# Questions

* In 0-1 distribution, or Bernoulli distribution, the probability of each elementary event is not same, so they are not belonging to classical probability model? Do we have other probability model examples?
* In the definition of ‘population’ in the p129, it seems similar to the definition of sample space. What are their differences?
* What is the difference between random variable and ordinary functions?
* How to always go from MLE to minimization of loss/cost function? We should always go from MLE to do minimize loss/cost function because MLE can be taken as the most fundamental part of machine learning.
* EM is also to maximize the likelihood, but with an iterative way. Therefore EM is also MLE? Or MLE is a simple form or a special case of EM?
* How to understand eliminating the correlation among features will not affect the predicting results. Particularly understand that correlation will lead to numerical difficulty in either SGD, or in matrix conversion. In other words, it will lead to ill-posed matrix.
* For now, supervised and unsupervised learning seems only differ in the following: in unsupervised case, the element of MLE is the likelihood of the whole event (whole row of observation), while in supervised learning, the element is conditional probability. However, for unsupervised seems more complicated, not just the whole event as above, because there is only one mean there. That is only one cluster to cluster.
* Be clear why in MLE there is no hidden variable while in EM there is.

# Key requirement of Machine learning training data set

* As the data generator works, the number of data should be big enough. Note just big enough, the data should also well spread under the PDF profile. Otherwise we cannot restore the correct u and sigma for example. Because like in normal distribution case, the u is just the mean of all data, if we have smaller number of data, then u will not be accurate. How to make sure this point?
* The data should be independent. This independence indicates the sampling independence. In other words, each row of data in the dataset is independently sampled. This is way we can use the **product** in MLE to calculate joint probability. **Note the independence here has nothing to do with the independence among features** (different columns of the data set).

# Easy Errors

* When plotting the conditional probability p(yi|xi;theta), be careful that xi it is not x-axis at all.
* ‘y|x;theta’ is not mysterious at all but just another random variable. We can define X = y|x;theta, and thus calucate p(y|x;theta) = p(X).

# Chapter 1 Fundamental concepts of probability

## Summary

* Random event is subset of sample space. Random variable is a function.

## 1.Random Experiment

* **Random experiment**: can repeat under the same condition; the outcome of experiment is not only one, but we know its all possible outcomes; don’t know the outcome before the experiment.

## 2.Sample space, random event

* **Sample space** (set of **all possible outcomes** of random experiment). Each outcome in the sample space is a **sample point**.
* Sample space understanding 1: [1] The sample space for different experiments are different. For example tossing on coin, the sample space is {H, T}; [2] tossing three coins the sample space has 8 elements {HHH, HHT, ….}; [3] For experiments that flip a random number of coins, then the sample space can be {H, HHT, HH, HT, HHHH,….}.
* Sample space understanding 2: [1] For whatever sample spaces mentioned above, each sample point is an elementary point? See later the definition of elementary event. [2] **The so called sample point is just the elementary event**?
* On page 1: What are the sample space of random experiments E1 to E7. Particularly note **the E3 and E7**.
* **Random event** is a **subset** of a sample space. This is true for the sample space of finite elements or countable infinite elements. Otherwise, there might be some exceptions, although in practice we are almost always OK. My comments: so random event can be any subset of a sample space, but not just subset containing one element.
* An random **event occurs** if and only if a sample point occur in the subset. Note it seems that we don’t need the appearance of all the sample points within the subset to define the event occurring.
* A **single sample point** in the sample space is an **elementary random event**. See the following from Internet: Using our SET terminology, a random event is a SUBSET of a SET of all elementary events (in other words, a SUBSET of a sample space). Any such SUBSET is called a random event.
* Elementary random event: Take coin flipping as example, the elementary event is not necessarily a single-coin flipping. This is depended on what type of random experiments you are doing. If you only flipping one coin in each experiment, then {H, T} is the sample space, and H, T are the elementary event. But if you flip three coins in one experiment, then the elementary event is {HHH, HHT,….}. If your experiment is flipping a random number of coins in EACH experiment, then the elementary event could be {H, HT, HHT, T, HHTH….}, although this type of experiment is seldom done, right?
* From Wikipedia: An **event**, however, is any subset of the sample space, including any singleton set (an **elementary event**), the empty set (an **impossible** **event**, with probability zero) and the sample space itself (a **certain event**, with probability one). Other **events** are proper subsets of the sample space that contain multiple elements. Note empty space and whole sample space are also subset of sample space, and thus are **two special events**, as mentioned earlier.
* Relation between events and their operations. Because even is a set. Thus the operations among events are same as those of sets. Be familiar with the mapping between event occurring and set relations. First example: set B contains set A, the occurring of even A must lead to the occurring of event B. See the even occurring definition earlier, particularly the example of life time of a light bulb. Second example: A-B = {x| x belongs to A but not to B}, A-B occurs if and only if A occurs but B not.

## 3.Frequency and Probability

* In the definition of probability, there are no frequency-like stuff, but only three other conditions. [1] p(A) >= 0; [2] P(S) = 1; [3] P(A1UA2U…) = P(A1) + P(A2)+…. That is all, **no frequency at all**! We use this to map a real number P(A) to the event A and use it to describe the likelihood of the occurring of event A in a single experiment.
* From section 3 and 4 chapter 1, and chapter 5, it can be shown that when n approaches infinity, frequency fn(A) approaches P(A) in some sense. Thus we can use P(A) to describe the possibility of the occurring of event A in a single experiment. However, people often use the concept of frequency to replace the concept of probability. **Why bother we define the probability that way**?
* When proving P(A1 U A2 U A3) =…., I made a mistake: (AB)&(AC) = A&B&C but my answer was A at first, very bad mistake.
* When proving probability basic problems, always try to construct the relation of **union** of subsets. Thus we can use the formal definition of probability.

## 4. Classical Probability Model

* P(e1) = P(e2) = P(e3)….
* Exercise 1: Flip three coins, the probability of at least one head. I should do this type of problems with three ways. [1] write out all possible ways and count the relevant ones. this is only for not so many ways. [2] Use binomial formula B(n, p, k). nCk\*p^k\*(1-p)^(n-k) [3] The way I must know and be familiar with. The size of sample space = 2^3 = 8. Now choose one H and two TT. First choose one H with 3C1 ways. Then two TT has 2C2. So the total is 3C1\*2C2 = 3 ways. The probability is then 3/8. Note we can also choose two tails and then one head. The result 3C2\*1C1, which is same as choosing H first. For the second problem: Finding probability of at least one Head, always transfer to 1-P(A), where P(A) is the probability of all tails.
* Exercise 2: For replacement case. Use the formal ways. 4C1\*4C1/6^2…..
* Exercise 3. The model of assigning n balls to N slots (n<=N) is same as the birthday problem. When choosing n (<=365) persons, then what is the probability of each person has different birthday? Another problem: What is the probability that at least two persons have the same birthday?
* Exercise 4. N products where D products are bad. Now we choose n products. Find the probability of there is k (<=D) bad product within n. Note consider only the case without replacement. Answer: total sample space NCn. Next the number of k bad product which is DCk. Then the probability is DCk/NCn. However this is wrong. It is not complete. The right answer is: DCk\*(N-D)C(n-k)/NCn.
* Exercise 5. Here is an example showing that the probability for experiment with and without replacement is same. Thus it provides a proof of that people buy lottery earlier and later has the same chance. **Make sure what is the key that makes it same with and without replacement**?
* Exercise 6. Use the operations on sets to transfer one problem to other problems. For example, find the probability of a number cannot be divided by both 6 and 8. Very beautiful. p(!A & !B) = p(!(AUB)) = 1-p(AUB) = 1 – [p(A) +p(B) – p(AB)].
* Exercise 7. 15 new students (3 are best students) are assigned into three classes (5 each). [1] Find the probability of one best student in each class, and [2] probability of 3 best student in one class. Answer: total events. N = 15C5\*10C5\*5C5. Or 15!/(5!\*5!\*5!). [1] [3!\*12!/(4!\*4!\*4!)]/N or [3C1\*12C4\*2C1\*8C4\*1C1\*4C4]/N. [2] [3\*12!/(2!\*5!\*5!)]/N, or 3\*12C2\*10C5\*5C5/N. Tricks. I found that many problems can be done with factorials denoting step by step selection but not necessarily with C and P. Furthermore, C is just P by dividing a m! Try use factorial only in the future.
* Exercise 8. Seems wrong?

## 5.Conditional Probability

* In this section, conditional probability, multiplication theorem and Bayes formula are actually all from conditional probability. So an intuitive understanding of conditional probability is extremely important. Moreover, I should be very familiar both the formula and the intuition. For example I should **instantly and directly tell that p(B|A)p(A) is a joint distribution** without explicitly showing it equal to p(BA). From above we know that for a partition B1+B2+B3+…+Bn =S. p(A) = p(A|B1)p(B1)+p(A|B2) + … = p(AB1)+p(AB2) + ….
* **Be very familiar with** the coin-flipping example to quickly write the formula of probability. Flip one coin twice, event A: at least one H; event B: two coins are HH or TT. Find the probability of B when A already occurs. See the understanding below.
* The total sample space is { HH, HT, TH, TT}=S. First do A experiment we have {HH, HT, TH}. When A is done, what is the probability of B. Note the sample space after A is done becomes {HH, HT, TH}= Sr. In other words the sample space is reduced after A experiment. When we calculate the probability of B, we should use this reduced sample space. Here B will have only HH in the reduced space, thus p(B|A) = 1/3.
* The conditional probability p(B|A) = p(AB)/p(A) can be exactly mapped to the above understanding. The events of B after the occurring of A is proportional to P(AB), and reduced space is proportional to p(A). Thus p(AB)/p(A) is the same as the understanding above. Using the example above to understand p(B|A) = p(AB)/p(A) = p(A|B)p(B)/p(A). In other words, the same coin-flipping experiment to understand conditional probability and Bayes formula.
* Don’t be scared by the condition in the conditional probability. All the equations for normal probability apply exactly to the conditional probability, as the formal definition of probability is exactly same as the conditional probability. **Here we just consider a reduced sample space**. Therefore, we can use any properties of normal probabilities. For example, p(B1UB2|A) = P(B1|A) + p(B2|A) = p(B1B2|A), etc.
* For now I can calculate conditional probability with two nice ways: Take the example in P15-16. [1] Use the definition of conditional probability. [2] Using the condition, reduce the original sample space from (3good, 1 bad) to (2good, 1bad) after event A draw a good one. In the reduced sample space, the probability of drawing a good one (event B) is obviously to be 2/3.
* From conditional probability we obtain p(AB) = p(B|A)p(A). This can be generalized to the so-called multiplication theorem, with which we can write for example p(ABCD) = p(D|ABC)p(ABC) = p(D|ABC)p(C|AB)p(AB) = p(D|ABC)p(C|AB)p(B|A)p(A). It should have many ways to write it. But keeping an order is easy to write. For example, we separate first D, then C, B, and finally A.
* If we know p(A) = p(A|B), then we should deduce A and B are independent. First, the B event occurs do not affect the probability of the occurring of A. It is easy to show P(AB) = p(A)\*p(B) from the earlier equation.
* In [**Bayesian statistics**](https://en.wikipedia.org/wiki/Bayesian_statistics), the posterior probabilityof is the [**conditional probability**](https://en.wikipedia.org/wiki/Conditional_probability)that is assigned after the relevant [evidence](https://en.wikipedia.org/wiki/Scientific_evidence) or background is taken into account. Posterior probability is proportional to the prior probability and likelihood (except a denominator). This can be shown with that tumor machine learning example. p(y=1|x ) = p(x|y=1)p(y=1)/p(x). p(y=1) is the prior probability of positive results. p(x|y=1) is the likelihood, which can be interpreted as the probability of the x given that y is true. p(y=1|x) is the posterior probability when we know the size of the tumor x. The p(x) is equal to p(x|y=1)p(y=1) + p(x|y=0)p(y=0).
* For multi-variate Gaussian distribution (e.g. 2D), the “direction” of the 3D Gaussian mountain is not necessarily align with the x-y axes. If there are correlations within the two random variables, then the “direction” can be in any direction. The off-diagonal covariance matrix is not zero at this time. Whatever direction the bump, the marginal and conditional distribution of multivariate Gaussians are still Gaussians. [2] Marginal probability. The picture is the integrate one variable to obtain a 1D normal distribution (in 2D case). [3] Conditional probability, fix one variable and cut a surface, which will give a Gaussian distribution at the fixed one variable. Note the pictures of marginal and conditional 2D Gaussians are very different.

# Chapter 2 Random variable and its distribution

* Details in another word file named PMF\_PDFs.docx. very important.
* Apart from the distributions studied in the above file, also refer to the following link for many other PDF or PMFs: <https://en.wikipedia.org/wiki/List_of_probability_distributions>
* Check <https://en.wikipedia.org/wiki/Moment-generating_function> for the moment-generating function of a random variable. This is just for easily calculating moments of a random variable, for example, expectation value, variance, …. This just provides an alternative way of calculating statistical results from PDF/PMF. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. However, not all random variables have generating functions. For this reason, it is not so urgent to learn them carefully. Anyway it is an alternative way.
* Check Wikipedia for another alternative to PDF/PMF is there Fourier transform, called characteristic function. It might also provide simplification in many calculations.
* Check Wikipedia for entropy and cross-entropy. However, it is not so urgent for now. For example, in the logistic regression, the cost function can also be obtained from alternative way without using cross-entropy. So it is not so urgent to learn for now.

## Random variable

* Random variable is a real-value **function** mapping each element in a sample space to a real value. Formally we have: S= {e} is the sample space, X= X(e) is the **single-value function defined on S**, then X is a random variable.
* Sometimes e is itself is a real number, then we let X = X(e) = e, then X is a random variable.
* Random variable X ϵ L, where L is a subset of R. P{X ϵ L} = P(B) = P{e|X(e) ϵ L}. e is sample point. X(e) is random variable. The reasoning is: When X has the value of a subset of R, then these values must correspond to a subset of sample points in the sample space. We call this subset of sample space event B. This event B has its probability, which is same as the probability of X ϵ L. In other words, because ‘random event’ (subset of sample space) has probability distribution, random variable also has probability distribution as random variable and random event are mapped.
* So **although called a variable, it is actually a function**. Then how about the normal variable or the variable in algebra? Is it also a function? Check Google: “edit: More rigorous Definition of a Variable” we obtain: variable is symbols … The [interpretation](http://en.wikipedia.org/wiki/Structure_(mathematical_logic)#Definition) of the symbols defines what they actually mean. So for instance the interpretation of x above would be a function which maps {2,3,5}↦{2,3,5} such that each element is mapped to itself i.e. x(2)=2, x(3)=3 and x(5)=5).
* What is the difference between random variable and ordinary functions? Before an experiment we don’t know what value a random variable (it is a function) will take. For a normal function as in algebra, we know this? How?

## Discrete Random variable and its distribution

## Random experiment of one coin

* **One experiment** only involves **ONE unbiased coin**, each coin has two possible outcome {0,1}. p(y=0) = p(y=1) = 0.5.
* **One experiment** only involves **ONE biased coin**, each coin has two possible outcome {0, 1}. Three ways of writing out the probability of each outcome

p(y=1) = φ, p(y=0) = 1 - φ .

p(y) =

p(y) =

* **Each experiment** involves **n unbiased coin**, each coin has two possible outcomes {0,1}. Here we can think that n coins are flipped simultaneously for a single experiment. Or we can think we did it one by one independently. Note because a single experiment involves n coins, and each coin has two outcomes, so the size of sample space is 2^n. Now calculate the probability of exactly k coins up. This can be calculated as for unbiased coins. Because here n coins in the experiment are independent, we can directly calculate the probability as (see many books) , where . In the first version of calculating p, we separate two steps, choosing k and n-k respectively and multiply them together. In the second version, we mainly use independence property, the is completely different from p. So the multiplication of has nothing to do with the multiplication of .
* Note when , the results from above two calculations are same. When , then it is the general form for **binomial distribution**. It is often written as B(n,). In the Machine learning course for binomial case, it is B() where n = 1 is eliminated. So the binomial case in ML course is just the simplest case B(n=1, )?
* The general binomial distribution comes from the fact that the formula is just the coefficient of . The multinomial comes from the coefficient of (a1+a2+a3…+ ak. An intuitive example for the general multinomial distribution M(n, ) is that we flip a biased die with k faces, and the probabilities for appearing the number of x1,x2,…xk times for each face respectively are , respectively. Also, we can describe the multinomial in other forms of drawing different colors of balls as shown in Wikipedia or Wolfram Mathworld. There we can also check the definitions of (m, n)! = , which is simpler case of multinomial coefficient (n1, n2,… nk)!.
* Similar to the binominal case, in the multinomial case we normally only consider the n=1 case. That is p(y) = . (also note in multinomial case, in Andrew’s course, we use y=1, 2,…k, but not y=0, 1, …k-1.

## Random experiment with colored balls for multinomial distribution

* Understand the multinomial with drawing n colored balls with k different colors, Xi = xi is the number of balls with color index i. Note sum\_i\_to\_k(Xi) = n. This is same as binomial case where n-k+k = n. M(n,p1,x1; p2,x…) indicate that drawing n balls, the probability of get k1 balls of color index 1. Note p1 is the probability of drawing once. For binomial distribution B(n,p,k) we can draw a curve B = B(X=k) (n and p are fixed). For multinomial distribution with two variables M = M(X1= x1i, X2i= x2i), we still can visualize the PMF with 3D plots. However, we cannot visualize with higher dimensions. It is sort of binomial distribution curve extended to 3D. Each slice of the “mountain” is a binomial distribution.
* In the ML course, it seems that for multinomial case, we still have n=1 ball for each color.

# Chapter 5 Big number and central limit theorem

* Big number law: frequency -> probability.
* For center limit theorem to be valid, we need X = X1+X2…… where Xi can have any type of distribution as long as they are independent and identical.

# Chapter 10 Bootstrap method

## General

* Normally we have only one sample (X1, X2,….Xn). Thus if we calculate sample mean, we only have one sample and thus one mean calculated from it. We thus cannot obtain a sample mean distribution and its standard deviation (here it is standard error).
* The key for sampling by replacement is like treat the existing sample as the population. So every sample obtained is like a real sample. Or we may think that we are doing real experiments even though the experiments are very hard to do or impossible to do.
* There are two types of bootstrap methods, as explained in the book.

# Chapter 12 Stochastic Process

## temp

* stochastic process vs sample function ⬄ population vs sample
* Infinite number of random variable is a stochastic process. For fixed t, X(t) is a random variable. For t ϵ T, x(t) is a sample function.
* Notation simplification: {X(t), t ϵT} -> X(t), t ϵT -> X(t).
* Compare the Pearson correlation I calculated with some new definitions defined on stochastic process. See notes in the stochastic folder.

## First way of understanding stochastic process

* Random variable -> can randomly take a value. That is a realization of a random variable is a single value. This is of course just one realization of the random variable. The variable can also take other values, one value for one realization.
* Now we first see what a process mean? It is a varying behavior described either by a discrete sequence or by a function (usually time-dependent function). In other words, a process needs a multiple values not single value to describe. For clarity, we assume a ‘fixed’ process is described by a function x1(t)—a realization.
* Now a random process, which means a process can take many random realization. The above x1(t) is one realization. It can also be another realization x2(t), x3(t)…..each process need many values to describe. This is exactly the diagrams in the beginning of chapter 12 for thermal noise in circuit, which is a prefect example to describe random process. In other words, each diagram is the random realization of the random process. The key is: formerly we have a random ‘variable’ which takes a single value in each realization, now we have a random ‘process’ which takes a single process (need many values to describe) in each realization**.**
* Anyway, we can still think a process as a ‘big’ random variable and each process realized is a single value. This is just like in functional analysis, a function (here the process) is taken as a single dependent value in the range, when the functional is in the domain and take the function as a variable.

## Second way of understanding stochastic process

* A population corresponds to a random variable X, and (X1, X2, …Xn) is a sample and (x1, x2,. …xn) is a realization of the sample. For random process, the sample should be (X1(t), X2(t), ….), and its implementation is (x1(t), x2(t), ….), where xi(t) is a single function. However, conceptually they are similar.
* I need make sure whether the following is correct? As t can take any values (infinite number), x1(t) thus can take infinite number of values. Rearranging the above random process as [(x1(t1), x2(t1), ….xn(t1),) ((x1(t2), x2(t2), ….xn(t2)…)…]. We can think this way, at arbitray time instant t1, we have a random varaible X(t1), which randomly takes the values of [(x1(t1), x2(t1), ….xn(t1),…). Similarly we have another random variable X(t2)…. Therefore, we have infinite number of random varaibles X(t), t >=0. Random process is thus a multi-variate random process. In fact it is infinite-variate random process. Much bigger than the 2D, 3D random process. Anyway, we can still obtain insight from the multi-variate chapters.

## Third way of understanding stochastic process

* In Newtonian dynamics, we study the dynamic behavior of a variable X(t). At each time instant t, we can for sure what X(t) will be. We can call it a deterministic process. However, in the stochastic process, X(t) in each time instant t is not determined (not repetitive), see the diagrams in the beginning of chapter 12. This is because in each time t, X(t) is a random variable, and thus might take different values if we repeat the experiments.
* However, like deterministic Newtonian dynamics, we can still study the dynamics of X(t). So stochastic process is just dynamic part of statistics. Therefore, like Newtonian dynamics where calculus is involved, we have stochastic calculus here. The key difference is: in Newtonian dynamics, X(t) is a deterministic variable and thus take a determined single value. In stochastic process however, X(t) at time = t is a random variable and thus will take random values.

## Time Series

* As mentioned before, for random process, the sample should be (x1(t), x2(t), ….) where xi(t) is a single function. A time series is just a realization of a random process. x1(t) is a time series. x2(t) is also a time series. They are just two realization of a randomly changed process.
* The correlation I calculated before is actually the correlation of two time-series. We may also think x1(t) describes the values of a random variable X, and x2(t) describe the values of a random variable Y, so the correlation is also similar to that of X and Y random variables.