Order Statistics and Distributions

1 Some Preliminary Comments and Ideas

In this section we consider a random sample $X_1, X_2, \ldots, X_n$ common continuous distribution function F and probability density f. Thus X_i, $i = 1, 2, \ldots, n$ are i.i.d. F (or f).

The order statistics are then the random variables $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(k)} \leq \ldots \leq X_{(n)}$. The random variable $X_{(k)}$ is thus the k-th largest of the n random variables $X_1, X_2, \ldots, X_n$.

These are interesting for various reasons in statistics. These are the basis of how QQ plots are constructed and why they *work* as a statistical diagnostic tool.

Since F is continuous then $P(X_1 = X_2) = 0$. To see this, notice that for any any $\epsilon > 0$

$$\begin{aligned} P\left(|X_1 - X_2| \leq \epsilon\right) &= \int_{-\infty}^{\infty} \int_{x_1-\epsilon}^{x_1+\epsilon} f(x_1) f(x_2) dx_2 dx_1 \\ &\to 0 \text{ as } \epsilon \to 0 \ . \end{aligned}$$

Therefore

$$P\left(|X_1 - X_2| = 0\right) = \lim_{\epsilon \to 0} P\left(|X_1 - X_2| \leq \epsilon\right) = 0 \ .$$

This will imply that for any sample size n the r.v.s $X_1, X_2, \ldots, X_n$ are all distinct with probability 1. We will not investigate this property further in this course. Note this would not occur if the X_i's were discrete r.v.'s.

For iid r.v.s from a continuous cdf F the r.v.s are all distinct and therefore no two of them are equal.

In this section we study the distribution of the order statistics. For any n the mapping

$$g(x_1, \ldots, x_n) = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$$

is a many to one mapping. In particular we see this quite easily when $n = 2$ where the mapping is

$$g(x, y) = (\min(x, y), \max(x.y)) \ .$$

Thus for any $a < b$ both (a, b) and (b, a) map to the same pair (a, b). Therefore the mapping or transformation is not a one to one invertible differential mapping so we cannot use the calculus method for these types of mappings.

This means that we will need to consider some special way of finding the distribution of the order statistics, and their one and two dimensional marginal distributions. As in our earlier study of such transformations the first attempt will be to find the cdf and from this obtain the pdf.

2 Marginal Distribution of $X_{(1)}$

The cdf, say F_1, of $X_{(1)}$ is easy to obtain.

$$\begin{aligned} F_1(x) &= P\left(X_{(1)} \leq x\right) \\ &= 1 - P\left(X_{(1)} > x\right) \\ &= 1 - P\left(X_1 > x, X_2 > x, \ldots, X_n > x\right) \\ &= 1 - \prod_{i=1}^{n} P\left(X_i > x\right) \\ &= 1 - (1 - F(x))^n \end{aligned}$$

Differentiating w.r.t. x gives f_1 the pdf of $X_{(1)}$ as

$$f_1(x) = nf(x)(1 - F(x))^{n-1} \ .$$

The student should try using a similar argument to find the pdf, f_n, of $X_{(n)}$. The answer is

$$f_n(x) = nf(x)F(x)^{n-1} \ .$$

Notice that in both cases the support of f_1 and of f_n is inherited from the support of f.

3 Joint Distribution of $X_{(1)}, X_{(2)}$

In this section we find the joint cdf of $X_{(1)}, X_{(2)}$, say $F_{1,2}$ and then obtain the joint pdf of $X_{(1)}, X_{(2)}$, say $f_{1,2}$.

Even if $n = 2$ the transformation $(X_1, X_2, \ldots, X_n) \mapsto (X_{(1)}, X_{(2)})$ is not 1 to 1. Thus we cannot find the pdf $f_{1,2}$ directly. We obtain $F_{1,2}$ and then obtain from this $f_{1,2}$.

Let $x < y$. Consider the event

$$B = \{X_{(1)} \leq x, X_{(2)} > y\} \ .$$

In words the event B occurs if and only if 1 of the n "trials" X_i takes on a value less than or equal to x and $n - 1$ of the trials takes on values greater than y. Notice there are $\binom{n}{1}$ ways of this occurring. Since the X_i's are iid thus

$$P(B) \quad = \quad \binom{n}{1} F(x) \left(1 - F(y)\right)^{n-1} \ .$$

Notice that we can decompose the event

$$\{X_{(1)} \leq x\} = \{X_{(1)} \leq x, X_{(2)} \leq y\} \cup \{X_{(1)} \leq x, X_{(2)} > y\} \ .$$

Thus by the axioms by of probability we have, for $x < y$

$$\begin{aligned} F_1(x) &= P(X_{(1)} \leq x) \\ &= F_{1,2}(x, y) + P(B) \end{aligned}$$

Also

$$\begin{aligned} P(X_{(1)} > x) &= P(X_i > x \text{ for all } i = 1, \ldots, n) \\ &= (1 - F(x))^n \end{aligned}$$

and hence $P(X_{(1)} \leq x) = 1 - (1 - F(x))^n$.

This yields, for $x < y$

$$F_{1,2}(x, y) = 1 - (1 - F(x))^n - nF(x)\left(1 - F(y)\right)^{n-1}$$

Taking the second partial derivative with respect to x, y gives

$$\begin{aligned} f_{1,2}(x, y) &= \frac{\partial^2 F_{1,2}(x, y)}{\partial x \, \partial y} \\ &= -nf(x)(n-1)\left(1 - F(y)\right)^{n-2}(-1)f(y) \\ &= n(n-1)f(x)f(y)\left(1 - F(y)\right)^{n-2} \ . \end{aligned}$$

For $x > y$ we also have $f_{1,2}(x, y) = 0$. The student should think about why this is the case. Putting these together we have

$$f_{1,2}(x, y) = \begin{cases} n(n-1)f(x)f(y)\left(1 - F(y)\right)^{n-2} & \text{if } x \leq y \\ 0 & \text{otherwise} \end{cases}$$

This method works well for finding the joint pdf of $X_{(n-1)}, X_{(n)}$.

This direct method becomes increasingly more complicated in terms of combinatorics, and inclusion exclusion arguments. For example it is already complicated for finding the pdf of $X_{(1)}, X_{(3)}$. Thus the *heuristic* method is used to obtain the correct answer. It is helpful to notice that

$$\frac{n!}{1! \; 1! \; (n-2)!} = n(n-1)$$

which is a combinatorial expression that comes up in the heuristic method below.

A similar method of adding and subtracting appropriate regions can also be used to derive the marginal pdf of $X_{(k)}$ and other bivariate marginals. However this method is very complicated to keep track of all the regions.

4 Heuristic Method

This is based on the idea of implementing the multinomial model for observations falling into various *bins*. The term heuristic refers to a process or method for attempting the solution of a problem, while giving a reasonable rule for obtaining the result may not give a technical justification for the steps involved.

4.1 Marginal Distribution

Consider $X_{(k)}$. For a given small $\Delta > 0$

$$\begin{aligned} & P(X_{(k)} \in [x, x + \Delta)) \\ & \quad \approx \frac{n!}{(k-1)!1!(n-k)!}\left\{F(x)^{k-1}\left(F(x + \Delta) - F(x)\right)\left(1 - F(x + \Delta)\right)^{n-k}\right\} \end{aligned}$$

There is an error in the approximation. We do not study it in this course, but this error is small compared with Δ. The error is due to ignoring that perhaps more than one observation might fall into the interval $[x, x+\Delta)$. It is for this reason that the method is *heuristic*. Let f_k be the pdf of $X_{(k)}$. Recall that for a pdf

$$P(X_{(k)} \in [x, x+\Delta)) = \int_x^{x+\Delta} f_k(y)dy$$

and therefore

$$\frac{P(X_{(k)} \in [x, x+\Delta))}{\Delta} = \frac{1}{\Delta}\int_x^{x+\Delta} f_k(y)dy \to f_k(x)$$

as $\Delta \to 0^+$. A similar property holds of course for F and its pdf f. (*There was a typo in the two lines above, now corrected.*) Therefore

$$\begin{aligned} f_k(x) &= \lim_{\Delta\to 0^+} \frac{P(X_{(k)} \in [x, x+\Delta))}{\Delta} \\ &= \frac{n!}{(k-1)!1!(n-k)!}\left\{F(x)^{k-1} \lim_{\Delta\to 0^+} \frac{(F(x+\Delta)-F(x))}{\Delta} \lim_{\Delta\to 0^+} (1-F(x+\Delta))^{n-k}\right\} \\ & \quad \text{(uses limit of product is product of limits)} \\ &= \frac{n!}{(k-1)!1!(n-k)!} F(x)^{k-1} f(x)\,(1-F(x+))^{n-k} \end{aligned}$$

Notice that the method used here is to recognize that we are calculating the probability of random variables falling into one of three bins, that is for each trial of a random variable, the outcome falls into one of three bins. This of course in the multinomial distribution, which the student should have reviewed in readings earlier in the course. The method or argument is *heuristic* since possibly more than one r.v. can fall into the bin $[x, x+\Delta)$, but this last part can be shown to be small relative to Δ, a property which we will not work through in this course.

4.2 Bivariate Marginal Distribution

Consider $X_{(k)}, X_{(\ell)}$, where $1 \le k < \ell \le n$. Let the joint pdf be $f_{k,\ell}$. Consider the event such that for small Δ the k and ℓ order statistics fall into the intervals $[x, x+\Delta)$ and $[y, y+\Delta)$ respectively, with $x < y$. Thus approximately

$$\begin{aligned} & P(X_{(k)} \in [x, x+\Delta), X_{(\ell)} \in [y, y+\Delta)) \\ & \approx \frac{n!}{(k-1)!1!(\ell-k-1)!1!(n-\ell)!}\Big\{F(x)^{k-1}(F(x+\Delta)-F(x))(F(y)-F(x+\Delta))^{\ell-k-1} \\ & \qquad (F(y+\Delta)-F(y))(1-F(y+\Delta))^{n-\ell}\Big\} \end{aligned}$$

The error in this approximation is that events with 2 or more ties are omitted. We will not study this further in this course, and take it as given that this error in the approximation is small relative to limits needed below.

The combinatorial term is obtained from counting the number of ways n indices in $\{1, 2, \ldots, n\}$ can be arranged into the 5 groups with $k-1$ of $X_1, X_2, \ldots, X_n$ fall into $(-\infty, x]$, 1 in $[x, x+\Delta)$, $\ell-k-1$ into the interval $[x+\Delta, y)$, 1 in $[y, y+\Delta)$ and the remaining $n-\ell$ into the interval $[y+\Delta, \infty)$.

Since the LHS is

$$P(X_{(k)} \in [x, x+\Delta), X_{(\ell)} \in [y, y+\Delta)) = \int_x^{x+\Delta} \int_y^{y+\Delta} f_{k,\ell}(u,v) dv du \approx f_{k,\ell}(x,y)\Delta\Delta$$

then dividing both sides by Δ^2, and taking the limit as $\Delta \to 0$, yields

$$\begin{aligned} & f_{k,\ell}(x,y) \\ = \quad & \frac{n!}{(k-1)!1!(\ell-k-1)!1!(n-\ell)!} \left\{ F(x)^{k-1} f(x) \left(F(y) - F(x)\right)^{\ell-k-1} f(y) \left(1 - F(y)\right)^{n-\ell} \right\} \end{aligned}$$

The coefficient in front is a generalization of the binomial coefficient and is sometimes written as

$$\binom{n}{(k-1)\ 1\ (\ell-k-1)\ 1\ (n-\ell)}$$

5 Examples

Example 1

Let $n = 4$, and that X_i are iid Uniform(0,1) random variables. Obtain the distribution of the sample median. Notice the sample median is thus $M = \frac{1}{2}\left(X_{(2)} + X_{(3)}\right)$. Thus we first need to obtain the density $f_{2,3}$ in the above notation.

The Uniform(0,1) pdf is

$$f(x) = \mathrm{I}_{(0,1)}(x) = \begin{cases} 1 & \text{if} \quad 0 < x < 1 \\ 0 & \text{if} \quad x \le 0 \text{ and } x \ge 1 \end{cases}$$

and the Uniform(0,1) CDF is

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ x & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x \ge 1\ . \end{cases}$$

Thus

$$f_{2,3}(x,y) = \begin{cases} 4!x(1-y) & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The CDF, say G, of M is thus given by

$$G(u) = P(M \le u) = P(X_{(2)} + X_{(3)} \le 2u)$$

and hence M has pdf

$$g(u) = 2\int_{-\infty}^{\infty} f_{2,3}(x, 2u - x) dx\ . \tag{2}$$

The student should review the method that was used to derive the convolution formula.

Aside : $X_{(2)}, X_{(3)}$ has pdf $f_{2,3}$. Thus $T = X_{(2)} + X_{(3)}$ has cdf F_T and M has pdf $F_M(u) = F_T(2u)$. Recall we have a formula for the pdf f_T in terms of an integral involving $f_{2,3}$.

$$\begin{aligned} f_M(u) &= \frac{dF_T(2u)}{du} \\ &= 2f_T(2u) \\ &= 2\int_{-\infty}^{\infty} f_{2,3}(x, 2u-x)dx \end{aligned}$$

End of Aside

To obtain the correct limits of integration we now have to use the specific properties of the uniform pdf, specifically it support as needed in (1). For a given value of u, let A_u be the support of the integrand in (2) so that $g(u) = 2\int_{A_u} f_{2,3}(x, 2u - x)dx$.

$$\begin{aligned} A_u &= \{x : 0 < x < 1, x < 2u - x \text{ and } 2u - x < 1\} \\ &= \{x : 0 < x < 1, 2x < 2u \text{ and } 2u - 1 < x\} \end{aligned}$$

We obtain these three pairs of inequalities from the support of $f_{2,3}$.

If $0 < u < \frac{1}{2}$ then $2u - 1 < 0$ and hence

$$A_u = \{x : 0 < x < 1, x < u\} = (0, u).$$

If $\frac{1}{2} < u < 1$ then $1 < 2u < 2 = 0 < 2u - 1 < 1$, and hence

$$A_u = \{x : 0 < x < 1, 2u - 1 < x < u\} = (2u - 1, u).$$

The student should complete the integral at home and find the the distribution of the sample median M in this case.

Example 2

Find the marginal distribution of the k-th order statistic from a sample of iid Uniform(0,1) random variables.

By the heuristic method the pdf f_k of $U_{(k)}$ is

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} f(x)(1-F(x))^{n-k}$$

where F and f are the cdf and pdf of the Uniform(0,1) distribution. There were calculate in the previous example.

Thus we have

$$f_k(x) = \begin{cases} \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Remark Since pdf's integrate to 1 we obtain from example 2 the following formula for integers $n > 0$ and $1 \le k \le n$:

$$\int_0^1 x^{k-1}(1-x)^{n-k} dx = \frac{(k-1)!(n-k)!}{n!} \ . \tag{3}$$

Example continued : Expected Value of k-th order statistic from a sample of n iid Uniform(0,1) random variables. Suppose that $U_1, U_2, \ldots, U_n$ are iid $\text{Unif}(0,1)$ r.v.s.

$$\begin{aligned} \mathrm{E}(U_{(k)}) &= \int_0^1 x f_k(x) dx \\ &= \int_0^1 x \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} dx \\ &= \frac{n!}{(k-1)!(n-k)!} \int_0^1 x^{k+1-1}(1-x)^{n-k} dx \\ &= \frac{n!}{(k-1)!(n-k)!} \int_0^1 x^{k+1-1}(1-x)^{(n+1)-(k+1)} dx \\ &= \frac{n!}{(k-1)!(n-k)!} \times \frac{(k+1-1)!(n+1-k-1)!}{(n+1)!} \\ &= \frac{k}{n+1} \end{aligned}$$

In the second last line we are using (3) with integer values $k+1$ and $n+1$.

Example 3

Find the marginal distribution of the 2-nd order statistic from a sample of $n = 4$ iid exponential paramter 1 random variables.

By the heuristic method the pdf f_k of $X_{(k)}$ is

$$f_2(x) = \frac{4!}{(2-1)!(4-2)!} F(x)^{2-1} f(x)(1-F(x))^{4-2}$$

where F and f are the cdf and pdf of the exponential parameter 1 distribution.

Specifically this gives $f_2(x) = 0$ when $x < 0$, and for $x \geq 0$ we have

$$\begin{aligned} f_2(x) &= 12\left(1 - e^{-x}\right) e^{-x} \left(1 - 1 + e^{-x}\right)^2 \\ &= 12\left(1 - e^{-x}\right) e^{-3x} \ . \end{aligned}$$

The student should now calculate for example

$$P(X_{(2)} > 1)$$

and

$$\mathrm{E}(X_{(2)}) \ .$$