Probability

Jan-April 2014

Chennai Mathematical Institute

Professor Nandini Kannan

Chapter 6

Order Statistics

Definition: The sequence of random variables $X_1, \ldots, X_n$ is said to form a **random sample** of size n from the population $F(.)$ if

1. $X_1, \ldots, X_n$ are independent;
2. $X_1, \ldots, X_n$ are identically distributed (i.e. have the same cdf).

We also refer to $X_1, \ldots, X_n$ as independent and identically distributed (iid) random variables.

For discrete random variables, the joint pmf of $X_1, \ldots, X_n$ is

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

For continuous random variables, the joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

Definition: A **statistic** $T = T(X_1, \ldots, X_n)$ is a function of $X_1, \ldots, X_n$ that does not depend on any unknown parameters.
The statistic T can be real or vector valued. We have

$$T : \mathcal{R}^n : \mathcal{R}^m \qquad m < n.$$

Ideally, m will represent the number of parameters.

Example: The following are statistics:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{Sample Mean.}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{Sample Variance.}$$

Since a statistic is also a random variable, we can talk about its probability distribution.
Definition: Let $T = T(X_1, \ldots, X_n)$ be a statistic. The probability distribution of T is called the **sampling distribution** of T.

Lemma 6.0.1. *Let $X_1, \ldots, X_n$ be iid random variables and $g(.)$ be a function such that $E[g(X_1)]$ and $Var[g(X_1)]$ exist. Then*

$$\begin{aligned} E[\sum_{i=1}^{n} g(X_i)] &= nE[g(X_1)] \\ Var[\sum_{i=1}^{n} g(X_i)] &= nVar[g(X_1)]. \end{aligned}$$

■

Theorem 6.0.2. *Let $X_1, \ldots, X_n$ be iid random variables from a population with mean μ and variance $\sigma^2 < \infty$. Then*

$$M_{\bar{X}}(t) = [M_X(t/n)]^n .$$

Proof: We have

$$\begin{aligned} M_{\bar{X}}(t) &= E[e^{t\bar{X}}] = E[e^{\frac{t}{n}(X_1+\ldots+X_n)}] \\ &= E\left[e^{\frac{t}{n}X_1} e^{\frac{t}{n}X_2} \ldots e^{\frac{t}{n}X_1}\right] \\ &= [M_X(t/n)]^n . \end{aligned}$$

■

6.1 Order Statistics

Let $(X_1, \ldots, X_n)$ be an $n-$dimensional random vector and $(x_1, \ldots, x_n)$ be an $n-$tuple assumed by $(X_1, \ldots, X_n)$.
Arrange $x_1, \ldots, x_n$ in increasing order of magnitude so that

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Definition: The function $X_{(k)}$ of $(X_1, \ldots, X_n)$ that takes on the value $x_{(k)}$ in each possible sequence

$(x_1, \ldots, x_n)$ of values assumed by $(X_1, \ldots, X_n)$ is known as the **k−th order statistic**.

$\{X_{(1)}, \ldots, X_{(n)}\}$ is called the set of order statistics for $(X_1, \ldots, X_n)$.

Example: Let X_1, X_2, X_3 be 3 discrete random variables. Let X_1, X_3 take values $0, 1$ and X_2 take values 1,2, and 3. Then the random vector $(X_1.X_2, X_3)$ assumes the following triples of values:

$$(0,1,0),(0,2,0),(0,3,0),(0,1,1),(0,2,1),(0,3,1),$$
$$(1,1,0),(1,2,0),(1,3,0),(1,1,1),(1,2,1),(1,3,1).$$

$X_{(1)}$ takes values $0, 1$, $X_{(2)}$ takes values $0, 1$ and $X_{(3)}$ takes values $1, 2, 3$.

Example: The sample range involves the order statistics $X_{(n)}$ and $X_{(1)}$:

$$R = X_{(n)} - X_{(1)}.$$

Example: The median involves the order statistics

$$M = \begin{cases} X_{((n+1)/2)}, & n \text{ odd}; \\ \frac{X_{(n/2)}+X_{(n/2+1)}}{2}, & n \text{ even}. \end{cases}$$

For any number p between 0 and 1, the $100p$-th sample percentile is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ of the observations are greater.

Definition: The notation $[b]$ is defined to be the number b rounded to the nearest integer.

Definition: The $100p$-th sample percentile is $X_{([np])}$ if $1/2n < p < .5$ and $X_{(n+1-[n(1-p)])}$ if $.5 < p < 1 - 1/2n$.
The cases $p < .5$ and $p > .5$ are defined separately so that the sample percentiles exhibit the following symmetry: if the $100p$-th sample percentile is the $i-$th smallest observation, then the $100(1-p)$-th sample percentile should be $i-$th largest observation.

Theorem 6.1.1. *Let $(X_1, \ldots, X_n)$ be a random sample from a discrete distribution with pmf*

$$p_X(x_i) = p_i,$$

where $x_1 < x_2 < \ldots$ are the possible values of X in ascending order. Define

$$\begin{aligned} P_0 &= 0 \\ P_1 &= p_1 \\ P_2 &= p_1 + p_2 \\ &\vdots \\ P_i &= p_1 + \ldots + p_i \\ &\vdots \end{aligned}$$

Then

$$P[X_{(j)} \leq x_i] = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{6.1}$$

$$P[X_{(j)} = x_i] = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right]. \tag{6.2}$$

Proof: For a fixed i, let Y be the number of $X_1, \ldots, X_n$ that are less than or equal to x_i. Define the event $\{X_j \leq x_i\}$ as a success. We have

$$P(Success) = P(X_j \leq x_i) = P_i.$$

The trials are independent because $X_1, \ldots, X_n$ are independent. Therefore

$$Y \sim Bin(n, P_i).$$

We also have

$$\{X_j \leq x_i\} = \{Y \geq j\},$$

i.e. at least j of the sample values are less than or equal to x_i. Therefore

$$\begin{aligned} P(X_j \leq x_i) &= P(Y \geq j) \\ &= \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}. \\ P(X_j \leq x_i) &= P(X_j \leq x_i) - P(X_j \leq x_{i-1}). \end{aligned}$$

■

Suppose $X_1, \ldots, X_n$ are iid continuous random variables with pdf $f(.)$. With probability one,

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

Theorem 6.1.2. *The joint pdf of* $(X_{(1)}, \ldots, X_{(n)})$ *is given by*

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i) \qquad -\infty < x_1 < \ldots < x_n < \infty. \tag{6.3}$$

Proof: The transformation from $X_1, \ldots, X_n$ to $(X_{(1)}, \ldots, X_{(n)})$ is not one-to-one.
For any set of n values $x_1, \ldots, x_n$, there are $n!$ possible arrangements of $x_1, \ldots, x_n$ in increasing order of magnitude. This implies there are $n!$ inverses to the transformation. Therefore,

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i) \qquad -\infty < x_1 < \ldots < x_n < \infty.$$

■

Theorem 6.1.3. *The marginal pdf of* $X_{(r)}$ *is given by*

$$f_{X_{(r)}}(x_r) = \frac{n!}{(r-1)!(n-r)!}[F_X(x_r)]^{r-1}[1 - F_X(x_r)]^{n-r} f_X(x_r). \tag{6.4}$$

Proof: We have

$$\begin{aligned} f_{X_{(r)}}(x_r) &= n! f_X(x_r) \int_{-\infty}^{x_r} \int_{-\infty}^{x_{r-1}} \cdots \int_{-\infty}^{x_2} \\ &\quad \int_{x_r}^{\infty} \int_{x_{r+1}}^{\infty} \cdots \int_{x_{n-1}}^{\infty} \prod_{i \neq r} f_X(x_i) dx_n \ldots dx_{r+1} dx_1 \ldots dx_{r-1} \\ &= n! f_X(x_r) \frac{[1 - F_X(x_r)]^{n-r}}{(n-r)!} \int_{-\infty}^{x_r} \int_{-\infty}^{x_{r-1}} \cdots \int_{-\infty}^{x_2} \prod_{i=1}^{r-1} f_X(x_i) dx_i \\ &= \frac{n!}{(r-1)!(n-r)!} f_X(x_r)[F_X(x_r)]^{r-1}[1 - F_X(x_r)]^{n-r}. \end{aligned}$$

■

Theorem 6.1.4. *The joint pdf of* $X_{(j)}$ *and* $X_{(k)}$, $1 \leq j < k \leq n$ *is given by*

$$\begin{aligned} f_{X_{(j)},X_{(k)}}(x_j, x_k) &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!}[F_X(x_j)]^{j-1}[F_X(x_k) - F_X(x_j)]^{k-j-1} \\ &\quad [1 - F_X(x_k)]^{n-k} f_X(x_j) f_X(x_k), \quad x_j < x_k. \end{aligned} \tag{6.5}$$

■

Example: Let $X_1, \ldots, X_n$ be iid $U(0,1)$. We have

$$F(x) = \int_0^x dt = x.$$

Therefore

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

which is the pdf of a Beta random variable. Therefore

$$X_{(j)} \sim Beta(j, n-j+1).$$

Let $R = X_{(n)} - X_{(1)}$ be the range, and $V = (X_{(1)} + X_{(n)})/2$ be the midrange.
The joint pdf of $X_{(1)}$ and $X_{(n)}$ is

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = n(n-1)[x_n - x_1]^{n-2} \qquad 0 < x_1 < x_n < 1.$$

The inverse transformations are

$$X_{(1)} = V - \frac{R}{2} \qquad X_{(n)} = V + \frac{R}{2},$$

and the Jacobian is 1. We know that $0 < r < 1$. We have the following inequalities

$$v - r/2 > 0 \Rightarrow v > r/2.$$

$$v + r/2 < 1 \Rightarrow v < 1 - r/2.$$

This implies $r/2 < v < 1 - r/2$. Therefore

$$f_{R,V}(r, v) = n(n-1)r^{n-2} \qquad 0 < r < 1, r/2 < v < 1 - r/2.$$

The marginal pdf of R is

$$f_R(r) = \int_{r/2}^{1-r/2} n(n-1)r^{n-2} dv = n(n-1)r^{n-2}(1-r) \quad 0 < r < 1.$$

To find the pdf of V, we consider the two cases: If $v < 1/2$, then $0 < r < 2v$. If $v > 1/2$, then $0 < r < 2(1-v)$. Therefore,

$$f_V(v) = \begin{cases} n(2v)^{n-1}, & 0 < v < 1/2; \\ n[2(1-v)]^{n-1}, & 1/2 < v < 1. \end{cases}$$

■

Example: Let $X_1, \ldots, X_n$ be iid with cdf

$$F(x) = x^\alpha \quad 0 < x < 1, \alpha > 0.$$

Show that $\frac{X_{(i)}}{X_{(n)}}, i = 1, \ldots, n-1$ and $X_{(n)}$ are independent.

■