Section 4.4

Day 7 (1/29)

1 Order statistics

Suppose that $(x_1, x_2, x_3) = (10.5, 3.7, 6.0)$. The order statistics are $(y_1, y_2, y_3) = (3.7, 6.0, 10.5)$, i.e. the order statistics are $(Y_1, Y_2, \ldots, Y_n)$ such that

$$Y_1 \leq Y_2 \leq \cdots \leq Y_n$$

Theorem 4.4.1 Let $X_1, \ldots, X_n$ be a random sample from a population with continuous density $f(\cdot)$. The joint density of the sample is

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n), \quad x_1 \epsilon \mathcal{D}, x_2 \epsilon \mathcal{D}, \ldots, x_n \epsilon \mathcal{D}$$

The joint density of the order statistics is

$$f(y_1, y_2, \ldots, y_n) = n!\, f(y_1)f(y_2)\cdots f(y_n), \quad y_1 \leq y_2 \leq \cdots \leq y_n$$

Proof: Skip

Example Let $X_1, X_2 \sim \text{Unif}(0,1)$

1. Calculate $P[X_1 > 0.5]$ and $P[Y_1 > 0.5]$
2. On the average, Y_1 should be around ________ give or take ________ or so.
   (R simulation here.)

Solution:

1. $P[X_1 > .5] = \int_0^{.5} 1\, dx = 0.5$

   $P[Y_1 > .5] = \int_{.5}^1 \int_{.5}^{y_2} 2\, dy_1 dy_2 = \int_{.5}^1 2(y_2 - .5)\, dy_2 = 2\left(\frac{y_2^2}{2} - .5y_2\right)\Big]_{.5}^1 = 0.25$

   or

   $P[Y_1 > .5] = P[\min\{X_1, X_2\} > .5] = P[\text{Both} > .5] = (.5)(.5) = .25$

2. To find the marginal pdf of Y_1, integrate the joint pdf of (Y_1, Y_2). For $0 \leq t \leq 1$,

$$g_{y_1}(t) = \int g(t, y_2)\, dy_2 = \int_t^1 2!\, dy_2 = \begin{cases} 2(1-t) & 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   Then

$$E(Y_1) = \int_0^1 t\, 2(1-t)dt = 2\left[\frac{t^2}{2} - \frac{t^2}{3}\right]_0^1 = 1/3$$

$$E(Y_1^2) = \int_0^1 t^2\, 2(1-t)dt = 2\left[\frac{t^3}{3} - \frac{t^4}{4}\right]_0^1 = 1/6$$

   so that

$$\sigma_{Y_1} = \sqrt{E(Y_1^2) - (E(Y_1)^2} = \sqrt{1/6 - (1/3)^2} = .2357$$

Theorem Marginal density of Y_k

Let $Y_1, Y_2, \ldots, Y_n$ be order statistics from a population with continuous density $f(\cdot)$ and cdf $F(\cdot)$. For $k = 1, \ldots, n$, the order statistic Y_k has marginal density

$$g_k(y_k) = \int \cdots \int g(y_1, y_2, \ldots, y_n)\, dy_1 dy_2 \ldots dy_{k-1} dy_{k+1} \ldots dy_n$$

$$\vdots$$

$$\vdots$$

$$= \begin{cases} \frac{n!}{(k-1)!(n-k)!} [F(y_k)]^{k-1}\ f(y_k)\ [1 - F(y_k)]^{n-k} & \text{if } y_k \epsilon \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Proof: Skip

Example

Let $Y_1 \le Y_2 \le Y_3 \le Y_4$ be order statistics from $f(x) = \begin{cases} 2x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$

Find $P[Y_3 > .5]$.

Solution:

$P[X \le t] = \int_0^t 2x\, dx = 2\ \frac{x^2}{2}\Big]_0^t = t^2$ so that $F(t) = \begin{cases} 0, & t < 0 \\ t^2, & 0 \le t \le 1 \\ 1, & 1 < t \end{cases}$

Using the theorem, Y_3 has marginal density

$$g_{y_3}(t) = \frac{4!}{2!1!} \left[t^2\right]^2\ 2t\ \left[1 - t^2\right] = 24\ (t^5 - t^7)$$

for $0 \le t \le 1$ so that

$$P[Y_3 > .5] = \int_{.5}^1 24(t^5 - t^7)dt = \cdots = \frac{243}{256}$$

Theorem Joint density of (Y_i, Y_j)

Let $Y_1, Y_2, \ldots, Y_n$ be order statistics from a population with continuous density $f(\cdot)$ and cdf $F(\cdot)$. For $k = 1, \ldots, n$, the order statistics (Y_j, Y_k) have joint density

$$g(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(y_i)]^{i-1}\ f(y_i)\ [F(y_j) - F(y_i)]^{j-i-1} f(y_j)[1 - F(y_j)]^{n-j}$$

if $(y_i, y_j) \epsilon \mathcal{D}$ and is 0 otherwise.

Proof: Skip

Example

Let $Y_1 \le Y_2 \le Y_3$ be order statistics from Unif(0,1).

1. Find the pdf of the range $R = Y_3 - Y_1$.

2. The value of the range should be around ________ give or take ________ or so.

   (R simulation here)

Solution:

1. First we find the joint distribution of (Y_1, Y_3).

$$g(y_1, y_3) = \frac{3!}{(0)!(1)!(0)!}[F(y_1)]^0 \; f(y_1) \; [F(y_3) - F(y_1)]^1 f(y_3)[1 - F(y_3)]^0$$

For the Unif(0,1) distribution $f(t)) = 1$ and $F(t) = t$ for $0 \leq t \leq 1$), so

$$g(y_1, y_3) = 6 \; (y_3 - y_1)$$

for $0 \leq y_1 \leq y_3 \leq 1$ and 0 otherwise. Then the range R has cdf

$$\begin{aligned} F_R(t) &= P[Y_3 - Y_1 \leq t] = 1 - P[Y_3 - Y_1 \geq t] = 1 - \int_t^1 \int_0^{y_3 - t} 6 \; (y_3 - y_1) \; dy_1 dy_3 \\ &= 1 - \int_t^1 3(y_3^2 - t^2) dy_3 \\ &= 1 - \left(2t^3 - 3t^2 + 1\right)\big) \\ &= 3t^2 - 2t^3 \end{aligned}$$

and pdf

$$f_R(t) = \frac{d}{dt} F_R(t) = \begin{cases} 6t(1-t), & 0 \leq t \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

2. The mean and standard deviation of R are

$$E(R) = \int_0^1 t f_R(t) \; dt = \int_0^1 t \; 6t(1-t) \; dt = \int_0^1 6(t^2 - t^3) \; dt = 6 \left[\frac{t^3}{3} - \frac{t^4}{4}\right]_0^1 = .50$$

$$E(R^2) = \int_0^1 t^2 \; 6t(1-t) \; dt = \int_0^1 6(t^3 - t^4) \; dt = 6 \left[\frac{t^4}{4} - \frac{t^5}{5}\right]_0^1 = .30$$

$$\sigma_R = \sqrt{E(R^2) - [E(R)]^2} = \sqrt{.30 - .50^2} = .2236$$

```
> # Simulating the minimum of two order statistics using replicate()
>
> runif(2,0,1)
[1] 0.6084405 0.5570129
> min(runif(2,0,1))
[1] 0.3109178
> a<-replicate(50, min(runif(2,0,1)))
> a
 [1] 0.764304088 0.149802179 0.215961420 0.377162105 0.375036888 0.543548201
 [7] 0.769198446 0.534621224 0.015304092 0.514847347 0.493023768 0.123526617
[13] 0.027832255 0.761034889 0.067350545 0.038292981 0.347922798 0.221848028
[19] 0.029371039 0.255063091 0.587119930 0.125518252 0.678753041 0.314997050
[25] 0.543066891 0.029599997 0.581367477 0.306087146 0.056177725 0.318066571
[31] 0.352271149 0.083896117 0.635335865 0.074246769 0.196533200 0.388689839
```

```
[37] 0.613574251 0.300420767 0.614536128 0.003827966 0.098701683 0.506094062
[43] 0.539256820 0.419528263 0.154893861 0.259439785 0.274287163 0.320029600
[49] 0.044214869 0.011988704
> mean(a)
[1] 0.3211515
> sd(a)
[1] 0.2315983
```

Simulating the range of three order statistics using replicate()

```
> temp1<-runif(3,0,1)
> r1<-max(temp1)-min(temp1)
> temp1
[1] 0.7369862 0.6046055 0.6482580
> r1
[1] 0.1323807
> a2<-replicate(50,
+ {temp1<-runif(3,0,1)
+ r1<-max(temp1)-min(temp1)
+ }
+ )
> mean(a2)
[1] 0.5380863
> sd(a2)
[1] 0.2281919
```

Simulating the range of three order statistics using matrix and apply()

```
> set.seed(4321)
> m1<-matrix(runif(3*50,0,1),ncol=3)
> a3<-apply(m1,1,max)-apply(m1,1,min)
> mean(a3)
[1] 0.4856879
> sd(a3)
[1] 0.2171406
```

Simulating the range of three order statistics using for() loop

```
> a4<-c(rep(0,50))
> for (i in 1:50){
+ temp1<-runif(3,0,1)
+ a4[i]<-max(temp1)-min(temp1)
+ }
> mean(a4)
[1] 0.5030461
> sd(a4)
[1] 0.2243892
```