Lecture 8: Random Variables and their distributions

Binomial Distribution $X \sim \text{Bin}(n, p)$

n → any positive integer
p → real number between 0 & 1

(1) Story: X is # successes in n independent $\text{Bern}(p)$ trials

p → prob. of success

$X \sim \text{Bin}(n,p)$ → X is a random variable that has this distribution

(2) Sum of indicator random variables: $X = X_1 + X_2 + \dots + X_n$,

where $X_j = \begin{cases} 1, & \text{if } j\text{th trial success} \\ 0, & \text{otherwise} \end{cases}$

$X_1, X_2, \dots X_n$ → R.V.s is mathematically a function, but intuitively, X_1 is 1 if first trial is success, 0 otherwise that depends on first trial

$X_1, \dots, X_n$ i.i.d. $\text{Bern}(p)$

i.i.d. → independent & identically distributed

Trials are independent and $X_1, X_2, \dots, X_n$ are the indicators of success for each trial, so those should be independent.

identically — All of these X's (i.e., $X_1, X_2, \dots, X_n$) have the same distribution. In other words, they are all $\text{Bern}(p)$.

(3) PMF (What's the prob. that X takes on any particular value)

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad q = (1-p)$$

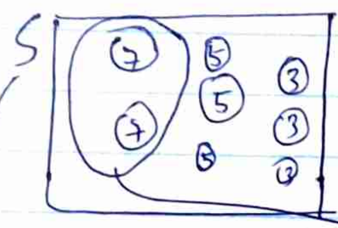
$$k \in \{0, 1, \dots, n\}$$

$P(X=k)$ → n trials, exactly k successes

Random variables (R.V.s.)

→ different from distributions
Distribution — saying what is the prob. that X will behave in different ways. $X_1, X_2, \ldots X_n$ → different random variables but all may have the same distribution.

S



→ Could be incredibly complicated possibly infinite space that we could never draw.
→ could be very high dimensional, infinite.

- Random variable is a function that assigns a number to each pebble (in pebble world case).

$(X = 7)$ is an event
↳ subset of sample space

X is a function.
Here, $X = 7$ is not a const. function, it's just a notation for an event that $X = 7$.

CDF (Cumulative distribution Function)

→ one way to describe a distribution

→ telling prob. of different possible values of X

$X \leq x$ is an event
$F(x) = P(X \leq x)$,
then F is called CDF of X.

PMF (for discrete random variables)

↳ possible values $a_1, a_2, a_3, \ldots, a_n$ or $a_1, a_2, \ldots$

$P(X = a_j)$, for all j
$= P_j$

$$\boxed{P_j \geq 0, \sum_j P_j = 1}$$ (When does a PMF valid?)

$$P(X=k) = \binom{n}{k} p^k q^{n-k}, \quad q = 1-p$$
$$k \in \{0, 1, \ldots, n\}$$

$$\left[\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1 \right]$$
(By Binomial theorem)

$X \sim Bin(n,p)$ [no. of successes in n trials]
$Y \sim Bin(m,p)$ [No. of successes in m trials]
} Independent

X, Y are separate set of trials

$\Rightarrow (X+Y) \sim Bin(n+m, p)$

$\hookrightarrow$ Adding two functions

(1) $X = X_1 + X_2 + \dots + X_n$

$Y = Y_1 + Y_2 + \dots + Y_m$, X_j and Y_j are indep. Bernoulli random variables

$$X+Y = \sum_{j=1}^{n} X_j + \sum_{i=1}^{m} Y_i$$

Sum of $(n+m)$ i.i.d. Bern(p) $\Rightarrow$ Bin$(n+m, p)$

(2) $P(\underbrace{X+Y}_{\text{convolution}} = K) = \sum_{j=0}^{K} P(X+Y=K \mid X=j)\, P(X=j)$ [Condition on X. wishful thinking. It would be easier if we know the value of X]

(Law of Total Prob.)

check $X+Y$ is Binomial by computing its PMF

$$= \sum_{j=0}^{K} P(Y=K-j \mid X=j) \underbrace{\binom{n}{j} p^j q^{n-j}}_{\substack{\text{Binomial PMF} \\ \text{since } X \text{ is Binomial}}}$$

we can also condition on Y

Since, X & Y are independent, so Conditioning on X gives no information about Y.

$$= \sum_{j=0}^{K} P(Y=K-j) \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{K} \binom{m}{K-j} p^{K-j} q^{m-k+j} \binom{n}{j} p^j q^{n-j}$$

$$= p^K q^{m+n-K} \sum_{j=0}^{K} \binom{m}{K-j}\binom{n}{j}$$

Vandermonde's identity

$$= p^K q^{m+n-K} \binom{m+n}{K} = \binom{m+n}{K} p^K q^{m+n-K}$$

Example 5 card hand, find distribution of number of aces.

discrete problem as #aces so, PMF (or, CDF) can be either 0, 1, 2, 3 or 4

Solu:- Let $X = \#$ aces

find $P(X=k)$.

This is 0, except if $k \in \{0,1,2,3,4\}$.

- Distribution is not Binomial as trial is not independent, because we can think of each card as trial, but those trials are not independent because if first card is an ace, it is less likely that second card is an ace. The more aces we have in the earlier cards, it is less likely to have more aces.

(PMF just by thinking about it)

$$P(X=k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}, \text{ for } k \in \{0,\dots,4\}$$

(Like the elk problem)

↳ same as tagged elks & untagged elks

Example ~~we~~ Suppose we have a jar ful of marbles. We have b black, w white marbles. Pick simple random sample of size n. Find distribution of number of white marbles in sample.

Solu:- $X = \#$ number of white marbles in sample

$$P(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

choose k white marbles

$w+b \rightarrow$ total population

$n \rightarrow$ sample

(Hyper geometric distribution)

$\boxed{0 \le k \le w}$, $\boxed{0 \le n-k \le b}$

Not binomial (sampling without replacement)

Hypergeometric Distribution is different from $\underbrace{\text{Binomial Distribution}}$,

↳ sampling with replacement

↓

picking marble
without replacement
(sampling without replacement)

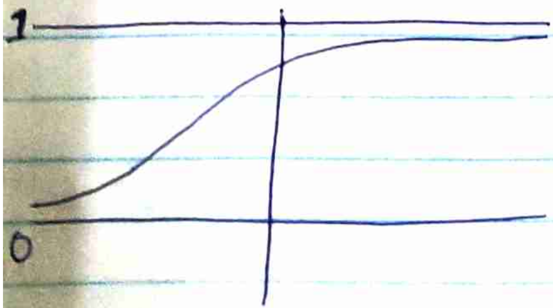
- When size of the population is very large, then both behave almost the same.

- Suppose number of marbles is billion and our sample is very small compared to a billion (let's say 10). We are picking 10 marbles out of a billion, it is extremely unlikely that we pick the same marble more than 1. So, it must be that sampling with or without replacement should behave very similarly there.

$$\frac{\sum_{k=0}^{w} \binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = \frac{1}{\binom{w+b}{n}} \underbrace{\sum_{k=0}^{w} \binom{w}{k}\binom{b}{n-k}}_{\text{VanderMonde's Identity}} = 1$$

CDF $P(X \leq x)$

Continuous



discrete