Probability for Data Science - Formula Sheet 1

tbimbato

Combinatorial Analysis

- Unordered
  $nCr = \binom{n}{r} = \frac{n!}{r!(n-r)!}$
  (with repeats):
  $\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$
- Ordered
  $nPr = P(n, r) = \frac{n!}{(n-r)!}$
  (with repeats): n^r

Set Operations and Probability Rules

- $P(A \cap B) = P(B|A) \cdot P(A)$
  If independent:
  $P(A \cap B) = P(A) \cdot P(B)$
  $P(A \cap B \cap C) = P(A)P(B)P(C)$
  For three independent events.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
- $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$
- **Complement Rules:** $P(A^C) = 1 - P(A)$
  $P(A^C \cap B^C) = P(A \cup B)^C = 1 - P(A \cup B)$
- **Bayes' TH:** $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$
  If multiple hypotheses:
  $P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$ with $k = 1, 2, \dots, n$
- **Law of Total Probability:**
  $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$
  $P(A) = P(A \cap B) + P(A \cap B^c)$

Discrete Random Variables

- $P(X = x) = p_x(x)$
- $P(X < x) = \sum_{k<x} P(X = k)$
- $\mathbb{E}(X) = \sum_i x_i \cdot p_i$ (weighted average)
- $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sum_i (x_i - \mathbb{E}(X))^2 \cdot p_i$
- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$
- If $X_1..X_n$ are indipendent:
  $$\text{E}(X_1 + ..X_n) = \sum_{i=1}^{n} E(X_i)$$
  $$\text{Var}(X_1 + ..X_n) = \sum_{i=1}^{n} \text{Var}(X_i)$$
- $J = X + Y \rightarrow P(J = k) = \sum_x P(X = x)P(Y = k - x)$

Continuous Random Variables

- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(x_t)dx_t$
  $P(X < x) = P(X \leq x)$ for all Continuous RV
- $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$
- $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mathbb{E}(X)^2$
- $P(X \in [a, b]) = \int_a^b f(x)dx$
- if J = X + Y $\rightarrow f_J(j) = \int_{-\infty}^{\infty} f_X(x)f_Y(j - x)dx$
- $f_X(x) \geq 0$ and $P(X = x) = 0$ $\forall$ Continuous RVs
- Convolution where J = X + Y $f_J(j) = \int_{-\infty}^{\infty} f_X(x)f_Y(j - x)dx$

Generic Joint Distributions

- $P(X = x, Y = y) = p_{X,Y}(x, y)$
- $[f \approx P]$ $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$
- $\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y])$
- $\text{Cov}(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$
- $\text{Var}[X, Y] = \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$
- **ind:** $f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \forall x, y$
- correlation k: $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$

Discrete Joint Distributions

- **Marginal PMF:** $P_X(x) = \sum_y p_{X,Y}(x, y)$
- **Conditional PMF:** $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$
- **Expectation:** $\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y)p_{X,Y}(x, y)$
- **Mean:** $\mathbb{E}[X] = \sum_x \sum_y x \cdot p_{X,Y}(x, y)$
- **Conditional E:** $\mathbb{E}[X \mid Y = y] = \sum_x xP_{X|Y}(x|y)$
- **Iterated Expectations:** $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$
- **Joint Expectation:** $\mathbb{E}[XY] = \sum_x \sum_y xyp_{X,Y}(x, y)$
- **Covariance:** $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- **Independence:** $P_{X,Y}(x, y) = P_X(x)P_Y(y) \quad \forall x, y$
- **MGF:** $M_{X,Y}(t_1, t_2) = \sum_x \sum_y e^{t_1 x + t_2 y} p_{X,Y}(x, y)$
- **Variance Formula:** $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Continuous Joint Distributions

- **Marginal:** $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$
- **Conditional:** $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- $\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$
- $\mathbb{E}[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf_{X,Y}(x, y)dxdy$
- **Conditional E:** $\mathbb{E}[X|Y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx$
- **Iterated Expectations:** $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$
  Where: $\mathbb{E}[X \mid Y = y] = \int xf_{X|Y}(x|y)dx$
- **Joint E:** $\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_{X,Y}(x, y)dxdy$
- **Covariance:** $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, the PDF of $J = X + Y$ is: $f_J(j) = \int_{-\infty}^{\infty} f_X(x)f_Y(j - x)dx$

LLN and CLT - Practical Guide

Assumptions: i.i.d. RVs: $X_1, X_2, \dots, X_n$, with $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$.

- **LLN**: Sample Mean $= (\frac{1}{n} \sum_{i=1}^{n} X_i) \rightarrow \mu$ as $n \rightarrow \infty$
- **CLT**:
  $$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$$
  if $S_n = X_1 + X_2 + \cdots + X_n \rightarrow S_n \approx \mathcal{N}(n\mu, n\sigma^2)$
  If mean $\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- $P(a \leq Sum \leq b)$:
  $$P(a \leq S_n \leq b) \approx P\left(\frac{a - n\mu}{\sigma\sqrt{n}} \leq Z \leq \frac{b - n\mu}{\sigma\sqrt{n}}\right)$$
  $\mathbb{E}[S_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$, if equal $= n\mathbb{E}[X]$
  $\sigma_{S_n} = \sqrt{Var(X_1) + Var(X_2) \dots}$, if equal $\sqrt{n \cdot \text{Var}(X)}$
- **Heuristics**:
  - **Bin**: $np \geq 5, \quad n(1 - p) \geq 5$ $(\rightarrow \quad \mathcal{N}(np, np(1 - p)))$.
  - **Poisson**: $\lambda > 10$ $(\rightarrow \quad \mathcal{N}(\lambda, \lambda))$.
  - **Sum of RVs**: $n \geq 30$.

Remember!

$\bigcap$ $\approx$Exclusive AND $\wedge$	$\bigcup$ $\approx$Inclusive OR $\vee$