Cumulative Distribution Function

CDF $\qquad X \leq x$

$$F(x) = P(\underbrace{X \leq x}_{\text{event}})\,.$$

$$\{ s : X(s) \leq x \}$$

Probability Mass Function.

PMF - only for discrete distributions.

discrete RV takes a set of possible values $a_1, a_2 \ldots$

PMF $\quad P(X = a_j)$ for all j

let $\quad p_j = P(X = a_j)$

$$p_j \geq 0$$

$$\sum_j p_j = 1$$

< 1 - we have omitted some possibilities

> 1 - not a probability.

Discrete RV - PMF is often more convenient than CDF.

but CDF works for a continuous RV.

Bin (n, p)

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad 0 \leq k \leq n$$

$$= 0 \qquad \text{otherwise}$$

Is this a valid PMF?

$$P(X=k) \geq 0$$

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1.$$

Sum of 2 Binomials.

$X \sim \text{Bin}(n, p)$ $\qquad$ X, Y are independent

$Y \sim \text{Bin}(m, p)$

$$Z = \underset{\substack{n \\ \text{trials}}}{X} + \underset{\substack{m \\ \text{trials}}}{Y} \sim \text{Bin}(n+m, p).$$

X, Y are independent $\Rightarrow$ Z is # successes in $m+n$ trials where each trial has prob p of success.

$$X = X_1 + X_2 + X_3 + \dots X_n$$

Binomial RV is the sum of a set of Bernoulli RV

$$Y = Y_1 + Y_2 + \dots + Y_m$$

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

sum of $(m+n)$ Bernoulli (p) RV

$$\Rightarrow \quad \text{Bin}(m+n, p)$$

---

using the PMF

$$P(X+Y = k).$$

$$\sum_{j=0}^{k} P(X+Y = k \mid X = j)\, P(X=j).$$

$$\sum_{j=0}^{k} P(Y = k-j)\, P(X=j).$$

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

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

what is $\sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j}$.

we have $m+n$ things of two different types.

chose k of the $m+n$ things

where $(k-j)$ are of type 1 and

j are of type 2.

The summation is over all the ways of splitting the k objects into some from type 1 and some from type 2.

i.e. it is $\binom{m+n}{k}$

$$P(X+Y=k) = \binom{m+n}{k} p^k q^{m+n-k}.$$

ie sum of two Binomial distributions is itself Binomial.

(same p ; X, Y independent).

Random 5 card hand.

what's the distribution of the number of aces?

X = # aces

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

$P(X = k) = 0$ unless $k \in \{0, 1, 2, 3, 4\}$

trials are not independent.

$$P(X = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}$$

$$\frac{\text{4 aces choose } k \quad \text{48 othe cards choose } 5-k.}{\text{\# of 5 card hands.}}$$

Generalize.

Deer on campus.

tagged deer $= t$

untagged deer $= d$

Pick a simple random sample of size n (all subsets of size n are equally likely).

k - number of tagged deer in the sample of size n

$P(X=k)$ - Prob of getting k tagged deer in a sample of size n where there are t tagged deer and d untagged deer in the population

$$= \frac{\binom{t}{k} \times \binom{d}{n-k}}{\binom{t+d}{n}} \qquad \begin{array}{l} 0 \le k \le t \\ 0 \le n-k \le d \end{array}$$

Hypergeometric Distribution.

- sampling <u>without</u> replacement.

In terms of white + black marbles

P(k white marbles in a set of n chosen from a container with w white + b black marbles) $= \dfrac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$

Is this a valid PMF?

$$P(X=k) \geq 0$$

$$\sum_k P(X=k) = 1$$

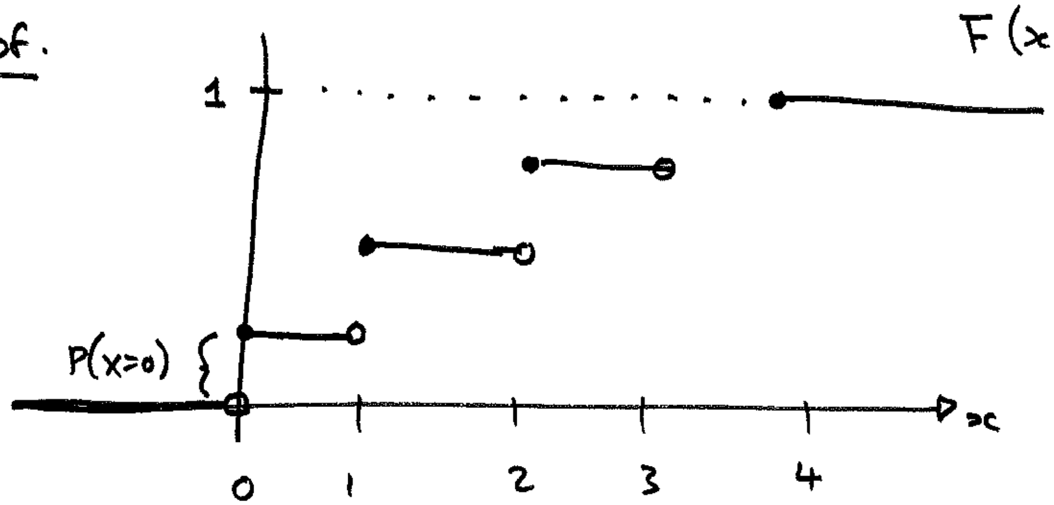
cf sum of Binomials.

$$\sum_{k=0}^{t} \frac{\binom{t}{k}\binom{d}{n-k}}{\binom{t+d}{n}} = \frac{1}{\binom{t+d}{n}} \sum_{k=0}^{t} \binom{t}{n}\binom{d}{n-k}$$

$$= \frac{\binom{t+d}{n}}{\binom{t+d}{n}} = 1$$

<u>CDF.</u>

$F(x) = P(X \leq x)$



jump sizes are given by PMF

jumps are ≥ 0
sum to 1.

$$P(1 < x \leq 3)$$

$$\underbrace{P(X \leq 1)}_{F(1)} + P(1 < x \leq 3) = \underbrace{P(X \leq 3)}_{F(3)}$$

$$P(1 < x \leq 3) = F(3) - F(1)$$

Discrete RV - be careful with $<$, $\leq$

Continuous RV - doesn't matter.

<u>Important properties</u> of CDF.

1/ increasing (not strictly increasing).

2/ right continuous

3/ $F(x) \to 0$ as $x \to -\infty$

$F(x) \to 1$ as $x \to \infty$

a valid CDF satisfies all 3 conditions

Independence of RV.

$$P(X \leq x, Y \leq y) = P(X \leq x)\, P(Y \leq y)$$

for all x, y

event $X \leq x$ event $Y \leq y$.   ← these 2 events are independent.

this definition is also valid for continuous RVs.

(Recall that for discrete RVs

$$P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

$$P(\overbrace{X=k}^{A} \mid \cancel{t}, \cancel{Tet}\ t, \overbrace{d}^{B}, n).$$

actually, we're often interested in estimating d.

→ use Bayes Theorem

get a PMF over d

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}.$$

$$p(d \mid k, t, n) = \frac{P(k \mid t, d, n) \times P(d \mid t, n)}{P(k)}$$