Lecture 5

Statistics

Distribution

Discrete Distributions

What do we need:

$m \longrightarrow$ Size of Ω

$\forall i \in \{1, \dots, m\}$

$p_i = P[\{i\}]$ - prob. of elementary events

$E[X]$

$Var(X)$

Examples:

$m = 2$ $\quad \Omega = \{1, 2\}$ coin flip

$= \{h, t\}$

$p_h \Rightarrow p_t = 1 - p_h$

$h \to 1 \qquad X(h) = 1$

$t \to 0 \qquad X(t) = 0$

$E[X] = 1 \cdot P[\{h\}] + 0 \cdot P[\{t\}]$

$= p_h$

$Var(X) = E[X^2] - E[X]^2$

$= p_h - p_h^2$

$= p_h(1 - p_h)$

0 1 Random variable

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

$E[X^2] = E[X] = P[X=1]$

$\Leftrightarrow 0$

$X^2 = X$

Binomial Distribution

Arg n

$X \Rightarrow$ # of heads in n coin flips (independent)

Let X_i be the 0-1 RV that models a Bernoulli trial (coin flip)

$$X = \sum_{i=1}^{n} X_i$$

$$E[X] = \sum_{i=1}^{n} E[X_i] = n p_h$$

$$E[X^2] = E\left[\sum_{i=1}^{n} X_i \cdot \sum_{j=1}^{n} X_j\right]$$

independence

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} E[X_i X_j]$$

underlying prob. distr:

$P[X = k] = P[\{\underbrace{h, \ldots, h}_{k}, \underbrace{t \ldots t}_{n-k}\} + \text{shuffling}]$

$0 \le k \le n$

$P[\{\underbrace{h, \ldots h}_{k}, \underbrace{t \ldots t}_{n-k}\}] = P[c_1 = h, \ldots, c_k = h, c_{k+1} = t, \ldots, c_n = t]$

$= P[c_1 = h] \ldots P[c_n = t] = p_h^k (1-p_h)^{n-k}$

$$P[X = k] = C_n^k \, p_h^k (1-p_h)^{n-k}$$

$$E[X] \overset{?}{=} \sum_{k=0}^{n} P[X = k] \cdot k$$

$$\vdots$$

$$= n p_h$$

Sampling:

General Definition (Problem)

We have some Data $\boxed{D}$

Ex: D is relation

Problem: Compute $f(D)$ for some function f

$$f : \mathbb{D} \to \mathbb{R}$$

set of possible datasets

f is hard (tedious) to compute on D

Sampling solution:

- Pick D' a random subset of D
- Compute $f(D')$ and use it to estimate $f(D)$

Important question:
What is the probability distribution of random set D'?

Ex:

1° $P[D' = \mathbb{D}] = 1$
$P[D' \neq \mathbb{D}] = 0$
for some $\mathbb{D} \subseteq D$
$\mathbb{D}$ - first 10 elements of D

not very useful

2° $D' = \langle x_1, \dots, x_n \rangle$ for fixed n (given)

$P[X_i = y_k] = \frac{1}{|D|}$

y_k is the k'th element of D

$D = \langle y_1, \dots, y_{|D|} \rangle$

$$P[D'] = P[\langle x_1, \dots, x_n \rangle] = P[X_1] \dots P[X_n] = \left(\frac{1}{|D|}\right)^{|D'|}$$

Sampling with replacement

Sampling without Replacement

Once an element is picked, it is not returned in D (cannot be picked again)

$$P(D') = P(\{D_1\}) \cdot P(D_2 \mid D_1) \cdots P(D_n \mid D_1, \ldots, D_{n-1})$$

$$\boxed{P(A, B) = P(A) \cdot P(B \mid A)}$$

$$= \frac{1}{|D|} \cdot \frac{1}{|D|-1} \cdots \frac{1}{2} \cdot \frac{1}{1}$$

$$= \frac{1}{|D|!}$$

Pick a random arrangement for D and let D' be the first n elements