Lec 3. Sampling distributions of order statistics

Zhang W.P.

Contents

1.1 Order statistics 1
1.2 Distribution of sample range 8
1.3 Gaps distribution 10
1.4 Distribution of sample quantile 15
1.5 Quantile 17

Learning objectives

After this section, you should be able to ...

- derive the distribution of order statistics
- understand the distribution of sample quantile

1.1 Order statistics

These statistics have in recent times come to play an important role in statistical inference partly because some of their properties do not depend upon the distribution from which the random sample is obtained.

Def.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the continuous type having a pdf $f(x)$ that has support $\mathcal{S} = (a, b)$, where $-\infty \leq a < b \leq \infty$. Let $X_{(1)}$ be the smallest of these X_i, $X_{(2)}$ the next X_i in order of magnitude, ..., and $Y_{(n)}$ the largest of X_i. That is, $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ represent $X_1, X_2, \ldots, X_n$ when the latter are arranged in ascending order of magnitude. We call $X_{(i)}$, $i = 1, 2, \ldots, n$, the ith order statistic of the random sample $X_1, X_2, \ldots, X_n$.

Theorem 1. *Using the above notation, let* $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *denote the* n *order statistics based on the random sample* $X_1, X_2, \ldots, X_n$ *from a continuous distribution with pdf* $f(x)$ *and support* (a, b)*. Then the joint pdf of* $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *is given by*

$$g(y_1, y_2, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n) & a < y_1 < y_2 < \cdots < y_n < b \\ 0 & \textit{elsewhere.} \end{cases}$$

Let X denote a random variable of the continuous type with a pdf $f(x)$ that is positive and continuous, with support $\mathcal{S} = (a, b), -\infty \leq a < b \leq \infty$. The distribution function $F(x)$ of X may be written

$$F(x) = \int_a^x f(w)dw, \quad a < x < b$$

If $x \leq a, F(x) = 0$; and if $b \leq x, F(x) = 1$. Thus there is a unique median m of the distribution with $F(m) = \frac{1}{2}$. Let X_1, X_2, X_3 denote a random sample from this distribution and let $Y_1 < Y_2 < Y_3$ denote the order statistics of the sample. Note that Y_2 is the sample median. We compute the probability that $Y_2 \leq m$.

↑Exam
↓Exam

The joint pdf of the three order statistics is

$$g(y_1, y_2, y_3) = \begin{cases} 6f(y_1)f(y_2)f(y_3) & a < y_1 < y_2 < y_3 < b \\ 0 & \text{elsewhere} \end{cases}$$

The pdf of Y_2 is then

$$\begin{aligned} h(y_2) &= 6f(y_2) \int_{y_2}^{b} \int_{a}^{y_2} f(y_1)f(y_3)dy_1 dy_3 \\ &= \begin{cases} 6f(y_2)F(y_2)\left[1 - F(y_2)\right] & a < y_2 < b \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

Accordingly,

$$\begin{aligned} P(Y_2 \leq m) &= 6 \int_{a}^{m} \left\{ F(y_2)f(y_2) - \left[F(y_2)\right]^2 f(y_2) \right\} dy_2 \\ &= 6 \left\{ \frac{\left[F(y_2)\right]^2}{2} - \frac{\left[F(y_2)\right]^3}{3} \right\}_a^m = \frac{1}{2} \end{aligned}$$

Hence, for this situation, the median of the sample median Y_2 is the population median m.

Once it is observed that

$$\int_a^x [F(w)]^{\alpha-1} f(w)dw = \frac{[F(x)]^\alpha}{\alpha}, \quad \alpha > 0$$

and that

$$\int_y^b [1-F(w)]^{\beta-1} f(w)dw = \frac{[1-F(y)]^\beta}{\beta}, \quad \beta > 0$$

it is easy to express the marginal pdf of any order statistic, say Y_k, in terms of $F(x)$ and $f(x)$. This is done by evaluating the integral

$$g_k(y_k) = \int_a^{y_k} \cdots \int_a^{y_2} \int_{y_k}^b \cdots \int_{y_n-1}^b n! f(y_1) f(y_2) \cdots f(y_n) dy_n \cdots dy_{k+1} dy_1 \cdots dy_{k-1}$$

The result is

$$g_k(y_k) = \begin{cases} \frac{n!}{(k-1)!(n-k)!} [F(y_k)]^{k-1} [1-F(y_k)]^{n-k} f(y_k) & a < y_k < b \\ 0 & \text{elsewhere.} \end{cases} \tag{1.1}$$

Finally, the joint pdf of any two order statistics, say $Y_i < Y_j$, is easily expressed in terms of $F(x)$ and $f(x)$. We have

$$g_{ij}(y_i, y_j) = \int_a^{y_i} \cdots \int_a^{y_2} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{b} \cdots \int_{y_{n-1}}^{b} n! f(y_1) \times \cdots \\ \times f(y_n) dy_n \cdots dy_{j+1} dy_{j-1} \cdots dy_{i+1} dy_1 \cdots dy_{i-1}.$$

Since, for $\gamma > 0$

$$\int_x^y [F(y) - F(w)]^{\gamma-1} f(w) dw = -\left. \frac{[F(y) - F(w)]^\gamma}{\gamma} \right|_x^y \\ = \frac{[F(y) - F(x)]^\gamma}{\gamma},$$

it is found that

$$g_{ij}(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(y_i)]^{i-1} [F(y_j) - F(y_i)]^{j-i-1} \\ \times [1 - F(y_j)]^{n-j} f(y_i) f(y_j) I(a < y_i < y_j < b)$$

Heuristic Derivation. There is an easy method of remembering the pdf of a vector of order statistics such as the one given in formula (1.1). The probability $P(y_i < Y_i < y_i + \Delta_i, y_j < Y_j < y_j + \Delta_j)$, where Δ_i and Δ_j are small, can be approximated by the following multinomial probability. In n independent trials, $i-1$ outcomes must be less than y_i [an event that has probability $p_1 = F(y_i)$ on each trial] ; $j-i-1$ outcomes must be between $y_i + \Delta_i$ and y_j[an event with approximate probability $p_2 = F(y_j) - F(y_i)$ on each trial]; $n-j$ outcomes must be greater than $y_j + \Delta_j$ [an event with approximate probability $p_3 = 1 - F(y_j)$ on each trial]; one outcome must be between y_i and $y_i + \Delta_i$ [an event with approximate probability $p_4 = f(y_i)\Delta_i$ on each trial ; and, finally, one outcome must be between y_j and $y_j + \Delta_j$ [an event with approximate probability $p_5 = f(y_j)\Delta_j$ on each trial]. This multinomial probability is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!1!1!} p_1^{i-1} p_2^{j-i-1} p_3^{n-j} p_4 p_5$$

which is $g_{i,j}(y_i, y_j)\Delta_i\Delta_j$, where $g_{i,j}(y_i, y_j)$ is given in expression (1.1)

1.2 Distribution of sample range

For sample range $\mathrm{R}_n = X_{(n)} - X_{(1)}$, because the joint density of $X_{(1)}$ and $X_{(n)}$ is given by

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)(F(y)-F(x))^{n-2} f(x) f(y) I(x<y),$$

the density of R_n can be shown as

$$f_{\mathrm{R}_n}(r) = n(n-1) \int (F(z+r)-F(z))^{n-2} f(z) f(z+r) dz \qquad (1.2)$$

If $X \sim \mathcal{U}(0,1)$, the density (1.2) becomes

$$f_{\mathrm{R}_n}(r) = n(n-1)\int_0^{1-r} (z+r-z)^{n-2}dz = n(n-1)r^{n-2}(1-r)$$

for $0 < r < 1$ and $f_{\mathrm{R}_n}(r) = 0$ otherwise.

↑Exam
↓Exam

1.3 Gaps distribution

There are many interesting (and important for applications) questions about order statistics. For example, given an n -sample from $\mathcal{U}(0,1)$ with the order statistic

$$0 \equiv X_{(0)} < X_{(1)} < X_{(2)} < \cdots < X_{(n)} < X_{(n+1)} \equiv 1$$

one is often interested in distribution of

- the extreme order variables $X_{(1)}$ and $X_{(n)}$,
- the kth gap
$$\Delta_{(k)}X \stackrel{\text{def}}{=} X_{(k)} - X_{(k-1)}$$
- the maximal gap $\max \Delta_{(k)}X = \max(X_{(k)} - X_{(k-1)})$,
- the minimal gap $\min \Delta_{(k)}X = \min(X_{(k)} - X_{(k-1)})$

where k ranges from 1 to $n+1$.

Lemma 1. *For a given n-sample from $\mathcal{U}(0,1)$, all gaps*

$$\Delta_{(k)}X \stackrel{def}{=} X_{(k)} - X_{(k-1)}, \quad k = 1, \ldots, n+1$$

have the same distribution with $\mathrm{P}(\Delta_{(k)}X > r) = (1-r)^n$ *for all* $r \in [0,1]$

Proof. Notice that

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

this implies that

$$f_{\Delta_{(k)}X}(r) = \int_0^{1-r} f_{X_{(k-1)},X_{(k)}}(x, x+r)dx$$

Changing the variables $x \mapsto y$ using $x = y(1-r)$ and comparing the result to the beta distribution, we deduce that $f_{\Delta_{(k)}X}(r) = n(1-r)^{n-1}I(0 < r < 1)$, equivalently, that $\mathrm{P}(\Delta_{(k)}X \geq r) = (1-r)^n$ for all $r \in [0,1]$.

□

Lemma 2. *Let an n-sample from $\mathcal{U}(0,1)$ be fixed. Then for all m, k, satisfying $1 \leq m < k \leq n+1$ we have*

$$\mathrm{P}\left(\Delta_{(m)}X \geq r_m, \Delta_{(k)}X \geq r_k\right) = (1 - r_m - r_k)^n$$

for all positive r_m, r_k satisfying $r_m + r_k \leq 1$.

Corollary 1. *Let $(X_k)_{k=1}^n$ be a fixed n-sample from $\mathcal{U}(0,1)$. Then for all positive r_k satisfying $\sum_{k=1}^{n+1} r_k \leq 1$ we have*

$$\mathrm{P}\left(\Delta_{(1)}X \geq r_1, \ldots, \Delta_{(n+1)}X \geq r_{n+1}\right) = \left(1 - \sum_{k=1}^{n+1} r_k\right)^n$$

In particular, the distribution of the vector of gaps $\left(\Delta_{(1)}X, \ldots, \Delta_{(n+1)}X\right)$ is exchangeable for every $n \geq 1$.

For an n-sample from $\mathcal{U}(0,1)$, let $D_n \stackrel{\text{def}}{=} \min_k \Delta_{(k)} X$. Then $\mathrm{P}\left(D_n \geq r\right) = (1-(n+1)r)^n$. In particular, for every $y \geq 0$

$$\mathrm{P}_n\left(n^2 D_n > y\right) = \left(1 - \frac{n+1}{n^2} y\right)^n \approx e^{-y}$$

ie., for large n the distribution of $Y_n \stackrel{\text{def}}{=} n^2 D_n$ is approximately $\mathrm{Exp}(1)$.

↑Exam

↓Exam

1.4 Distribution of sample quantile

For $0 < p < 1$, the p th quantile or fractile of F is defined as

Def.

$$\xi(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

- If F is continuous and strictly increasing, F^{-1} is defined by

$$F^{-1}(y) = x \quad \text{when } y = F(x)$$

- If F has a discontinuity at x_0, suppose that $F(x_0-) < y < F(x_0)$. In this case, although there exists no x for which $y = F(x)$, $F^{-1}(y)$ is defined to be equal to x_0
- If $U \sim U(0,1)$, then $X = F_X^{-1}(U) \sim F_X(x)$.

Let $\{X_1, X_2, \ldots, X_n\}$ be a sample from F, the sample pth quantile $\hat{\xi}_{pn}$ (which is $X_{([np])}$ or $X_{([np]+1)}$ depending on whether np is an integer, here $[np]$ denotes the integer part of np) is defined as

Def.

$$\hat{\xi}_{pn} = F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$$

where F_n is the corresponding empirical distribution function.

- $X_{(i)} = F_n^{-1}\left(\frac{i}{n}\right), i = 1, \ldots, n$

Theorem 2. *Let $0 < p < 1$. Suppose that F possesses a density f in a neighborhood of ξ_p and f is positive and continuous at ξ_p, then*

$$\sqrt{n}\left(\hat{\xi}_{pn} - \xi_p\right) \xrightarrow{d} N\left(0, \frac{p(1-p)}{[f(\xi_p)]^2}\right)$$