Lecture - 5

04-AUG-2009

In the last class we observed that whenever $P[X=x]\neq 0$, the distribution function $F_X(x)$ is not continuous at x. Now we want to study such class of distribution functions which are discontinuous at countable number of points. Here goes the formal definition:

DISCRETE RANDOM VARIABLE:

A random variable X is called a discrete random variable if there exists a countable set E such that:

$$P[X\in E]=1$$

In other words $P[X\notin E]=0$.

Let the set E be $\{x_1, x_2, \ldots\}$. Now without loss of generality we can assume $x_1\leq x_2\leq\cdots$ and also assume $P[X=x_i]\neq 0 \ \forall x_i\in E$.

Now by the very defn. of discrete r.v. we have: $\sum_{x_i\in E} P[X=x_i]=1$.

It is easy to see that the distribution function of X can now be written in terms of $P[X=x_i]$ as follows:

$$F_X(x)=\sum_{\substack{x_i: x_i\leq x,\\ x_i\in E}} P[X=x_i]$$

This shows that given the values of $P[X=x_i]$, the distribution function gets uniquely determined.

~~Hence~~ Also, $P[X = x_i] = \underbrace{P[X \le x_i]}_{F_X(x_i)} - \underbrace{P[X < x_i]}_{F_X(x_i^-)}$ $\rightarrow$ (the left limit)

Hence specifying $P[X = x_i]$ $\forall\, x_i \in E$ is equivalent to specifying the distribution function $F_X(x)$ and vice-versa

~~The~~ We give a name to the $P[X = x_i]$ as <u>probability mass function</u> :

$$\cancel{f(x_i) = P[X = x_i] \text{ if } x_i \in E, \quad 0 \text{ if } x_i \notin E}$$

$$f_X(x) = \begin{cases} P[X = x_i] & \text{if } x = x_i \in E \\ 0 & \text{if } x \notin E \end{cases}$$

In case of discrete r.v. we always specify the prob. mass function (p.m.f.) i.e. $f_X(x)$ instead of distribution function $F_X(x)$.

Now, the only constraints on f_X are as follows:

$$\left.\begin{array}{lll} \text{(i)} & 0 \le f_X(x_i) \le 1 \quad \forall\, x_i \in E & (\because \text{defn. of } f_X) \\ \text{(ii)} & \displaystyle\sum_{x_i \in E} f_X(x_i) = 1 & (\because P[X \in E] = 1) \end{array}\right\} \quad (\mathrm{I})$$

Recall that this is exactly what we did when we looked for "valid" probability functions on countable sets! (So we already know some eg. like geometric series etc. do the job)

We explore some special discrete r.v. now:

(Discrete) Uniform R.V.

Our first eg. is as follows:

Consider the set $E = \{1, 2, \ldots, n\}$ & the pmf: $f_X(i) = \frac{1}{n}, \forall i \in E$

It is easy to verify f_X is a valid p.m.f. (n is a parameter)

Now this can be applied to any situation where we know the outcomes are "equally likely". This basically "models" the classical probability. eg. coin toss, throwing die etc.

Bernoulli R.V.

Consider the set $E = \{0, 1\}$ & the pmf: $f_X(1) = p$, $f_X(0) = 1-p$

Here $0 \leq p \leq 1$ is a parameter. Again f_X satisfies (I) and hence is a valid p.m.f.

This r.v. models all random expts. with two outcomes. For eg. coin toss, manufacture of good/bad parts etc. Such expts. are known as Bernoulli trials (i.e. rand. expts. with two outcomes). Usually in a Bernoulli trial one of the two outcomes is called success $(X=1)$ and the other failure $(X=0)$. Hence prob. of success is p.

Binomial R.V.

Consider the set $E = \{0, 1, 2, \ldots, n-1, n\}$ & the pmf: $f_X(i) = {}^nC_i \, p^i (1-p)^{n-i}$ $\forall i \in E$

Here $n \in \mathbb{N}$ & $0 \leq p \leq 1$ are two parameters to the Binomial random variable. Now lets check if f_X is a p.m.f:

It is obvious that $f_X(i) \geq 0 \quad \forall i \in E$

So we need to verify if $\sum_{i=0}^{n} {}^{n}C_i \, p^i (1-p)^{n-i} = 1$. This is indeed true because $\sum_{i=0}^{n} {}^{n}C_i \, p^i (1-p)^{n-i} = (p+(1-p))^n = 1^n = 1$. Hence f_X is a valid p.m.f.

Binomial r.v. can be employed to "model" probability of 'k' successes in 'n' independent Bernoulli trials. Recall the defn. of Bernoulli trial: it is a rand. expt with two outcomes: success (prob. p) and failure (prob. $1-p$). Now let us denote ~~why a binomial r.v.~~ the prob. space of the i^{th} Bernoulli trial by $(\Omega_i, \mathcal{F}_i, P_i)$. Here $\Omega_i = \{\text{Success}(S), \text{Failure}(F)\}$. $P_i(\{S\}) = p$ & $P_i(\{F\}) = 1-p$. $\forall$ $i = 1$ to n Bernoulli trials.

Now consider the combined expt of all the n Bernoulli trials. The sample space of this is $\Omega_1 \times \Omega_2 \times \dots \times \Omega_n$. Now consider this singleton event of the combined expt.: $\{(\underbrace{S, S, \dots, S}_{k}, \underbrace{F, F, \dots, F}_{n-k})\} = A$.

In words this event is nothing but the event where first 'k' trials were a success & the remaining 'n-k' trials were failures.

Now consider the event: $\boxed{\text{We want to calculate prob. of events in combined expts. using probabilities } P_1, P_2, \dots, P_n.}$

$$A_1 = \{(S, \omega_2, \omega_3, \dots, \omega_n) \mid \omega_2 \in \Omega_2, \omega_3 \in \Omega_2, \dots, \omega_n \in \Omega_n\}$$

In words, this event is the event of a success in '1st' trial. Similarly define A_i for $i = 1$ to n. Note that A_i $i = 1$ to k represent success in i^{th} trial.
A_i $i = k+1$ to n " failures in i^{th} trial.

It is easy to see that $A = A_1 \cap A_2 \cap \dots \cap A_n$

Now let P be a prob. function of the sample space $\Omega_1 \times \Omega_2 \times \dots \times \Omega_n$

So, $P(A) = P(A_1 \cap A_2 \cap \ldots \cap A_n)$

$= P(A_1) P(A_2) \ldots P(A_n) \rightarrow$ we assume each trial is independent of others so the event A_i are independent of each other

This is the assumption of Independent trials.

~~A~~ Now we take $P(A_1) = P_1(\{S\})$, $P(A_2) = P_2(\{S\}) \ldots$, and so on ~~$P(A_n) = P_n(\{S\})$~~

i.e. we construct the prob. in the combined expt. with $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$ such that it is "consistent" with the probabilities $P_1, P_2, \ldots, P_n$ in the individual trials!.

~~Using similar argument~~

Hence we have $P(A) = P(A_1 \cap A_2 \ldots \cap A_n)$ } Assumption of Independent trials

$= P(A_1) P(A_2) \ldots P(A_n)$ Assumption of consistency

$= \underbrace{P_1(\{S\}) P_2(\{S\}) \ldots P_k(\{S\})}_{k} \underbrace{P_{k+1}(\{F\}) \ldots P_n(\{F\})}_{n-k}$

Assumption of identical Trials $\Big\{ \; = p^k (1-p)^{n-k}$.

Now ~~we~~ we have that prob. of ~~k~~ first 'k' trials a success & next 'n-k' trials failure $= p^k (1-p)^{n-k}$. Now actually prob. of any 'k' trials success & remaining failures is again $p^k (1-p)^{n-k}$. But there are exactly ${}^nC_k$ ways 'k' successes can happen in 'n' trials.

Hence prob. of 'k' successes in 'n' trials $= \underbrace{p^k (1-p)^{n-k} + \ldots + p^k (1-p)^{n-k}}_{{}^nC_k}$

$= {}^nC_k \, p^k (1-p)^{n-k}$

Now let us ask a slightly different but related question: "what is prob. that the number of trials ~~that need to be done to~~ for realizing the first success is 'k'." As we shall see below a geometric r.v. helps us ~~to do~~ model this:

Geometric R.V.

Consider the set $E = \{1, 2, \ldots\}$ i.e. $E = \mathbb{N}$. (Note that this is the first eg. for the countably infinite discrete r.v). Define pmf as $f_X(x_i) = p(1-p)^{x_i - 1}$ $\forall x_i \in E$. It is routine to verify this f_X is indeed a valid pmf.

This random variable models the trial at which the first success occurs in a sequence of ~~Bern~~ identical and ~~independent~~ independent Bernoulli trials. (IIB trials)

X = trial at which the first success occured

It is easy to X can take values $\{1, 2, \ldots\}$ which is exactly the set E for geometric r.v. Also, using the ideas discussed in the previous section for analyzing IIB trials, we have:

$$P[X = k] = p(1-p)^{k-1}.$$

Hence a geometric variable is suitable to model "trail at which first success occurs". Now, let us look at:

$$P[X > m] = 1 - P[X \le m] = 1 - \sum_{i=1}^{m} p(1-p)^{i-1} = 1 - p\,\frac{1-(1-p)^m}{1-(1-p)} = (1-p)^m.$$

Also,

$$P[X > k+m / X > k] = P_X((k+m,\infty) / (k,\infty)) = \frac{P_X((k+m,\infty) \cap (k,\infty))}{P_X((k,\infty))} = \frac{P_X((k+m,\infty))}{P_X((k,\infty))} = \frac{(1-p)^{m+k}}{(1-p)^k} = (1-p)^m = P[X > m]$$

This important observation that, $P[X > k+m / X > k] = P[X > m]$ is called the 'memory loss' property. In words it says that ~~for~~ prob. of achieving success after m steps is same immaterial of how many trials have been performed.

In this lecture we will first complete our discussions on discrete r.v. by presenting the Poisson r.v. Then we will move on to continuous r.v. — their definition, examples and applications.

POISSON R.V.

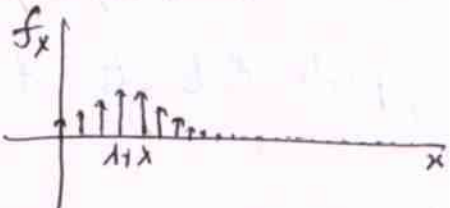
$E = \{0, 1, 2, \ldots\}$, the set of whole numbers.

pmf: $f_X(x_i) = e^{-\lambda} \dfrac{\lambda^{x_i}}{x_i!} \quad \forall\, x_i \in E$, where $\lambda > 0$ is a parameter.

Its easy to verify f_X is non-negative & $\displaystyle\sum_{x_i=0}^{\infty} e^{-\lambda}\left(\frac{\lambda^{x_i}}{x_i!}\right) = e^{-\lambda}\underbrace{\left(1+\lambda+\frac{\lambda^2}{2!}+\cdots\right)}_{e^{\lambda}} = 1$

Hence f_X is indeed a valid pmf. Now lets plot this pmf. for that lets try to look at the ratio of pmf values at two consecutive numbers:

$$\frac{f_X(k+1)}{f_X(k)} = \frac{\lambda}{k+1}$$



In other words, f_X increases till $\lambda-1$ & then decreases. This plot is similar to the binomial case; the difference being that this extends to all (whole) numbers. The distribution function again is an (infinite) staircase of equal length steps by heights proportional to f_X values.

Infact, after the study of concept of convergence of r.v., one can show that the binomial distribution "converges" to the poisson distribution in the case $n \to \infty$ (no. bernoulli trials is large), $p \to 0$ (prob. of success in each trial is low) such that $np = \lambda$.

In other words,

$$\begin{array}{ccc} P[X_b = k] & & P[X_p = k] \\ \downarrow & & \downarrow \\ {}^{n}C_k\, p^k (1-p)^{n-k} & \xrightarrow[np=\lambda]{n\to\infty,\, p\to 0} & e^{-\lambda}\dfrac{\lambda^k}{k!} \end{array}$$

$X_b \to$ binomial r.v.
$X_p \to$ poisson r.v.

Recall that Binomial r.v. can model 'no. successes in 'n' Bernoulli trials'.

Hence, by the above relation, we can say that the Poisson r.v. can be used to model 'no. of ~~successes~~ occurrences of a rare event in large no. B. trials'.

eg: A person keeps ~~buy~~ buying lottery tickets. The no. times ~~he~~ he wins a lottery follows Poisson distribution (why?)

eg No. words written by IPL instructor which are perfectly legible :)
(on board & this notes)

Till now we have looked at random variable which took discrete values and had discontinuous distribution functions. Now lets turn our attention to r.v. whose distribution functions are continuous (infact absolutely conts.) As we already discussed:

$$P[X \le x] = P[X < x] + P[X = x]$$

$$\underset{\text{(right limit)}}{F_X(x)} = \underset{\text{(left limit)}}{F_X(x^-)} \iff P[X=x]=0 \quad \forall x \in \mathbb{R}.$$

In other words r.v. with continuous distribution functions cannot have $P[X=x] \neq 0$ for any $x \in \mathbb{R}$! Hence we cannot have a "pmf" function in this case. The idea is to have a prob. density function (pdf) instead ~~which~~ and finding area under the density function would give probabilities.

More formally we define continuous r.v. as follows:

CONTINUOUS R.V.

A r.v. X is called a continuous r.v. if there exists a function $f_X : \mathbb{R} \to \mathbb{R}$, called the probability density function (pdf), such that:

$$\underbrace{P_X(B)}_{\text{induced prob. fuction with r.v. } X} = \underbrace{\int_B f_X(x)\,dx}_{\text{we know how to calculate integrals when B are intervals etc.}} \quad \forall B \in \underbrace{\mathcal{B}}_{\text{borel σ-algebra.}}$$

Here goes an intuition why f_X is called a pdf:

Suppose we consider $B = (x - \varepsilon/2, x + \varepsilon/2)$ where ε is tiny. i.e. B is a small interval around x. Then $P_X((x-\varepsilon/2, x+\varepsilon/2)) = \int_{x-\varepsilon/2}^{x+\varepsilon/2} f_X(y)\,dy = \varepsilon f_X(x)$
since ε is tiny f_X does not change.

In other words $f_X(x) = \dfrac{P_X((x-\varepsilon/2, x+\varepsilon/2))}{\varepsilon}$ for $\varepsilon \to 0$. ~~Hence~~ Since f_X is ratio of prob. & length it is called as 'prob. density'.

Now, $F_X(x) = P_X((-\infty, x]) = \int_{-\infty}^{x} f_X(y)\,dy$ (I)

In other words, given the p.d.f, the dist. func. $F_X(x)$ is fixed. Functions like F_X which are expressible as integral over functions like f_X are known as absolutely continuous functions. Absolute continuity is a stricter condition than continuity. In fact, we even know that F_X is differentiable:

$$\frac{dF_X(x)}{dx} = f_X(x) \quad \forall x \text{ at which } f_X \text{ is continuous.} \quad \text{(II)}$$

Now since one can obtain the dist. function F_X given f_X (pdf) and vice-versa, we characterize continuous r.v. using pdfs.

Let's look at some properties of the pdf:

pdf (f_X) satisfies:

(i) Non-negativity: i.e. $f_X(x) \geq 0 \ \forall x \in \mathbb{R}$. This follows from the monotonicity of F_X. Since F_X is area under f_X, there is no way the F_X (area) can monotonically increase if $f_X < 0$. Mathematically:

$$x_1 \leq x_2 \Rightarrow F_X(x_2) - F_X(x_1) = P_X((x_1, x_2]) = \int_{x_1}^{x_2} f_X(y)\,dy \geq 0 \quad \forall x_1 \leq x_2$$

$$\Rightarrow f_X(x) \geq 0 \ \forall x.$$

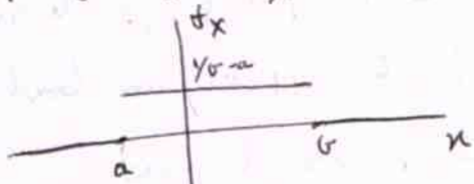
(ii) Unit-area: We have, $1 = P_X(\mathbb{R}) = \int_{-\infty}^{\infty} f_X(x)\,dx$, Hence the area under f_X must be unity.

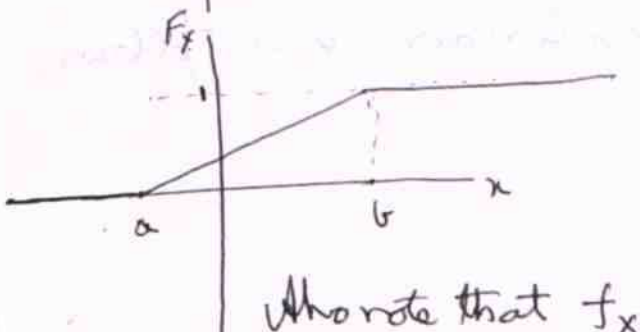
Any function which satisfies these two conditions we called it a prob. density function (pdf). Let us look at some eg. of conts. r.v.

(conts.) Uniform R.V.

pdf: $f_X(x) = \begin{cases} \frac{1}{b-a} & \forall x \in [a,b] \\ 0 & \forall x \notin [a,b] \end{cases}$. Here $a<b$ ~~a & b~~ are two parameters.

It is trivial to check f_X is indeed a pdf. Now the plots of pdf & F_X are:





Note the relations (I), (II) from these PLOTS ~~graphs~~. The points where F_X is not differentiable is exactly where f_X is discontinuous.

Observe that $1/b-a$ can be >1. So there is no reason to believe in general that $f_X(x) \leq 1$. (So $f_X(x)$ need not be ≤ 1)

Also note that $f_X(a)$, $f_X(b)$ ~~can be defined as any number~~ can be changed to arbitrary values without changing F_X!

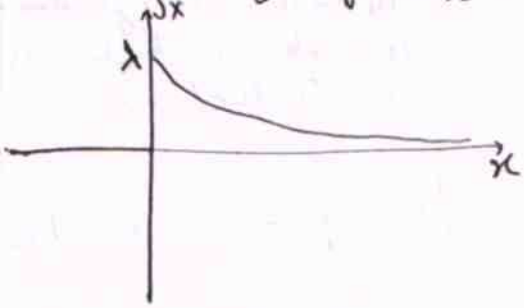
Exponential R.V.

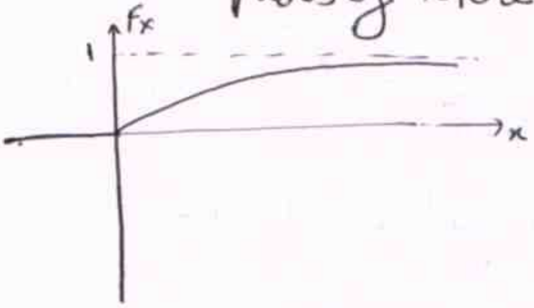
pdf: $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$ $\lambda > 0$ is a parameter

Again $f_X(x) \geq 0$. Also, $\int_{-\infty}^{\infty} f_X(x)dx = \int_0^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^{\infty} = 1$. Hence $f_X(x)$ is a valid pdf.

Now, $F_X(x) = P_X((-\infty, x]) = \int_{-\infty}^{x} f_X(y)dy = \int_0^x \lambda e^{-\lambda y}dy = e^{-\lambda y}\Big|_0^x = 1 - e^{-\lambda x}$.

$f_X(x)$ is an exponential decay function on $x \geq 0$ (no convex on $x \geq 0$), whereas F_X is a negative exp. decay (no concave for $x \geq 0$). Here are the plots of these functions:





Now, similar to the geometric r.v in discrete case, exponential r.v. satisfies the memory-less property:

TST $P[X > x+y \mid X > y] = P[X > x]$ $\forall x, y \geq 0$.

Proof: $P[X > x] = 1 - F_X(x) = e^{-\lambda x}$ $\forall x \geq 0$.

$$P[X > x+y \mid X > y] = \frac{P[X > x+y \cap X > y]}{P[X > y]} = \frac{P[X > x+y]}{P[X > y]} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x} = P[X > x]$$

$\longrightarrow$

One can show the converse also i.e if X is a conts. r.v. which is non-negative and satisfies memory-less property then X <u>MUST</u> be exponential r.v.

Proof: From above proof we have that memory less property is same as:

$$P[X > x+y] = P[X > x]P[X > y] \quad \forall x, y \geq 0$$

Now use the fact that X is a r.v., this gives:

$$[1 - F_X(x+y)] = [1 - F_X(x)][1 - F_X(y)]$$

Note this does not say the events $X > x$, $X > y$ are "independent"

Now call $G_X(x) \equiv 1 - F_X(x)$. We know that $F_X(x)$ is a continuous function, hence $G_X(x)$ is a conts. function which satisfies:

$$G_X(x+y) = G_X(x)\, G_X(y) \qquad \forall\, x, y \geq 0. \quad — (1)$$

Now,

$$G_X\left(\tfrac{m}{n}\right) = G_X\Big(\underbrace{\tfrac{1}{n} + \cdots + \tfrac{1}{n}}_{m \text{ times}}\Big) = G_X^{m}(1/n) \qquad (\because \text{repeated application of (1)})$$

Also, if $m = n$, we have $G_X(1) = G_X^{n}(1/n) \Rightarrow G_X(1/n) = (G_X(1))^{1/n}$

Hence, $\quad G_X\left(\frac{m}{n}\right) = \left(G_X(1)\right)^{m/n}$.

So we proved that G_X is ~~an~~ a power function for all rationals $\frac{m}{n}$ (≥ 0)

By continuity of G_X, G_X must be a power function for all reals (≥ 0)

$$\Rightarrow G_X(x) = \left(G_X(1)\right)^{x} \qquad \forall\, x \geq 0.$$

Now $G_X(1) = P_X[X > 1]$. Hence $0 < G_X(1) < 1$. Because of this I can choose a $\lambda > 0$ such that $\lambda = -\log(G_X(1))$.

$$\Rightarrow G_X(x) = e^{-\lambda x}, \quad \lambda > 0. \quad \forall\, x \geq 0.$$

$$\Rightarrow F_X(x) = 1 - e^{-\lambda x} \quad \forall\, x \geq 0 \quad (\lambda > 0).$$ which is nothing but the distribution function of an exponential r.v. Hence proved.

Thus in non-negative conts. r.v., memory-less property is unique to Exponential r.v. Now, ~~we can~~ encouraged by this, we can apply exp. r.v. to model all cases (conts. versions) for which geometric r.v. was applicable. (Recall that geometric r.v. also is the only memory-less discrete r.v.)

$\therefore$ Exp r.v. can model waiting time for a succesful event.

However care needs to taken (as in case of geometric r.v.) that the physical situation make not support the memory-less property.

<u>Folly.</u> Suppose we model the time to failure of a T.V. by an exponential random variable. Then we will be saying an absurd ~~th~~ statement as follows:

"Let prob. of (new) T.V. working for 10 yrs. be 0.7. Then given that it already worked for 10 years, the prob. that it works for 10 more yrs. is again 0.7".

Normal R.V.

pdf: $$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad x \in \mathbb{R}$$

Appears in many many applications. And needs no introduction.

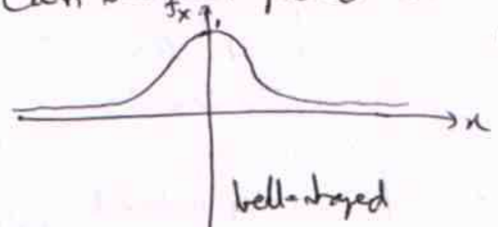
f_X is indeed non-negative. To show that f_X is a valid pdf, we need to show that:
$$I \equiv \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1.$$

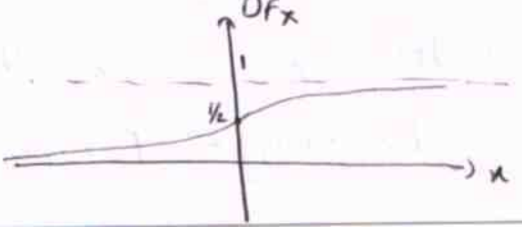
<u>Proof</u>: Consider the integral,
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dx\,dy = \int_{-\infty}^{\infty} \underbrace{\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right]}_{I} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = I^2.$$

Now transform the integral in I^2 using polar coordinates:
$$I^2 = \int_0^{2\pi}\left[\int_0^{\infty} e^{-r^2/2} r\,dr\right]\frac{1}{2\pi} d\theta = \int_0^{\infty} e^{-r'} dr' = -e^{-r'}\Big|_0^{\infty} = 1$$

$\Rightarrow I = 1$ ~~because~~ ($I \neq -1$ because I is integral of non-negative function).

Unfortunately $F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ cannot be computed in closed form. However can be computed using numerical integration. Here are the graphs:





Consider the following random experiment: "Choose a random circle centered at origin and having radius between 0 & 1. Assume all radii are equally likely". Let the probability space for this expt. be $\mathbb{P} = (\Omega, \mathcal{F}, P)$.

Now, consider a random variable "R" ~~defined~~ on this probability space which in words is "radius of the circle". In other words, R is a r.v following uniform distribution between [0,1].

Consider another mapping from Ω onto $\mathbb{R}$, which is "A": "area of the circle". ~~In this~~ ~~Now~~ Note that for each circle $\omega \in \Omega$, we have:

$$A(\omega) = \pi (R(\omega))^2.$$

~~Getting~~ Defining a new function $g : \mathbb{R} \to \mathbb{R}$ such that $g(x) = \pi x^2$, it is easy to see that $A = g \circ R$, where '$\circ$' denotes composition of functions. In other words, $A(\omega) = \cancel{g(} g \circ R(\omega) = g(R(\omega)) \; \forall \omega \in \Omega$.

We denote this as $\underline{A = g(R)}$ (abuse of notation?)

Now obvious questions are:

(i) Given that R is a r.v., Is $A = g(R)$ ~~always~~ a "valid" r.v?

(ii) If so, what is the distribution of the new random variable A, which is defined in terms of ~~a r.v~~ another r.v R? (In particular, what is the distribution of the r.v "Area of circle", given that "R: radius of circle" follows uniform distribution in [0,1]?)

In this lecture, we try a answer the above questions and in general, study the notion of "functions of R Vs".

~~Now~~ First, lets answer question (i):

The only thing we need to check is ~~the~~ whether:

$$A^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}$$

$$\Leftrightarrow (g \circ R)^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}$$

$$\Leftrightarrow R^{-1}(g^{-1}(B)) \in \mathcal{F} \quad \forall B \in \mathcal{B}$$

Now suppose $g^{-1}(B) \in \mathcal{B} \ \forall B \in \mathcal{B}$, then since $R^{-1}(B) \in \mathcal{B} \ \forall B \in \mathcal{B}$, by the very fact that R is a r.v, we have that

(I) $\underline{\underline{g^{-1}(B) \in \mathcal{B} \ \forall B \in \mathcal{B}}}$ is sufficient for A being a valid r.v.

Now $g: \mathbb{R} \to \mathbb{R}$. One can show (not in this class) that if g is conts. then the above contd. is meet. In other word if g is conts then A is assured to be r.v.

(In our case, $g(x) = \pi x^2$, so indeed A is a r.v.)

Also note that contd. (I) itself implies that g is a valid r.v. with initial probability space as $\mathbb{P}_{\mathbb{R}} = (\mathbb{R}, \mathcal{B}, P_{\mathbb{R}})$!

So now we exploit this fact and attempt defining prob. wrt. A i.e. P_A using $P_{\mathbb{R}}$ (same thing as we did in case ~~of~~ while defining r.v!)

We define $\quad \underline{\underline{P_A(B) = P_R(g^{-1}(B)) \quad \forall B \in \mathcal{B}}}$

~~It is easy to check~~ Note that P_A is a valid prob. function because g is itself a r.v. on $\mathbb{P}_{\mathbb{R}}$ (as noted above).

To give an overall picture:

$$\begin{tikzcd} \mathbb{P} = (\Omega, \mathcal{F}, P) \arrow[d, "X"] \arrow[dd, bend left, "Y=g(X)"] \\ \mathbb{P}_X = (\mathbb{R}, \mathcal{B}, P_X) \arrow[d, "g"] \\ \mathbb{P}_Y = (\mathbb{R}, \mathcal{B}, P_Y) \end{tikzcd}$$

$\mathbb{P} = (\Omega, \mathcal{F}, P)$ $\longrightarrow$ something like picking circles at random

$\mathbb{P}_X = (\mathbb{R}, \mathcal{B}, P_X)$ $\longrightarrow$ X is something like "radius of circle". X is a r.v.

$\mathbb{P}_Y = (\mathbb{R}, \mathcal{B}, P_Y)$ $\longrightarrow$ Y is something like "area of circle". g is like πx^2. g relates Y & X thru: $Y = g(X)$.

and the idea is to compute $P_Y(B) \equiv P_X(g^{-1}(B))$ $\rightarrow$ this is already known!

Now lets answer the question "what is the distribution of "Area of circle" given radius follows uniform dist. between $[0,1]$?"

~~Eg.~~ Let dist. function of A be F_A & that of R be F_R.

$$F_A(a) = P[A \le a] = P[g(R) \le a]$$

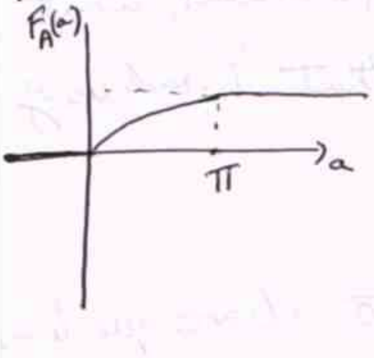
$$= P[\pi R^2 \le a] \quad \left[= P_R(\{r \in \mathbb{R} / \pi r^2 \le a\})\right]$$

$$= \begin{cases} P\left[-\sqrt{\frac{a}{\pi}} \le R \le \sqrt{a/\pi}\right] & a \geq 0 \\ 0 & a < 0 \end{cases}$$

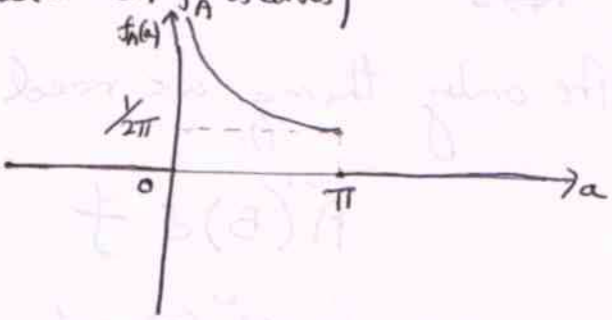
$$= \begin{cases} F_R(\sqrt{a/\pi}) & a \geq 0 \quad (\because R \geq 0) \\ 0 & a < 0 \end{cases}$$

$$= \begin{cases} 0 & a < 0 \\ \sqrt{a/\pi} & 0 \le a < \pi \\ 1 & a \geq \pi \end{cases}$$

Here's the plot:



Now $f_A(a) = \dfrac{d\, F_A(a)}{da}$ ($\forall$ a at which f_A is conts)



$$= \begin{cases} 0 & a < 0 \\ \frac{1}{2\sqrt{\pi a}} & 0 \le a < \pi \\ 0 & a \geqslant \pi \end{cases}$$

Note that the distribution of "A: area of circle" is nowhere near uniform distribution. Also, from the pdf it looks like the values near zero are "preferred" i.e. have more prob. density. This is also intuitive as R is uniform & more importantly ≤ 1! (By now, the Bertrand's Paradox also must be resolved!)

Note that the only trick is in writing the dist. function of $Y = g(X)$ in terms of dist. function of X. The above example would have showed that the ease in doing this really depends on how "simple" is $g^{-1}((-\infty, x])$ for any $x \in \mathbb{R}$.

This immediately hints on considering cases where 'g' is monotonic; because if g is monotonic, then:

$$g(X) \le a$$
$$\Leftrightarrow \begin{cases} X \le g^{-1}(a) & \text{if } g \uparrow \text{ (monotonically increasing g)} \\ X \geqslant g^{-1}(a) & \text{if } g \downarrow \text{ (" decreasing g)} \end{cases}$$

The following result is immediate:

<u>Result 1</u>: If X is a r.v & g is conts, monotonic, then the following is true for the r.v $Y = g(X)$:

$$F_Y(y) = P[Y \le y] = P[g(X) \le y]$$
$$= \begin{cases} P[X \le g^{-1}(y)] & \text{if } g \uparrow \\ P[X \geqslant g^{-1}(y)] & \text{if } g \downarrow \end{cases}$$

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if } g\uparrow \\ 1 - F_X(g^{-1}(y)) + P[X = g^{-1}(y)] & \text{if } g\downarrow \end{cases}$$

Also, the following result is true:

<u>Result 2</u>: Suppose further ~~that~~ g is diff. & X is conts. r.v., then:

$$f_Y(y) = \frac{dF_Y(y)}{dy} \qquad (\forall\, y \text{ at which } f_Y \text{ is conts.})$$

$$= \begin{cases} \dfrac{d}{dy} F_X(g^{-1}(y)) & \text{if } g\uparrow \\ -\dfrac{d}{dy} F_X(g^{-1}(y)) & \text{if } g\downarrow \end{cases} \qquad \begin{matrix}(\because X \text{ is conts}) \\ P[X = g^{-1}(y)] = 0\end{matrix}$$

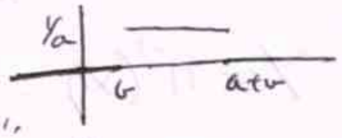
$$= \begin{cases} f_X(g^{-1}(y)) \dfrac{d\, g^{-1}(y)}{dy} & \text{if } g\uparrow \\ -f_X(g^{-1}(y)) \dfrac{d\, g^{-1}(y)}{dy} & \text{if } g\downarrow \end{cases} \qquad (\because \text{chain rule})$$

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

One can apply these results to various 'g'. Let us take the case of $g(x) = ax + b \quad (a > 0)$, which is a monotonically increasing diff. function.

$$Y = aX + b$$

(i) X is uniform between $(0,1]$ $\longrightarrow f_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{a} = \begin{cases} 0 & y < b \\ 1/a & b \le y \le a+b \\ 1 & y \ge a+b \end{cases}$



Again models "equally likely". So we can call Y as uniform distribution between $[b, a+b]$.

(ii) If X is Normal, then $f_Y(y) = \frac{1}{a\sqrt{2\pi}} e^{-\frac{(y-b)^2}{2a^2}}$.

(iii) X is uniform between $[0,1]$ & $g = H^{-1}$, where H is any distribution function ~~of a contin~~ of a contin r.v.

Note that, indeed g is monotonically increasing (since H is dist func.) ~~and in fact differential (at most points)~~ and g is also contn.

$$\boxed{Y = H^{-1}(X)}$$

From result 1, $F_Y(y) = F_X(g^{-1}(y)) = F_X(H(y))$

$= H(y)$ ($\because$ X is uniform between $[0,1]$ & H is a distn. function & hence $0 \le H(x) \le 1 \ \forall x$)

In other words, if $Y = H^{-1}(X)$ where H is a distn. func., then distn. of Y is itself H! This is way of generating random variables from a uniform r.v. itself. This can be used to generate random numbers with diff. distributions using random numbers from uniform distribution. This technique of random number generation is called "~~Inverse~~ Prob. Inverse" Technique.

However it may not be useful in practice always because H^{-1} may not be easily computable (for eg. for Normal dist.)

Here is an eg. where H^{-1} is ~~has~~ closed form solution:

Consider H as dist. of exponential r.v.

$$H(y) = 1 - e^{-\lambda y} \quad y \geq 0 \quad (\lambda > 0).$$

$$\text{Now } H^{-1}(x) = \frac{-1}{\lambda} \log(1-x)$$

$\therefore$ If one takes $Y = H^{-1}(X) = -\frac{1}{\lambda} \log(1-X)$ & X is uniform between $[0,1]$, then

$$F_Y(y) = H(y) = 1 - e^{-\lambda y} \quad !$$

This lecture introduces the concept of expectation of a r.v or expected value or mean value of a r.v. It is denoted as $E[X]$ for a r.v X.

Intuition: We know that in a random expt one cannot predict which exact value a random variable takes. However ~~in practice one usually talks about~~ one can talk about an average value or an expected value for the r.v. The ~~notion~~ concept of expected value will help us to relate to notions like "if we toss a fair coin repeatedly on an average we will see heads for half no. times" etc.

Here goes the definition:

$$E[X] = \sum_{x_i \in E} x_i \, f_X(x_i) \qquad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} x \, f_X(x) \, dx \qquad \text{(Conts. case)}$$

I

Note that, $E[X]$ is either a series sum (probably infinite sum) ~~of~~ numbers which are not necessarily +ve or an improper integral over functions which are " " +ve. ~~Our analysis knowledge~~ In such cases, the value of sum/improper integral might depend on the way we compute them. For eg. consider the Cauchy ~~dist~~ r.v. defined in Assignment problem (7a). There we showed that if one computes $\int_{-\infty}^{\infty} x f_X(x) dx$ by splitting it into $\int_{-\infty}^{0} x f_X(x) dx + \int_{0}^{\infty} x f_X(x) dx$, then the value is undefined. Whereas if we compute it taking $\int_{-\infty}^{\infty} x f_X(x) dx = \lim_{a\to\infty} \int_{-a}^{a} x f_X(x) dx = 0$.

Hence one additionally puts the condition that $E[X]$ is defined if the corresponding sum/integral is absolutely convergent.

In other words, if $\sum_{x_i \in E} |x| f_X(x)$ converges then $E[X] = \sum_{x_i \in E} x_i f_X(x_i)$

if $\int_{-\infty}^{\infty} |x| f_X(x) dx$ converges then $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$.

Once a sum/integral is absolutely convergent many properties satisfied by "usual" sums/integrals also get satisfied. We will indicate these as and when they are used.

eg: Consider $f_X(x) = \frac{1}{x^2}$, $x \geq 1$.

Here $\int_{-\infty}^{\infty} x f_X(x) dx = \int_{1}^{\infty} \frac{1}{x} dx = \infty$.

In this case the improper integral is defined and is equal to ∞ (unlike the case of Cauchy r.v, where integral itself was undefined!) However we may choose to consider whether to include the case $E[X] = \infty$ as "well-defined" or not. For the purposes of this class we can choose $E[X] = \infty$ as being "well-defined".

It is a straight-forward exercise to show that: (pls. do this exercise)

(i) $E[X] = np$ for binomial r.v.
(ii) $E[X] = \frac{1}{p}$ for geometric r.v
(iii) $E[X] = \lambda$ for Poisson r.v.
(iv) $E[X] = 1/2$ for Uniform $(0,1]$
(v) $E[X] = \frac{1}{\lambda}$ for exponential
(vi) $E[X] = 0$ for Normal r.v.

In all cases, note the intuition behind each value: (let success prob. be p)

(i) Avg. no. of success in 'n' trials is np
(ii) Avg. no. trials needed for a success is $1/p$
(iii) λ is the avg. no. successes
(iv) Center of gravity of uniform rod is at midpoint
(v) ~~Avg waiting time~~ λ is the success rate
(vi) avg. error in measurements are zero.

Now say $Y = g(X)$. One way to compute $E[Y]$ is to use the defn (I) after computing the pmf/pdf of Y. Assignment prob. 9(b) shows this can be tedious (and unnecessary).

One can directly compute $E[Y]$ using the distn. of X itself:

Theorem: If X is ~~conts~~ discrete, ~~$E[Y]$~~ $E[g(X)] = \sum_{x_i \in E} g(x_i) f_X(x_i)$

X is conts, $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

The proof is simple in the discrete case:

$$E[Y] = E[g(X)] = \sum_{y_i} y_i f_Y(y_i) \qquad \text{(let } y_i \text{ be the values } Y \text{ takes on)}$$

$$= \sum_{y_i} y_i \, P[g(X) = y_i]$$

$$= \sum_{y_i} y_i \sum_{x_i : g(x_i) = y_i} P[X = x_i]$$

$$= \sum_{y_i} \sum_{x_i : g(x_i) = y_i} g(x_i) \, f_X(x_i)$$

$$= \sum_{x_i} g(x_i) f_X(x_i)$$

$\because$ We assume $E[g(X)]$ whenever defined, the sum is absolutely convergent so order in which summation is done, does not matter!

Properties of $E[X]$: (we show them for conts r.v but also true for discrete r.v)

(1) E is a linear operator: $E[aX+b] = aE[X] + b$

$$\underbrace{E[aX+b] = \int_{-\infty}^{\infty} (ax+b) f_X(x)\,dx}_{\text{by above thm.}} \underbrace{= a\int_{-\infty}^{\infty} x f_X(x)\,dx + b\int_{-\infty}^{\infty} f_X(x)\,dx}_{\text{abs. convergence of improper integral \& integral is linear operator}} = aE[X] + b$$

where $\int_{-\infty}^{\infty} x f_X(x)\,dx = E[X]$ and $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

(2) $L \leq X \leq U \implies L \leq E[X] \leq U$

$P[L \leq X \leq U] = 1$

Proof of

$X \leq U \implies E[X] \leq U$:

$$E[X] = \underbrace{\int_{-\infty}^{\infty} x f_X(x)\,dx \leq \int_{-\infty}^{\infty} U f_X(x)\,dx}_{\text{abs. conv. of improper integral \& monotonicity of integral prop.}} = U$$

|||ly $L \leq X \implies L \leq E[X]$.

hence $L \leq X \leq U \implies L \leq E[X] \leq U$

Now its easy to prove:

$\underbrace{P[L \leq X \leq U] = 1}_{X \text{ is almost surely between } [L, U]} \implies L \leq E[X] \leq U$

(3) The "best" (optimal) constant value approximation of a rv X which minimizes the "average squared error" in approximation is $E[X]$

In other words $E[X] = \underset{c}{\operatorname{argmin}}\ E\left[(X-c)^2\right]$

Proof

$$\underset{c}{\operatorname{argmin}}\ E\left[(X-c)^2\right] = \underset{c}{\operatorname{argmin}}\ E\left[(X^2 - 2cX + c^2)\right]$$

$$= \underset{c}{\operatorname{argmin}}\ E[X^2] - 2cE[X] + c^2$$

repeated application of linearity property of E

$$= \underset{c}{\operatorname{argmin}}\ (c - E[X])^2 + E[X^2] - (E[X])^2$$

$$= E[X]$$

Now, the minimized error is called as variance of X denoted by $var(X)$.

i.e. $var(X) = \min_{c} E\left[(X-c)^2\right] = E\left[(X - E[X])^2\right]$

$= E[X^2] - (E[X])^2$ } by above proof

Now one can compute $var(X)$ of various rvs discussed in this course:

(i) $var(X) = np(1-p)$ for binomial

(ii) $var(X) = \frac{1-p}{p^2}$ for geometric

(iii) $var(X) = \lambda$ for Poisson

(iv) $var(X) = 1/3$ for Uniform $[0,1]$

(v) $var(X) = 1/\lambda^2$ for exponential rv

(vi) $var(X) = 1$ for Normal rv.

Properties of var(X):

① By the very definition of var(x), it is E of a non negative r.v (which is $(X-E[X])^2$). Hence $\underline{\text{var}(X) \geq 0}$.

Now, $\text{var}(X) = E\left[(X-E[X])^2\right] = E[X^2] - (E[X])^2 \geq 0$

$$\Rightarrow \quad \underline{E[X^2] \geq (E[X])^2} \qquad \text{—— (A)}$$

This is a very important inequality and is a specific case of the following inequality:

$$E[g(X)] \geq g(E[X]) \quad \forall\, g \text{ convex}$$

which is ~~for~~ known as the Jensen's inequality. (In (A), $g(x)=x^2$)

(fundamental inequalities like AM ≥ GM are special cases of this inequality)

② $\text{var}(X+b) = E\left[\left(X+b-E[X+b]\right)^2\right] = E\left[(X-E[X])^2\right] = \text{var}(X)$

$\underbrace{}_{\text{linearity of } E}$

③ $\text{var}(aX) = E\left[(aX-E[aX])^2\right] = E\left[a^2(X-E[X])^2\right] = a^2\,\text{var}(X)$

$\underbrace{}_{\text{linearity of } E}$

Variance is not a linear operator

Now let $Y = \sigma X + \mu$ where X is a Normal r.v. (σ, μ are some numbers)

$E[Y] = E[\sigma X + \mu] = \sigma E[X] + \mu = \mu \quad (\because E[X]=0)$

$\text{var}(Y) = \text{var}(\sigma X+\mu) = \text{var}(\sigma X) = \sigma^2\,\text{var}(X) = \sigma^2 \quad (\because \text{var}(X)=1)$

We also know, $f_Y(y) = \dfrac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}.$

This is our new defn of Normal (Gaussian) r.v. with mean μ & variance σ^2.

We call the case $\mu=0$ & $\sigma^2=1$ as std. Normal r.v.

Median of a r.v. is that number for which $P[X \le M] = 1/2$.
Assignment problem shows that

$$M = \underset{c}{\text{argmin}}\ E[|X-c|]$$

In other words median is that value that minimizes the absolute error in approximating a r.v by a constant.

Mode of a r.v. is the "most frequently taken value" of a r.v.
Let Mode of X be m. Mathematically,

$$m = \underset{c}{\text{argmin}}\ E\left[1_{\{X \ne c\}}\right].$$

$$1_{\{X\ne c\}} = \begin{cases} 1 & \text{if } X \ne c \\ 0 & \text{if } X = c \end{cases}$$

mode minimizes the average number of times X does'nt take its value. In other words, maximizes the avg. no. times X takes the particular value.

Now, $E[1_{\{X\ne c\}}] = 1\,P[X \ne c] + 0\,P[X=c] = P[X\ne c]$

In discrete case,
$$m = \underset{c}{\text{argmin}}\ P[X\ne c] = \underset{c}{\text{argmax}}\ P[X=c]$$

i.e. Mode is the value with highest prob. of occuring.
Analogously for the conts. case we have $m = \underset{c}{\text{argmax}}\ f_X(c)$.
__For eg.__ std. Normal r.v. is "Unimodal" with '$m=0$'.
(has one max $f_X(c)$)
Each such peak in pdf/pmf is called mode (Loosely).
Mean, median need not be values taken by X, whereas mode must be a value taken by X.

This lecture completes our discussion on single r.v.s by discussing concepts of generic moments, moment generating function and few important inequalities like Jensen's, Markov and Chebyshev's inequalities.

Moments & Moment Generating Function

Encouraged by the notions of mean & variance etc. we now define some higher order moments of r.v. as follows:

$E[X^n]$ is called the 'n'th moment of X (first moment is mean (centre of gravity), second " is moment of inertia etc.)

$E[(X-E[X])^n]$ is called the 'n'th central moment of X ($n=2$ is variance (central moment of inertia etc.))

$E[(X-a)^n]$ is the 'n'th generalized moment of X about 'a' (eg. moment of inertia abt some axis)

We can define the absolute value versions of these:

$E[|X|^n] \longrightarrow n^{th}$ absolute moment

$E[|X-E[X]|^n] \longrightarrow n^{th}$ absolute central moment

$E[|X-a|^n] \rightarrow n^{th}$ absolute moment about 'a'. and so on...

Now consider a function M_X defined as follows:

$$M_X(s) = E[e^{sX}]$$

This function is known as the moment generating function (wherever it exists!) assuming $E[e^{sX}]$ exists. (eg. of conts. case:)

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx = \int_{-\infty}^{\infty} \left(1 + sx + \frac{s^2x^2}{2!} + \cdots\right) f_X(x)\,dx \quad \text{— because of abs. convergence of } E[e^{sX}]$$

$$= \int_{-\infty}^{\infty} f_X(x)\,dx + s\int_{-\infty}^{\infty} x f_X(x)\,dx + \frac{s^2}{2!}\int_{-\infty}^{\infty} x^2 f_X(x)\,dx + \cdots$$

$$\Rightarrow \quad M_X(s) = 1 + sE[X] + \frac{s^2}{2!}E[X^2] + \dots + \frac{s^n}{n!}E[X^n] + \dots$$ ②

(This is why it is called as moment generating function!)

Also it is easy to see that $\left.\frac{d}{ds}M_X(s)\right|_{s=0} = E[X]$

$$\left.\frac{d^2}{ds^2}M_X(s)\right|_{s=0} = E[X^2] \text{ and so on} \dots$$

In general, $\left.\frac{d^n M_X(s)}{ds^n}\right|_{s=0} = E[X^n]$.

In other words, mgf is a MacLaurin series with coeff. given by

In general, mgf may not exist (for eg. take case of Cauchy distribution where we know first moment itself doesn't exist!). But whenever it exists, by the above relations all moments exist (and are finite).

However the converse statement that if all moment exist then the mgf also exists may not be true in general.

(Take the log Normal dist defined as $Y = e^X$ where X is std. Normal and try!)
(You will notice that all moments exist but mgf does not.

Lets compute mgf for Poisson r.v:

$$M_X(s) = E[e^{sX}] = \sum_{k=0}^{\infty} e^{sk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^s\lambda)^k}{k!} = e^{-\lambda}e^{e^s\lambda} = e^{e^s\lambda - \lambda}$$

mgf for std. Normal r.v:

$$M_X(s) = E[e^{sX}] = \underbrace{\int_{-\infty}^{\infty} e^{sx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx}_{\text{completing quadratic}} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{s^2}{2} - \frac{(s-x)^2}{2}}dx = \underbrace{\frac{e^{s^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-t^2/2}dt}_{t = s-x} = e^{s^2/2}$$

$$\left(\because \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-t^2/2}dt = 1\right)$$

$$\therefore M_X(s) = e^{s^2/2}$$
for std. Normal

mgf for Normal rv: $Y = \sigma X + \mu$

$$M_Y(s) = E[e^{sY}] = E[e^{\sigma s X + s\mu}] = \underbrace{E[e^{s\sigma X}]}_{\text{mgf of std. Normal}} e^{s\mu} = e^{\sigma^2 s^2/2} e^{s\mu} = e^{s\mu + \frac{1}{2}s^2\sigma^2}$$

$$\Rightarrow \boxed{M_Y(s) = e^{s\mu + \frac{1}{2}s^2\sigma^2}} \quad \text{I}$$

Apart from fact that mgf "generates" moments there is an important application of it: mgf also characterize rv! In other words if somebody ~~proves that~~ proves/ascertains that a certain alien rv has mgf for eg as $e^{s\mu + \frac{1}{2}s^2\sigma^2}$, then certainly that alien rv must be a Normal rv.

[Intuitively, here's the reason why mgf characterizes a rv: Take $s = j\omega$ then mgf is nothing but the Fourier transform of $f_X(x)$! $\longrightarrow M_X(j\omega) = E[e^{j\omega X}] = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\,dx$. Hence charaterizing f_X is equivalent to characterizing mgf's]

Let's calculate n^{th} moment of a log-Normal rv: $Y = e^X$, X is Normal rv.

$$E[Y^n] = E[(e^X)^n] = \underbrace{E[e^{nX}] = e^{n\mu + \frac{1}{2}n^2\sigma^2}}_{\text{by I}}$$

This shows that log-Normal has all moments! (but as said earlier doesn't have an mgf)

Jensen's Inequality

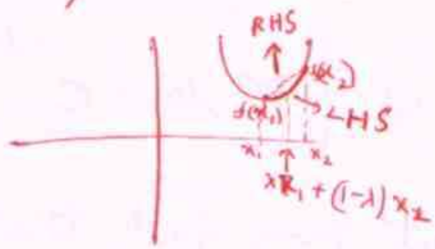
$$E[f(X)] \geq f(E[X]) \quad \forall X,\ f \text{ convex on } \mathbb{R}.$$

In assignment we saw a lengthy and restricted proof of this inequality following from the very defn. of a convex function. Now lets look at diff characterization of a convex function which leads to a simple proof.

Characterization of Convex functions

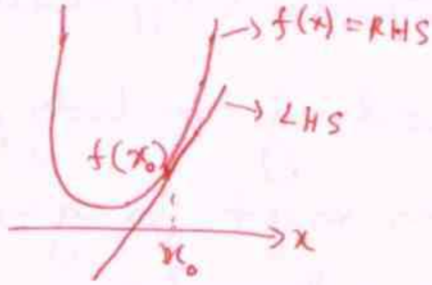
$\longrightarrow f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\quad \lambda \in [0,1]$

This is the definition. $\xrightarrow{\text{intuition}}$



$\longrightarrow f(x) \geq f(x_0) + \frac{d f(x_0)}{dx}(x - x_0)$

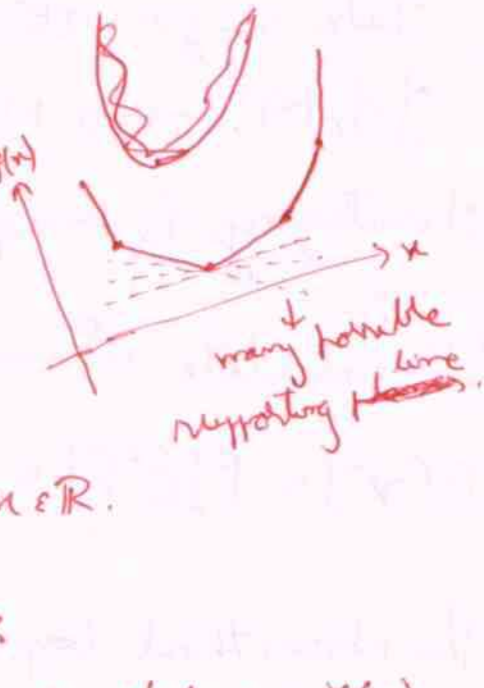
In words, the linear approx. at x_0 always under-estimates the function $\xrightarrow{\text{intuition}}$



But this view is limited to differentiable convex functions only.

$\rightarrow$ Above intuition holds in all convex function (even though not differentiable)

i.e. At every pt. there is a supporting line



Mathematically,

f is convex on $\mathbb{R}$ $\iff \exists$ a $\lambda(x)$ such that

$$f(x) \geq f(x_0) + \lambda(x_0)(x - x_0) \quad \forall x \in \mathbb{R}.$$

$\downarrow$

This is the characterization which we use now:

X is a r.v. $\rightarrow$ in other words a mapping from $\Omega \rightarrow \mathbb{R}$. Hence take $x = X(\omega)$.

Take $x_0 = E[X]$.

$\Rightarrow f(X(\omega)) \geq f(E[X]) + \lambda(E[X])(X(\omega) - E[X]) \quad \forall \omega \in \Omega.$

Now we know that expectation maintains order relations

$\Rightarrow E[f(X)] \geq E[f(E[X]) + \lambda(E[X])(X - E[X])]$

$= f(E[X]) + \cancel{\lambda E[X] (E[X] - E[X])} = f(E[X])$

Hence Proved.

Applications of Jensen's inequality:

(i) Take $f(x) = -\log(x) \to$ convex. Take X as discrete r.v. taking values $\underbrace{x_1, \dots x_n}_{\geq 0}$ & Uniform distribution.

$$f(E[X]) \leq E[f(X)]$$

$$\Rightarrow -\log\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \leq \frac{-\log(x_1) + \dots + -\log(x_n)}{n}$$

$$= -\frac{1}{n}\log(x_1 x_2 \dots x_n)$$

$$= -\log \sqrt[n]{x_1 x_2 \dots x_n}$$

$$\Rightarrow \frac{x_1 + x_2 + \dots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \dots x_n}$$ ($\because$ $-\log$ is a monotonically decr. function)

$$AM \geq GM.$$

So Jensen's inequality is a generalization of AM, GM inequality.

(ii) TST $|\mu - M| \leq \sigma$ (μ is mean, σ is std. div., M is median).

$$|\mu - M| = |E[X] - M| = \underbrace{|E[X-M]| \leq E[|X-M|]}_{\substack{\text{Jensen's Ineq.}\\ \text{with } |\cdot| \text{ as convex}\\ \text{function!}}} \cancel{= E[\sqrt{(X-M)^2}]} \; \cancel{\leq \sqrt{E[(X-M)^2]}}$$

$$= \min_{c} E[|X-c|] \to \text{we saw this in assignment}$$

$$\leq E[|X-c|] \quad \forall c \text{ in particular take } c = \mu = E[X]$$

$$= E[|X - E[X]|]$$

$$= \underbrace{E\left[\sqrt{(X-E[X])^2}\right] \leq \sqrt{E[(X-E[X])^2]}}_{\substack{\text{Jensen's Inequality}\\ \text{with } \sqrt{\cdot} \text{ as concave function!}}} = \sigma$$

Hence Proved.

Occurs frequently in many places for eg: Information theory (cross-entropy etc.

Till now we have been looking at random variables which take on real values. In other words, the range of r.v. was always $\mathbb{R}$. Now, we will generalize the notion of r.v.s to include ones taking on vectorial values i.e. rvs for which the range is $\mathbb{R}^n$, the n-dimensional Euclidian Space.

i.e. We define functions of the form $X : \Omega \rightarrow \mathbb{R}^n$. Such functions from some sample space to $\mathbb{R}^n$ (with additional restrictions as we shall see as we proceed) are called as Multi-variate Random variables. Other names are random vector, multi-valued random variable so on...

Intuitively a (usual) random ~~vector~~ variable quantifies outcomes in terms of numbers (scalars) whereas a ~~random vector~~ multi variate r.v. (m.r.v.) quantifies outcomes in terms of n-tuples (vectors). Once this view is clear, the applications where an m.r.v. can be employed are obvious: eg. wherever the outcome of a random experiment can be defined in terms of vectorial values rather than scalar values.

To give a realistic example, let us consider the random experiment where people in IITB are clinically examined for presence of Swine-Flu. Here it is immediate the each person's (an outcome in our case) health cannot be described by a single quantity such as temperature or cough etc., but can be described using a collection of all these data!

Let us run through this example:

Let Ω be the set of all people (living) in IITB (say N of them).
~~The set~~ An event in Ω is nothing but groups of people (take $\mathcal{F} = 2^{\Omega}$).
Now define $P(\{x_i\}) = \frac{1}{N}$ $\forall x_i \in \Omega$ (x_i is the i^{th} person).
This gives a valid prob. space $\mathbb{P} = (\Omega, \mathcal{F}, P)$.

Now define a r.v. X_1 which is nothing but a 'thermometer'.
(Thermometer takes input as a patient gives output as a number, specifically the body temperature of that patient). Let B_1 be the set of all "high temperatures" i.e. $B_1 = \{x \in \mathbb{R} / x \geq 103\}$

Now we saw, that, $P_{X_1}(B_1) = P(X_1^{-1}(B_1)) = P(\underbrace{\{\omega \in \Omega / X_1(\omega) \geq 103\}}_{\text{set of all people with high temp.}})$

This gives us the induced prob. space $\mathbb{P}_{X_1} = (\mathbb{R}, \mathcal{B}, P_{X_1})$.

IIIly ~~for~~ each symptom of swine flu $\left(\begin{array}{l}\text{which is of course}\\ \text{quantifiable}\\ \text{as a number}\end{array}\right)$ we can represent it with a r.v.

Let $X_1, X_2, \ldots, X_{n-1}$ be r.vs representing "clinical devices" quantifying each symptom of swine-flu disease. Now consider an expert doctor who looks at the diagnostic report of patient ω (i.e. looks at $X_1(\omega), \ldots, X_{n-1}(\omega)$) and certifies presence of swine flu or not. ~~In other words, X~~ Let X_n r.v. represent the expert doctor (again he takes as input a patient $\omega \in \Omega$ & gives as output a number 1 (if swine-flu presence) or 0 (~~else~~ for normal patient)).
In other words X_n is the indicator fuction of presence of disease.
(Note that X_n depends "implicitly" on all $X_1, \ldots, X_{n-1}$).

Let $B_n \in \mathcal{B}$ be an event of drawing swine flu i.e. $B_n = \{1\}$.

Now let us define a m.r.v. X as follows:

$$X:\Omega\to\mathbb{R}^n \text{ such that,}$$

$$X(\omega)=\big(X_1(\omega), X_2(\omega), X_3(\omega), \dots, X_{n-1}(\omega), X_n(\omega)\big) \quad \forall \omega\in\Omega.$$

(*Note that intuitively $X(\omega)$ is nothing but an analyzed diagnostic report of patient ω).

Note that X (which is an m.r.v.) not only helps in representing a "complicated" outcomes like health of a person, but also helps in analyzing relationships between X_i, X_j r.v.s !!

~~Now let us see if there is a concept of induced probability for a m.r.v. :~~ Let us see how do events in $\mathbb{R}^n$ look like:

~~In order to do that let~~ An event in $\mathbb{R}^n$ looks like:

$$B = B_1\times B_2\times\dots\times B_n \qquad \text{where } B_i\in\mathcal{B}\ \forall i$$

$$=\{(x_1,x_2,\dots,x_n) \mid x_i\in B_i,\ B_i\in\mathcal{B}\}$$

In our medical eg: B is nothing but high temperature values in first coordinate, ... , presence of swine flue i.e. 1 in the last coordinate.

Now collection of all such events B in $\mathbb{R}^n$ is the Borel-σ-algebra in $\mathbb{R}^n$.

$$\underset{\substack{\text{(Borel } \sigma\text{-algebra}\\ \text{in } \mathbb{R}^n)}}{\mathcal{B}^n} = \{B \mid B=B_1\times B_2\times\dots\times B_n,\ B_i\in\mathcal{B}\ \forall i\}$$

Let us now see if there exists a concept of induced probability for a m.r.v. :

i.e. can we define $P_X(B)$?

Following a strategy similar to the case of (usual) r.v. we have:

$$P_X(B) \equiv P(\{\omega\in\Omega \mid X(\omega)\in B\}) \xrightarrow{\text{Nothing but}} P(X^{-1}(B)) = P[X\in B]$$

square bracket notation

$$= P(\{\omega\in\Omega \mid \underbrace{(X_1(\omega),X_2(\omega),\dots,X_n(\omega))\in B}_{B_1\times B_2\times\dots\times B_n}\})$$

$$= P(\{\omega\in\Omega \mid X_1(\omega)\in B_1,\ X_2(\omega)\in B_2,\ \dots,\ X_n(\omega)\in B_n\})$$

$$= P\Big(\bigcap_{i=1}^{n}\{\omega\in\Omega \mid X_i(\omega)\in B_i\}\Big) \quad \text{Notation}$$

$$= P[X_1\in B_1,\ X_2\in B_2,\ \dots,\ X_n\in B_n] \qquad \text{(I)}$$

In words, (in case of medical eg.) $P_X(B)$ is nothing prob. of observing high temp (first coordinate), ..., swine flu (last co-ordinate).

$P[\cancel{X\in}$ $P(\bigcap_{i=1}^{n}\{\omega\in\Omega / X_i(\omega)\in B_i\})$ is nothing but the prob. of a person having high temp. , ..., & having swine flu. This is indeed intuitive. Since P models classical prob in our medical eg., this is exactly the "fraction of people in IITB having all symptoms of swine-flu & also have swine flu!". So In future classes, based on this we will try to ~~see~~ answer ~~for~~ ~~which~~ questions like which symptoms are crucial for swine-flu, how to predict presence of swine flu given a raw diagnostic report (i.e. guess values of X_n given raw values of $X_1,\dots,X_{n-1}$) and so on & so forth!

Math. detail ~~Tech. question~~:

Q: Why does $\bigcap_{i=1}^{n}\{\omega\in\Omega / X_i\in B_i\}\in\mathcal{F}$? (unless this happens my induced prob. defn. is invalid!)

A: I know each of ⤴ for fixed i $\in\mathcal{F}$ ($\because X_i$ is a RV)

So intersection of them also $\in\mathcal{F}$ ($\because \mathcal{F}$ is a σ-algebra)

Now that we have correctly defined Induced prob. of a m.r.v., let us Extend the concept of "distribution function" to an m.r.v.:

Distribution functions of m.r.v.

In case of r.v. we defined $F_X : \mathbb{R} \to \mathbb{R}$ such that

$$F_X(x) = P[X \leq x].$$

Now we will define in an analogous way for an m.r.v. (Note that range of m.r.v is $\mathbb{R}^n$):

$$F_X : \mathbb{R}^n \to \mathbb{R} \ni F_X(\underbrace{\underline{x}}_{\in \mathbb{R}^n}) = P[X \leq \underline{x}]$$

$\underline{x} = (x_1, x_2, \ldots, x_n)$

~~for~~ Using (I) we get:

$$F_X(\underline{x}) = P\{X \leq \underline{x}] = P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n].$$

called the prob. dist. func. of m.r.v. X & it is also called as joint prob. dist. func. of $X_1, X_2, \ldots, X_n$ (in this case it is represented as $F_{X_1 X_2 \ldots X_n}(x_1, x_2, x_3, \ldots, x_n)$)

Now as in case of ~~single~~ F_X of a r.v. we can show the following 4 prop. for F_X of a m.r.v. also:

i) $F_X(\underline{x}) \geq 0 \ \forall \underline{x}$ (after all its a prob.)

ii) $F_X(\underline{\infty}) = P[X_1 \leq \infty, X_2 \leq \infty, \ldots, X_n \leq \infty] = P[X \in \mathbb{R}^n] = P(\Omega) = 1.$

$F_X(-\underline{\infty}) = P[X_1 \leq -\infty, X_2 \leq -\infty, \ldots, X_n \leq -\infty] = P(\{\phi\}) = 0.$

represents vector with all values $\infty / -\infty$.

This not only shows that ~~P[X∈]~~ prob. of events can be computed in terms of dist function but also the extra cond. ~~that~~ $F_{X_1X_2}(b_1,b_2) - F_{X_1X_2}(a_1,b_2) - F_{X_1X_2}(b_1,a_2) + F_{X_1X_2}(a_1,a_2) \geq 0$.

Now (II) ~~can be proved this (cumbersome) inductively using these ideas.~~ is a generalization of this to a n-dimensional case.

Now using set algebra we can show (not in this class) that $P[X \in B]$ can be written in terms of F_X. So we from now onwards characterize a m.r.v. $X = [X_1 X_2, \dots X_n]$ using (F_X or $F_{X_1X_2\dots X_n}$) dist. function.

$B \in \mathcal{B}^n$

Lets look at the case where all X_i's are discrete. Then we can define a prob. mass function (pmf) for the m.r.v. X: (analogous to discrete r v's case):

$$f_X(\underline{x}) = \overset{\text{pmf of } X}{f_X(x_1, x_2, \dots, x_n)} = \overset{\text{joint pmf of } X_1, X_2 \dots, X_n}{f_{X_1X_2\dots X_n}(x_1, x_2, \dots, x_n)}$$

$$\equiv P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n]$$

Now let E be the set of (discrete) values in $\mathbb{R}^n$ taken by X, then following two properties of pmf are immediate:

$$f_X(\underline{x}) \geq 0 \qquad (\because f_X \text{ is after all a prob.})$$

$$\sum_{\underline{x} \in E} f_X(\underline{x}) = 1 \qquad (\because P(\Omega) = 1)$$

Now again any function ~~satisfying~~ $f_X : \mathbb{R}^n \to \mathbb{R}$ satisfying these two properties is called a pmf. & also given a pmf, of X, distr. of X (F_X) is fixed and vice-versa. So we can characterize discrete m.r.v. using pmf. (from now-onwards).

Lets. look at an eg. of a discrete m.r.v. {which is a generalization of the binomial r.v.}:

Multinomial R.V.

Suppose we ~~d~~ define the following $f_X : \mathbb{R}^n \to \mathbb{R}$,

$$f_X(\underline{x}) = P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \frac{n!}{x_1! x_2! \ldots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \quad \text{— III}$$

$$\forall \; x_i \geq 0, \sum_{i=1}^{n} x_i = n$$

and $(n, p_1, \ldots p_n)$ are parameters such that $n \in \mathbb{N}$ and $p_i \geq 0, \sum_{i=1}^{n} p_i = 1$.

(It is an exercise to first check if this is a valid pmf!)

At first look this might look weird but consider the following random expt: Suppose I throw a die n times. In each throw I have ($n = 6$) outcomes. Suppose p_i is probability of seeing no. 'i' ($i = 1$ to 6). Now the answer to the question: "what is the prob. of seeing x_1 1s, x_2 2s, ..., x_6 6s" is exactly given by III ! (why?)

Now that we know some 'physical' interpretation of multinomial r.v. lets see (if at all) what kind of r.v.'s are $X_1, X_2, \ldots, X_n$ individually?

In the die throwing case, $X_i = \#$ throws in which 'i' was observed

Now it is easy to see that X_i is a binomial r.v. with parameters (n, p_i) !.

So when $X = [X_1\ X_2 \ldots X_n]$ follows multinomial distribution then each of X_i $(i=1 \text{ to } n)$ follow binomial distri. (with diff. parameters) -

Now we ~~suppose~~ can compute prob like

$$P[X_i = x_i] \quad \text{using the pmf of } X_i.$$

But note that $P[X_1 = x_1, \ldots, X_n = x_n]$ (which is nothing but the joint pmf of $X_1, X_2, \ldots X_n$), cannot be computed merely from the knowledge of $P[X_i = x_i]$'s. (atmost you can give bounds on joint pmf using inequalities like Bonferroni's inequality you proved in the assigns.)

However the pmf of X_i can be computed given the pmf of X (i.e. the joint pmf of X_i's):

$$\sum_{\substack{x_2, x_3, \ldots, x_n \\ \ni\ x_i \geq 0,\ \sum_{i=1}^{n} x_i = n}} f_X(\underset{\text{fixed}}{x_1}, \underbrace{x_2, \ldots x_n}_{\text{all varied for all allowed values}}) = f_{X_1}(x_1).$$

So ~~pmf~~ specifying joint pmf is a "richer" information! (than specifying pmf of X_i's alone)

(iii) "F_X is right conts. & has left limit"
(we dont show this here)

Missing Page 6
from Lecture-10

(iv) Monotonicity: (but in all variables)

$$F_X(x_1+\underset{\geq 0}{\Delta x_1}, x_2+\underset{\geq 0}{\Delta x_2}, \ldots, x_n+\underset{\geq 0}{\Delta x_n}) - F_X(x_1, x_2, \ldots, x_n)$$

$$= P[X_1 \leq x_1+\Delta x_1, \ldots, X_n \leq x_n+\Delta x_n] - P[X_1 \leq x_1, \ldots, X_n \leq x_n]$$

$$= P[x_1 \leq X_1 \leq x_1+\Delta x_1, \ldots, x_n \leq X_n \leq x_n+\Delta x_n]$$

≥ 0 ($\because$ it is a prob. of some event).

$\longrightarrow$ Till this all prop. are ~~common~~ analogous to those in case of r.v.

But in case of m. r.v. an extra condition needs to be satisfied:

(v)

$$F_X(x_1+\epsilon_1, x_2+\epsilon_2, \ldots, x_n+\epsilon_n) - \sum_i F_X(x_1+\epsilon_1, \ldots, x_i, \ldots x_n+\epsilon_n)$$

$$+ \sum_i \sum_{j>i} F_X(x_1+\epsilon_1, \ldots, x_i, x_{i+1}+\epsilon_{i+1}, \ldots, x_j, x_{j+1}+\epsilon_{j+1}, \ldots, x_n+\epsilon_n)$$

$$\vdots$$

$$(-1)^n F_X(x_1, x_2, \ldots, x_n) \geq 0 \qquad \forall\, x_i, \epsilon_i > 0. \qquad \text{II}$$

We can get an intuition for this by looking at a r.v. taking values in $\mathbb{R}^2$:

Let $X = [X_1 \;\; X_2]$. Let $F_{X_1,X_2}(x_1,x_2)$ be the joint prob. dist. function at (x_1,x_2).

Now suppose I want to compute

$$0 \leq P[a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2]$$ in terms of $\uparrow$ then:

$$= F_{X_1,X_2}(b_1,b_2) - F_{X_1,X_2}(a_1,b_2) - F_{X_1,X_2}(b_1,a_2) + F_{X_1X_2}(a_1,a_2)$$

(how we get this was explained in class)

Lecture-11

Let $X = [X_1\ X_2 \ldots X_n]$ be a m.r.v. X is called a continuous m.r.v. (or equivalently $X_1, X_2, \ldots, X_n$ are said to be "jointly continuous r.v.s") iff there exists a function $f_X : \mathbb{R}^n \to \mathbb{R}$ such that:

$$P[X \in B] = \underbrace{\int_B f_X(\underline{x})\, d\underline{x}}_{\substack{\text{multidimensional}\\ \text{integral}}} \quad \forall\, B \in \mathcal{B}^n .$$

Such a function f_X is called the prob. density function of X or joint prob. density function of $X_1, X_2, \ldots, X_n$.

For eg. if $n=2$, $P[X \in B] = \iint_B f_{X_1 X_2}(x_1, x_2)\, dx_1\, dx_2$.

Now take $B = \{(a_1, a_2, \ldots a_n) \mid a_i \in (-\infty, x_i]\}$

$$\Rightarrow \underbrace{P[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n]}_{F_X(x_1, x_2, \ldots, x_n)} = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} f_X(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

Hence F_X is fixed if f_X is known. ~~One can show the converse also~~ Also we have:

$$\frac{\partial^n F_X(x_1, x_2, \ldots, x_n)}{\partial x_1\, \partial x_2 \cdots \partial x_n} = f_X(x_1, x_2, \ldots, x_n) \quad \text{(wherever } f_X \text{ is } \cancel{\text{conts}} \text{ continuous)}$$

As in case of 1-d r.v., the values of f_X (where f_X is disconts.) does'nt matter (they do not account for the area). One can show that (again, not here) that such points are "few", so we can make the statement "given F_X, we have f_X fixed and vice-versa".

So from now on we characterize X by f_X (pdf)

Now lets look at some properties of f_X:

we have, (i) $1 = P(\Omega) = P[X \in \mathbb{R}^n] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_X(x_1, x_2, \ldots x_n)\,dx_1\,dx_2\ldots dx_n$

(ii) we can show $f_X(x_1, x_2, \ldots, x_n) \geq 0 \;\; \forall \underline{x} \in \mathbb{R}^n$. Recall that an analogous statement in case of r.v. followed from "monotonicity property of F_X". Here it follows from the "(v)th prop. of F_X" which in 2-d case is illustrated below:

We know, $F_{X_1X_2}(b_1, b_2) - F_{X_1X_2}(a_1, b_2) - F_{X_1X_2}(b_1, a_2) + F_{X_1X_2}(a_1, a_2) \geq 0$.

$$\Leftrightarrow \underbrace{\int_{-\infty}^{b_1}\int_{-\infty}^{b_2} f_X(x_1,x_2)\,dx_1\,dx_2 - \int_{-\infty}^{a_1}\int_{-\infty}^{b_2} f_X(x_1,x_2)\,dx_1\,dx_2}_{} - \underbrace{\int_{-\infty}^{b_1}\int_{-\infty}^{a_2} f_X(x_1,x_2)\,dx_1\,dx_2 + \int_{-\infty}^{a_1}\int_{-\infty}^{a_2} f_X(x_1,x_2)\,dx_1\,dx_2}_{} \geq 0 \quad \forall\, a_1 \leq b_1,\ a_2 \leq b_2$$

$$\Leftrightarrow \int_{a_1}^{b_1}\int_{-\infty}^{b_2} f_X(x_1,x_2)\,dx_1\,dx_2 - \int_{a_1}^{b_1}\int_{-\infty}^{a_2} f_X(x_1,x_2)\,dx_1\,dx_2 \geq 0 \quad \forall a_1 \leq b_1,\ a_2 \leq b_2$$

$$\Leftrightarrow \int_{a_1}^{b_1}\int_{a_2}^{b_2} f_X(x_1,x_2)\,dx_1\,dx_2 \geq 0 \quad \forall a_1 \leq b_1,\ a_2 \leq b_2 \Leftrightarrow f_X(\underline{x}) \geq 0 \;\; \forall \underline{x}$$

Hence, we may pdf is any function that satisfies:

(i) $f_X(\underline{x}) \geq 0 \;\; \forall \underline{x}$

(ii) $\int\int\cdots\int_{-\infty}^{\infty} f_X(\underline{x})\,d\underline{x} = 1$

Now lets look at a particular eg. of a conts. m.r.v.:

$$f_X(\underline{x}) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}\underline{x}^T\underline{x}} \quad \forall \underline{x} \in \mathbb{R}^n. \qquad \text{(I)}$$

First lets check if it is pdf?

f_X is indeed non-negative. Only non-trivial thing to verify is if it integrates to unity:

$$\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\sum_{i=1}^{n}x_i^2}}_{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_i^2}}dx_1\,dx_2\cdots dx_n$$

$$=\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_1^2}dx_1\right)\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2}dx_2\right)\cdots\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_n^2}dx_n\right)=1$$

(each of integral is 1).

Now, $F_{X_1X_2\cdots X_n}(x_1,x_2,\ldots,x_n)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\sum x_i^2}dx_1\,dx_2\cdots dx_n$

$$=\left(\int_{-\infty}^{x_1}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_1^2}dx_1\right)\cdots\left(\int_{-\infty}^{x_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_n^2}dx_n\right)$$

$$=\underbrace{F_{X_1}(x_1)\,F_{X_2}(x_2)\cdots F_{X_n}(x_n)}_{\text{each is the distribution function of the std. Normal r.v.!!}}$$

There are two things to note about (I):

(i) Its distribution func. is product of dist. func. of individual r.v.s (Later on we will see that such r.v.'s are called as independent r.v.s)

(ii) ~~Each~~ Dist. func. of each individual r.v. is the std. Normal dist.

Now both in case of discrete and conts. m.r.v., we saw that the distribution functions or pmfs or pdfs of individual r.v.s can be obtained from their joint-distribution. This leads to the notion of ~~p~~ Marginal distributions:

~~Now $F_{X_1 X_2}$~~ From now onwards to simplify the notation we will consider collections of two r.v.s. However keep in mind that ~~the~~ analogous results do hold in the generic n-d case also.

So from now onwards consider two r.v. X, Y. F_{XY} is the joint dist. function of X and Y.

Now, $F_X(x) = P[X \le x] = P[X \le x, Y \le \infty] = F_{XY}(x, \infty) \quad \forall x$

|||ly, $F_Y(y) = F_{XY}(\infty, y) \quad \forall y$.

The dist. functions of X/Y are ~~k~~ also known as the marginal dist. of X/Y wrt the joint dist. function of X and Y.

Now if X, Y are discrete,

$$f_X(x) = P[X=x] = \sum_{\forall y} P[X=x, Y=y] = \sum_{\forall y} f_{XY}(x,y) \quad \forall x$$

$$|||\text{ly} \quad f_Y(y) = \sum_{\forall x} f_{XY}(x,y) \quad \forall y$$

Again f_X, f_Y are known as the marginal pmfs of X and Y wrt. to the joint pmf of X, Y i.e. f_{XY}.

Now suppose X, Y are jointly conts:

$$\Rightarrow \quad F_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x',y')\,dx'\,dy'.$$

Now
$$\left.\begin{aligned} F_{XY}(x,\infty) &= \int_{-\infty}^{x}\underbrace{\int_{-\infty}^{\infty} f_{XY}(x',y')\,dy'}_{\downarrow}\,dx' \\ &\;\; \| \\ \text{Also,}\quad F_X(x) &= \int_{-\infty}^{x} f_X(x')\,dx' \end{aligned}\right\} \text{(II)}$$

Since the pdf is fixed (except at "few" points) given the dist. fun., we have that
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y')\,dy' \quad \text{from (II)}$$

|||ly
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x',y)\,dx'$$

Again f_X, f_Y are known as the marginal pdfs of X, Y wrt. to the joint-pdf of X and Y i.e. f_{XY}.

Now, it is easy to see that in example (I), the following holds:
$$f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n) = \underbrace{f_{X_1}(x_1)\,f_{X_2}(x_2)\cdots f_{X_n}(x_n)}_{\substack{\text{each is pdf of}\\ \text{std. Normal.}}}$$

Now lets run through the calculation of marginal pdfs using a toy-example:

$$\text{Let } f_{XY}(x,y) = \begin{cases} 24xy & \text{if } 0<x,\ 0<y,\ 0<x+y<1 \\ 0 & \text{otherwise} \end{cases}$$

be the joint pdf of X and Y. Let us compute the marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$ ~~Now~~

$$= \begin{cases} \int_0^{1-x} 24xy \, dy & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$ (we know that $0<x, 0<y$, $0<x+y<1$)

$$= \begin{cases} 12x(1-x)^2 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Since expression of f_{XY} is symmetric in terms of x, y, we will get that

$$f_Y(y) = \begin{cases} 12y(1-y^2) & y \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Also, $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = \int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 12x(1-x)^2\,dx$$

$$= \int_0^1 12x^2(1-x)\,dx = 4x^3 - 3x^4 \Big|_0^1 = 1$$

This verifies that both the marginals and in turn the joint pdf are indeed "valid" pdfs!

Lecture-12

31-08-09

Now ~~suppose~~ recall the example of "swine-flu" analysis done in a previous lecture. Suppose I/we want to evaluate the truth in the statement that "~~the~~ given the symptoms of swine-flu are indeed good indicators of presence (or absence) of disease". To answer such question we would (say) consider the set of all ppl. who have the symptoms and then look at frac. of ppl. in that set who also have high swine-flu. (if this value is high, then symptoms are indeed indicators). In other words, we need to ask questions abt. probabilities ~~considering the~~ as if the original ~~reduced~~ Ω is shrinked to the set of all people having the symptoms. As we explained in one of the early lectures, ~~such~~ conditional probability is a mechanism which facilitates this "shrinkage of Ω".

Defn: → Lets now look at the concept of cond. prob. mass fure. defined in terms of usual conditional prob. over events

Consider two discrete RVs X, Y with joint pmf given by f_{XY}. Suppose ~~that~~ $f_Y(y_i) = P[Y=y_i] \neq 0$ for some fixed value of y_i. Now define a new prob. mass function:

$$\underbrace{f_{X/Y}(x_i/y_i)}_{\text{notation for conditional probability mass function of } X \text{ given } Y=y_i} = \frac{f_{XY}(x_i, y_i)}{f_Y(y_i)} \quad \forall x_i \rightarrow \text{values which random variable } X \text{ takes.}$$

First of all, lets check if (this) is a valid pmf?

First of all its a ratio of values of pmfs and hence is ≥ 0. Secondly,

$$\sum_{\forall x_i} f_{X/Y}(x_i/y_i) = \sum_{\forall x_i} \frac{f_{XY}(x_i, y_i)}{f_Y(y_i)} = \frac{\sum_{\forall x_i} f_{XY}(x_i, y_i)}{f_Y(y_i)} = \frac{\underbrace{f_Y(y_i)}_{\text{marginal pmf definition.}}}{f_Y(y_i)} = 1$$

Now for each value of y_i such that $f_Y(y_i) \neq 0$, we can define a different pmf. Hence we have a family of pmf. all derived from the joint pmf of X, Y.

Note that, $f_{X/Y}(x_i/y_i) = \dfrac{f_{XY}(x_i, y_i)}{f_Y(y_i)} = \underbrace{\dfrac{P[X=x_i, Y=y_i]}{P[Y=y_i]} = P[X=x_i / Y=y_i]}_{\text{defn. of cond. prob. over events which is familiar to all of us.}}$

In other words, the conditional pmf is defined in terms of cond. prob. over events.

Now once we have a cond. pmf we can also define conditional distribution function: $F_{X/Y}$

$$F_{X/Y}(x_0/y_i) = \sum_{\forall x_i \leq x} f_{X/Y}(x_i/y_i) = \sum_{\forall x_i \leq x} P[X=x_i / Y=y_i]$$
$$= P[X \leq x_0 / Y=y_i]$$

Lets try to put down the cond. prob. for a toy example:

(Note that, instead we could have started by defining cond. distr. ~~fn~~ and later discovered the same defn. of cond. pmf!)

Consider the ~~two~~ usual prob. space associated with tossing of a coin (with prob. of getting a head $= p$) for 'n' times (identical & independent Bernoulli trials), say ($n>2$).

Now define two r.v.s

X: trial at which first head appears (X takes values 1 to n)

Y: no. of heads in the 'n' tosses (Y takes values 0 to n)

Note that X is not a valid r.v. as per the defn. since there is "no trial id" which handles the case of all tails in n tosses. As a correcting factor lets ~~also~~ include a dummy value (say "0") which X takes on to represent the case of all tails.

Here is the marginal pmf of X:

$$f_X(x_i) = \begin{cases} (1-p)^n & x_i = 0 \text{ (dummy value representing all tails)} \\ (1-p)^{x_i - 1} p & x_i = 1 \text{ to } n \\ 0 & \text{otherwise} \end{cases}$$

→ Now this is a valid pmf

marginal pmf of Y: → Binomial r.v.

$$f_Y(y_i) = \begin{cases} {}^nC_{y_i}\, p^{y_i} (1-p)^{n-y_i} & y_i = 0 \text{ to } n. \end{cases}$$

~~joint pmf of~~ Lets write down the conditional prob. mass function of $X / Y = i$

(a) $\underline{i=0}$: $$f_{X/Y}(x_i/0) = \begin{cases} 1 & \text{if } x_i = 0 \\ 0 & \text{otherwise} \end{cases}$$ → valid pmf

(b) $\underline{i=1}$: $$f_{X/Y}(x_i/1) = \begin{cases} \dfrac{f_{XY}(x_i,1)}{f_Y(1)} = \dfrac{(1-p)^{n-1}p}{{}^nC_1\, p(1-p)^{n-1}} = \dfrac{1}{n} & \text{if } x_i = 1 \text{ to } n \\ 0 & \text{otherwise} \end{cases}$$

(c) $\underline{\underline{i=2}}$: $f_{X/Y}(x_i/2) = \begin{cases} \dfrac{{}^{n-x_i}C_1\, p^2 (1-p)^{n-2}}{{}^{n}C_2\, p^2 (1-p)^{n-2}} & \text{if } x_i = 1 \text{ to } n-1 \\ 0 & \text{otherwise} \end{cases}$

→ This is also a valid pmf.

so on. . . . .

~~Now we have already put down values of joint pmf for few values~~

In ~~the~~ the process of writing down the cond. ~~prob~~ pmf we have also put down the joint pmf for ~~this~~ few value pair of (x_i, y_i).

In the next lecture we will look at the case of cond. prob. for jointly conts. r.v. etc.

Lecture Notes-13

Suppose X, Y are jointly conts. r.v. Now we want to define notion of r.v. cond. distr. and cond. pdf (if possible), moreover using the familiar notion of cond. prob. over events (which is very familiar to us). Now pdf has "no direct link" with prob. So may be its better to start ~~from~~ by defining cond. distr. function using cond. prob. over events.

Suppose we attempt the following: $\underbrace{F_{X|Y}(x/y) = P\{X \le x / Y = y\}}_{\text{note this was the defn for discrete r.v. case.}}$

We are bound to fail since $P\{Y=y\} = 0 \ \forall y$. The work around is to define as follows: (which intuitively means the same!)

$$F_{X|Y}(x/y) = \lim_{\Delta y \downarrow 0} P[X \le x / y \le Y \le y+\Delta y] \longrightarrow \left(\text{this is the definition we go with}\right)$$

$$= \lim_{\Delta y \downarrow 0} \frac{P[X \le x, y \le Y \le y+\Delta y]}{P[y \le Y \le y+\Delta y]} \longrightarrow \left(\text{familiar notion of cond. prob. over events}\right)$$

$$= \lim_{\Delta y \downarrow 0} \frac{\int_{-\infty}^{x}\int_{y}^{y+\Delta y} f_{XY}(x', y')\, dy'\, dx'}{\int_{y}^{y+\Delta y} f_Y(y')\, dy'} \longrightarrow \left(\text{Since X, Y are jointly conts.}\right) \text{ there } \exists \text{ a pdf}$$

$$= \lim_{\Delta y \downarrow 0} \frac{\int_{-\infty}^{x} f_{XY}(x', y)\, \Delta y\, dx'}{f_Y(y)\, \Delta y} \longrightarrow \left(\text{over small intervals we assume pdf doesn't change}\right) \text{ so area is height} \times \text{interval length}$$

$$\Rightarrow F_{X/Y}(x/y) = \frac{\int_{-\infty}^{x} f_{XY}(x', y)\, dx'}{f_Y(y)} = \int_{-\infty}^{x} \underbrace{\left[\frac{f_{XY}(x', y)}{f_Y(y)}\right]}_{\text{this must be cond. pdf!!}} dx'$$

$$\Rightarrow \boxed{f_{X/Y}(x/y) = \frac{f_{XY}(x, y)}{f_Y(y)}} \quad \text{(Conditional pdf)}$$

→ Note the similarity in the expression even in the discrete r.v.s case!

Now ~~for~~ this defn. is first of all valid iff y is such that $f_Y(y) \neq 0$. (i.e. prob. density of Y at y is non-zero).

Now let us check ~~suppose Exist~~ if for a fixed y, the $f_{X/Y}$ is indeed a pdf or not. (This check will complete the defn.)

(i) firstly $f_{X/Y}$ is ≥ 0.

(ii) $$\int_{-\infty}^{\infty} f_{X/Y}(x/y)\, dx = \int_{-\infty}^{\infty} \frac{f_{XY}(x, y)}{f_Y(y)}\, dx = \frac{\int_{-\infty}^{\infty} f_{XY}(x, y)\, dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1$$

Hence for a fixed value of y such that $f_Y(y) \neq 0$, $f_{X/Y}$ is indeed a pdf and with different values of y (satisfying $f_Y(y) \neq 0$) we get different pdf's!

Let look at an eg. given in one of prev. lectures:

eg1 $f_{XY}(x,y) = \begin{cases} 24xy & x>0, y>0, x+y<1 \\ 0 & \text{otherwise} \end{cases}$

We already saw that

$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \begin{cases} 12x(1-x)^2 & 0<x<1 \\ 0 & \text{otherwise} \end{cases}$

IIIly $f_Y(y) = \begin{cases} 12y(1-y)^2 & 0<y<1 \\ 0 & \text{otherwise.} \end{cases}$

Now lets compute

$f_{X/Y}(x/y) = \begin{cases} \dfrac{24xy}{12y(1-y)^2} = \dfrac{2x}{(1-y)^2} & \text{if } 0<x<1-y \\ 0 & \text{otherwise} \end{cases}$

for some $y \in (0,1)$ where I am sure $f_Y(y) \neq 0$

(It is an easy exercise to check if ↑ is valid pdf)

For eg. $f_{X/Y}(x/0.25) = \begin{cases} \dfrac{32x}{9} & 0<x<0.75 \\ 0 & \text{otherwise} \end{cases}$

$f_{X/Y}(x/0.75) = \begin{cases} 32x & 0<x<0.25 \\ 0 & \text{otherwise} \end{cases}$

So we can get a family of pdf using different values of $y \in (0,1)$.

Now lets take another eg of a jointly conts pdf we saw in last class eg2 $f_{\underline{X}}(\underline{x}) = f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}$ $\underline{x} \in \mathbb{R}^n$.

We also say that (i) each of $X_1, X_2, \ldots X_n$ have std. Normal dist.

(ii) $F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$

(why?) $\Updownarrow$

$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$

Now lets take $n=2$, then calculate

$$f_{X_1 | X_2}(x_1 / x_2) = \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{f_{X_1}(x_1) \cancel{f_{X_2}(x_2)}}{\cancel{f_{X_2}(x_2)}} = f_X(x_1).$$

In other words knowledge abt. X_2 is not effect pdf of X_1 !

Similar notion was discussed while discussing notion of independent events!

Lets formalize this notion of Independence;

<u>Independence of rvs</u>

X, Y are said to be independent

$\iff \forall B_1, B_2 \in \mathcal{B}$ it happens that: (I).

$$\underbrace{P[X \in B_1, Y \in B_2]}_{\text{in other words the events}} = P[X \in B_1] P[Y \in B_2]$$

$\{X \in B_1\}$ & $\{Y \in B_2\}$

$\equiv$ $\equiv$

$\{\omega \in \Omega / X(\omega) \in B_1\}$ $\{\omega \in \Omega / Y(\omega) \in B_2\}$

are independent

Now lets choose $B_1 = (-\infty, x]$ & $B_2 = (-\infty, y]$

then (I) $\Longrightarrow$ $\underline{F_{XY}(x,y) = F_X(x)\,F_Y(y)}$

(the converse is also true & beyond scope of this course)

This means the joint dist factorizes into marginals of in case of independent r.v.s the marginals completely determine the joint-dist.!!

$\longrightarrow$ Now if X, Y are discrete r.v.s, then:

Take $B_1 = \{x\}$, $B_2 = \{y\}$: (I) $\Longrightarrow$ $\underline{f_{XY}(x,y) = f_X(x)\,f_Y(y)}$

$\longrightarrow$ If X, Y are jointly conts. r.v.s, then we anyway have:

$$F_{XY}(x,y) = F_X(x)\,F_Y(y)$$

$$\Rightarrow \frac{\partial F_{XY}(x,y)}{\partial y} = F_X(x)\frac{dF_Y(y)}{dy} = F_X(x) f_Y(y)$$

$$\Rightarrow f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = \frac{d}{dx}F_X(x) f_Y(y) = f_X(x) f_Y(y).$$

$$\Longrightarrow f_{XY}(x,y) = f_X(x)\, f_Y(y).$$

So dist. functions, pmf, pdf's factorize!!

~~We can extend~~

Also for independent r.v's : $\underbrace{f_{X|Y}(x/y) = \frac{f_{XY}(x,y)}{f_Y(y)}}_{\text{Conditional = marginal!}} = \frac{f_X(x)\,f_Y(y)}{f_Y(y)} = f_X(x)$

The notion of independence of r.v's can be extended to any $X_1, X_2, \ldots X_n$:

We say $X_1, \ldots, X_n$ are independent if ~~for~~ all sub-collection $X_i \ldots X_j$ are independent (in other words $X_i X_j$ are independent, $X_i X_j X_k$ are independent, so on ...) and $F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n)$ factorizes into marginals.

(So there are $2^n - n - 1$ conditions to be checked!)

One of the conditions is $F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n)$

$$= F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$

Note that the r.v's in eg 1 are not independent and those in eg 2 are independent r.v's!

Moreover the r.v's in eg 2 are also <u>identically distributed</u>

In other words, each of the $X_1, X_2, \ldots, X_n$ has the same std. Normal distribution.

Such a collection of r.v's which are <u>independent</u> and are <u>identically distributed</u> are called as <u>iid</u> r.v's.

Let's look at a quick eg:

eg Suppose X, Y are i.i.d. & conts. r.v's.

Calculate $P[X>Y]$.

Intuitive answer is:

$$\cancel{P[\mathbb{R}^2]}\ 1 = P\{(X,Y)\in\mathbb{R}^2\} = P[X>Y \cup X<Y \cup X=Y]$$
$$= P\{X>Y\} + P[X<Y] + P\{X=Y\}$$

each are mutually exclusive events

$P\{X=Y\} \to \cancel{P(\omega)}\ P\{X=x, Y=x\} \to$ zero since X, Y are conts. r.v's!

Now there is no reason to believe $X>Y$ or $Y<X$ (since they are independent & imitations of same distribution) so $P[X>Y]$ must be equal to $P[Y<X]$

$$\Rightarrow P[X>Y] = 1/2$$

|||ly argument shows $P\{X_1>X_2>\cdots>X_n\} = \frac{1}{n!}$ (now all are iid.)

$\underbrace{\qquad}_{\text{because this is just one ordering among all } n! \text{ orderings!!}}$

$\longrightarrow$

Now lets look at a more rigorous answer and we will realize that our intuition is right!

$$P[X>Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{XY}(x,y)\,dy\,dx = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{x} f_X(x)\,f_Y(y)\,dy\,dx}_{X,Y \text{ are independent}}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_X(x) f_X(y)\, dy\, dx \quad \left(\because X, Y \text{ are identically distributed} \right)$$

$$= \int_{-\infty}^{\infty} f_X(x) \underbrace{\left[\int_{-\infty}^{x} f_X(y)\, dy \right]}_{\longrightarrow F_X(x)\,!} dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\, F_X(x)\, dx$$

(two ways of computing it)

(by parts method)

$$= F_X(x) F_X(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_X(x) f_X(x)\, dx$$

$$\Rightarrow \int_{-\infty}^{\infty} f_X(x) F_X(x)\, dx = \frac{F_X(x) F_X(x) \Big|_{-\infty}^{\infty}}{2}$$

$$= 1/2$$

(substitution method)

Put $x \longrightarrow F_X(x)$

$$= \int_0^1 F_X(x)\, d\,F_X(x) = \frac{F_X(x)^2}{2}\Big|_0^1 = 1/2$$

$\longrightarrow$

BAYES Theorem (extension to r.v.).

Let X, Y are discrete / jointly conts. r.v.'s

Let f_{XY}, f_X, f_Y represent their joint & marginals pmf/pdf. (whichever is the case)

We know: $f_{X|Y}(x/y) = \dfrac{f_{XY}(x,y)}{f_Y(y)} \Rightarrow f_{XY}(x,y) = f_{X|Y}(x/y)\, f_Y(y)$

also, $f_{Y|X}(y/x) = \dfrac{f_{XY}(x,y)}{f_X(x)} = \dfrac{f_{X|Y}(x/y)\, f_Y(y)}{f_X(x)}$.

Now if X,Y are discrete r.v.s, then:

$$f_X(x) = \sum_{\forall y'} f_{XY}(x,y') = \sum_{\forall y'} f_{X/Y}(x/y')\, f_Y(y')$$

Now substituting back we get:

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)\, f_Y(y)}{\sum_{\forall y'} f_{X/Y}(x/y')\, f_Y(y')} \quad \Bigg\} \text{ Baye's theorem for discrete r.v.s.}$$

If X, Y are jointly conts., then:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y')\, dy' = \int_{-\infty}^{\infty} f_{X/Y}(x/y')\, f_Y(y')\, dy'$$

Substituting back we get:

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)\, f_Y(y)}{\int_{-\infty}^{\infty} f_{X/Y}(x/y')\, f_Y(y')\, dy'} \quad \Bigg\} \text{ Bayes theorem for jointly conts r.v.s.}$$

The advantage ~~In the next class~~ is that without knowing joint pmf or joint pdf we are able to calculate cond. prob on one side using cond. prob on the other side i.e. $f_{X/Y}$ $\downarrow$ $f_{Y/X}$

and of course we also are using $f_Y(y)$ or $f_X(x)$ i.e. marginals.

∴ Use wherever joint-pmf/pdf is difficult to compute !! (unnecessary)

This lecture begins with some applications of the Baye's theorem

eg1 Suppose we are told that a person picks up a coin at random from a set of 'm' coins. It is also given to us ~~that~~ that the prob. of picking the i^{th} coin is 'q_i' (i.e. $q_i \geq 0$, $\sum_{i=1}^{m} q_i = 1$). Now just given this information suppose we are asked to guess what coin was picked up by the person, intuitively our "answer" would be $\arg\max_i q_i$. i.e. the coin which has the maximum prob. of ~~be~~ being picked. Note that we are using ~~no~~ absolutely no information regarding the particular coin picked ~~up~~ up but some "generic" information ~~abt~~ about q_i.

Now suppose the person is generous to reveal some partial information regarding the coin in his hand and then asks us to guess. ~~otherwise, if the~~ In particular, suppose he reveals the number of heads he got by tossing the coin in his hand for 'n' times and he also reveals the prob. of getting heads with each of the coin (i.e. $p_i \geq 0$, $p_i \leq 1$, $i = 1$ to m).

Now, ~~the~~ a little bit of thinking will show that ~~a~~ given the partial information, we can come up with a "better" guess. (Think abt two extreme cases where all q_i are same & q_i highly distinct)

Let us formalize our ideas: define two rvs

X : # heads in 'n' tosses (with the coin picked up) $\rightarrow \in \{0,1,2,\ldots,n\}$

Y : 'id' of the coin picked up $\rightarrow \in \{1,2,\ldots,m\}$

Now ~~This~~ pmf of Y is given: $f_Y(y) = \begin{cases} q_y & y \in \{1,2,\ldots,m\} \\ 0 & \text{otherwise} \end{cases}$

~~Since this is an information given prior to the partial information regarding the coin picked up, this information is~~

Also, the following cond. pmf is given:

$$f_{X/Y}(x/y) = \begin{cases} {}^nC_x \, p_y^x (1-p_y)^{n-x} & x \in \{0,1,\ldots,n\} \\ 0 & \text{otherwise} \end{cases}$$

$y \in \{1,2,\ldots,m\}$

Now we wish to ~~cal~~ calculate the prob. that the coin picked up is 'y' given that there were 'x' heads:

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)\, f_Y(y)}{\sum_{\forall y'} f_{X/Y}(x/y')\, f_Y(y')} \quad (\because \text{Bayes theorem})$$

$x \in \{0,1,\ldots,m\}$

$$= \frac{{}^nC_x \, p_y^x (1-p_y)^{n-x} \, q_y}{\sum_{\forall y'} {}^nC_x \, p_{y'}^x (1-p_{y'})^{n-x} \, q_{y'}}.$$

Now given a 'n' by the person, for different values of y we can calculate this quantity. Again (intuitively) the guess is to pick the coin which maximizes $f_{Y/X}$!

Note that if all $q_y = \frac{1}{m}$ (for $y = 1$ to m), then $f_{Y/X}$ is same as f_Y (and our guess wouldn't change!) In other words, if all coins have same prob. of getting heads, the partial information regarding the coin picked up does not give any ~~made~~ additional insight to it! → i.e. no. heads in 'n' tosses

Since $f_{Y/X}$ is prob. of ~~picking~~ a coin being picked after looking ~~after~~ at the partial information abt X, it is usually called as posterior prob. & f_Y is called as prior probability.
(the same idea is extensively used in prob. model known as ~~Now to be seen through the~~ Hidden Markov model (HMM) which are ~~t~~ state of the art models for automated speech recognition systems!)

Hence the idea of cond. prob. & Baye's theorem have far reaching consequences.

Now lets look at another eg. illustrating the utility of the Baye's theorem. The reader is encouraged to see ~~t~~ at every step the analogy between these two examples.

eg 2 Suppose X represents the "raw" diagnostic report of a patient (for the sake of simplicity assume it represents the body temp. of the patient) and suppose Y represents whether he has a disease or not.

$Y \to 0$ (normal), 1 (disease)

Now (again) ~~suppose~~ the task is to predict (given) whether a patient has disease or not! Assume the following information is given:

(i) f_Y (pmf) is given. (This is the prior information). In words, the fraction of normal & diseased people in a population is given. (I)

(ii) $f_{X/Y}$ (pdf) is given. In words, the body temp. distribution of normal as well as patients with disease are given. Note that X/Y is a conts. r.v. $\Rightarrow F_{X/Y}(x/y) \equiv P\{X \le x / Y = y\} = \int_{-\infty}^{x} f_{X/Y}(x'/y)\,dx'$.

Again we ~~need~~ wish to compute $f_{Y/X}$ (pmf) which in words is the prob. of the patient is normal or has disease given his diagnostic report (body temp.)

Before this ~~suppose we~~ let us answer a simpler question "what is $f_X(x)$?". In words, what is the body temp. dist. of the entire population? Since we have not assumed anything abt X (conts. or discrete), let us compute its dist. function:

$$F_X(x) = P[X \le x] = \sum_{\forall y} P[X \le x, Y = y] \qquad (\because \text{marginal fundae})$$

$$= \sum_{\forall y} P[X \le x / Y = y] P[Y = y] \qquad (\because \text{cond. prob.})$$

$$= \sum_{\forall y} F_{X/Y}(x/y) f_Y(y) \qquad (\because \text{defn. of cond. distt.})$$

$$= \sum_{\forall y} \int_{-\infty}^{x} f_{X/Y}(x'/y)\, dx'\, f_Y(y) \qquad (\because X/Y \text{ is a conts rv})$$

$$= \int_{-\infty}^{x} \boxed{\sum_{\forall y} f_{X/Y}(x'/y) f_Y(y)}\, dx' \qquad \left(\because \text{interchange } \int \,\&\, \Sigma\right)$$

$$\Rightarrow f_X(x) = \sum_{\forall y} f_{X/Y}(x/y) f_Y(y) \qquad (\because \text{defn. of pdf})$$

Recall that this resembles the total prob. rule (for the case X, Y are both discrete). However note that f_X and $f_{X/Y}$ are pdfs and f_Y is a pmf. Also, it looks like f_X is a convex combination of conditional pdfs ($f_{X/Y}$). In other words it looks like X is a 'mixture' of two kinds of rvs ($X/Y=0$, $X/Y=1$ here!). f_X is also hence sometimes called as mixture density. f_Y are called as mixing prob. & $f_{X/Y}$ as class conditional density!

Models satisfying (i) (ii) in (I) hence are called as "Mixture Models".

Now lets try to compute:

$$f_{Y/X}(y/x) = \lim_{\Delta x \downarrow 0} P[Y=y / X \in [x, x+\Delta x]]$$

(!!! to define in case of X,Y conts r.v.)

↓ pmf for fixed x such that $f_X(x) \neq 0$

$$= \lim_{\Delta x \downarrow 0} \frac{P[x \le X \le x+\Delta x, Y=y]}{P[x \le X \le x+\Delta x]}$$

$$= \lim_{\Delta x \downarrow 0} \frac{P[x \le X \le x+\Delta x / Y=y]\, P[Y=y]}{P[x \le X \le x+\Delta x]}$$

$$= \lim_{\Delta x \downarrow 0} \frac{\int_x^{x+\Delta x} f_{X/Y}(x'/y)\, dx'\, f_Y(y)}{\int_x^{x+\Delta x} f_X(x')\, dx'}$$

($\because$ X/Y and therefore X are conts r.v.s)

$$= \lim_{\Delta x \downarrow 0} \frac{f_{X/Y}(x/y)\, \Delta x\, f_Y(y)}{f_X(x)\, \Delta x}$$

$$\Rightarrow f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)\, f_Y(y)}{f_X(x)} = \frac{f_{X/Y}(x/y)\, f_Y(y)}{\sum_{\forall y'} f_{X/Y}(x/y')\, f_Y(y')}$$

Again, this looks like Baye's theorem in case of X, Y both discrete r.v.s! But here $f_{Y/X}$ and f_Y are ~~pdfs~~ pmfs whereas $f_{X/Y}$ is a pdf!

Again $f_{Y/X}$ is posterior pmf (pmf after looking at partial information i.e. diagnostic report of the patient)

f_Y is simply the prior information.

In general, there are two ways to guess:

$\arg\max_y f_Y(y)$ → y that maximizes the prior prob. (without looking at the particular patient!)

$\arg\max_y f_{Y|X}(y/x)$ → given diagnostic report x of a patient guess his status of health

1) max. prior prob.

2) max. posterior prob. → it is easy to reason out this gives a better guess.

Suppose we want to design a chair which withstands the weight of people who sit on it as well as is not built from too costly or heavy (strong) material! One way is to design it for the heaviest person on earth. But this is too pessimistic and will lead to a chair perhaps too heavy to even move :)
On the other hand we want chair to be 'robust' enough to handle heavy people.

One way to put this is to design chair such that it withstands the weight of any 100 random 'a' 'n' people who sit on it. i.e. design for:

$$M = \max\{X_1, X_2, X_3, \ldots, X_n\}$$

Here $X_1, \ldots, X_n$ represents weights of 'n' people. It is easy to see that they are i.i.d. r.v.s (why??).
Also note that M is a r.v and is infact a function of collection of r.v.s $X_1, \ldots, X_n$!

Let compute the df of M:

$$F_M(x) = P[M \le x] = P[\max\{X_1, X_2, \ldots, X_n\} \le x]$$

$$= P[X_1 \le x, X_2 \le x, \ldots, X_n \le x]$$

$$= P\{X_1 \le x\} P\{X_2 \le x\} \cdots P\{X_n \le x\} \quad (\because \text{they are independent})$$

$$= F_{X_1}(x) F_{X_2}(x) \cdots F_{X_n}(x)$$

$$= (F(x))^n$$

(here F is the common dist. function of all the identically distributed r.v.s $X_1, \ldots X_n$)

$$\Rightarrow f_M(x) = n(F(x))^{n-1} f(x)$$

(here 'f' is the common pdf of $X_1, \ldots X_n$)

In words $F(\cdot)$ represent ~~represent~~ the dist. of body weight among humans!

Illy one can consider:

7

$N = \min\{X_1, X_2, \dots, X_n\}$. N is another function of collections of r.v.s.

$$F_N(x) = P[N \le x] = P[\min\{X_1, X_2, \dots, X_n\} \le x]$$
$$= 1 - P[\min\{X_1, X_2, \dots, X_n\} > x]$$
$$= 1 - P[X_1 > x, X_2 > x, \dots, X_n > x]$$
$$= 1 - \left(1 - F(x)\right)^n \quad \text{(again by } \underline{iid} \text{)}$$

$$\Rightarrow f_N(x) = n\left(1 - F(x)\right)^{n-1} f(x).$$

Now one can consider the joint dist. of collection of M, N r.v.s which are in turn function of collections of r.v.s!

$$F_{MN}(x,y) = P[M \le x, N \le y]$$
$$= P[M \le x] - P[M \le x, N > y]$$
$$= \left(F(x)\right)^n - P[y < X_1 \le x, y < X_2 \le x, \dots, y < X_n \le x]$$
$$= \left(F(x)\right)^n - \left(F(x) - F(y)\right)^n \quad (\because \text{iid arguments})$$

In future classes we will look into more eg. of functions of r.v.s.

Lecture Notes-15

7-Sep-09

This lecture formalizes the notion of functions of multivariate r.v.s (in other words functions of collections of r.v.s).

Suppose $X_1, X_2, \ldots, X_n$ are r.v.s defined on $\mathbb{P} = (\Omega, \mathcal{F}, P)$. Also, $g: \mathbb{R}^n \to \mathbb{R}$ is given. Consider a new function $Z: \Omega \to \mathbb{R}$ defined as

$$Z(\omega) = g(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \quad \forall \omega \in \Omega.$$

The short hand representation of $\downarrow$ is $Z = g(X_1, X_2, \ldots, X_n)$.

Now Z is indeed a function from $\Omega \to \mathbb{R}$, so it is a r.v. if:

$$Z^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}.$$

i.e.

$$\Rightarrow \{\omega \in \Omega \,/\, g(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{B}$$

Consider the condition $\underline{g^{-1}(B) \in \mathcal{B}^n}$ $\forall B \in \mathcal{B}$ in other words $g^{-1}(B) = B_1 \times B_2 \times \cdots \times B_n$ each $B_i \in \mathcal{B}$.

It is easy to see that $Z^{-1}(B) = \bigcap_{i=1}^{n} \underbrace{\{\omega \in \Omega \,/\, X_i(\omega) \in B_i\}}_{\text{since each } X_i \text{ is a valid r.v.!}} \in \mathcal{F}$

Hence $g^{-1}(B) \in \mathcal{B}^n$ is the condition on g^{-1} which makes Z a valid r.v.

Again (not in this class) we can show that if <u>g is a continuous function</u> then $Z = g(X_1, \ldots, X_n)$ is ~~also~~ a valid r.v.

Let's consider an example:

eg1 $\quad Z = X + Y$ $\quad$ (here $g(x,y) = x + y$)

Suppose joint pdf of X, Y is known. Compute f_Z.

$$F_Z(z) = P[Z \le z] = P[X + Y \le z]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x,y)\, dy\, dx$$

Now $$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} \frac{d}{dz} \int_{-\infty}^{z-x} f_{XY}(x,y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} f_{XY}(x, z-x)\, dx$$

$\longrightarrow$

Assume now that X, Y are independent r.v.s.

$$\Rightarrow f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx \quad \Big\}\ \text{nothing but convolution of } f_X, f_Y \text{ at } z$$

$$\equiv f_X(z) * f_Y(z)$$

Hence, pdf of sum of two r.v.s is the convolution of the individual pdfs!

Lets now take the special case $f_X(x) = f_Y(x) = \begin{cases} 1/2 & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

In other words we are assuming X, Y are both $\sim U[-1,1]$ (uniform r.v.s in $[-1,1]$).

$$f_Z(z) = f_X(z) * f_Y(z)$$

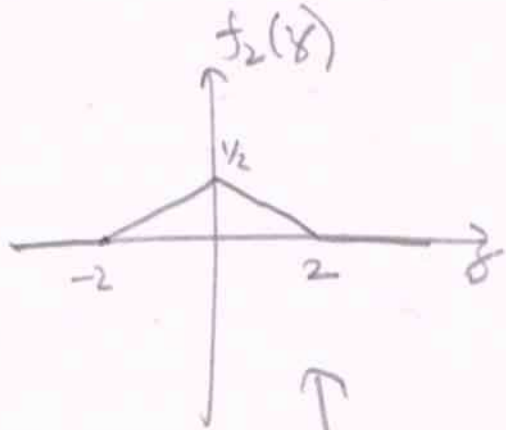
(X, Y are iid and are uniform r.v. in (-1,1])

$$= \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx$$

$f_X(x)$ and $f_Y(z-x)$ are non zero iff
$-1<x<1$
$-1<z-x<1$
i.e. $-1<x<1$
$z-1<x<z+1$

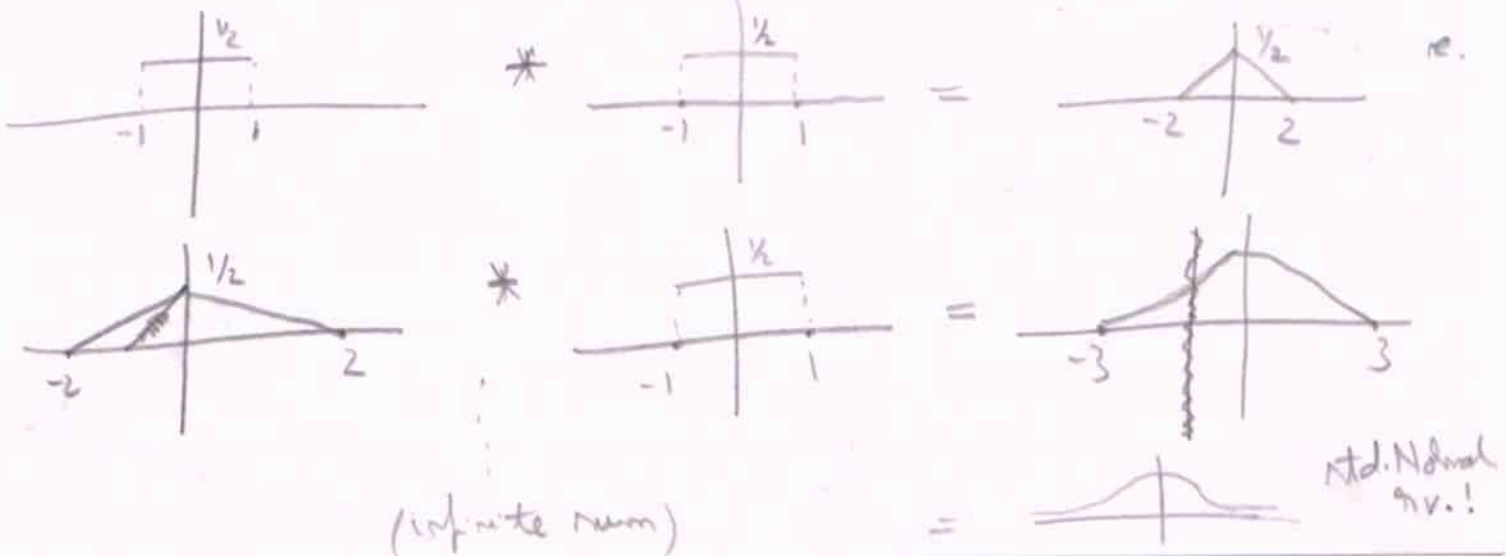
$$= \begin{cases} 0 & z \le -2 \\ \displaystyle\int_{-1}^{z+1} \tfrac{1}{4}\, dx & -2 < z \le 0 \\ \displaystyle\int_{z-1}^{1} \tfrac{1}{4}\, dx & 0 < z \le 2 \\ 0 & z > 2 \end{cases}$$

$$= \begin{cases} 0 & z \le -2 \\ \dfrac{z+2}{4} & -2 < z \le 0 \\ \dfrac{2-z}{4} & 0 < z \le 2 \\ 0 & z > 2 \end{cases}$$



So sum of two iid U[-1,1] r.v.s is not U[-1,1] but in fact.

IIIly we can sum three, four, ... iid r.v's which are U[-1,1]:



This was an intuition for a special case of Central Limit Theorem.
(Here sum of infinite iid rvs all $U[-1,1]$ is converging to std. Normal r.v.)

→ **eg2** Let X, Y be iid and ~~[illegible]~~ be Normal rvs.

by above argument ~~[illegible]~~ $Z = X+Y$ has the pdf as convolution of pdfs of X, Y. It is a well-known result that convolution of any two <u>Gaussian functions</u> is a Gaussian function.
Using this result we can say Z is <u>again a Normal rv</u>!
(we will see a generic result of this kind later)

(A Gaussian function is any function of the form:
$$f(x) = a e^{-(x-b)^2/c^2}$$
$a, c > 0$. Note that ~~the~~ Normal rv has a pdf as a Gaussian func. Hence Normal rv are also known as gaussian rvs!)

eg3 $Z = \dfrac{X}{Y}$.

$$F_Z(z) = P[Z \le z] = P\left[\frac{X}{Y} \le z\right]$$
$$= P[X \le zY, Y>0] + P[X \ge zY, Y<0]$$
$$= \int_0^\infty \int_{-\infty}^{zy} f_{XY}(x,y)\,dx\,dy + \int_{-\infty}^{0} \int_{zy}^{\infty} f_{XY}(x,y)\,dx\,dy$$

$$\Rightarrow f_Z(z) = \frac{dF_Z(z)}{dz} = \int_0^\infty \frac{d}{dz}\int_{-\infty}^{zy} f_{XY}(x,y)\,dx\,dy + \int_{-\infty}^{0}\frac{d}{dz}\int_{zy}^{0} f_{XY}(x,y)\,dx\,dy$$

$$= \int_0^\infty y\, f_{XY}(zy,y)\,dy + \int_{-\infty}^{0} -y\, f_{XY}(zy,y)\,dy$$

$$= \int_{-\infty}^{\infty} |y|\, f_{XY}(zy,y)\,dy$$

$\longrightarrow$ Now lets take X, Y as i.i.d and std. Normal r.v.s.

then $f_Z(z) = \int_{-\infty}^{\infty} |y| \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2y^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = \frac{1}{\pi}\int_0^\infty y\, e^{-(z^2+1)\frac{y^2}{2}} dy$

$$= \frac{1}{\pi}\left[\frac{e^{-(z^2+1)t}}{-(z^2+1)}\right]_0^\infty$$

$$= \frac{1}{\pi(1+z^2)}$$

$\therefore$ Z is a Cauchy r.v !

$\longrightarrow$

Now, Consider the collection of r.v's $Z_1, Z_2, \ldots, Z_n$ each of which are ~~two four continuous probability~~ in turn functions of the r.v.s: $X_1, X_2, \ldots, X_n$.

i.e. Consider
$$Z_1 = g_1(X_1, X_2, \ldots, X_n)$$
$$Z_2 = g_2(X_1, X_2, \ldots, X_n)$$
$$\vdots$$
$$Z_n = g_n(X_1, X_2, \ldots, X_n)$$

we already saw that each Z_i is a r.v (from the same initial $\mathbb{P}$). Hence the collection of $\{Z_1, Z_2 \ldots Z_n\}$ is indeed a 'valid' multivariate r.v. Hence we can talk about its dist. fuc. & equivalently, the joint distribution of $Z_1, Z_2, \ldots, Z_n$ which are collections of fuctions of $X_1, X_2, \ldots, X_n$ r.v.s.

$$F_Z(\underline{z}) = F_{Z_1 Z_2 \ldots Z_n}(z_1, z_2, z_3, \ldots, z_n)$$

$$= P[Z \leq \underline{z}]$$

$$= P[Z \in B] \qquad \text{where } B = (-\infty, z_1] \times (-\infty, z_2] \times \ldots \times (-\infty, z_n]$$

Now $Z = \underline{g}(X)$ where $\underline{g} : \mathbb{R}^n \to \mathbb{R}^n$

$$\underline{g}(x_1, x_2, \ldots, x_n) = \left(g_1(x_1, x_2, \ldots, x_n), g_2(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n)\right)$$

$$\Rightarrow F_Z(\underline{z}) = P[\underline{g}(X) \in B]$$

$$= P[X \in \underline{g}^{-1}(B)]$$

$$= \int\int_{\underline{g}^{-1}(B)} \cdots \int f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

Now suppose ~~we do the follow~~ that $\underline{g}$ is invertible:

$$\Rightarrow \exists\ \underline{h} \ni \mathbf{X} = \underline{h}(Z).$$

Also suppose $\underline{g}, \underline{h}$ are continuously differentiable.

Then

$$F_Z(z) = \underset{\underline{g}^{-1}(B)}{\iint\cdots\int} f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n$$

$$= \underset{B}{\iint\cdots\int} f_{X_1 \cdots X_n}\left(h_1(z_1, z_2, \ldots z_n), h_2(z_1, \ldots z_n), \ldots, h_n(z_1, \ldots z_n)\right)$$

$$|J|\ dz_1 dz_2 \cdots dz_n$$

$\rightarrow$ abs value of the Jacobian.

$$\Rightarrow f_Z(z) = f_{X_1 \cdots X_n}\left(h_1(z_1 \cdots z_n), \ldots, h_n(z_1 \cdots z_n)\right) |J|$$

$$|J| = \text{abs of det. of } \begin{bmatrix} \frac{\partial h_1}{\partial z_1} & \cdots & \frac{\partial h_1}{\partial z_n} \\ \vdots & & \\ \frac{\partial h_n}{\partial z_1} & \cdots & \frac{\partial h_n}{\partial z_n} \end{bmatrix}.$$

We will see an explanation in the next lecture.

We will continue the discussion at end previous lecture (now restricting ourselves to 2-d case):

Consider two rvs: $Z_1 = g_1(X_1, X_2)$ $\quad g_1 : \mathbb{R}^2 \to \mathbb{R}$

$Z_2 = g_2(X_1, X_2)$ $\quad g_2 : \mathbb{R}^2 \to \mathbb{R}$

Let Z be the multivariate rv representing Z_1, Z_2
X " " " X_1, X_2

Also let, $\underline{g} : \mathbb{R}^2 \to \mathbb{R}^2$ be defined as $\underline{g}(x,y) = (g_1(x,y), g_2(x,y))$

It is easy to see that $Z = \underline{g}(X)$.

Now assume:

(i) $\underline{g}$ is invertible ($\underline{g}$ is bijection). Let $\underline{h} = \underline{g}^{-1}$.
it is easy to see $X = \underline{h}(Z)$. Also let $\underline{h}(z_1, z_2) = (h_1(z_1,z_2), h_2(z_1,z_2))$

(ii) Assume $\underline{g}, \underline{h}$ are continuously differentiable.

~~Let~~ (iii) Assume Z, X are conts. (multivariate) rvs.

We wish to write down the joint pdf of Z_1, Z_2 (i.e. pdf of Z) in terms of joint-pdf of X_1, X_2 (ie, pdf of X). To this end:

$$F_Z(\underline{z}) = P[Z \le \underline{z}]$$
$$= P[Z \in B] \quad \text{where } B = (-\infty, z_1] \times (-\infty, z_2].$$

$$\Rightarrow F_Z(\underline{y}) = P[\underline{g}(X) \varepsilon B]$$
$$= P[X \varepsilon \underline{h}(B)]$$
$$= \iint_{\underline{h}(B)} f_{X_1X_2}(x_1, x_2)\, dx_1\, dx_2 \quad \text{I}$$

Now suppose I do change of dummy variables x_1, x_2 in the double integral:

$$x_1 = h_1(y_1, y_2)$$
$$x_2 = h_2(y_1, y_2)$$

(remember that $h_1 = g_1^{-1}$, $h_2 = g_2^{-1}$)

Now $(x_1, x_2) \varepsilon \underline{h}(B)$ from the integral limits

$$\Rightarrow \left(h_1(y_1, y_2), h_2(y_1, y_2)\right) \varepsilon \underline{h}(B)$$

$$\Rightarrow \underline{h}(y_1, y_2) \varepsilon \underline{h}(B) \Rightarrow (y_1, y_2) \varepsilon B \quad \text{II}$$

Now in order to proceed with change of variables I need to figure out how elevatory area in y_1, y_2 coordinate looks like! Before doing that say elementary area is $|J|\, dy_1\, dy_2$ (we will shortly show what is $|J|$)

Then: $$F_Z(\underline{y}) = \iint_{B} \boxed{f_{X_1X_2}(h_1(y_1,y_2), h_2(y_1,y_2))\, |J|}\, dy_1\, dy_2$$

from II ← B

must be $f_{Z_1Z_2}$!! (why?)

Since dist. of Z can be computed by integrating a function (over relevant 2-d interval), that same function must be the pdf of Z!

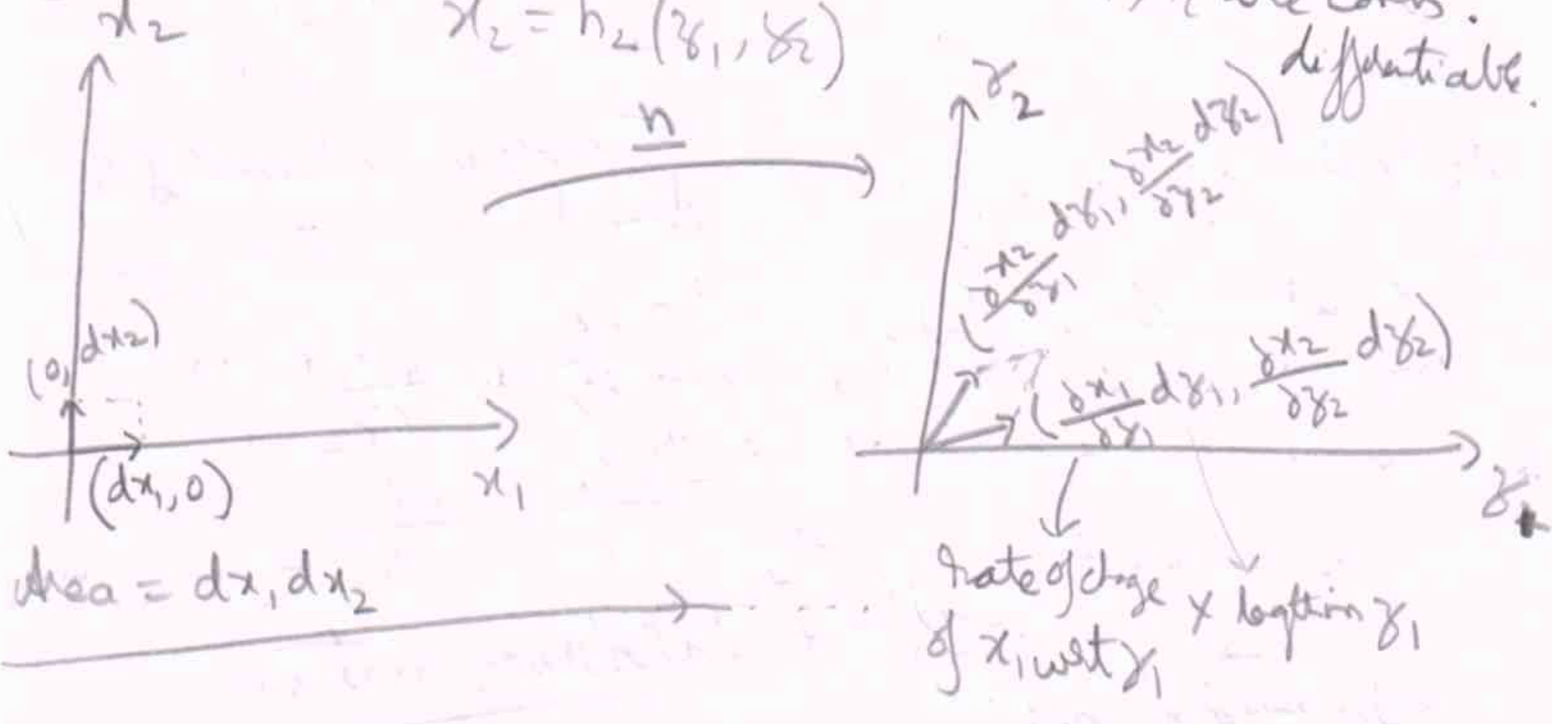
$$\Rightarrow \boxed{f_{Z_1,Z_2}(y_1,y_2) = f_{X_1,X_2}\big(h_1(y_1,y_2),\, h_2(y_1,y_2)\big)\,|J|}$$

Hence we are successful in the derivation. Now let us see how $|J|$ can be computed as:

Change of Variables in Multiple integrals

$$x_1 = h_1(y_1, y_2)$$
$$x_2 = h_2(y_1, y_2)$$

h_1, h_2 are conts. differentiable.



Area in y_1, y_2 coordinates is area of parallelogram (for which we know the vectors of sides!)

Area of ||gm is nothing but cross-product of the vectors of base sides:

$$\text{Area vector} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \frac{\partial x_1}{\partial y_1} dy_1 & \frac{\partial x_1}{\partial y_2} dy_2 & 0 \\ \frac{\partial x_2}{\partial y_1} dy_1 & \frac{\partial x_2}{\partial y_2} dy_2 & 0 \end{vmatrix}$$

$$= \left(\frac{\partial x_1}{\partial y_1} dy_1 \frac{\partial x_2}{\partial y_2} dy_2 - \frac{\partial x_1}{\partial y_2} dy_2 \frac{\partial x_2}{\partial y_1} dy_1 \right) \hat{k}$$

$$\text{Area} = |\quad| = \underbrace{\left| \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_2}{\partial y_1} \frac{\partial x_1}{\partial y_2} \right|}_{\text{is called Jacobian and denoted by } |J|} dy_1 dy_2$$

Note that,

$|J|$ is also abs. of det. of $\rightarrow \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}$

Called as Jacobian matrix

In n-dimensional case:

$$\text{Jacobian matrix} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Lets take an eg and work out details:

<u>eg1</u> Let $Z_1 = X + Y$

$Z_2 = X - Y$ joint pdf of X, Y is given

Compute joint pdf of Z_1, Z_2.

We know, $f_{Z_1 Z_2}(z_1, z_2) = f_{X_1 X_2}(h_1(z_1, z_2), h_2(z_1, z_2)) \, |J|$

1) we ~~get~~ first need to figure out what are h_1, h_2:

i.e. express X, Y in terms of Z_1, Z_2:

$$X = \frac{Z_1 + Z_2}{2} \qquad\qquad h_1(z_1, z_2) = \frac{z_1 + z_2}{2}$$

$$\Rightarrow$$

$$Y = \frac{Z_1 - Z_2}{2} \qquad\qquad h_2(z_1, z_2) = \frac{z_1 - z_2}{2}$$

Now $|J| = \text{abs} \begin{vmatrix} \dfrac{\partial\left(\frac{z_1 + z_2}{2}\right)}{\partial z_1} & \dfrac{\partial\left(\frac{z_1 + z_2}{2}\right)}{\partial z_2} \\ \dfrac{\partial\left(\frac{z_1 - z_2}{2}\right)}{\partial z_1} & \dfrac{\partial\left(\frac{z_1 - z_2}{2}\right)}{\partial z_2} \end{vmatrix}$

$$= \text{abs} \begin{vmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{vmatrix} = 1/2$$

$$\Rightarrow f_{Z_1 Z_2}(\gamma_1, \gamma_2) = \frac{1}{2} f_{XY}\left(\frac{\gamma_1+\gamma_2}{2}, \frac{\gamma_1-\gamma_2}{2}\right)$$

Now lets compute marginal Z_1:

$$f_{Z_1}(\gamma_1) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{\gamma_1+\gamma_2}{2}, \frac{\gamma_1-\gamma_2}{2}\right) d\gamma_2$$

$$= \int_{-\infty}^{\infty} f_{XY}(t, \gamma_1 - t)\, dt \qquad \left(\text{Put } t = \frac{\gamma_1+\gamma_2}{2}\right)$$

This expression is familiar from prev. lecture.

This shows we are consistent.

eg2

$$Z_1 = X/Y$$

$$Z_2 = Y$$

again $X = Z_1 Z_2$ $\quad$ $\left(\begin{array}{l}\text{i.e.} \\ h_1(\gamma_1, \gamma_2) = \gamma_1 \gamma_2 \\ h_2(\gamma_1, \gamma_2) = \gamma_2\end{array}\right)$

$Y = Z_2$

$$|J| = \text{abs.} \begin{vmatrix} \gamma_2 & \gamma_1 \\ 0 & 1 \end{vmatrix} = |\gamma_2|$$

$$\Rightarrow f_{Z_1 Z_2}(\gamma_1, \gamma_2) = |\gamma_2|\, f_{XY}(\gamma_1 \gamma_2, \gamma_2)$$

Now again

$$f_{Z_1}(z_1) = \underbrace{\int_{-\infty}^{\infty} |z_2|\, f_{XY}(z_1 z_2, z_2)}\, dz_2$$

This expression is also familiar from prev. lecture!

EXPECTATIONS

Now lets return to the topic of expectations.

Consider $Z = g(X, Y)$

we know Z is a r.v.

So we know: $E[Z] = \begin{cases} \int_{-\infty}^{\infty} z\, f_Z(z)\, dz & \text{if } Z \text{ is contin. r.v.} \\ \sum_{\forall z_i} z_i\, f_Z(z_i) & \text{if } Z \text{ is discrete r.v.} \end{cases}$

But one can also show:

Theorem: $E[Z] = E[g(X,Y)] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{XY}(x,y)\, dx\, dy \\ \sum_{\forall x_i} \sum_{\forall y_i} g(x_i, y_i)\, f_{XY}(x_i, y_i) \end{cases}$

Recall that we proved a similar theorem for $Z = g(X)$ also.

Again like prev. time we will show this only for the discrete case:

Proof: $E[Z] = \sum_{\forall \gamma_i} \gamma_i f_Z(\gamma_i)$

$$= \sum_{\forall \gamma_i} \gamma_i \sum_{\substack{(x_i, y_i): \\ g(x_i, y_i) = \gamma_i}} f_{XY}(x_i, y_i)$$

$$= \sum_{\forall \gamma_i} \sum_{(x_i, y_i) : g(x_i, y_i) = \gamma_i} g(x_i, y_i) f_{XY}(x_i, y_i)$$

$$= \sum_{\forall (x_i, y_i)} g(x_i, y_i) f_{XY}(x_i, y_i)$$

if $E[Z]$ exists then series sum is abs. convergent so it doesn't matter in which order we take the sum!

In summary, we know how to compute expectation of function of two (or in general n) rvs!

Lecture Notes - 17

10-Sep-2009

In this lecture we will proceed with discussion of expectation in case of collections of r.v.s.

We already showed that:

$$E[g(X,Y)] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{XY}(x,y)\,dx\,dy & \text{(if } X,Y \text{ are jointly conts)} \\ \sum_{\forall x_i}\sum_{\forall y_i} g(x_i, y_i)\, f_{XY}(x_i, y_i) & \text{(if } X,Y \text{ are discrete rvs)} \end{cases}$$

~~Lets~~ (All derivations from now on (unless specified explicitly) take care of X, Y jointly conts. and prove results on expectation using integrals, however the generic results do hold ~~of~~ for discrete r.v.'s case also).

Consider $Z = g(X,Y) = X+Y$

$$E[Z] = E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\, f_{XY}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f_{XY}(x,y)\,dy\,dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{XY}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} x \boxed{\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy}\,dx + \int_{-\infty}^{\infty} y \boxed{\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx}\,dy$$

$$= \int_{-\infty}^{\infty} x\, f_X(x)\,dx + \int_{-\infty}^{\infty} y\, f_Y(y)\,dy$$

$$= E[X] + E[Y]$$

In general, $E[X_1 + \dots + X_n] = \sum_{i=1}^{n} E[X_i]$.

i.e. ~~Sum~~ Expectation of sum of rvs = sum of expectation of rvs

(Note that we did NOT assume these rvs are independent) I

~~Also~~ ~~Also~~

eg1 Let X_i = indicator of success at i^{th} trial among 'n' independent and identical ~~Berno~~ Bernoulli trials.

i.e. each X_i is a (independent) Bernoulli rv with $P\{X_i=1\}=p$

$P\{X_i=0\}=1-p$

Now $E\{X_i\} = 1\,P\{X_i=1\} + 0\,P\{X_i=0\} = p$

Consider the rv $X = X_1+X_2+\dots+X_n$. In words, X is no. successes in 'n' iid Bernoulli trials. Of course X follows a binomial distribution with parameters (n, p).

Let compute $E\{X\}$ using I:

$$E\{X\} = \sum_{i=1}^{n} E\{X_i\} = \sum_{i=1}^{n} p = np$$ $\rightarrow$ we know this is $E\{X\}$ of binomial rv!

eg2 Let X_i = indicator of changeover at i^{th} interline (gap) between two consecutive tosses in 'n' independent coin tosses of the same coin.

We have already seen that $P\{X_i=1\} = 2p(1-p)$

$$\Rightarrow E\{X_i\} = P\{X_i=1\} = 2p(1-p)$$

Now consider $X = X_1+X_2+\dots+X_{n-1}$. In words, X is the number of changeovers in 'n' tosses!

$$\Rightarrow E\{X\} = \sum_{i=1}^{n} E\{X_i\} = 2(n-1)p(1-p).$$

(Note that here X_i are not independent Bernoulli rvs, so X is not binomial distributed. However the expectation matches to that of a binomial rv!)

Now nothing particular abt $g(X,Y) = X+Y$, this linearity prop. of E is ~~induced~~ followed from the linearity prop. of integrals and summations. So in general one has:

eg

Consider $g(X,Y,Z) = \sum_{i=1}^{l} a_i f_i(X,Y,Z) + \sum_{i=1}^{m} b_i g_i(X,Z) + \sum_{i=1}^{n} c_i h_i(X) + d$

linear combination of fuctions of X, Y, Z.

It is easy to see that:

$$E[g(X,Y,Z)] = \sum_{i=1}^{l} a_i E[f_i(X,Y,Z)] + \sum_{i=1}^{m} b_i E[g_i(X,Z)] + \sum_{i=1}^{n} c_i E[h_i(X)] + d,$$

to compute we will need joint dist of X, Y, Z

to compute we need joint dist of X, Z

we need only dist. of X.

(Also this can be further generalized to fuctions over n r.v.s)

This is the <u>linearity property of Expectation.</u>

One can show another property of Expectation:

Suppose X, Y are <u>independent</u> r.v.s. Then:

$$E[\underbrace{g(X)\,h(Y)}_{f(X,Y)}] = E[g(X)]\,E[h(Y)] \qquad \text{II}$$

Proof: $LHS = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\, f_{XY}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y) f_X(x) f_Y(y)\,dx\,dy$

$$= \left(\int_{-\infty}^{\infty} g(x) f_X(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y) f_Y(y)\,dy\right) = RHS.$$

LHS involves double integral/summation whereas RHS involves two single integrals/summations. So it is useful observation.

Also, in general, we can show if X, Y are independent then

$g(X), h(Y)$ are also independent (provided $g(X), h(Y)$ are well defined r.v.s!)

Proof: $g(X), h(Y)$ are independent r.v.s

$\Leftrightarrow [g(X) \in B_1], [h(Y) \in B_2]$ are independent events $\forall B_1, B_2 \in \mathcal{B}$

$\Leftrightarrow [X \in g^{-1}(B_1)], [Y \in h^{-1}(B_2)]$,,

but $g^{-1}(B_1)$ and $h^{-1}(B_2)$ are some Borel sets!

which is true since X, Y are themselves independent r.v.s!

Moments of Collections of r.v.s

While discussing r.v.s we defined moments, absolute moments etc. Now we can extend these definitions:

$$\mu_{m,n} = E[X^m Y^n] \rightarrow m,n^{th} \text{ moment of } X, Y$$

(this is some function of X, Y hence we can compute its expectation!)

eg $\mu_{10} = E\{X\} = \mu_X$, $\mu_{01} = E\{Y\} = \mu_Y$, $\mu_{11} = E\{XY\} \equiv \mu_{XY}$. so on...

Similarly, one can extend the concept of central moments:

$$\sigma_{m,n} = E\left[(X-E\{X\})^m (Y-E\{Y\})^n\right] \to mn^{th} \text{ central moment of } X, Y.$$

eg. $\sigma_{10} = 0 = \sigma_{01}$, $\sigma_{20} \doteq var(X) = \sigma_X^2$, $\sigma_{02} = var(Y) = \sigma_Y^2$,

$$\sigma_{11} = E[(X-E\{X\})(Y-E\{Y\})] \equiv \underbrace{Cov(X,Y)}_{\text{new symbol}}$$

σ_{11} is called as covariance of X, Y. (Ofcourse $Cov(X,Y) = Cov(Y,X)$).

$Cov(X,Y)$ has some connection with the notion of "how correlated two rvs X, Y are". Lets explore this connection now:

(i) Suppose X, Y are independent. Then we can show $Cov(X,Y) = 0$

Proof: $Cov(X,Y) = E[(X-E\{X\})(Y-E\{Y\})]$

$$= E\left[(XY + E\{X\}E\{Y\} - XE\{Y\} - YE\{X\})\right]$$

$$= E[XY] - E\{X\}E\{Y\} \quad \to \text{linearity prop. of } E$$

$$= E\{X\}E[Y] - E\{X\}E\{Y\} \quad \to \because X, Y \text{ are independent (by II)}$$

$$= 0$$

So X, Y are independent $\Longrightarrow Cov(X,Y) = 0$

However $Cov(X,Y) = 0 \not\Longrightarrow X, Y$ are independent.

Here is the counter eg:

Consider $Y = X^2$ and X is such that $E\{X\} = E\{X^3\} = 0$.

Note that Y, X are surely dependent (not independent!)

However, for this eg: $Cov(X,Y) = Cov(X,X^2)$

$$= E\{X^3\} - E\{X\}E\{X^2\} = 0$$

So $Cov(X,Y) = 0 \not\Rightarrow X, Y$ are independent.

(Note that an eg. of X such that $E\{X\} = E\{X^3\} = 0$ is the std. Normal r.v. In fact in assignment you showed that all odd moments of a ~~Nor~~ std. Normal r.v are zero.) Also you showed that $Y = X^2$ has Chi-square distribution if X is std. Normal)

$\longrightarrow$ For X, Y Normal r.v ~~independent it turns out that~~ such that $Cov(X,Y) = 0$ it turns out that indeed X, Y are independent! So Normal r.v's are an exception! and the converse holds!!

We say X, Y are uncorrelated if $Cov(X,Y) = 0$.

(uncorrelated is hence in some sense weaker cond. than independence)

In fact we can ~~d~~ quantify the "correlation" between two r.v's using what is known as the correlation coefficient defined as:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

~~One~~ Of course $\rho_{XY} = 0 \Rightarrow Cov(X,Y) = 0 \Rightarrow X, Y$ are uncorrelated

Also one can show that:

$|\rho_{XY}| \le 1$ & $\rho_{XY} = \pm 1$ ~~implies~~ implies "perfect" correlation in the sense that X, Y are linearly dependent!

Proof: TST $|\rho_{XY}| \le 1$

i.e. TST $(\rho_{XY})^2 \le 1$

i.e. TST $(Cov(X,Y))^2 \le \sigma_X^2 \sigma_Y^2$

i.e. TST $\left(E[(X-E[X])(Y-E[Y])]\right)^2 \le E[(X-E[X])^2]\, E[(\cancel{X}-E[Y])^2]$

(Lets put $X' = X - E[X]$, $Y' = \cancel{X}Y - E[Y]$)

i.e. TST $\left(E[X'Y']\right)^2 \le E[X'^2]\, E[Y'^2]$ (III)

This is known as Cauchy-Schwartz inequality

This inequality also appears in linear algebra (vector spaces) and is a fundamental inequality. In fact this being satisfied in (III) form motivates the study ~~for~~ of vector spaces of r.v.s!!

$\rightarrow$ Here is some intuition:

Suppose there exists ~~for~~ some vector space in which inner product is ~~to~~ given by $E[X'Y']$ i.e. $\langle v_1, v_2 \rangle = E[X'Y']$

It is easy to see, $\langle v_1, v_1 \rangle = E[X'^2]$

$\langle v_2, v_2 \rangle = E[Y'^2]$

but we know that $(\langle v_1, v_2 \rangle)^2 \le \langle v_1, v_1 \rangle \langle v_2, v_2 \rangle$

$(\|v_1\| \|v_2\| \cos\theta)^2 \le \|v_1\|^2 \|v_2\|^2$

$\rightarrow |\cos\theta| \le 1$ which is true of course.

So (III) is extension of Cauchy-Schwartz inequality in case of Euclidean vectors!

Proof of III is simple (& typical wherever Cauchy-Swartz appears!)

Proof Consider $E\left[(aX'+Y')^2\right]$. we know it is ≥ 0.

$$\Rightarrow a^2 E\{X'^2\} + 2a\,E\{X'Y'\} + E\{Y'^2\} \geq 0$$

$$\Leftrightarrow \quad \left(E\{X'Y'\}\right)^2 \leq E\{X'^2\}E\{Y'^2\}$$

(discriminant ≤ 0)

Also note that strict equality appears if and only if $aX'+Y'=0$ (with prob. 1) i.e. X, Y are linearly dependent!

This proves overall claim that $|\rho_{XY}| \leq 1$ &

$$\rho_{XY} = \begin{cases} 0 & \text{in case where } X, Y \text{ are uncorrelated} \\ \pm 1 & \text{in case of "highest" correlation} \\ & \text{i.e. } X, Y \text{ are linearly related!} \end{cases}$$

Now go back to the example of swine flu. We want to know which of symptoms $X_1, X_2, \ldots, X_n$ is most important symptom that characterizes Y (presence of swine flu or not).

$\rightarrow$ (One answer) Compute $|\rho_{X_1Y}|, |\rho_{X_2Y}|, \ldots, |\rho_{X_nY}|$

whichever symptom has max $\nearrow$ we can say it has "highest correlation" with disease and we can hence declare it to be the most important symptom for the disease!

Now lets compute $\mathrm{var}(Z)$ where $Z = X+Y$, in terms of var & cov. of X, Y:

$$
\begin{aligned}
\mathrm{var}(Z) = \mathrm{var}(X+Y) &= E\left[\left(X+Y-E[X+Y]\right)^2\right] \\
&= E\left[\left((X-E[X]) + (Y-E[Y])\right)^2\right] \\
&= E\left[(X-E[X])^2\right] + E\left[(Y-E[Y])^2\right] + 2\,E\left[(X-E[X])(Y-E[Y])\right] \\
&= \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y). \quad \text{IV}
\end{aligned}
$$

~~Note that var(X+Y) = var(X) + var(Y) + 2 cov(X,Y) can be written in the following (weird) ways:~~

~~var(X+X~~

$\longrightarrow$

Apart from this sometimes "vectorial" versions of mean & variance are defined. Here's the motivation:

$\to$ X is an n-dimensional multivariate r.v.

Suppose we want to find $E[a^T X]$. $a^T X = \sum_{i=1}^{n} a_i X_i$.

$$E[a^T X] = E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i\, E[X_i] \quad \to \text{by linearity prop. of } E.$$

$$= a^T \underbrace{E[X]}_{\text{new notation } E[X] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}}$$

This $E[X]$ (exp. of multivariate r.v) is nothing but the vector of expectations of the individual r.v.s. This is also sometimes called as the 'mean vector' of the multivariate r.v X.

Now suppose we wish to find $\mathrm{var}(a^T X)$:

$$
\begin{aligned}
\mathrm{var}(a^T X) = E\left[\left(a^T X - E[a^T X]\right)^2\right] &= E\left[\left(a^T X - a^T E[X]\right)^2\right] \\
&= E\left[a^T(X - E[X])\, a^T(X - E[X])\right]
\end{aligned}
$$

since transpose of a number is the number itself we get

$$\mathrm{var}(a^T x) = E\Big[\, a^T \underbrace{(x - E\{x\})(x - E\{x\})^T}_{n\times n \text{ matrix}}\, a \,\Big]$$

by linearity prop. of E one can show $= a^T \Sigma a$ where

Σ is called as the covariance matrix whose entries are given by $\cancel{E}\ \Sigma_{ij} = \mathrm{Cov}(x_i, x_j)$

ij^{th} element of the covariance matrix.

→ for a 2-d case we can go through the steps easily:

$$\mathrm{var}(a_1x_1 + a_2x_2) = E\left[\left(a_1x_1 + a_2x_2 - E[a_1x_1 + a_2x_2]\right)^2\right]$$

$$= a_1^2\, \mathrm{var}(x_1) + a_2^2\, \mathrm{var}(x_2) + 2a_1a_2\, \mathrm{Cov}(x_1, x_2)$$

(by repeated appl. of linearity prop. of E, similar to (IV))

$$= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_1, x_2) & \mathrm{var}(x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$= a^T \Sigma a$$

→ Hence sometimes instead ~~talking~~ of talking about moments and central moments of collections of r.v.s, people talk abt mean vector and covariance matrix of the corresponding multivariate r.v.

CONDITIONAL EXPECTATION

Suppose X, Y are two random variables. Now in all cases
(i) X, ~~Y~~ Y are discrete (ii) X, Y are jointly conts (iii) one of them is conts. other is discrete,

we defined $f_{X/Y}(x/y) \longrightarrow$ either conditional pmf or conditional pdf given $Y=y$
↓ if X is discrete ↓ if X is conts.

Now the rv for which ↑ is the pmf or pdf is denoted by:

$$Z = X/Y=y$$

We already now that Z exactly takes those values which X takes and its pmf/pdf is given by $f_{X/Y}(x/y)$.

Since Z is a random variable we can talk about its expectation:

$$E[Z] = E[X/Y=y] = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x\, f_{X/Y}(x/y)\, dx & \text{if } X \text{ is conts rv.} \\ \displaystyle\sum_{\forall x_i} x_i\, f_{X/Y}(x_i/y) & \text{if } X \text{ is discrete rv.} \end{cases}$$

This is called as conditional expectation of X given that $Y=y$.

⟶ Now we can further extend this concept:

X, Y are rvs & say $Z=g(X,Y)$

we can talk abt. $f_{X/Z}(x/z)$ i.e. $f_{X/g(X,Y)}(x/z)$

and in turn talk abt. $E[X/g(X,Y)=z]$ and so on....