5장. 표본분포의 근사

5.3 극한분포의 계산

정리 5.3.1: 슬럿츠키(Slutsky)의 정리

확률변수 $X_n, Y_n (n=1,2,\cdots), Z$와 실수의 상수 c에 대하여

$$X_n \underset{n\to\infty}{\overset{d}{\to}} Z, \qquad \underset{n\to\infty}{\mathrm{p}\lim} Y_n = c$$

이면 다음이 성립한다.

(a) $X_n + Y_n \underset{n\to\infty}{\overset{d}{\to}} Z + c$

(b) $X_n - Y_n \underset{n\to\infty}{\overset{d}{\to}} Z - c$

(c) $Y_n X_n \underset{n\to\infty}{\overset{d}{\to}} cZ$

(d) $X_n / Y_n \underset{n\to\infty}{\overset{d}{\to}} Z/c \ \ (c \neq 0)$

[증명의 Key] $\underset{n\to\infty}{\mathrm{p}\lim} Y_n = c$ 이므로 $|Y_n - c| < \epsilon$ 인 경우와 $|Y_n - c| \geq \epsilon$ 인 경우로 나누어

$$\lim_{n\to\infty} P(|Y_n - c| < \epsilon) = 1, \ \lim_{n\to\infty} P(|Y_n - c| \geq \epsilon) = 0$$

임을 활용한다.

예 5.3.1 스튜던트화된 표본평균의 극한분포:

모평균이 μ이고 모표준편차가 $\sigma \ (0 < \sigma < +\infty)$인 모집단에서의 랜덤표본 n개로부터의 표본평균과 표본표준편차를 각각 $\overline{X_n}, S_n$이라고 할 때

$$\frac{\overline{X_n} - \mu}{S_n / \sqrt{n}}$$

를 스튜던트화(Studentized)된 표본평균이라고 한다. 한편 중심극한 정리와 예 5.2.5로부터

$$\frac{\overline{X_n} - \mu}{\sigma / \sqrt{n}} \underset{n\to\infty}{\overset{d}{\to}} Z, \ Z \sim N(0,1) \ \text{이고} \quad \underset{n\to\infty}{\mathrm{p}\lim} S_n = \sigma$$

이므로 슬럿츠키의 정리로부터

$$\frac{\overline{X_n} - \mu}{S_n / \sqrt{n}} = \frac{\overline{X_n} - \mu}{\sigma / \sqrt{n}} / \frac{S_n}{\sigma} \underset{n\to\infty}{\overset{d}{\to}} Z/1 = Z, \ Z \sim N(0,1)$$

모집단의 분포가 무엇이든 양수의 분산이 정의될 수만 있으면

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{S_n / \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$\therefore \ P\left\{\overline{X_n} - z_{\alpha/2} S_n / \sqrt{n} \leq \mu < \overline{X_n} + z_{\alpha/2} S_n / \sqrt{n}\right\} \simeq 1 - \alpha, \ \ n \to \infty$$

이므로, 모평균 μ에 관한 $100(1-\alpha)\%$ 점근(漸近 asymptotic) 신뢰구간이 다음과 같이 주어진다.

$$\mu \in [\overline{X_n} - z_{\alpha/2} S_n / \sqrt{n}, \ \overline{X_n} + z_{\alpha/2} S_n / \sqrt{n})$$

예 5.3.2 표본분산의 극한분포:

모평균이 μ이고 모표준편차가 $\sigma\,(0<\sigma<+\infty)$인 모집단에서의 랜덤표본 n개로부터의 표본분산 S_n^2의 표본분포에 대하여 생각해보자. 표본평균을 $\overline{X_n}$라고 하면

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}[(X_i-\mu)-(\overline{X_n}-\mu)]^2=\frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\overline{X_n}-\mu)^2\right\}$$

한편 $Y_i=(X_i-\mu)^2\,(i=1,\cdots,n)$이라고 하면 이들은 서로 독립이고 동일한 분포를 따르므로, 중심극한정리로부터 $E[(X_1-\mu)^4]<+\infty$일 때

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right\}\underset{n\to\infty}{\overset{d}{\to}}W,\quad W\sim N(0,E[(X_1-\mu)^4]-\sigma^4)$$

또한 중심극한정리와 대수의 법칙으로부터

$$\sqrt{n}(\overline{X_n}-\mu)\underset{n\to\infty}{\overset{d}{\to}}Z,\ Z\sim N(0,\sigma^2),\quad \mathrm{p}\lim_{n\to\infty}(\overline{X_n}-\mu)=0$$

이므로 슬럿츠키의 정리와 정리 5.2.1로부터

$$\sqrt{n}(\overline{X_n}-\mu)(\overline{X_n}-\mu)\underset{n\to\infty}{\overset{d}{\to}}0\times Z=0,\ Z\sim N(0,\sigma^2)$$

$$\therefore\ \mathrm{p}\lim_{n\to\infty}\sqrt{n}(\overline{X_n}-\mu)^2=0$$

따라서 $E[(X_1-\mu)^4]<+\infty$일 때, $W\sim N(0,E[(X_1-\mu)^4]-\sigma^4)$라고 하면

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X_n})^2-\sigma^2\right)=\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right\}-\sqrt{n}(\overline{X_n}-\mu)^2\underset{n\to\infty}{\overset{d}{\to}}W-0=W$$

그러므로 $E[(X_1-\mu)^4]<+\infty$일 때 $\rho_4=E[(\frac{X_1-\mu}{\sigma})^4]-3$이라고 하면[1)]

$$\sqrt{n}(S_n^2-\sigma^2)=\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2-\sigma^2\right\}+\frac{1}{\sqrt{n}}S_n^2\underset{n\to\infty}{\overset{d}{\to}}W,\ W\sim N(0,(\rho_4+2)\sigma^4)$$

다차원 경우의 극한분포:

다차원 확률변수의 열 $X_n=(X_{n1},\cdots,X_{nk})^t\ (n=1,2,\cdots)$과 $Z=(Z_1,\cdots,Z_k)^t$에 대하여

$$\lim_{n\to\infty}P(X_{n1}\le x_1,\cdots,X_{nk}\le x_k)=P(Z_1\le x_1,\cdots,Z_k\le x_k)^{2)}$$

가 $Z=(Z_1,\cdots,Z_k)^t$의 결합누적분포함수 $cdf_Z(x_1,\cdots,x_k)$가 연속인 모든 점 $x=(x_1,\cdots,x_k)^t$에서 성립할 때, $Z=(Z_1,\cdots,Z_k)^t$의 분포를 $X_n=(X_{n1},\cdots,X_{nk})^t$ 분포들의 극한분포 또는 점근분포라고 하며 기호로는 일차원의 경우와 마찬가지로 다음과 같이 나타낸다.

$$X_n\underset{n\to\infty}{\overset{d}{\to}}Z\ \text{ 또는 }\ (X_{n1},\cdots,X_{nk})^t\underset{n\to\infty}{\overset{d}{\to}}(Z_1,\cdots,Z_k)^t$$

---

1) 여기에서 정의된 ρ_4를 모집단의 첨예도(尖銳度 kurtosis)라고 하며 이는 모집단 분포가 평균 부근에 밀집되어 있는 정도를 나타내는 측도이고, 모집단의 분포가 정규분포인 경우에 첨예도는 0이다.

2) 이러한 결합확률을 간략히 $P(Z_1\le x_1,\cdots,Z_k\le x_k)=P(Z\le x)$로 나타내기로 한다.

정리 5.3.2: 연속함수와 극한분포

다차원 확률변수 $X_n = (X_{n1},\cdots,X_{nk})^t\,(n=1,2\cdots)$과 $Z=(Z_1,\cdots,Z_k)^t$에 대하여

$$(X_{n1},\cdots,X_{nk})^t \underset{n\to\infty}{\overset{d}{\to}} (Z_1,\cdots,Z_k)^t$$

일 때, 연속함수 g에 대하여

$$g(X_n) \underset{n\to\infty}{\overset{d}{\to}} g(Z)$$

[증명] 이 정리의 증명은 이 책의 수준을 넘으므로 생략한다.

[유의] 정리 5.3.1이 이 정리의 특별한 경우임에 유의...

예 5.3.3 이항분포와 카이제곱근사:

이항분포의 정규근사로부터 $X_n \sim B(n,p)\,(0<p<1)$일 때

$$\frac{X_n - np}{\sqrt{np(1-p)}} \underset{n\to\infty}{\overset{d}{\to}} Z,\ \ Z\sim N(0,1)$$

또한 제곱 함수는 연속함수이므로 정리 5.3.2로부터

$$\left(\frac{X_n - np}{\sqrt{np(1-p)}}\right)^2 \underset{n\to\infty}{\overset{d}{\to}} Z^2,\ \ Z\sim N(0,1)$$

$$\therefore\ \frac{(X_n - np)^2}{np(1-p)} \underset{n\to\infty}{\overset{d}{\to}} W,\ W\sim \chi^2(1)$$

한편, $X_{n1}=X_n, X_{n2}=n-X_n, p_1=p, p_2=1-p$라고 하면

$$(X_{n1},X_{n2})^t \sim \mathrm{Multi}(n,\,(p_1,p_2)^t)$$

이고 다음 등식이 성립함을 알 수 있다.

$$\frac{(X_{n1}-np_1)^2}{np_1}+\frac{(X_{n2}-np_2)^2}{np_2}=\frac{(X_n-np)^2}{np}+\frac{(n-X_n-n(1-p))^2}{n(1-p)}=\frac{(X_n-np)^2}{np(1-p)}$$

따라서 $(X_{n1},X_{n2})^t \sim \mathrm{Multi}(n,\,(p_1,p_2)^t)$이고

$$\sum_{j=1}^{2}\frac{(X_{nj}-np_j)^2}{np_j} \underset{n\to\infty}{\overset{d}{\to}} W,\ \ W\sim\chi^2(1)$$

예 5.3.4 다항분포와 카이제곱근사:

예 5.3.3을 일반화 하여

$$(\mathrm{X}_{n1},\cdots,\mathrm{X}_{nk})^t \sim \mathrm{Multi}(n,(p_1,\cdots,p_k)^t)\,(p_1+\cdots+p_k=1,\, p_i>0, i=1,\cdots,k)$$

일 때 다음과 같은 카이제곱근사가 성립하는 것을 알아보자.

$$\sum_{j=1}^{k}\frac{(\mathrm{X}_{nj}-np_j)^2}{np_j}\xrightarrow[n\to\infty]{d}\mathrm{W},\ \mathrm{W}\sim\chi^2(k-1)$$

여기에서 $\mathrm{X}_n=(\mathrm{X}_{n1},\cdots,\mathrm{X}_{nk})^t, p=(p_1,\cdots,p_k)^t$라고 하면 예 5.1.2에서와 같이 다차원 경우의 중심극한정리로부터

$$(\mathrm{X}_n-np)/\sqrt{n}\xrightarrow[n\to\infty]{d}\mathrm{X},\ \mathrm{X}\sim\mathrm{N}_k(0,D(p_j)-pp^t)$$

이제 정리 5.3.2를 이용하여 이 결과를 밝히기 위하여

$$\mathrm{Y}_{nj}=\mathrm{X}_{nj}\,,\mu_{\mathrm{j}}=p_j\,(j=1,\cdots,r), \mathrm{Y}_n=(\mathrm{Y}_{n1},\cdots,\mathrm{Y}_{nr})^t\,,\mu=(\mu_1,\cdots,\mu_r)^t\,, r=k-1$$

이라고 하고, $\mu_1,\cdots,\mu_r$ 을 대각원소로 하는 대각행렬을 $D(\mu_j)$라고 하면

$$(\mathrm{Y}_n-n\mu)/\sqrt{n}\xrightarrow[n\to\infty]{d}\mathrm{Z},\ \mathrm{Z}\sim\mathrm{N}_r(0,\Sigma),\ \Sigma=D(\mu_j)-\mu\mu^t$$

또한 이차형식 $z^t\Sigma^{-1}z$은 연속함수이므로 정리 5.3.2로부터

$$\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}^t\Sigma^{-1}\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}\xrightarrow[n\to\infty]{d}\mathrm{Z}^t\Sigma^{-1}\mathrm{Z},\ \mathrm{Z}\sim\mathrm{N}_r(0,\Sigma)$$

정리 4.4.5로부터 $\mathrm{Z}\sim\mathrm{N}_r(0,\Sigma)$이면 $\mathrm{Z}^t\Sigma^{-1}\mathrm{Z}\sim\chi^2(r), r=k-1$이므로

$$\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}^t\Sigma^{-1}\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}\xrightarrow[n\to\infty]{d}\mathrm{W},\ \mathrm{W}\sim\chi^2(k-1)$$

한편 예 4.4.4에서와 같이 특수한 형태의 행렬에 관한 연산[3]을 이용하여 계산하면

$$\Sigma^{-1}=(D(\mu_j)-\mu\mu^t)^{-1}=(\mathrm{I}-D^{-1}(\mu_j)\mu\mu^t)^{-1}D^{-1}(\mu_j)=D^{-1}(\mu_j)+\frac{1}{p_k}11^t$$

이므로

$$\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}^t\Sigma^{-1}\left\{(\mathrm{Y}_n-n\mu)/\sqrt{n}\right\}=\sum_{j=1}^{k-1}\frac{(\mathrm{Y}_{nj}-n\mu_j)^2}{n\mu_j}+\frac{(\mathrm{X}_{nk}-np_k)^2}{np_k}$$

$$\therefore\ \sum_{j=1}^{k}\frac{(\mathrm{X}_{nj}-np_j)^2}{np_j}\xrightarrow[n\to\infty]{d}\mathrm{W},\ \mathrm{W}\sim\chi^2(k-1)$$

---

3) 벡터 a,b에 대하여 $1+b^ta\neq 0$이면

$$(\mathrm{I}+ab^t)^{-1}=\mathrm{I}+c\,ab^t,\ \ c=-1/(1+b^ta)$$

정리 5.3.3: 일차근사를 이용한 극한분포 계산
다차원 확률변수 $X_n = (X_{n1},\cdots,X_{nk})^t\ (n=1,2,\cdots)$, $Z=(Z_1,\cdots,Z_k)^t$와 벡터 $\theta=(\theta_1,\cdots,\theta_k)^t$에 대하여

$$\sqrt{n}(X_n-\theta) \underset{n\to\infty}{\overset{d}{\to}} Z$$

이고 함수 $g(\theta)$의 일차편도함수[4] $\dot{g}(\theta)$가 연속함수이면 다음이 성립한다.

$$\sqrt{n}(g(X_n)-g(\theta)) \underset{n\to\infty}{\overset{d}{\to}} (\dot{g}(\theta))^t Z$$

[증명] 다차원의 경우도 일차원 경우와 마찬가지이므로 일차원의 경우만 증명하기로 한다.
함수 $g(\theta)$가 미분가능하므로 다음을 만족하는 $r(x)(=r_\theta(x))$가 존재한다.

$$g(x)=g(\theta)+\{\dot{g}(\theta)+r(x)\}(x-\theta),\quad r(x)\underset{x\to\theta}{\to}0$$

여기에서 잉여항의 $r(x)$에 대한 조건으로부터 다음이 성립하는 것을 알 수 있다.

$$\forall\epsilon>0,\ \exists\delta>0:(|x-\theta|<\delta\Rightarrow|r(x)|<\epsilon)$$

따라서 임의의 양수 ϵ에 대하여 이러한 양수 δ를 선택하면

$$(|X_n-\theta|<\delta)\subseteq(|r(X_n)|<\epsilon),\ \text{즉}\ (|r(X_n)|\ge\epsilon)\subseteq(|X_n-\theta|\ge\delta)$$

$$\therefore P(|r(X_n)|\ge\epsilon)\le P(|X_n-\theta|\ge\delta)$$

한편 $\sqrt{n}(X_n-\theta) \underset{n\to\infty}{\overset{d}{\to}} Z$ 이므로 슬럿츠키의 정리와 정리 5.2.1로부터

$$X_n-\theta=\frac{1}{\sqrt{n}}\times\sqrt{n}(X_n-\theta)\underset{n\to\infty}{\overset{d}{\to}}0\times Z=0,$$

$$\therefore \operatorname*{plim}_{n\to\infty} X_n=\theta,\ \text{즉}\ \lim_{n\to\infty}P(|X_n-\theta|\ge\delta)=0$$

$$\therefore 0\le\lim_{n\to\infty}P(|r(X_n)|\ge\epsilon)\le\lim_{n\to\infty}P(|X_n-\theta|\ge\delta)=0,\quad \text{즉}\ \operatorname*{plim}_{n\to\infty} r(X_n)=0$$

따라서

$$\sqrt{n}(g(X_n)-g(\theta))=(\dot{g}(\theta)+r(X_n))\sqrt{n}(X_n-\theta),\ \operatorname*{plim}_{n\to\infty}(\dot{g}(\theta)+r(X_n))=\dot{g}(\theta)$$

그러므로 슬럿츠키의 정리로부터 $\sqrt{n}(g(X_n)-g(\theta)) \underset{n\to\infty}{\overset{d}{\to}} \dot{g}(\theta)Z$

예 5.3.5 표본표준편차의 극한분포:
모평균이 μ이고 모표준편차가 $\sigma\ (0<\sigma<+\infty)$인 모집단에서의 랜덤표본 n개로부터의 표본분산을 S_n^2 이라고 할 때, 예 5.3.2로부터 $E[(X_1-\mu)^4]<+\infty$ 라는 조건하에서

$$\sqrt{n}(S_n^2-\sigma^2)\underset{n\to\infty}{\overset{d}{\to}}W,\ W\sim N(0,(\rho_4+2)\sigma^4),\quad \rho_4=E[(\frac{X_1-\mu}{\sigma})^4]-3$$

표본표준편차 S_n은 표본분산의 제곱근 $\sqrt{S_n^2}$ 이고 제곱근함수 $g(x)=\sqrt{x}$의 도함수가 $\dot{g}(x)=1/(2\sqrt{x})$ 이므로, 정리 5.3.3으로부터 $E[(X_1-\mu)^4]<+\infty$ 라는 조건하에서

$$\sqrt{n}(\sqrt{S_n^2}-\sqrt{\sigma^2})\underset{n\to\infty}{\overset{d}{\to}}\frac{1}{2\sqrt{\sigma^2}}W,\ W\sim N(0,(\rho_4+2)\sigma^4)$$

$$\therefore\ \sqrt{n}(S_n-\sigma)\underset{n\to\infty}{\overset{d}{\to}}Z,\ Z\sim N(0,(\rho_4+2)\sigma^2/4)$$

4) 일차원 경우에는 $\dot{g}(\theta)=dg(\theta)/d\theta$, 다차원 경우에는 $\dot{g}(\theta)=(\partial g(\theta)/\partial\theta_1,\cdots,\partial g(\theta)/\partial\theta_k)^t$

예 5.3.6 표본상관계수의 극한분포:

모상관계수가 $\rho\,(-1<\rho<1)$인 이변량의 모집단에서의 랜덤표본을 $(\mathrm{X}_1,\mathrm{Y}_1)^t,\cdots,(\mathrm{X}_n,\mathrm{Y}_n)^t\,(n>2)$이라고 할 때, 모상관계수의 추측에 사용되는 표본상관계수(標本相關係數 sample correlation coefficient)는 다음과 같이 정의된다.

$$\widehat{\rho_n}=\frac{\sum_{i=1}^{n}(\mathrm{X}_i-\overline{\mathrm{X}})(\mathrm{Y}_i-\overline{\mathrm{Y}})}{\sqrt{\sum_{i=1}^{n}(\mathrm{X}_i-\overline{\mathrm{X}})^2}\sqrt{\sum_{i=1}^{n}(\mathrm{Y}_i-\overline{\mathrm{Y}})^2}}$$

표본상관계수의 표본분포를 구할 때에는

$$\mu_1=0,\mu_2=0,\sigma_1=1,\sigma_2=1$$

이라고 가정해도 좋다. 또한 간략한 표현을 위하여

$$\overline{\mathrm{X}_n}=\sum_{i=1}^{n}\mathrm{X}_i/n,\ \overline{\mathrm{Y}_n}=\sum_{i=1}^{n}\mathrm{Y}_i/n,\ \overline{(\mathrm{XY})_n}=\sum_{i=1}^{n}\mathrm{X}_i\mathrm{Y}_i/n,\ \overline{(\mathrm{X}^2)_n}=\sum_{i=1}^{n}\mathrm{X}_i^2/n,\ \overline{(\mathrm{Y}^2)_n}=\sum_{i=1}^{n}\mathrm{Y}_i^2/n$$

라고 하면, 함수 $g(t_1,t_2,t_3,t_4,t_5)=(t_3-t_1t_2)(t_4-t_1^2)^{-1/2}(t_5-t_2^2)^{-1/2}$ 에 대하여

$$\widehat{\rho_n}=\frac{\overline{(\mathrm{XY})_n}-\overline{\mathrm{X}_n}\,\overline{\mathrm{Y}_n}}{\sqrt{\overline{(\mathrm{X}^2)_n}-(\overline{\mathrm{X}_n})^2}\sqrt{\overline{(\mathrm{Y}^2)_n}-(\overline{\mathrm{Y}_n})^2}}=g(\overline{\mathrm{X}_n},\overline{\mathrm{Y}_n},\overline{(\mathrm{XY})_n},\overline{(\mathrm{X}^2)_n},\overline{(\mathrm{Y}^2)_n})$$

와 같이 표본상관계수를 $(\mathrm{X}_i,\mathrm{Y}_i,\mathrm{X}_i\mathrm{Y}_i,\mathrm{X}_i^2,\mathrm{Y}_i^2)^t\ (i=1,\cdots,n)$의 평균의 함수로 나타낼 수 있다.

또한 $\mathrm{Z}_i=(\mathrm{X}_i,\mathrm{Y}_i,\mathrm{X}_i\mathrm{Y}_i,\mathrm{X}_i^2,\mathrm{Y}_i^2)^t\ (i=1,\cdots,n)$이라고 하면 $\mathrm{Z}_i\,(i=1,\cdots,n)$는 서로 독립이고 동일한 분포를 따르는 5차원의 확률변수이므로 그 분산행렬이 존재할 때 다음이 성립한다.

$$\sqrt{n}\,(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Z}_i-\mathrm{E}(\mathrm{Z}_1))\ \underset{n\to\infty}{\overset{d}{\to}}\ \mathrm{V},\ \ \mathrm{V}\sim\mathrm{N}_5(0,\mathrm{Var}(\mathrm{Z}_1))$$

한편 추가된 가정으로부터

$$\mathrm{E}(\mathrm{Z}_1)=(0,0,\rho,1,1)^t,\ \rho=g(0,0,\rho,1,1)=g(\mathrm{E}(\mathrm{Z}_1))$$

이고 함수 $g(t_1,t_2,t_3,t_4,t_5)=(t_3-t_1t_2)(t_4-t_1^2)^{-1/2}(t_5-t_2^2)^{-1/2}$의 편도함수들이 연속이므로, 정리 5.3.3으로부터 분산행렬 $\mathrm{Var}(\mathrm{Z}_1)$이 존재할 때[5] $\theta=\mathrm{E}(\mathrm{Z}_1)=(0,0,\rho,1,1)^t$라고 하면

$$\sqrt{n}\,(\widehat{\rho_n}-\rho)=\sqrt{n}\,(g(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Z}_i)-g(\theta))\ \underset{n\to\infty}{\overset{d}{\to}}\ (\dot{g}(\theta))^t\mathrm{V},\ \ \mathrm{V}\sim\mathrm{N}_5(0,\mathrm{Var}(\mathrm{Z}_1))$$

여기에서 함수 g의 일차편도함수를 구하여 계산하면

$$\dot{g}(\theta)=\dot{g}(0,0,\rho,1,1)=(0,0,1,-\rho/2,-\rho/2)^t$$

$$(\dot{g}(\theta))^t\mathrm{V}\sim\mathrm{N}(0,(\dot{g}(\theta))^t\mathrm{Var}(\mathrm{Z}_1)(\dot{g}(\theta)))=\mathrm{N}(0,\mathrm{Var}[(\dot{g}(\theta))^t\mathrm{Z}_1])$$

$$(\dot{g}(\theta))^t\mathrm{Z}_1=\mathrm{X}_1\mathrm{Y}_1-\frac{\rho}{2}\mathrm{X}_1^2-\frac{\rho}{2}\mathrm{Y}_1^2$$

이므로, $\mathrm{E}(\mathrm{X}_1^4)<+\infty,\mathrm{E}(\mathrm{Y}_1^4)<+\infty$ 일 때

$$\sqrt{n}\,(\widehat{\rho_n}-\rho)\ \underset{n\to\infty}{\overset{d}{\to}}\ \mathrm{W},\ \ \mathrm{W}\sim\mathrm{N}(0,\mathrm{Var}(\mathrm{X}_1\mathrm{Y}_1-\frac{\rho}{2}\mathrm{X}_1^2-\frac{\rho}{2}\mathrm{Y}_1^2))$$

---

5) $\mathrm{Z}_1=(\mathrm{X}_1,\mathrm{Y}_1,\mathrm{X}_1\mathrm{Y}_1,\mathrm{X}_1^2,\mathrm{Y}_1^2)^t$ 이므로 정리 1.6.2 으로부터 $\mathrm{E}(\mathrm{X}_1^4)<+\infty,\mathrm{E}(\mathrm{Y}_1^4)<+\infty$ 이면 Z_1의 분산행렬이 존재하는 것을 알 수 있다.

예 5.3.7 표본상관계수와 분산안정변환:

예 5.3.6에서 이변량 정규분포 $N(\mu_1,\mu_2;\sigma_1^2,\sigma_2^2,\rho)$ $(\sigma_1>0,\sigma_2>0,\ -1<\rho<1)$가 모집단의 분포일 때 표본상관계수의 극한분포에 대하여 알아보자. 이 경우에는 예 4.4.3으로부터

$$Y_1-\rho X_{1|X_1=x_1}\sim N(0,1-\rho^2)$$

으로서 조건부 분포가 조건 $X_1=x_1$에 의존하지 않으므로 $Y_1-\rho X_1$은 X_1과 서로 독립이다.

따라서 $\mu_1=\mu_2=0,\sigma_1=\sigma_2=1$ 인 경우에

$$T=\frac{Y_1-\rho X_1}{\sqrt{1-\rho^2}}$$

이라고 하면, X_1과 T는 서로 독립이고 각각 표준정규분포 $N(0,1)$을 따른다.

한편 $Y_1=\rho X_1+\sqrt{1-\rho^2}\,T$ 를 대입하여 정리하면

$$X_1Y_1-\frac{\rho}{2}X_1^2-\frac{\rho}{2}Y_1^2=\frac{\rho}{2}(1-\rho^2)X_1^2+(1-\rho^2)^{3/2}X_1T-\frac{\rho}{2}(1-\rho^2)T^2$$

이고, 이 표현을 이용하여 계산하면 $\mu_1=\mu_2=0,\sigma_1=\sigma_2=1$ 인 경우에

$$\mathrm{Var}(X_1Y_1-\frac{\rho}{2}X_1^2-\frac{\rho}{2}Y_1^2)=(1-\rho^2)^2$$

$$\therefore\ \sqrt{n}(\hat{\rho}_n-\rho)\underset{n\to\infty}{\overset{d}{\to}}W,\ W\sim N(0,(1-\rho^2)^2)$$

여기에서 표본상관계수 $\hat{\rho}_n$의 함수 $g(\hat{\rho}_n)$의 극한분포를 구해보면 정리 5.3.3으로부터

$$\sqrt{n}(g(\hat{\rho}_n)-g(\rho))\underset{n\to\infty}{\overset{d}{\to}}\dot{g}(\rho)W,\ \dot{g}(\rho)W\sim N(0,[\dot{g}(\rho)]^2(1-\rho^2)^2)$$

이 때 극한분포의 분산을 ρ에 의존하지 않게 하는 $g(\hat{\rho}_n)$을 $\hat{\rho}_n$의 분산안정변환(分散安定變換 variance stabilizing transformation)이라고 한다. 특히

$$g(\hat{\rho}_n)=\frac{1}{2}\log\frac{1+\hat{\rho}_n}{1-\hat{\rho}_n}$$

은 피셔의 변환이라고 불리우며 $[\dot{g}(\rho)]^2(1-\rho^2)^2=1$이 되어 극한분포가 $N(0,1)$이 된다. 즉

$$\sqrt{n}(g(\hat{\rho}_n)-g(\rho))\underset{n\to\infty}{\overset{d}{\to}}Z,\ Z\sim N(0,1)$$

예 5.3.8 표본분위수의 극한분포:

예 5.2.7에서는 표본분위수 $X_{(r_n)}$ ($r_n \underset{n\to\infty}{\sim} \alpha n,\ 0<\alpha<1$)의 확률수렴에 대하여 알아보았다. 여기에서는 예 5.2.7에서의 조건이 만족될 때 표본분위수의 극한분포에 대하여 생각해보자.

모집단의 누적분포함수의 역함수를 F^{-1}이라 하고 $h(y)=F^{-1}(1-e^{-y})\,(y>0)$라고 하면 표본분위수의 분포는 정리 4.3.4로부터

$$X_{(r_n)} \overset{d}{\equiv} h\left(\frac{1}{n}Z_1+\cdots+\frac{1}{n-r_n+1}Z_{r_n}\right),\quad Z_i \overset{iid}{\sim} \mathrm{Exp}(1)\,(i=1,\cdots,n)$$

한편

$$Y_n=\frac{1}{n}Z_1+\cdots+\frac{1}{n-r_n+1}Z_{r_n}$$

이라고 하면 그 평균과 분산이 다음과 같이 근사되는 것을 알고 있다.

$$E(Y_n)=\frac{1}{n}+\cdots+\frac{1}{n-r_n+1}\underset{n\to\infty}{\sim}-\log(1-\alpha)$$

$$\mathrm{Var}(Y_n)=\frac{1}{n^2}+\cdots+\frac{1}{(n-r_n+1)^2}\underset{n\to\infty}{\sim}\frac{1}{n}\frac{\alpha}{1-\alpha}$$

이를 이용하여

$$W_n=\sqrt{n}\,\frac{Y_n-(-\log(1-\alpha))}{\sqrt{\alpha/(1-\alpha)}}$$

의 적률생성함수를 근사해보면[6] 다음이 성립하는 것을 알 수 있다.

$$\lim_{n\to\infty} mgf_{W_n}(t)=\exp\left(\frac{1}{2}t^2\right)=mgf_W(t),\ \ W\sim N(0,1)$$

$$\therefore\ W_n \underset{n\to\infty}{\overset{d}{\to}} W,\ \ W\sim N(0,1)$$

$$\therefore\ \sqrt{n}(Y_n-(-\log(1-\alpha)))\underset{n\to\infty}{\overset{d}{\to}}\sqrt{\alpha/(1-\alpha)}\,W$$

따라서 정리 5.3.3으로부터 함수 h가 미분가능할 때[7]

$$\sqrt{n}\{h(Y_n)-h(-\log(1-\alpha))\}\underset{n\to\infty}{\overset{d}{\to}}\dot{h}(-\log(1-\alpha))\sqrt{\alpha/(1-\alpha)}\,W$$

그런데 이러한 조건하에서 $\dot{h}(-\log(1-\alpha))=(1-\alpha)/f(F^{-1}(\alpha)),\ f=\dot{F}$ 이므로

$$\sqrt{n}\{X_{(r_n)}-F^{-1}(\alpha)\}\underset{n\to\infty}{\overset{d}{\to}}Z,\quad Z\sim N(0,\alpha(1-\alpha)/[f(F^{-1}(\alpha)]^2)$$

---

6) 연습문제 5.15에 이러한 적률 생성함수를 근사하는 과정이 소개되어 있다.

7) 확률밀도함수가 $f(x)=dF(x)/dx$ 로 주어지고 $f(F^{-1}(\alpha))>0$ 이면 이 조건이 만족된다.

5.4 모의실험을 이용한 근사

특정한 분포를 따르는 확률변수의 관측값을 흔히 난수(亂數 random number)라고 부르고, 최근에는 이러한 난수들을 생성할 수 있는 기능이 많은 통계패키지에 주어져 있다.

이러한 난수의 생성에 기본이 되는 것은 균등분포 $\mathrm{U}(0,1)$에서의 관측값으로서 이를 균등난수(均等亂數 uniform random number) 라고 하며, 부록 III에서는 패키지 R에서 균등난수를 생성하는 기능이 소개되어 있다. 또한 이러한 균등난수로부터 확률적분변환에 관한 정리 4.3.3을 이용하여 임의의 분포로부터의 난수를 생성할 수 있다.

예 5.4.1 로지스틱분포에서의 난수 생성:
예 4.1.5에서 소개된 로지스틱분포 $\mathrm{L}(0,1)$의 확률밀도함수와 누적분포함수는 각각

$$f(z)=\frac{e^z}{(1+e^z)^2},\ \mathrm{F}(z)=1-\frac{1}{(1+e^z)},\ -\infty<z<+\infty$$

로 주어진다. 이로부터 누적분포함수의 역함수는 다음과 같이 주어지는 것을 알 수 있다.

$$\mathrm{F}^{-1}(u)=\log\frac{u}{1-u}$$

따라서 정리 4.3.3으로부터 균등분포 $\mathrm{U}(0,1)$을 따르는 확률변수 U에 대하여

$$\mathrm{Z}=\log\frac{\mathrm{U}}{1-\mathrm{U}}\sim\mathrm{L}(0,1)$$

임을 알 수 있고, 이를 이용하여 로지스틱분포 $\mathrm{L}(0,1)$에서의 난수를 구할 수 있다. 또한

$$\sigma\mathrm{Z}+\mu=\sigma\log\frac{\mathrm{U}}{1-\mathrm{U}}+\mu\sim\mathrm{L}(\mu,\sigma)$$

임을 이용하여 일반적인 로지스틱분포 $\mathrm{L}(\mu,\sigma)$에서의 난수를 생성할 수 있다.

예 5.4.1에서와 마찬가지로, 균등분포 $\mathrm{U}(0,1)$을 따르는 확률변수 U에 대하여

$$-\log(1-\mathrm{U})\sim\mathrm{Exp}(1),\ \sigma\{-\log(1-\mathrm{U})\}\sim\mathrm{Exp}(\sigma)$$

임을 이용하여 지수분포 $\mathrm{Exp}(\sigma)$에서의 난수를 생성할 수 있고, 표준정규분포의 누적분포함수의 역함수 $\Phi^{-1}(u)$에 대하여

$$\Phi^{-1}(\mathrm{U})\sim\mathrm{N}(0,1),\ \sigma\Phi^{-1}(\mathrm{U})+\mu\sim\mathrm{N}(\mu,\sigma^2)$$

임을 이용하여 정규분포 $\mathrm{N}(\mu,\sigma^2)$에서의 난수를 생성할 수 있다. 이러한 방법을 이용하여 여러 가지 분포에서의 난수를 생성하는 기능이 부록 III에 소개된 R에 패키지화 되어 있다.

정리 5.4.1: 난수를 이용한 정적분의 근사

서로 독립이고 균등분포 $\mathrm{U}(0,1)$을 따르는 $\mathrm{U}_1,\cdots,\mathrm{U}_n$과 구간 $[a,b]$에서 연속인 함수 $g(x)$에 대하여 다음이 성립한다.

$$\mathrm{X}_i=(b-a)\mathrm{U}_i+a\overset{iid}{\sim}\mathrm{U}(a,b)\ (i=1,\cdots,n)$$

$$\mathrm{p}\lim_{n\to\infty}\frac{b-a}{n}\sum_{i=1}^{n}g(\mathrm{X}_i)=\int_a^b g(x)dx$$

[증명] $\mathrm{X}_1,\cdots,\mathrm{X}_n$이 서로 독립이고 확률밀도함수가

$$pdf_{\mathrm{X}_1}(x)=\frac{1}{b-a}\mathrm{I}_{(a,b)}(x)$$

로 주어지는 균등분포 $\mathrm{U}(a,b)$를 따르고

$$\mathrm{E}[g(\mathrm{X}_1)]=\frac{1}{(b-a)}\int_a^b g(x)\,dx$$

로 주어진다. 따라서 대수의 법칙으로부터

$$\mathrm{p}\lim_{n\to\infty}\frac{b-a}{n}\sum_{i=1}^{n}g(\mathrm{X}_i)=(b-a)\mathrm{E}[g(\mathrm{X}_1)]=\int_a^b g(x)dx$$

정리 5.4.1에서의 확률수렴을 이용하여 정적분의 근사값을 구하는 방법을 **몬테칼로적분**(Monte Carlo integration) 이라고 한다. 즉 서로 독립적으로 생성된 균등난수 $u_1,\cdots,u_n$을
이용하여

$$y_i=(b-a)g((b-a)u_i+a)\ (i=1,\cdots,n)$$

$$\int_a^b g(x)dx\simeq\frac{1}{n}\sum_{i=1}^{n}y_i,\quad n\to\infty$$

와 같이 정적분의 근사값을 구하는 방법을 몬테칼로적분이라고 한다. 이는

$$\mathrm{Y}_i=(b-a)g((b-a)\mathrm{U}_i+a)\ \ (i=1,\cdots,n)$$

의 관측값을 이용하여 $\mathrm{E}(\mathrm{Y}_1)$에 대한 추측을 하는 것이므로 $\mathrm{E}(\mathrm{Y}_1)$에 대한 점근신뢰구간으로 이러한 근사의 정밀도를 나타낼 수 있다.

예 5.4.2

적분 공식을 이용하면 정적분 $\int_1^3 x^2dx$ 의 값이 26/3임은 잘 알고 있다. 한편 독립적으로 생성된 $\mathrm{U}(0,1)$에서의 난수 $u_1,\cdots,u_n$에 대하여

$$x_i=2u_i+1,\,y_i=2x_i^2\ \ (i=1,\cdots,n)$$

이라고 하여

$$\int_1^3 x^2dx\simeq\frac{2}{n}\sum_{i=1}^{n}x_i^2\ =\frac{1}{n}\sum_{i=1}^{n}y_i$$

와 같이 정적분의 근사값을 구할 수 있다.

R 패키지를 이용하여 생성된 난수들을 이용하여 이 정적분의 근사값과 95% 점근신뢰구간을 구해보면 다음과 같이 주어진다. 이 결과에서 볼 수 있듯이 난수의 개수 n이 커질수록 이러한 근사의 정밀도가 좋아진다.

n	100	1000	10000	100000
$\overline{y}$	8.7612	8.5633	8.6515	8.6473
$\overline{y}-1.96\,s_y/\sqrt{n}$	7.8241	8.2728	8.5604	8.6185
$\overline{y}+1.96\,s_y/\sqrt{n}$	9.6983	8.8538	8.7426	8.6762

표 5.4.1 몬테칼로적분에 의한 정적분 $\int_1^3 x^2dx\,(=8.666\cdots)$의 근사값과 95% 오차한계

예 5.4.3 로지스틱분포의 분산:
로지스틱분포 $\mathrm{L}(0,1)$은 $x=0$에 관해 대칭인 분포로서 그 분산은[8)]

$$\int_{-\infty}^{+\infty} x^2 \frac{e^x}{(1+e^x)^2} dx = \frac{\pi^2}{3}$$

임이 알려져 있다. 한편 예 5.4.1로부터 균등분포 $\mathrm{U}(0,1)$을 따르는 확률변수 U에 대하여

$$\mathrm{X} = \log \frac{\mathrm{U}}{1-\mathrm{U}} \sim \mathrm{L}(0,1)$$

따라서 독립적으로 생성된 $\mathrm{U}(0,1)$에서의 난수 $u_1, \cdots, u_n$에 대하여

$$x_i = \log \frac{u_i}{1-u_i}, \;\; y_i = x_i^2 \;\; (i=1,\cdots,n)$$

이라고 하여 다음과 같이 로지스틱분포 $\mathrm{L}(0,1)$의 분산의 근사값을 구할 수 있다.

$$\int_{-\infty}^{+\infty} x^2 \frac{e^x}{(1+e^x)^2} dx = \mathrm{E}(\mathrm{X}^2) \simeq \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

R 패키지를 이용하여 생성된 난수들을 이용하여 이 정적분의 근사값과 95% 점근신뢰구간을 구해보면 다음과 같이 주어진다.

n	100	1000	10000	100000
$\overline{y}$	2.9775	3.3375	3.3201	3.2949
$\overline{y} - 1.96\, s_y / \sqrt{n}$	2.0043	2.9799	3.2034	3.2584
$\overline{y} + 1.96\, s_y / \sqrt{n}$	3.9506	3.6951	3.4369	3.3313

표 5.4.2 로지스틱분포 $\mathrm{L}(0,1)$의 분산 $\pi^2/3 (= 3.289\cdots)$의 근사값과 95% 오차한계

예 5.4.4 표본비율의 극한분포:
모비율이 $p\,(0<p<1)$인 베르누이시행을 독립적으로 n번 관측한 결과를 $\mathrm{X}_1, \cdots, \mathrm{X}_n$이라고 할 때, 대수의법칙과 중심극한정리로부터 표본비율 $\widehat{p_n} = (\mathrm{X}_1 + \cdots + \mathrm{X}_n)/n$ 에 대하여 다음이 성립하는 것을 알고 있다.

$$\mathrm{p}\lim_{n\to\infty} \widehat{p_n} = p, \quad \frac{\widehat{p_n} - p}{\sqrt{p(1-p)/n}} \xrightarrow[n\to\infty]{d} \mathrm{Z}, \;\; \mathrm{Z} \sim \mathrm{N}(0,1)$$

따라서 슬럿츠키의 정리와 정리 5.3.2로부터 $\dfrac{\widehat{p_n} - p}{\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}} \xrightarrow[n\to\infty]{d} \mathrm{Z}, \;\; \mathrm{Z} \sim \mathrm{N}(0,1)$ 이므로

$$\lim_{n\to\infty} \mathrm{P}\left(-z_{\alpha/2} \le \frac{\widehat{p_n} - p}{\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$\therefore \lim_{n\to\infty} \mathrm{P}\left\{\widehat{p_n} - z_{\alpha/2}\sqrt{\widehat{p_n}(1-\widehat{p_n})/n} \le p \le \widehat{p_n} + z_{\alpha/2}\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}\right\} = 1-\alpha$$

이로부터 모비율 p에 관한 $100(1-\alpha)\%$ 점근(漸近 asymptotic) 신뢰구간을 다음과 같이 나타낼 수 있다.

$$p \in \left[\widehat{p_n} - z_{\alpha/2}\sqrt{\widehat{p_n}(1-\widehat{p_n})/n},\; \widehat{p_n} + z_{\alpha/2}\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}\right]$$

---

8) 부분적분을 이용하고 무한급수로 나타내어 적분하면

$$\int_{-\infty}^{+\infty} x^2 \frac{e^x}{(1+e^x)^2} dx = 4\int_0^{+\infty} \frac{e^{-x}}{1+e^{-x}} x\, dx = 4\int_0^{+\infty} \sum_{n=1}^{\infty} (-e^{-x})^{n-1} e^{-x} x\, dx = 4\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2}$$

한편 후리에급수를 이용한 다음의 항등식에서 $x=0$을 대입하면 위의 무한급수의 값을 얻을 수 있다.

$$x^2 = \frac{1}{3}\pi^2 + 4\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos(nx) \;\; (-\pi \le x \le \pi)$$

예 5.4.5 표본비율의 극한분포를 이용한 원주율 π의 근사:

서로 독립이고 균등분포 $\mathrm{U}(0,1)$을 따르는 두 확률변수 $\mathrm{U}_1,\mathrm{U}_2$에 대하여 이들을 좌표로 하는 점 $(\mathrm{U}_1,\mathrm{U}_2)$을 관측한다는 것은 각 변의 길이가 1인 정사각형 내에 한 점을 랜덤하게 떨어뜨리는 것으로 생각할 수 있다. 이 실험에서 떨어뜨린 점이 사분원내에 떨어질 확률은

$$\mathrm{P}(\mathrm{U}_1^2+\mathrm{U}_2^2 \le 1)=\pi/4$$

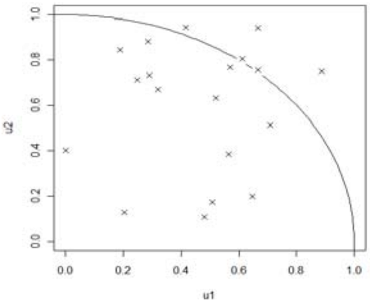

그림 5.4.1랜덤하게 떨어뜨린 점들(u_{i1},u_{i2})

이제 이러한 실험을 독립적으로 반복하여 랜덤하게 떨어뜨린 점들이 사분원내에 떨어지는 상대도수가 이 사건의 확률 $\pi/4$에 가까워진다는 것이 대수의 법칙의 뜻이다. 즉 서로 독립이고 균등분포 $\mathrm{U}(0,1)$을 따르는 두 확률변수의 순서쌍들인 $(\mathrm{U}_{i1},\mathrm{U}_{i2})\,(i=1,\cdots,n)$를 독립적으로 관측하는 실험에서 상대도수

$$\widehat{p_n}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}_{(\mathrm{U}_{i1}^2+\mathrm{U}_{i2}^2 \le 1)}$$

은 $p=\mathrm{P}(\mathrm{U}_1^2+\mathrm{U}_2^2 \le 1)=\pi/4$에 확률수렴한다.

따라서 독립적으로 생성된 $\mathrm{U}(0,1)$에서의 난수 $u_{11},u_{12};\cdots;u_{n1},u_{n2}$에 대하여

$$x_i=u_{i1}^2+u_{i2}^2 \ \ (i=1,\cdots,n)$$

이라고 하면 표본비율의 관측값

$$\widehat{p_n}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}_{(x_i \le 1)}$$

은 확률 $p=\mathrm{P}(\mathrm{U}_1^2+\mathrm{U}_2^2 \le 1)=\pi/4$의 근사값이며 $4\widehat{p_n}$을 $4p=\pi$의 근사값으로 사용할 수 있다. 또한 예 5.4.4로부터 $4p$에 관한 점근신뢰구간으로 근사의 정밀도를 나타낼 수 있다.

R 패키지를 이용하여 생성된 난수들을 이용하여 $4p=\pi$의 근사값과 95% 점근신뢰구간을 구해보면 다음과 같이 주어진다.

n	100	1000	10000	100000
$4\widehat{p_n}$	3.3600	3.1320	3.1344	3.1452
$4\widehat{p_n}-4\times 1.96\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}$	3.0755	3.0308	3.1024	3.1351
$4\widehat{p_n}+4\times 1.96\sqrt{\widehat{p_n}(1-\widehat{p_n})/n}$	3.6445	3.2331	3.1664	3.1552

표 5.4.3 $4p=\pi(=3.14159\cdots)$의 근사값과 95% 오차한계