Statistics for Data Science - 2

Notes - Week 1 and 2

Multiple Random Variables

1. **Joint probability mass function:** Suppose X and Y are discrete random variables defined in the same probability space. Let the range of X and Y be T_X and T_Y , respectively. The joint PMF of X and Y , denoted f_{XY} , is a function from $T_X \times T_Y$ to $[0, 1]$ defined as

$$f_{XY}(t_1, t_2) = P(X = t_1 \text{ and } Y = t_2), t_1 \in T_X, t_2 \in T_Y$$

   - Joint PMF is usually written as table or a matrix.
   - $P(X = t_1 \text{ and } Y = t_2)$ is denoted $P(X = t_1, Y = t_2)$

2. **Marginal PMF:** Suppose X and Y are jointly distributed discrete random variables with joint PMF f_{XY}. The PMF of the individual random variables X and Y are called as marginal PMFs. It can be shown that

$$f_X(t_1) = P(X = t_1) = \sum_{t_2 \in T_Y} (f_{XY}(t_1, t_2))$$

$$f_Y(t_2) = P(X = t_2) = \sum_{t_1 \in T_X} (f_{XY}(t_1, t_2))$$

   Note: Given the joint PMF, the marginal is unique.

3. **Conditional distribution given an event:** Suppose X is a discrete random variable with range T_X , and A is an event in the same probability space. The conditional PMF of X given A is defined as the PMF

$$f_{X|A}(t) = P(X = t|A)$$

   where $t \in T_X$
   We will denote the conditional random variable by $X|A$. (Note that $X|A$ is a valid random variable with PMF $f_{X|A}$).

   - $f_{X|A}(t) = \dfrac{P((X = t) \cap A)}{P(A)}$
   - Range of $(X|A)$ can be different from T_X and will depend on A.

4. **Conditional distribution of one random variable given another:**
Suppose X and Y are jointly distributed discrete random variables with joint PMF f_{XY} . The conditional PMF of Y given $X = t$ is defined as the PMF

$$f_{Y|X=x}(y) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f_{XY}(x, y)}{f_X(x)}$$

We will denote the conditional random variable by $Y|(X = x)$. (Note that $Y|(X = x)$ is a valid random variable with PMF $f_{Y|(X=x)}$.

- Range of $(Y|X = t)$ can be different from T_Y and will depend on t.
- $f_{XY}(x, y) = f_{Y|X=x}(x, y).f_X(x) = f_{X|Y=y}(x, y).f_Y(y)$
- $\sum\limits_{y \in T_Y} f_{Y|X=x}(y) = 1$

5. **Joint PMF of more than two discrete random variables:**
Suppose $X_1, X_2, \ldots, X_n$ are discrete random variables defined in the same probability space. Let the range of X_i be T_{X_i}. The joint PMF of X_i, denoted by $f_{X_1X_2\ldots X_n}$, is a function from $T_{X_1} \times T_{X_2} \times \ldots \times T_{X_n}$ to [0, 1] defined as

$$f_{X_1X_2\ldots X_n}(t_1, t_2, \ldots, t_n) = P(X_1 = t_1, X_2 = t_2, \ldots, X_n = t_n); t_i \in T_{X_i}$$

6. **Marginal PMF in case of more than two discrete random variables:**
Suppose $X_1, X_2, \ldots, X_n$ are jointly distributed discrete random variables with joint PMF $f_{X_1X_2\ldots X_n}$. The PMF of the individual random variables $X_1, X_2, \ldots, X_n$ are called as marginal PMFs. It can be shown that

$$f_{X_1}(t_1) = P(X_1 = t_1) = \sum_{t_2 \in T_{X_2}, t_3 \in T_{X_3}, \ldots, t_n \in T_{X_n}} f_{X_1X_2\ldots X_n}(t_1, t_2, \ldots, t_n)$$

$$f_{X_2}(t_2) = P(X_2 = t_2) = \sum_{t_1 \in T_{X_1}, t_3 \in T_{X_3}, \ldots, t_n \in T_{X_n}} f_{X_1X_2\ldots X_n}(t_1, t_2, \ldots, t_n)$$

$$\vdots$$

$$f_{X_n}(t_n) = P(X_n = t_n) = \sum_{t_1 \in T_{X_1}, t_2 \in T_{X_2}, \ldots, t_{n-1} \in T_{X_{n-1}}} f_{X_1X_2\ldots X_n}(t_1, t_2, \ldots, t_n)$$

7. **Marginalisation:** Suppose $X_1, X_2, \ldots, X_n$ are jointly distributed discrete random variables with joint PMF $f_{X_1X_2\ldots X_n}$. The joint PMF of the random variables $X_{i_1}, X_{i_2}, \ldots X_{i_k}$, denoted by $f_{X_{i_1}X_{i_2}\ldots X_{i_k}}$ is given by

$$f_{X_{i_1}X_{i_2}\ldots X_{i_k}}(t_{i_1}, t_{i_2}, \ldots t_{i_k}) = \sum f_{X_1X_2\ldots X_n}(t_1, \ldots t_{i_1-1}, t_{i_1}, t_{i_1+1}, \ldots t_{i_k-1}, t_{i_k}, t_{i_k+1} \ldots, t_n)$$

- Sum over everything you don't want.

8. **Conditioning with multiple discrete random variables:**
   - A wide variety of conditioning is possible when there are many random variables. Some examples are:
   - Suppose $X_1, X_2, X_3, X_4 \sim f_{X_1X_2X_3X_4}$ and $x_i \in T_{X_i}$, then
     - $f_{X_1|X_2=x_2}(x_1) = \dfrac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}$
     - $f_{X_1,X_2|X_3=x_3}(x_1, x_2) = \dfrac{f_{X_1X_2X_3}(x_1, x_2, x_3)}{f_{X_3}(x_3)}$
     - $f_{X_1|X_2=x_2,X_3=x_3}(x_1) = \dfrac{f_{X_1X_2X_3}(x_1, x_2, x_3)}{f_{X_2X_3}(x_2, x_3)}$
     - $f_{X_1X_4|X_2=x_2,X_3=x_3}(x_1, x_4) = \dfrac{f_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4)}{f_{X_2X_3}(x_2, x_3)}$

9. **Conditioning and factors of the joint PMF:**
   Let $X_1, X_2, X_3, X_4 \sim f_{X_1X_2X_3X_4}, X_i \in T_{X_i}$.

   $$\begin{aligned} f_{X_1X_2X_3X_4}(t_1, t_2, t_3, t_4) =& P(X_1 = t_1 \text{ and } (X_2 = t_2, X_3 = t_3, X_4 = t_4)) \\ =& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1)P(X_2 = t_2 \text{ and } (X_3 = t_3, X_4 = t_4)) \\ =& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1)f_{X_2|X_3=t_3,X_4=t_4}(t_2)P(X_3 = t_3 \text{ and } X_4 = t_4) \\ =& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1)f_{X_2|X_3=t_3,X_4=t_4}(t_2)f_{X_3|X_4=t_4}(t_3)f_{X_4}(t_4). \end{aligned}$$

   - Factoring can be done in any sequence.

10. **Independence of two random variables:**
    Let X and Y be two random variables defined in a probability space with ranges T_X and T_Y , respectively. X and Y are said to be independent if any event defined using X alone is independent of any event defined using Y alone. Equivalently, if the joint PMF of X and Y is f_{XY}, X and Y are independent if

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

    for $x \in T_X$ and $y \in T_Y$

    - X and Y are independent if

      $$f_{X|Y=y}(x) = f_X(x)$$
      $$f_{Y|X=x}(y) = f_Y(y)$$

      for $x \in T_X$ and $y \in T_Y$

    - To show X and Y independent, verify

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

      for **all** $x \in T_X$ and $y \in T_Y$

- To show X and Y dependent, verify

$$f_{XY}(x,y) \neq f_X(x)f_Y(y)$$

  for **some** $x \in T_X$ and $y \in T_Y$
  - **Special case:** $f_{XY}(t_1,t_2) = 0$ when $f_X(t_1) \neq 0, f_Y(t_2) \neq 0$.

11. **Independence of multiple random variables:**
Let $X_1, X_2, \ldots, X_n$ be random variables defined in a probability space with range of X_i denoted T_{X_i}. $X_1, X_2, \ldots, X_n$ are said to be independent if events defined using different X_i are mutually independent. Equivalently, $X_1, X_2, \ldots, X_n$ are independent iff

$$f_{X_1X_2\ldots X_n}(t_1,t_2,\ldots,t_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

for all $x_i \in T_{X_i}$

   - All subsets of independent random variables are independent.

12. **Independent and Identically Distributed (i.i.d.) random variables:**
Random variables $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributed (i.i.d.), if
(i) they are independent.
(ii) the marginal PMFs f_{X_i} are identical.
Examples:

   - Repeated trials of an experiment creates i.i.d. sequence of random variables
     - Toss a coin multiple times.
     - Throw a die multiple times.
   - Let $X_1, X_2, \ldots X_n \sim$ i.i.d.X (Geometric(p)).
     X will take values in $\{1, 2, \ldots\}$
     $P(X = k) = p^{k-1}p$

     Since X_i's are independent and identically distributed, we can write

$$\begin{aligned} P(X_1 > j, X_2 > j, \ldots, X_n > j) =& P(X_1 > j)P(X_2 > j)\ldots P(X_n > j) \\ =& [P(X > j)]^n \end{aligned}$$

$$\begin{aligned} P(X > j) =& \sum_{k=j+1}^{\infty} (1-p)^{k-1}p \\ =& (1-p)^j p + (1-p)^{j+1}p + (1-p)^{j+2}p + \ldots \\ =& (1-p)^j p[1 + (1-p) + (1-p)^2 + \ldots] \\ =& (1-p)^j p \left(\frac{1}{1-(1-p)}\right) \\ =& (1-p)^j \end{aligned}$$

$\Rightarrow P(X_1 > j, X_2 > j, \ldots, X_n > j) = [P(X > j)]^n = (1-p)^{jn}$

13. **Functions of a random variable:** X : random variable with PMF $f_X(t)$.
$f(X)$: random variable whose PMF is given as follows.

$$\begin{aligned} f_{f(X)}(a) &= P(f(X) = a) = P(X \in \{t : f(t) = a\}) \\ &= \sum_{t:f(t)=a} f_X(t) \end{aligned}$$

- PMF of $f(X)$ can be found using PMF of X.

14. **Functions of multiple random variables ($g(X_1, X_2, \ldots, X_n)$):**

Suppose $X_1, X_2, \ldots, X_n$ have joint PMF $f_{X_1 X_2 \ldots X_n}$ with T_{X_i} denoting the range of X_i. Let $g : T_{X_1} \times T_{X_2} \times \ldots \times T_{X_n} \rightarrow R$ be a function with range T_g . The PMF of $X = g(X_1, X_2 ..., X_n)$ is given by

$$f_X(t) = P(g(X_1, X_2 ..., X_n) = t) = \sum_{(t_1,\ldots,t_n):g(X_1,X_2...,X_n)=t} f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n)$$

- **Sum of two random variables taking integer values:**
  $X, Y \sim f_{XY}, Z = X + Y$.
  Let z be some integer,

$$\begin{aligned} P(Z = z) =& P(X + Y = z) \\ &= \sum_{x=-\infty}^{\infty} P(X = x, Y = z - x) \\ &= \sum_{x=-\infty}^{\infty} f_{XY}(x, z - x) \\ &= \sum_{y=-\infty}^{\infty} f_{XY}(z - y, y) \end{aligned}$$

- **Convolution:** If X and Y are independent, $f_{X+Y}(z) = \sum\limits_{x=-\infty}^{\infty} f_X(x) f_Y(z - x)$

- Let $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$
  - X and Y are independent.
  - $Z = X + Y$, $z \in \{0, 1, 2, \ldots\}$

  $f_Z(z) \sim \text{Poisson}(\lambda_1 + \lambda_2)$
  $(X = k \mid Z = n) \sim \text{Binomial}\left(n, \dfrac{\lambda_1}{\lambda_1 + \lambda_2}\right), (Y = k \mid Z = n) \sim \text{Binomial}\left(n, \dfrac{\lambda_2}{\lambda_1 + \lambda_2}\right)$

15. **CDF of a random variable:**
Cumulative distribution function of a random variable X is a function $F_X : R \rightarrow [0, 1]$ defined as

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

16. **Minimum of two random variables:**
Let $X, Y \sim f_{XY}$ and let $Z = \min\{X, Y\}$, then

   - $$\begin{aligned} f_Z(z) = P(Z = z) &= P(\min\{X, Y\} = z) \\ &= P(X = z, Y = z) + P(X = z, Y > z) + P(X > z, Y = z) \\ &= f_{XY}(z, z) + \sum_{t_2 > z} f_{XY}(z, t_2) + \sum_{t_1 > z} f_{XY}(t_1, z) \end{aligned}$$

   - $$\begin{aligned} F_Z(z) = P(Z \le z) &= P(\min\{X, Y\} \le z) \\ &= 1 - P(\min\{X, Y\} > z) \\ &= 1 - [P(X > z, Y > z)] \end{aligned}$$

17. **Maximum of two random variables:**
Let $X, Y \sim f_{XY}$ and let $Z = \max\{X, Y\}$, then

   - $$\begin{aligned} f_Z(z) = P(Z = z) &= P(\max\{X, Y\} = z) \\ &= P(X = z, Y = z) + P(X = z, Y < z) + P(X < z, Y = z) \\ &= f_{XY}(z, z) + \sum_{t_2 < z} f_{XY}(z, t_2) + \sum_{t_1 < z} f_{XY}(t_1, z) \end{aligned}$$

   - $$\begin{aligned} F_Z(z) = P(Z \le z) &= P(\max\{X, Y\} \le z) \\ &= [P(X \le z, Y \le z)] \end{aligned}$$

18. **Maximum and Minimum of n i.i.d. random variables**

   - Let $X \sim \text{Geometric}(p), Y \sim \text{Geometric}(q)$
     X and Y are independent.
     $Z = \min(X, Y)$
     $$Z \sim \text{Geometric}(1 - (1-p)(1-q))$$
   - Maximum of 2 **independent** geometric random variables is not geometric.

Important Points:

1. Let $N \sim \text{Poisson}(\lambda)$ and $X|N = n \sim \text{Binomial}(n, p)$, then $X \sim \text{Poisson}(\lambda p)$

2. Memory less property of Geometric(p)
   If $X \sim$ Geometric(p), then

$$P(X > m+n | X > m) = P(X > n)$$

3. Sum of n **independent** Bernoulli(p) trials is Binomial(n, p).

4. Sum of 2 **independent** Uniform random variables is not Uniform.

5. Sum of **independent** Binomial(n, p) and Binomial(m, p) is Binomial($n + m, p$).

6. Sum of r **i.i.d.** Geometric(p) is Negative-Binomial(r, p).

7. Sum of **independent** Negative-Binomial(r, p) and Negative-Binomial(s, p) is Negative-Binomial($r + s, p$)

8. If X and Y are independent, then $g(X)$ and $h(Y)$ are also independent.