Order Statistics

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §5.4)

Introduction

- A useful statistic of a random sample is to order the sample values in ascending order.
- This is called order statistics, denoted by $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, distinguishing from the original values $x_1, x_2, \cdots, x_n$.
- The *sample minimum*, $x_{(1)}$, and the *sample maximum*, $x_{(n)}$, are also order statistics.
- The *sample median* is the middle order statistic, $x_{(m+1)}$, if $n = 2m + 1$ (n is odd).
- If n is even, the sample median is usually taken to be the average of the two middle order statistics, $(x_{(n/2)} + x_{(n/2+1)})/2$.

Introduction (cont'd)

- The *sample range*, $R = x_{(n)} - x_{(1)}$, is the distance between the smallest and largest observations.
- $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are not independent since

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

- They are not identically distributed as well since

$$EX_{(1)} < EX_{(2)} < \cdots < EX_{(n)}.$$

Sample Maximum

- The distribution of the sample maximum can be easily derived since

$$\{X_{(n)} \le x\} = \{X_1 \le x, \cdots, X_n \le x\}$$

- This implies

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(X_1 \le x, \cdots, X_n \le x) = \{F(x)\}^n$$

- If X is continuous,

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = nf(x)\{F(x)\}^{n-1}.$$

- **Example** If f is the uniform(0,1) pdf, then

$$f_{X_{(n)}}(x) = nx^{n-1}, \quad x \in (0, 1).$$

Sample Minimum

- Similarly,

$$\{X_{(1)} > x\} = \{X_1 > x, \cdots, X_n > x\}$$

- This implies

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x) = 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \{1 -$$

- If X is continuous,

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = nf(x)\{1 - F(x)\}^{n-1}.$$

- **Example** If f is the exp(β) pdf, then

$$f_{X_{(1)}}(x) = n\beta^{-1} e^{-x/\beta}\{1 - 1 + e^{-x/\beta}\}^{n-1} = (\beta/n)^{-1} e^{-x/(\beta/n)}.$$

Joint Distribution of Order Statistics

- The vector of order statistics is a function of the sample values, $(x_{(1)}, \dots, x_{(n)}) = g(x_1, \cdots, x_n)$.
- The inverse transformation, from order statistics to sample values, does not exist (not 1-to-1).
- What did we learn from “not 1-to-1" previously? Partition!!!
- Restrict the sample to, for example, the set

$$\{(x_1, x_2, x_3) : x_2 < x_3 < x_1\}.$$

  We would be able to compute the inverse of $(x_{(1)} = 2, x_{(2)} = 5, x_{(3)} = 9)$ as $(x_1 = 9, x_2 = 2, x_3 = 5)$.
- How may such sets? It's $3! = 6$.

Joint Distribution of Order Statistics (cont'd)

- Keep in mind that the order statistics are a permutation of the sample values.
- Partition: $A_1, \cdots, A_{n!}$. Let g_i be the transformation on A_i and g^{-1} be its inverse.
- Each row and column of Jacobian matrix (or called *permutation matrix* here) consists of 1 one and $n-1$ zeros, so $|J| = 1$.
- The joint pdf of the order statistics is

$$f_{X_{(1)},\cdots,X_{(n)}}(y_1, \cdots, y_n) = \sum_{j=1}^{n!} f_{X_1,\cdots,X_n}(g_j^{-1}(y_1, \cdots, y_n)) = n! \prod_{i=1}^{n} f_X(y_i),$$

for $y_1 < \cdots < y_n$.

Distribution of $X_{(j)}$

- $\{X_{(j)} \leq x\} = \{$at least j of the sample vales are $\leq x\}$.
- If $Z_i = I(X_i \leq x)$ and $Y_i = \sum_{i=1}^{n} Z_i$, then $\{X_{(j)} \leq x\} = \{Y \geq j\}$.
- Let $A = F(x)$ and $a = f(x)$. We have

$$\begin{aligned} F_{X_{(j)}}(x) &= P(X_{(j)} \leq x) = P(Y \geq j) \\ &= \sum_{k=j}^{n} P(Y = k) = \sum_{k=j}^{n} \binom{n}{k} A^k (1 - A)^{n-k}. \end{aligned}$$

- The pdf of $X_{(j)}$ is

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\ &= \sum_{k=j}^{n} \binom{n}{k} k a A^{k-1} (1 - A)^{n-k} - \sum_{k=j}^{n} \binom{n}{k} A^k (n-k) a (1 - \\ &= C - D \end{aligned}$$

Distribution of $X_{(j)}$ (cont'd)

- C can be expressed by $C = C_1 + C_2$, where

$$C_1 = \binom{n}{j} jaA^{j-1}(1-A)^{n-j}$$

$$\begin{aligned} C_2 &= \sum_{k=j+1}^{n} \binom{n}{k} kaA^{k-1}(1-A)^{n-k} \\ &= \sum_{t=j}^{n-1} \binom{n}{t+1} (t+1)aA^{t}(1-A)^{n-t-1} \\ &= \sum_{t=j}^{n-1} \binom{n}{t} (n-t)aA^{t}(1-A)^{n-t-1}. \end{aligned}$$

- One can show $C_2 = D$ since the last term in D ($j = n$) is 0.
- $f_{X_{(j)}}(x) = C_1 = \binom{n}{j} jaA^{j-1}(1-A)^{n-j}$

Distribution of $X_{(j)}$ (cont'd)

$$
\begin{aligned}
f_{X_{(j)}}(x) = C_1 &= \binom{n}{j} jaA^{j-1}(1-A)^{n-j} \\
&= \frac{n!}{(j-1)!(n-j)!} f(x)\{F(x)\}^{j-1}\{1-F(x)\}^{n-j}
\end{aligned}
$$

- Intuitive interpretation: $(j-1)$ observations are on the left of $X_{(j)}$, contributing $\{F(x)\}^{j-1}$, $X(j)$ itself, contributing $f(x)$, and $(n-j)$ observations are on the right of $X_{(j)}$, contributing $\{1-F(x)\}^{n-j}$.
- The combinatorial factor is the number of ways in which n observations can be grouped into three sets containing $j-1$, 1, and $n-j$ observations.

Distribution of $X_{(j)}$ (cont'd)

- **Example** Suppose that $X_1, \cdots, X_n$ are iid from the uniform density on $(0, 1)$. Then for $1 \le j \le n$,

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j}$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{n-j}, \quad x \in (0, 1)$$

- This is the pdf of $Beta(j, n-j+1)$ with $EX_{(j)} = \frac{j}{n+1}$ and $VarX_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}$.
- If $n = 2m + 1$ (n is odd), it follows that the sample median, $X_{(m+1)}$, has a $Beta(m+1, m+1)$ density with mean 1/2 and variance $1/\{4(n+2)\}$.
- The expected value of sample mean is 1/2 and variance $1/(12n)$.

Distribution of $(X_{(i)}, X_{(j)})$

- This follows the same lines as the derivation of $f_{X_{(j)}}$.
- The joint distribution of $(X_{(i)}, X_{(j)})$ is

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(u)f(v) \times F(u)^{i-1}\{F(v)-F(u)\}^{j-i-1}\{1-F(v)\}^{n-j}$$

- **Example** Suppose that $X_1, \cdots, X_n$ are iid from the uniform density on $(0, a)$, $a > 0$. For $0 < x < y < a$,

$$f_{X_{(1)},X_{(n)}}(x,y) = \frac{n(n-1)(y-x)^{n-2}}{a^n}.$$

- One may be interested in the distribution of the range variable $R = X_{(n)} - X_{(1)}$ and midrange variable $V = (X_{(n)} + X_{(1)})/2$,

Distribution of $(X_{(i)}, X_{(j)})$ (cont'd)

- One has $X_{(n)} = V + R/2$, $X_{(1)} = V - R/2$, and $|J| = 1$. The joint pdf of (R, V) is

$$f_{R,V}(r, v) = f_{X_{(1)},X_{(n)}}(v + r/2, v - r/2) = \frac{n(n-1)r^{n-2}}{a^n},$$

for $0 < r < a$ and $r/2 < v < a - r/2$ since $0 < x_{(1)} < x_{(n)} < a$.

- The support region of (R, V) is a triangle.
- The marginal pdf of R can be obtained as

$$f_R(r) = \int_{r/2}^{a-r/2} f_{R,V}(r, v)dv = \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a.$$

Distribution of $(X_{(i)}, X_{(j)})$ (cont'd)

- If $Z = R/a$, then $Z \sim Beta(n-1, 2)$ since

$$
\begin{aligned}
f_Z(z) &= n(n-1)z^{n-2}(1-z) \\
&= \frac{1}{B(n-1,2)} z^{n-2}(1-z), \quad z \in (0,1).
\end{aligned}
$$