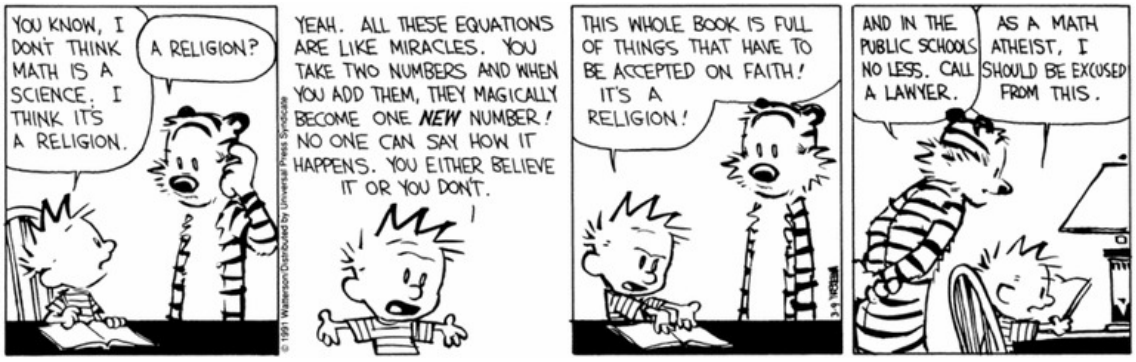

CMI/BVR Probability notes 3

independent repetitions:

We can think of the earlier example as a general method of repeating independently any experiment, not necessarily tossing a coin.

Let (Ω, p) be an experiment. Thus Ω is a countable set (could be finite or countably infinite) and p associates a non-negative number with each point of Ω and these numbers add to one.

Independent repetitions of this experiment n times means the following.

Sample space is

$$\Omega^* = \Omega^n = \{(\omega_1, \omega_2, \cdots, \omega_n) : \omega_i \in \Omega \quad \forall i\}.$$

$$p^*(\omega_1, \omega_2, \cdots, \omega_n) = \prod_1^n p(\omega_i)$$

We multiplied because we want the repetitions to be independent, thus it is dictated that chance of ω_i for $i \leq n$ should be the product of their individual chances.

Note that these are non-negative numbers and they add to one (why?).

binomial probabilities:

Example: Toss a coin independently n times; chance of Heads in a toss is p $(0 < p < 1)$.

Thus the experiment is tossing a coin. If you do once the space is

$$\Omega = \{H, T\}; \quad P(H) = p; \;\; P(T) = 1 - p.$$

Instead of denoting probabilities of outcomes with $p(H)$ etc which confuses with given data about chance of heads; we used $P(H)$.

You need to repeat this experiment n times.

$\Omega^* =$ sequences of length n consisting of the letters H, T. If ω is such a sequence then

$$P^*(\omega) = p^{\#H \text{ in } \omega} \;\; (1-p)^{\#T \text{ in } \omega}$$

Let us calculate some probabilities of events. Let X be defined on Ω^* as follows; For each n-tuple ω we put $X(\omega)$ to be the number of H in that outcome.

The event $(X = 0)$ has only one outcome $(T, T, \cdots, T)$ and so

$$P(X = 0) = (1-p)^n$$

In general if you take $0 \leq k \leq n$ then the event

$$\{\omega : X(\omega) = k\}$$

has $\binom{n}{k}$ outcomes and each such outcome has probability $p^k(1-p)^{n-k}$. Thus

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}; \qquad k = 0, 1, 2, \cdots, n.$$

These probabilities are called binomial probabilities. As you see if you do binomial expansion of $[p + (1-p)]^n$, you get the numbers above.

geometric probabilities:

Toss the above coin independently till you get a Head and then stop.

$$\Omega = \{H, \quad TH, \quad TTH, \quad TTTH, \cdots\}$$

The outcome $T^k H$ has probability $(1-p)^k p$. Let us define X as the number of failures before success, that is, number of T before H. Then X takes values $0, 1, 2, \cdots$. For each k the event $\{\omega : X(\omega) = k\}$ has only one outcome, namely $T^k H$. Thus

$$P(X = k) = p(1-p)^k; \qquad k = 0, 1, 2, \cdots$$

These are called geometric probabilities, because these are terms of a geometric series.

There is a subtle point. You can doubt if we have forgotten to write one outcome namely 'all T'. It is theoretically conceivable that in tossing the coin you may never get H. In such a case the tossing never ends and so it continues for ever. However, this theoretical possibility has probability zero. Argue that chances of getting all k tails is $(1-p)^k$. So getting all tails has probability smaller than $(1-p)^k$ for each k and hence must be zero. Remember $0 < p < 1$.

You can also argue that the sum of the probabilities of all the outcomes we have written above add to one. So chances of anything else must be zero.

You can also not pay attention to this discussion and accept on faith (not really advisable) that the above written things are the only outcomes.

negative binomial probabilities:

Fix an integer $r \geq 1$. Toss the above coin independently till you get r Heads and stop. It is difficult to write all the outcomes. But we can describe. An outcomes consists of a sequence of of letters H, T; and H must be exactly r times; and last letter must be H and before this last occurrence, H should appear exactly $r-1$ times. Any such outcome must have probability $p^r(1-p)^k$ where k is the number of T in the outcome.

To calculate probabilities of some events, put X as the number of failures (T) in that outcome. Possible values of X are again $\{0, 1, 2 \cdots\}$. Now of course the event $(X = k)$ would in general have more than one outcome. In fact if you want k tails, the number of tosses must be $k+r$, since the last one must be H, we must have the k tails somewhere in the first $k+r-1$ tosses. Each such outcome has the same probability, namely, $p^r(1-p)^k$. Thus

$$P(X = k) = \binom{k+r-1}{k} p^r(1-p)^k; \qquad k = 0, 1, 2, \cdots .$$

These are called negative binomial probabilities because the terms above occur in a specific binomial expansion with negative exponent.

In fact, by binomial expansion

$$[1-(1-p)]^{-r} = \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (1-p)^k$$

but

$$\begin{aligned}\binom{-r}{k} &= \frac{(-r)(-r-1)(-r-2)\cdots[-r-(k-1)]}{1\cdot 2\cdot 3\cdots k} \\ &= (-1)^k \binom{r+k-1}{k}\end{aligned}$$

hence

$$[1-(1-p)]^{-r} = \sum_{k=0}^{\infty} \binom{r+k-1}{k} (1-p)^k.$$

Hypergeometric probabilities:

We have a box of items of which N_1 are good and N_2 are defective. We select a subset of size n at random. Thus the number of outcomes is

$$\binom{N_1+N_2}{n}.$$

Let us define X to be the number of defective items in the selected subset. The variable X takes values $0, 1, 2\cdots, n$.

$$P(X=k) = \binom{N_2}{k}\binom{N_1}{n-k} \Big/ \binom{N_1+N_2}{n}\ ; \qquad k = 0, 1, 2\cdots n.$$

There is a subtle point: Suppose there are 10 good and 3 defective items. We take a subset of 5 items. Can the number of defectives be 4 or 5? No because there are only 3 defectives. But in our above calculation we said that X can take values $0, 1, 2\cdots n$. But do not worry the unnecessary numbers get probability zero automatically. If $k > N_2$ or $n-k > N_1$, then the above quantities reduce to zero.

But then why did we list outcomes which may not occur? Just to avoid notational complications (or ugly expressions).

These are called Hypergeometric probabilities. Because these arise as coefficients in Hypergeometric function (or actually, series) discussed by Gauss.

Incidentally, consider the experiment above, instead of picking a set of n items; you take a sample without replacement of size n. Then ask the same question. You get the same answer as above.

Random variables, distribution, moments:

Most of the time we perform an experiment and measure some quantity as in the above examples. Such measurements are called random variables. Thus **a random variable** on a probability space (Ω, p) is simply a real valued function defined on Ω.

For example, we may select a person in this room and measure his/her height. Thus the person is unimportant, it is the height that is important. Sample points of this experiment are all persons in this room. For each outcome, height of that outcome is the measurement or random variable — more precisely, value of the random variable for the outcome.

For example in tossing a coin 50 times we have 2^{50} outcomes. Most probably the interest could be the number of heads obtained rather than the place of their occurrence. Then the experiment can be summarised by 51 numbers rather than hanging on to the 2^{50} outcomes. For each outcome, the number of heads in that outcome is the measurement or value of our random variable for that outcome. In some sense the random variable summarizes the experiment.

A table giving values of the random variable along with the probability of assuming that value is called **distribution** of the random variable. Some important distributions that we come across in practice, have names.

Returning to the above examples, here are their names.

In the example of tossing a coin n times; $X =$ the number of heads in the outcome, is a random variable. We say its distribution is **binomial distribution** and the random variable X is **Binomial random variable**. More precisely,

$$X \sim B(n, p);$$

X is binomial with parameters n and p.

In the example of tossing a coin till H is obtained (and then stop), $X =$ the number of failures; is a r.v. Its distribution is called **geometric dis-**

tribution and the random variable is called **geometric random variable**. More precisely

$$X \sim G(p)$$

X is geometric with parameter p.

In the example of tossing a coin till r many H are obtained (and then stop), X = the number of failures is a r.v. Its distribution is called **negative binomial distribution** and the random variable is called **negative binomial random variable**. More precisely

$$X \sim NB(p).$$

X is negative binomial with parameters r and p.

In the example of X = the number of defectives in a sample of size n from a lot of N_1 good and N_2 defective items; the distribution is called **Hypergeometric distribution** and the random variable is called **Hypergeometric random variable**. More precisely

$$X \sim HG(N_1, N_2; n)$$

X is hypergeometric with parameters N_1, N_2 and n.

A random variable X which takes two values zero and one with probabilities $1-p$ and p is called a **Bernoulli random variable**; Bernoulli with parameter p. Sometimes a random variable X which takes two values ± 1 with probabilities $1-p$ and p is called a Bernoulli random variable. Some people refer to any random variable that takes two values as Bernoulli random variables.

mean and variance:

Consider the following game:
Game 1: We toss a fair coin. If Heads up you get one rupee; tails up you give me one rupee or equivalently you get (-1) rupee. is this a fair game?

Even though one of us is sure to loose in a single game; if we play this game a large number of times approximately half of the time you loose a rupee and half of the time you gain a rupee; so on the average it appears fair.

Game 2: We toss a fair coin. If Heads up you get 1000 rupees; tails up you give me 1000 rupees or equivalently you get (-1000) rupees. is this a fair game?

Arguing as earlier, if we play this game a large number of times approximately half of the time you loose and half of the time you gain a certain amount; so on the average it appears fair.

Both are fair, which one are you willing to play?

Clearly you will hesitate to play the second game because it is risky. It may so happen the first ten game you loose (and later you may win twelve games). In other words the stakes in the second game are too spread out.

Thus if X is the profit then the average value of X and the spread of X are important numbers that give an idea of the random variable. This is what we capture now.

If X is a random variable taking values

$$x_1, \quad x_2, \quad x_3, \cdots, \quad x_k, \cdots$$

with probabilities

$$p_1, \quad p_2, \quad p_3, \cdots, \quad p_k, \cdots$$

then we define mean value of X as follows:

$$E(X) = p_1x_1 + p_2x_2 + p_3x_3 + \cdots \quad = \quad \sum x_k p_k$$

provided this sum is convergent.

Is this a good definition? No. What is bad about it? No one told us how to list the values of the variable. Suppose some one writes that the values are

$$x_2, \quad x_1, \quad x_4, \quad x_3, \quad \cdots\cdots$$

with probabilities

$$p_2, \quad p_1, \quad p_4, \quad , p_3, \cdots\cdots .$$

Then he/she has written the same thing as we did. So they calculate

$$x_2p_2 + x_1p_1 + x_4p_4 + x_3p_3 + \cdots\cdots .$$

Will they get the same answer as we did? If they do not get the same answer then the definition depends on the order in which the values are listed and

is no good.

In fact any permutation of values and the corresponding permutation of probabilities should lead to the same answer. You realize that a series $\sum a_k$ of real numbers converges irrespective of the ordering of the terms of the series **iff** the series $\sum |a_k|$ is convergent. More precisely if $\sum a_{\pi(k)}$ converges for every permutation π of $\{1, 2, \cdots\cdots\}$ then the series must be absolutely convergent.

If in our case, for example $\sum x_k p_k$ converges and $\sum |x_k p_k|$ does not converge then given any number c we can produce a permutation π of natural numbers so that $\sum x_{\pi(k)} p_{\pi(k)}$ converges to the number c.

With this background we make the following definition.

If a random variable X takes values $\{x_k : k \geq 1\}$ (finitely many or infinitely many) with probabilities $\{p_k : k \geq 1\}$ respectively, then we say expectation of X is defined in case $\sum |x_k p_k|$ converges. In that case we define **expected value** or **average value** or **mean value** of X by

$$E(X) = \sum x_k p_k.$$

Note that the same intuitive feeling as we had earlier still holds, namely, if you observe the variable X many times you will see the value x_k a proportion p_k times (approximately). thus average of your observations is approximately the above number.

Regarding spread, you realize that a variable which takes values $-1, +1$ has the same spread as the variable taking values $99, 101$. The first one is spread around zero where as the second one is spread around 100 but spread is only two units. Keeping this in mind we make the following definition.

If a random variable X takes values $\{x_k : k \geq 1\}$ (finitely many or infinitely many) with probabilities $\{p_k : k \geq 1\}$ and if its mean value is θ the **variance** of the random variable X is defined by

$$\text{variance } (X) = \sum (x_k - \theta)^2 p_k$$

Why did we square: if you did not then you see

$$\sum (x_k - \theta) p_k \quad = \quad 0.$$

Of course, instead of squaring, you could have taken $|x_k - \theta|$ or $|x_k - \theta|^6$ etc. The first alternative uses modulus function which is not differentiable and will cause problems in analyzing this quantity. We shall see more on this later. Regarding the second alternative, you realize that sixth power is more complicated than second power.

Usually mean of a random variable X is denoted μ_X and variance by σ_X^2. If you know that a quantity is always non-negative, then in statistics it is customary to notate it by square of something. So the notation draws your attention to the fact that the quantity is non-negative.

Actually one can define 'moments' of higher order too as follows. For $n \geq 1$ the n-th moment of the random variable X is

$$\mu_n = \sum x_k^n p_k; \quad \text{provided} \quad \sum |x_k^n p_k| < \infty.$$

Binomial:

Let us consider $X \sim B(n,p)$ and calculate its mean and variance.

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$= np \sum_{k=1}^{n} \binom{n-1}{k} p^{k-1} (1-p)^{n-k} = np[p + (1-p)]^{n-1} = np.$$

This is expected, because if chance of heads is $1/5$ then you expect roughly $n/5$ heads in n tosses.

$$\text{variance}(X) = \sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \sum_{k=0}^{n} [k(k-1) + k + n^2p^2 - 2npk] \binom{n}{k} p^k (1-p)^{n-k}$$

using similar calculation as above

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

WLLN for coin tossing:

Before we calculate means and variances for other random variables, let us ask ourselves:

is our maths on the right track?

When we said chance of heads is p, our intuitive idea is that in a large number of tosses a proportion p will be heads. We have developed the concept of 'independence' and 'tossing a coin n times' purely mathematically. Can we prove that our feeling is reflected in our Maths? Yes.

Chebyshev's inequality:

■ 1) Let X be a random variable which takes non-negative values. For any real number $a > 0$

$$P(X \geq a) \leq E(X)/a.$$

2) Let X be a random variable with finite mean μ and variance σ^2. For any number $a > 0$

$$P(|X - \mu| \geq a) \leq \sigma^2/a^2. \blacksquare$$

Proof is simple and will be done soon. But what is its importance for us?

A Weak Law of Large Numbers:

■ Suppose we have a coin with chance of heads p $(0 < p < 1)$. Whatever $\epsilon > 0$ be given, the chances that the proportion of heads in n tosses lies between $(p - \epsilon, p + \epsilon)$ gets closer and closer to one as the number of tosses increases.

More precisely, Let X_n be the number of heads in n independent tosses of a coin whose chance of heads is p. Then for any $\epsilon > 0$

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \epsilon\right) \to 0; \quad n \to \infty. \blacksquare$$

Observe that X_n/n is indeed the proportion of heads in n tosses. The chances that the proportion differs from p by an amount larger than prescribed ϵ gets closer and closer to zero; more precisely, it converges to zero.

To prove this result, first note $E(X_n) = np$ so that Chebychev tells you the required probability is same as

$$P(|X_n - np| \geq n\epsilon) \leq \frac{np(1-p)}{n^2\epsilon^2} = \frac{p(1-p)}{\epsilon^2} \ \frac{1}{n} \to 0.$$

Returning to Chebyshev, second part follows from first because

$$P(|X - \mu| > \epsilon) = P(|X - \mu|^2 > \epsilon^2)$$

and first part tells

$$\leq E(X-\mu)^2/a^2 = \sigma^2/a^2.$$

(did I use something here?)

Proof of first part is also easy:

$$E(X) \;=\; \sum x_k p_k \;\geq\; \sum_{x_k \geq a} x_k p_k \;\geq\; a \sum_{x_k \geq a} p_k \;=\; aP(X \geq a)$$

as required.

$G(p)$**:**

Let $X \sim G(p)$. what is expected value of X? Intuitively you feel that if chance of heads is $1/7$ then roughly, you need 7 tosses to see Heads. In general you need roughly $1/p$ many tosses to see one head. Does our math show it? Yes. For ease in writing let us use q for $(1-p)$.

$$E(X) \;=\; \sum_0^\infty kpq^k \;=\; pq \sum_1^\infty kq^{k-1} \;=\; pq \sum_0^\infty \frac{d}{dq} q^k$$

$$= \; pq \frac{d}{dq} \frac{1}{(1-q)} \;=\; pq \frac{1}{(1-q)^2} \;=\; \frac{q}{p}.$$

Remember X is the number of tails. So number of tosses equals $1+X$. Thus expected number of tosses needed to get a Head equals

$$E(1+X) \;=\; 1 + \frac{q}{p} \;=\; \frac{1}{p}$$

as expected.

We shall now calculate variance.

$$\text{variance } (X) = \sum_0^\infty (k - \frac{q}{p})^2 q^k p = \sum_0^\infty \left[k^2 - 2\frac{q}{p}k + \frac{q^2}{p^2} \right] q^k p$$

$$= \sum_0^\infty [k(k-1)+k] q^k p \;\; - 2\frac{q}{p}\frac{q}{p} + \frac{q^2}{p^2}$$

$$= q^2 p \sum_2^\infty \frac{d}{dq} q^{k-2} \;\; + \frac{q}{p} - \frac{q^2}{p^2}$$

$$= q^2 p \frac{d}{dq} \frac{1}{1-q} \quad + \frac{q}{p} \quad - \frac{q^2}{p^2}$$

$$= q^2 p \frac{2}{p^2} \quad + \frac{q}{p} - \frac{q^2}{p^2}$$

$$\frac{q^2}{p^2} + \frac{q}{p} = \frac{q}{p}\left[1 + \frac{q}{p}\right]$$

$$= \frac{q}{p^2}.$$

Things to think about:
is term-by-term differentiation justified?
Instead of justifying this, can we bypass it?

Probability and conditional probability:

Let us collect at one place the properties that probability has.

Let (Ω, p) be an experiment, that is, Ω is accountable set and p is non-negative real valued function on Ω and $\sum p(\omega) = 1$. This is given to us.
Then we define for subsets $A \subset \Omega$;

$$P(A) = \sum_{\omega \in A} p(\omega).$$

We make the convention that empty sum is zero.

■ **Theorem: Probability properties**

1. $0 \leq P(A) \leq 1$ for all events A. Further $P(\Omega) = 1$ and $P(\emptyset) = 0$.

2. If $A \subset B$, then $P(A) \leq P(B)$.

3. For a sequence of disjoint events $A_1, A_2, A_3, \cdots$ (finite or infinite)

$$P(\cup A_n) = \sum P(A_n).$$

(4) the inclusion-exclusion formula holds for the union of any finitely many events. ■

Proof:

1. $P(A) = \sum_{\omega \in A} p(\omega)$ sum of non-negative numbers and hence (?) non-negative.

$$P(A) = \sum_{\omega \in A} p(\omega) \leq \sum_{\omega \in \Omega} p(\omega)$$

The inequality is because on the right side there are more terms and all are non-negative.

$P(\Omega) = 1$ follows from definition of p; and $P(\emptyset) = 0$ is by definition of empty sum.

2. is already included in the above argument: $P(B)$ adds more non-negative numbers.

3. If $\cup A_n = \{\omega_1, \omega_2, \cdots\}$ then by definition

$$P(\cup A_n) = p(\omega_1) + p(\omega_2) + \cdots\cdots .$$

by definition of sum of series this means

$$P(\cup A_n) = \lim_k \sum_{i=1}^{k} p(\omega_i)$$

If you take any k, then these finitely many points $\{\omega_i : i \leq k\}$ will be already in finitely many of the $\{A_n : n \leq n_0\}$ (say). Then easy to see, by monotonicity (part 2 above) that

$$\sum_1^k p(\omega_i) \leq \sum_1^{n_0} P(A_n) \leq \sum_1^{\infty} P(A_n)$$

This being true for every k we conclude

$$\sum_1^{\infty} p(\omega_i) \leq \sum_n P(A_n)$$

That is

$$P(\cup A_n) \leq \sum P(A_n)$$

To complete the proof, we shall now show the other inequality

$$\sum P(A_n) \leq P(\cup A_n)$$

Le

$$A_n = \{\omega_1^n, \omega_2^n, \cdots\}$$

so that

$$P(A_n) = \sum_{i=1}^{\infty} p(\omega_i^n) = \lim_m \sum_{i=1}^{m} p(\omega_i^n)$$

Thus

$$\{\omega_i^n : n \leq k, i \leq m\} \subset \cup A_n$$

and the elements listed left side are distinct because A_n are disjoint. Observe left side is a finite set. Hence

$$\sum_{n=1}^{k} \sum_{i=1}^{m} p(\omega_i^n) \leq P(\cup A_n)$$

Taking limit as $m \to \infty$ we see

$$\sum_{n=1}^{k} P(A_n) \leq P(\cup A_n)$$

This being true for every k we see

$$\sum P(A_n) \leq P(\cup A_n)$$

Proof of (3) is complete. We had earlier proved (4). ∎

In the proof above we pretended that our sets are all infinite. You can easily modify it if some sets are finite. Actually it is better to make a convention that even for a set consisting of finitely many out comes, its probability is: sum of probabilities of outcomes in it followed by infinitely many zeros if needed. Since I do not want to spend time on unimportant point, I leave it here.

Whatever is true of probabilities is also true bout conditional probabilities.

∎ **Theorem: conditional probability properties**

Let $P(D) > 0$.
1. $0 \leq P(A|D) \leq 1$ for all events A.
Further $P(\Omega|D) = 1$ and $P(\emptyset|D) = 0$.

2. If $A \subset B$, then $P(A|D) \leq P(B|D)$.

3. For a sequence of disjoint events $A_1, A_2, A_3, \cdots$ (finite or infinite)

$$P(\cup A_n|D) = \sum P(A_n|D).$$

(4) the inclusion-exclusion formula holds for the union of any finitely many events even for conditional probabilities. ∎

All these follow by simply writing the definition $P(A|D) = P(A \cap D)/P(D)$ and using earlier result.

∎ **Theorem: Random variables Properties**

Let (Ω, p) be a probability space and L be the collection of all random variables. Then L is a real vector space under usual operations. ∎

Proof is simple. Probability has no role to play. L is nothing but the collection of real valued functions defined on the countable set Ω. ∎

∎ **Theorem: Expectation properties**

Let (Ω, p) be a probability space.

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

2. Let L_1 be the collection of all random variables for which expectation is defined. Then L_1 is a vector space with usual operations. The map $X \mapsto E(X)$ is a linear map on L_1. ∎

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

Note that any sum where all terms are non-negative can be rearranged and added in any manner and we get the same answer. For example, suppose you are adding

$$\sum_1^\infty y_n$$

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

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

You can add those y_n such that $n \in A_k$ only and obtain the sum. Let us say z_k. Now add these z_k to get z. No matter how you make the partition $\{A_k\}$

and calculate you will get the same final answer. Thus by one method if you get 25 (or ∞) then you will get 25 (or ∞) by any other method. You must keep in mind that terms (y_n) are non-negative.

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

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

Also when a sum is absolutely convergent then you can rearrange the series in any manner and add, you get the same value. We shall use these two facts now.

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

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

Note that

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

Also

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

From the observation made above

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

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

But if the expectation is defined, that is, left side is finite, then the series $\sum X(\omega) p(\omega)$ is absolutely convergent. Now use the second observation made

above to see you can make subtotals over each A_k and then add them. Thus you see

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

Proof of 2. Let X and Y be random variables and $Z = X + Y$, that is $Z(\omega) = X(\omega) + Y(\omega)$. Note that if

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

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

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

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

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

or by using part 1 above,

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

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

This proves the second part of the theorem. ∎

You must appreciate part 1, which allowed you to do linearity. After all if you know distribution of X and Y then there is **no** way of calculating distribution of $X + Y$. Thus if you depended on the definition of expectation via distribution ($\sum x_k p_k$) then there is no way to conclude the above.

Scenario 1:
Let us toss a fair coin independently twice. Then outcomes are

$$HH, HT, TH, TT$$

each probability 1/4.

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

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

$X + Y$ is always 2.

Thus $X + Y$ takes only one value 2 with probability one.

Scenario 2:

Now consider another experiment. tossing a coin four times independently. Then there are 16 outcomes each with probability 1/16.

Let X be the number of heads in the first two tosses and Y be the number of heads in the last two tosses. then these have the same distributions as the X, Y of scenario 1. However $X + Y$ now takes values

$$\{0, 1, 2, 3, 4\}$$

with probabilities

$$\{\frac{1}{16}, \quad \frac{4}{16}, \quad \frac{6}{16}, \quad \frac{4}{16}, \quad \frac{1}{16}\}$$

respectively. this is different from the distribution of $X + Y$ in scenario 1.

Thus knowing distributions of X and Y separately will not allow you to calculate distribution of $X + Y$.

Expected number of matches:

The fact sum of expectations equals expectation of sum is a very very useful result. This is true not only for two but any finite number of random variables.

Let us consider the matching problem with 52 cards and 52 envelopes.

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

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

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

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

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

and

$$E(X) = 1.$$

A hopeless situation is rescued by the linearity of expectation.

Also note that whatever be the number of letters and envelopes the expected number of matches is one!

joint distribution:

We saw that from the distribution of random variables X and Y we can not get distribution of $X+Y$. We can not understand any relationships that exist between the random variables. In scenario one there is perfect relation between the variables, $Y = 2 - X$. In the second scenario, they are actually independent, in the following sense: for any two values i, j the events $(X = i)$ and $(Y = j)$ are independent, that is,

$$P(X = i \ \& \ Y = j) = P(X = i)P(X = j).$$

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