Lecture 12

giovedì 30 ottobre 2025 17:28

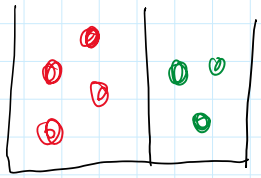
MOODLE — DATA SCIENCE

EXAM — THURSDAY 18 DECEMBER

↓

for DS 16:30 - 19:30 ROOM P300 and LUF1

HYPERGEOMETRIC RANDOM VARIABLES



URN

D RED BALLS

N-D GREEN BALLS

DRAW n balls

- WITH REPLACEMENT

X = # of red balls

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

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball is red} \\ 0 & \text{if } i\text{-th ball is green} \end{cases}$$

$X_i \sim Ber(p)$ $\quad p = P(X_i = 1) = \frac{D}{N}$

$X_i \perp\!\!\!\perp X_j$ for $i \neq j$ (independent)

$X \sim Bin(n, p)$ BINOMIAL DISTRIBUTION

- WITHOUT REPLACEMENT

Y = # of RED balls in n draws without replacement

$Y \sim Hp(n, N, D)$ HYPERGEOMETRIC DISTRIBUTION

density of Y, $k = 0, \dots, n$

$$p_Y(k) = P(X = k) = \frac{\text{favorable cases}}{\text{possible cases}} = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

k red $\quad n-k$ green

n balls

Expectation of Y? $\quad$ Var(Y)?

Consider $Y = Y_1 + Y_2 + \dots + Y_n$

Consider $Y = Y_1 + Y_2 + \dots + Y_n$

Y_i random variable $\quad Y_i = \begin{cases} 1 & \text{if } i\text{-th ball is RED} \\ 0 & \text{if } i\text{-th ball is green} \end{cases}$

$Y_i \sim Ber(p_i)$

$p_i = P(Y_i = 1) = P(i\text{-th ball is RED - drawn WITHOUT replacement})$

It can be proved that $p_i = \frac{D}{N} \quad \forall i = 1, \dots, n$

$Y_i \sim Ber\left(\frac{D}{N}\right)$

but $Y_i \not\perp Y_j$ for $i \neq j$ (NOT independent)

We can compute $E[Y] = n\frac{D}{N}$

$E[Y] = E[Y_1] + E[Y_2] + \dots + E[Y_n] = n \cdot \frac{D}{N}$

$Var(Y) = ?$

$Var(Y) = Var(Y_1 + \dots + Y_n)$

$= Var(Y_1) + \dots + Var(Y_n) + 2Cov(Y_1, Y_2) + 2Cov(Y_1, Y_3)$

$+ \dots Cov(Y_1, Y_n) + \dots + Cov(Y_{n-1}, Y_n) = \dots$

$$Var(Y) = n\frac{D(N-D)}{N^2}\frac{N-n}{N-1} = Var(X)\frac{N-n}{N-1}$$

recall $Var(X) = n\,Var(X_1) = n\frac{D(N-D)}{N^2}$

if $X \sim Bin\left(n, \frac{D}{N}\right)$

Note $\frac{N-n}{N-1} < 1 \Rightarrow Var(Y) < Var(X)$

POISSON DISTRIBUTION

$X_n \sim Bin(n, p)$

in general, it is difficult to compute density $p_X(k)$ when n is large

Consider $X \sim Poi(\lambda)$ will be a possible approximation for $Bin(n, p)$ when n is large, but np is bounded

$\boxed{\lambda = np}$

$X : \Omega \to \mathbb{N}$ POISSON r.v. parameter λ

$X \sim \text{Poi}(\lambda)$ $\quad p_X(n) = e^{-\lambda} \frac{\lambda^n}{n!} \quad \forall n \in \mathbb{N} = \{0, 1, \dots\}$

Recall $e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$

Hence $\sum_n p_X(n) = 1$

$n! = n(n-1)(n-2)\cdots 1 = n(n-1)!$ ($= (n-1)!$)

$$E[X] = \sum_{n=0}^{\infty} n\, p_X(n) = \sum_{n=0}^{\infty} n\, e^{-\lambda} \frac{\lambda^n}{n(n-1)!}$$

$$= \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda \cdot \lambda^{n-1}}{(n-1)!} = \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} \qquad k = n-1$$

$$= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

$E[X] = \lambda$

$\text{Var}(X) = \lambda$ $\qquad E[X^2] = \lambda^2 + \lambda$

$= \sum_{n=1}^{\infty} n^2 P(X = n)$

Thm LAW OF SMALL NUMBERS

Let $X_1, \dots, X_n$ independent such that $X_i \sim \text{Ber}(p_i)$

Set $S_n = X_1 + \dots + X_n$ and let $W_n \sim \text{Poi}(p_1 + \dots + p_n)$

Then, for any $A \subseteq \mathbb{N}$

$$\left| P(S_n \in A) - P(W_n \in A) \right| \le \sum_{i=1}^{n} p_i^2$$

↳ distance between distribution of S_n and W_n

Corollary

In case $p_i = \frac{\lambda}{n}$ for any $i = 1, \dots, n$

$S_n \sim \text{Bin}(n, \frac{\lambda}{n})$ $\qquad W_n \sim \text{Poi}(\lambda)$

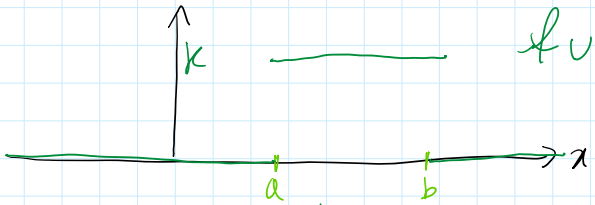
Then

$$\left| P(S_n \in A) - P(W_n \in A) \right| \le \sum_{i=1}^{n} \left(\frac{\lambda}{n}\right)^2 = n \frac{\lambda^2}{n^2} = \frac{\lambda^2}{n}$$

ABSOLUTELY CONTINUOUS DISTRIBUTION

- UNIFORM DISTRIBUTION

$U \sim \mathcal{U}(a,b) \qquad -\infty < a < b < \infty$



$$f_U(x) = \begin{cases} k & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{-\infty}^{+\infty} f_U(x) = 1 = \int_a^b k\,dx = 1 = k(b-a) = 1 \iff k = \frac{1}{b-a}$$

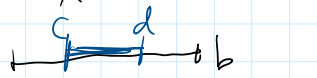
$$f_U(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

$U \sim \mathcal{U}(a,b)$

$$f_U(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x) \qquad P(a \le U \le b) = 1$$

- distribution $P(U \in A) = \int_A f_U(x)\,dx$

if $a < c < d < b$



$$P(U \in [c,d]) = P(c \le U \le d) = \int_c^d f_U(x)\,dx$$

$$= \int_c^d \frac{1}{b-a}\,dx = \frac{d-c}{b-a} = P(c \le U \le d)$$

$$E[U] = \int_{-\infty}^{+\infty} x\, f_U(x)\,dx = \int_a^b x \frac{1}{b-a}\,dx$$

$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_{x=a}^{x=b} = \frac{b^2-a^2}{2} \cdot \frac{1}{b-a} = \frac{(b-a)(b+a)}{2}\frac{1}{b-a}$$

$$E[U] = \frac{a+b}{2} = \text{middle point}$$

EXPONENTIAL r.v. $X \sim Exp(\lambda)$

$$F_X(x) = P(X \le x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

$E[X] = \frac{1}{\lambda} \qquad V(X) = \frac{1}{\lambda^2}$

$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$

$E[X] = \frac{1}{\lambda}$ $\quad$ $Var(X) = \frac{1}{\lambda^2}$

- MINIMUM and MAXIMUM of r.v.

Consider X, Y r.v. $\quad X \perp\!\!\!\perp Y$

What is distribution of

$Z = \max(X,Y)$ $\quad$ $W = \min(X,Y)$?

$[Z(\omega) = \max(X(\omega), Y(\omega))]$

We compute distribution function of Z and W

- $P(Z \le z) = F_Z(z) = P(\max(X,Y) \le z)$

$\{\max(X,Y) \le z\} = \{X \le z \text{ and } Y \le z\}$

$= P(X \le z, Y \le z)$ $\quad$ $= \{X \le z\} \cap \{Y \le z\}$

$\overset{\text{INDEPENDENCE}}{=} P(X \le z) P(Y \le z) = F_X(z) F_Y(z)$

$\Rightarrow F_{MAX(X,Y)}(z) = F_X(z) F_Y(z)$

- if X and Y are absolutely continuous, the density of $Z = MAX(X,Y)$ is $\quad [\frac{d}{dz} F_X(x) = f_X(x)]$

$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz}\left(F_X(z) \cdot F_Y(z)\right)$

- $P(W > w) = 1 - F_W(w) = P(MIN(X,Y) > w)$

$= P(X > w, Y > w) \overset{\text{INDEP.}}{=} P(X > w) P(Y > w)$

$F_{MIN(X,Y)} = 1 - (1 - F_X(w))(1 - F_Y(w))$

Example

Let $X \sim Exp(\lambda)$, $Y \sim Exp(\mu)$, $X \perp\!\!\!\perp Y$

What is distribution of $MIN(X,Y)$

$\Rightarrow MIN(X,Y) \sim Exp(\lambda + \mu)$

$P(X > x) = \begin{cases} e^{-\lambda x} & \text{if } x \ge 0 \end{cases}$

$$P(X>x) = \begin{cases} e^{-\lambda x} & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \end{cases}$$

- $P(MIN(X,Y) > z) = P(X>z)\,P(Y>z)$ if $z \geq 0$
$= e^{-\lambda z} \cdot e^{-\mu z} = e^{-(\lambda+\mu)z}$
- $P(MIN(X,Y) > z) = 1$ if $z < 0$

$\Rightarrow MIN(X,Y) \sim Exp(\lambda+\mu)$

because we computed the distribution function

Thm A random variable $T : \Omega \to [0,+\infty)$ has exponential distribution if and only if it has the following MEMORYLESS PROPERTY

(*) $P(T > s+t \mid T > s) = P(T > t)$ $\forall\, s,t > 0$

(and it holds $P(T>t) = e^{-\lambda t}$)

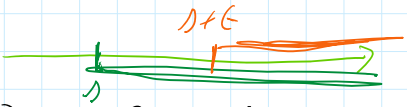
Proof

($\Rightarrow$) $T \sim Exp(\lambda)$. We show that (*) holds.

$P(T>t) = e^{-\lambda t}$

$$P(T > s+t \mid T > s) = \frac{P(T>s+t,\ T>s)}{P(T>s)}$$

$\{T > s+t\} \subseteq \{T > s\}$



$T > s+t > s \Rightarrow \omega \in \{T > s+t\} \Rightarrow \omega \in \{T > s\}$

then $\{T>s+t\} \cap \{T>s\} = \{T>s+t\}$

$$= \frac{P(T>s+t)}{P(T>s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

($\Leftarrow$) Given T r.v. satisfying (*),
we have to show that $T \sim Exp(\lambda)$, that is,
there exists λ such that $P(T>t) = e^{-\lambda t}$

there exists λ such that $P(T>\epsilon) = e^{-\lambda\epsilon}$

$$P(T>s+\epsilon \mid T>s) = \frac{P(T>s+\epsilon, T>s)}{P(T>s)} = \frac{P(T>s+\epsilon)}{P(T>s)}$$

$$\overset{\text{(*)}}{=} P(T>\epsilon)$$

If we call $g(\epsilon) = P(T>\epsilon)$, function of ϵ, we derive the equality

$$g(s+\epsilon) = g(\epsilon)\, g(s)$$

CAUCHY FUNCTIONAL INEQUALITY

If g is continuous, the unique solutions to this equation are the exponential ▨