PROBABILITY THEORY

4° LECTURE - 04/12/2018

ORDER STATISTICS: Let $X_1, \dots, X_n$ be iid RV

Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ the same r.v. placed in ascending order, that is.

$$X_{(1)} = \min\{X_1 \dots X_n\} \qquad 1° \text{ smallest}$$

$$X_{(2)} = \min\{X_1 \dots X_n\} \setminus \{X_{(1)}\} \qquad 2° \text{ smallest}$$

$$\vdots$$

$$X_{(n)} = \max\{X_1, \dots, X_n\} \qquad n° \text{ smallest (or 1° biggest)}$$

ORDER STATISTICS

Hence $\mathbb{P}(X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}) = 1$ (and if they are continuous, also $\mathbb{P}(X_{(1)} < \dots < X_{(n)}) = 1$)

Remark: the r.v. $X_{(1)}, \dots, X_{(n)}$ are now strongly correlated. (indeed, if $\mathbb{P}(X_{(n)} \leq a) = 1 \Rightarrow \mathbb{P}(X_{(k)} \leq a \ \forall k) = 1$)

We want to determine their joint (and marginal) distribution

Application: For example we are interested in measuring the variability of data $\{X_1, \dots, X_n\}$, for iid X_k's, that is the sample range: $X_{(n)} - X_{(1)}$.

Simple cases:

- $\mathbb{P}(X_{(n)} \leq t) = \mathbb{P}(X_1 \leq t, \dots, X_n \leq t) = \prod_{k=1}^{n} \mathbb{P}(X_k \leq t) = F(t)^n$
- $\mathbb{P}(X_{(1)} \leq t) = 1 - \mathbb{P}(X_{(1)} > t) = 1 - \mathbb{P}(X_1 > t, \dots, X_n > t) = 1 - (1 - F(t))^n$

Thus we show that

$$F_{(n)}(t) = \mathbb{P}(X_{(n)} \leq t) = F(t)^n$$

$$F_{(1)}(t) = \mathbb{P}(X_{(1)} \leq t) = 1 - (1-F(t))^n$$

Similarly

$$F_{(n-1)}(t) = \underbrace{F(t)^n}_{\text{all } X_i\text{'s are } \leq t} + \underbrace{n F(t)^{n-1}(1-F(t))}_{\text{all but one of the } X_i\text{'s are } \leq t}$$

then it holds

<u>Theorem 1</u> For $X_1, \dots X_n$ iid with distr. F, and letting $F_{(k)}$ the distr. of $X_{(k)}$, it holds

$$F_{(k)}(t) = \sum_{i=k}^{n} \binom{n}{i} F(t)^i (1-F(t))^{n-i} \qquad \forall t \in \mathbb{R}, \ \forall k \in \{1, \dots n\}$$

If moreover X_j's are absolutely continuous with density f, then also the order statistics are absol. continuous with density

$$f_{(k)}(t) = n \binom{n-1}{k-1} F(t)^{k-1} (1-F(t))^{n-k} \cdot f(t), \quad t \in \mathbb{R}, \ \forall k \in \{1, \dots, n\}$$

↳ Explicit formula also for discrete r.v., but more involved

<u>Examples</u>:

1] Let $X_1, \dots, X_n$ iid $\sim \mathcal{U}[0,1]$.

Then by theorem 1, also recalling that

$$f(x) = \mathbb{1}_{[0,1]}(x) \quad \text{and} \quad F(x) = x \mathbb{1}_{[0,1]}(x) + \mathbb{1}_{(1,+\infty)}(x),$$

we get

$$f_{(k)}(x) = n\binom{n-1}{k-1} \cdot x^{k-1}(1-x)^{n-k} \cdot \mathbb{1}_{[0,1]}(x)$$

costant

that is, $X_{(k)} \sim Beta(k, n-k+1)$

Recall: For all $\alpha, \beta > 0$
$X \sim Beta(\alpha, \beta)$ s.t.

$$f_X(x) = x^{\alpha-1}(1-x)^{\beta-1} \cdot c \; \mathbb{1}_{[0,1]}(x)$$

where $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

$E(X) = \alpha/\alpha+\beta$

By the exchangeability of $X_1, \dots, X_n$, and assuming that they are absolutely continuous ~~so~~ that $\mathbb{P}(X_1 < \dots < X_n) = \frac{1}{n!}$, one can show that the joint density of $X_{(1)}, \dots, X_{(n)}$, denoted by f_{OS}, is

Theorem 2: $$f_{OS}(x_1, \dots, x_n) = n! \underbrace{f(x_1) \cdots f(x_n)}_{= f_{X_1,\dots,X_n}(x_1,\dots,x_n)} \mathbb{1}(x_1 < x_2 < \dots < x_n)$$

Starting from theorem 2, and with some computations, one can express the joint density of any couple of order statistic $X_{(l)}, X_{(m)}$. If $l < m$, then

(*)

$$f_{X_{(l)},X_{(m)}}(x_l, x_m) = \frac{n!}{(l-1)!(m-l-1)!(n-m)!} F(x_l)^{l-1} \cdot [F(x_m) - F(x_l)]^{m-l-1} \cdot (1-F(x_m))^{n-m} \cdot f(x_m) \cdot f(x_l)$$

Exercises:

1) Let $X_1, \dots, X_n$ iid $\sim Exp(\lambda)$ with $f(x) = \lambda e^{-\lambda x} \mathbb{1}(x>0)$

Then $$f_{OS}(x_1, \dots, x_n) = n! \lambda^n e^{-\lambda \sum_{j=1}^n x_j} \cdot \mathbb{1}(0 < x_1 < x_2 < \dots < x_n)$$

Let $Y_1 = X_{(1)}$, $Y_2 := X_{(2)} - X_{(1)}$ ... $Y_n := X_{(n)} - X_{(n-1)}$

We want to compute the joint density of $Y = (Y_1, \dots, Y_n)$.

Notice that $Y = (Y_1, \dots, Y_n) = g(X_{(1)}, \dots, X_{(n)})$

with $g(x_1, \dots, x_n) = (x_1, x_2 - x_1, x_3 - x_2, \dots, x_n - x_{n-1})$.

thus $g^{-1}(y_1, \dots, y_n) = (y_1, y_1 + y_2, \dots, y_1 + y_2 + \dots y_n)$

and $J_{g^{-1}}(y_1 \dots y_n) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 1 \end{pmatrix}$ with $|\det J_{g^{-1}}| = 1$

$$\Rightarrow f_Y(y_1, \dots y_n) = f_{OS}(y_1, y_1 + y_2, \dots, y_1 + \dots + y_n) = n! \, \lambda^n e^{-\lambda \sum_{j=1}^{n} (n-j-1) y_j} \cdot \prod_{j=1}^{n} \mathbb{1}(y_j > 0)$$

$\Longrightarrow Y_1, \dots, Y_n$ are independent s.t

$Y_j \sim Exp((n-j-1)\lambda)$ $\#$

---

2) Range of uniform r.v.

Let $X_1, \dots, X_n$ iid $\sim U[0,1]$

We want to determine the distribution of $X_{(n)} - X_{(1)}$.

Using $\#$, we get

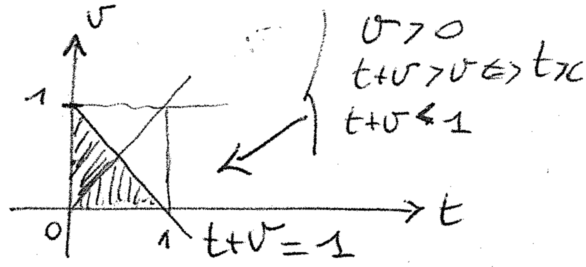
$$f_{X_{(1)}, X_{(n)}}(x_1, x_n) = n(n-1)(x_n - x_1)^{n-2} \mathbb{1}(0 < x_1 < x_n < 1)$$

Let $T = X_{(n)} - X_{(1)}$ and $V = X_{(1)}$ so that

$(T, V) = g(X_{(1)}, X_{(n)})$ with $g^{-1}(t, v) = (v, t + v)$, $J_{g^{-1}}(t, v) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$

$$|\det J_{g^{-1}}| = 1 \Longrightarrow f_{T,V}(t, v) = n(n-1)(t + v - v)^{n-2} \cdot \mathbb{1}(0 < v < t + v < 1)$$
$$= n(n-1) t^{n-2} \mathbb{1}(0 < v < t + v < 1)$$

$$f_T(t) = \int_0^{1-t} n(n-1)t^{n-2}\,dv \overset{\mathbb{1}_{[0<t<1)}}{=} \underline{n(n-1)t^{n-2}(1-t)}\cdot \mathbb{1}_{[0<t<1)}$$



$\Rightarrow T \sim Beta(n-1, 2)$

---

3) Let X_1, X_2, X_3 iid $\sim \mathcal{U}[0,1]$.

Compute the marginal densities of the order statistics, their averages, and $\mathbb{P}(X_{(1)} \leq \frac{1}{2}, X_{(2)} > \frac{3}{4})$.

Sol: Recall that $F(x) = x\,\mathbb{1}_{[0,1)}(x) + \mathbb{1}_{[1,+\infty)}(x)$

and $f(x) = \mathbb{1}_{[0,1]}(x)$, for X_1, X_2, X_3.

Using theorem 1, we get

$f_{(1)}(x) = 3(1-x)^2\,\mathbb{1}_{[0,1]}(x) \Rightarrow X_{(1)} \sim Beta(1,3)$

$f_{(2)}(x) = 6x(1-x)\,\mathbb{1}_{[0,1]}(x) \Rightarrow X_{(2)} \sim Beta(2,2)$

$f_{(3)}(x) = 3x^2\,\mathbb{1}_{[0,1]}(x) \Rightarrow X_{(3)} \sim Beta(3,1)$

$$\mathbb{E}(X_{(1)}) = 3\int_0^1 \underbrace{x(1-x)^2}_{\substack{\text{inverse normalization}\\ \text{of Beta}(2,3)}}\,dx = \frac{\Gamma(2)\Gamma(3)}{\Gamma(5)}\cdot 3 = \frac{1\cdot 2\cdot 3}{4!} = \frac{1}{4} \left(= \frac{\alpha}{\alpha+\beta}\right)$$

$\mathbb{E}(X_{(2)}) = \frac{2}{4} = \frac{1}{2}$ and $\mathbb{E}(X_{(3)}) = \frac{3}{4}$

From formula $\star$, it can be shown, for iid $\sim \mathcal{U}[0,1]$,

$$\mathbb{P}(X_{(k)} \leq x, X_{(k+1)} > y) = \int_0^x dx_k \int_y^1 dx_{k+1} \cdot f_{X_{(k)}X_{(k+1)}}(x_k, x_{k+1}) = \binom{n}{k} x^k (1-y)^{n-k}$$

Then we get

$$\mathbb{P}\left(X_{(1)} \leq \frac{1}{2}, X_{(2)} > \frac{3}{4}\right) = \binom{3}{1}\frac{1}{2}\cdot\left(\frac{1}{4}\right)^2 = \frac{3}{32}$$

#

Notice that given $X_1, \dots, X_n \implies X_{(1)}, \dots, X_{(n)}$ is determined; the viceversa is false. The piece of information that is missing is the vector of ranks of the X_i's.

Def: The rank of X_i among $X_1, \dots, X_n$ is the value $R_i \in \{1, \dots, n\}$ s.t. $X_i = X_{(R_i)}$. (unique, if X_i's are continuous)

Then $\underline{R} = (R_1, \dots, R_n)$ is the vector of ranks.

Remarks

a: Then $(\underline{R}, X_{(1)}, \dots, X_{(n)}) \xleftrightarrow{1-1} (X_1, \dots, X_n)$

b: $\underline{R} = (R_1, \dots, R_n) \sim \mathcal{U}[P(n)]$ where $P(n) = \{\pi \text{ permutation of } \{1, \dots, n\}\}$ $\hookrightarrow$ with $|P(n)| = n!$

Indeed, $R_k \in \{1, \dots, n\}$ $\forall k$ and

$$P(R_1 = \pi_1, \dots, R_n = \pi_n) = \mathbb{P}(X_{\pi_1} < X_{\pi_2} < \dots < X_{\pi_n}) = \frac{1}{n!}$$

Moreover it holds:

Theorem 3: in the above notation:

(1) $(X_{(1)}, \dots, X_{(n)})$ and $(R_1, \dots, R_n)$ are independent

(2) If $g: \mathbb{R}^n \to \mathbb{R}$ and $Y = g(X_1, \dots, X_n)$ with finite mean, then

$$\mathbb{E}(Y \mid \underline{R} = \pi) = \mathbb{E}(g(X_{(1)}, \dots, X_{(n)}))$$

for $\pi = (\pi_1, \dots, \pi_n)$

Ex: Compute the marginal densities of the rank R_j, $j=1,\dots,n$, its average and variance.

Sol: Since $\underline{R} \sim \mathcal{U}[\mathcal{S}(n)]$, we have $\forall j=1,\dots n$

$$\mathbb{P}(R_k = k) = \sum_{\substack{\pi\in\mathcal{S}(n):\\ \pi_j = k}} \mathbb{P}(\underline{R}=\pi) = \sum_{\substack{\pi\in\mathcal{S}(n):\\ \pi_j=k}} \frac{1}{n!} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

$\Longrightarrow R_k \sim \mathcal{U}\{1,2,\dots n\}$ $\forall k=1,\dots n$ (all i.i.d.)

Then $\mathbb{E}(R_k) = \sum_{j=1}^{n} j\cdot\frac{1}{n} = \frac{n+1}{2}$

$$\mathbb{E}(R_k^2) = \sum_{j=1}^{n} j^2\frac{1}{n} = \frac{(n+1)(2n+1)}{6}$$

and $\mathrm{Var}(R_k) = \frac{n^2-1}{12}$ $\cancel{*}$

Application: let $T = \sum_{j=1}^{n} a_j R_j$ (linear rank statistics)

for constants $a_1,\dots,a_n$. Compute $\mathbb{E}(T)$ and $\mathrm{Var}(T)$.

$$\mathbb{E}(T) = \sum_{j=1}^{n} a_j\,\mathbb{E}(R_j) = \frac{n+1}{2}\sum_{j=1}^{n} a_j$$

$$\mathrm{Var}(T) = \sum_{j=1}^{n} a_j\,\mathrm{Var}(R_j) + 2\sum_{i<j=1}^{n} a_i a_j\,\mathrm{Cov}(R_i,R_j)$$

To compute the covariance $\mathrm{Cov}(R_i;R_j)$ we may notice that

$$\sum_{j=1}^{n} R_j = \frac{n+1}{2} \quad \text{(deterministic)}$$

$$\Longrightarrow \mathrm{Var}\left(\sum_{j=1}^{n} R_j\right) = 0 = \sum_{j=1}^{n}\mathrm{Var}(R_j) + 2\sum_{i<j=1}^{n}\mathrm{Cov}(R_i R_j)$$

$\longrightarrow$

Since $Cov(R_i, R_j)$ does not depends on i and j by the symmetricity of problem, we get altogether

$$n(n-1)\, Cov(R_i R_j) = -n\, Var(R_k) = -n\frac{(n^2-1)}{12}$$

$$\Rightarrow Cov(R_i, R_j) = -\frac{n^2-1}{12(n-1)} = -\frac{n+1}{12}$$

[otherwise, compute marginal density of (R_i, R_j) and do explicit computation]

Finally $\quad Var(T) = \frac{n^2-1}{12}\sum_{j=1}^{n} a_j^2 - \frac{n+1}{6}\sum_{i<j=1}^{n} a_j a_i$.

Notice that if $X_1, \dots, X_n$ are iid, $E(T)$ and $Var(T)$ should agree with the computed values. These quantities are then used to test the data (and verify they are iid), in particular time trend data.

Exercise: Let $X_1, \dots, X_n$ iid RV's $\sim U(0,1)$
$a_1, \dots, a_n$ constant. Show that

$$\underbrace{\mathbb{E}\left[\sum_{j=1}^{n} a_j X_j \mid R_1, \dots, R_n\right)}_{h(R_1, \dots, R_n)} \text{ is a linear rank statistics}$$

Sol: By theorem 3 (2), if $\pi = (\pi_1, \dots, \pi_n)$ is permutation $\mathcal{S}_n$

$$\mathbb{E}\Big(\sum_{j=1}^{n} a_j X_j \mid \underline{R} = \pi\Big) = \mathbb{E}\Big[\sum_{j=1}^{n} a_j X_{(\pi_j)}\Big] = \sum_{j=1}^{n} a_j E(X_{(\pi_j)})$$

We've already shown that $X_{(j)} \sim Beta(j, n-j+1)$

and thus has average $\mathbb{E}(X_{(j)}) = \frac{j}{n+1}$

$\longrightarrow$

$$\Rightarrow \mathbb{E}\left[\sum_{j=1}^{n} a_j X_j \,\middle|\, \underline{R} = \pi\right] = \sum_{j=1}^{n} a_j \frac{\pi_j}{n+1}$$

and $$\mathbb{E}\left[\sum_{j=1}^{n} a_j X_j \,\middle|\, \underline{R}\right] = \sum_{j=1}^{n} \frac{a_j}{n+1} R_j \quad \#$$

<u>MARTINGALES</u>: let $X_1, X_2, \dots$ be (an infinite) sequence of RV's s.t. $\mathbb{E}(X_k) < \infty$ $\forall k \in \mathbb{N}$

We say that $(X_n)_{n\in\mathbb{N}}$ is a <u>martingale</u> if

$$* \quad \mathbb{E}(X_{n+1} \mid X_1, X_2, \dots, X_n) = X_n \qquad \forall n \geq 1$$

<u>Idea</u>: X_n = amount of a gambler after n-th game, $\forall n \geq 1$ starting from X_0

the ~~relation~~ $*$ expresses that games are fair

<u>Prop</u>: $(X_n)_{n\in\mathbb{N}}$ is a martingale $\Leftrightarrow$ $\forall$ $m \leq n$

$$\mathbb{E}(X_n \mid X_1, X_2, \dots, X_m) = X_m$$

<u>Dim</u>: $\Leftarrow$ obvious (take $m = n-1$)

$$\Rightarrow \mathbb{E}(X_n \mid X_1, \dots, X_m) = \mathbb{E}\left[\mathbb{E}(X_n \mid X_1, \dots, X_m, X_{n-1}) \,\middle|\, X_1 \dots X_m\right]$$

$$\overset{\text{martingale property}}{=} \mathbb{E}\left[X_{n-1} \mid X_1 \dots X_m\right) = \overset{\text{iterate}}{\dots} =$$

$$= \mathbb{E}(X_{m+1} \mid X_1 \dots X_m) = X_m \quad \#$$

<u>Example 1</u>: let $Y_1, \dots, Y_n, \dots = (Y_n)_{n\in\mathbb{N}}$ sequence of independent RV's s.t. $\mathbb{E}(Y_k) = 0$ $\forall k \in \mathbb{N}$

Define $$X_n = \sum_{j=1}^{n} Y_j \quad , n \in \mathbb{N}$$

Then $(X_n)_{n\in\mathbb{N}}$ is a martingale. Indeed

$$\mathbb{E}(X_{n+1} \mid X_1 + X_2 \dots + X_n) = \mathbb{E}[X_n + Y_{n+1} \mid X_1, \dots, X_n]$$

$$= X_n + \mathbb{E}(Y_{n+1} \mid X_1, \dots X_n) \overset{\text{indep}}{=} X_n + \underbrace{\mathbb{E}(Y_n)}_{=0} = X_n$$

Example 2: let $(Y_n)_{n\in\mathbb{N}}$ indep. RV's s.t. $\mathbb{E}(Y_k) = 1$ $\forall k \in \mathbb{N}$

Define $X_n = \prod_{k=1}^{n} Y_k$, $\forall n \in \mathbb{N}$

Then $$\mathbb{E}(X_{n+1} \mid X_1, \dots, X_n) = \mathbb{E}[Y_{n+1} \cdot X_n \mid X_1, \dots, X_n]$$

$$= X_n \, \mathbb{E}(Y_{n+1} \mid X_1, \dots, X_n) = X_n \cdot \underbrace{\mathbb{E}(Y_{n+1})}_{=1} = X_n$$

that is $(X_n)_{n\in\mathbb{N}}$ is a martingale

$\downarrow$

Game "double or nothing": Set $X_0 = 1$ (starting stake) the gambler at each step double the stake as long he loses, and stop to play as soon as he wins. So that

$$X_{n+1} = \begin{cases} 2X_n & \text{prob } 1/2 \\ 0 & \text{prob } 1/2 \end{cases}$$

$$\Longrightarrow \mathbb{P}(X_n = 2^n) = \frac{1}{2^n}, \quad \mathbb{P}(X_n = 0) = 1 - \frac{1}{2^n} \quad \forall n \geq 1$$

Notice that $X_n = \prod_{k=1}^{n} Y_k$ with $\mathbb{P}(Y_k = 0) = \mathbb{P}(Y_k = 2) = \frac{1}{2}$ and $\mathbb{E}(Y_k) = 1$

$\Longrightarrow (X_n)_{n\in\mathbb{N}}$ is a martingale

Let us compute:

i * the distribution and the average of the number (N) of games played (up to the first winning)

ii * the average amount spent (S) during the total play-

Sol: i. this is clearly a geometric distribution with parameter $1/2$ so that $\mathbb{P}(N=k) = \frac{1}{2^k}$ and $\mathbb{E}(N) = 2$.

ii If the gambler stop at the k-th game, then the spent amount is $1+2+\dots+2^{k-1} = 2^k - 1$

that is
$$\mathbb{P}(S=n) = \begin{cases} \frac{1}{2^k} & \text{if } n = 2^k - 1, \ \forall k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(S) = \sum_{k=1}^{\infty} \frac{1}{2^k} (2^k - 1) = \infty$$

Likelihood ratio statistic

Let $(X_n)_{n \in \mathbb{N}}$ iid RV's with density f depending on a parameter θ. To test the null hypothesis $H_0 : \theta = \theta_0$ against alternative $H_1 : \theta = \theta_1$, by the Neyman-Pearson lemma we can use the likelihood ratio statistic

$$L_n = \prod_{k=1}^{n} \frac{f(X_k; \theta_1)}{f(X_k; \theta_0)}$$ densities under H_1 and H_0 respectively

Notice that $Y_k = \frac{f(X_k; \theta_1)}{f(X_k; \theta_0)}$, $k \in \mathbb{N}$ are iid RV's s.t. →

$$\mathbb{E}_{\theta_0}\left(\frac{f(X_k;\theta_1)}{f(X_k;\theta_0)}\right) = \int_{\mathbb{R}} \frac{f(x,\theta_1)}{f(x,\theta_0)} f(x;\theta_0)\,dx = \int_{\mathbb{R}} f(x,\theta_1)\,dx = 1$$

$$\Longrightarrow L_n = \prod_{k=1}^{n} Y_k \quad , n\in\mathbb{N} \text{ is a martingale} \quad \#$$

A general fact regarding martingales (easily derived from $*$), is that $\mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) = \dots = \mathbb{E}(X_1)$

(for example is 0 if $X_n = \sum_{j=1}^{n} Y_j$
and is 1 if $X_n = \prod_{j=1}^{n} Y_j$)
