{k:|k-nx|\geq n\delta} [\text{binomial prob}]$$

(Used choice of δ and M) First sum, being addition of some binomial probabilities, is at most one. Second sum is nothing but $P(|Z - nx| \geq \delta)$ for a $B(n, x)$ variable and is hence at most

$$\frac{nx(1-x)}{n^2\delta^2} \leq \frac{1}{4\delta^2 n}$$

(used that $x(1-x) \leq 1/4$ for each $x \in [0, 1]$)

$$|f(x) - P(x)| \leq \frac{\epsilon}{2} + 2M\frac{1}{4\delta^2 n} \leq \epsilon$$

(choice of N and $n > N$).
This completes the proof. Note that our choices of N, δ. and M have nothing to do with x and so the above holds for all $x \in [0, 1]$.

Graphs with large χ:

Most of the applications that we saw involve some calculations, not too many. If you are clever and use the results you know carefully; you can achieve more complicated things. I want to show you one such calculation of Erdos. You should get ready and pull your energies to climb a peak.

Recall that a graph is a pair $G = (V, E)$ where V is a finite set of points, called vertices. E is a collection of unordered pairs of distinct vertices. Pictorially you can imagine V to be a set of points; join a pair of points if that pair is in E — join by a line (or curve), called edge.

The pairs are unordered and hence there is no direction for the edge (Remember Tournment, it is ordered). These are called undirected graphs. Also the edge has distinct vertices, thus a vertex is not joined to itself by a curve. Thus our graphs have no loops. since we do not have two edges joining any pair of vertices, one says, there are no multiple edges. thus we are considering undirected graphs without loops and without multiple edges.

For a graph G its chromatic number is the smallest number of colours needed to colour the vertices so that no edge gets the same colour for its two end vertices. This is denoted by $\chi(G)$. Thus if $\chi(G) = 6$ you can partition the set of vertices V into six disjoint sets so that for any edge its two vertices are in different sets; equivalently, if you take one of these six sets and take two points from it, then they are not joined.

Let $k \geq 3$. A sequence of vertices $(v_1, v_2, \cdots, v_k)$ is called a cycle if all the pairs $(v_1, v_2); (v_2, v_3); \cdots, (v_{k-1}, v_k); (v_k, v_1)$ are edges. It is called a cycle of length k. Note that v_k, v_1 are joined. Also note $k \geq 3$. Thus edges are not cycles.

Graphs with few edges can be coloured with few colours. If the chromatic number is very large then there must be lots of connections in the graph and some believed that the graph must then possess small cycles. However this is not the case. there is no relations between these two.

Theorem Erdös:
Let $k \geq 1$ and $l \geq 3$ be integers. then it is possible to make a graph whose chromatic number is at least k and has no cycles of length smaller than l.

This is one of the first results that led to the rich theory of Random Graphs or Erdös-Renyi graphs. In a sense this is the beginning of random graph theory.

Chromatic number $\chi(g)$ is very difficult to handle. We define the notion of independent set. For a graph g, a set S of its vertices is called independent set if no two vertices in S are joined. By $\alpha(g)$ we denote the cardinality of largest possible independent set. More precisely, if $g = (V, E)$ is the graph, then

$$\alpha(g) = \max\{|S| : S \subset V; S \text{ independent}\}$$

Since we are considering graphs with finite number of vertices, the above quantity is well defined. Before we proceed we make a simple observation to convince you that information about α tells you some thing about χ.

We claim that for any graph $g = (V, E)$

$$|V| \leq \alpha(g)\chi(g); \quad i.e., \quad \chi(g) \geq |V|/\alpha(g) \quad (\bigstar)$$

Indeed, if $\chi(g) = c$ then colour the vertices with c colours and put V_i to be the set of vertices receiving colour i for $1 \leq i \leq c$. Thus we have a partition

of V into c sets, in particular $|V| = \sum |V_i|$. Each V_i is an independent set because no two vertices of the same colour are joined. Thus $|V_i| \leq \alpha(g)$ for each i. adding over i we get $|V| \leq c\,\alpha(g)$ as stated.
The plan of the proof is the following.

For $n \geq 4$, consider the collection of all graphs on n the vertices $\{1, 2, \cdots, n\}$. Let G_n be the bag containing these graphs. Line up all these bags.

$$G_4; \quad G_5, \quad \cdots, \cdots, G_n, \cdots\cdots$$

(♣). We show that after some stage every bag contains plenty of graphs without small cycles.
(♢). We show that after some stage every bag contains plenty of graphs which have only small independent sets. Remember that (★) implies these graphs have large χ value.
(♡). We make sure that both the above hold for some graphs.
(♠). Take one such graph g as above. Kill all unwanted cycles by removing one vertex from each such cycle, you will be removing only few vertices so that the resulting graph g^* is still large and see its chromatic number. Since independent sets will still be small, hopefully chromatic number will be large.

How do we plan to estimate above quantities? — by using probability. On each G_n we put a probability and use the inequalities we are familiar with – make each edge with some probability. Let us consider G_n. Suppose someone gives us a number $0 < p < 1$. Here is a probability on the set G_n: choose each edge, independent of others, with probability p.
Equivalently take any pair of vertices, toss a coin whose chance of heads is p, if heads join these two vertices, if tails do not join. do this for each of the $n(n-1)/2$ pairs.
Equivalently, take any graph $g \in G_n$. If it has a edges then probability of this outcome g is

$$\text{Prob}(g) = p^a q^b; \qquad b = \frac{n(n-1)}{2} - a.$$

equipped with this probability, this bag is denoted $(G_n; p)$, Erdös-Renyi random graph model (with parameter p). But remember unless I tell you what p you should take, you have no model. To execute the plan mentioned, fix a number θ with $0 < \theta < 1/l$. In other words fix any number strictly between zero and $1/l$ and name it θ. this number will not change till the proof ends. Given $n \geq 4$ we take $p = n^{\theta - 1}$. Thus we have a probability space (G_n, p) with this p as mentioned above. Strictly speaking we should have named

it as (G_n, p_n). No need to get confused, but as long as you remember that p depends on n, we need not tax our notation with suffixes. Whenever we calculate probabilities of events in G_n we use this probability.

for any graph g, let $X(g)$ = number of cycles of length at most l in the graph g. These are unwanted cycles. Observe that X can be regarded as a random variable on each G_n. Strictly speaking we should restrict definition of X to G_n, name it X_n and regard this as a random variable on G_n. But no need to tax our brains as long as we understand.

The first step is made precise by showing the following: There is an n_0 such that for $n > n_0$ we have

$$\text{on } G_n; \qquad P(X \geq n/2) < 1/2. \qquad (\clubsuit)$$

The second step is made precise as follows. define $x_n = \lfloor \frac{3}{p_n} \log(2n) \rfloor$. We show there is an integer n_1 such that if $n > n_1$, then

$$\text{on } G_n \qquad P(\alpha \geq x_n) < 1/2. \qquad (\diamondsuit)$$

Third step is executed as follows. Since the sum of the above two probabilities is less than one,

$$n > \max(n_0, n_1) \longrightarrow \exists g \in G_n, \quad X(g) < n/2, \quad \alpha(g) < x_n. \qquad (\heartsuit)$$

The final step is executed as follows: Note that g has n number of vertices; it has at most $n/2$ many unwanted cycles (cycles of length at most l). We shall now destroy these unwanted cycles. Take each such cycle and remove one vertex that appears in that cycle. We have thus removed at most $n/2$ many vertices. Let g^* be the graph g restricted to these vertices. it has the undeleted vertices as the vertex set and any pair here is joined iff it was joined in g. This graph g^* has no unwanted cycles – Then it would have been a cycle earlier too and how come no vertex in this cycle is removed?

Moreover, an independent set here is independent in g as well. Thus

number of vertices in g^* is at least $n/2$;
g^* has no unwanted cycles;
$\alpha(g^*) \leq \alpha(g) \leq x_n \leq \frac{3}{p_n} \log(2n) = 3n^{1-\theta} \log(2n)$.

Let us now see $\chi(g^*)$. By $(\bigstar)$

$$\chi(g^*) \geq \frac{n/2}{x_n} \geq \frac{n}{6n^{1-\theta} \log(2n)} = \frac{n^\theta}{6 \log(2n)}$$

Remember this is true for any $n > \max(n_0, n_1)$. Since the last quantity in the above display increases to infinity, we can choose such a large n and $g \in G_n$ so that that $\chi(g^*) > k$.
This completes the proof. Need to execute the first two steps as made precise above.

Let us understand X in simpler terms. If $N_i(g)$ is the number of cycles of length i in g, then we have

$$X(g) = \sum_{i=3}^{l} N_i(g).$$

In turn, N_i can be expressed in simpler terms as follows. Take a sequence $s = (v_1, \cdots, v_i)$ of i vertices. Define $I_s(g)$ to be one or zero according as $(v_1, \cdots, v_i)$ is a cycle in g or not. Then we have

$$N_i(g) \leq \sum_{s:|s|=i} I_s(g)$$

Here the sum is over all sequences s of length $|s| = i$ consisting of distinct vertices. The reason for inequality is that on right side one i-cycle is counted several times. For example $(v_1, v_2, \cdots, v_i)$ is the same cycle as $(v_2, \cdots, v_n, v_1)$.

$$E(I_s) = P(I_s = 1) = p_n^i$$

joining each of the pairs (v_j, v_{j+1}), including (v_i, v_1) is done with probability p_n and independently. Also the number of summands is less than n^i. Hence

$$E(N_i) \leq n^i p_n^i = n^i n^{\theta i - i} = n^{\theta i}.$$

Finally,

$$E(X) \leq \sum_{3}^{l} n^{\theta i} \leq l n^{\theta l}.$$

In particular, by Chebyshev, on G_n,

$$P(X \geq n/2) \leq \frac{E(X)}{n/2} \leq 2l n^{\theta l - 1} \to 0$$

because $\theta l - 1 < 0$. Thus there is an n_0 such that

$$n \geq n_0 \quad \Rightarrow \quad P(X \geq n/2) < 1/2. \qquad (\clubsuit)$$

Now let us understand α. *Keep in mind we fix n and discuss G_n.* Thus α is random variable on G_n. Let us take an integer $x > 1$. We shall suggest,

later, what x to take. Note that $\alpha(g) \geq x$ iff there is an independent set of size x in g.

$$\{g : \alpha(g) \geq x\} = \bigcup_{S:|S|=x} A_S; \quad A_S = \{g : S \text{ independent in } g\}$$

$$P(A_S) = (1-p_n)^{x(x-1)/2}$$

(each of the $x(x-1)/2$ pairs in S are not joined)

$$\leq e^{-p_n x(x-1)/2}$$

(used the inequality $1-p \leq e^{-p}$) How many terms are there in the union? At most $\binom{n}{x} \leq n^x$. Using $P(\cup A_S) \leq \sum P(A_S)$, we get

$$P(\alpha \geq x) \leq n^x e^{-p_n x(x-1)/2} = [ne^{-p_n(x-1)/2}]^x.$$

I want to make a choice of x so that the quantity in brackets above is smaller than half. That is want

$$ne^{-p_n(x-1)/2} < 1/2; \quad i.e. \quad x > 1 + \frac{2}{p_n}\log(2n).$$

Take

$$x_n = \lfloor \frac{3}{p_n}\log(2n) \rfloor \qquad \lfloor c \rfloor = \text{greatest integer } \leq c. \quad (\bullet)$$

to convince you this will do, I need to show

$\lfloor \frac{3}{p_n}\log(2n) \rfloor > 1 + \frac{2}{p_n}\log(2n)$, this is true because

$\frac{3}{p_n}\log(2n) > 2 + \frac{2}{p_n}\log(2n)$ this is true because

$\log(2n) > 2p_n$ (for all large n). Remember $0 < p_n < 1$.

Thus, with our choice of x_n, we have on G_n;

$$P(\alpha \geq x_n) \leq (1/2)^{x_n}$$

Since $x_n \uparrow \infty$ the above quantity converges to zero executing ($\diamondsuit$) This completes proof of the Theorem.

Note that as n increases p_n gets smaller, giving only few connections in G_n.

joint distribution:

If we have two random variables on a probability space and look at their distributions separately, We can not understand any relationships that my exist between the random variables.

Scenario 1:

Let us toss a fair coin independently twice. Then outcomes are

$$HH, HT, TH, TT$$

each probability 1/4.

Let X be number of Heads and Y be number of tails. Then X and Y have the same distribution:

Values: 0, 1, 2. probabilities: 1/4, 1/2, 1/4.

The fact that $X + Y = 2$ is not clear when we look at the distributions.

To understand two random variables fully, we need to consider their joint distribution. Just as distribution of one random variable is a table giving values along with respective probabilities; joint distribution of (X, Y) is table giving values of the pair and corresponding probabilities.

Method 1:

Just like one r.v., suppose we plot the values of the pair (X, Y) along with their corresponding probabilities.

values	probabilities
$(0, 2)$	1/4
$(1, 1)$	1/2
$(2, 0)$	1/4

Then it is clear that numbers in each pair add to two.

Method 2:

You can present the same table as a bivariate table as follows. the values of X are in the left vertical margin. The values of Y are in the top horizontal margin. The (i, j)-th entry in the table is the probability that $(X = i, Y = j)$. The bottom horizontal margin are the column totals and the right vertical margin are the row sums.

If you read the top and bottom margins then you see the distribution of Y. If you read the left and right vertical margins then you see the distribution of X.

$X\backslash Y$	0	1	2	total
0	0	0	1/4	1/4
1	0	1/2	0	1/2
2	1/4	0	0	1/4
total	1/4	1/2	1/4	1

Even now it is clear that only those pairs of values that add to two have positive probability and others have probability zero.

Method 1 has the advantage that only those pairs which have non-zero probability are listed, unnecessary pairs are not listed. However if some one asks you what are the possible values of X alone; you need to see each pair and pick up the first coordinate, time consuming. Of course, in the above example it is simple but in general it is not so.

Method 2 appears a neater presentation, though there are several zeros in the matrix of probabilities. However you can immediately understand, without any calculation, the possible values of X and their probabilities: just read the top and bottom margins. Similarly for Y. Thus by looking at this table, you can not only understand the pair (X, Y) but also X alone and Y alone too without any further work. You will be able to detect some interesting phenomena too by just staring at the table. We see now.

Consider tossing a fair coin four times. let X denote the number of heads in the first two tosses and Y the number of heads in the last two tosses.

We shall not describe method 1; it has nine tuples with corresponding probabilities. here is method 2.

$X\backslash Y$	0	1	2	total
0	1/16	1/8	1/16	1/4
1	1/8	1/4	1/8	1/2
2	1/16	1/8	1/16	1/4
total	1/4	1/2	1/4	1

The table immediately tells you another interesting feature: product of the marginals gives the corresponding entry in the matrix. More precisely

i-th row j-th column entry $=$ (i-th row total) $\cdot$ (j-th column total).

Such a feature would not be easily detectable if we used method 1.

joint distribution of two rvs X, Y defined on a space is a bivariate table giving values of X along the first vertical margin and values of Y along the top horizontal margin and the (a, b)-th entry of the table is the probability $p_{ab} = P(\omega : X(\omega) = a, Y(\omega) = b)$

If you look at the row sums $\sum_b p_{ab}$ you get the probability $P(X = a)$. Thus if you enlarge your table by adding row and column totals, then reading the vertical margins gives you the distribution of X, called marginal distribution of X. Similarly the horizontal margins of the table gives you the distribution of Y, called the marginal distribution of Y.

You must keep in mind that the marginal distributions are indeed distributions. The adjective 'marginal' draws your attention to the fact that there are other variables and they had a joint distribution and by looking at the 'appropriate' marginal we got the distribution of X.

Independence of rv:

This is a new definition but there is no new idea.

Two rvs X, Y defined on a space are independent if they have nothing to do with each other; meaning answer to a question about one rv is not influenced when information about the other is revealed. Equivalently, if A is an event described by X and B is an event described by Y, then A, B are independent. Here is the precise definition.

Two rvs X, Y defined on a space are independent if for every values a of X and b of Y; $P(X = a, Y = b) = P(X = a)P(Y = b)$

We do not have to say values of X and value of Y, because in the contrary case both sides are zero, equality holds. Of course, did we put down what we have in mind? After all, if you take any set of values of X, say S and any set of values of Y, say T then we can define two events: $A = \{\omega : X(\omega) \in S\}$ and $B = \{\omega : Y(\omega) \in T\}$. Does sour definition show that these two events

are independent? Yes

$$P(X \in S; Y \in T) = \sum_{a \in S, b \in T} P(X = a, Y = b) = \sum_{a \in S, b \in T} P(X = a)P(Y = b)$$

$$= \sum_{a \in S} \sum_{b \in T} P(X = a)P(Y = b) = \sum_{a \in S} P(X = a)P(Y \in T)$$

$$= P(X \in S)P(Y \in T)$$

finitely many rvs:

Suppose we have finitely many rvs $X_1, \cdots, X_k$ on a probability space. We can talk about the joint distribution of $(X_1, \cdots, X_k)$. It is a table giving all the k-tuples $(a_1, \cdots, a_k)$ of possible values of these rvs and against such a tuple, its probability $P(X_1 = a_1, \cdots, X_k = a_k)$.

The rvs $(X_1, \cdots, X_k)$ defined on a space are independent if for every k-tuple as above

$$P(X_1 = a_1, \cdots, X_k = a_k) = P(X_1 = a_1) \cdots P(X_k = a_k)$$

. Of course, even if a k-tuple of numbers is not a possible value of our rvs, the above equation remains true because both sides then are zero.

Just as in the case of two rvs, we can show that if $S_1, \cdots, S_k$ are subsets of the values of the rvs then independence implies

$$P(X_1 \in S_1, \cdots, X_k \in S_k) = P(X_1 \in S_1) \cdots P(X_k \in S_k)$$

. Indeed,

■ The r.v. $X_1, X_2, \cdots, X_k$ are independent iff for any $S_1, S_2, \cdots, S_k \subset R$ where S_i is a subset of the possible values of X_i

$$P(X_i \in S_i \ \forall \ i) = \prod_1^k P(X_i \in S_i) \qquad \blacksquare$$

■ If $X_1, \cdots, X_k$ are independent then any sub collection is also independent. For example X_2, X_4, X_7 are independent. ■

This follows by taking some of the sets S_i above to be Ω, the sample space.

Digression: table

We have been using the word table to describe distributions. It is not necessary to use this word.

As in the case of one rv, if we have k random variables $X_1, \cdots, X_k$ then their distribution is the function defined on R^k as follows:

$$f(a_1, \cdots, a_k) = P\{\omega : X_i(\omega) = a_i \ \ \forall \ \ i\} \qquad (a_1, \cdots, a_k) \in R^k$$

This is called *joint probability mass function* (in order not to be confused with another function we come across later) of the k variables $X_1, \cdots, X_k$. Since Ω is countable, we see that for only countably many k-tuples, the event in the display above is non-empty. Our earlier description used only these tuples in the table. In the second method we used a 'box' of values for the tuple. Think about it.

expectation of $\varphi(X_1, \cdots, X_k)$:

Suppose we have r.v. $X_1, \cdots, X_k$ defined on a probability space (Ω, p). Suppose that $\varphi : R^k \to R$ is a function. We can define a random variable Z on Ω as follows.

$$Z(\omega) = \varphi(X_1(\omega), \cdots, X_k\ (\omega)) = \varphi o(X_1, \cdots, X_k)(\omega).$$

We can think of calculating $E(Z)$ in several ways.

In what follows we assume that all sums, we are dealing with, are absolutely convergent.

■

(**A**) We can calculate the distribution of Z; say takes values (z_n) with respective probabilities (α_n). Then compute

$$E(Z) = \sum_n z_n \alpha_n$$

(**B**) You already have the joint distribution of $(X_1, \cdots, X_k)$ before you — say this tuple takes values $\{(a_1, \cdots, a_k)\}$ with respective probabilities $\{p(a_1, \cdots, a_k)\}$. No need to calculate distribution of Z. Just compute

$$E(Z) = \sum_{(a_1, \cdots, a_k)} \varphi(a_1, \cdots, a_k) p(a_1, \cdots, a_k)$$

(**C**). Do not calculate any distribution whatsoever. Look at sample space and compute

$$E(Z) = \sum_\omega Z(\omega) p(\omega) = \sum_\omega Z(\omega) p(\omega)$$

■

Are these equivalent? Do they give same answer? Note that (**A**) is definition and the others are not.
Yes. they all give the same answer. We had already seen, in one variable set-up that (**A**) and (**C**) are equivalent. Similar proof shows that (**B**) and (**C**) are also equivalent. Here is how. Start with (**C**). Do the sum in two ways. First for each fixed tuple $(a_1, \cdots, a_k)$; sum or subtotal only over those ω in the set

$$\{X_1 = a_1, \cdots, X_k = a_k)$$

Then add all these subtotals, you get (**B**). ∎

if you know a little analysis, you can actually show that if one series above converges absolutely, then so do the others. Thus absolute convergence of any one of the above three series can be taken towards the existence of $E(Z)$ and then any one of the three series above can be used as definition of $E(Z)$.

The equivalence of the above three will be referred to as **substitution formula** or **change of variable formula**.

∎ If X and Y are independent random variables on a probability space, $E(X)$ and $E(Y)$ defined; Then $E(XY)$ is also defined and

$$E(XY) = E(X)E(Y)$$

Let us not spend time on existence part. Towards the proof of the equality, use the function $\varphi(a, b) = a.b$ in substitution rule to see

$$E(XY) = \sum_{a,b} \varphi(a, b) P(X = a, Y = b)$$

Use independence

$$= \sum_{a,b} a\ b\ P(X = a)P(Y = b)$$

sum w.r.t. b;

$$= \sum_{a} aP(X = a)E(Y) \quad = \quad E(X)E(Y).$$

Covariance, correlation:

if two random variables X, Y are independent then we have

$$E(XY) = E(X)E(Y)$$

It appears reasonable to take $E(XY) - E(X)E(Y)$ as a measure of non-independence or relatedness.

Definition: covariance between X, Y is defined as

$$Cov(X, Y) = \sigma_{X,Y} = E(XY) - E(X)E(Y)$$

Correlation between X and Y is defined as

$$\text{correlation}(X, Y) = \rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

ρ is defined only when. $\sigma_X \neq 0$ and $\sigma_Y \neq 0$. The relation between two quantities should not depend on the units used to measure the quantities. Suppose I measure heights X in inches and weights Y n Kg and find correlation. You find heights in feet and weights in Kg. Then all my quantities X are 12 times yours and you will quickly see, if I did not have the denominator we get different answers.

If X, Y are independent then $cov(X, Y) = 0$. However $cov(X, Y) = 0$ does not mean that they are independent.

theorem:
1. $\sigma_X^2 = 0$ iff there is a number a such that $P(X = a) = 1$
2.If $X_1, \ldots, X_n$ are independent then $Var(\sum X_i) = \sum Var(X_i)$
3. More generally, for any random variables $X_1, \ldots, X_n$ defined on a space $Var(\sum_i X_i) = \sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$
4. If ρ is the correlation between X and Y, then $-1 \leq \rho \leq 1$.

Remember when we talk about variances etc, we assume they exist.

Proof:
(1) Let $E(X) = \mu$. Since $(X - \mu)^2$ takes nonnegative values, the hypothesis $E[(X - \mu)^2] = \sigma^2 = 0$ implies $P\{(X - \mu)^2 = 0\} = 1$. Thus $P(X = \mu) = 1$. Conversely, if $P(X = a) = 1$, then $\mu = E(X) = a$ and $(X - \mu)^2 = 0$ showing $\sigma^2 = 0$.
(2) If $E(X_i) = \mu_i$, then $E(\sum X_i) = \sum \mu_i$. Also by independence, $E(X_i X_j) = \mu_i \mu_j$ for $i \neq j$, so that

$$E\{(X_i - \mu_i)(X_j - \mu_j)\} = 0$$

$$Var(\sum X_i) = E\{(\sum X_i - \sum \mu_i)^2\} = E\{[\sum (X_i - \mu_i)]^2\}$$
$$= E\{\sum (X_i - \mu_i)^2 \ + \ 2\sum_{i<j} (X_i - \mu_i)(X_j - \mu_j)\}$$
$$= \sum Var(X_i)$$

because the crossproduct terms have expectation zero by independence of the random variables.

$$E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - E(X_i \mu_j) - E(X_j \mu_i) + \mu_i \mu_j = 0$$

(3) The same proof above shows this also, use $E\{(X_i - \mu_i)(X_j - \mu_j)\} = Cov(X_i, X_j)$
(4) Suppose $P(X = x_i, Y = y_j) = p_{ij}$. Asume $E(X) = E(Y) = 0$. By change of variable formula and Cauchy-Schwarz

$$|E(XY)| \ = \ |\sum_{i,j} x_i y_j p_{ij}| \ \leq \ \sum_{i,j} |x_i| \ |y_j| p_{ij} \ \leq \ \sum_{i,j} (|x_i| \sqrt{p_{ij}})(|y_j| \sqrt{p_{ij}})$$

$$\leq \ \sqrt{\sum_i \sum_j x_i^2 p_{ij}} \sqrt{\sum_j \sum_i y_j^2 p_{ij}} \ = \ \sigma_X \sigma_Y.$$

Thus $|Cov(X,Y)| \leq \sigma_X \sigma_Y$ showing $|\rho| \leq 1$. In the general case whern means are not zero, $E(X) = \mu$ and $E(Y) = \nu$, apply this result to $(X - \mu)$ and $(Y - \nu)$ to complete the proof. ∎
Here are some consequences

Theorem WLLN

Suppose $X_1, \ldots, X_n$ are independent random variables with common mean μ and common variance $\sigma^2 > 0$, then

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

In particular suppose we have an infinite sequence $\{X_i, i \geq 1\}$ such that for each n, the variables $(X_i, 1 \leq i \leq n)$ satisfy the above hypothesis. Then we have, for each $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \quad \text{as} \quad n \to \infty \qquad \blacksquare$$

This is known as the Weak Law of Large numbers. Of course if the random variables all have same distribution (which has finite variance) then they have the same mean and variance.

The content of the theorem, just as in coin tossing case, is: the average gets as close to the mean as you want with as high probability as you want provided n is large.

In particular if you take sample (with replacement) $X_1, \ldots, X_n$ from a population, then the average $\sum X_i/n$ has the interpretation of observed average. Thus if the sample size is large then the observed average is close to population mean. In particular, if you did not know the mean, then the observed average is a good estimate of the population mean.

The earlier WLLN for coin tossing is a special case of this, by taking each X_i as one with probability p and 0 with probability $q = 1 - p$.

Definition: An infinite sequence of random variables $(X_i, i \geq 1\}$ defined on a space is said to be an independent sequence if for each n the variables $(X_i, 1 \leq i \leq n)$ are independent.

Thus the hypothesis of the theorem amounts to saying that the infinite sequence of variables are independent. Unfortunately we can not give good examples now because we are considering only discrete spaces, that is, countable spaces Ω. Later we will see. However, we can reformulate the theorem in

terms of sampling and sample size (in place of tossing and number of tosses) along the lines of earlier WLLN.

Of course proof of WLLN is simply Chebyshev. Note that if $\sigma^2 = 0$ then, your random variables are 'essentially' constant and the average is also 'essentially' constant and the WLLN holds, you need not use Chebyshev.

Negative Binomial:
Let us consider X, number of tails before r-th head in coin tossing. We know $X \sim NB(r,p)$. Here is an easy way to obtain its mean and variance. First define X_i to be the number of Tails between $(i-1)$-th and i-th Head. Thus X_1 is the number of Tails before the first Head and $X = \sum X_i$.
We claim $(X_i, 1 \leq i \leq r)$ are independent $G(p)$ random variables.
Since we know for such a geometric variable expectation is q/p and variance is q/p^2 we immediately get

$$E(X) = r\frac{q}{p} \qquad Var(X) = r\frac{q}{p^2}$$

To prove the stated independence is easy. The event $(X_1 = m_1, \ldots, X_r = m_r)$ has only one out come $T^{m_1}HT^{m_2}H \ldots T^{m_r}H$. Thus

$$P(X_1 = m_1, \ldots, X_r = m_r) = q^{m_1}pq^{m_2}p \ldots q^{m_r}p \qquad (\spadesuit)$$

Adding over all the $m_2, \ldots m_r$ we get $P(X_1 = m_1) = q^{m_1}p$ showing $X_1 \sim G(p)$. Similarly adding over all m_j except m_i we get $X_i \sim G(p)$ The equation $(\spadesuit)$ now in just

$$P(X_1 = m_1, \ldots, X_r = m_r) = P(X_1 = m_1) \cdots P(X_r = m_r)$$

Hypergeometric:
Consider a box of items of which N_1 are defective and N_2 are good and $N_1 + N_2 = N$. We take a sample of size n without replacement from this box and count the number X of defectives in the sample. here is the distribution of X:

$$P(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}}; \qquad 0 \leq k \leq n$$

Observe that if we have 10 defective, 30 good and take a sample of size 13, then number of defectives in the sample can not be 13. But there is no need to worry because the probability we have written against it above is zero. This distribution is called **Hypergeometric distribution** and a random variable having this distribution as **Hypergeometric random variable**

We shall now calculate its mean and variance. Let X_i denote one or zero according as the i-th item of the sample is defective or good. Thus $X = \sum X_i$. Clearly

$$P(X_i = 1) = \frac{N_1\ (N-1)(N-2)\cdots(N-n+1)}{N(N-1)\cdots(N-n+1)} = \frac{N_1}{N}$$

For $i \neq j$

$$P(X_i = 1, X_j = 1) = \frac{N_1(N_1-1)(N-2)(N-3)\cdots(N-n+1)}{N(N-1)\cdots(N-n+1)}$$

$$= \frac{N_1(N_1-1)}{N(N-1)}$$

Thus for each i

$$E(X_i) = \frac{N_1}{N}; \qquad .Var(X_i) = \frac{N_1}{N}(1 - \frac{N_1}{N})$$

Also for $i \neq j$

$$Cov(X_i, X_j) = \frac{N_1(N_1-1)}{N(N-1)} - \frac{N_1}{N}\frac{N_1}{N} = \frac{N_1}{N}\ \frac{N_1 - N}{N(N-1)}$$

Thus

$$E(X) = n\frac{N_1}{N}$$

from the formulae for variance of sum we have

$$Var(X) = n\frac{N_1}{N}(1 - \frac{N_1}{N})\ +\ 2\binom{n}{2}\frac{N_1}{N}\ \frac{N_1 - N}{N(N-1)}$$

$$= n\frac{N_1}{N}(1 - \frac{N_1}{N})\ -\ n(n-1)\frac{N_1}{N}\ (1 - \frac{N_1}{N})\frac{1}{N-1}$$

$$= n\frac{N_1}{N}(1 - \frac{N_1}{N})\ \{1 - \frac{n-1}{N-1}\}$$

$$= n\frac{N_1}{N}(1 - \frac{N_1}{N})\ \frac{N-n}{N-1}$$

sums of independent random variables:

On a particular day, If the number X of accidents in Chennai is $P(\lambda)$ and the number Y of accidents in Bombay is $P(\mu)$; one would like to know the distribution of the total number of accidents: $X + Y$. It is reasonable to

assume that X and Y are independent.

If $X \sim P(\lambda)$, and $Y \sim P(\mu)$ are independent then $X + Y \sim P(\lambda + \mu)$
Indeed

$$P(X + Y = n) = P(X = k, Y = n - k \quad \text{for some } 0 \leq k \leq n)$$

$$\sum_{k=0}^{n} e^{-\lambda} \frac{\lambda^k}{k!}\ e^{-\mu} \frac{\mu^{n-k}}{(n-k)!} \ = \ e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}$$

If $X \sim B(n, p)$; $Y \sim B(m, p)$ independent then $X + Y \sim B(n + m, p)$
Indeed, proceeding as above, for $k \leq m + n$

$$P(X+Y = k) = P(X = i, Y = j \quad \text{for some} \quad 0 \leq i \leq n; 0 \leq j \leq m; i+j = k)$$

$$= \sum_{i+j=k} \binom{n}{i} p^i q^{n-i} \binom{m}{j} p^j q^{m-j} \ = \ p^k q^{m+n-k} \binom{m+n}{k}$$

We have already seen
$X \sim G(p)$; $Y \sim G(p)$ *independent, then $X + Y \sim NB(2, p)$.*

These results extend to finite sums. For example If X_1, X_2, X_3 are independent then use, $X_1 + X_2, X_3$ are independent. In fact if $X_1, \ldots, X_n$ are independent and Y depends on (X_1, X_9) and Z depends on (X_3, X_5) and W depends on (X_2, X_4, X_8) then Y, Z, W are independent. Here is a sample of such a result.

If X_1, X_2, X_3, X_4 are independent and if f and g are functions of two real variables then the random variables $Y = f(X_1, X_2)$ and $Z = g(X_3, X_4)$ are independent.
Instead of denoting values by x_i, y_j, z_k, w_l etc we just denote by i, j, k, l. If you do not like please feel free to use as you wish.

$$P(Y = m, Z = n) =$$

$$P\left[\bigcup_{ijkl} (X_1 = i, X_2 = j, X_3 = k, X_4 = l) : f(i, j) = m; g(k, l) = n\right]$$

$$= \sum \{P(X_1 = i, X_2 = j, X_3 = k, X_4 = l) : f(i, j) = m; g(k, l) = n\}$$

sum is over i, j, k, l satisfying the conditions mentioned in the brackets.

$$= \sum \{P(X_1 = i, X_2 = j)P(X_3 = k, X_4 = l) : f(i, j) = m; g(k, l) = n\}$$

by independence of (X_i)

$$= P(Y = m)P(Z = n)$$

sum over i, j and then over k, l.

Conditional distributions and Conditional expectations:

You now have all the basic material: probability and conditional probability; random variables; expectation and variance; joint distribution of random variables.
What comes next is combination of conditional probability with other concepts leading to conditional distribution, conditional expectation. This is a truly probabilistic concept and is indeed fundamental to several topics in probability (as well as its applications).

Consider the experiment: Take sample of size two with replacement from $\{0, 1, 2\}$, equivalently, pick two numbers independently at random from these three. Let X be the maximum of the sample and Y is the second chosen point. Sample space is

$$\Omega = \{00;\ 01;\ 02;\ 10;\ 11;\ 12;\ 20;\ 21;\ 22\}$$

Each outcome has probability 1/9.

The joint distribution of (X, Y) is given below.

$X\backslash Y$	0	1	2	total
0	1/9	0	0	1/9
1	1/9	2/9	0	3/9
2	1/9	1/9	3/9	5/9
total	3/9	3/9	3/9	1

Sometimes we have partial information. For example you have info that $X = 2$. You would wonder what could Y be? Of course, when you are dealing with chance variables such a question does not mean you should tell a particular value of Y. You would like to know the distribution of Y. Of course Y takes values $0, 1, 2$ and against each value we should now list the conditional probabilities: $P(Y = 0|X = 2); (Y = 1|X = 2); P(Y = 2|X = 2)$.

Thus conditional distributions of Y are as follows:

Given $X = 0$:

values of Y:	0	1	2
Conditional probabilities:	1	0	0

Given $X = 1$:

values of Y:	0	1	2
Conditional probabilities:	1/3	2/3	0

Given $X = 2$:

values of Y:	0	1	2
Conditional probabilities:	1/5	1/5	3/5

These are called conditional distributions of Y for the values of X given as shown. Note that these conditional probabilities do add to one in each case. You can calculate conditional expectations which means expectations under these conditional distributions. Thus $E(Y \mid X = 0)$, conditional expectation of Y given $X = 0$ is 0.

Similarly $E(Y \mid X = 1) = 2/3$ and $E(Y \mid X = 2) = 7/5$.

Remember distribution of Y is only one; whereas conditional distributions are many; one for each given value of X. Similarly expectation is just a number where as conditional expectations of Y are many numbers; one for each given value of X.

We want to summarize all these conditional expectations into one quantity; not a number but a function Z. It will be defined as a function of X. When X takes a value x this function takes the number $E(Y \mid X = x)$ as its value. Thus for all sample points ω for which $X(\omega) = a$ we have the same value for the function Z. This Z is called conditional expectation of Y given X, denoted $E(Y|X)$. You think of it as a function of X; say, $\varphi o X$.

So keep in mind $E(Y \mid X)$ is not a number; it is a function of X. There is again no confusion because we only said 'expectation of Y given X' (sounds like an incomplete sentence, we only said given X but did not say what is it that is given about X) and did not say any specific value of X. Hence it must encode all the information and hence it is the function Z rather than any one of the earlier numbers.

Since it is a function of X, it is a random variable. So it makes sense to take its expectation once again!

Let us quickly do all these considerations in generality.

Suppose (Ω, p) is a probability space and X, Y are random variables. Suppose their joint distribution, given in a bivariate table is as follows.

X takes values $\{x_i : i = 1, 2, \cdots\}$ may be finite or infinite set.
Y takes values $\{y_j : j = 1, 2, \cdots\}$ may be finite or infinite set.

$$p_{ij} = P(X = x_i, Y = y_j)$$

The so called marginal totals are the following.

$$p_{i\bullet} = \sum_j p_{ij}; \quad p_{\bullet j} = \sum_i p_{ij}; \quad i, j \geq 1$$

Thus the distribution of X and Y are given by

$$P(X = x_i) = p_{i\bullet} : \quad i = 1, 2, \cdots$$

$$P(Y = y_j) = p_{\bullet j} : \quad j = 1, 2, \cdots .$$

Now we can define the conditional distribution of Y given $X = x_i$:

$$P(Y = y_j \,|\, X = x_i) = \frac{p_{ij}}{p_{i\bullet}} : \quad j = 1, 2, \cdots .$$

this is defined for each $i = 1, 2, \cdots$ and whatever be i all the numbers above add to one when summed over j. You must remember i and j are on different footing here. In the above display, i is fixed. j varies.

$$E(Y \,|\, X = x_i) = \sum_j y_j \frac{p_{ij}}{p_{i\bullet}}$$

Finally $E(Y \,|\, X)$ is the following function: When X takes value x_i, it takes the value $E(Y \,|\, X = x_i)$. If you want what exactly is the function, it is the following on the sample space.

$$E(Y \,|\, X)\ (\omega) = E(Y \,|\, X = a); \quad \text{where} \quad a = X(\omega)$$

Thus the function $E(Y \,|\, X)$ takes same value at two sample points ω and η if $X(\omega) = X(\eta)$. As a consequence it is a function of X, namely

$$E(Y \,|\, X) = \varphi o X; \quad \text{where} \quad \varphi(x_i) = E(Y \,|\, X = x_i)$$

Of course we have not defined the function φ at real numbers other than the x_i s. But if you are worried about that put $\varphi(x) = 0$ if x is not any of the numbers $\{x_i\}$. These do not matter. After all when when you compose two functions $f(g(x))$, values of f on points in the range of g matter. Think about it.

As a consequence of the change of variable rule

$$E[E(Y \mid X)] = \sum_i \varphi(x_i) P(X = x_i)$$

$$= \sum_i \sum_j y_j \frac{p_{ij}}{p_{i\bullet}} p_{i\bullet} = \sum_j \sum_i y_j p_{ij} = \sum_j y_j p_{\bullet j}$$

$$= \sum_j y_j P(Y = y_j) = E(Y)$$

Thus we have proved

$$\blacksquare \qquad E[E(Y \mid X)] = E(Y) \qquad \blacksquare$$

This is a very useful formula. It allows you to calculate expected values in complicated situations. here is an example.

Example:

Imagine there is a maze with five doors: 1,2,3,4,5.

If rat exits through 1, it is out in one minute. If exits through 2, then it it is out in 2 minutes. If exits through 3, it travels for 3 minutes and then returns to the starting place. similaly, If it exits through 4 (or 5), then returns to the starting place after 4 (or 5) minutes.

Assume that the rat always chooses one of the four doors at random. What is the expected time to exit from the maze?

You can write down the sample points, for each sample point calculate its probability $p(\omega)$ and the time $Y(\omega)$ to exit for that sample point ω and calculate $\sum Y(\omega)p(\omega)$.
Every sample point is a finite (possibly empty) sequence of symbols 3,4,5 followed by a last symbol which is either 1 or 2.
Here is a simpler way. Let Y be the time to exit and X be the first choice of door. Suppose $E(Y) = a$. Clearly

$$E(Y \mid X = 1) = 1; \quad E(Y \mid X = 2) = 2. \quad E(Y \mid X = 3) = 3 + a$$

because the conditional distribution of Y given $X = 3$ is same as the distribution of $3 + Y$. similarly $E(Y|X = 4) = 4 + a$. $E(Y|X = 5) = 5 + a$. Thus using the earlier formula,

$$a = E(Y) = E[E(Y|X)] =$$

$$\frac{1 + 2 + 3 + a + 4 + a + 5 + a}{5} = (15 + 3a)/5$$

or

$$a = 15/2$$

Of course you might wonder if this is an artificial example. No. If some one says that rats learn from experience, how do you test if this is right? Here is one way. you conduct above experiment on several rats several times on each. If they always take time close to above value then obviously they are not learning. If the time is visibly shorter after some repetitions, then you have reason to believe that they are learning.

The conditional distribution of Y given $X = x_i$ is a new definition, not a new concept. it is just concept of distribution and conditional probability of events. Conditional expectation $E(Y|X = x_i)$ is a new definition, not a new concept; it is just expectation w.r.t. the conditional; distribution. However $E(X|Y)$ is a new concept, we are thinkling of this as a function on the sample space. Unless you are careful, this will cause confusion.

pgf:

Let X be a non-negative integer valued random variable, say takes the value k with probability p_k for $k = 0, 1, 2 \ldots$.

Definition: The probability generating function (pgf) is defined by

$$\varphi_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots = \sum p_n s^n \qquad 0 \leq s \leq 1.$$

This is an infinite degree polynomial or power series. Note it converges for at least $-1 \leq s \leq 1$. Of course if the random variable takes only finitely many values, then this is a polynomial and defined for all s.

This is a way of remembering the sequence of probabilities. you can see that

$$p_n = \varphi^{(n)}(0)/n!$$

Thus the function φ generates the probabilities.

$$X \sim B(n, p); \qquad \varphi_X(s) = (q + ps)^n$$

$$X \sim P(\lambda); \qquad \varphi(s) = e^{-\lambda(1-s)}$$

$$X \sim G(p); \qquad \varphi(s) = 1/(1-qs)$$

Since

$$(p_0 + p_1 s + p_2 s^2 + \cdots)(a_0 + a_1 s + a_2 s^2 + \cdots) =$$

$$p_0 a_0 \ + \ (p_0 a_1 + p_1 a_0) s \ + \ (p_0 a_2 + p_1 a_1 + p_2 a_0) s^2 \ + \cdots$$

we see that if X, Y are independent (non-negative integer valued) then

$$\varphi_{X+Y} = \varphi_X \ \varphi_Y$$

This is one way of computing the distribution of sum of independent random variables. You can identify the distribution of $X + Y$ by looking at $\varphi_X \varphi_Y$.

Here is an example from Manjunath Krishnapur and Persi Diaconis

$X \backslash Y$	0	1	2	total
0	1/9	1/9	1/9	3/9
1	1/9	1/9	1/9	3/9
2	1/9	1/9	1/9	3/9
total	3/9	3/9	3/9	1

Fix any ϵ, $0 < \epsilon < 1/9$,

$X^* \backslash Y^*$	0	1	2	total
0	1/9	$1/9 - \epsilon$	$1/9 + \epsilon$	3/9
1	$1/9 + \epsilon$	1/9	$1/9 - \epsilon$	3/9
2	$1/9 - \epsilon$	$1/9 + \epsilon$	1/9	3/9
total	3/9	3/9	3/9	1

Then X, X^*, Y, Y^* all have the same distribution; uniform on $\{0, 1, 2\}$. X, Y are independent. But X^*, Y^* are not independent. However $X + Y$ and $X^* + Y^*$ have same distribution. Thus $\varphi_{X^*} \ \varphi_{Y^*} = \varphi_{X^*+Y^*}$ but X^*, Y^* are

not independent. You can even take $\epsilon = 1/9$.

$X^*\backslash Y^*$	0	1	2	total
0	1/9	0	2/9	3/9
1	2/9	1/9	0	3/9
2	0	2/9	1/9	3/9
total	3/9	3/9	3/9	1

mgf:
Let now X be any random variable not necessarily integer valued. Say, takes values $x_i, i \geq 1$ with respective probabilities $p_i, i \geq 1$. We have defined $E(X) = \sum x_i p_i$. and $E(X^2) = \sum x_i^2 p_i$. These are called the first and second moments, μ_1 and μ_2 . These were, to some extent, give a feeling for the distribution – mean and spread. We can define n-th moment.

$$\mu_n = E(X^n) = \sum x_i^n p_i \qquad \text{provided} \qquad \sum |x_i^n| p_i < \infty$$

This is called n-th moment. We put $\mu_0 = 1$.

The moment generating function, mgf, of a random variable X is the function defined by

$$M_X(t) = E(e^{tX}) \quad \textit{defined for} \quad t \in \{E(e^{tX}) < \infty\}$$

Unlike the pgf which is defined for at least $-1 \leq s \leq 1$, it may so happen that M is defined only for $t = 0$, even if X is integer valued. Note that $M(0) = 1$ always. Thus if X takes values $\{x_i\}$ with corresponding probabilities $\{p_i\}$ then

$$M(t) = \sum e^{tx_i} p_i.$$

For example,

$$X \sim B(n,p); \qquad M(t) = (q + pe^t)^n$$
$$X \sim P(\lambda); \qquad M(t) = e^{-\lambda(1-e^t)}$$
$$X \sim G(p); \qquad M(t) = 1/(1 - qe^t) \qquad \text{if} \quad qe^t < 1$$

When X and Y are independent then so are e^{tX} and e^{tY}. Since $e^{t(X+Y)} = e^{tX} e^{tY}$ we see

$$X, Y \quad \text{independent}: \qquad M_{X+Y} = M_X M_Y \quad t \in Dom(M_X) \cap Dom(M_Y)$$

where *Dom* is domain of the function.
mgf does actually generate the moments. Suppose a random variable has mgf defined for t in an interval $(-a, +a)$. Then we can show

$$\mu_n = M^{(n)}(0)$$

where $M^{(n)}$ denotes the n-th derivative. One can show that under the assumed hypothesis, the function M is differentiable any number of times and the above equality holds.
If X has mgf $M_X(t)$, then $Y = aX$ has $M_Y(t) = M_X(at)$ These functions play fundamental role in certain contexts. However our purpose is not to discuss these in detail. My main purpose is to prove a beautiful result.

Chernoff bound:

Theorem (Chernoff bound): Let X be a random variable and $a > 0$. Then for any $t > 0$ such that $M_X(t)$ is finite, we have

$$P(X \geq a) \leq M_X(t)\, e^{-ta}$$

Proof is simple, the events $(X \geq a)$ and $(e^{tX} \geq e^{ta})$ are same because $t > 0$ and exponential function is increasing. Now Chebyshev does it. ∎

But then why is it so important. You have a handle t with you and you can make best use of it. Here is an illustration.

Let us consider $X_1, X_2, \ldots, X_n$ independent and each assuming values ± 1 with equal probability. Thus mean is zero. Also variance is one. Let $S_n = \sum X_i$. Thus S_n has variance n, Thus variance of S_n/n is $1/n$. Usual Chebyshev gives, for any $a > 0$

$$P(|\frac{S_n}{n}| \geq a) \leq \frac{1}{na^2} \qquad (\clubsuit)$$

Let us see how our handle helps us. Let $Y = S_n/n$. Then

$$M_Y(t) = M_{S_n}(t/n) \qquad M_{S_n}(t) = [M_{X_1}(t)]^n$$

the last equality is by independence.

$$M_{X_1}(t) = \frac{(e^t + e^{-t})}{2} = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots + \frac{t^{2n}}{(2n)!} + \cdots$$

Since

$$(2n)! \geq 2 \times 4 \times 6 \times \cdots \times 2n = 2^n\, n!$$

we get

$$M_{X_1}(t) \leq 1 + \frac{t^2/2}{1!} + \frac{(t^2/2)^2}{2!} + \frac{(t^2/2)^3}{3!} + \cdots + \frac{(t^2/2)^n}{n!} + \cdots$$

$$= e^{t^2/2}$$

Thus

$$M_{S_n}(t) \leq e^{nt^2/2} \qquad M_Y(t) \leq e^{t^2/2n}$$

Chernoff gives

$$P(\frac{S_n}{n} \geq a) \leq e^{-ta}\ e^{t^2/2n}$$

We choose $t = na$ so that $ta = na^2$ and $(t^2/2n) = na^2/2$ giving

$$P(\frac{S_n}{n} \geq a) \leq e^{-na^2/2}$$

Since $-S_n$ and S_n have same distribution we get

$$P(\frac{S_n}{n} \leq -a) \leq e^{-na^2/2}$$

Thus

$$P(|\frac{S_n}{n}| \geq a) \leq 2e^{-na^2/2} \qquad (\spadesuit)$$

This is an exponential bound compared to (♣). This precision is very important especially when you simulate you need to know after how many steps you should stop. This has other uses too.

Chernoff bound along with the following (we met earlier) belong to the standard tool kit.

First moment method: $P(|X| \geq a) \leq E(|X|)/a$ for $a > 0$.

Second moment method: $P(|X| \geq a) \leq E(X^2)/a^2$ for $a > 0$.

Union bound: $P(\cup A_i) \leq \sum P(A_i)$, for events $\{A_i\}$

Experiments with toooo many outcomes:

We have been discussing experiments/rvs that take countably many values, usually called discrete rvs. We shall now discuss experiments that have tooooo many outcomes and rvs that take tooooo many values.

For example, when you want to model life time of this bulb, any number $0 < x < \infty$ is a possible value. The same when you want to model the time between two disintegrations of a radio active material. That is, imagine a radio active material placed near a Geiger counter. As soon as a particle is emitted, the counter beeps. Consider the time between two beeps. It is not fixed. This can be anything in $(0, \infty)$.

an experiment:

to make life simple, let us consider the following experiment: pick a point at random from the interval $(0, 1]$. How should we build a model for this experiment.

First step is this: we should understand as to what happens when you do the experiment – understand the set of outcomes. Obviously any number in this interval is a possible outcome. Thus the sample space is

$$\Omega = \{x : 0 < x \leq 1\} = (0, 1].$$

How should we go about assigning probabilities?

imitation being the first choice here is the **first try**. Associate probability for each outcome and then define probabilities of events as sum of probabilities of outcomes in that event. But there are two problems. Firstly, event could now contain an uncountable number of outcomes, how are you going to define uncountable sums? (well, can be done; but in our context useless; let us not bother).

Secondly, what could be probability for an outcome? The picking is random, so every outcome must have same probability; no bias! but if chance of every outcome is, say $1/10^{10}$, then taking a subset with $10^{10}+1$ outcomes you will get an event with probability larger than one. But chances of anything should be between zero and one. This argument shows that chance of an outcome can not be *any* strictly positive number; it has to be zero.

Thus the imitation fails. We have to reconcile to the fact that there are too many outcomes and chance of every single outcome must be zero. so what do we do? Take a clue from the physicists. If you take this piece of paper;

they talk about mass density. mass at each single point is zero; but it is distributed across; we do not take mass of a region of the paper to be sum of masses of points. We assign masses to regions at one stroke.

Thus here is **second try**. Associate for every event A; that is, every subset of $(0,1]$ a number $P(A)$ to 'represent' the chances that the selected number falls in A. What do we expect such an association to satisfy?

(i) $P(\Omega) = 1$ and $P(\emptyset) = 1$ and $0 \leq P(A) \leq 1$.
(ii) If $A_1, A_2, \cdots$ is a sequence of disjoint events then $P(\cup A_n) = \sum P(A_n)$.
(iii) $P(\frac{i}{2^k}, \quad \frac{i+1}{2^k}] = \frac{1}{2^k}; \qquad 0 \leq i < 2^k; \quad k \geq 1.$

The first condition is our belief that chances of anything should be non-negative. Chance of something or other happening should be one; chance of nothing happening should be zero.
Second condition is just what we used in the discrete setup and we should not forego that. The chances that a Poisson variable X takes an even integer value is obtained by summing $P(X = n)$ for $n = 0, 2, 4, 6, \cdots$.
The third conditions just reflects the fact that we are selecting point X at random. For example chances that $(X \in (0, 1/2])$ should be same as the chances that $(X \in (1/2, 1])$. Hence each must have probability $1/2$. Similarly the events $(X \in (0, 1/4])$ and $(X \in (1/4, 1/2])$ should have the same probability and hence each must be $1/4$. And so on.

Is there an assignment $A \mapsto P(A)$ for all subsets of Ω which satisfies the three conditions above. Unfortunately answer to this question is not straight forward. [Under two assumptions — namely, Axiom of Choice and Continuum Hypothesis — the answer is in the negative. You need not understand these]. Our attempt fails.

Where did we go wrong? After all, any demand that is reasonable should be satisfiable. All the three demands listed are reasonable. the unreasonableness lies at an unexpected place. We wanted assignment of probability for **every** subset of Ω. Is it necessary? What is the purpose of doing probability? To answer questions regarding chances of certain things happening. In other words, calculate probabilities of events that we come across. Should we unnecessarily burden ourselves with assigning probabilities for sets that we never come across?

In other words; not every set should be an event. Events are just those subsets in which we will be interested; not every subset of Ω.

So what subsets of Ω should be called events? Intervals occur in practice, they are the simplest sets and they should be called events. For example, we will be interested in the simple question $P(0.258 < X < 0.349)$.

If we are interested in some thing happening, then we will also be interested in the chances of that not happening. Thus if A is an event then its complement A^c should also be an event.

If we are interested in: What are the chances that $(X \in A_1)$?; what are the chances that $(X \in A_2)$?; etc then we will be interested in: what are the chances that X is in 'one of those sets'. In other words if each A_n is an event then their union $\cup A_n$ should be an event. Since this is what we were using in experiments with countably many outcomes, we require this here. More over condition (ii) for probability already suggests that if each A_n is an event then $\cup A_n$ must be an event. Of course we do not demand that uncountable union of events be an event.

It is unnecessary for you to remember these conditions. You just remember that the collection of events is a bag $\mathcal{B}$ of subsets of Ω satisfying some reasonable conditions.

So here is a **third try** for modelling the experiment of picking a point at random from $(0, 1]$.

Is there a bag $\mathcal{B}$ of subsets of $\Omega = (0, 1]$ satisfying 'reasonable' conditions and for each set A in the bag a number $P(A)$ satisfying conditions (i,ii,iii) above.

Finally **success**: Yes, there is such a $\mathcal{B}$ and P.

Of course, you can continue this discussion: how many such things are there; if there are many; which one should we take as model etc. But we stop this discussion here by just noting that there is only one such 'with proper formulation' and so there is no confusion.

In passing let me assure you that such a bag contains *all subsets you can think of.* In other words, whatever be such a bag, it is hard to think of sets which are NOT in the bag! This does not mean most of the subsets of Ω are in such a bag; far from it. There are many many more sets which are not in the bag than sets which are in the bag. However all subsets you can think of are here. Thus for all practical purposes, assigning probability for sets in such a bag of is enough for answering all questions of practical importance.

Probability models:

Thus the upshot of all this is the following: No need to allow every subset of sample space as an event. make up your mind as to which sets you would

like to be events and then assign probability. Here is the precise definition:

We define a probability model or a probability space to consist of a triple $(\Omega, \mathcal{B}, P)$.

(•) Ω is a non-empty set.

Ideally it is the set of outcomes of your experiment.

(••) $\mathcal{B}$ is a bag of subsets of Ω such that (a) empty set and Ω are in the bag; (b) if a set is in the bag then so is its complement; and finally (c) if you take a sequence of sets in the bag then their union is also in the bag.

Ideally sets in the bag are events in which you will be interested. Sets not in the bag are NOT events.

(• • •) P is a map that associates with every A in the bag a number $P(A)$ in such a way that (i) $P(\emptyset) = 1$ and $P(\Omega) = 1$ and $0 \leq P(A) \leq 1$; (ii) for a sequence $(A_n : n \geq 1)$ of disjoint sets in the bag $P(\cup A_n) = \sum P(A_n)$.

Ideally $P(A)$ denotes the probability of ending up with an outcome in A when you perform the experiment.

As usual a random variable is a measurement — associates with every outcome a real number. In other words it is a real valued function X defined on Ω. Remember we should be able to answer questions concerning our measurement. if some one asks what are the chances that value of the measurement is at most 29; we should be able to answer. How do we propose to answer? The obvious way, collect all outcomes for which the required condition holds, that is; $A = \{\omega : X(\omega) \leq 29\}$. Then $P(A)$ is the required answer.
There is one catch, how do we know that the above set A is in the bag? If it were not, then $P(A)$ is meaningless. Thus we could answer questions about our measurement only when such sets are in the bag. With this in mind we make the following definition.

A random variable X is a real valued function on Ω such that for every real number a, the set $\{\omega : X(\omega) \leq a\}$ is an event. That is, this set is in the bag $\mathcal{B}$.

Of course, you might still wonder if answering such simple questions is good enough. We might be interested in more complicated questions concerning our random variable, would it be possible to answer them with the above definition? Yes, we can answer any question you can think of, if only you can answer the simple questions described above.

You probably would now think life is getting complicated because we talked about bag of sets etc. But you must keep in mind the following.

(i) There is **just one** simple difference between experiments with countably many outcomes and experiments with uncountably many outcomes: not every subset of outcomes is an event. The discrete experiments are also covered with this formulation because you can always take the bag of events to be the collection of all subsets of the sample space. (ii) All the above discussion is lie foundation of building whichj we need not worry about. We only use the rooms and pathways in the buildind but not the foundation. [Remember: If good foundation does not exist, the building collapses].

Density:
How do we hope to answer questions concerning random variables? In the discrete case we defined distribution of the random variable. This is table: value of the random variable along with probability of taking that value. We used this to answer all questions regarding the variable, we need not look at the sample space.

In the present case there is a function which answers questions about the random variable. This function is like a lawyer for the random variable.

We say that a function f on the real line is a density function if it takes non-negative values and $\int_{-\infty}^{\infty} f(x)dx = 1$. we say that a random variable X obeys a density function f if for any number a; $P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.

This is enough to answer all questions (we can think of) about the random variable. For example if $a < b$;

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

This is because

$$P(X \leq a) + P(a < X \leq b) = P(X \leq b)$$

Similarly if you take disjoint union of a sequence of intervals and ask for the chances of X being in one of these intervals, you can use additivity and answer. Thus answers regarding probabilities of events described using the rv can be obtained by looking at the area under the curve above appropriate interval (or union of intervals). This is just similar to discrete case where we used sums of mass function over appropriate sets. Think.

Incidentally, the densities we consider, in our course, are piecewise continuous and the integrals above are to be interpreted as Riemann integrals – as learnt in your calculus course. But, in practice one allows a more general concept of integration; we shall not. **we use only Riemann integral. Relax.**

Sometimes density function is also called probability density function. Any way we are doing probability and we do not need this adjective.

We shall handle questions regarding random variable via its density function, not going to the sample space Ω and the bag of sets $\mathcal{B}$. Relax.

The doubt arises: Then why did you tell us all this story, bag of sets and nonsense. If we did not go over it, you would not know what are the problems associated with modelling; what exactly is a model and what exactly is to be done to get a model; you can not even define a random variable. You are left with the impression that random variable is something hanging in the air; it is a fuzzy object; it is random etc. it is neither random nor variable; it is as concrete as you and me. It is a function defined on Ω – measurement associated to each outcome.

[There are historical reasons for using the word 'random variable'. The word 'random' draws your attention to the fact that this is not any function on an arbitrary set; there is a probability floating in the background. The word 'variable' suggests that you can use it as variable and talk, for example about $\sin X$ and e^X etc — like functions of real variable etc.]

Of course, it is a different matter that we finally decided to use density function to answer questions about the random variable. This only reflects the fact that we are not mature enough to handle probability spaces.

Just as there are several rvs; binomial, Poisson, geometric, etc in the discrete case here too there are several that raise in practice and we start with some of them to get concrete examples of rvs. You must sketch all these curves and see them explicitly.

unif$(0,1)$**:**
Consider the function $f(x) = 1$ for $0 < x < 1$ and $f(x) = 0$ for $x \notin (0,1)$.
You can see that the area under the curve is one.
This is called **uniform** $(0,1)$ **density** and a random variable that obeys this density is called a **uniform**$(0,1)$ **random variable**.

It is easy to see that if X is such a random variable then the following is

true: for any interval $(a, b) \subset (0, 1)$ we have $P(X \in (a, b)) = b - a$. further, $P(X \in (0, 1)^c) = 0$.

In other words this random variable is precisely a point picked at random from $(0, 1)$.

More generally, fix any interval (α, β) where $-\infty < \alpha < \beta < \infty$. Let f be the function defined as

$$f(x) = \frac{1}{\beta - \alpha}, \quad x \in (\alpha, \beta); \qquad f(x) = 0, \quad x \notin (\alpha, \beta)$$

This is a density function named **unif**(α, β) **density** and a random variable having this density is called **unif**(α, β) **random variable**. This models a point selected at random from this interval.

Exp(λ):

Consider the following function: $f(x) = e^{-x}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$.

you can see area under the curve is one. This density function is called **exponential density**, more precisely exp(1) density function and a random variable having this density is called **exponential random variable**, more precisely, exponential rv with parameter 1; exp(1) random variable.

If X is such a random variable then $P(X \leq 0) = 0$ where as for any $0 \leq a < b$, we have

$$P(a < X < b) = e^{-a} - e^{-b}.$$

More generally fix $\lambda > 0$. Consider

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0; \qquad f(x) = 0, \quad x \leq 0.$$

This is called exp(λ) **density** and a random variable having this density is called exp(λ) **random variable**. Here $\lambda > 0$ is a parameter and so this is called exponential variable with parameter λ. Remember we had binomial with parameter p, Poisson with parameter λ etc.

This density is useful in modelling life time of electric bulbs. This is also a good model for inter-disintegration time for radio active material. This is also good model for inter arrival times of customers at a service station.

Gamma density:

Fix $a > 0$. Let

$$\varphi(x) = e^{-x} x^{a-1} \quad x > 0; \qquad \varphi(x) = 0; \text{ for } x \leq 0$$

Then we claim that $\int_{-\infty}^{\infty} \varphi(x)dx$ is finite.

First we claim that $\int_1^\infty \varphi(x)dx$ is finite. Indeed if $0 < a < 1$ then

$$e^{-x}x^{a-1} \leq e^{-x} \qquad x > 1$$

since the area under the function on the right side is finite, it is so for the left side function as well. In case $a \geq 1$, fix an integer k such that $a < k+1$. Now you use the fact that

$$e^{-x}x^{a-1} \leq e^{-x}x^k \qquad x > 1$$

As above area under right side curve is finite etc.

Now we shall argue $\int_0^1 \varphi(x)dx$ is finite. if $a \geq 1$ then integrand is a continuous function on $[0,1]$ and is hence integrable. Let now $0 < a < 1$. Then

$$e^{-x}x^{a-1} \leq x^{a-1} \qquad 0 < x < 1$$

As earlier comparison takes over.

Since both $\int_0^1 \varphi(x)dx$ and $\int_1^\infty \varphi(x)dx$ are finite we conclude that $\int_0^\infty \varphi(x)dx$ is finite. We denote

$$\int_0^\infty \varphi(x)dx = \Gamma(a).$$

$f(x) = \dfrac{1}{\Gamma(a)}\varphi(x)$ is a density function. This is called **Gamma density**, more precisely gamma density with parameter (a). A random variable which obeys this density is called a **gamma random variable**.

It is instructive to sketch the curves $f(x)$ for $a = 1/2$; for $a = 1$; for $a = 2$.
Integration by parts immediately gives that $\Gamma(a+1) = a\Gamma(a); \qquad a > 0.$
This will immediately gives $\Gamma(n) = (n-1)!; \qquad n \geq 1.$

Beta density:
Fix $a > 0, b > 0$. Let

$$\varphi(x) = x^{a-1}(1-x)^{b-1} \quad 0 < x < 1; \quad \varphi(x) = 0; \text{ if } x \notin (0,1).$$

We claim that $\int_{-\infty}^{\infty} \varphi(x)dx$ is finite. Of course the integrand being zero outside unit interval this amounts to showing $\int_0^1 \varphi(x)dx$ is finite.

We first show that $\int_0^{1/2} \varphi(x)dx$ is finite. This is immediate if you observe that $(1-x)^{b-1}$ is a bounded continuous function on this interval and a being strictly positive, x^{a-1} is integrable.

We now show that $\int_{1/2}^1 \varphi(x)dx$ is finite. This is immediate if you observe that x^{a-1} is a bounded continuous function on this interval and b being strictly positive, $(1-x)^{b-1}$ is integrable.

Since both $\int_0^{1/2} \varphi(x)dx$ and $\int_{1/2}^1 \varphi(x)dx$ are finite we conclude that $\int_0^1 \varphi(x)dx$ is finite. We denote

$$\int_0^1 \varphi(x)dx = \beta(a,b).$$

Thus

$$f(x) = \frac{1}{\beta(a,b)}\varphi(x)$$

is a density. This is called **beta density**. More precisely it is called beta density with parameters a and b. A random variable obeying this density is called **beta random variable**.

It is instructive to sketch these curves for $a = 1/2, b = 1/2$; for $a = 1, b = 1$; for $a = 2; b = 1/2$; for $a = 1/2, b = 2$; for $a = 2, b = 3$.

The gamma and beta functions arise in several contexts, not only probability but also in Physics, number theory, differential equations, special functions and so on. So does the normal density that we discuss next.

Double exponential:

$$f(x) = \frac{1}{2}e^{-|x|} \qquad -\infty < x < \infty$$

Then $\int f = 1$. This is called double exponential or Laplace density. A rv having this density is called a double exponential rv.

Normal:

There are several densities that arise in practice. But we discuss only one more. Let

$$f(x) = e^{-x^2/2}; \qquad -\infty < x < \infty.$$

Then $\int f(x)dx < \infty$. Indeed on $[-1, 1]$ this is bounded continuous and hence integrable. On $(-\infty, 1)$ and $(1, \infty)$ we have $f(x) \leq e^{-|x|}$ and hence integrable. We later show that $\int f = 1/\sqrt{2\pi}$. Thus

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty.$$

is a density. This is called **standard normal density** and a random variable obeying this density is called a **standard normal variable**.

Normal integral:

You can use double integrals to show, in a painless way,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \;=\; 1.$$

You need to use polar coordinates, change variables $x = r\cos\theta$ and $y = r\sin\theta$ and use Jacobian formula.

Since we are not sure of these techniques, we shall follow elementary high school methods to derive the same result. Integrand being symmetric, we only need to show

$$\int_0^{\infty} e^{-x^2/2} dx \;=\; \sqrt{\pi/2}$$

Denote

$$a_n = \int_0^{\infty} x^n e^{-x^2/2} dx; \quad n = 0, 1, 2, \cdots$$

$$a_0 = ?; \qquad a_1 = 1.$$

Integration by parts gives $a_k = (k-1)a_{k-2}$ for $k > 1$ leading to

$$a_{2n} = (2n-1)(2n-3)\cdots(3)(1)a_0 = \frac{(2n)!}{n!\, 2^n} a_0$$

$$a_{2n+1} = (2n)(2n-2)\cdots(2)1 = n!\, 2^n$$

Fix $k \geq 1$. Note that for any $\lambda \in R$ we have

$$\int_0^{\infty} x^{k-1} \left(\lambda x - 1\right)^2 e^{-x^2/2} dx > 0$$

because integrand is non-negative, continuous and is not the zero function. Thus

$$\lambda^2 a_{k+1} - 2\lambda a_k + a_{k-1} > 0; \qquad \forall \lambda \in R$$

hence

$$a_k^2 \leq a_{k+1} a_{k-1}; \qquad \text{or} \qquad a_k \leq \sqrt{a_{k+1} a_{k-1}}$$

Now $a_{2n} \leq \sqrt{a_{2n+1} a_{2n-1}}$ gives you

$$\frac{(2n)!}{n!\ 2^n} a_0 \ \leq \ \sqrt{n!\ (n-1)! 2^n\ 2^{n-1}} \ = \ n!\ 2^n \frac{1}{\sqrt{2n}}$$

Or

$$a_0 \ \leq \ \frac{n!\ n!\ 2^n\ 2^n}{(2n)!\ \sqrt{2n}}$$

This being true for all n, we get

$$a_0 \ \leq \ \lim_n \frac{n!\ n!\ 2^n\ 2^n}{(2n)!\ \sqrt{2n}} \qquad (\spadesuit)$$

Now use $a_{2n+1} \leq \sqrt{a_{2n+2} a_n}$ to see

$$n!\ 2^n \ \leq \ \sqrt{\frac{(2n)!}{n!\ 2^n}\ a_0\ \frac{(2n+2)!}{(n+1)!\ 2^{n+1}}\ a_0} \ = \ a_0\ \frac{(2n)!}{n!\ 2^n} \sqrt{(2n+1)}$$

Or

$$a_0 \geq \frac{n!\ n!\ 2^n\ 2^n}{(2n)!} \frac{1}{\sqrt{2n+1}}$$

This being true for every n, we get

$$a_0 \geq \lim_n \frac{n!\ n!\ 2^n\ 2^n}{(2n)!} \frac{1}{\sqrt{2n+1}} \qquad (\clubsuit)$$

Of course, in both ($\spadesuit$) and ($\clubsuit$) we assumed that the limits on the right exist. the limits are same because $(2n)$ or $(2n+1)$ makes little difference (to whom?).

We shall now show that those limits exist and equal $\sqrt{\pi/2}$ as required.

Walli's product:
We start with

$$\int_0^{\pi/2} (\sin x)^0 dx = \frac{\pi}{2}; \qquad \int_0^{\pi/2} (\sin x)^1 dx = \cos 0 - \cos(\pi/2) = 1.$$

If $m > 1$,

$$\int_0^{\pi/2} (\sin x)^m dx = \int_0^{\pi/2} (\sin x)^{m-1} (-\cos x)' dx$$

integration by parts

$$= \int_0^{\pi/2} \cos x \;\; (m-1) \sin^{m-2} x \cos x dx$$

$$= (m-1) \int_0^{\pi/2} \sin^{m-2} x dx - (m-1) \int_0^{\pi/2} \sin^m x dx$$

so that

$$\int_0^{\pi/2} \sin^m x dx = \frac{m-1}{m} \int_0^{\pi/2} \sin^{m-2} x dx.$$

Thus

$$\int_0^{\pi/2} \sin^{2m} x dx = \frac{2m-1}{2m} \frac{2m-3}{2m-2} \frac{2m-5}{2m-4} \cdots \frac{3}{4} \frac{1}{2} \frac{\pi}{2}.$$

$$\int_0^{\pi/2} \sin^{2m+1} x dx = \frac{2m}{2m+1} \frac{2m-2}{2m-1} \frac{2m-4}{2m-3} \cdots \frac{4}{5} \frac{2}{3} 1.$$

So (check you are not dividing by zero)

$$\frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} = \frac{(2m-1)(2m+1)}{(2m)^2} \frac{(2m-3)(2m-1)}{(2m-2)^2} \cdots$$

$$\cdots \frac{3 \times 5}{4^2} \frac{1 \times 3}{2^2} \frac{\pi}{2}.$$

$$\frac{\pi}{2} = \frac{2^2}{1 \cdot 3} \frac{4^2}{3 \cdot 5} \frac{6^2}{5 \cdot 7} \cdots \frac{(2m-2)^2}{(2m-3)(2m-1)} \frac{(2m)^2}{(2m-1)(2m+1)}$$

$$\times \frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx}.$$

We shall now show that as $m \to \infty$;

$$\frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} \to 1. \qquad (\bullet)$$

It will then follow that

$$\frac{\pi}{2} = \lim_{m \to \infty} \frac{2^2}{1 \cdot 3} \frac{4^2}{3 \cdot 5} \frac{6^2}{5 \cdot 7} \cdots \frac{(2m-2)^2}{(2m-3)(2m-1)} \frac{(2m)^2}{(2m-1)(2m+1)}.$$

This is called Walli's product.

$$\frac{\pi}{2} = \lim_{m\to\infty} \frac{2^{2m}(m!)^2}{3^2\cdot 5^2\cdots(2m-1)^2(2m+1)} = \lim_{m\to\infty} \frac{2^{4m}(m!)^4}{[(2m)!]^2(2m+1)}.$$

Or

$$\sqrt{\frac{\pi}{2}} = \lim_{m\to\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{(2m+1)}}$$

Or

$$\sqrt{\pi} = \lim_{m\to\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{(m+1/2)}}$$

Since $\sqrt{m}/\sqrt{m+1/2} \to 1$. we can also write the above neatly as

$$\sqrt{\pi} = \lim_{m\to\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{m}}$$

This is called Walli's formula for $\sqrt{\pi}$.

Let us now prove $(\bullet)$.
Observe that for $0 \le x \le \pi/2$, we have $0 \le \sin x \le 1$; so that

$$\sin^{2m+1} x \le \sin^{2m} x \le \sin^{2m-1} x$$

Hence

$$\int_0^{\pi/2} \sin^{2m+1} x dx \;\le\; \int_0^{\pi/2} \sin^{2m} x dx \;\le\; \int_0^{\pi/2} \sin^{2m-1} x dx$$

All quantities being positive,

$$1 \;\le\; \frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} \;\le\; \frac{\int_0^{\pi/2} \sin^{2m-1} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx}$$

Using the recurrence relation obtained at the beginning, the above is same as saying

$$1 \;\le\; \frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} \;\le\; \frac{2m}{2m-1}$$

proving $(\bullet)$.

Stirling formula for $n!$:

There is another useful mathematical formula, better to discuss while we are visiting Walli.

it says that $n!$ is like $\sqrt{2\pi}e^{-n}n^{n+1/2}$.
This is to be interpreted in the following sense. Their ratio converges to one. When we say that a sequence (a_n) is like another sequence (b_n) (both are sequences of strictly positive numbers) there are two ways of understanding. Let us assume that our sequences are increasing to infinity.

Either $(a_n - b_n) \to 0$ or $(a_n/b_n) \to 1$.

Of course when the first happens, then the second also happens. However the other way is not in general true. [it is important that our sequences are diverging to infinity. Otherwise, you can take $a_n = 1/n^2$ and $b_n = 1/n^3$. Since both converge to zero you see $a_n - b_n \to 0$ as well. However a_n/b_n equals n.]

For example (n) is like $(n + 1/n)$ in the first sense and hence also in the second sense. the sequence (n^2) is like $(n^2 + n)$ in the second sense, but not so in the first sense. In fact their difference is n which becomes larger and larger. But then in what sense are they like each other? well, Both numbers are becoming large, when you replace one by the other, the relative error (relative to the quantity you are measuring) is small.

If you are measuring length of this room, if you are off by hundred meters then the error is indeed very huge. On the other hand if you are measuring distance (of earth) to sun, if you are off by a mile or even hundred miles, the error is very very small. So the absolute error is many times unimportant and it is the relative error that matters. Think.

Returning to our problem, we need to show

$$\frac{n!}{\sqrt{2\pi}e^{-n}n^{n+(1/2)}} \to 1 \qquad \text{or} \qquad \frac{n!}{e^{-n}n^{n+(1/2)}} \to \sqrt{2\pi}$$

This is achieved in two steps: first show limit exists and non-zero. Next show (using a test case) that it must the right side.

You see we have expression like $e^{-n}n^n$. To understand it, take logarithm, then this becomes: $n \log n - n$. This should remind you $x \log x - x$ which is integral of $\log x$. The proof of Stirling in just a clever approximation of the area under the curve $f(x) = \log x$ from $x = 1$ to $x = n$.

The function $\log x$ is a concave function so that any chord is below the curve and any tangent is above the curve. Thus, for $k \geq 1$,

the area under the curve $y = \log x$ from k to $k+1$ is in between
the area under the chord joining $(k, \log k)$, $(k+1, \log(k+1))$ and
area under the tangent at $x + (1/2)$ above $(k, k+1)$.

Draw graphs and see. Thus

$$\frac{1}{2}\log(k+1) + \frac{1}{2}\log k \leq \int_k^{k+1} \log x \, dx \leq \log(k+1/2).$$

Adding these for $k = 1, 2, \cdots, n-1$ and remembering that $x \log x - x$ is a primitive for $\log x$ we get

$$\log(n!) - \frac{1}{2}\log n \leq n \log n - n + 1 \leq \sum_1^{n-1} \log(k+1/2).$$

Let

$$a_n = n \log n - n + 1 - [\log(n!) - \frac{1}{2}\log n] = \log\left\{\frac{e^{-n} n^{n+1/2}}{n!}\right\} + 1.$$

Then a_n is the area between the curve $y = \log x$ and the 'chords' explained above, from $x = 1$ to $x = n$ Thus we see

$$a_n \geq 0; \qquad a_n \uparrow . \qquad\qquad (\spadesuit)$$

Also

$$\begin{aligned}
a_n &\leq \sum_1^{n-1} \{\log(k+1/2) - \frac{1}{2}\log(k+1) - \frac{1}{2}\log k\}. \\
&= \frac{1}{2}\sum_1^{n-1} \left\{\log\frac{(k+1/2)}{k} - \log\frac{(k+1)}{k+1/2}\right\} \\
&\leq \frac{1}{2}\sum_1^{n-1} \left\{\log(1+\frac{1}{2k}) - \log(1+\frac{1}{2(k+1/2)})\right\} \\
&\leq \frac{1}{2}\sum_1^{n-1} \left\{\log(1+\frac{1}{2k}) - \log(1+\frac{1}{2(k+1)})\right\} \\
&= \frac{1}{2}\log\frac{3}{2} - \frac{1}{2}\log(1+\frac{1}{2n}) \;\leq\; \frac{1}{2}\log\frac{3}{2}.
\end{aligned}$$

As a consequence a_n is bounded above . So (♠) implies that a_n converges to a finite limit. Say $a_n \uparrow c$

$$\log\left\{\frac{e^{-n}n^{n+1/2}}{n!}\right\} = a_n - 1 \uparrow c - 1.$$

Or

$$\frac{e^{-n}n^{n+1/2}}{n!} \to e^{c-1}.$$

Or

$$\frac{n!}{e^{-n}n^{n+1/2}} \to e^{1-c} = k \text{ say} \quad k \neq 0.$$

Or

$$\frac{n!}{ke^{-n}n^{n+1/2}} \to 1.$$

We shall now evaluate the constant k by using the above limit in a known case, namely, Walli's product. We know

$$\frac{2^{2n}(n!)^2}{(2n)!\sqrt{n}} \to \sqrt{\pi}.$$

Suppose that we have strictly positive numbers a_n and b_n and $a_n/b_n \to 1$. If $\alpha_n \times a_n \to c$ then $\alpha_n \times b_n \to c$. This is because

$$\alpha_n \times b_n = \alpha_n \times a_n \times \frac{b_n}{a_n} \to c \times 1.$$

Similarly if $\alpha_n/a_n \to c$ then $\alpha_n/b_n \to c$. In other words we can replace a_n by b_n. As a consequence the above result of Walli implies

$$\frac{2^{2n}k^2e^{-2n}n^{2n+1}}{ke^{-2n}(2n)^{2n+1/2}\sqrt{n}} \to \sqrt{\pi}.$$

That is,

$$k = \sqrt{2\pi}.$$

Thus

$$\frac{n!}{\sqrt{2\pi}e^{-n}n^{n+1/2}} \to 1.$$

This completes proof of Stirling.

Here is proof about chord and tangent.

Let us first make a few observations which depend on the fact that $f'' = -1/x^2 \leq 0$.

Let g be a twice differentiable function on an interval (a, b) with $g'' \leq 0$.

Consider any two points $u < v$ in the interval (a, b). We claim that the chord (or secant) joining the two points $(u, f(u))$ and $(v, f(v))$ lies below the graph of f. There are several ways of seeing this. Here is one way.

First observe that the equation of the chord is

$$y \;=\; f(u) + \frac{f(v) - f(u)}{v - u}(x - u).$$

Consider any point $w \in [u, v]$. We need to show

$$f(u) + \frac{f(v) - f(u)}{v - u}(w - u) \leq f(w).$$

That is,

$$f(u) - f(w) + \frac{f(v) - f(u)}{v - u}(w - u) \leq 0.$$

Or

$$[f(u) - f(w)](v - u) + [f(v) - f(u)](w - u) \leq 0.$$

$$[f(u) - f(w)](v - u) + [f(v) - f(w) + f(w) - f(u)](w - u) \leq 0.$$

$$[f(v) - f(w)](w - u) - [f(w) - f(u)](v - w) \leq 0.$$

By MVT, there are points $\theta \in (u, w)$ and $\eta \in (w, v)$ such that $f(w) - f(u) = f'(\theta)(w - u)$ and $f(v) - f(w) = f'(\eta)(v - w)$. So we need to show

$$f'(\eta)(v - w)(w - u) - f'(\theta)(v - w)(w - u) \leq 0.$$

That is,

$$[f'(\eta) - f'(\theta)](v - w)(w - u) \leq 0.$$

First factor above is $f''(\zeta)(\eta - \theta)$ for some ζ. Now $f'' \leq 0$ and $\theta < \eta$ tell you that the first factor above is negative. Since $u < w < v$, the other two factors are positive and hence the inequality is true.

Consider any point u in the interval (a, b). We claim that the tangent (to the graph of f) at u lies above the graph.

The equation of the tangent is

$$y \;=\; f(u) + f'(u)(x - u).$$

Let us take any other point $w \in (a, b)$. We need to show

$$f(w) \leq f(u) + f'(u)(w - u). \qquad (\clubsuit)$$

$\underline{u < w}$. Need to show

$$\frac{f(w) - f(u)}{w - u} \leq f'(u).$$

But the left side is $f'(\theta)$ for some $u < \theta < w$ and since f' is decreasing (remember $f'' \leq 0$), (♣) is verified.

$\underline{w < u}$. Need to show

$$f(w) \leq f(u) - f'(u)(u - w) \qquad i.e \qquad f'(u)(u - w) \leq f(u) - f(w)$$

$$f'(u) \leq \frac{f(u) - f(w)}{u - w}$$

The left side is f' at some point below u, decreasing nature of f' now shows the above inequality and verifies (♣).

Digression: Area

There is one subtle point. I have been using arguments like: $0 \leq f \leq g$, area under g is finite and so area under f is also finite. Of course, if you draw curves you see it is immediate; simply because you daw curves that you *can*. Consider the function $g(x) = 1$ for $x \in (0,1)$ and $g(x) = 0$ for $x \notin (0,1)$. Consider $f(x) = 1$ only for irrational numbers x in $(0,1)$; and $f(x) = 0$ for all other points. Thus $f(x) = 0$ for all rational numbers in $(0,1)$ and also for every number outside of $(0,1)$. Clearly $0 \leq f \leq g$; area under g is finite. Do you think area under f is finite? Definitely not, because area under f does not make sense — f is not Riemann integrable. We decided to use only Riemann integral.

Is our earlier argument incorrect? No, it is correct, we used piece-wise continuous functions. A function f on R is piecewise continuous if R can be divided into finitely many disjoint intervals such that f is continuous on each interval except possibly at the end points. If $g \geq 0$ is Riemann integrable and if $0 \leq f \leq g$ and f is piecewise continuous, then f is Riemann integrable and $\int f \leq \int g$.

Digression: densities

We said that f is density for a random variable X if the following holds: $P(a < X < b) = \int_a^b f(x)dx$ for every $-\infty < a < b < \infty$. For example Unif$(0,1)$ random variable X, has the density $f(x) = 1$ for $0 < x < 1$ and $f(x) = 0$ for $x \notin (0,1)$. Let us consider $g(x) = 1$ for $0 \leq x \leq 1$ and $g(x) = 0$ for $x \notin [0,1]$. Clearly this is different from f. However If you recall properties of Riemann integral, then you notice the following: $\int_a^b g(x)dx = \int_a^b f(x)dx$ for every $a < b$. In other words we could have said g is also **a** density for X.

Similarly if X is a standard normal variable, then we have its density to be

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty.$$

Suppose we define

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty, x \neq 0; \qquad g(0) = 200.$$

Then you can see that $\int_a^b \varphi(x)dx = \int_a^b \psi(x)dx$ for every $a < b$. Thus you can

regard ψ also as density for the standard normal. We shall not do so because if there is a continuous density, then we can show it is unique. In such a case we take **the** continuous function as the density. Thus for standard normal, φ is the density.

In the first example of uniform$(0,1)$ it is unclear what to take, both f and g are piecewise continuous. Actually you can also consider $h(x) = 1$ for $[0 < x < 1$ and $x \neq 1/2, 1/3, 1/4, \cdots]$ and $h(x) = 0$ for all other values. You can show that h is not piecewise continuous; h is Riemann integrable; and $\int_a^b h(x)dx = \int_a^b f(x)dx$ and thus h is also a density for the uniform $(0,1)$ random variable. However we take piecewise continuous function as density if there is one. Of course, as mentioned above either f or g ca be taken as density in this case. But as you see, the only difference between them is only at the finitely many end points of appropriate intervals. But you need not worry about this because you can choose any one of them and do your calculations. Your end result of the calcularton does not depend on what you take.

Remember, in our course, we consider only piecewise continuous densities. Thus we have a partition $-\infty = a_0 < a_1 < \cdots < a_k = \infty$ such that in each interval (a_i, a_{i+1}) the density is continuous. The only vagueness is at the end points of these intervals: a_i for $0 < i < k$. There you choose what is convenient.

Expectation, variance:

We can imitate everything we did with discrete variables.

definition: Suppose X has density f. Then the expected value/mean/average of X is

$$E(X) = \int xf(x)dx \qquad \text{if} \qquad \int |x|f(x)dx < \infty$$

More generally,The n-th moment of X is

$$\mu_n = E(X^n) = \int x^n f(x)dx \qquad \text{if} \qquad \int |x|^n f(x)dx < \infty.$$

Also Variance is defined as $var(X) = E(X^2) - [E(X)]^2$.

Thus instead of multiplying value and probability and adding $(\sum x_i p_i)$ we are multiplying value and density $(xf(x))$ and integrating. Earlier interpretations of mean and variance remain valid. Moreover as earlier,

$$var(X) = \int (x-\mu)^2 f(x)dx = \int x^2 f(x)dx - \mu^2 \quad \text{where} \quad \mu = E(X).$$

Example: $X \sim Unif(a,b)$
Thus $f(x) = \frac{1}{b-a}$ for $a < x < b$ and zero otherwise. Hence

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{b+a}{2}$$

$$Var(X) = \int_a^b \frac{x^2}{b-a} dx - (\frac{b+a}{2})^2 = \frac{(b-a)^2}{12}$$

Chebyshev:

For any nonnegative random variable X and $a > 0$;

$$P(X \geq a) \leq E(X)/a$$

Proof is exactly same as in the discrete case. Note that the random variable being non-negative we can take $f(x) = 0$ for $x < 0$.

$$E(X) = \int_0^\infty x f(x) dx \geq \int_a^\infty x f(x) dx \geq a \int_0^\infty f(x) dx = aP(X \geq a)$$

Change of variables, dimension one:

We start with a theorem in integration well known to you that goes by the name of change of variable formula or substitution rule.

■ Suppose $(a.b)$ is an open interval and φ be a function on this interval onto an interval (c,d). To fix ideas and not to get worried about irrelevant things, imagine that these are bounded intervals (though they need not be).

We assume three things about φ.
(i) it is continuously differentiable, this means, it is differentiable and its derivative φ' is a continuous function on (a,b).
(ii) φ' is never zero, that is $\varphi'(x) \neq 0$ for all $x \in (a,b)$.
(iii) φ is a one-one function on (a,b) onto (c,d).
Then conclusion:
(A) φ has inverse $\psi : (c,d) \to (a,b)$ which is continuously differentiable.
(B) Let now f be a nice function on (a,b), for example bounded continuous function. Let $g(y) = f(\psi(y))$ for $y \in (c,d)$. Then the following holds:

$$\int_{(a,b)} f(x) dx = \int_{(c,d)} g(u) |\psi'(u)| du. \qquad (\spadesuit)$$

∎

This is precisely substitution rule. You look at the right side and usually say, put $\psi(u) = x$ so that $\psi'(u)du = dx$ and $f(\psi(u)) = f(x)$ to give you left side.

We want to see its use for us and generalise this to several variables. Before we do this three comments are in order.

First, I said you can imagine bounded intervals, simply because integration over unbounded intervals is defined via bounded intervals increasing to that unbounded interval. We do apply this result for unbounded intervals when needed.

Second, i said f is bounded because for unbounded functions integration is via taking increasing sub intervals on each of which the function is bounded. This does not mean general case is trivial, one needs to do work, but if you understand this case, you are in good shape. We apply for unbounded functions too.

Third point is the following. having said that φ' is continuous and never zero why did we say that the function φ is one-one. is this not a consequence? yes. By continuity, you see that φ' is either through out strictly positive or through out strictly negative. Hence it has to be one-one. I am looking into the future and did not want to take advantage of the fact that we are on R.

One trouble with calculus is that you have a luxurious life in one dimension. If you do not accept those luxuries, you see that several variable results are exactly the same as one variable case.

First, let us see the importance of the above formula for us? Suppose X is a random variable which has a density $f(x)$ which is zero for $x \notin (a,\ b)$. suppose φ is a function as stated above on $(a,\ b)$ onto $(c,\ d)$. let us define a new random variable $Y = \varphi(X)$. It is better to keep in mind that X is defined on a sample space Ω and Y is defined on the same sample space $Y(\omega) = \varphi(X(\omega))$, simply composition
Then the density of Y is given by

$$h(y) = f(\psi(y))|\psi'(y)|; \qquad y \in (c,\ d).$$

$h(y)$ is zero for points y not in $(c,\ d)$.

Why is this so? Enough to show that for any bounded interval $(\gamma,\ \delta)$

$$P\{Y \in (\gamma, \delta)\} = \int_{(\gamma,\delta)} h(y)dy = \int_{(\gamma,\delta)} f(\psi(y))|\psi'(y)|dy.$$

Since surely $X \in (a,b)$ we know $Y \in (c,d)$ surely. It is enough to take $(\gamma,\delta) \subset (c,d)$. In fact we can even take $\gamma, \delta \in (c,d)$. If for example $\gamma = c$, then you can take $c < \gamma_n < \delta$ such that $\gamma_n \downarrow \gamma = c$ and apply the result with γ_n and take limits. Do not bother about such subtleties.

Let α and β be such that $\varphi(\alpha) = \gamma$ and $\varphi(\beta) = \delta$ To fix ideas, let us assume that φ' is positive so that φ is increasing; hence $\alpha < \beta$ and (α,β) is mapped onto (γ,δ). (Similar argument applies if φ' is negative) Thus

$$\{\omega : Y(\omega) \in (\gamma,\delta)\} = \{\omega : X(\omega) \in (\alpha,\beta)\}$$

so that

$$P\{Y \in (\gamma,\delta)\} = \int_{(\alpha,\beta)} f(x)dx = \int_{(\gamma,\delta)} f(\psi(y))|\psi'(y)|dy.$$

as required. The last equality here is simply (♠).

Note that it is not necessary that φ be defined on all of R, enough if it is defined on an open set outside of which your density is zero. A subtle point arises and can not be left unanswered. *But you can safely ignore.* Suppose that there is a sample point ω such that $X(\omega)$ is not in this interval (a,b) where φ is defined. Then the above definition $Y(\omega) = \varphi(X(\omega))$ makes no sense for this sample point. However, note that

$$P\{X \notin (a,b)\} = 0$$

Thus in your sample space (which we need not see) the set

$$N = \{\omega : X(\omega) \notin (a,b)\}$$

has $P(N) = 0$. We can define $Y(\omega) = \varphi(X(\omega))$ for $\omega \notin N$ because then $X(\omega)$ belongs to the interval where φ is defined. For $\omega \in N$ put $Y(\omega) = 0$. It makes no difference (as far as density calculation is concerned) what value you put because this event is of probability zero.

Here is an example: Let X be Unif$(0,1)$ and $\varphi(x) = -\log x$ defined on $(0,1)$. Then $(a,b) = (0,1)$; $(c,d) = (0,\infty)$; φ is decreasing; $\psi(y) = e^{-y}$; $f(x) = 1$; $g(y) = f(\psi(y)) = 1$; $\psi'(y) = -e^{-y}$. All this leads to

$$g(y) = e^{-y} \quad y > 0$$

. Thus $Y \sim \exp(1)$

simple change of variables:

Suppose X is a random variable with pdf f. let $\mu \in R$. Set $Y = X + \mu$. Then the density of Y is given by

$$g(y) = f(x - \mu); \quad x \in R$$

This follows from observing that

$$P(Y \leq a) = P(X + \mu \leq a) = P(X \leq a - \mu)$$

$$= \int_{-\infty}^{a-\mu} f(x)dx = \int_{-\infty}^{a} f(x - \mu)dx.$$

where we changed the variable for the last equality. This shows that g is density of Y.
Equivalently, you can use the above change of variable formula, taking
$(a, b) = R$; $(c, d) = R$; $\varphi(x) = x + \mu$; $\psi(y) = y - \mu$; $\psi' = 1$.

More generally, suppose that X has density f. Let $\mu \in R$ and $\sigma > 0$ be numbers. Set $Y = \sigma X + \mu$. then density of Y is given by

$$g(y) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right).$$

This follows exactly as earlier taking $\varphi(x) = \sigma x + \mu$.

change of variable, dimension two:

Exactly the same result holds in two dimensions too.
Suppose $\Omega \subset R^2$ is an open region and φ is a function on this region onto a region Ω'. To fix ideas and not to get worried about irrelevant things, as earlier, imagine that these are bounded regions (though they need not be).

Let us clearly understand that now φ associates with every point of Ω a point of Ω'; which means there are two real valued functions $\varphi_1(x_1, x_2)$ and $\varphi_2(x_1, x_2)$ on Ω such that

$$\varphi(x_1, x_2) = (\varphi_1(x_1, x_2), \varphi_2(x_1, x_2)).$$

In other words $\varphi_1(x_1, x_2)$ is the first coordinate of the point $\varphi(x_1, x_2)$ whereas $\varphi_2(x_1, x_2)$ is the second coordinate of $\varphi(x_1, x_2)$. That is all. Note φ_1 and φ_2 are real valued functions on Ω.

We assume three things about φ as earlier.

(i) φ is continuously differentiable. This means the following. The real valued function φ_1 has both partial derivatives

$$\frac{\partial \varphi_1}{\partial x_1}; \qquad \frac{\partial \varphi_1}{\partial x_2}$$

at each point of Ω and they are continuous (real valued) functions on Ω. Similarly the real valued function φ_2 has both partial derivatives

$$\frac{\partial \varphi_2}{\partial x_1}; \qquad \frac{\partial \varphi_2}{\partial x_2}$$

at each point of Ω and they are continuous (real valued) functions on Ω. We denote

$$\varphi'(x_1, x_2) = \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_1}(x_1, x_2) & \frac{\partial \varphi_1}{\partial x_2}(x_1, x_2) \\ \frac{\partial \varphi_2}{\partial x_1}(x_1, x_2) & \frac{\partial \varphi_2}{\partial x_2}(x_1, x_2) \end{pmatrix}$$

Thus now φ' at each point of Ω is a 2×2 matrix.
(ii) φ' is non-singular, that is $\varphi'(x_1, x_2)$ is a nonsingular matrix for all points $(x_1, x_2) \in \Omega$. remember this is same as saying that its determinant is non zero.
(iii) φ is a one-one function on Ω onto Ω'.

Then conclusion:
(A) φ has an inverse $\psi : \Omega' \to \Omega$ which is again continuously differentiable. In other words if

$$\psi(y_1, y_2) = (\psi_1(y_1, y_2), \psi_2(y_1, y_2))$$

then both partial derivatives of each of these ψ_1 and ψ_2 are continuous functions. We denote

$$\psi'(y_1, y_2) = \begin{pmatrix} \frac{\partial \psi_1}{\partial y_1}(y_1, y_2) & \frac{\partial \psi_1}{\partial y_2}(y_1, y_2) \\ \frac{\partial \psi_2}{\partial y_1}(y_1, y_2) & \frac{\partial \psi_2}{\partial y_2}(y_1, y_2) \end{pmatrix}$$

We denote $|\psi'|$ to be the modulus of the determinant of the 2×2 matrix ψ'.
(B). Let now f be a nice real valued function on Ω, for example bounded

continuous function. Define the composed function $g(y) = f(\psi(y))$ *on* Ω' *Then the following holds:*

$$\int_{\Omega} f(x)dx = \int_{\Omega'} g(y) \ \ |\psi'(y)|dy. \qquad (\spadesuit)$$

That is

$$\int\int_{\Omega} f(x_1, x_2)dx_1dx_2 = \int\int_{\Omega'} g(y_1, y_2) \ \ |\psi'(y_1, y_2)|dy_1dy_2. \qquad (\spadesuit)$$

This is called change of variable formula or substitution rule or Jacobian formula. ψ' is called the Jacobian matrix.
You may find the change of variable formula to be a little boring, because I made the statement longer than what it is. This is just to make clear to you what exactly the hypothesis is and the meaning of the terms involved. Remember, we shall not prove everything we use. But we should be clear about what we are using. You can not say 'by such and such a theorem' without even knowing the statement of the theorem. You are excused if you do not know the proof. But you must know the meaning of what you are saying.

OK, what is the use of the change of variable?

Review of integration:

First review of bivariate integrals. You will learn, possibly with a different definition, integration of functions of two variables. But what we do now is equivalent to what you will learn in calculus **for functions we deal with in our course**. However that definition may not be same as this for all functions.

Suppose you have a function $f(x_1, x_2)$ of two variables. By definition (our definition) $\iint f(x_1, x_2)dx_1dx_2$ or simply $\int f(x)dx$ is the following number. Fix x_2, then $x_1 \mapsto f(x_1, x_2)$ is a function of just one variable x_1; integrate it using usual functions of one variable; you get a number, say $g(x_2)$ depending on the x_2 which you fixed; now you integrate this function of one variable $x_2 \mapsto g(x_2)$, you will get a number. This number is by definition $\int f(x)dx$. Natural question is whether you can fix x_1 and perform integration w.r.t. x_2 first of the function $x_2 \mapsto f(x_1, x_2)$ to get a number $h(x_1)$ and then perform $\int h(x_1)dx_1$. Yes, you can. Of course like interchange of order of summation, this interchange is not always permitted, but in our course we have only such functions for which this is valid. Do not worry.

Integrating a function φ over a rectangle or disc or a region means integrating the function f which is φ over the region under consideration and zero outside the region.

Example: Integrate the constant function 1 over the rectangle $(a,b)\times(c,d)$ bounded rectangle.

If you fix y(not using x_2) then integral is $(b-a)$ if $y \in (c,d)$; otherwise integral is zero. Now integrate this function to get the answer $(b-a)(d-c)$ — area of the rectangle.

Example: integrate the constant function 1 on the disc $D = (x^2+y^2 \leq 1)$.

If you fix $-1 \leq y \leq 1$, then the function is 1 on $-\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}$ and zero otherwise and so the integral is $2\sqrt{1-y^2}$ whereas for $y \notin [-1,+1]$ the integral is zero. Thus required integral is

$$\int f dxdy = \int_{-1}^{+1} 2\sqrt{1-y^2}dy = 4\int_0^1 \sqrt{1-y^2}dy = \pi$$

(put $y = \sin\theta$, $dy = \cos\theta d\theta$ etc) we get area of the disc.

Normal integral again:

Let

$$a = \int_{-\infty}^{\infty} e^{-x^2/2}dx$$

Then

$$a = \int_{-\infty}^{\infty} e^{-x^2/2}dx$$

$$a^2 = \int_{-\infty}^{\infty} e^{-x^2/2}dx \int_{-\infty}^{\infty} e^{-y^2/2}dx$$

Remembering $dxdy$ is integrating one by one we see

$$a^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2/2+y^2/2)}dxdy = \int f(x,y)dxdy$$

Now consider

$$\Omega = R^2 \setminus \{(x,0) : x \geq 0\}; \qquad \Omega' = (0,\infty)\times(0,2\pi)$$

$$\varphi(x,y) = (r,\theta); \qquad r = \sqrt{x^2+y^2}; \qquad \theta = \tan^{-1}(y/x)$$

$$\psi(r,\theta) = (r\cos\theta, r\sin\theta); \quad \psi'(r,\theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

$$|\psi'(r,\theta) = r; \qquad g(r,\theta)|\psi'(r,\theta) = re^{-r^2/2}$$

we see

$$a^2 = \int f = \int g|\psi'| = \int_0^\infty \int_0^{2\pi} re^{-r^2/2} = 2\pi\, 1$$

giving

$$a = \sqrt{2\pi}$$

as discovered earlier.

Comment: Here $r, \theta.$ are called polar coordinates. We removed part of x-axis to make φ well defined (what is theta at zero?) and to make it smooth (what are limits of theta as you go from above/below x-axis towards the point $(5, 0)$?).

Gamma and beta:

We shall now prove

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}; \quad i.e \quad \Gamma(a)\Gamma(b) = \Gamma(a+b)\beta(a,b) \quad (\clubsuit)$$

Take

$$\Omega = (0,\infty)\times(0,\infty); \qquad \Omega' = (0,\infty)\times(0,1)$$

$$\varphi(x_1,x_2) = (x_1+x_2, \frac{x_1}{x_1+x_2}) = (y_1,y_2); \quad \psi(y_1,y_2) = (y_1y_2, y_1(1-y_2))$$

$$\psi'(y_1,y_2) = \begin{pmatrix} y_2 & y_1 \\ (1-y_2) & -y_1 \end{pmatrix} \qquad |\psi'| = y_1$$

$$f(x_1,x_2) = e^{-x_1}x_1^{a-1}e^{-x_2}x_2^{b-1}; \qquad g(y_1,y_2) = e^{-y_1}y_1^{a-1}y_2^{a-1}y_1^{b-1}(1-y_2)^{b-1}$$

$$\iint f = \Gamma(a)\Gamma(b); \qquad \iint g|\psi'| = \Gamma(a+b)\beta(a,b)$$

Change of variable formula shows ($\clubsuit$).

Definition: Joint density of two random variables X_1, X_2 is a function $f(x_1,x_2)$ such that for any rectangle $\square = (a,b)\times(c,d)$;

$$P\{(x_1,x_2)\in\square\} = \iint\limits_{\square} f = \int_{x_2=c}^{d}\int_{x_1=a}^{b} f(x_1,x_2)dx_1dx_2.$$

This is same as saying $P(X\in\Omega) = \int\limits_{\Omega} f$ for any region Ω, not only for rectangles. Here we used $X = (X_1, X_2)$.

Transformation of densities:

Here is analogue of the result in one dimension concerning change of densioties for functions of random variables.
Suppose (X_1, X_2) have joint density given by $f(x_1, x_2)$ which is zero for points not in an open region Ω. Suppose $\varphi : \Omega \to \Omega'$ as in the change variable theorem satisfying those three conditions. Then $g(y_1, y_2)|\psi'(y_1, y_2)|$ for points in Ω' and zero for points not in Ω' is joint density of $(Y_1, Y_2) = \varphi(X_1, X_2)$.

Definition: X_1, X_2 defined on a space are independent if their joint density equals product of marginals, that is, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. More precisely the function $f_1(x_1)f_2(x_2)$ is a joint density of (X_1, X_2).
Here we used f_1 for density of X_1 and f_2 for density of X_2.

This is same as saying for any intervals (a, b) and (c, d)

$$P\{X_1 \in (a, b); X_2 \in (c, d)\} = P\{X_1 \in (a, b)\} \times P\{X_2 \in (c, d)\}$$

This is also same as saying for any two numbers u, v

$$P(X \leq u, Y \leq v) = P(X \leq u) \times P(Y \leq v)$$

Beta, Gamma again:

The earlier argument showing $\Gamma(a)\Gamma(b) = \Gamma(a + b)\beta(a, b)$ actually shows the following:
If $X_1 \sim \Gamma(a)$, $X_2 \sim \Gamma(b)$ and are independent, then the following holds: $Y_1 = X_1 + X_2$, $Y_2 = X_1/(X_1 + X_2)$ are independent and $Y_1 \sim \Gamma(a + b)$ and $Y_2 \sim \beta(a, b)$.

Orthogonal transforms and bivariate normals

suppose that X_1, X_2 are independent standard normal. Let

$$Y_1 = \frac{X_1 + X_2}{\sqrt{2}}; \qquad Y_2 = \frac{X_1 - X_2}{\sqrt{2}}.$$

By independence we see that joint density of (X_1, X_2) is

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-(x_1^2 + x_2^2)/2}$$

Here $\Omega = R^2 = \Omega'$.

$$\varphi_1(x_1, x_2) = \frac{x_1 + x_2}{\sqrt{2}}; \qquad \varphi_2(x_1, x_2) = \frac{x_1 - x_2}{\sqrt{2}}$$

$$\psi_1(y_1, y_2) = \frac{y_1 + y_2}{\sqrt{2}}; \qquad \psi_2(y_1, y_2) = \frac{y_1 - y_2}{\sqrt{2}}$$

$$\psi'(y_1, y_2) = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}; \qquad |\psi'| = 1$$

Finally

$$h(y_1, y_2) = \frac{1}{2\pi} e^{-(y_1^2+y_2^2)/2}$$

Thus Y_1, Y_2 are independent standard normal again.

In fact we can take any orthogonal transformation: $Y = AX$ where $AA^t = A^tA = I$. For example

$$Y_1 = X_1 \cos 17^o + X_2 \sin 17^o; \quad Y_2 = -X_1 \sin 17^o + X_2 \cos 17^o$$

are also standard normal and are independent
Just as in one dimension we have the following special case of transformations:

translation/scaling:

Suppose $X = (X_1, X_2)$ has joint density $f(x) = f(x_1, x_2)$. Let $\mu = (\mu_1, \mu_2) \in R^2$. Define $Y = (Y_1, Y_2) = (X_1 + \mu_1, X_2 + \mu_2)$. Then density of Y is given by $g(y_1, y_2) = f(x_1 - \mu_1, x_2 - \mu_2)$

More generally, suppose A is a non-singular 2×2 matrix and $Y = AX + \mu$ then Y has density $g(y) = f(A^{-1}(y - \mu))|\det A^{-1}|$

In particular, if we take any symmetric positive definite 2×2 matrix Σ, then taking X_1, X_2 independent standard normals and taking a positive definite symmetric matrix A such that $A^2 = AA^t = \Sigma$ and any $\mu \in R^2$; we see the following is a density function:

$$g(y) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-(y-\mu)^t\Sigma^{-1}(y-\mu)/2}; \qquad y \in R^2.$$

Random variable $Y = (Y_1, Y_2)$ with above density function is called bivariate normal with mean vector μ and covariance matrix Σ; denoted $Y \sim N_2(\mu, \Sigma)$.

change of variable, dimension k:

Exactly the same result as in one and two dimensions holds. Even though there is nothing new, I shall spell out so that (there is a non-zero chance that) you will read again and understand better.

Suppose $\Omega \subset R^k$ is an open region and φ is a function on this region onto a region $\Omega' \subset R^k$.

Now φ associates with every point of Ω a point of Ω'; which means there are k real valued functions $\varphi_1(x)$, $\varphi_2(x)$ $\cdots$ $\varphi_k(x)$ for $x \in \Omega$ such that $\varphi(x) = (\varphi_1(x) \ldots, \varphi_k(x))$. In other words $\varphi_i(x)$ is the i-th coordinate of the point $\varphi(x)$. Note φ_i are real valued functions on Ω.
We assume three things about φ as earlier.

(i) φ is continuously differentiable. This means the following. For each i, The real valued function φ_i has all k partial derivatives $\frac{\partial \varphi_i}{\partial x_j}$ for $1 \leq j \leq k$, at each point of Ω and they are continuous (real valued) functions on Ω. We denote $\varphi'(x)$ the $k \times k$ matrix of functions: (i,j)-th element is $\frac{\partial \varphi_i}{\partial x_j}$. Thus i-th row consists of the poartial derivatives of i-th function φ_i. Thus now φ' at each point of Ω is a $k \times k$ matrix.
(ii) φ' is non-singular, that is $\varphi'(x)$ is a nonsingular matrix for all points $x \in \Omega$ — equivalently, its determinant is non zero.
(iii) φ is a one-one function on Ω onto Ω'.

Then conclusion:
(A) φ has an inverse $\psi : \Omega' \to \Omega$ which is again continuously differentiable. In other words if $\psi(y) = (\psi_1(y), \ldots, \psi_k(y))$ then all the k partial derivatives of all the k functions, ψ_i are continuous functions. We denote $\psi'(y)$ the $k \times k$ matrix whose (i,j)-th element is $\frac{\partial \psi_i}{\partial x_j}$. Thus i-th row consists of the partial derivatives of i-th function ψ_i. We denote $|\psi'|$ to be the modulus of the determinant of the $k \times k$ matrix ψ'.
(B). Let now f be a nice real valued function on Ω, for example bounded continuous function. Define the composed function $g(y) = f(\psi(y))$ on Ω' Then the following holds:

$$\int_{\Omega} f(x)dx = \int_{\Omega'} g(y) \ \ |\psi'(y)|dy. \qquad (\spadesuit)$$

integration in k dimension:

Suppose $f(x) = f(x_1, x_2, \cdots, x_k)$ is a nice real valued function of k variables. We define its integral as follows. First calculate for each fixed $x_1, \ldots, x_{k-1}$:

$$\int_{-\infty}^{\infty} f(x_1, \cdots, x_{k-1}, x_k)dx_k = f_{k-1}(x_1, \cdots, x_{k-1}); \quad \text{(say)}$$

Then for each fixed $(x_1, \cdots, x_{k-2})$ calculate

$$\int_{-\infty}^{\infty} f_{k-1}(x_1, \cdots, x_{k-2}, x_{k-1})dx_{k-1} = f_{k-2}(x_1, \cdots, x_{k-2}); \quad \text{(say)}$$

Continue, continue. to get

$$\int_{-\infty}^{\infty} f_1(x_1)dx_1 = f_0 \qquad (say)$$

Observe that each integral above is integral of function of one variable only which we know. Further f_0 is a number. This number is called integral of f denoted

$$\int f \quad \text{or} \quad \int f(x)dx \quad \text{or} \quad \int\limits_{R^k} f(x)dx \quad \text{or}$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots)dx_1dx_2 \cdots dx_k$$

If you have to integrate f over a region Ω, then you iconsider $g = f$ on Ω and $g = 0$ outside Ω and integrate g as above. Thus

$$\int_{\Omega} f = \int_{R^k} g$$

where $g(x) = f(x)$ for $x \in \Omega$ and $g(x) = 0$ for $x \notin \Omega$.

You can try when f is indicator of a cube (?). As in two dimensions, for functions we ever come across in our course, the order does not matter, you can integrate variables one by one — your convenience!

translation in dimension k;

Let f be a nice function on R^k. Let $\mu \in R^k$. Define $g(x) = f(x-\mu)$. Then $\int f = \int g$. Use the change of variable with $\Omega = \Omega' = R^k$ and $\varphi(x) = x + \mu$ so that $\psi(y) = y - \mu$ and $\psi' = I$, the identity matrix. Thus

$$\int f(x)dx = \int f(x - \mu)dx$$

scaling:

Let f be a nice function on R^k. Let A be a non-singular $x \times k$ matrix. Define $g(x) = f(Ax)$. Then $\int f = |A| \int g$. Use the change of variable with $\Omega = \Omega' = R^k$ and $\varphi(x) = A^{-1}x$ so that $\psi(y) = Ay$ and $\psi' = A$. Thus

$$\int f(x)dx = \int f(Ax)|A|dx$$

Usually you say, put $Ax = y$ on right side so that $|A|dx = dy$ to get left side.

positive definite matrices:

We consider only symmetric matrices in what follows. A symmetric $k \times k$ matrix Σ is called positive definite if all its eigen values are strictly positive. You can easily show that this is same as saying the following: For each $v \in R^k$; $\langle v, \Sigma v \rangle \geq 0$ with equality iff $v = 0$. In other words Denoting $\Sigma = ((\sigma_{ij}))$; the quadratic form

$$\sum_{i,j} \sigma_{ij} v_i v_j \geq 0; \quad \text{and} \quad = 0 \text{ iff } v = 0$$

Given any such matrix, there is diagonalization: there is an orthogonal matrix P (change of basis) such that

$$\Sigma = PDP' = PDP^{-1}$$

where D is the diagonal matrix with entries, the eigen values. Consider the matrix

$$A = P\sqrt{D}P' = P\sqrt{D}P^{-1}$$

where $\sqrt{D}$ is the diagonal matrix with diagonal entries being Positive square-roots of diagonal entries of D, thus, square roots of eigen values. makes sense because these are positive. Also this is nonsingular and symmetric. This is denoted $\sqrt{\Sigma}$, notation justified because

$$A^2 = AA' = \Sigma; \quad A \text{ symmetric, positive definite.}$$

Multivariate normal:

Let now Σ be any $k \times k$ positive definite symmetric matrix. Let $\mu \in R^k$. Then we claim

$$\int_{R^k} f(x) = 1; \qquad f(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}.$$

This is very simple. By translation it suffices to prove

$$\int \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}x'\Sigma^{-1}x} = 1 \quad \text{or} \quad \int |\Sigma|^{-1/2} e^{-\frac{1}{2}x'\Sigma^{-1}x} = (2\pi)^{k/2}$$

Apply change of variable: $\Omega = \Omega' = R^k$ $\varphi(x) = Ax$ where $A = \sqrt{\Sigma}$ so that $\psi(y) = A^{-1}y$ and $\psi' = A^{-1}$. Note $\det A = (\det \Sigma)^{1/2}$. Apply to the function

$$f^*(x) = e^{-\frac{1}{2}x_1^2} e^{-\frac{1}{2}x_2^2} \cdots e^{-\frac{1}{2}x_k^2}$$

where we already know

$$\int f^* = (2\pi)^{k/2}$$

Did you realize you have evaluated a complicated integral of million variables if k is a million. and Σ is any million by million positive definite matrix.!

This function f is called multivariate normal density; $N_p(\mu, \Sigma)$-density.

It appears to me that most of you have given up (because you do not appear in class). I do not know if you are paid to learn, but I am paid to make you understand! Let me see if we should continue with this theme or change track!

I give some time so that you can familiarize with functions of several variables, integration, and get a feel for change of variable formula.

We turn to discrete random variables. Sometimes we allow random variables to take values in a set, not necessarily real values.

Definition: A random variable X is discrete if there is a countable set D such that $P(X \in D) = 1$. A random variable X is continuous if for every point x in the range of X, we have $P(X = x) = 0$. A **real valued** random variable X has density function f if for every $a < b \in R$ we have

$$P(a < X < b) = \int_a^b f(x)dx. \qquad \blacksquare$$

Thus if X is discrete there is a countable set $\{d_1, d_2, \ldots\}$ such that for each i, $P(X = d_i) > 0$ and $\sum P(X = d_i) = 1$. Indeed if $D = \{x_1, x_2, \ldots\}$ is as in the definition, then $\sum P(x = x_i) = P(X \in D) = 1$. Enumerate only those x_i with $P(x = x_i) > 0$ to get $\{d_1, d_2, \ldots\}$ as stated.
A discrete random variable may be defined on a countable set or on a uncountable sample space.

frequency of binary digits:
Let $Y \sim Unif(0, 1)$. This means its density equals 1 for $0 < x < 1$ and equals zero for other values. In particular, for any $(a, b) \subset (0, 1)$ the chances that Y belongs to this interval equals length of the interval. Thus you can think of Y as a point picked at random from the unit interval.

Recall the following theorem (and prove it):

Let $0 \leq x \leq 1$.

(i). There is a sequence $\{\epsilon_i : i \geq 1\}$ of numbers; each number is either zero or one such that

$$x = \frac{\epsilon_1}{2} + \frac{\epsilon_2}{2^2} + \frac{\epsilon_3}{2^3} + \cdots\cdots$$

(ii) Further if there are two expansions as above $(\epsilon_i : i \geq 1)$ and $(\eta_i; i \geq 1)$ then there is a $k \geq 1$ such that the following holds:
$\epsilon_i = 1$ for $i = k$; $\epsilon_i = 0$ for $i > k$; AND $\eta_i = 0$ for $i = k$; $\eta_i = 1$ for $i > k$; AND $\epsilon_i = \eta_i$ for $i < k$.
(Or $\eta_i = 1$ for $i = k$; $\eta_i = 0$ for $i > k$; AND $\epsilon_i = 0$ for $i = k$; $\epsilon_i = 1$ for $i > k$; AND $\epsilon_i = \eta_i$ for $i < k$.)

(iii) There do not exist more than two expansions. ∎

When there are two expansions, then the one ending with zeros is called terminating expansion and the other ending with all ones is called non-terminating expansion. In what follows, you choose. once and for all, one of these so that there is no ambiguity. The digits ϵ_i are binary digits of x and more specifically ϵ_i is the i-th digit of the number x.

Now let us pick a number Y at random from $(0,1)$. Let Y_i be the i-th binary digit of Y. Since Y is random, we conclude that each Y_i is also random. If you knew the proof of the above theorem on expansion, then the following is immediate.

Theorem:
Y_i takes values zero and one, each with probability 1/2. Further $\{Y_i; i \geq 1\}$ are independent random variables. ∎

As a result the WLLN applies. The average $\sum_1^n Y_i/n$ get closer and closer to $E(Y_1) = 1/2$. The SLLN tells that those sample points $\omega \in (0,1)$ for which $\frac{1}{n}\sum_1^n Y_i(\omega) \not\to 1/2$ has probability zero. This will be restated now. Let us denote $\epsilon_i(x)$ for the i-th binary digit of $x \in (0,1)$

Theorem: Borel's SLLN:
If we denote b y P the uniform$(0,1)$ probability, then

$$P\left\{x \in (0,1) : \frac{1}{n}\sum_1^n \epsilon_i(x) \to \frac{1}{2}\right\} = 1$$

Recall uniform probability is the probability with density function 1 on $(0,1)$ and zero outside. This gives length as probability for interval $(a,b) \subset (0,1)$. That is why this is also called length (even if the event is NOT interval), instead of probability. Thus length of the above set equals one. In other words, for 'almost all' numbers the frequency of digit zero is 1/2.
You can state similar result with decimal expansion: For almost all numbers in $(0,1)$ the frequency of each decimal digit equals 1/10.

Example 1: Random Walk:

Consider the set Z of all integers and a walk in this set as follows. Here is the rule: we start at zero. Every day we toss a fair coin and move one step

forward (if at 35, move to 36) if Heads up; one step backward (if at 35, move to 34) if Tails up. Of course, there is a non-zero chance of returning to zero. For example HT or TH will bring you to zero on day two. The question is: Are we sure to return to zero? is the chance of 'ever returning' one?

This is called simple symmetric random walk in one dimension. One dimension because we are moving in Z. Random walk because the motion is not deterministic and is governed by tossing coin, random mechanism. Symmetric because there is equal chance of moving in either direction. Simple because only one unit is moved at a time.

We can consider the set Z^2, pairs of integers. We start at the origin $(0, 0)$, which we still denote by 0. We select one of the axes at random and move in that direction forward or backward at random. Thus, from (a, b) we move to one of the four points selected at random: $(a-1, b)$, $(a+1, b)$, $(a, b-1)$, $(a, b+1)$. Again it is clear that there is a chance of returning to origin. Question: Are you sure to return? Is the chance of returning one?.

This is called simple symmetric random walk in two dimensions.

We can consider the set Z^3, triples of integers. We start at the origin $(0, 0, 0)$, which we still denote by 0. We select one of the axes at random and move in that direction forward or backward at random. Thus, from (a, b, c) we move to one of the six points selected at random: $(a-1, b, c)$, $(a+1, b, c)$, $(a, b-1, c)$, $(a, b+1, c)$, $(a, b, c-1)$, $(a, b, c+1)$. Again it is clear that there is a chance of returning to origin. Question: Are you sure to return? Is the chance of returning one?.

This is called simple symmetric random walk in three dimensions. In our course we refer to these as simply random walks. The above questions are dificult because the question depends on infinitely many rvs: entire future rvs.

Let us denote our position, state, on day n by X_n which obviously is a rv. Of course $X_0 = 0$. The event of interest is

$$A = \{X_n = 0 \quad \text{for some} \quad n \geq 1\}.$$

Let $p^{(n)} = P(X_n = 0)$ for $n \geq 1$. We put $p^{(0)} = 1$ simply because $P(X_0 = 0) = 1$. Clearly our event of interest A, is union of all these events over $n \geq 1$, thus the question is: Is $P(X_n = 0 \quad \text{for some} \quad n \geq 1) = 1$? Unfortunately these events are not disjoint. Let us disjointify. We define $f^{(0)} = 0$ and for $n \geq 1$,

$f^{(n)} = P(X_m \neq 0, \forall \ 1 \leq m < n; \ \ X_n = 0)$.

Thus $f^{(n)}$ is the chances of returning to zero for the first time on day n. Clearly, on day zero game started and no return has yet taken place and hence $f^{(0)} = 0$. Our event of interest is union of all these disjoint events.

Thus the question is: Is $\sum f^{(n)} = 1$?

Here is a theorem which we shall prove soon:
Theorem: $\sum f^{(n)} = 1$ iff $\sum p^{(n)} = \infty$.

The point to note is that the quantities $p^{(n)}$ are easy to calculate where as the quantities $f^{(n)}$ are not easy to calculate. The question is about $\sum f^{(n)}$ and this theorem converts the problem to a problem about the easily computable quantities. Accept this for now. Let us see how this helps us.

This reduces our problem, in one dimension, to decide whether the following series converges or not.

$$\sum \binom{2n}{n} \frac{1}{2^{2n}}$$

But how do we understand these numbers? This is where we take the help of James Stirling. Stirling's formula says $n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}$.

What is the notation? Suppose that (a_n) and (b_n) are sequences of strictly positive numbers. We use

$$a_n \sim b_n \iff \frac{a_n}{b_n} \to 1.$$

From the definition you can show the following.

$$a_n \sim b_n; c_n \sim d_n \Rightarrow a_n c_n \sim b_n d_n; \ \frac{a_n}{c_n} \sim \frac{b_n}{d_n}$$

As a result

$$\frac{2n!}{n!n!} \ \frac{1}{2^{2n}} \sim \frac{1}{\sqrt{\pi n}}$$

It is also easy to see that

$$a_n \sim b_n \implies \left(\sum a_n < \infty \leftrightarrow \sum b_n < \infty \right)$$

This can be seen by just noting that after some stage

$$\frac{1}{2} < \frac{a_n}{b_n} < 2; \quad \frac{1}{2} b_n < a_n < 2 b_n$$

Now recall $\sum \frac{1}{\sqrt{n}} = \infty$. As a result we conclude that $\sum p^{(n)} = \infty$ This solves our problem. Thus

We are sure to return to zero.

Two dimensions:

As earlier, we can argue to see

$$p^{(2n+1)} = 0$$

$$p^{(2n)} = \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!} \quad \frac{1}{4^{2n}}.$$

This is because to be at origin on day $2n$, we should have made a certain number (may be zero, may be n, but not more) of right moves and same number of left moves and in the remaining $(2n-2k)$ days we should make $(n-k)$ up moves and same number of down moves. Thus

$$p^{(2n)} = \binom{2n}{n} \sum_{k=0}^{n} \frac{n!n!}{k!k!(n-k)!(n-k)!} \quad \frac{1}{4^{2n}}.$$

$$= \binom{2n}{n}\binom{2n}{n} \frac{1}{4^{2n}} \quad \sim \quad \frac{1}{\pi n}.$$

Thus again the sum $\sum p^{(n)} = \infty$. Thus

We are sure to return to the origin.

Three dimensions:

To be at origin on day $2n$ you need to make certain number k of moves in X^+ direction and same number in X^- direction; a certain number l in Y^+ direction and same number in Y^- direction; make sure $k+l \leq n$ and in the remaining $2n-2k-2l$ days half Z^+ and half Z^- moves. Thus

$$p^{(2n)} = \sum_{0 \leq k,l,k+l \leq n} \frac{(2n)!}{k!\; k!\; l!\; l!\; (n-k-l)!\; (n-k-l)!} \quad \frac{1}{6^{2n}}$$

$$= \binom{2n}{n} \frac{1}{2^{2n}} \frac{1}{3^n} \sum_{0 \leq k,l,k+l \leq n} \left[\frac{n!}{k!\; l!\; (n-k-l)!} \right]^2 \quad \frac{1}{3^n}$$

If you have positive numbers a_i and $\max a_i \leq M$ then

$$\sum a_i^2 \leq M \sum a_i$$

Thus if

$$\max \left\{ \frac{n!}{k!\; l!\; (n-k-l)!} : 0 \leq k, l, k+l \leq n \right\} = M_n$$

then

$$\sum_{0\le k,l,k+l\le n}\left[\frac{n!}{k!\ l!\ (n-k-l)!}\right]^2\ \frac{1}{3^n}\le M_n\sum_{0\le k,l,k+l\le n}\left[\frac{n!}{k!\ l!\ (n-k-l)!}\right]\ \frac{1}{3^n}$$

But the last sum is just trinomial expansion

$$\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{3}\right)^n=1$$

Thus

$$p^{(2n)}\le\binom{2n}{n}\frac{1}{2^{2n}}\ \ \frac{1}{3^n}M_n$$

We know

$$\binom{2n}{n}\frac{1}{2^{2n}}\sim\frac{1}{\sqrt{\pi n}}$$

Remembering that the multinomial terms are the largest at

$$k=l=(n-k-l)=n/3$$

(we are careless here) we see

$$\frac{1}{3^n}M_n\le\frac{1}{3^n}\frac{n!}{(\frac{n}{3})!\ (\frac{n}{3})!\ (\frac{n}{3})!}$$

$$\sim\ \frac{1}{3^n}\ \ [\sqrt{2\pi}e^{-n}n^{n+(1/2)}]\ \ [\sqrt{2\pi}e^{-n/3}(n/3)^{(n/3)+(1/2)}]^{-3}$$

$$=\frac{1}{2\pi}\ \ 3^{3/2}\ \ \frac{1}{n}$$

where we used Stirling. Thus ultimately

$$p^{(2n)}\le\alpha_n;\qquad\qquad\alpha_n\sim Cn^{-3/2}$$

for some constant C. Since $\sum n^{-3/2}$ and hence $\sum\alpha_n$ and hence $\sum p^{(2n)}$ converges we conclude the following.

chance of NOT returning to the origin is strictly positive.

We were careless at one point: $n/3$ may not be integer and hence saying that $(n/3,\ n/3,\ n/3)$ term is largest does not make sense. How do you rectify it? Well, the argument above definitely shows that sum of all $p^{(2n)}$ where n is multiple of 3, is finite.

Consider the sum of $p^{(2n)}$ over n of the form 1 mod(3). That is, over all n of the form $3m+1$. Now the max term is attained at $(m, m, m+1)$ Do similar Stirling analysis and show sum is finite.

Then consider n of the form 2 mod(3). The trinomial term for this case attains maximum at $(m, m+1, m+1)$ and again show this sum is finite.

Thus you deduce that $\sum p^{(2n)}$ is finite. AAha, done!

We shall now prove the theorem thus completing the discussion of Random Walks. This involves interesting ideas which are valid in more generality.

Renewal Equation:

$$p^{(n)} \;=\; \sum_{k=0}^{n} f^{(k)} p^{(n-k)}; \qquad n \geq 1.$$

The first term on right side ($k = 0$) is zero simply because $f^{(0)} = 0$. for aesthetic reasons we keep it. Also later it helps in relating to convolution.

Recall $p^{(n)} = P(X_n = 0)$. In case n is odd both sides are zero and you can as well assume that n is even. However it is instructive to note that the renewal equation holds more generally, with the same argument as below.

The event $A = (X_n = 0)$ can be expressed as disjoint union of events depending on first return to zero – apart from return to zero on day n.

$$A = \bigcup_{1}^{n} A_k; \quad A_k = (X_i \neq 0, 1 \leq i < k; \quad X_k = 0; X_n = 0).$$

If you take any sample point for which $X_n(\omega) = 0$ then there must be a first $k \leq n$ such that $X_k(\omega) = 0$ and hence the above equation holds. By rules about conditional probability,

$$P(A_k) = P(X_i \neq 0, 1 \leq i < k;\ X_k = 0) P(X_n = 0 | X_i \neq 0, i < k;\ X_k = 0).$$

$$= f^{(k)}\ P(X_n = 0 | X_i \neq 0, i < k;\ X_k = 0). \qquad (\clubsuit)$$

Note that under the given condition, irrespective of whatever be X_i for $i < k$; if we know $(X_k = 0)$ holds then chances of $(X_n = 0)$ is just the following: chances of reaching zero on day n starting at zero on day k which is same as chances of reaching zero on day $(n-k)$; having started at zero initially. Thus

$$P(A_k) = f^{(k)}\ p^{(n-k)}.$$

If you do not like the conditional probability argument ($\clubsuit$), you can directly calculate the number of outcomes in the event A_k: Consider any sequence of

H/T of length k having equal number of H and T but never equal before k (there are $f^{(k)}2^k$ such) AND folow it up with any sequence of H/T of length $(n-k)$ having equal number of H and T (there are $p^{(n-k)}2^{n-k}$ such). Thus, noting $f^{(0)} = 0$;

$$p^{(n)} \;=\; \sum_{k=1}^{n} f^{(k)} p^{(n-k)} \;=\; \sum_{k=0}^{n} f^{(k)} p^{(n-k)}$$

generating functions:

Let us consider the generating functions:

$$P(s) = p^{(0)} + p^{(1)}s + p^{(2)}s^2 + \cdots + p^{(k)}s^k + \cdots$$

$$F(s) = f^{(0)} + f^{(1)}s + f^{(2)}s^2 + \cdots + f^{(k)}s^k + \cdots$$

which are defined at least for $0 \leq s < 1$. Of course $f^{(n)}$ for $n \geq 1$, being probabilities of disjoint events, F is defined for $s = 1$ as well.

Observe that, irrespective of whether $\sum p^{(n)}$ is finite or not we have,

$$\lim_{s\uparrow 1} P(s) = \sum_n p^{(n)}; \qquad \lim_{s\uparrow 1} F(s) = \sum_n f^{(n)}.$$

Since limit can be interchanged with finite sums and not always with infinite sums let us argue as follows to justify the above equality. First note that $\lim P(s)$ exists because P is increasing on $[0,1)$ Fix any $k \geq 1$. Since everything is non-negative we have $P(s) \geq \sum_{n=0}^{k} p^{(n)}s^n$. Since we have finite sum on right side, we conclude $\lim P(s) \geq \sum_{n=0}^{k} p^{(n)}$. This being true for every k, we conclude $\lim P(s) \geq \sum p^{(n)}$. Of course, for every $s < 1$, we have $P(s) \leq \sum p^{(n)}$ so that $\lim P(s) \leq \sum p^{(n)}$. Both these inequalities prove the stated equality.

Cauchy product of series:
Recall that if $\sum a_n$ and $\sum b_n$ are two series of numbers then we can define another series, Cauchy product, as $\sum c_n$ where

$$c_n = a_0b_n + a_1b_{n-1} + a_2b_{n-2} + \cdots + a_nb_0$$

Here then is Cauchy's theorem: If $\sum a_n = A$ and $\sum b_n = B$ and if at least one of these series is absolutely convergent, then the series $\sum c_n$ converges and converges to the product AB.

Now fix an s $(0 \leq s < 1)$ and consider the two series defining $F(s)$ and $P(s)$. Note they these being of positive terms they are absolutely convergent and thus we see by Cauchy's theorem and renewal equation,

$$F(s)P(s) = P(s) - 1.$$

We have here used that renewal equation is valid for $n \geq 1$, not for $n = 0$. $(f^{(0)}p^{(0)} = 0; p^{(0)} = 1.)$ Thus

$$P(s) = \frac{1}{1 - F(s)}; \quad 0 \leq s < 1. \qquad (\spadesuit)$$

Note that $F(s) < 1$ for $0 \leq s < 1$, so the above makes sense. Moreover, if $\sum f^{(n)} = 1$ then $\lim_{s\uparrow 1} F(s) = 1$ and $(\spadesuit)$ shows that $\lim_{s\uparrow 1} P(s) = \infty$ or $\sum p^{(n)} = \infty$. On the other hand if $\sum f^{(n)} = c < 1$ then $\lim_{s\uparrow 1} F(s) = c$ and $(\spadesuit)$ shows that $\lim_{s\uparrow 1} P(s) = \frac{1}{1-c} < \infty$ or $\sum p^{(n)} < \infty$.

This completes proof of the theorem.

Digression: Brownian motion:

What happens if we move faster but smaller distance; is there a limit which is in some sense a continuous motion? Yes, need to formualate carefully and this limiting continuous motion is called Brownian Motion.

Let us consider one dimension again. Let us move every $1/n$ unit of time distance of $\pm 1/\sqrt{n}$. Thus we have rvs indexed $\{X_{\frac{k}{n}} : k \geq 0\}$. Thus for each n, we have a process indexed by t running over all k/n, where $k \geq 0$. YES, we can show this has a limit and gives us a continuous process $\{X_t : t \geq 0\}$.

Assuming water is composed of molecules; assuming the pollen molecules are exhibiting motion due to the bombardment by water molecules; even-though each single hit of pollen by one water molecule shows no visible displacement, there are a large number of hits so that there is a visible displacement over a period of time. One of the fundamental discoveries of Einstein and Smoluchowski is the following: displacement during an interval is proportional to the square root of the length of the time interval!
If you recall the CLT, you will see that, in the above continuous process, the time one random variable is Normal – in the limit.
You can ask many many other questions, but we shall not.

Example 2: Chandrasekhar Model

I have a huge supply of balls, as many as I want. I have two numbers: $0 < p < 1$ and $\lambda > 0$. Here is a game i play.

I start with a box having a certain number of balls.

Every morning, I take the balls in the box and for each ball decide to keep it or throw out. Having done that I add a certain number of balls.
Question: what happens in the long run?
of course, to make sense of the question and answer, you should know the mechanism of removing and adding balls. Here is how I remove:

Take a ball, toss coin; Heads up decide to remove, Tails up decide to keep. Do this for each of the balls in the box. Thus if there are 100 balls, you toss coin 100 times, keep as many balls as the tails obtained and delete as many balls as the number of Heads obtained.
Here is the mechanism of adding the balls:
Add $P(\lambda)$ many balls. This means, select an integer in such a way that chance of selecting n is

$$e^{-\lambda}\frac{\lambda^n}{n!}; \quad n = 0, 1, 2, \cdots .$$

Having selected an integer as above add so many balls. remember there is no limit for the number of balls being added but it is some finite number. So there never are infinitely many balls in the box.
Now the mechanism is completely specified. The only thing that needs to be told is: how did the game start? With how many balls did I start on day zero?
For simplicity let us assume that we started with zero balls.
later you will see that it does not matter. Even if you decide to roll a die and start with as many balls as the face that shows up, you will have exactly the same answer as we get below.

so let $p_k(n)$ be the probability of having k balls on day n. Thus if X_n is the number of balls on day n then you agree that it is a random variable.

$$p_k(n) = P(X_n = k).$$

Thus $p_k(0)$ is one if $k = 0$ and zero if $k \neq 0$.

$$p_k(1) = e^{-\lambda}\lambda^k/k!$$

because there is nothing to remove on day one, all the balls are those added that morning.

$$p_k(2) = P(X_2 = k) = \sum_{i=0}^{\infty} P(X_2 = k, X_1 = i)$$

$$= \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} P(X_2 = k | X_1 = i)$$

Let us denote

$$p_{ij} = P(X_2 = j | X_1 = i)$$

To have j balls tomorrow you should keep certain number of balls; this number can not exceed i (what you have) and also can not exceed j (what you want to have) and then add some to make total j. Thus

$$p_{ij} = \sum_{l=0}^{i \wedge j} \binom{i}{l} q^l p^{i-l} e^{-\lambda} \lambda^{j-l} \frac{1}{(j-l)!}.$$

In passing let us note that p_{ij} is also $P(X_{n+1} = j | X_n = i)$ whatever be n; because we use same mechanism each day.

Returning to earlier calculation

$$p_k(2) = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \sum_{l=0}^{i \wedge k} \binom{i}{l} q^l p^{i-l} e^{-\lambda} \lambda^{k-l} \frac{1}{(k-l)!}.$$

$$= e^{-\lambda}\, e^{-\lambda}\, \lambda^k\, \frac{1}{k!} \sum_{l=0}^{k} \binom{k}{l} q^l \sum_{i=l}^{\infty} p^{i-l} \lambda^{i-l} \frac{1}{(i-l)!}$$

$$= e^{-\lambda}\, e^{-\lambda}\, \lambda^k\, \frac{1}{k!}\, (1+q)^k\, e^{\lambda p}$$

$$= e^{-\lambda(1+q)}\ [\lambda(1+q)]^k / k!.$$

Exactly the same argument shows

$$p_k(3) = e^{-\lambda(1+q+q^2)}\ [\lambda(1+q+q^2)]^k / k!.$$

$$p_k(n+1) = e^{-\lambda(1+q+\cdots+q^n)}\ [\lambda(1+q+\cdots+q^n)]^k / k!.$$

Lo and behold

$$p_k(n) \to e^{-\lambda/p}\, (\lambda/p)^k\, \frac{1}{k!}.$$

Thus in the long run you expect to have $P(\lambda/p)$ many balls in the box. In the long run, as usual means: rigorously 'as $n \to \infty$' and in practice 'for all large n'.

This model is called Chandrasekhar model. What is this model for and why are we removing and adding balls?

Well, actually balls are not balls and box is not box and day is not day and we are doing nothing either. Then is this all a hoax?

Imagine a Huge glass vessel filled with liquid. You put some coloured suspended particles (simply referred to as particles) into it. You fix, a small piece of volume in the middle of the vessel. This small piece of volume is our box (not the big thing). Day is not a day, but (1/200)th of a second. During every time interval, some of the particles (in the small volume, I fixed my attention on) leave that volume. Of course, some particles, from the Jar outside this volume, enter during this period.
What we discussed is precisely modelling this phenomenon. This model is due to Astrophysicist S Chandrasekhar. Data was actually collected, he was testing Brownian motion calculations in the theory of molecular fluctuations. Why Poisson for the number of particles entering the volume under focus? There are so many particles in the Huge jar, each having a small chance of entering our region. Now you know what should be a good model for the number of particles entering.

Example Card Shuffling:
Consider usual deck of 52 cards arranged in a stack, say top to bottom cards numbered 1 to 52. Pick two numbers (with replacement) at random, $1 \leq i, j \leq 52$. Interchange the cards at positions i, j. You have a new arrangement of the stack. Of course, if $i = j$ nothing changes. So be it. Repeat for ever.
Question: What happens in the long run?
It will be a well shuffled deck! Let us denote by X_n the stack arrangement on day n. Thus values of X_n are not real numbers, but Permutations of the deck, equivalently, elements of S_{52}. We can show

$$P(X_n = \pi) \to \frac{1}{52!} \qquad \forall\, \pi \in S_{52}$$

Thus the deck will be in random order.
You can select i, j without replacement. Result is same.
If you want to change only one card, you can do too as follows: Given a stack, select $1 \leq i \leq 52$ at random. Put top card (position 1) at position i; do NOT change others. the final conclusion is same!
This technique is very useful because many times you want to pick an element at random from a given finite set.

Example: Ehrenfests
Here is a very instructive dynamics of historical importance. Start with a box having two compartments H, C. I have 2000 balls numbered 1 to 2000.

I toss a fair coin. Heads up I put 1900 balls in H and others in C. If Tails, I put 1901 balls in H and others in C. Here is how we continue the game. We pick a number from $S = \{1, 2, \cdots, 2000\}$ at random, see where ball of that number is and move it to the other compartment. Here X_n is the number of balls in the compartment H after n exchanges. Thus $X_n \in S$. Once you know how many balls are in H, you know $2000 - X_n$ are in the compartment C.
From a state i we move to $(i-1)$ (if you have selected one of these i balls and so) with probability $i/(2R)$; you move to $(i+1)$ (if you have selected ball from other compartment and so) with probability $(1 - \frac{i}{2R})$. Repeat.
Question: What happens in the long run?
This model was proposed by the physicists, Ehrenfests (Paul Ehrenfest and Tatiana Ehrenfest) to clarify and explain subtle points in the phenomenon of heat exchange. The compartment H has hot milk and C has cold water. The balls are the molecules (and hence R is a huge number). Exchanging ball from M to W signifies a fast moving milk molecule colliding with a slow moving water molecule and thus increasing its momentum. Similarly exchanging ball from W to M signifies a water molecule doing the same. Keep in mind the exchange of heat is not a one way process; a water molecule which gained momentum earlier may now bump into a slow moving milk molecule and impart momentum. The number of balls in a compartment signifies its temperature.

Just as there was a steady state in the Chandrasekhar model, here too you can show (in some sense) there is a steady state: In the limit it appears as if the balls are distributed at random into the two compartments. Thus limiting distribution of number of balls in H is binomial$(2000, 1/2)$.

$$P(X_n = k) \to \binom{2000}{k} 2^{-2000} \qquad\qquad 0 \le k \le 2000$$

Hence on the average there are 1000 balls in each compartment. Remembering number of balls signifies the temperature of the compartment, we see both water and milk reach a common temperature.

In all these examples there is a set Z or Z^2 or Z^3 or N or S_{52} etc. There is dynamics/motion in the set. We are told how to start and we are told how to move. These are the systems we study now.

All the previous examples have the following commonality.

There is a finite or countable set, to be called, State Space S. There is a probability μ on S – $\{\mu(i), i \in \text{S}\}$. This tells you how to start on day zero. Start at i with probability $\mu(i)$. This is called initial distribution. For each state i, there is a probability $\{p_{ij} : j \in S\}$. This tells that on any day if you are at i, then (do not ask how did you come to i) move to j with probability p_{ij}. These numbers p_{ij} are arranged as a $S \times S$ matrix: $P = ((p_{ij}))_{i,j \in S}$. This P is called transition matrix. We sometimes write $p(i,j)$ instead of p_{ij}.
We use P for probability of events as well as for transition matrix but there would be no confusion.

Definition: A sequence of random variables $\{X_n, n \geq 0\}$ is called markov chain with initial distribution μ and transition matrix P if

$$P(X_0 = i) = \mu(i) \qquad i \in S$$

and

$$P(X_{n+1} = j \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = p_{ij}, \qquad n \geq 0;\ i, j \in S$$

Thus the dynamics is that on any day if you are at i you move to j with probability p_{ij}.

♦ For one dimensional random walk,
$S = Z$; and $\mu(0) = 1$.
$p_{i,\imath\pm1} = 1/2$ $p_{ij} = 0$ for $j \neq i \pm 1$.

♦ For two dimensional random walk
$S = Z^2$. $\mu\{(0,0)\} = 1$
$p_{(i,j),(k,l)} = 1/4$ if $|i-k| + |j-l| = 1$ and zero otherwise.

♦ For Chandrasekhar model
$S = \{0, 1, 2, \ldots\}$ $\mu(0) = 1$

$$p_{ij} = \sum_{l=0}^{i \wedge j} \binom{i}{l} q^l p^{i-l}\ e^{-\lambda} \lambda^{j-l} \frac{1}{(j-l)!}$$

♦ For Ehrenfests model
$S = \{0, 1, 2, \ldots, 2000\}$ $\mu(1900) = \frac{1}{2} = \mu(1901)$

$$p_{i,i+1} = \frac{2000 - i}{2000} \qquad p_{i,i-1} = \frac{i}{2000}$$

♦ For card shuffling
$Z = S_{52}$ the set of permutations of $\{1, 2, \ldots, 52\}$ OR of the 52 cards.
understanding is π means the vertical stack:
top card $\pi(1)$ and bottom card $\pi(52)$.
$\mu(\pi_0) = 1$ for some permutation.
For the i, j interchange shuffle when i, j are selected with replacement, the transition matrix is:
$p_{\pi,\pi} = 1/52$
$p_{\pi,\eta} = 2/(52 \times 51)$ if there is $i_0 \neq j_0$ such that $\eta(i_0) = \pi(j_0)$; $\eta(j_0) = \pi(i_0)$ and $\eta(i) = \pi(i)$ for $i \neq i_0, j_0$.
For the top to random shuffle
$p_{\pi,\eta} = 1/52$ if there is an i_0 $(1 \leq i_0 \leq 52)$ such that
$\eta(i_0) = \pi(1)$ and $\eta(i) = \pi(i+1)$ for $i < i_0$ and $\eta(i) = \pi(i)$ for $i > i_0$.
Note that if $i_0 = 1$ then there is no i with $i < i_0$ and $\eta = \pi$.

We do not worry as to where the random variables are defined. Our interest is to calculate probabilities and say something interesting (and of immense use). Here is how we calculate.

(i) $P(X_0 = i_0) = \mu(i_0)$. this follows from definition.

(ii) $P(X_0 = i_0, X_1 = i_1) = \mu(i_0)p(i_0, i_1)$ because
left side equals $P(X_0 = i_0)P(X_1 = i_1 \,|\, X_0 = i_0)$

(iii) $P(X_0 = i_0, X_1 = i_1, X_2 = i_2) = \mu(i_0)p(i_0, i_1)p(i_1, i_2)$
because left side equals
$P(X_0 = i_0)P(X_1 = i_1 \,|\, X_0 = i_0)P(X_2 = i_2 \,|\, X_0 = i_0, X_1 = i_1)$

(iv) In general for any n

$$P(X_0 = i_0, \cdots X_{n-1} = i_{n-1}, X_n = i_n) = \mu(i_0) \prod_{m=1}^{n} p(i_{m-1}, i_m)$$

(v) For any $n \geq 0$ we have $P(X_{n+1} = j \,|\, X_n = i) = p(i, j)$.
You can dispose this by saying that it is a consequence of the definition of the dynamics. But remember the dynamics only tells you
$P(X_{n+1} = j \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = p(i, j)$.
However, since we explained how probabilities are to be calculated, it is better to see directly that the definition of conditional probability gives the

formula above. Left side equals by definition

$$\frac{P(X_{n+1}=j, X_n=i)}{P(X_n=i)}$$

$$= \frac{\sum P(X_{n+1}=j, X_n=i, X_{n-1}=i_{n-1}, \dots X_0=i_0)}{\sum P(X_n=i, X_{n-1}=i_{n-1}, \dots X_0=i_0)}$$

where the sum in both numerator and denominator is over $(i_0, \dots, i_{n-1})$.

$$\frac{\sum \mu(i_0)p(i_0,i_1)\cdots p(i_{n-1},i)\ \ p(i,j)}{\sum \mu(i_0)p(i_0,i_1)\cdots p(i_{n-1},i)} = p(i,j)$$

For this reason the matrix $P = ((p_{ij}))$ is also called the one step transition matrix. Let us denote the powers of this matrix by $P^2, P^3, \dots$. Also let us denote the entries of P^m by $p^{(m)}(i,j)$ or $p_{ij}^{(m)}$. We see now that P^2 is the two step transition matrix and in general P^m is the m-step transition matrix.

(vi) For any $n \geq 1$ we have $P(X_{n+1}=j \,|\, X_n=i, X_{n-1}=i_{n-1}) = p(i,j)$. More generally if $m_1 < m_2 < \cdots < m_l < n < n+1$ then
$P(X_{n+1}=j \,|\, X_{m_1}=i_1, X_{m_2}=i_2, \dots X_{m_l}=i_{m_l}, X_n=i) = p(i,j)$.
The proof is exactly as above. Thus any exact information about the past is irrelevant if you have yesterday's information. You can in fact strengthen this to say among all past information, the most recent information is enough. Just keep in mind we need exact information about the state, not like: the sate on day 2 is this or that. Suppose $m_1 < m_2 < \cdots < m_l < m < n+1$
$P(X_{n+1}=j \,|\, X_{m_1}=i_1, X_{m_2}=i_2, \dots X_{m_l}=i_{m_l}, X_m=i)$
$= P(X_{n+1}=j \,|\, X_m=i)$.

(vii) For any $n \geq 0$ we have $P(X_{n+2}=j \,|\, X_n=i) = p^{(2)}(i,j)$.
This is immediate from previous result

$$P(X_{n+2}=j \,|\, X_n=i) = \sum_k P(X_{n+2}=j, X_{n+1}=k \,|\, X_n=i)$$

$$= P(X_{n+1}=k \,|\, X_n=i)P(X_{n+2}=j \,|\, X_n=i, X_{n+1}=k) = \sum_k p_{ik}p_{kj}$$

as stated.

(viii) For any $m \geq 1$ and any n; $P(X_{n+m}=j \,|\, X_n=i) = p_{ij}^{(m)}$, the ij-th element of the matrix P^m.
in other words P^m is the m-step transition matrix; its i-th row tells you where you are after m days if you are at i today. The proof is exactly as

above. We define P^0 to be the identity matrix.

(ix) Chapman-Kolmogorov equation:

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)} \qquad m, n \geq 0; \quad i, j \in S$$

We shall now define two important concepts.
This follows from definition.

Definition:
A state i leads to j, if for some $m \geq 1$, $p^{(m)}(i,j) > 0$; symbols: $i \rightsquigarrow j$.
A chain is irreducible if $i \rightsquigarrow j$ for all i, j.
For a state i, its period $d(i) = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$, if this set is non-empty.
The chain is aperiodic if period equals one for all states.

Thus a chain is irreducible if any two states communicate $i \rightsquigarrow j$ and $j \rightsquigarrow i$. It is enough to say any to different states communicate; because then Chapman-Kolmogorov tells that this happens even when $i = j$. Also by the C-K equations the set defining $d(i)$ has the property that if m, n are in the set then $m + n$ is also in the set.

Here are two basic theorems (we do no more):

Theorem 1: For a finite state chain with transition matrix P

$$\frac{I + P + \cdots + P^{n-1}}{n} \to Q$$

for some stochastic matrix Q. Further $QP = PQ = QQ = Q$. ∎

Theorem 2: For a finite state irreducible aperiodic chain with transition matrix P,

$$P^n \to \Pi$$

for some stochastic matrix Π, with all rows same. Thus if π denotes its first row, then all rows equal π. Further
(i) $\pi P = \pi$.
(ii) $P(X_n = j \,|\, X_0 = i) \to \pi(j)$ whatever be i.
(iii) If $X_0 \sim \pi$ then $X_n \sim \pi$ for all n. More generally for all n, m, we have $(X_0, \ldots, X_m) \sim (X_m, X_{m+1}, \ldots, X_{n+m})$ if $X_0 \sim \pi$.
(iv) For almost all sample points,

$$\frac{\#\{0 \leq m \leq n-1 : X_m = j\}}{n} \to \pi(j)$$

(v) If $P(X_0 = i) = 1$, then there is surely there is an m such that $X_m = i$ and the expected return time equals $1/\pi(i)$. ∎

Here is explanation of theorem 2.

(i) says that π is a left eigen vector for P corresponding to the eigen value 1. Recall P being stochastic, 1 is an eigen value with right eigen vector e consisting of all ones. This π is 'the' left eigen vector. Thus this is a purely linear algebra result.
(ii) says that π is the final equilibrium distribution, no matter where you start. Thus there is no dependence on initial data. This is also called steady state distribution.
(iii) says that if your initial distribution is π then it remains so for any day, In other words it is stationary. Thus it is called stationary initial distribution. Not only every day it remains so, any stretch appears stationary. The distribution on any consecutive five days looks same – does not depend on when you start counting these five days.
(iv) says that the proportion of time you spend in each state is given by this vector. Thus $\pi(i)$ is the proportion of time you spend in state i, no matter how you started.
(v) says that if you started in state i, you will surely return. Suppose that $f^{(n)}$ for $n \geq 1$ denotes the probability of first return on day n; that is

$$f^{(n)} = P(X_n = i, X_m \neq i \ \ \forall 1 \leq m < n \,|\, X_0 = i$$

Then the expected return time: $\sum nf^{(n)}$. equals $1/\pi(i)$.

Now you see the importance of the vector. Also, from practical point, if you simulate and run the chain for a long time, say N, then what you see as X_N will be a sample from the distribution π. That is $P(X_N = i)$ is 'nearly' $\pi(i)$. In particular, if π is uniform probability $\pi(i) = 1/|S|$, then what you see as X_N is a point picked at random from S.

Of course in the irreducible, aperiodic case you can get theorem 1 from theorem 2. This simply Cesaro's theorem or numbers: if $a_n \to a$ then $(a_1 + a_2 + \cdots + a_n)/n \to a$. The interesting point is that theorem 1 has no hypothesis. Obviously, we can not say that the limit matrix has identical rows in theorem 1, it is false in general. Both theorems are 'largely' true even in the infinite state space situation also. Unfortunately, you can not then say Q is a stochastic matrix; it could be the all-zero matrix. Similarly with π.
In case of finite state space the hypothesis of theorem 2 is same as saying that some power of P has all entries strictly positive.

Proof of theorem 1 is simple.

Let us denote

$$Q_n = \frac{I + P + \cdots + P^{n-1}}{n}$$

Since there are finitely many sequences, $\{q_n(i,j) : n \geq 1\}$, one for each pair (ij) in the finite state space we can take common convergent subsequence. Thus select $n_1 < n_2 < n_3 < \cdots$ such that

$$Q_{n_1}, Q_{n_2}, Q_{n_3}, \cdots$$

converges to say Q. Note that

$$PQ_n = \frac{P + P^2 + P^3 + \cdots P^n}{n} = Q_n - \frac{I}{n} + \frac{P^n}{n}$$

writing the above equation for our subsequence

$$PQ_{n_r} = Q_{n_r} - \frac{I}{n_r} + \frac{P^{n_r}}{n_r}$$

and taking limits we get $PQ = Q$. We used the fact that entries of P^n are all bounded between zero and one.

Similarly $QP = Q$. This in turn tells $QP^n = Q$ and taking averages we get $QQ_{n_r} = Q$, now taking limits we see $QQ = Q$. Thus

$$PQ = QP = QQ = Q \quad (\spadesuit)$$

It remains to show that the entire sequence (Q_n) converges. Suppose that for some (i^*j^*), the entries $q_n(i^*, j^*)$ do not converge to $q_{i^*j^*}$ where $Q = (q_{ij})$ is the above sub-sequential limit matrix. So there is an $\epsilon > 0$ such that infinitely many terms are outside the interval $(q_{i^*j^*} - \epsilon, q_{i^*j^*} + \epsilon)$. Again boundedness of the sequence tells us that we can get a convergent subsequence of those infinitely many that lie outside this interval. of course they can not converge to $q_{i^*j^*}$. The upshot is that we can get a subsequence $m_1 < m_2 < m_3 < \cdots$, such that $\{q_{i^*j^*}^{(m_r)}; r \geq 1\}$ converges to a number different from $q_{i^*j^*}$. By taking a further subsequence of this, we can safely assume that the matrices $Q_{m_1}, Q_{m_2}, \cdots, Q_{m_r}, \cdots$ converge, say to Q^*.

Of course Q^* has also similar properties as Q, namely,

$$PQ^* = Q^*P = Q^* \quad (\clubsuit)$$

Using $(\spadesuit)$, we see $P^mQ = Q$ and taking averages we see $Q_{m_r}Q = Q$ and taking limits we see

$$Q^*Q = Q \quad (\bullet).$$

Using (♣), we see $Q^*P^n = Q^*$ and taking averages we see $Q^*Q_{n_r} = Q^*$ and taking limits we see

$$Q^*Q = Q^* \quad (\bullet\bullet).$$

Now (•) and (••) tell that $Q = Q^*$ which is a contradiction because they differ in the (i^*j^*)-th term.

This shows that the entire sequence Q_n converges. Since we have a finite matrix, row sum of limit equals limit row sum and hence equals one. Non-negativity is clear. Thus Q is stochastic matrix.

Before proving theorem 2, we recall some simple facts.

(1^o) The only subgroups of Z are $G = \{0\}$ and $G = gZ$ for some $g \geq 1$.

(2^o) Let $S \subset \{1, 2, \ldots\}$, finite or infinite, but non-empty. Then $gcdS$ make sense (exists).

Indeed, if S is finite you knew in high school. In any case, let m be an element of S. If it divides all elements of S, then done, remember $m \in S$. Otherwise try $m - 1$. Continue this way, there are only finitely many integers below m and you will surely stop and that will be $\gcd S$. If you reached one already then one is $\gcd S$.

(3^o) if $\gcd S = a$ then there are elements $s_1, \ldots, s_k \in S$ and integers $x_1, \ldots, x_k$ such that $a = \sum x_i s_i$.

Indeed the collection of all such elements $\sum x_i s_i$ is a group and both S and this have same gcd, but if $G = gZ$ $(g \geq 1)$ then clearly g is gcd of G.

(4^0) Suppose $S \subset \{1, 2, \ldots\}$ non-empty and has gcd 1. Assume that S has the property: $m, n \in S$ implies $m + n \in S$. Then there exists n_0 such that $n > n_0$ implies $n \in S$.

get the gcd $1 = \sum x_i s_i$ let us put x as sum of those $x_i s_i$ where $x_i > 0$ and $-y$ as sum of those $x_i s_i$ where $x_i < 0$. Thus $x - y = 1$. Because of the assumed condition on S, we see $x, y \in S$. Now let $n > y^2$ then we can write $n = dy + r$ where $0 \leq r < y$ and $d \geq y$. This last one is because $n > y^2$. Thus

$$n = dy + r1 = dy + r(x - y) = (d - r)y + rx$$

Thus n is a non-negative combination of x, y and is hence in S.

(5^o) For an irreducible chain period does not depend on the state, that is, $d(i) = d(j)$ for all i, j.

Indeed fix $i \neq j$. By irreducibility fix a, b such that $p_{ij}^{(a)} > 0$ and $p_{ji}^{(b)} > 0$. Observe that by C-K equations, $p_{jj}^{(b+a)} > 0$ and $p_{ii}^{(a+b)} > 0$. Thus $d(i)$ and

$d(j)$ both divide $a+b$. Again by C-K, we see that whenever $p_{ii}^{(n)} > 0$, then $p_{jj}^{(b+n+a)} > 0$ so that $d(j)$ divides $b+n+a$ and hence divides n. Thus $d(j)$ divides all n with $p_{ii}^{(n)} > 0$. In other words $d(j) \leq d(i)$. Similarly $d(i) \leq d(j)$. completes proof of statement.

(6^o) In an irreducible aperiodic chain, given (ij) there is n_0 such that $n > n_0$ implies $p_{ij}^{(n)} > 0$.
Indeed fix a such that $p_{ij}^{(a)} > 0$, possible since $i \rightsquigarrow j$. Fix m such that $n > m$ implies $p_{jj}^{(n)} > 0$, by earlier result. Now C-K implies $p_{ij}^{(a+n)} > 0$ and this is true for all $n > n_0$.

(7^o) A finite state chain is irreducible and aperiodic iff some power of P has all entries strictly positive
Indeed if it is irreducible and aperiodic then for fixed (ij) there is n_0 so that $p_{ij}^{(n)} > 0$ for all $n > n_0$. State space being finite there are only finitely many such pairs and so done. Conversely, if some power of P has strictly positive entries then we immediately see $i \rightsquigarrow j$ for all i, j. Also if you take any i, then the set $\{n \geq 1 : p_{ii}^{(n)} > 0\}$ includes all integers after some stage and hence must have gcd one.

In view of the above, here is a restatement of Theorem 2:

Theorem 2^*: If a finite state chain with transition matrix P is such that for some n all entries of P^n are strictly positive then $P^n \to \Pi$.

Proof of Theorem 2 is a hands-on calculation.

Let us fix any one n_0 with all entries of P^{n_0} strictly positive.and let $\epsilon > 0$ be strictly below minimum of all the entries of P^{n_0}.
Define for each n, column-wise maximum and minimums of P^n matrix.

$$m_n(j) = \min_i p_{ij}^{(n)}, \quad M_n(j) = \max_i p_{ij}^{(n)}.$$

Let us fix a state j till further orders and denote $m_n(j)$ and $M_n(j)$ by just m_n and M_n. Clearly $m_n \leq M_n$ for each n.
Claim:

$$m_1 \leq m_2 \leq \cdots \leq m_n \leq \cdots\cdots \leq M_n \leq \cdots \leq M_2 \leq M_1 \quad (\spadesuit)$$

This is seen as follows. For any i,

$$p_{ij}^{(n+1)} = \sum_k p_{ik} p_{kj}^{(n)} \leq M_n \sum_k p_{ik} = M_n$$

This being true for each i, $\max_i p_{ij}^{(n+1)} \leq M_n$. In other words $M_{n+1} \leq M_n$ for each n. Similarly

$$p_{ij}^{(n+1)} = \sum_k p_{ik} p_{kj}^{(n)} \geq m_n \sum_k p_{ik} = m_n$$

This being true for each i, $\min_i p_{ij}^{(n+1)} \geq m_n$. In other words $m_{n+1} \geq m_n$ for each n. Thus claim is proved.

Clearly $M_n \leq 1$ and $m_{n_0} \geq \epsilon$. Thus

$$M_{n_0} - m_{n_0} \leq (1 - \epsilon). \qquad (\bullet)$$

$p_{ij}^{(2n_0)} = \sum_k p_{ik}^{(n_0)} p_{kj}^{(n_0)}$
$= \sum_k (p_{ik}^{(n_0)} - \epsilon p_{jk}^{(n_0)}) p_{kj}^{(n_0)} + \sum_k \epsilon p_{jk}^{(n_0)} p_{kj}^{(n_0)}$ (add and subtract)
$\geq m_{n_0}(j) \sum_k (p_{ik}^{(n_0)} - \epsilon p_{jk}^{(n_0)}) + \epsilon p_{jj}^{(2n_0)}$ by definition of $m_{n_0}(j)$
$= (1 - \epsilon) m_{n_0}(j) + \epsilon p_{jj}^{(2n_0)}$ row sums are one.
This being true for every i, taking minimum over i, we see

$$m_{2n_0}(j) \geq (1 - \epsilon) m_{n_0}(j) + \epsilon p_{jj}^{(2n_0)} \qquad (\blacklozenge)$$

Similarly,
$p_{ij}^{(2n_0)} = \sum_k (p_{ik}^{n_0} - \epsilon p_{jk}^{(n_0)}) p_{kj}^{(n_0)} + \sum_k \epsilon p_{jk}^{(n_0)} p_{kj}^{(n_0)} \leq (1 - \epsilon) M_r(j) + \epsilon p_{jj}^{(2n_0)}$.
This being true for every i, taking maximum over i, we see

$$M_{2n_0}(j) \leq (1 - \epsilon) M_{n_0}(j) + \epsilon p_{jj}^{(2n_0)} \qquad (\blacktriangle)$$

The two inequalities ($\blacktriangle$) and ($\blacklozenge$) prove

$$M_{2n_0} - m_{2n_0} \leq (1 - \epsilon)^2. \qquad (\bullet)$$

It is. now easy to see that for every $t \geq 1$

$$M_{tn_0} - m_{tn_0} \leq (1 - \epsilon)^t. \qquad (\bullet)$$

Because of the inequality ($\spadesuit$) the above suffices to show that $p_{ij}^{(n)} \to \pi_j$, limit not depending on i. Thus there is a limit matrix and it has same rows. ∎

1. We have proved the basic theorem 2. The proof gives rate of convergence also, though not a good one.

Theorem:

$$|p_{ij}^{(n)} - \pi_j| \leq (1-\epsilon)^{\lfloor \frac{n}{n_0} \rfloor}$$

Indeed if $k = \lfloor \frac{n}{n_0} \rfloor$, then $kn_0 \leq n$. From proof of the theorem we know

$$M_{kn_o} - m_{kn_0} \leq (1-\epsilon)^k$$

and $m_{kn_0} \leq m_n \leq M_n \leq M_{kn_0}$ now tells that $M_n - m_n \leq (1-\epsilon)^k$. Since $p_{ij}^{(n)}$ and π_j are both in the interval $[m_n, M_n]$ we have the stated inequality.

2. π is left eigen vector corresponding to the eigen value one.

$$\Pi P = (\lim P^n)P = \lim P^n P = \lim P^{n+1} = \Pi$$

and hence $\pi P = \pi$. Let $\eta = (\eta_1, \ldots)$ be any vector with $\eta P = \eta$. We show $\eta = c\pi$ where $c = \sum \eta_i$. Hypothesis implies $\eta P^n = \eta$ so that $\eta \Pi = \eta$. Direct multiplication of left side shows that $(\sum \eta_i)\pi = \eta$.

3. All entries of π are strictly positive. First, since row sums of P equal one, so is it for P^n for each n. State space being finite, same is true for the limit and thus $\sum \pi_i = 1$. In particular you can fix $s \in S$ with $\pi_s > 0$. Now take any $j \in S$. Since $s \rightsquigarrow j$ Fix m such that $p_{sj}^{(m)} > 0$. Use $\pi P^m = \pi$ to see

$$\pi_j = \sum_i \pi_i p_{ij}^{(m)} \geq \pi_s p_{sj}^{(m)} > 0$$

4. We know $P(X_n = j \,|\, X_0 = i) = p_{ij}^{(n)}$, the (ij)-th element of P^n and hence converges to π_j. Thus distribution of X_n, your position on day n converges to π. More generally, let $X_0 \sim \mu$. Then $X_n \sim \mu P^n$ simply because

$$P(X_n = j) = \sum_i P(X_n = j, X_0 = i) = \sum_i P(X_0 = i) P(X_n = j \,|\, X_0 = i)$$

$$= \sum_i \mu(i) p_{ij}^{(n)} = (\mu P^n)_j$$

Thus even if $X_0 \sim \mu$ then distribution of X_n converges to π. So π is the equilibrium/steady state distribution of the chain no matter how it started.

(5) If $X_0 \sim \pi$, then as seen above $X_m \sim \pi P^m = \pi$. More generally, in this case, this leads to

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \pi(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} i_n}$$

$$P(X_m = i_0, X_{m+1} = i_1, \ldots, X_{m+n} = i_n) = \pi(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} i_n}$$

Thus whatever be m, n we have

$$(X_0, \ldots, X_n) \sim (X_m, \ldots, X_{m+n})$$

(6) Let us start from j. We show that you surely return to j. This is as follows. Get as earlier N and ϵ such that all entries of $P^N$are larger than ϵ. Recall that $P(X_{m+n} = j \,|\, X_m = i) = p_{ij}^{(n)}$ for any m, n.

$$P(X_N \neq j) = 1 - P(X_N = j) \leq (1 - \epsilon)$$

$$P(X_N \neq j, X_{2N} \neq j) = \sum_{i \neq j} P(X_N = i) P(X_{2N} \neq j \,|\, X_N = i)$$

$$\leq \sum_{i \neq j} P(X_N = i)\ (1 - \epsilon) \leq (1 - \epsilon)^2$$

Thus we can show by induction that for each m

$$P(X_N \neq j; X_{2N} \neq j; \ldots, X_{mN} \neq j) \leq (1 - \epsilon)^m \qquad (\spadesuit)$$

Define events

$$A_m = (X_N \neq j; X_{2N} \neq j; \ldots, X_{mN} \neq j) \qquad m \geq 1$$

$$A = (X_{pN} \neq j \ \ \forall p \geq 1)$$

But then $A_m \downarrow A$ and so $P(A_m) \downarrow P(A)$ and $(\spadesuit)$ implies $P(A) = 0$. In particular $P(X_n \neq j\ \forall n) = 0$.

Here $A_m \downarrow A$ means the events A_m are decreasing: $A_1 \supset A_2 \supset A_3 \supset \cdots$ and $\cap A_m = A$. When this happens, we can show $P(A_m) \downarrow P(A)$.

(7) When the chain starts from i, that is, $P(X_0 = i) = 1$ then it makes sense to define the first return time T as:
$T = n$ iff $\{X_n = i; \quad \forall (1 \leq m < n)\ X_m \neq i\}$.
and $P(T = n) = f_{ii}^{(n)}$ as was done in the Random walk discussion. Now from what was proved above, we conclude $\sum f_{ii}^{(n)} = 1$. Thus T is a legitimate random variable taking value n with probability $f_{ii}^{(n)}$ and $E(T) = \sum n f_{ii}^{(n)}$. This is denoted by m_{ii}.

One can show $m_{ii} = 1/\pi_i$ using the uniqueness of the invariant probability, but we shall not do.

(8) One can use SLLN to show that, no matter how the chain starts, the proportion of time spent in state i is π_i. That is

$$\frac{\#\{0 \le m \le n-1 : X_m = i\}}{n} \to \pi_i$$

Since X_m are random variables, they are not states themselves but functions defined on some probability space the above is interpreted as follows: for almost all sample points the above proportion converges to π_i. Thus

$$P\left\{\frac{\#\{0 \le m \le n-1 : X_m = i\}}{n} \to \pi_i\right\} = 1$$

[Recall: Decimal digits depend on the number you picked from unit interval. But for almost all choices the proportion of each digit is 1/10.] The above statement, in particular, implies that you visit each state infinitely many times.

(9) if the hypotheses of the theorem fail what would happen. Either the chain is irreducible OR it is irreducible but period is $d > 1$.

Theorem: For a finite state irreducible chain with Transition matrix P, there is a unique probability vector π with $\pi P = \pi$. Each entry of π is strictly positive. No matter how the chain starts the proportion of time spent in state i equals π_i Starting from any state i we surely return to i and the expected return time is $m_{ii} = 1/\pi_i$.

What we can not say is that $P^n \to \Pi$, a matrix. We can partition the state space $S = S_0 \cup \cdots \cup S_{d-1}$ such that

if $i \in S_r$ then $p_{ij} = 0$ unless $j \in S_{r+1}$. (♣)

Thus if you are in a state in S_r today, you can only go to some state in S_{r+1} tomorrow. We interpret $(d-1)+1 = 0$ – addition modulo d. Obviously

then the matrix P looks like

$$\begin{array}{cccccccc} & S_0 & S_1 & S_2 & \cdots & S_{d-2} & S_{d-1} \\ S_0 & 0 & P_1 & 0 & \cdots & 0 & 0 \\ S_1 & 0 & 0 & P_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{d-2} & 0 & 0 & 0 & \cdots & 0 & P_{d-1} \\ S_{d-1} & P_0 & 0 & 0 & \cdots & 0 & 0 \end{array}$$

and P^d is block diagonal matrix. Note that the matrices P_r may not be square matrices. In fact order of P_r is $|S_r| \times |S_{r+1}|$

These sets are called cyclically moving subclasses and they are uniquely determined in the following sense: If you decompose
$S = S_0^* \cup S_1^* \cup \cdots \cup S_{d-2}^* \cup S_{d-1}^*$ satisfying (♣)
then there is an r such that
$S_0^* = S_r; S_1^* = S_{r+1}; S_2^* = S_{r+2}; \ldots, S_{d-1}^* = S_{r+d-1}$
Thus your sets are same as the above, but in a different 'cyclic' order.

We shall not prove the above.

The other possibility is the chain is not irreducible. Then there are two possibilities. First is you may be able to decompose the state space into disjoint sets such that, if the chain starts in one of these sets, then it stays there and the chain is irreducible there. Then you can understand the full chain, by understanding how it behaves in each of these sets – noninteracting disjoint chains. The other possibility is that apart from the irreducible sets as above, there may be others sets with the following property: if you start from that set then eventually you leave that set and enter one of the other irreducible sets and stay there from then on. We shall not get into the details and examples for possibilities are given in the class.

Ehrenfest revisited:
All the above discussion tells you that
(i) the Ehrenfest chain is heading to a steady state: as if each ball is placed at random in the two compartments.
(ii) Each state is visited infinitely often: That is you see k balls in H infinitely many times in (almost) any run of the chain for $0 \leq k \leq 2000$
if you are not careful, you think that these two conclusions are contradictory: a steady state in (i) and a chaotic behaviour in (ii). The point is that (i) is macroscopic behaviour, distribution of the state and not any specific run of

the chain; whereas (ii) is a microscopic description in each run of the game. There is no contradiction. This was one of the sources of confusion in the initial days of the Boltzmann Theory. There are others too like: have you seen atoms/molecules? In mechanics if a force makes a particle move from A to B, then you can make it move from B to A by applying same force in reverse direction. So if wish to explain heat through mechanics of motion then can you reverse from steady state of common temperature to original mixture of hot milk and cold water?

Graph Walk:
Consider a connected (undirected) graph G on a finite set of vertices V. Eventhough what we are going to describe works perfectly well even with loops and multiple edges, let us assume that there are no loops and multiple edges. Say that two vertices are neighbours if they are joined. Here is the walk on V: From any vertex move to one of its neighbours chosen at random. Thus if d_v is the degree of v, then $p_{vw} = 1/d_v$ if w is a neighbour of v and $p_{vw} = 0$ otherwise. This is irreducible. Its invariant distribution is given by

$$p\pi_v = d_v/d; \qquad\qquad d = \sum d_v$$

That this is invariant can be easily verified by using a simple useful fact.:
For a chain with transition matrix P, if you can find a vector π with positive entries with $\pi_i p_{ij} = \pi_j p_{ji}$ for all i, j, then $\pi P = \pi$. Thus if you. normalize π then it is an invariant probability vector. Proof of this is easy.

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j$$

In the present case, $\pi_v p_{vw} = \pi_w p_{wv} = 1/d$ or zero.

Bose-Einstein again:
Consider a finite set X and a finite group G acting on X. That is, for every $g \in G$, there is a bijection $x \mapsto g \cdot x$ on X such that $g_1 g_2 \cdot x = g_1 \cdot (g_2 \cdot x)$. In particular $e \cdot x$ is the identity. map and $x \mapsto g^{-1} \cdot x$ is the inverse map to $x \mapsto g \cdot x$. Given $x \in X$ we define $G_x = \{g \in G : g \cdot x = x\}$. Gen $g \in G$ we define $X_g = \{x \in X : g \cdot x = x\}$. Thus $G_x \subset G$ and $X_g \subset X$.

Here is the chain with state space X. if at x, pick a h at random from G_x and then pick a point y at random from X_h. Move to y. What is the transition matrix? p_{xy} is the chances of picking some h from G_x multiplied by chances of picking y from X_h. Note that chances of h is zero unless $h \in G_x$.

But then the chances of y is zero unless $h \in G_y$. Thus

$$p_{xy} = \sum_{h \in G_x \cap G_y} \frac{1}{|G_x|} \frac{1}{|X_h|}$$

If $O_x = \{g \cdot x : g \in G\}$ is the orbit of x, then you know $|G| = |G_x||O_x|$ Thus

$$p_{xy} = \frac{|O_x|}{|G|} \sum_{h \in G_x \cap G_y} \frac{1}{|X_h|}$$

If we take

$$\pi_x = \frac{1}{|O_x|}$$

then you see $\pi_x p_{xy} = \pi_y p_{yx}$. Thus if we define $Z = \sum \pi_x$ and the vector $\pi = (\pi_x)$ then π/Z is the invariant probability. What will this do? This picks an orbit at random. In other words if for each x, $\pi(O_x) = \sum\{\pi(y) : y \in O_x\}$ is just $1/N$ where N is the number of orbits. Thus if you run the chain for a long time and get x, pick the orbit of x, then you have picked a point at random from the space of orbits.

Here is a special case:

$$X = \{1, 2, \ldots, 19\}^{50}$$

space of all sequences $(x_1, \ldots, x_{50})$ where each x_i is in the set $\{1, 2, \ldots, 19\}$ Take $G = S_{50}$, group of permutations of $\{1, 2, \ldots, 19\}$. Action is the following:

$$\pi \cdot (x_1, \ldots, x_{50}) = (x_{\pi(1)}, \ldots, x_{\pi(50)})$$

You can think of X as space of possible placings of 50 numbered balls in 19 numbered boxes (energy levels). But if the balls are photons then B-E tells that it is not the possible placings that are equally likely, but the orbits are equally likely. So if you want to pick a B-E configuration at random, then you run the chain above for a while and get a point x and pick the orbit of x. This does it.
This chain, called Burnside Chain was introduced by the computer scientists Jerrum-Sinclair. Its rate of convergence and properties were studied by Persi Diaconis.

Back to densities:

We shall now return to densities and discuss normal densities first.

If $X \sim N(\mu, \sigma^2)$, then it has density

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

We know $E(X) = \mu$ and $var(X) = \sigma^2$.

Consider two dimensional or bivariate normal. Have a positive definite matrix Σ and $\mu \in R^2$

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}; \quad x = (x_1, x_2) \in R^2$$

where prime stands for transpose and x is the column vector but we write in row for typographical reasons and NOT put transpose for ease in reading. It is more convenient, in fact necessary in higher dimensions, to use matrix notations. However we shall do some explicit calculations for two reasons. Firstly, you should know how the density 'looks like' (and not keep it hiding behind matrix notation) . Secondly, I want to explain conditional densities.

$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$. Since the matrix is positive definite we must have $\sigma_{11} > 0$ and we denote it by σ_1^2 (with $\sigma_1 > 0$); similarly $\sigma_{22} > 0$ and we denote it by σ_2^2 (with $\sigma_2 > 0$). Further being symmetric, $\sigma_{12} = \sigma_{21}$. Also determinant being positive, we have $\sigma_1^2\sigma_2^2 > \sigma_{12}^2$. Thus if we denote

$$\frac{\sigma_{12}}{\sigma_1\sigma_2} = \rho$$

then $-1 < \rho < 1$ and $\sigma_{12} = \rho\sigma_1\sigma_2$. Thus

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Hence

$$|\Sigma| = (1-\rho^2)\sigma_1^2\sigma_2^2$$

$$\Sigma^{-1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

Thus

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1^2\sigma_2^2\sqrt{(1-\rho^2)}}\times\ e^{-\frac{1}{2(1-\rho^2)}Q(x_1,x_2)}\qquad(\clubsuit)$$

where

$$Q(x_1,x_2)=(\frac{x_1-\mu_1}{\sigma_1})^2-2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2})+(\frac{x_2-\mu_2}{\sigma_2})^2$$

This is the bivariate normal density with means μ_1,μ_2 and variances σ_1^2,σ_2^2 and correlation ρ.
For simplicity, we take $\mu_1=\mu_2=0$ from now on. Thus

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1^2\sigma_2^2\sqrt{(1-\rho^2)}}\times\ e^{-\frac{1}{2(1-\rho^2)}Q(x_1,x_2)}\qquad(\spadesuit)$$

where

$$Q(x_1,x_2)=\frac{x_1^2}{\sigma_1^2}-2\rho\frac{x_1}{\sigma_1}\frac{x_2}{\sigma_2}+\frac{x_2^2}{\sigma_2^2}$$

This is bivariate normal density with means zero and variances σ_1^2,σ_2^2 and correlation ρ.
If variances are one

$$f(x_1,x_2)=\frac{1}{2\pi\sqrt{(1-\rho^2)}}\times\ e^{-\frac{1}{2(1-\rho^2)}Q(x_1,x_2)}\qquad(\heartsuit)$$

where

$$Q(x_1,x_2)=x_1^2-2\rho x_1x_2+x_2^2$$

This is standard bivariate normal density with correlation ρ.
If we take correlation to be zero, then

$$f(x_1,x_2)=\frac{1}{2\pi}\times\ e^{-\frac{1}{2}(x_1^2+x_2^2)}\qquad(\diamondsuit)$$

This is bivariate normal density with zero means, unit variances and zero correlation – or independent standard normals.

We shall from now on consider the bivariate with means zero, variances σ_1^2,σ_2^2 and correlation ρ – thus density ($\spadesuit$) – and justify that means etc are as stated.

Recall that in the discrete case we had joint distribution of two random variables:
X taking values $(x_i;i\geq 1)$ and Y taking values $(y_j,j\geq 1)$.

$p_{ij} = P(X = x_i, Y = y_j)$
From this we can read marginal distributions:
$p_{i\bullet} = \sum_j p_{ij} \;=\; P(X = x_i)$
$p_{\bullet j} = \sum_i p_{ij} \;=\; P(Y = y_j)$
Also for each i we can talk of the conditional distribution of Y given $X = x_i$:

$$\frac{p_{ij}}{p_{i\bullet}} = P(Y = y_j \,|\, X = x_i) \quad j = 1, 2, 3, \ldots$$

Conditional expectation of Y given $X = x_i$ is nothing but the expectation of Y w.r.t. the above conditional probability.
Similarly the conditional distribution and conditional expectation of X is defined for each given value y_j of Y.

We shall now imitate those:
Definition: If $f(x, y)$ is the joint density of X, Y; then the marginal density of X is defined as

$$f(x\bullet) = \int f(x, y)dy \qquad x \in R$$

The conditional density of Y given $X = x$ is defined as

$$f(y \,|\, x) = \frac{f(x, y)}{f(x\bullet)} \qquad y \in R$$

whenever $f(x\bullet) \neq 0$. When $f(x\bullet) = 0$ this is not defined. The conditional expectation of Y given $X = x$ is the expectation of Y w.r.t. the conditional density, that is

$$E(Y \,|\, X = x) = \int y f(y \,|\, x)dy.$$

Similarly the marginal of Y and conditional density, conditional expectation of X given $Y = y$ are defined.

Observe that marginal density of X is indeed density of X:

$$P\{X \in (a, b)\} = P\{(X, Y) \in (a, b) \times R\}$$

$$= \int_{x=a}^{b} \int_R f(x, y)dydx = \int_a^b f(x\bullet)dx$$

The marginal $f(x\bullet)$ is usually denoted as $f_1(x)$ and the marginal of Y is denoted $f_2(y)$.

Note that in the discrete the concept of conditional distribution is just an application of the notion of conditional probability:

$$P(Y = y_j \,|\, X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i\bullet}}$$

In the present case it is not so: $P(X = x, Y = y)/P(X = x)$ makes no sense. So the above definition, in the density case, is NOT a consequence of anything, it is just defined so by agreement. There are strong reasons for such an agreement, but we shall not go into because we shall not discuss much about these.
Also for each x for which $(x\bullet) \neq 0$; the function $f(y \,|\, x)$ is indeed a density function: non-negative and $\int f(y \,|\, x)dy = 1$.

Let us return to the bivariate normal density (♠). By the usual method of completing squares of the exponent, we see

$$f(x_1\bullet) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-x_1^2/2\sigma_1^2}$$

Thus $X_1 \sim N(0, \sigma_1^2)$. In particular it has mean zero and variance σ_1^2. Similarly, you see $X_2 \sim N(0, \sigma_2^2)$; has mean zero and variance σ_2^2.
The conditional distribution of X_2 given $X_1 = x_1$ is

$$\frac{f(x_1, x_2)}{f(x_1\bullet)} = \frac{1}{\sqrt{2\pi}\sqrt{(1-\rho^2)}\,\sigma_2} e^{-\frac{1}{2\sigma_2^2(1-\rho)^2}(x_2 - \rho\sigma_2\frac{x_1}{\sigma_1})^2}$$

which shows that the conditional distribution of X_2 given $X_1 = x_1$ is Normal with mean $\rho\sigma_2\frac{x_1}{\sigma_1}$ and variance $\sigma_2^2(1-\rho^2)$. In particular, given X_1 the variance of X_2 has become smaller by a factor $(1 - \rho^2)$. Further

$$\int x_1 x_2 f(x_1, x_2)dx_1 dx_2 = \int \left[\int x_2 \frac{f(x_1, x_2)}{f(x_1\bullet)} dx_2\right] x_1 f(x_1\bullet)dx_1$$

$$= \int \rho\sigma_2 \frac{x_1}{\sigma_1} x_1 f(x_1\bullet)dx_1 = \rho\sigma_2\sigma_1$$

Hence covariance between X_1, X_2 is $\rho\sigma_1\sigma_2$. As in the discrete case correlation is defined as

$$\text{correlation}(X, Y) = \rho_{XY} = \frac{cov(X, Y)}{SD(X)SD(Y)}$$

Thus correlation between X_1, X_2 is ρ.
This justifies the terminology. As a result the matrix Σ is usually called

covariance matrix, or sometimes, variance-covariance matrix. Same calculations hold good in higher dimensions. But you need to use matrix algebra. You would like to know, for example, conditional density of (X_1, X_2, X_3) given (X_4, X_5). But all this can be done with clever matrix algebra.
Incidentally what we did above can also be painlessly obtained by taking square root of Σ and constructing the bivariate normal density from independent normals.

exponentials:

Lifetime of bulbs are independent $\exp(\lambda)$. I replaced the bulb immediately after it died. The second bulb also died now at time t. I forgot the time when the earlier bulb died (that is when the present bulb was installed). When did I do this replacement?
Customers enter a bank and their inter arrival times are in independent $\exp(\lambda)$. Second customer arrived at time t. I wonder: when did the first customer arrive?
The time between successive accidents (say at Sholinganallur) are independent $\exp(\lambda)$. Now at time t second accident occurred and I wonder; when did the first accident occur?

All these problems are exactly the same. Let $X \sim exp(\lambda)$ and $Y \sim \exp(\lambda)$ be independent. Find the conditional of X given $X + Y = t$.
Joint density of (X, Y) is

$$f(x, y) = \lambda^2 e^{-\lambda(x+y)} \qquad x > 0, y > 0$$

Let us put $U = X$ and $V = X + Y$. Then by change of variable formula

$$g(u, v) = \lambda^2 e^{-v} \qquad 0 < u < v < \infty.$$

[$\Omega = \{(x, y) : x > 0, y > 0\}$
$u = \varphi_1(x, y) = x; \qquad v = \varphi_2(x, y) = x + y$
$\Omega' = \{(u, v) : 0 < u < v < \infty\}$
$\psi_1(u, v) = u; \qquad \psi_2(u, v) = v - u$
Mod determinant of Jacobian =1.]

Marginal of V is

$$g(\bullet v) = \lambda^2 v e^{-\lambda v} \qquad 0 < v < \infty.$$

Conditional density of U given $V = t$ is

$$\frac{g(u, t)}{g(\bullet t)} = \frac{1}{t} \quad 0 < u < t$$

Thus if you heard the second beep on Gieger counter at time t, and wondering when the first occurred, then the answer is: it occurred at a time chosen uniformly below t.

Suppose $X_1, X_2, \ldots$ are independent $\exp(\lambda)$ random variables – time gaps between successive beeps of Gieger counter. It makes sense to ask: when did the n-th beep occur? That is, distribution of $X_1 + \cdots + X_n$. Let us denote its density by f_n. Thus

$$f_1(x) = \lambda e^{-\lambda x} \quad x > 0; \qquad\qquad f_2(x) = \lambda^2 x e^{-\lambda x} \quad x > 0$$

You can now repeat earlier argument with $X = X_1 + X_2$ whose density is f_2 and $Y = X_3$ to get density of $X_1 + X_2 + X_3$ and so on. By induction, you see

$$f_n(x) = \lambda^n \frac{x^{n-1}}{(n-1)!} e^{-\lambda x} \qquad x > 0$$

You can wonder at time t: How many beeps were there till now? If N_t is the number of beeps upto time t, then it takes non-negative integer values. Clearly, $N_t = 0$ iff $(X_1 > t)$. Thus

$$P(N_t = 0) = \int_t^\infty f_1(x)dx = e^{-\lambda t}$$

$N_t = 1$ iff $(X_1 \leq t; X_1 + X_2 > t)$. Thus

$$P(N_t = 1) = P(X_1 \leq t) - P(X_1 + X_2 \leq t)$$

$$= \int_0^t f_1(x)dx - \int_0^t f_2(x)dx = \lambda t e^{-\lambda t}$$

(integrate f_2 by parts to end up with f_1 integral). In general,

$$P(N_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \qquad k = 0, 1, 2, 3 \ldots$$

We started with a sequence of exponential random variables and ended up with simpler discrete Poisson variables but uncountably many, one for each $t > 0$. There is a nice useful story that can be developed here but that is for an advanced course.

moments, mgf:

Just as in the discrete case we can define moments and moment generating function if they exist. The n-th moment of X with density f is

$$\mu_n = \int x^n f(x)dx; \qquad \text{defined when} \qquad \int |x|^n f(x)dx < \infty$$

When $n = 0$ we take $\mu_0 = 1$.

$$M_X(t) = E(e^{tX}) = \int e^{tx} f(x)dx$$

$M(0) = 1$ and M may not exist for other values of t. For example for $X \sim N(0, 1)$ usual completing square (in exponent) leads to

$$M_X(t) = e^{t^2/2}$$

As in the discrete case one can get Chernoff bound using mgf.

digression cf:

Since mgf need not always exist one defines characterstic function $\varphi_X(t)$ of a random variable with density f as the following complex valued function:

$$\varphi_X(t) = E(e^{itX}) = E(\cos(tX)) + i\ E(\sin(tX)) \quad t \in R$$

Remember this is defined for $t \in R$ and then $|\cos(tx)|$ and $|\sin(tx)|$ are bounded, so have expectations. and hence the defining integral always exists. The change of t to it makes a drastic difference. If $X \sim N(0, 1)$ one can show

$$\varphi_X(t) = e^{-t^2/2}$$

The observation that both density and characterstic function are (upto constant) same type: $e^{-u^2/2}$ is very interesting. It appears in both probability, Mathematics and physics (uncertainty principle).

Properties of Expectation:

In the discrete case we have proved expectation is additive and in the independent situation, it is multiplicative. It is true in density case also (it is true in all cases!). Suppose X, Y have joint density $f(x, y)$.

$$E(X + Y) = \int \int (x + y)f(x, y)dxdy$$

$$= \int x(\int f(x,y)dy)dx + \int y(\int f(x,y)dx)dy$$

$$= \int xf(x\bullet)dx + \int yf(\bullet y)dy = E(X) + E(Y)$$

Remember, this holds without independence.

Now suppose X, Y are independent. Then $f(x,y) = f_1(x)f_2(y)$ so that

$$E(XY) = \int\int xyf(x,y)dxdy = \int(\int xf_1(x)dx)yf_2(y)dy$$

$$= \int E(X)yf_2(y)dy = E(X)E(Y)$$

This in turn yields that for sum of independent random variables variances add up. Indeed if $E(X) = \mu$ and $E(Y) = \nu$

$$E[(X-\mu)(Y-\nu)] = E(XY) - \mu\nu = 0$$

so that

$$var(X+Y) = E[(X+Y-\mu-\nu)^2]$$

$$= E[(X-\mu)^2] + E[(Y-\nu)^2] + 2E[(X-\mu)(Y-\nu)]$$

$$= var(X) + var(Y)$$

distribution functions:

Are there two different theories: discrete and density? Are there others? Is there any common link? Yes. All are parts of **one** theory. For any random variable X we define distribution function:

$$F_X(a) = P(X \leq a) \qquad a \in R$$

Basic Theorem:

1. Suppose X is a random variable on a probability space. Then its distribution function $F_X = F$ satisfies

(i) monotone: $a \leq b$ implies $F(a) \leq F(b)$

(ii) right-continuous: $a_n \downarrow a$ implies $F(a_n) \to F(a)$.

(iii) $\lim F(a) = 0$ as $a \to -\infty$ and $\lim F(a) = 1$ as $a \to \infty$.

2. Conversely, given any function F satisfying the above three properties, we can make a probability space and a random variable X on that space such that the given F is F_X.

3. Let $X \sim F$. Then $P(X = a) = F(a) - F(a-)$ for any $a \in R$. The quantity $F(a) - F(a-)$ is denoted $J_F(a)$, jump of F at a.

4. Let $X \sim F$. Then X is discrete random variable iff there is a sequence of points $a_1, a_2, \ldots$ such that $\sum J_F(a_i) = 1$. In that case the distribution of X is: value a_i with probability $J_F(a_i)$ for $i \geq 1$. Count only those a_i with $J_F(a_i) > 0$.

5. Let $X \sim F$. Then X has density iff for all a we have

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

where $f(x) = F'(x)$ when F is differentiable at x and zero otherwise. In that case f is a density for X. ∎

Thus by knowing F you can understand if X is discrete or has density. If discrete you can get the values and probabilities (using jumps of F); if it has density, you can get density (using derivative of F). Remember, it is not enough to differentiate F, you must verify that the equation displayed in 5 holds.

Here is an example of F:
$F(a) = 0$ for $a < 1$;
$F(a) = 1/2$ for $1 \leq a \leq 4$;
$F(a) = \frac{1}{2} + \frac{a-4}{4}$ for $4 \leq a \leq 6$
and $F(a) = 1$ for $a \geq 6$.

You can verify this is a distribution function. If $X \sim F$, then F is not discrete random variable because sum of jumps is $1/2$ and not 1. X does not have a density, because F is not continuous.

Digression:
We discussed Cantor distribution function F which is constant in each deleted interval, $F(0) = 0$ and $F(1) = 1$ and F is continuous. If you followed the prescription in 5 above, you get $f \equiv 0$ and the displayed equation fails. Thus this is an continuous distribution function with no density.

CLT:
Recall that a sequence of random variables $(X_n, n \geq 1)$ is independent if for each n the variables $X_1, X_2, \ldots, X_n$ are independent. A sequence of random variables $(X_n, n \geq 1)$ is identically distributed if they all have the same distribution. A sequence is iid – independent identically distributed – if both

hold.
They may have density or they may all be discrete or neither. They have the same distribution function. In what follows, you keep in mind only the first two cases because we have not defined means etc in the general case.

Here is Central Limit Theorem.

Theorem:
Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with mean μ and variance $\sigma^2 > 0$. Then for any $b \in R$

$$P(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq b) \to \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

For any $a < b \in R$

$$P(a < \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < b) \to \int_{a}^{b} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \qquad \blacksquare$$

You must appreciate that no serious assumption is made about the distribution except that the mean and variance are finite. You should understand that the quantity $Z_n = (\sum X_i - n\mu)/\sigma\sqrt{n}$ is simply the sum standardized; which means make mean zero and variance one. Since $E(X_i) = \mu$ we see $E(\sum X_i) = n\mu$ so that $E(\sum X_i - n\mu) = 0$. By independence variance adds up so that $var(\sum X_i) = n\sigma^2$ and $var(\sum X_i - n\mu) = n\sigma^2$. Thus $E(Z_n) = 0$ and $var(Z_n) = 1$

There are three basic peaks in elementary probability. First is the WLLN and SLLN. They say that partial averages approach the expectation. The difference between the two versions is in the mode of approach (as explained in the coin tossing). We have proved and seen uses of WLLN. We do not see SLLN in elementary course (because, the probability space is generally unseen, we see only random variables). But you have seen one example of SLLN. when you pick a point at random from the unit interval and look at its binary digits we showed that they are iid taking values 0/1 with equal probability so that expectation equals 1/2. We have shown the averages of the digits (which is proportion of ones in the expansion) converges to 1/2 for almost all points of the unit interval. this is an instance of SLLN.

The second peak is the CLT. The third peak is LIL, Law of iterated logarithm which tells you 'exact magnitude of fluctuations (from expectation) in the averages'. This is rather complicated.

Returning to CLT, what it says is that the 'normalized partial sum' is approximately standard normal. This is very useful theorem, Firstly, as

explained in the discrete case (with coin tossing) this helps in calculating probabilities – approximating probabilities. This also appears in various statistical techniques like testing. When you are measuring something there are measurement errors. Errors occur due to several reasons and each reason causing a small error. You assume errors are not just one sided and so have mean zero. You can rewrite Z_n as

$$Z_n = \frac{X_1}{\sigma\sqrt{n}} + \frac{X_2}{\sigma\sqrt{n}} + \cdots + \frac{X_n}{\sigma\sqrt{n}}$$

You can think of Z_n as a large sum of small 'errors'. In model building, you can assume that error, which is sum of a large number of small errors caused due to diverse reasons, is approximately normal. Thus this helps justify certain assumptions in model building.
We leave this topic here.

Before going to a new idea (the last we discuss in our course), here are some diverse comments. The first three are about CLT, We have not proved it because it is an advanced topic. But Why should you believe it? Can we take a peep?

1. CLT To start with Suppose (X_n) are iid taking values ± 1 with equal probability. Thus $\mu = 0$ and $\sigma^2 = 1$. Let

$$Z_n = \frac{X_1 + \cdots X_n}{\sqrt{n}} = \frac{X_1}{\sqrt{n}} + \frac{X_2}{\sqrt{n}} + \cdots + \frac{X_n}{\sqrt{n}}$$

Now $X_1/\sqrt{n}$ has mgf

$$\frac{1}{2}[e^{t/\sqrt{n}} + e^{-t/\sqrt{n}}] = [1 + \frac{t^2}{2n} + o(1/n)]$$

where

$$o(1/n) \quad = \quad [\frac{t^4}{4!n^2} + \frac{t^6}{4!n^2} + \cdots]$$

which when multiplied by n goes to zero (as $n \to \infty$). thus mgf of Z_n is

$$M_n(t) = [1 + \frac{t^2}{2n} + o(1/n)]^n \to e^{t^2/2}$$

mgf of the standard normal.

2 CLT: Suppose (X_n) are iid taking values $0, 1$ with probabilities $q = 1-p$ and p respectively. Then $\mu = p$ and $\sigma^2 = pq$. Proceeding as above

$$Z_n = \frac{X_1 + \cdots X_n - np}{\sqrt{npq}} = \frac{X_1 - p}{\sqrt{npq}} + \frac{X_2 - p}{\sqrt{npq}} + \cdots + \frac{X_n - p}{\sqrt{npq}}$$

mgf of $(X_1 - p)/\sqrt{npq}$ equals

$$qe^{-pt/\sqrt{npq}} + pe^{qt/\sqrt{npq}} = q\exp\{-t\frac{\sqrt{p/q}}{\sqrt{n}}\} + p\exp\{t\frac{\sqrt{q/p}}{\sqrt{n}}\}$$

$$= q[1 - t\frac{\sqrt{p/q}}{\sqrt{n}} + t^2\frac{p/q}{2n} - t^3\frac{(\sqrt{p/q})^3}{3!n\sqrt{n}} + \cdots]$$

$$+p[1 + t\frac{\sqrt{q/p}}{\sqrt{n}} + t^2\frac{q/p}{2n} + t^3\frac{(\sqrt{q/p})^3}{3!n\sqrt{n}} + \cdots]$$

$$= [1 + \frac{t^2}{2n} + o(1/n)]$$

Thus mgf of Z_n equals

$$M_n(t) = [1 + \frac{t^2}{2n} + o(1/n)]^n \to e^{t^2/2}$$

mgf of the standard normal.

3 CLT: Suppose (X_n) are iid uniform $(-1, 1)$ variables so that $\mu = 0$ and $\sigma^2 = 1/3$.

$$Z_n = \frac{X_1 + \cdots X_n}{\sqrt{n/3}} = \frac{X_1}{\sqrt{n/3}} + \frac{X_2}{\sqrt{n/3}} + \cdots + \frac{X_n}{\sqrt{n/3}}$$

mgf of $X_1/\sqrt{n/3}$ equals

$$\frac{1}{2}\int_{-1}^{1} e^{tx/\sqrt{n/3}}\ dx = \frac{1}{2}\frac{e^{t/\sqrt{n/3}} - e^{-t/\sqrt{n/3}}}{t/\sqrt{n/3}}$$

denoting $t/\sqrt{n/3}$ by u

$$= \frac{e^u - e^{-u}}{2u} = [u + \frac{u^3}{3!} + \frac{u^5}{5!} + \cdots]/u$$

$$= 1 + \frac{u^2}{3!} + \frac{u^4}{5!} + \cdots$$

using $u^2 = 3t^2/n$ and $u^4 = 3^2t^2/n^2$ etc

$$= 1 + \frac{t^2}{2n} + \frac{3^2t^2}{5!n^2} + \cdots = [1 + \frac{t^2}{2n} + o(1/n)]$$

Thus mgf of Z_n is

$$M_n(t) = [1 + \frac{t^2}{2n} + o(1/n)]^n \to e^{t^2/2}$$

In each case we are calculating mgf explicitly. Firstly the mgf may not exist; secondly we may not be able to explicitly evaluate. One uses characterstic function φ_X which always exists. Moreover we dio not need all of the cf; only Taylor expansion up to second order and an 'estimate' of the error with $o(1/n)$ term. Thus you need not calculate the cf in full.
One needs a theorem that normal distribution is identified by its cf. This

shows the convergence of cf of your mormalized partial sum to the cf of standard normal. To conclude CLT, you need to know that this convergence does imply convergence of $P(a < Z_n < b)$ as stated. These involve work and are advanced topics.

4. SLLN:

As a consequence of mgf and Chebyshev inequality we have derived the very useful Chernoff bound: Let (X_n) be iid random variables taking values ± 1 with equal probability. Let

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then

$$P\left(|\frac{S_n}{n}| \geq a\right) \leq 2e^{-na^2/2} \qquad (\spadesuit)$$

Having come this far, (by the way, Chernoff bound is a bread and butter tool in CS too) It will be a shame if I did not tell you SLLN:

$$\frac{S_n}{n} \to 0 \qquad \text{for almost all sample points.} \qquad (\clubsuit)$$

Here is the catch: I said that we never ask where our variables are defined. We calculate ALL needed probabilities using the densities of random variables (that is why densities are there, as far as our course is concerned). But to prove $(\clubsuit)$ we need to pick up all sample points from our space (which we do not know) for which the above convergence fails and show that this event has probability zero, so that the complement event has probability one.

Luckily two mathematicians devised a link for this which you can implement without looking at your space!

Borel-Cantelli Lemma: Suppose (A_n) are events in a space. Suppose $\sum P(A_n) < \infty$. Define an event B As follows: all points which belong to infinitely many of the A_n. Then $P(B) = 0$.

Proof of this is so trivial you will laugh. If a point is in infinitely many A_n, then whatever m you take it must be in $\cup\{A_n : n \geq m\}$. In other words

$$B \subset \cup\{A_n : n \geq m\} \qquad \forall m$$

But

$$P[\cup\{A_n : n \geq m\}] \leq \sum_{n=m}^{\infty} P(A_n) \qquad \text{(see below (*))}$$

So

$$P(B) \leq P[\cup\{A_n : n \geq m\}] \leq \sum_{n=m}^{\infty} P(A_n) \qquad \forall m$$

But $\sum P(A_n)$ finite tells that the above tail sum can be made as small as we please. Thus

$$P(B) = 0$$

[$(*)$: We know that For countably many disjoint events $P(\cup C_i) = \sum P(C_i)$.
We know $C \subset D$ implies $P(C) \leq P(D)$.
If you take events $A_m, A_{m+1}, A_{m+2}, \ldots$, then
$C_m = A_m$
$C_{m+1} = A_{m+1} \setminus A_m$
$C_{m+2} = A_{m+2} \setminus (A_m \cup A_{m+1}), \ldots$
are disjoint and

$$C_i \subset A_i; \qquad \cup_{i \geq m} C_i = \cup_{i \geq m} A_i; \qquad C_i \quad \text{disjoint}$$

so that

$$P(\cup_{i \geq m} A_i) = P(\cup_{i \geq m} C_i) = \sum_{i \geq m} P(C_i) \leq \sum_{i \geq m} P(A_i)$$

as required]

Let us return to (♠). For every number $a > 0$ we know

$$\sum e^{-na^2/2} < \infty$$

Take $a = 1/2$ and A_n be the event $\left(|\frac{S_n}{n}| \geq 1/2\right)$ and B_1 be those points which are in infinitely many of these A_n. Then $P(B_1) = 0$ and for each point not in B_1 we have $|S_n/n| < 1/2$ after some stage (depending on the point).

Take $a = 1/2^2$ and A_n be the event $\left(|\frac{S_n}{n}| \geq 1/2^2\right)$ and B_2 be those points which are in infinitely many of these A_n. Then $P(B_2) = 0$ and for each point not in B_2 we have $|S_n/n| < 1/2^2$ after some stage (depending on the point) .

By taking $a = 1/2^k$ you get a set B_k such that $P(B_k) = 0$ and for each point not in B_k we have $|S_n/n| < 1/2^k$ after some stage.

Now let $B = \cup B_k$. Then $P(B) = 0$ and for points not in $B = \cup B_k$: whatever be k; we have $|S_n/n| < 1/2^k$ after some stage. In other words for points not in B, we have $|S_n/n| \to 0$. ∎

Observe that the above shows that for almost all points picked at random from $(0,1)$ the proportion of each binary digit is $1/2$. If, to change notation, Y_n is the n-th binary digit then apply the above to $X_n = 2Y_n - 1$.

5. Bose-Einstein revisited:

Let us return to distributing n balls again into boxes through an interesting experiment: by Tossing a random coin/dice.

I pick a number p at random from $(0,1)$. This is done once and for all. You can also say it as follows: from all coins pick one at random — that is, for each $0 < p < 1$ there is one coin with chance of heads p and we pick one at random.

For each of n balls, I toss the coin and put that ball in box 1 or 2 according as Heads or Tails. Tell me the number X of balls in box 1.

Clearly the event $(X = k)$ occurs iff you get k heads in n tosses. It is $\binom{n}{k}p^k(1-p)^{n-k}$. But unfortunately this is conditional probability, If you knew that the coin picked has chance of heads p. But we are not looking for conditional probability.

In the discrete case if we have two random variables X, Y and if we knew $P(X = x \,|\, Y = y_i)$ then we calculated

$$P(X = x) = \sum P(X = x \,|\, Y = y_i)P(Y = y_i) \qquad (\spadesuit)$$

Even in the case when we have joint density $f(x,y)$ for (X,Y) we did the same: Marginal density of y is $f(\bullet y)$ and the conditional density of X given $Y = y$ is $f(x,y)/f(\bullet y)$ so that conditional probability is

$$P(a < X < b \,|\, Y = y) = \int_a^b \frac{f(x,y)}{f(\bullet y)}dx$$

$$\int P(a < X < b \,|\, Y = y)f(\bullet y)dy = \int_y \int_{x=a}^b \frac{f(x,y)}{f(\bullet y)}dx f(\bullet y)dy$$

$$= \int_{x=a}^b \int_y \frac{f(x,y)}{f(\bullet y)}f(\bullet)dydx = \int_{x=a}^b f(x\bullet)dx$$

$$= P(a < X < b)$$

Thus we again have

$$P(a < X < b) = \int P(a < X < b \,|\, Y = y)f(\bullet y)dy \qquad (\spadesuit)$$

with sum replaced by integral and conditional probabilities multiplied by the density.

Unfortunately, in the present situation we have X is discrete and Y takes values in $(0,1)$ and has uniform density. We make the same definition.

Definition: If $X, Y_1, \ldots, Y_r$ are random variables and X is discrete taking values $x_1, x_2, \ldots$ and $Y = (Y_1, \ldots, Y_r)$ has density $f(y_1, \ldots, y_r)$ and if we know $P(X = x_i \,|\, Y = y_1, \ldots, Y = y_r)$ then the unconditional probabilities for X are given by

$$P(X = x_i) =$$

$$\int \cdots \int P(X = x_i \,|\, Y = y_1, \ldots, Y = y_r) f(y_1, \ldots, y_r) dy_1 \cdots dy_r \qquad (\spadesuit)$$

Incidentally all these ($\spadesuit$) are consequences of just the same one definition, but we need to use distribution functions — understandable, because we have discrete and continuous random variables all mixed up in a single context. In our course we had discrete random variables and then random variables with densities. But in practice you have both at the same time on stage! We saw that when we started with a sequence of exponential variables and ended up with a huge collection of Poisson variables. The present problem is another instance.

Returning to our problem

$$P(X = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1}$$

In other words, as if the n balls are put in two boxes and obey B-E statistics.

Let us now consider the same problem of distributing n balls into three boxes as follows: Pick a three faced die at random and use it distribute the balls. What does this mean? Let

$$\Delta = \{(p_1, p_2) : 0 < p_1, p_2, 1 - p_1 - p_2 < 1\}$$

Pick a point at random from this set and take a die with chance of faces $(p_1, p_2, 1 - p_1 - p_2)$. Use it. We would like to know $P(k_1, k_2, k_3)$, probability of the event that there are k_i balls in box i. We assume that $\sum k_i = n$. We know (multinomial probabilities, Exercise 87?)

$$P\{(k_1, k_2.k_3) \,|\, (p_1, p_2)\} = \binom{n}{k_1, k_2, k_3} p_1^{k_1} p_2^{k_2} (1 - p_1 - p_2)^{k_3}.$$

Here

$$\binom{n}{k_1, k_2, k_3} = \frac{n!}{k_1!\, k_2!\, k_3!}$$

Thus

$$P(k_1, k_2, k_3) = \iint_{\Delta} P\{(k_1, k_2.k_3) \,|\, (p_1, p_2)\} 2 \, dp_1 \, dp_2$$

Here we have used the fact that the set Δ has area $1/2$ (the region is a triangle) and so density equals 2 on the set and zero outside the set. After simplification we get

$$P(k_1, k_2, k_3) \;=\; \frac{2}{(n+1)(n+2)} \;=\; \frac{1}{\binom{n+2}{2}}$$

Again as if the n balls being placed in two boxes obey B-E statistics.

In general if n particles distribute themselves into r energy levels as follows: they pick a r-faced die at random and obey what the die dictates. More precisely, let

$$\Delta = \{(p_1, \ldots, p_{r-1}) : 0 < p_1, \ldots, p_{r-1}, 1 - \sum p_i < 1\}$$

subset of $(r-1)$ dimensional space. We pick a point at random from this set. Then

$$P(X = (k_1, \ldots, k_r) \,|\, p) = \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r}$$

where p_r is abbreviation for $1 - p_1 - p_2 - \cdots p_{r-1}$. Finally letting v denote the volume of Δ;

$$P(X = (k_1, \ldots, k_r)) =$$

$$\int \cdots \int_{\Delta} \frac{n!}{k_1! \ldots k_r!} \, p_1^{k_1} \;\ldots\; p_r^{k_r} \, \frac{1}{v} \, dp_1 \ldots dp_{r-1} \;=\; \frac{1}{\binom{n+r-1}{r-1}}$$

Again the familiar expression.
Thus God does not play dice but BE statistics do. This interpretation is due to the statistician Sudhakar Kunte, though many physicists have discovered it independently.

We discussed a little about Entropy and Information, Since it is just for fun and not part of course we shall not record.

FINAL EXAM: 26th April Wednesday.
DETAILS are already in Exercise set.