Basic Theories On Order Statistics

Po-Ning Chen, Professor

Institute of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

Distribution of the two-end order statistic

Assumption $X_1, \ldots, X_n$ are i.i.d. with marginal cdf $F(\cdot)$.

Then

$$\begin{aligned} F_{(n)}(x) &= \Pr[X_{(n)} \le x] \\ &= \Pr[\max_{1\le i\le n} X_n \le x] \\ &= \Pr[X_1 \le x \wedge \cdots \wedge X_n \le x] \\ &= \Pr[X_1 \le x] \cdots \Pr[X_n \le x] \\ &= F^n(x). \end{aligned}$$

Likewise,

$$\begin{aligned} F_{(1)}(x) &= \Pr[X_{(1)} \le x] \\ &= 1 - \Pr[X_{(1)} > x] \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

How about the distribution of $X_{(r)}$?

Density of $X_{(r)}$

$$\begin{aligned} F_{(r)}(x) &= \Pr[X_{(r)} \leq x] \\ &= \Pr[\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x] \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} \\ &= I_{F(x)}(r, n - r + 1). \end{aligned}$$

For $a > 0$, $b > 0$ and $0 \leq p \leq 1$,

$$I_p(a, b) = \frac{\int_0^p t^{a-1}(1-t)^{b-1} dt}{\int_0^1 t^{a-1}(1-t)^{b-1} dt}$$

is the *incomplete beta function.*

A well-known result for the incomplete beta function is:

$$\sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} = I_p(r, n - r + 1).$$

Density of $X_{(r)}$

If X has density, then so does $X_{(r)}$.

The density of $X_{(r)}$ is equal to:

$$\begin{aligned} f_{(r)}(x) &= \frac{dI_{F(x)}(r, n-r+1)}{dx} \\ &= \frac{1}{\int_0^1 t^{r-1}(1-t)^{n-r}dt} \frac{d}{dx} \int_0^{F(x)} t^{r-1}(1-t)^{n-r}dt \\ &= \frac{1}{B(r, n-r+1)} F^{r-1}(x)[1-F(x)]^{n-r} f(x), \end{aligned}$$

where

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

is the Euler beta function, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the Euler gamma function.

Joint distribution of several order statistics

Denote the joint density function of $X_{(r)}$ and $X_{(s)}$ by the assumed existing $f_{(r,s)}(x,y)$, where $1 \le r < s \le n$.

$f_{(r,s)}$ can be derived in an explicit way for $x < y$.
($f_{(r,s)}(x,y) = 0$ for $x > y$!)

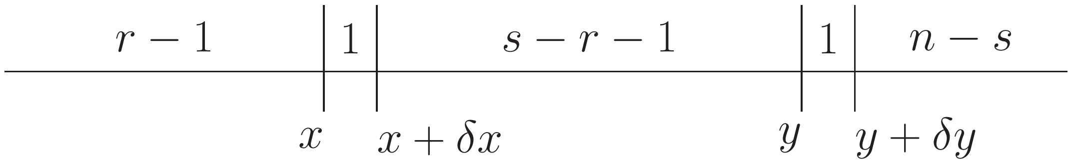
In other words,

$$\Pr\left[\left(x < X_{(r)} \le x + \delta x\right) \wedge \left(y < X_{(s)} \le y + \delta y\right)\right]$$

The above event can be described as:

1. $(r-1)$ of X's are less than x;
2. one X's lies between x and $x + \delta x$;
3. $(s-r-1)$ of X's lies between $x + \delta x$ and y;
4. one X's lies between y and $y + \delta y$;
5. $(n-s)$ of X's is larger than $y + \delta y$.

Joint distribution of several order statistics



Hence, we can estimate $f_{(r,s)}(x,y)$ for $x<y$ through:

$$\frac{n!}{(r-1)!\cdot 1!\cdot (s-r-1)!\cdot 1!\cdot (n-s)!}\times$$

$$F^{r-1}(x)[F(x+\delta x)-F(x)][F(y)-F(x+\delta x)]^{s-r-1}[F(y+\delta y)-F(y)][1-F(y)]^{n-s}.$$

By dividing δx and δy and letting them approaching 0, we obtain that for $x<y$,

$$\begin{aligned} & f_{(r,s)}(x,y) \\ = \; & \frac{n!}{(r-1)!\cdot (s-r-1)!\cdot (n-s)!}F^{r-1}(x)f(x)[F(y)-F(x)]^{s-r-1}f(y)[1-F(y)]^{n-s}. \end{aligned}$$

Joint distribution of several order statistics

As a result, for $x_1 < x_2 < \cdots < x_k$,

$$\begin{aligned}
& f_{(n_1,\ldots,n_k)}(x_1,\ldots,x_k) \\
= \; & \frac{n!}{(n_1-1)!(n_2-n_1-1)!\cdots(n-n_k)!} \\
& F^{n_1-1}(x_1)f(x_1)[F(x_2)-F(x_1)]^{n_2-n_1-1}f(x_2)[F(x_3)-F(x_2)]^{n_3-n_2-1}\cdots f(x_k)[1-F(x_k)]^{n-n_k} \\
= \; & n!\left[\prod_{j=1}^{k} f(x_j)\right]\left[\prod_{j=0}^{k}\frac{[F(x_{j+1})-F(x_j)]^{n_{j+1}-n_j-1}}{(n_{j+1}-n_j-1)!}\right],
\end{aligned}$$

where $x_0 = -\infty$, $x_{k+1} = \infty$, $n_0 = 0$ and $n_{k+1} = n+1$.

Joint distribution of several order statistics

The joint cdf of $X_{(r)}$ and $X_{(s)}$, where $1 \leq r < s \leq n$, can be derived directly (or by integrating $f_{(r,s)}(x,y)$) as that for $x < y$:

$$\begin{aligned}
F_{(r,s)}(x,y) &= \Pr\left[\text{at least } r \text{ of } X's \leq x \text{ and at least } s \text{ of } X's \leq y\right] \\
&= \sum_{j=s}^{n}\sum_{i=r}^{j} \Pr\left[\text{exactly } i \text{ of } X's \leq x \text{ and } \text{ exactly } j \text{ of } X's \leq y\right] \\
&= \sum_{j=s}^{n}\sum_{i=r}^{j} \frac{n!}{i!(j-i)!(n-j)!} F^{i}(x)[F(y)-F(x)]^{j-i}[1-F(y)]^{n-j},
\end{aligned}$$

and for $x \geq y$,

$$\begin{aligned}
F_{(r,s)}(x,y) &= \Pr\left[\text{at least } r \text{ of } X's \leq x \text{ and at least } s \text{ of } X's \leq y\right] \\
&= \Pr\left[\text{at least } s \text{ of } X's \leq y\right] \\
&= F_{(s)}(y).
\end{aligned}$$

Distribution of the range

Since we have the joint distribution of $X_{(r)}$ and $X_{(s)}$, we can derive the distribution of range $W_{(r,s)} = X_{(s)} - X_{(r)}$.

$$f_{(r,s)}(x,y) = \frac{n!}{(r-1)! \cdot (s-r-1)! \cdot (n-s)!} F^{r-1}(x) f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}.$$

$$\omega_{(r,s)}(w) = \int_{-\infty}^{\infty} f_{(r,s)}(x, w+x) dx \quad (\text{for } w > 0)$$

$$f_{(1,n)}(x,y) = \frac{n!}{(n-2)!} f(x) [F(y) - F(x)]^{n-2} f(y).$$

$$\omega_{(1,n)}(w) = \int_{-\infty}^{\infty} \frac{n!}{(n-2)!} f(x) f(w+x) [F(w+x) - F(x)]^{n-2} dx$$

Distribution of the range

The cdf of $W_{(1,n)}$ is:

$$\begin{aligned}
\Omega_{(1,n)}(w) &= \int_0^w \int_{-\infty}^{\infty} \frac{n!}{(n-2)!} f(x) f(z+x)[F(z+x) - F(x)]^{n-2} dx dz \\
&= \int_{-\infty}^{\infty} \frac{n!}{(n-2)!} f(x) \int_0^w f(z+x)[F(z+x) - F(x)]^{n-2} dz dx \\
&= \int_{-\infty}^{\infty} \frac{n!}{(n-2)!} f(x) \left(\frac{1}{n-1} [F(z+x) - F(x)]^{n-1} \Big|_0^w \right) dx \\
&= \int_{-\infty}^{\infty} n f(x) [F(w+x) - F(x)]^{n-1} dx.
\end{aligned}$$

Distribution of the range

Example Suppose $f(x) = 1$ for $0 \le x < 1$, and 0, otherwise.

- $F_{(r)}(x) = I_{F(x)}(r, n-r+1) = I_x(r, n-r+1)$ for $0 \le x < 1$, and 0, otherwise.
- For $0 \le x < 1$,

$$f_{(r)}(x) = \frac{1}{B(r, n-r+1)} F^{r-1}(x)[1 - F(x)]^{n-r} f(x)$$
$$= \frac{1}{B(r, n-r+1)} x^{r-1}(1-x)^{n-r}$$

  So, $X_{(r)}$ is beta distributed.

- For $1 \le r < s \le n$ and $0 \le x \le y \le 1$,

$$f_{(r,s)}(x, y)$$
$$= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F^{r-1}(x) f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}$$
$$= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} x^{r-1}(y-x)^{s-r-1}(1-y)^{n-s}.$$

Distribution of the range

- $(0 \leq x \leq y = w + x \leq 1)$

$$
\begin{aligned}
\omega_{(r,s)}(w) &= \int_0^{1-w} \frac{n!}{(r-1)!(s-r-1)!(n-s)!} x^{r-1}((w+x)-x)^{s-r-1}(1-(w+x))^{n-s}dx \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{s-r-1} \int_0^{1-w} x^{r-1}((1-w)-x))^{n-s}dx \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{(s-r)-1}(1-w)^{n-(s-r)} \int_0^1 z^{r-1}(1-z)^{n-s}dz \\
&\quad \text{(Let } x = z(1-w).) \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{(s-r)-1}(1-w)^{n-(s-r)} B(r, n-s+1) \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{(s-r)-1}(1-w)^{n-(s-r)} \frac{\Gamma(r)\Gamma(n-s+1)}{\Gamma(n+r-s+1)} \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{(s-r)-1}(1-w)^{n-(s-r)} \frac{(r-1)!(n-s)!}{(n+r-s)!} \\
&= \frac{n!}{((s-r)-1)!(n-(s-r))!} w^{(s-r)-1}(1-w)^{n-(s-r)},
\end{aligned}
$$

which is also beta distributed, and is only dependent on $s - r$ and not on individual r and s. □

Order statistics for discrete parents

Discrete parents

Assume (without loss of generality) that X takes values over $\{0, 1, 2, \ldots\}$.

The distribution of $X_{(r)}$ is still:

$$\begin{aligned} F_{(r)}(x) &= \Pr[X_{(r)} \leq x] \\ &= \Pr[\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x] \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} \\ &= I_{F(x)}(r, n - r + 1). \end{aligned}$$

So for non-negative integer x,

$$\begin{aligned} \Pr[X_{(r)} = x] &= F_{(r)}(x) - F_{(r)}(x - 1) \\ &= I_{F(x)}(r, n - r + 1) - I_{F(x-1)}(r, n - r + 1). \end{aligned}$$

Order statistics for discrete parents

The joint cdf of $X_{(r)}$ and $X_{(s)}$, where $1 \le r < s \le n$, can be derived directly as that for non-negative integers $x < y$:

$$\begin{aligned}
F_{(r,s)}(x,y) &= \Pr\left[\text{at least } r \text{ of } X's \le x \text{ and at least } s \text{ of } X's \le y\right] \\
&= \sum_{j=s}^{n} \sum_{i=r}^{j} \Pr\left[\text{exactly } i \text{ of } X's \le x \text{ and } \text{ exactly } j \text{ of } X's \le y\right] \\
&= \sum_{j=s}^{n} \sum_{i=r}^{j} \frac{n!}{i!(j-i)!(n-j)!} F^{i}(x)[F(y)-F(x)]^{j-i}[1-F(y)]^{n-j},
\end{aligned}$$

and for non-negative integers $x \ge y$,

$$\begin{aligned}
F_{(r,s)}(x,y) &= \Pr\left[\text{at least } r \text{ of } X's \le x \text{ and at least } s \text{ of } X's \le y\right] \\
&= \Pr\left[\text{at least } s \text{ of } X's \le y\right] \\
&= F_{(s)}(y).
\end{aligned}$$

This gives that for non-negative integers $x \le y$,

$$\Pr[X_{(r)} = x \wedge X_{(s)} = y] = \begin{cases} F_{(r,s)}(x,y) - F_{(r,s)}(x-1,y) - F_{(r,s)}(x,y-1) + F_{(r,s)}(x-1,y-1), & \text{if } x \le y; \\ 0, & \text{if } x > y. \end{cases}$$

Order statistics for discrete parents

An alternative but equivalent expression for $\Pr[X_{(r)} = x \wedge X_{(s)} = y]$ is as follows.

For non-negative integers x and y with $x < y$,

$$\boxed{(r-i) \le (r-1)} \qquad \boxed{(s-u) \le (s-1)}$$

$(r-i)$	$(i+t)$	$(s-u-(r+t))$	$(u+j)$	$(n-(s+j))$
$< x$	$x \cdots x$	$> x$ and $< y$	$y \cdots y$	$> y$

Denote $\Pr[X = x]$ by p_x.

Then the probability of the above snapshot case is equal to:

$$F^{r-i}(x-1)p_x^{i+t}[F(y-1) - F(x)]^{s-u-r-t}p_y^{u+j}[1 - F(y)]^{n-s-j}$$

Order statistics for discrete parents

Therefore,

$$\Pr[X_{(r)} = x \wedge X_{(s)} = y]$$

$$= \sum_{\substack{(r-i)\le(r-1),(s-u)\le(s-1),j\ge 0,t\ge 0 \\ r-i\ge 0,i+t\ge 1,s-u-(r+t)\ge 0,u+j\ge 1,n-(s+j)\ge 0}} A_{i,j,u,t} F^{r-i}(x-1) p_x^{i+t} [F(y-1) - F(x)]^{s-u-r-t} p_y^{u+j} [1 - F(y)]^{n-s-j}$$

$$= \sum_{i=1}^{r} \sum_{j=0}^{n-s} \sum_{u=\max\{1-j,1\}=1}^{s-r} \sum_{t=\max\{1-i,0\}=0}^{s-r-u} A_{i,j,u,t} F^{r-i}(x-1) p_x^{i+t} [F(y-1) - F(x)]^{s-u-r-t} p_y^{u+j} [1 - F(y)]^{n-s-j},$$

where

$$A_{i,j,u,t} = \frac{n!}{(r-i)!(i+t)!(s-u-r-t)!(u+j)!(n-s-j)!}.$$

Order statistics for discrete parents

Observe that

$$
\begin{aligned}
A_{i,j,u,t} &= \frac{n!}{(r-i)!(i+t)!(s-u-r-t)!(u+j)!(n-s-j)!} \\
&= \left(\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\right)\left(\frac{(r-1)!}{(i-1)!(r-i)!}\right)\left(\frac{(n-s)!}{j!(n-s-j)!}\right) \\
&\quad \left(\frac{(s-r-1)!}{(s-u-r)!(u-1)!}\right)\left(\frac{(s-u-r)!}{(s-u-t-r)!t!}\right)\left(\frac{(i-1)!t!}{(i+t)!}\right)\left(\frac{j!(u-1)!}{(u+j)!}\right) \\
&= C_{rs}\binom{r-1}{i-1}\binom{n-s}{j}\binom{s-r-1}{u-1}\binom{s-u-r}{t} \\
&\qquad \times\left(\int_0^1 z^{i-1}(1-z)^t dz\right)\left(\int_0^1 \theta^j(1-\theta)^{u-1}d\theta\right),
\end{aligned}
$$

where $C_{rs} = \dfrac{n!}{(r-1)!(s-r-1)!(n-s)!}$.

For nonnegative integers a and b,

$$\int_0^1 t^a(1-t)^b dt = \frac{a!b!}{(a+b+1)!}$$

Order statistics for discrete parents

Hence,

$$
\begin{aligned}
& \Pr[X_{(r)} = x \wedge X_{(s)} = y] \\
= \; & C_{rs} \sum_{i=1}^{r} \sum_{j=0}^{n-s} \sum_{u=1}^{s-r} \sum_{t=0}^{s-r-u} \binom{r-1}{i-1} \binom{n-s}{j} \binom{s-r-1}{u-1} \binom{s-u-r}{t} \\
& F^{r-i}(x-1) p_x^{i+t} [F(y-1) - F(x)]^{s-u-r-t} p_y^{u+j} [1 - F(y)]^{n-s-j} \\
& \left(\int_0^1 \int_0^1 z^{i-1} (1-z)^t \theta^j (1-\theta)^{u-1} dz d\theta \right) \\
= \; & C_{rs} \int_0^1 \int_0^1 \sum_{i=1}^{r} \sum_{j=0}^{n-s} \sum_{u=1}^{s-r} \sum_{t=0}^{s-r-u} \binom{r-1}{i-1} \binom{n-s}{j} \binom{s-r-1}{u-1} \binom{s-u-r}{t} \\
& F^{r-i}(x-1) p_x^{i+t} [F(y-1) - F(x)]^{s-u-r-t} p_y^{u+j} [1 - F(y)]^{n-s-j} z^{i-1} (1-z)^t \theta^j (1-\theta)^{u-1} dz d\theta \\
= \; & C_{rs} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} \left[\sum_{u=1}^{s-r} \binom{s-r-1}{u-1} [v - F(y-1)]^{u-1} \right. \\
\times \; & \sum_{t=0}^{s-r-u} \binom{s-u-r}{t} [F(y-1) - F(x)]^{s-u-r-t} [F(x) - w]^t \\
\times \; & \left. \sum_{i=1}^{r} \binom{r-1}{i-1} F^{r-i}(x-1) [w - F(x-1)]^{i-1} \sum_{j=0}^{n-s} \binom{n-s}{j} [1 - F(y)]^{n-s-j} [F(y) - v]^j \right] dw dv,
\end{aligned}
$$

where $v = F(y) - \theta p_y$ and $w = F(x-1) + z p_x$.

Order statistics for discrete parents

$$\Pr[X_{(r)} = x \wedge X_{(s)} = y]$$

$$= C_{rs} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} \left[\sum_{u=1}^{s-r} \binom{s-r-1}{u-1} [v - F(y-1)]^{u-1} \right.$$

$$\sum_{t=0}^{s-r-u} \binom{s-u-r}{t} [F(y-1) - F(x)]^{s-u-r-t} [F(x) - w]^t$$

$$\left. \sum_{i=1}^{r} \binom{r-1}{i-1} F^{r-i}(x-1)[w - F(x-1)]^{i-1} \sum_{j=0}^{n-s} \binom{n-s}{j} [1 - F(y)]^{n-s-j} [F(y) - v]^j \right] dw dv$$

$$= C_{rs} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} \left[\underbrace{\sum_{u=1}^{s-r} \binom{s-r-1}{u-1} [v - F(y-1)]^{u-1} [F(y-1) - w]^{s-r-u}}_{(v-w)^{s-r-1}} w^{r-1} (1-v)^{n-s} \right] dw dv$$

$$= C_{rs} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} (v-w)^{s-r-1} (1-v)^{n-s} w^{r-1} dw dv$$

$$= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} w^{r-1} (v-w)^{s-r-1} (1-v)^{n-s} dw dv.$$

Order statistics for discrete parents

Interesting though, the pmf $\Pr[X_{(r)} = x \wedge X_{(s)} = y]$ is the integration over the region $(F(x-1), F(x)) \times (F(y-1), F(y))$ for the density:

$$\begin{cases} \dfrac{n!}{(r-1)!(s-r-1)!(n-s)!} w^{r-1}(v-w)^{s-r-1}(1-v)^{n-s}, & \text{for } 0 \le w \le v < 1; \\ 0, & \text{otherwise.} \end{cases}$$

This density is the $f_{(r,s)}(x, y)$ in the aforementioned example (cf. Slide OR2-11).

This is similar to do the **qnantiles** on the cdf of $f_{(r,s)}(x, y)$
(Recall that $f_{(r,s)}(x, y)$ is the joint density of the order statistics, denoted by $U_{(r)}$ and $U_{(s)}$, for uniform-over-$[0, 1)$ parent distribution).

Order statistics for discrete parents

Can we establish a parent-distribution-free theory on order statistics? For example, x is the medium satisfying $F(x) = 1/2$. Then,

$$
\begin{aligned}
&\Pr[X_{(r)} \le x < X_{(s)}] \\
&= \sum_{i=0}^{x} \sum_{j=x+1}^{\infty} \Pr[X_{(r)} = i \wedge X_{(s)} = j] \\
&= \sum_{i=0}^{x} \sum_{j=x+1}^{\infty} \Pr[F(i-1) \le U_{(r)} < F(i) \wedge F(j-1) \le U_{(s)} < F(j)] \\
&= \sum_{i=0}^{x} \Pr[F(i-1) \le U_{(r)} < F(i) \wedge U_{(s)} \ge F(x)] \\
&= \Pr[U_{(r)} < F(x) \wedge U_{(s)} \ge F(x)] \\
&= \Pr\left[U_{(r)} < \frac{1}{2} \le U_{(s)}\right],
\end{aligned}
$$

which has nothing to do with the shape of function F.

Distribution-free confidence intervals for quantiles

Define the quantile of random variable X as:

$$Q(p) \triangleq \sup\{x \in \Re : F(x) \leq p\}.$$

Observation The probability of $Q(p)$ belonging to $[X_{(r)}, X_{(s)})$ for $1 \leq r < s \leq n$, namely

$$\Pr\left[X_{(r)} \leq Q(p) < X_{(s)}\right],$$

is *independent* of the distribution of X !

- This observation allows us to construct the **distribution-free confidence intervals for** $Q(p)$.

Distribution-free confidence intervals for quantiles

Observe that

$$\begin{aligned}\Pr[X_{(r)} \le Q(p)] &= \Pr[X_{(r)} \le Q(p) \wedge X_{(s)} > Q(p)] + \Pr[X_{(r)} \le Q(p) \wedge X_{(s)} \le Q(p)] \\ &= \Pr[X_{(r)} \le Q(p) < X_{(s)}] + \Pr[X_{(s)} \le Q(p)],\end{aligned}$$

which implies that if $F(\cdot)$ has inverse function,

$$\begin{aligned}\Pr[X_{(r)} \le Q(p) < X_{(s)}] &= \Pr[X_{(r)} \le Q(p)] - \Pr[X_{(s)} \le Q(p)] \\ &= I_{F(Q(p))}(r, n-r+1) - I_{F(Q(p))}(s, n-s+1) \\ &= I_p(r, n-r+1) - I_p(s, n-s+1) \\ &= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i},\end{aligned}$$

which is **independent** of $F(\cdot)$.

Distribution-free confidence intervals for quantiles

In case $F(\cdot)$ has no inverse function,

$$\Pr[X_{(r)} < Q(p) < X_{(s)}] \le \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \le \Pr[X_{(r)} \le Q(p) \le X_{(s)}].$$

Observation The probability that $[X_{(r)} \le a$ and $X_{(s)} > a]$ is still dependent on the distribution of $F(\cdot)$.

For example, if $F(\cdot)$ has inverse function,

$$\Pr[X_{(r)} \le a < X_{(s)}] = \sum_{i=r}^{s-1} \binom{n}{i} F^i(a)(1-F(a))^{n-i}.$$

Distribution-free confidence intervals for quantiles

Define $\pi(r, s, n, p) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$.

Definition Confidence intervals with confidence coefficient $\geq 1 - \alpha$.

- For given n and p, make $(s - r)$ as small as possible subject to $\pi(r, s, n, p) \geq 1 - \alpha$.

Example For given $p = 1/2$ (and any n),

$$\pi(r, s, n, 1/2) = \sum_{i=r}^{s-1} \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \sum_{i=r}^{s-1} \binom{n}{i}.$$

Then for fixed $d = (s - r)$, $\pi(r, s, n, 1/2)$ is largest, if $r = \left\lfloor \frac{n+1}{2} - \frac{d}{2} \right\rfloor$ and $s = \left\lfloor \frac{n+1}{2} + \frac{d}{2} \right\rfloor$.

Notably, $Q(1/2)$ is the **median**.

Distribution-free confidence intervals for quantiles

Some researchers approximate $(1-\alpha)$ confident interval for the median in terms of normal approximation of binomial distribution, which is accurate at n large.

$B_1, \ldots, B_n$ are i.i.d., and take values from $\{0, 1\}$.
Suppose $\Pr[B_1 = 1] = p$.
Then $B_1 + \cdots + B_n$ is binomial distributed with

$$\Pr[B_1 + \cdots + B_n = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

The central limit theorem says that

$$\frac{(B_1 + \cdots + B_n) - np}{\sqrt{p(1-p)n}} \Rightarrow N.$$

So

$$\begin{aligned}
&\pi\left(\left\lfloor\frac{n+1}{2}-\frac{d}{2}\right\rfloor,\left\lfloor\frac{n+1}{2}+\frac{d}{2}\right\rfloor,n,\frac{1}{2}\right)\\
=\ &\Pr\left[\left\lfloor\frac{n+1}{2}-\frac{d}{2}\right\rfloor\le B_1+\cdots+B_n<\left\lfloor\frac{n+1}{2}+\frac{d}{2}\right\rfloor\right]\\
\approx\ &\Pr\left[\frac{n}{2}-\frac{d}{2}\le B_1+\cdots+B_n<\frac{n}{2}+\frac{d}{2}\right]\\
=\ &\Pr\left[-\frac{d}{\sqrt{n}}\le\frac{(B_1+\cdots+B_n)-n/2}{\sqrt{(1/4)n}}<\frac{d}{\sqrt{n}}\right]\\
\approx\ &\Phi\left(\frac{d}{\sqrt{n}}\right)-\Phi\left(-\frac{d}{\sqrt{n}}\right).
\end{aligned}$$

Hence,

$$\Phi\left(\frac{d}{\sqrt{n}}\right)-\Phi\left(-\frac{d}{\sqrt{n}}\right)=2\Phi\left(\frac{d}{\sqrt{n}}\right)-1\ge 1-\alpha \text{ implies } \frac{d}{\sqrt{n}}\ge\Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

or equivalently,

$$d=r-s\ge\sqrt{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

Distribution-free confidence intervals for quantiles

In other words, to have $(1 - \alpha)$-confident interval for the median is obtained by:

- Obtain n random samples.
- Calculate $d = \sqrt{n}\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.
- Let
$$r = \left\lfloor \frac{n+1}{2} - \frac{d}{2} \right\rfloor$$
and
$$s = \left\lfloor \frac{n+1}{2} - \frac{d}{2} \right\rfloor.$$
- Then the **median** should be between $X_{(r)}$ and $X_{(s)}$ with $(1 - \alpha)$ confidence. Namely, (in terms of normal approximation)
$$\Pr\left[X_{(r)} \le \text{median} < X_{(s)}\right] \ge 1 - \alpha.$$

Distribution-free confidence intervals for quantiles

Example

- Obtain 100 random samples.
- Calculate $d = 10 \cdot \Phi^{-1}\left(1 - \frac{0.05}{2}\right) = 10 \cdot 1.96 = 19.6.$
- Let

$$r = \left\lfloor \frac{101}{2} - \frac{19.6}{2} \right\rfloor = 40$$

  and

$$s = \left\lfloor \frac{101}{2} - \frac{19.6}{2} \right\rfloor = 60.$$

- Then the **median** should be between $X_{(40)}$ and $X_{(60)}$ with 95% confidence. Namely, (in terms of normal approximation)

$$\Pr\left[X_{(40)} \leq \text{median} < X_{(60)}\right] \geq 0.95.$$

Distribution-free confidence intervals for quantiles

We usually estimate mean by $(X_1 + \cdots + X_n)/n$.

But how confident is this estimate?

Rigorously, one should say the mean should lie between

$$\frac{X_1 + \cdots + X_n}{n} - \varepsilon \quad \text{and} \quad \frac{X_1 + \cdots + X_n}{n} + \varepsilon$$

with confidence level at least $(1 - \alpha)$, where

$$\Pr\left[\frac{X_1 + \cdots + X_n}{n} - \varepsilon \leq m < \frac{X_1 + \cdots + X_n}{n} + \varepsilon\right] \geq 1 - \alpha,$$

where m is the true mean.

Distribution-free confidence intervals for quantiles

How to estimate the standard deviation of a distribution?

Answer: In term of *quantile interval* estimate.

Lemma For $q > p$,

$$\Pr\left[X_{(s)} - X_{(r)} \geq Q(q) - Q(p)\right] \geq I_p(r, n-r+1) - I_q(s, n-s+1)$$

and

$$\Pr\left[X_{(v)} - X_{(u)} \leq Q(q) - Q(p)\right] \geq I_q(v, n-v+1) - I_p(u, n-u+1).$$

Proof:

$$\begin{aligned}
\Pr\left[X_{(s)} - X_{(r)} \geq Q(q) - Q(p)\right] &\geq \Pr\left[X_{(s)} \geq Q(q) \wedge X_{(r)} \leq Q(p)\right] \\
&\geq \Pr[X_{(s)} \geq Q(q)] + \Pr[X_{(r)} \leq Q(p)] - 1 \\
&= \Pr[X_{(r)} \leq Q(p)] - \Pr[X_{(s)} < Q(q)] \\
&\geq I_p(r, n-r+1) - I_q(s, n-s+1),
\end{aligned}$$

$$\Pr[X_{(r)} \leq Q(p)] = I_{F(Q(p))}(r, n-r+1) \geq I_p(r, n-r+1) \geq \Pr[X_{(r)} < Q(p)].$$

and therefore,

$$\begin{aligned}\Pr\left[X_{(v)} - X_{(u)} \leq Q(q) - Q(p)\right] &= \Pr\left[X_{(u)} - X_{(v)} \geq Q(p) - Q(q)\right] \\ &\geq I_q(v, n - v + 1) - I_p(u, n - u + 1).\end{aligned}$$

□

Distribution-free confidence intervals for quantiles

Observation For any α, where $0 < \alpha < 1$, there exists one set of integers r, s, u and v for which

$$\Pr\left[X_{(s)} - X_{(r)} \geq Q(q) - Q(p)\right] \geq 1 - \frac{1}{2}\alpha$$

and

$$\Pr\left[X_{(v)} - X_{(u)} \leq Q(q) - Q(p)\right] \geq 1 - \frac{1}{2}\alpha.$$

Therefore,

$$\begin{aligned}
& \Pr\left[X_{(v)} - X_{(u)} \leq Q(q) - Q(p) \leq X_{(s)} - X_{(r)}\right] \\
\geq\ & \Pr\left[X_{(s)} - X_{(r)} \geq Q(q) - Q(p)\right] + \Pr\left[X_{(v)} - X_{(u)} \leq Q(q) - Q(p)\right] - 1 \\
\geq\ & \left(1 - \frac{1}{2}\alpha\right) + \left(1 - \frac{1}{2}\alpha\right) - 1 \\
=\ & 1 - \alpha.
\end{aligned}$$

Distribution-free confidence intervals for quantiles

In the proof of the previous lemma, we actually require:

$$\Pr\left[X_{(s)} \geq Q(q) \wedge X_{(r)} \leq Q(p)\right] \ \geq \ 1 - \frac{1}{2}\alpha.$$

and

$$\Pr\left[X_{(u)} \geq Q(p) \wedge X_{(v)} \leq Q(q)\right] \ \geq \ 1 - \frac{1}{2}\alpha.$$

This can be re-written as:

$$\Pr\left[X_{(s)} \geq Q(q) > Q(p) \geq X_{(r)}\right] \ \geq \ 1 - \frac{1}{2}\alpha.$$

and

$$\Pr\left[Q(q) \geq X_{(v)} \geq X_{(u)} \geq Q(p)\right] \ \geq \ 1 - \frac{1}{2}\alpha.$$

This is why $[X_{(r)}, X_{(s)}]$ and $[X_{(u)}, X_{(v)}]$ are named *outer* and *inner* confidence intervals for the quantile interval $[Q(p), Q(q)]$.

Distribution-free tolerance intervals

Then for any two constants $0 \le \beta, \gamma \le 1$, tolerance interval seeks random variables L and V such that

$$\Pr\left[F(V) - F(L) \ge \gamma\right] \ge \beta.$$

Lemma $\Pr\left[F(V) - F(L) \ge \gamma\right]$ is independent of the parent distribution $F(\cdot)$ if, and only if, L and V are order statistics (such as $X_{(r)}$ and $X_{(s)}$).

In this lemma, L and V are allowed to be $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$.

Idea of the proof.

- $F(X_{(r)})$ and $F(X_{(s)})$ can be viewed as $U_{(r)}$ and $U_{(s)}$, where $U_{(r)}$ and $U_{(s)}$ are simply the order statistics corresponding to a uniform parent distribution in $[0, 1)$.
- As a consequence, (if $F(\cdot)$ has inverse function)

$$\begin{aligned}\Pr[F(X_{(s)}) - F(X_{(r)}) \ge \gamma] &= \Pr[U_{(s)} - U_{(r)} \ge \gamma] \\ &= \Pr[W_{(s,r)} \ge \gamma] \\ &= 1 - I_\gamma(s - r, n - (s - r) + 1).\end{aligned}$$

Distribution-free tolerance intervals

Example Suppose that F has inverse function. For $r = 1$ and $s = n$, we have

$$\begin{aligned}
\Pr[F(X_{(n)}) - F(X_{(1)}) \geq \gamma] &= \Pr[U_{(n)} - U_{(1)} \geq \gamma] \\
&= \Pr[W_{(1,n)} \geq \gamma] \\
&= 1 - I_\gamma(n-2, 2) \\
&= 1 - \frac{\int_0^\gamma z^{n-2}(1-z)dz}{\int_0^1 z^{n-2}(1-z)dz} \\
&= 1 - \frac{\frac{1}{n-1}\gamma^{n-1} - \frac{1}{n}\gamma^n}{\frac{1}{n-1} - \frac{1}{n}} \\
&\geq \beta,
\end{aligned}$$

which is equivalent to:

$$n\gamma^{n-1} - (n-1)\gamma^n \leq 1 - \beta.$$

With the above inequality, we can solve "**how large** n **should be to satisfy it**?" For example, $\gamma = 0.95$ and $\beta = 0.9$, the minimum n to satisfy the above inequality is 77.

Conditional distribution of order statistics

Premise: $1 \le r < s \le n$

We already know that for $y \ge x$,

$$\begin{aligned} & f_{(r,s)}(x,y) \\ = \; & \frac{n!}{(r-1)! \cdot (s-r-1)! \cdot (n-s)!} F^{r-1}(x) f(x) [F(y)-F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s}, \end{aligned}$$

and

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x) [1-F(x)]^{n-r}.$$

This implies that

$$\begin{aligned} f_{X_{(s)}|X_{(r)}}(y|x) &= \frac{f_{X_{(r,s)}}(x,y)}{f_{X_{(r)}}(x)} = \frac{(n-r)!}{(s-r-1)!(n-s)!} \frac{[F(y)-F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s}}{[1-F(x)]^{n-r}} \\ &= \frac{(n-r)!}{((s-r)-1)!((n-r)-(s-r))!} \\ & \quad \times \left(\frac{F(y)-F(x)}{1-F(x)} \right)^{(s-r)-1} \left(1 - \frac{F(y)-F(x)}{1-F(x)} \right)^{(n-r)-(s-r)} \left(\frac{f(y)}{1-F(x)} \right) \end{aligned}$$

Conditional distribution of order statistics

Observation $f_{X_{(s)}|X_{(r)}}(y|x)$ over population of size n with parent density $f(\cdot)$ is nothing but $f_{\bar{X}_{(s-r)}}(\cdot)$ over population of size $(n-r)$ with parent density

$$f^{\diamond}(y) = \begin{cases} \dfrac{f(y)}{1-F(x)}, & \text{for } y \geq x; \\ 0, & \text{for } y < x \end{cases}$$

Marcovian of order statistics

Premise: $1 \leq n_1 < n_2 < \cdots < n_k \leq n$

We already know that for $x_1 \leq x_2 \leq \cdots \leq x_k$,

$$f_{(n_1,\ldots,n_k)}(x_1,\ldots,x_k) = n! \left[\prod_{j=1}^{k} f(x_j)\right] \left[\prod_{j=0}^{k} \frac{[F(x_{j+1}) - F(x_j)]^{n_{j+1}-n_j-1}}{(n_{j+1}-n_j-1)!}\right],$$

where $x_0 = -\infty$, $x_{k+1} = \infty$, $n_0 = 0$ and $n_{k+1} = n+1$.

We can similarly prove that:

$$f_{X_{(s)}|X_{(r)},X_{(r-1)},\ldots,X_{(1)}}(y|x_{(r)},x_{(r-1)},\ldots,x_{(1)}) = f_{X_{(s)}|X_{(r)}}(y|x_{(r)})$$

Observation $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ forms a first-order Markov chain for a parent distribution with density.

Marcovian of order statistics

Example (implication of Markovian) Suppose the parent density is e^{-x} for $x \geq 0$.

Then the joint distribution of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ for $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n$ is given by:

$$\begin{aligned}
f_{(1,\ldots,n)}(x_1, \ldots, x_n) &= n! \left[\prod_{j=1}^{n} f(x_j)\right] \left[\prod_{j=0}^{n} \frac{[F(x_{j+1}) - F(x_j)]^{(j+1)-j-1}}{((j+1)-j-1)!}\right] \\
&= n! \left[\prod_{j=1}^{n} e^{-x_j}\right] \\
&= n! \exp\left\{-\sum_{j=1}^{n} x_j\right\}.
\end{aligned}$$

Observe that with $x_0 = 0$,

$$\sum_{j=1}^{n}(n-j+1)(x_j - x_{j-1}) = \left\{\begin{array}{lcc} & n & (x_1 - x_0) \\ + & (n-1) & (x_2 - x_1) \\ + & (n-2) & (x_3 - x_2) \\ & & \cdots \\ + & & (x_n - x_{n-1}) \end{array}\right\} = \sum_{j=1}^{n} x_j.$$

Marcovian of order statistics

Hence,

$$
\begin{aligned}
f_{(1,\ldots,n)}(x_1,\ldots,x_n) &= n!\,\exp\left\{-\sum_{j=1}^{n}(n-j+1)(x_j-x_{j-1})\right\} \\
&= n!\prod_{j=1}^{n}\exp\left\{-(n-j+1)(x_j-x_{j-1})\right\}
\end{aligned}
$$

By defining $Y_j = (n-j+1)(X_{(j)} - X_{(j-1)})$, where $X_{(0)} = 0$. I.e.,

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{n-1} \\ Y_n \end{bmatrix}
=
\begin{bmatrix}
n & 0 & 0 & \cdots & 0 & 0 \\
-(n-1) & (n-1) & 0 & \cdots & 0 & 0 \\
0 & -(n-2) & (n-2) & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
0 & 0 & 0 & \cdots & 2 & 0 \\
0 & 0 & 0 & \cdots & -1 & 1
\end{bmatrix}
\begin{bmatrix} X_{(1)} \\ X_{(2)} \\ X_{(3)} \\ \vdots \\ X_{(n-1)} \\ X_{(n)} \end{bmatrix},
$$

which gives that

$$
f(y_1,\ldots,y_n) = \prod_{i=1}^{n}\exp\{-y_i\} \text{ for each } y_i \in [0,\infty).
$$

This immediately implies that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with exponential parent density.

Marcovian of order statistics

Notably,

$$f_{(1,\ldots,n)}(x_1,\ldots,x_n) = \begin{cases} n! \prod_{j=1}^{n} \exp\{-x_j\}, & \text{for } x_1 \le x_2 \le \cdots \le x_n; \\ 0, & \text{otherwise.} \end{cases}$$

does not mean that $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are i.i.d., even if the pdf is a "product form".

Observation 1 In this example, first-order Markovian of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ allows us to transform it to an i.i.d. sequence $Y_1, Y_2, \ldots, Y_n$, where

$$Y_i = (n-i+1)(X_{(i)} - X_{(i-1)})$$

or equivalently

$$X_{(i)} = X_{(i-1)} + \frac{Y_i}{n-i+1}.$$

This indicates that $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ forms an **additive** Markov chain.

Observation 2 In this example,

$$X_{(r)} = \sum_{i=1}^{r}(X_{(i)} - X_{(i-1)}) = \sum_{i=1}^{r} \frac{Y_i}{n-i+1}.$$

Marcovian of order statistics

Example Suppose the parent density of $U_{(1)}, \ldots, U_{(n)}$ is uniformly distributed over $(0, 1]$.

Then $-\log U_{(n)}, \ldots, -\log U_{(1)}$ forms order statistics with exponential parent density, where $-\log U_{(n)} \le \ldots \le -\log U_{(1)}$.

$$\Pr[-\log U \le x] \;=\; \Pr[U \ge e^{-x}] = 1 - e^{-x}.$$

The previous example then suggests:

$$Y_{n-i+1} = i\left[\left(-\log U_{(i)}\right) - \left(-\log U_{(i+1)}\right)\right] = i \log \frac{U_{(i+1)}}{U_{(i)}}$$

is i.i.d., where $U_{(0)} = 1$.

This implies that

$$\left(\frac{U_{(i+1)}}{U_{(i)}}\right)^{i} = \exp\left\{Y_{n-i+1}\right\}$$

is also i.i.d.

Observation $U_{(i+1)} = U_{(i)} \cdot \sqrt[i]{Z_i}$ forms a multiplicative Markov chain, where $\{Z_i\}$ is i.i.d.

Marcovian of order statistics

Example Suppose the parent distribution of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is standard normal distributed.

Then the joint distribution of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ for $0 \le x_1 \le x_2 \le \cdots \le x_n$ is given by:

$$
\begin{aligned}
f_{(1,\ldots,n)}(x_1, \ldots, x_n) &= n! \left[\prod_{j=1}^{n} f(x_j)\right] \left[\prod_{j=0}^{n} \frac{[F(x_{j+1}) - F(x_j)]^{(j+1)-j-1}}{((j+1)-j-1)!}\right] \\
&= n! \left[\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_j^2/2}\right] \\
&= \frac{n!}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n} x_j^2\right\}.
\end{aligned}
$$

Observe that with $x_0 = 0$,

$$
\sum_{j=1}^{n}(n-j+1)(x_j^2 - x_{j-1}^2) = \left\{
\begin{array}{rcc}
 & n & (x_1^2 - x_0^2) \\
+ & (n-1) & (x_2^2 - x_1^2) \\
+ & (n-2) & (x_3^2 - x_2^2) \\
 & & \cdots \\
+ & & (x_n^2 - x_{n-1}^2)
\end{array}
\right\} = \sum_{j=1}^{n} x_j^2.
$$

Marcovian of order statistics

Hence,

$$f_{(1,\ldots,n)}(x_1,\ldots,x_n) = \frac{n!}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(n-j+1)(x_j^2-x_{j-1}^2)\right\}$$
$$= n!\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(n-j+1)}{2}(x_j^2-x_{j-1}^2)\right\}$$

By defining

$$Y_j = S_j\sqrt{\frac{(X_{(j)}^2-X_{(j-1)}^2)}{1/(n-j+1)}},$$

where $X_{(0)} = 0$ and $\Pr[S_j = +1] = \Pr[S_j = -1] = 1/2$ and $\{S_j\} \perp\!\!\!\perp \{X_j\}$, we obtain:

$$f(y_1,\ldots,y_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\{-y_i^2/2\}.$$

This immediately implies that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with standard normal parent density.

Marcovian of order statistics

Well-known property of i.i.d. (standard normal) Gaussian:

1. $\bar{X} \perp\!\!\!\perp (X_i - \bar{X})$ for every i.

   Proof: $X_i - \bar{X}$ and $\bar{X}$ are jointly Gaussian distributed. Hence, uncorrelation implies independence between them.

$$\begin{aligned} E\left[(X_i - \bar{X})\bar{X}\right] &= E\left[\left(X_i - \frac{1}{n}\sum_{j'=1}^{n} X_{j'}\right)\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)\right] \\ &= E\left[\frac{1}{n}\sum_{j=1}^{n} X_i X_j - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n} X_j X_{j'}\right] \\ &= \frac{1}{n} - \frac{1}{n^2} n \\ &= 0 = E[(X_i - \bar{X})]E[\bar{X}]. \end{aligned}$$

2. $\bar{X}$ is independent of any function of $\{(X_i - \bar{X})\}_{i=1}^{n}$, such as range $W_{(1,n)} = \max_{1 \le j \le n}(X_j - \bar{X}) - \min_{1 \le j \le n}(X_j - \bar{X})$.

3. $\bar{X}$ is independent of $W_{(r,s)} = X_{(s)} - X_{(r)}$.

Independent non-identically distributed variables

Assumption Now suppose $X_1, X_2, \ldots, X_n$ are only *independent*, but not necessarily identically distributed.
Denote their distributions by $F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot)$, respectively.

Then

$$\begin{aligned}
F_{(n)}(x) &= \Pr[X_{(n)} \leq x] \\
&= \Pr[\max_{1 \leq i \leq n} X_n \leq x] \\
&= \Pr[X_1 \leq x \wedge \cdots \wedge X_n \leq x] \\
&= \Pr[X_1 \leq x] \cdots \Pr[X_n \leq x] \\
&= \prod_{i=1}^{n} F_i(x).
\end{aligned}$$

Likewise,

$$\begin{aligned}
F_{(1)}(x) &= \Pr[X_{(1)} \leq x] \\
&= 1 - \Pr[X_{(1)} > x] \\
&= 1 - \Pr[\min_{1 \leq i \leq n} X_n > x] \\
&= 1 - \prod_{i=1}^{n} (1 - F_i(x)).
\end{aligned}$$

Independent non-identically distributed variables

$$
\begin{aligned}
F_{(r)}(x) &= \Pr[X_{(r)} \leq x] \\
&= \Pr[\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x] \\
&= \sum_{i=r}^{n} \sum_{\{(j_1,\ldots,j_n)\in\mathbb{P}_n \,:\, j_1<\cdots<j_i \text{ and } j_{i+1}<\cdots<j_n\}} \prod_{\ell=1}^{i} F_{j_\ell}(x) \prod_{\ell=i+1}^{n} [1 - F_{j_\ell}(x)],
\end{aligned}
$$

where the set $\mathbb{P}_n$ consists of all permutations of $(1, 2, \ldots, n)$.

Independent non-identically distributed variables

Theorem (Sen 1970) Define $\bar{F}(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$.

1. For all real y,

$$\Pr\left[X_{(1)} \le y \middle\| (F_1, F_2, \ldots, F_n)\right] \ge \Pr\left[X_{(1)} \le y \middle\| (\bar{F}, \bar{F}, \ldots, \bar{F})\right]$$

   with equality holding only if $F_1(y) = F_2(y) = \cdots = F_n(y) = \bar{F}(y)$.

2. For integer $2 \le r \le n-1$, real x satisfying $\bar{F}(x) \le (r-1)/n$ and real y satisfying $\bar{F}(y) \ge r/n$,

$$\Pr\left[x < X_{(r)} \le y \middle\| (F_1, F_2, \ldots, F_n)\right] \ge \Pr\left[x < X_{(r)} \le y \middle\| (\bar{F}, \bar{F}, \ldots, \bar{F})\right],$$

   with equality holding only if $F_1(x) = F_2(x) = \cdots = F_n(x) = \bar{F}(x)$ and $F_1(y) = F_2(y) = \cdots = F_n(y) = \bar{F}(y)$.

3. For all real y,

$$\Pr\left[X_{(n)} \le y \middle\| (F_1, F_2, \ldots, F_n)\right] \le \Pr\left[X_{(n)} \le y \middle\| (\bar{F}, \bar{F}, \ldots, \bar{F})\right]$$

   with equality holding only if $F_1(y) = F_2(y) = \cdots = F_n(y) = \bar{F}(y)$.

Independent non-identically distributed variables

Lemma (Hoeffding) Let p_i be the probability of success at the ith trial, and suppose each trial is independent.
Denote by S the number of success after n trials.
Then

$$\Pr\left[S \le c \| (p_1, p_2, \ldots, p_n)\right] \le \Pr\left[S \le c \| (\bar{p}, \bar{p}, \ldots, \bar{p})\right] \quad \text{if } 0 \le c \le n\bar{p} - 1,$$

and

$$\Pr\left[S \le c \| (p_1, p_2, \ldots, p_n)\right] \ge \Pr\left[S \le c \| (\bar{p}, \bar{p}, \ldots, \bar{p})\right] \quad \text{if } n\bar{p} \le c \le n,$$

where $\bar{p} = (p_1 + p_2 + \cdots + p_n)/n$, provided that c is an integer.
• Notably, $E[S] = n\bar{p}$ is the margin point.

Proof of Sen's Theorem: We first prove Case 2 (in terms of Hoeffding's Lemma).

Define a success at the ith trial to be $[X_i \le y]$.

Independent non-identically distributed variables

Then

$$\begin{aligned}
&\Pr[X_{(r)} \le y \| (F_1, \ldots, F_n)] \\
= \;& \Pr[S > r-1 \| (F_1(y), \ldots, F_n(y))] \\
= \;& 1 - \Pr[S \le r-1 \| (F_1(y), \ldots, F_n(y))] \\
&\begin{cases} \ge 1 - \Pr[S \le r-1 \| (\bar{F}(y), \ldots, \bar{F}(y))], & \text{if } 0 \le r-1 \le n\bar{F}(y) - 1 \\ \le 1 - \Pr[S \le r-1 \| (\bar{F}(y), \ldots, \bar{F}(y))], & \text{if } n\bar{F}(y) \le r-1 \le n \end{cases} \\
&\begin{cases} \ge \Pr[S > r-1 \| (\bar{F}(y), \ldots, \bar{F}(y))], & \text{if } \underbrace{1 \le r}_{\text{always valid}} \le n\bar{F}(y) \\ \le \Pr[S > r-1 \| (\bar{F}(y), \ldots, \bar{F}(y))], & \text{if } n\bar{F}(y) + 1 \le \underbrace{r \le n+1}_{\text{always valid}} \end{cases} \\
&\begin{cases} \ge \Pr[X_{(r)} \le y \| (\bar{F}, \ldots, \bar{F})], & \text{if } \bar{F}(y) \ge r/n \\ \le \Pr[X_{(r)} \le y \| (\bar{F}, \ldots, \bar{F})], & \text{if } \bar{F}(y) \le (r-1)/n \end{cases} . \qquad (1)
\end{aligned}$$

Hence, when $\bar{F}(x) \le (r-1)/n$ and $\bar{F}(y) \ge r/n$ and $r = 2, \ldots, n-1$,

$$\begin{aligned}
&\Pr[x < X_{(r)} \le y \| (F_1, \ldots, F_n)] \\
= \;& \Pr[X_{(r)} \le y \| (F_1, \ldots, F_n)] - \Pr[X_{(r)} \le x \| (F_1, \ldots, F_n)] \\
\ge \;& \Pr[X_{(r)} \le y \| (\bar{F}, \ldots, \bar{F})] - \Pr[X_{(r)} \le x \| (\bar{F}, \ldots, \bar{F})] \\
= \;& \Pr[x < X_{(r)} \le y \| (\bar{F}, \ldots, \bar{F})].
\end{aligned}$$

Independent non-identically distributed variables

Inequality (1) has already proved that

$$\Pr[X_{(1)} \le y \| (F_1, \ldots, F_n)] \ge \Pr[X_{(1)} \le y \| (\bar{F}, \ldots, \bar{F})] \text{ for } \bar{F}(y) \ge 1/n$$

and

$$\Pr[X_{(n)} \le y \| (F_1, \ldots, F_n)] \le \Pr[X_{(n)} \le y \| (\bar{F}, \ldots, \bar{F})] \text{ for } \bar{F}(y) \le (n-1)/n.$$

Here, we need to further prove their validity for all $y \in \Re$.

The other two cases can be proved as follows.

$$\begin{aligned}
\Pr[X_{(n)} \le y \| (F_1, \ldots, F_n)] &= \prod_{i=1}^{n} F_i(y) \\
&\le \left[\frac{1}{n} \sum_{i=1}^{n} F_i(y) \right]^n \quad \text{(Geometric mean} \le \text{arithmetic mean)} \\
&= \bar{F}^n(y) \\
&= \Pr[X_{(n)} \le y \| (\bar{F}, \ldots, \bar{F})],
\end{aligned}$$

Independent non-identically distributed variables

and

$$\begin{aligned}
\Pr[X_{(1)} \le y \| (F_1, \ldots, F_n)] &= 1 - \Pr[X_{(1)} > y \| (F_1, \ldots, F_n)] \\
&= 1 - \prod_{i=1}^{n} (1 - F_i(y)) \\
&\ge 1 - \left[\frac{1}{n} \sum_{i=1}^{n} (1 - F_i(y)) \right]^n \\
&= 1 - \left[1 - \bar{F}(y) \right]^n \\
&= \Pr[X_{(1)} \le y \| (\bar{F}, \ldots, \bar{F})].
\end{aligned}$$

□

Lemma (Sen 1970)

$$\left| \text{median}(X_{(r)} \| (F_1, \ldots, F_n)) - \text{median}(X_{(r)} \| (\bar{F}, \ldots, \bar{F})) \right| \le q_r - q_{r-1},$$

provided that q_r and q_{r-1} are uniquely defined by $\bar{F}(q_r) = r/n$ and $\bar{F}(q_{r-1}) = (r-1)/n$, where median$(Z)$ denotes the median of random variable Z.

$f_{(r,s)}(x, y)$ for independent non-identical densities

Suppose $F_1, \dots, F_n$ have densities $f_1, \dots, f_n$.

$$f_{(r,s)}(x, y \| (F_1, \dots, F_n)) = \frac{1}{(r-1)!(s-r-1)!(n-s)!}$$

$$\times \begin{vmatrix} F_1(x) & F_2(x) & \cdots & F_n(x) \\ \vdots & \vdots & \cdots & \vdots \\ F_1(x) & F_2(x) & \cdots & F_n(x) \\ f_1(x) & f_2(x) & \cdots & f_n(x) \\ F_1(y)-F_1(x) & F_2(y)-F_2(x) & \cdots & F_n(y)-F_n(x) \\ \vdots & \vdots & \cdots & \vdots \\ F_1(y)-F_1(x) & F_2(y)-F_2(x) & \cdots & F_n(y)-F_n(x) \\ f_1(y) & f_2(y) & \cdots & f_n(y) \\ 1-F_1(y) & 1-F_2(y) & \cdots & 1-F_n(y) \\ \vdots & \vdots & \cdots & \vdots \\ 1-F_1(y) & 1-F_2(y) & \cdots & 1-F_n(y) \end{vmatrix},$$

where
there are $(r-1)$ rows of $F_1(x), F_2(x), \dots, F_n(x)$,
there are $(s-r-1)$ rows of $F_1(y) - F_1(x), F_2(y) - F_2(x), \cdots, F_n(y) - F_n(x)$, and
there are $(n-s)$ rows of $1 - F_1(y), 1 - F_2(y), \cdots, 1 - F_n(y)$.