Chap 5. Random Sample (랜덤샘플)

정의. 랜덤샘플(random sample) :
확률변수의 모임 $\{X_1, X_2, \ldots, X_n\}$이 서로 독립이며 동일한 분포를 따를 때 이를 랜덤샘플이라 함. That is, random sample is independently and identically distributed. (*i.i.d*)

5.1 Sample mean and sample variance (표본평균과 표본분산)

표본평균(sample mean) : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

표본분산(sample variance) : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Thm 5.1 $X_1, X_2, \ldots, X_n$이 $E(X_i) = \mu$, $Var(X_i) = \sigma^2$인 모집단으로부터 뽑은 랜덤표본

(i) $E(\overline{X}) = \mu$, $Var(\overline{X}) = \frac{\sigma^2}{n}$

(ii) $E(S^2) = \sigma^2$, i.e., S^2 은 σ^2 의 불편추정량(unbiased estimator)

(proof)

Since $X_1, X_2, \ldots, X_n$ are i.i.d,

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$Var(\overline{X}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

Note that

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

Then,

$$E(S^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right)$$

$$E(X_i^2) = Var(X_i) + (E(X_i))^2 = \sigma^2 + \mu^2,\quad E(\overline{X}^2) = Var(\overline{X}) + (E(\overline{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Hence,

$$E(S^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(\mu^2+\sigma^2) - n\left(\mu^2+\frac{\sigma^2}{n}\right)\right)$$
$$= \sigma^2$$

5.2 Some probability distributions

1) Chi-square distribution (카이제곱분포)

Two important subfamilies of Gamma distribution ($X \sim \text{Gamma}(\alpha, \frac{1}{\lambda})$)

① 카이제곱분포($X \sim \chi^2(r) \equiv Gamma(\frac{r}{2}, 2)$)

When $\alpha = \frac{r}{2}$, $\frac{1}{\lambda} = \frac{1}{2}$ (r : positive integer, degree of freedom)

$$f(x|r) = \frac{1}{\Gamma(\frac{r}{2})2^{r/2}} x^{r/2-1} e^{-x/2},\ x>0$$

$E(X) = r/2 \times 2 = r$

$Var(X) = r/2 \times 2^2 = 2r$

② 지수분포($X \sim \text{Exp}(\frac{1}{\lambda})$)

When $\alpha = 1$, we called exponential distribution.

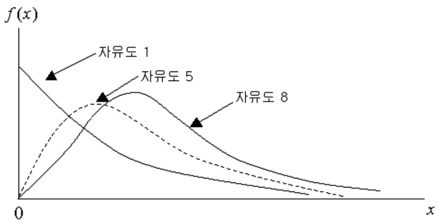
$$f(x) = \lambda e^{-\lambda x},\ x > 0$$

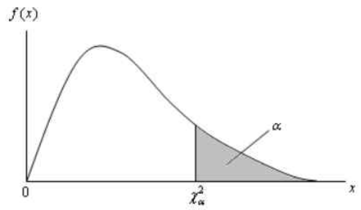
카이제곱분포

- 정의 : $Z \sim N(0,1)$일 때, $Z^2 \sim \chi^2(1)$ ([예 2.16] 참고).
- 성질 : $Z_1, Z_2, \ldots, Z_k \overset{iid}{\sim} N(0,1)$일 때, $\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$ (가법성. additivity)
- 활용 예 : $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$이라 할 때,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \text{ (See Thm 5.6)}$$

- 카이제곱분포의 확률밀도함수 형태 및 상위 $(100\times\alpha)$백분위수 :





$\chi^2_\alpha(r)$: $X \sim \chi^2(r)$일 때, $P(X \geq x)=\alpha$를 만족하는 x를 자유도가 r인 카이제곱분포의 상위 $(100\times\alpha)$ 백분위수

Thm 5.2 Additivity of Chi-square distn. (카이제곱분포의 가법성의 일반화):

$X_1 \sim \chi^2(r_1),\ X_2 \sim \chi^2(r_2)$: indep. Then, $X_1+X_2 \sim \chi^2(r_1+r_2)$

(Proof) Method 1. Recall the distribution of X_1+X_2
when $X_1 \sim \mathrm{Gamma}(\alpha_1,\beta),\ X_2 \sim \mathrm{Gamma}(\alpha_2,\beta)$

Method 2. Use mgf technique.
Let $Y=X_1+X_2$.

The mgf of X_1 & X_2 : $M_{X_1}(t)=(1-2t)^{\frac{r_1}{2}}$

$M_{X_2}(t)=(1-2t)^{\frac{r_2}{2}}$.

Then, the mgf of Y : $M_Y(t)=(1-2t)^{\frac{r_1+r_2}{2}}$

$\therefore\ Y\sim\chi^2(r_1+r_2)$

Thm 5.3 Auxiliary result

X_1 & X_2 are indep. Let $Y=X_1+X_2$ with $Y\sim\chi^2(m),\ X_1 \sim \chi^2(n)$.

Then, $X_2 \sim \chi^2(m-n)$.

2) t-distribution ($X \sim t(r)$)

Def 5.2 $Z \sim N(0,1)$, $U \sim \chi^2(r)$: indep.

$$\Rightarrow T = \frac{Z}{\sqrt{U/r}} \sim t(r)$$

Properties of t-distribution :

- pdf of $t(r)$: $f(x) = \dfrac{\Gamma((r+1)/2)}{\sqrt{r\pi}\,\Gamma(r/2)}\left(1+\dfrac{x^2}{r}\right)^{-\frac{r+1}{2}}, \quad -\infty < x < \infty$
- Symm. w.r.t 0, $t(1) \sim C(0,1)$ (코쉬분포)
- Shape : See Figure 5.1 (Close to standard normal as d.f increases)

Derivation of pdf of $t(r)$

$$f(z,u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{\Gamma(\frac{r}{2})2^{r/2}} u^{r/2-1} e^{-u/2}$$

Let $T = \dfrac{Z}{\sqrt{U/r}}$ and $W = U$ $(-\infty < T < \infty,\ W > 0)$

Then, $Z = T\sqrt{W/r}$, $U = W$

$$J = \begin{pmatrix} \sqrt{w/r} & \frac{1}{2}t\frac{1}{\sqrt{w}\sqrt{r}} \\ 0 & 1 \end{pmatrix} \Longrightarrow |J| = \sqrt{w/r}.$$

$$\therefore f(t,w) = \frac{1}{\sqrt{2\pi}} exp(t^2 w/2r) \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} \exp(-w/2)\sqrt{w/r}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(r/2)2^{r/2}} \frac{1}{\sqrt{r}} w^{\frac{r+1}{2}-1} \exp[-(\frac{t^2}{2r}+\frac{1}{2})w]$$

Hence, marginal dist of T :

$$f(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(r/2)2^{r/2}} \frac{1}{\sqrt{r}} \int_0^\infty w^{\frac{r+1}{2}-1} \exp\left[-(\frac{t^2}{2r}+\frac{1}{2})w\right] dw$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(r/2)2^{r/2}} \frac{1}{\sqrt{r}} \Gamma(\frac{r+1}{2})(\frac{2}{1+t^2/r})^{\frac{r+1}{2}}$$

$$= \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})} \frac{1}{\sqrt{\pi r}} \frac{1}{(1+\frac{t^2}{r})^{\frac{r+1}{2}}}, \quad -\infty < t < \infty.$$

3) F-distribution ($X \sim F(m,n)$)

Def 5.2 $U \sim \chi^2(m),\ V \sim \chi^2(n)$: indep.

$$\Rightarrow F = \frac{U/m}{V/n} \sim F(m,n)$$

Properties of F-distribution :

- pdf of $F(m,n)$:

$$f(x) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} x^{m/2-1} \left(1+\frac{m}{n}x\right)^{-(m+n)/2}, \ x \geq 0$$

- Support : positive real number
- Asymmetric
- $X \sim t(r) \Rightarrow X^2 \sim F(1,r)$

$\because$

- $X \sim F(m,n) \Rightarrow \frac{1}{X} \sim F(n,m)$

$\because$

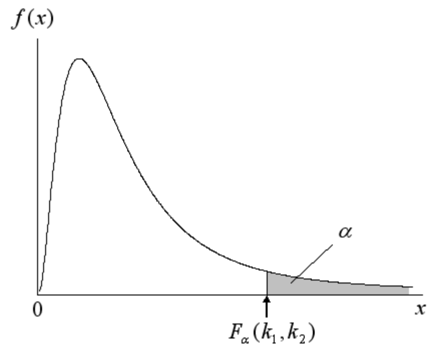
- Ex : $X_1,\cdots,X_n \sim N(\mu_1,\sigma_1^2)$, $Y_1,\cdots,Y_m \sim N(\mu_2,\sigma_2^2)$: indep.

  $S_1^2,\ S_2^2$: sample variance. Then,

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1),\quad \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1) : \text{indep.}$$

$$\therefore\ \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1,m-1)$$

- Density shape of F-distn & upper $(100\times\alpha)$ percentile



5.3 Normal Random Sample (정규 랜덤샘플)

Thm 5.3 $X_i \sim N(\mu_i,\sigma_i^2)$, $i=1,\cdots,n$ and indep.

$$\Rightarrow \sum_{i=1}^n a_iX_i \sim N\left(\sum_{i=1}^n a_i\mu_i,\ \sum_{i=1}^n a_i^2\sigma_i^2\right)$$

(Proof) Use mgf technique.

Let $Y=\sum_{i=1}^n a_iX_i$ $M_Y(t)$: mgf of Y. Then,

$$M_Y(t)=E\left(e^{t\sum_{i=1}^n a_iX_i}\right)=E(e^{ta_1X_1})E(e^{ta_2X_2})\cdots E(e^{ta_nX_n})$$

$$=e^{a_1\mu_1t+\frac{1}{2}a_1^2\sigma_1^2t^2}\,e^{a_2\mu_2t+\frac{1}{2}a_2^2\sigma_2^2t^2}\cdots e^{a_n\mu_nt+\frac{1}{2}a_n^2\sigma_n^2t^2}$$

$$=e^{(\sum_{i=1}^n a_i\mu_i)t+\frac{1}{2}(\sum_{i=1}^n a_i^2\sigma_i^2)t^2}$$

: mgf of $N\left(\sum_{i=1}^n a_i\mu_i,\ \sum_{i=1}^n a_i^2\sigma_i^2\right)$

Thm 5.5 (Distribution of the sample mean of normal sample)

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \;\Rightarrow\; \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(Proof)

Now, consider the distribution of the sample variance of normal sample

Thm 5.6 Let $X_1, X_2, \cdots, X_n$ be a r.s. from $N(\mu, \sigma^2)$ and

$$\overline{X} = \sum_{i=1}^{n} X_i / n, \quad S^2 = \sum_{i=1}^{n} (X_i - \overline{X}) / (n-1). \text{ Then,}$$

a) $\overline{X}$ and S^2 are independent

b) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(Proof of a)

$\overline{X}$ and $(X_k - \overline{X})$ are normally distributed.

(Since it is the linear combination of normal random variables.)

Hence,

$\overline{X}$ and $(X_k - \overline{X})$ indep. $\Leftrightarrow$ $Cov(\overline{X}, X_k - \overline{X}) = 0$

(Recall the definition of independence, expecially normal case)

$$Cov(\overline{X}, X_k - \overline{X}) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i,\ X_k - \frac{1}{n}\sum_{j=1}^{n} X_j\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Cov(X_i, X_k) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

$$= \frac{\sigma^2}{n} - \frac{n\sigma^2}{n^2} = 0$$

$\therefore$ $\overline{X}$ and $(X_k - \overline{X})$ indep.

& S^2 is a function of $(X_1 - \overline{X}, \cdots, X_n - \overline{X})$.

$\Rightarrow$ $\overline{X}$ and S^2 are indep.

(Proof of b)

Note that

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2$$

and

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n),\ \frac{n(\overline{X}-\mu)^2}{\sigma^2} \sim \chi^2(1)$$

Then, by Thm 5.3

$$\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2} \ \sim\ \chi^2(n-1)$$

Reference : Another Proof of a) : Not essential

Let $\frac{X_i-\mu}{\sigma}=Z_i$, then $Z_i \overset{iid}{\sim} N(0,1)$ and $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\sigma\overline{Z}+\mu$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n-1}\sigma^2\sum_{i=1}^{n}(Z_i-\overline{Z})^2$$

$$\Longrightarrow (n-1)\frac{S^2}{\sigma^2}=\sum_{i=1}^{n}(Z_i-\overline{Z})^2$$

$\therefore\ \overline{Z}$ is independent of $\sum_{i=1}^{n}(Z_i-\overline{Z})^2$

$\Leftrightarrow\ \overline{X}$ is independent of S^2

Consider the Z_i's.

$$\sum_{i=1}^{n}(Z_i-\overline{Z})=0 \Longrightarrow Z_1-\overline{Z}=-\sum_{i=2}^{n}(Z_i-\overline{Z})$$

So,

$$\sum_{i=1}^{n}(Z_i-\overline{Z})^2=(Z_1-\overline{Z})^2+\sum_{i=2}^{n}(Z_i-\overline{Z})^2=(-\sum_{i=2}^{n}(Z_i-\overline{Z}))^2+\sum_{i=2}^{n}(Z_i-\overline{Z})^2$$

Now consider the vector $(\overline{Z},\ Z_2-\overline{Z},\ \cdots,\ Z_n-\overline{Z}.)$

Let $Y_1=\overline{Z},\ Y_2=Z_2-\overline{Z},\ \cdots,\ Y_n=Z_n-\overline{Z}$

We need joint p.d.f of $Y_1,Y_2,\cdots,Y_n$ from $Z_1,Z_2,\cdots,Z_n$

$$Z_1 = Y_1 - \sum_{i=2}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} Z_i - \sum_{i=2}^{n} Z_i + \frac{n-1}{n}\sum_{i=1}^{n} Z_i$$
$$Z_2 = Y_2 + Y_1$$
$$\vdots$$
$$Z_n = Y_n + Y_1$$

Jacobian:

$$J = \begin{pmatrix} 1 & -1 & -1 & \cdots & & -1 \\ 1 & 1 & 0 & \cdots & & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & \ddots & & \\ 1 & 0 & & \cdots & 0 & 1 \end{pmatrix} \qquad J = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \text{ form}$$

$$\Rightarrow |J| = |D|\,|A - BC| = |1| \cdot |1 + (n-1)| = n$$

$$f_Z(z_1, z_2, \cdots, z_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} z_i^2}$$

$$f_Y(y_1, y_2, \cdots y_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{\frac{-(y_1 - \sum_{i=2}^{n} y_i)^2}{2}} e^{\frac{-\sum_{i=2}^{n}(y_i + y_1)^2}{2}} \cdot n$$

$$= n\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left\{-\frac{1}{2}(ny_1^2)\right\} exp\left\{-\frac{1}{2}\left(\sum_{i=2}^{n} y_i\right)^2 - \frac{1}{2}\sum_{i=2}^{n} y_i^2\right\}$$

$$= g(y_1)\, h(y_2, \cdots, y_n)$$

Hence, Y_1 and $(Y_2, \cdots, Y_n)$ are independent.

So, $\overline{Z} = Y_1$ is independent of $S^2 =$ function of $(Y_2, \cdots, Y_n)$.

Thm 5.7 $X_1, X_2, \cdots, X_n$: r.s. from $N(\mu, \sigma^2)$. Then, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$.

(Proof) We know that $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ Distn. of $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$?

Note that $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} = \dfrac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/(n-1)}}$

Check : (i) Distn of numerator
(ii) Distn of denominator
(iii) Independence of numerator and denominator

5.4 Order Statistics (순서통계량)

Def. (Order Statistics)

$X_1, X_2, \cdots, X_n$: r.s. from a population.

Order Statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are sample values put into increasing order. i.e.,

$X_{(1)}$: $\min X_i$

$X_{(2)}$: Second smallest X_i

$\vdots$

$X_{(n)} = \max X_i$

$X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$

if it is continuous => $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.

Read text p.195 : Example of finding probability of O.S.

Question1. $X_1, X_2, \cdots, X_n$: r.s. from a population with pdf $f(x)$.

What is the distn of $X_{(n)}$?

Question2. $X_1, X_2, \cdots, X_n$: r.s. from a population with pdf $f(x)$.

What is the distn of $X_{(1)}$?

Ex 5.1 $X_1, X_2, \cdots, X_n$: r.s from $U(0,1)$. What is the distn of $X_{(n)}$?

Ex 5.2 $X_1, X_2, \cdots, X_n$: r.s from $U(0,1)$. What is the distn of $X_{(1)}$?

Ex 5.3

The p.d.f for $X_{(i)}$? (Thm 5.8)

c.d.f : $F_{X(i)}(x) = P(X_{(i)} \leqq x)$

Define Y = # of X_j's that are $\leqq x$

Let $\{X_j \leqq x\}$ be a success. Then Y = # of success in n trials.

And the trials are independent because X_j's are independent.

So, $\{X_{(i)} \leqq x\}$ implies $Y \geqq i$.

Probability of success = $P(X_i \leqq x) = F(x)$.

Then, $Y \sim B(n, F(x))$

$$F_{X(i)}(x) = P(X_{(i)} \leqq x) = P(Y \geqq i) = \sum_{j=i}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j}$$

So, the p.d.f of $X_{(i)}$ is

$$f_{X_{(i)}}(x) = \frac{d}{dx} F_{X_{(i)}}(x) = \sum_{j=i}^{n} \binom{n}{j} j F(x)^{j-1} (1-F(x))^{n-j} f(x)$$

$$+ \sum_{j=i}^{n} \binom{n}{j} F(x)^j (n-j)(1-F(x))^{n-j-1}(-f(x))$$

$$==> f_{X_{(i)}}(x) = \frac{n!}{i!(n-i)!} iF(x)^{i-1}(1-F(x))^{n-i} f(x) \quad (\text{ In case of } \ j=i\ )$$

$$+ \sum_{j=i+1}^{n} \binom{n}{j} jF(x)^{j-1}(1-F(x))^{n-j} f(x)$$

$$- \sum_{j=i}^{n-1} \binom{n}{j} (n-j)F(x)^{j}(1-F(x))^{n-j-1} f(x)$$

$(j = n$ term vanishes $)$

Hence,

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}(1-F(x))^{n-i} f(x)$$

$$+ \sum_{j=i}^{n-1} \binom{n}{j+1} (j+1)F(x)^{j}(1-F(x))^{n-j-1} f(x)$$

$$- \sum_{j=i}^{n-1} \binom{n}{j} (n-j)F(x)^{j}(1-F(x))^{n-j-1} f(x)$$

Note that $\binom{n}{j+1}(j+1) = \frac{n!}{j!(n-j-1)!} = \frac{n!}{j!(n-j)!}(n-j) = \binom{n}{j}(n-j)$.

Therefore, $f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}(1-F(x))^{n-i} f(x)$

Note. Easy comprehension

$(i-1)$개 $\quad$ 1개 $\quad$ $(n-i)$개

$X_{(i)} = x$

Ex 5.4 $X_1, X_2, \cdots, X_n$: r.s from $U(0,1)$. What is the distn of $X_{(k)}$?

Joint p.d.f of $X_{(i)},\ X_{(j)}$ for $1 \leqq i < j \leqq n$? (Thm 5.9)

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(u)^{i-1}(F(v)-F(u))^{j-i-1}F(v)^{n-j}f(u)f(v),$$

where $-\infty < u < v < \infty$

Joint p.d.f of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$?

$$f_{X_{(1)},X_{(2)},\cdots,X_{(n)}}(x_1,x_2,\cdots,x_n) = n!f(x_1)f(x_2)\cdots f(x_n)$$

where $-\infty < x_1 < \cdots < x_n < \infty$.

Ex 5.5 $X_1, X_2, \cdots, X_n$: r.s from $U(0,1)$.

(a) What is the joint distn of $X_{(1)}$ and $X_{(n)}$?

(b) What is the distn of $X_{(n)} - X_{(1)}$?