Order Statistics and Probability

Varenya Upadhyaya

IITH

May 5, 2021

Introduction

Definition and Notation

- For a given statistical sample, the k^{th} order statistic is the k^{th} smallest value in said sample.
- This means that if the sample is arranged in ascending order, the k^{th} order statistic is nothing but the k^{th} element from the left.
- Often denoted as $X_{(k)}$
- For a sample $\{X_1, X_2, \cdots X_n\}$ of size n:

$$X_{(1)} = \min\{X_1, X_2, \cdots X_n\} \tag{1}$$
$$X_{(n)} = \max\{X_1, X_2, \cdots X_n\} \tag{2}$$

Introduction

Range and Median

For a sample $\{X_1, X_2, \cdots X_n\}$, the range is given by:

$$\text{Range}\{X_1, X_2, \cdots X_n\} = X_{(n)} - X_{(1)} \tag{3}$$

For Median two cases arise:

1. $n = 2k$

$$\text{Median} = \frac{1}{2}(X_{(k)} + X_{(k+1)}) \tag{4}$$

2. $n = 2k + 1$

$$\text{Median} = X_{(k)} \tag{5}$$

Example

Consider a set of values $\{5, 2, 9, 16, 8\}$. For this the order statistics will be:

$$X_{(1)} = 2, X_{(2)} = 5, X_{(3)} = 8, X_{(4)} = 9, X_{(5)} = 16 \tag{6}$$

The Range and Median for this will be:

$$\text{Range} = X_{(5)} - X_{(1)} = 16 - 2 = 14 \tag{7}$$

$$\text{Median} = X_{(3)} = 8 \tag{8}$$

Density Functions

- Consider a collection of i.i.d. random variables $\{X_1, X_2, \cdots X_n\}$ with cdf $F(x)$ and pdf $f(x)$.
- Let $F_{X_{(k)}}$ and $f_{X_{(k)}}$ be the Cumulative and Probability Density Functions of $X_{(k)}$ respectively.

$$F_{X_{(k)}}(x) = \Pr\left(X_{(k)} \leq x\right) \tag{9}$$

When we say $X_{(k)} \leq x$, the following condition must be satisfied:

$$X_{(i)} \leq x \quad \forall i < k \tag{10}$$

however for $i > k$, no such restriction exists and thus separate cases must be taken,

CDF

Assume that when put in ascending order, the sample becomes $\{X_1', X_2' \cdots X_n'\}$, Eq.(9) can then be rewritten as:

$$F_{X_{(k)}} = {}^n\mathrm{C}_k \Pr\left(X_1' \leq x \cdots X_k' \leq x\right) \Pr\left(X_{k+1}' > x, \cdots X_n' > x\right) + {}^n\mathrm{C}_{k+1} \Pr\left(X_1' \leq x, \cdots X_{k+1}' \leq x\right) \Pr\left(X_{k+2}' > x, \cdots X_n' > x\right) + \cdots + {}^n\mathrm{C}_n \Pr\left(X_1' \leq x, \cdots X_n' \leq x\right) \quad (11)$$

$$F_{X_{(k)}}(x) = \sum_{j=k}^{n} {}^n\mathrm{C}_j \times \prod_{i=1}^{j} \Pr\left(X_i' \leq x\right) \times \prod_{i=j+1}^{n} \Pr\left(X_i' > x\right) \quad (12)$$

$$= \sum_{j=k}^{n} {}^n\mathrm{C}_j \times (F(x))^j \times (1 - F(x))^{n-j} \quad (13)$$

CDF

Two special cases

Equation (13) takes two simple forms for $X_{(1)}$ and $X_{(n)}$:

$$F_{X_{(1)}} = 1 - (1 - F(x))^n \tag{14}$$

$$F_{X_{(n)}} = (F(x))^n \tag{15}$$

PDF

The pdf for a general $X_{(k)}$ can be calculated by differentiating Eq.(13):

$$f_{X_{(k)}}(x) = \frac{d}{dx}\sum_{j=k}^{n} {}^{n}\mathrm{C}_{j} \times (F(x))^{j} \times (1 - F(x))^{n-j} \tag{16}$$

$$= \sum_{j=k}^{n} {}^{n}\mathrm{C}_{j} \times \frac{d}{dx}\left((F(x))^{j} \times (1 - F(x))^{n-j}\right) \tag{17}$$

$$= \sum_{j=k}^{n} {}^{n}\mathrm{C}_{j} \times \left(j\,(F(x))^{j-1} f(x) \times (1 - F(x))^{n-j}\right)$$

$$- \sum_{j=k}^{n} {}^{n}\mathrm{C}_{j} \times \left((F(x))^{j} \times (n - j)\,(1 - F(x))^{n-j-1} f(x)\right) \tag{18}$$

$$= f(x)\sum_{j=k}^{n} {}^{n}\mathrm{C}_{j} \left((F(x))^{j-1} \times (1 - F(x))^{n-j-1} \times (j - nF\,(x))\right) \tag{19}$$

PDF

(19) can be expanded and simplified to give:

$$\implies f_{X_{(k)}}(x) = nf(x) \times {}^{n-1}\mathrm{C}_{k-1} \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} \tag{20}$$

A simpler approach to calculate the pdf is as follows, for some $i \in \{1, 2 \cdots n\}$:

$$\Pr\left(X_{(k)} \in [x, x + dx]\right) = \Pr\left(X_i \in [x, x + dx], \text{exactly } k - 1 \text{ terms} < x\right)$$

$$= {}^{n}\mathrm{C}_1 \Pr\left(X'_k \in [x, x + dx]\right) \times {}^{n-1}\mathrm{C}_{k-1} \Pr\left(X'_1 < x \cdots X'_{k-1} < x\right) \times \Pr\left(X'_{k+1} > x \cdots X'_n > x\right)$$

$$\implies f_{x_{(k)}}(x) = nf(x) \times {}^{n-1}\mathrm{C}_{k-1} \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} \tag{21}$$

PDF

Eq.(21) takes relatively simpler forms for $X_{(1)}$ and $X_{(n)}$

Two special cases

For $k = 1$ and $k = n$ it becomes:

$$f_{X_{(1)}}(x) = n \times (1 - F(x))^{n-1} \times f(x) \tag{22}$$

$$f_{X_{(n)}}(x) = n \times (F(x))^{n-1} \times f(x) \tag{23}$$

Example

Consider an exponential distribution with pdf:

$$f(x) = \begin{cases} e^{-x} & \text{when } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

For a random sample of size 3, the pdfs of the order statistics will be:

$$f_{X_{(1)}} = 3e^{-3x} \tag{25}$$

$$f_{X_{(2)}} = 6e^{-2x}\left(1 - e^{-x}\right) \tag{26}$$

$$f_{X_{(3)}} = 3e^{-x}\left(1 - e^{-x}\right)^2 \tag{27}$$

Example(Figure)

The plots for the three pdfs can be seen in Fig. 1:

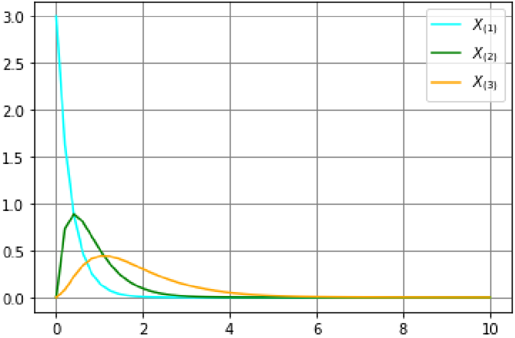


Figure: PDF for $X_{(i)}$

gov/stats/2015/statistics-I(1), Q.3(b)

Let Y_1 denote the first order statistic in a random sample of size n from a distribution that has the pdf,

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{when } \theta < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Obtain the distribution of $Z_n = n(Y_1 - \theta)$

Solution

The problem involves the first order statistic, we begin the solution by finding the cdf of the given distribution:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \tag{28}$$

$$= \int_{-\infty}^{\theta} 0\,dt + \int_{\theta}^{x} e^{\theta - t}\,dt \tag{29}$$

$$= \left(e^{\theta - t}\right)_{x}^{\theta} \tag{30}$$

$$= 1 - e^{\theta - x} \tag{31}$$

$$= 1 - f(x) \tag{32}$$

Solution-CDF

Let $F_{Z_n}(z)$ and $f_{Z_n}(z)$ be the cdf and pdf for Z_n. Calculating the cdf for Z_n:

$$F_{Z_n}(z) = \Pr\left(n(Y_1 - \theta) \le z\right) \tag{33}$$

$$= \Pr\left(Y_1 \le \frac{z}{n} + \theta\right) \tag{34}$$

$$= 1 - \Pr\left(Y_1 \ > \frac{z}{n} + \theta\right) \tag{35}$$

Let $\dfrac{z}{n} + \theta = z'$

$$F_{Z_n} = 1 - \Pr\left(X_1 > z', X_2 > z', ..., X_n > z'\right) \tag{36}$$

$$= 1 - \prod_{i=1}^{n} \Pr\left(X_i > z'\right) \tag{37}$$

$$= 1 - \left(1 - F(z')\right)^n \tag{38}$$

Solution-CDF

Substituting the expression for z' back,

$$\implies F_{Z_n}(z) = 1 - \left(1 - F\left(\frac{z}{n} + \theta\right)\right)^n \tag{39}$$

$$= 1 - \left(f\left(\frac{z}{n} + \theta\right)\right)^n \tag{40}$$

The expression for the cdf can thus be written as:

$$F_{Z_n}(z) = \begin{cases} 1 - e^{-n\left(\frac{z}{n}+\theta-\theta\right)} & \text{when } \theta < \frac{z}{n} + \theta < \infty \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

$$= \begin{cases} 1 - e^{-z} & \text{when } 0 < z < \infty \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

Solution-PDF

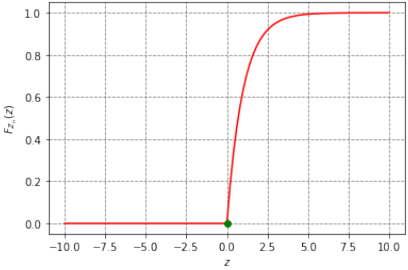
Using the cdf in (42) to calculate the pdf:

$$f_{Z_n}(z) = \frac{d}{dz} F_{Z_n}(z) \tag{43}$$

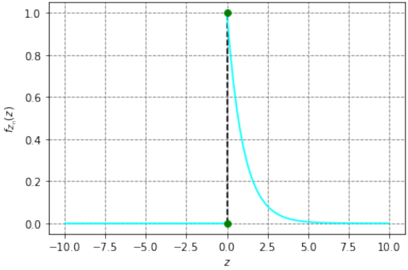
$$= \begin{cases} e^{-z} & \text{when } 0 < z < \infty \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

Solution-Plots

The plots for the cdf in (42) and the pdf in (44) are shown below:



(a) cdf of Z_n



(b) pdf of Z_n

Figure: Plots