한은 통계직렬 스터디: 풀이편

2021-15115 권이태

July 3, 2024

Contents

2024년

Problem. 전수조사(census)와 표본조사(sampling)의 차이에 대해서 약술하시오.

Solution. 전수조사(census)는 조사하고자 하는 모집단 전체를 조사하는 반면, 표본조사는 해당 모집단에서 유한한 크기의 표본을 뽑아 표본을 조사함으로써 모집단의 성질을 추정한다. 전수조사는 모집단 전체를 조사하기에 정확한 정보를 얻을 수 있다는 장점이 있는 반면 시간과 비용이 많이 들고, 표본조사는 이를 절약할 수 있지만 추정이 불확실하다는 약점이 있다.

Problem. 회귀분석의 변수선택방법 중 후진제거법(backward elimination)에 대해서 약술하시오.

Solution. 회귀분석은 변수 간의 상관관계를 확인할 수 있는 간단한 방법이지만, 설명변수가 많은 경우 그 해석이 어렵고 과적합 등이 발생할 수 있다. 따라서 다양한 변수선택방법을 통해 적절한 설명변수를 채택하는 방법이 회귀분석에서 자주 이용된다. 그 중 후진제거법(backward elimination)은 전체 변수에 대한 회귀모형을 적합한 뒤 필요없는 변수를 제거해나가면서 변수선택을 진행하는 법을 일컫는다. 그 과정은 아래와 같다.

Step 1 모든 설명변수를 이용한 full model을 적합한다.

Step 2 각 step에서, 변수 중 가장 작은 F값을 가진 변수 하나를 제거한다. 이는 곧 해당 변수가 적합 수준에 큰 변화를 불러오지 않음을 의미한다.

Step 3 미리 정해놓은 변수의 개수 p^*개에 도달하거나, 모든 변수의 F통계량이 미리 정해놓은 임계값 F_{OUT}에 비해 크면 종료한다.

이로써 Full model로부터 원하는 수준의 복잡도나 유의도를 가지는 reduced model을 얻을 수 있다.

Problem. 1부터 6까지의 숫자가 적힌 정육면체인 주사위를 한 번 던졌을 때 값이 1이 나오는 경우의 오즈(odds)를 구하시오.

Solution. 주사위를 한 번 던졌을 때 1이 나올 성공 확률은 1/6이다. 따라서 그 오즈는

$$odds = \frac{1/6}{1-1/6} = \frac{1}{5}$$

으로 주어진다. 문제에서 오즈비를 구하라고 하였는데, 이는 실수로 보인다.

Problem. 완비통계량(complete statistic)에 대해서 약술하시오.

Solution. 모형 $X \sim f(x;\theta), \theta \in \Theta$를 고려할 때, 만약 모든 θ에 대하여

$$\mathbb{E}_\theta[g(Y)] = 0$$

인 함수 g가 영함수와 확률 1로 같다면, Y를 θ의 완비통계량이라 부른다.

Problem. 중심극한정리(central limit theorem)에 대해서 약술하시오.

Solution. 중심극한정리는 $X_1, X_2, \cdots, X_n$이 $\mathbb{E}[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$인 IID 표본일 때,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

라는 정리를 일컫는다. 이는 IID 표본이 아니더라도 여러 좋은 조건을 만족하는 표본 $X_1, \cdots, X_n$에 대해 적용할 수 있다. 이는 좋은 조건을 만족하는 표본들의 표준화된 표본평균이 정규분포로 수렴함을 의미하며, 이에 따라 실제 분포가 정규분포가 아니더라도 표본의 크기가 충분히 크다면 표본평균의 분포를 정규분포로 근사하여 통계적 추정 및 추론을 해낼 수 있다.

Problem. $P(X = x) = kx^3$, $x \in \{1, 2, 3, 4, 5\}$으로 주어진 확률질량함수(probability mass function)에서 X를 이산확률변수라고 하자. $P(X = x)$이 확률질량함수가 되기 위한 적절한 k값을 구하시오.

Solution. $P(X = x)$가 확률질량함수가 되기 위해서는 X의 support가 $\{1, 2, 3, 4, 5\}$이므로

$$1 = \sum_{x=1}^{5} P(X = x)$$

를 만족해야 한다. 그렇다면

$$\begin{aligned}\sum_{x=1}^{5} P(X = x) &= \sum_{x=1}^{5} kx^3 \\ &= k \times \left(\frac{5(5+1)}{2}\right)^2 = 225k\end{aligned}$$

이므로, $k = \frac{1}{225}$여야 한다.

Problem. 앞의 문제에서 구한 값에 따라서 확률변수 X가 $1 \leq X \leq 2$ 구간에 있을 확률을 구하시오.

Solution. 아래와 같이 구할 수 있다.

$$\begin{aligned}P(1 \leq X \leq 2) &= P(X = 1) + P(X = 2) \\ &= \frac{1}{225}1^3 + \frac{1}{225}2^3 \\ &= \frac{9}{225} = \frac{1}{25}\end{aligned}$$

Problem. 이산확률변수 Y, Z의 결합확률분포가 다음과 같다. Y와 Z의 주변확률분포를 구하시오.

		Z		
		10	20	30
Y	10	0.3	0.1	0.2
	20	0.2	0.1	0.1

Solution. 아래와 같이 구할 수 있다. 먼저 Y의 주변확률분포에 대한 확률질량함수 $P_Y(Y=y)$는

$$P_Y(Y=10) = \sum_{z\in\{10,20,30\}} P(Y=10, Z=z) = 0.3+0.1+0.2 = 0.6$$
$$P_Y(Y=20) = \sum_{z\in\{10,20,30\}} P(Y=20, Z=z) = 0.2+0.1+0.1 = 0.4$$

으로 주어지고, support는 $\{10,20\}$이다. Z의 주변확률분포에 대한 확률질량함수 $P_Z(Z=z)$는

$$P_Z(Z=10) = \sum_{y\in\{10,20\}} P(Y=y, Z=10) = 0.3+0.2 = 0.5$$
$$P_Z(Z=20) = \sum_{y\in\{10,20\}} P(Y=y, Z=20) = 0.1+0.1 = 0.2$$
$$P_Z(Z=30) = \sum_{y\in\{10,20\}} P(Y=y, Z=30) = 0.2+0.1 = 0.3$$

이다.

Problem. 앞선 문제에서 $\mathbb{E}[Y]$와 $\mathbb{E}[Z]$를 구하고, Y와 Z는 서로 독립인지 여부와 그 이유를 설명하시오.

Solution. 아래처럼 구할 수 있다.

$$\mathbb{E}[Y] = \sum_{y\in\{10,20\}} yP_Y(Y=y) = 10\times 0.6 + 20\times 0.4 = 14$$
$$\mathbb{E}[Z] = \sum_{z\in\{10,20,30\}} zP_Z(Z=z) = 10\times 0.5 + 20\times 0.2 + 30\times 0.3 = 12$$

한편 YZ의 기대값은

$$\mathbb{E}[YZ] = \sum_{\substack{y\in\{10,20\}\\ z\in\{10,20,30\}}} yzP(Y=y, Z=z) = 30+20+60+40+40+60 = 250$$

으로 계산된다. 만약 Y와 Z가 독립이라면 $\mathbb{E}[YZ] = \mathbb{E}[Y]\times\mathbb{E}[Z]$여야 하지만, 실제로 250과 $168 = 14\times 12$는 다르다. 따라서 둘은 독립이 아니다.

Problem. 앞선 문제에서 $Y=20$일 때 Z의 조건부 기대값을 구하시오.

Solution. 아래처럼 구할 수 있다.

$$\begin{aligned}\mathbb{E}[Z|Y=20] &= \sum_{z\in\{10,20,30\}} zP_Z(Z=z|Y=20)\\ &= \sum_{z\in\{10,20,30\}} z\frac{P(Z=z, Y=20)}{P_Y(Y=20)}\\ &= \frac{1}{0.4}(10\times 0.2 + 20\times 0.1 + 30\times 0.1)\\ &= 17.5\end{aligned}$$

Problem. 확률변수 X가 아래와 같은 균등분포(uniform distribution)을 따를 때, 평균과 분산이 각각

$\mathbb{E}[X] = \frac{1}{2}(a+b)$, $\mathrm{Var}(X) = \dfrac{1}{12}(b-a)^2$임을 보이시오.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{o.w.} \end{cases}$$

Solution. 아래처럼 보일 수 있다.

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b x f_X(x) dx \\ &= \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b \\ &= \frac{1}{b-a} \times \frac{b^2-a^2}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \int_a^b x^2 f_X(x) dx \\ &= \frac{1}{b-a}\left[\frac{1}{3}x^3\right]_a^b \\ &= \frac{1}{3(b-a)}(b^3-a^3) \\ &= \frac{1}{3}(a^2+ab+b^2) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{1}{3}(a^2+ab+b^2) - \frac{1}{4}(a+b)^2 \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Problem. 시행횟수 n이 커짐에 따라 이항분포가 포아송분포로 근사하게 됨을 보이시오.

Solution. 이를 보이기 위해서는 n이 충분히 클 때 성공 확률이 p인 이항분포 $Binom(n,p)$를 따르는 확률변수 X의 확률질량함수 $P_X(X=x)$가 어떤 λ에 대해

$$P_X(X=x) \approx \frac{e^{-\lambda}\lambda^x}{x!}$$

임을 확인하면 충분하다. 그렇다면 아래처럼 전개할 수 있다. $np = \lambda$로 쓸 때, $p = \frac{\lambda}{n} \to 0$이므로,

$$\begin{aligned} P_X(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n \times (n-1) \times \cdots (n-x+1)}{x!} p^x (1-p)^{n-x} \\ &\approx \frac{n^x}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \end{aligned}$$

$$= \frac{e^{-\lambda}\lambda^x}{x!}$$

이다. 따라서 $\lambda = np$가 고정되었을 때 $n \to \infty, p = \frac{\lambda}{n} \to 0$임에 따라 이항분포는 평균이 λ인 푸아송 분포로 근사할 수 있다.

Problem. $n \times 1$ 벡터 $i = (1, 1, \cdots, 1)^T$에 대하여 행렬 $M = I_n - \frac{ii^T}{i^T i}$의 랭크가 $n-1$이고 대칭인 멱등행렬임을 보이시오. 또한 $Mi = 0$임을 보이시오.

Solution. 먼저 행렬 M의 랭크가 $n-1$임을 보이자. rank-nulity theorem에 의하여, $\text{nulity}(M) = n - (n-1) = 1$임을 보이기만 하면 충분하다. 만약 $\mathbf{x} \in \ker(M)$이라면, $M\mathbf{x} = 0_n$이다. 그러면

$$0_n = M\mathbf{x} = I_n\mathbf{x} - \frac{ii^T}{i^T i}\mathbf{x} = \mathbf{x} - \frac{1}{n}(i^T\mathbf{x})i$$

이므로, $\mathbf{x} \parallel i$여야 함을 알 수 있다. 즉 $\ker(M) = \text{span}(\{i\})$이고, $\text{nulity}(M) = \dim(\ker(M)) = 1$이다. 따라서 그 rank는 $n-1$이 된다.
다음으로는 M이 대칭행렬임을 보이자. 이는 간단하게

$$M^T = \left(I_n - \frac{ii^T}{i^T i}\right)^T = I_n^T - \frac{ii^T}{i^T i} = I_n - \frac{ii^T}{i^T i} = M$$

으로 보일 수 있다.
멱등행렬임은

$$\begin{aligned} MM &= \left(I_n - \frac{ii^T}{i^T i}\right)\left(I_n - \frac{ii^T}{i^T i}\right) \\ &= I_n - 2\frac{ii^T}{i^T i} + \frac{ii^T ii^T}{(i^T i)^2} \\ &= I_n - 2\frac{ii^T}{i^T i} + \frac{ii^T}{i^T i} \\ &= I_n - \frac{ii^T}{i^T i} = M \end{aligned}$$

으로 보일 수 있다. 또한

$$Mi = \left(I_n - \frac{ii^T}{i^T i}\right)i = i - i\frac{i^T i}{i^T i} = i - i = 0_n$$

을 얻는다.

Problem. 임의의 표본 $y = (y_1, \cdots, y_n)^T$에 대해 표본평균과 표본분산이 다음과 같음을 보이시오.

$$\bar{y} - \frac{i^T y}{i^T i}, \quad s^2 = \frac{y^T M y}{n-1}$$

Solution. 아래처럼 보일 수 있다.

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{i^T i} i^T y = \frac{i^T y}{i^T i} \end{aligned}$$

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
&= \frac{1}{n-1}(y-\bar{y}i)^T(y-\bar{y}i) \\
&= \frac{1}{n-1}\left(I_n y - \frac{ii^T}{i^Ti}y\right)^T\left(I_n y - \frac{ii^T}{i^Ti}y\right) \\
&= \frac{1}{n-1}(My)^T(My) \\
&= \frac{1}{n-1}y^T(M^TM)y \\
&= \frac{1}{n-1}y^T(MM)y \quad (M:\text{symmetric}) \\
&= \frac{y^TMy}{n-1} \quad (M:\text{idempotent})
\end{aligned}
$$

Problem. $y \sim N(0_n, \sigma^2 I_n)$이면, $\frac{(n-1)s^2}{\sigma^2} = \frac{y^TMy}{\sigma^2} \sim \chi^2_{(n-1)}$임을 보이시오.

Solution. 앞선 문제에서 $s^2 = \frac{y^TMy}{n-1}$임을 보였으므로,

$$\frac{(n-1)s^2}{\sigma^2} = \frac{y^TMy}{\sigma^2}$$

임은 당연하다. 이제 이것이 $\chi^2_{(n-1)}$을 따름을 보이기만 하면 충분하다. 한편 M은 대칭행렬이기에 orthogonal diagonalizable이며, idempotent matrix이기에 eigenvalue로 0 혹은 1만 가진다. 이때 rank가 $n-1$이므로, 1의 geometric multiplicity는 $n-1$, 0의 geometric multiplicity는 dimension이 1이다. 따라서

$$M = Q^T\Lambda Q$$

로 분해 가능하며, Q는 직교행렬, $\Lambda = \text{diag}(1, 1, \cdots, 1, 0)$이다. 따라서

$$y^TMy = y^TQ^T\Lambda Qy = (Qy)^T\Lambda(Qy)$$

이고, $z = Qy$는

$$z \sim N(0_n, \sigma^2 I_n)$$

을 따른다. 이는 $Q0_n = 0_n$, $Q(\sigma^2 I_n)Q^T = \sigma^2 QQ^T = \sigma^2 I_n$이기 때문이다. 그렇다면

$$y^TMy = z^T\Lambda z = \sum_{i=1}^{n-1} z_i^2 = \sigma^2\sum_{i=1}^{n-1}\left(\frac{z_i}{\sigma}\right)^2$$

으로 쓸 수 있으며, 각 z_i/σ는 $i = 1, 2, \cdots, n-1$에 대하여 IID인 표준정규분포를 따르는 확률변수이기에

$$\frac{(n-1)s^2}{\sigma^2} = \frac{y^TMy}{\sigma^2} = \sum_{i=1}^{n-1}\left(\frac{z_i}{\sigma}\right)^2 \sim \chi^2_{(n-1)}$$

을 얻는다.

Problem. $\bar{y}$과 s^2이 서로 독립이라고 가정하고, 다음을 도출하시오.

$$t = \frac{\bar{y}}{\sqrt{\frac{s^2}{n}}} \sim t_{(n-1)}$$

Solution. $\bar{y} = \frac{1}{n} i^T y$이므로,

$$\bar{y} \sim N\left(0, \frac{\sigma^2}{n}\right)$$

를 얻는다. 그렇다면 표준정규분포를 따르는 확률변수 Z에 대하여 $\bar{y} = \frac{\sigma}{\sqrt{n}} Z$처럼 쓸 수 있다. 동일하게 $\chi^2_{(n-1)}$를 따르는 확률변수를 U라 쓸 때

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{\sigma^2 U}{n(n-1)}}$$

으로 쓸 수 있다. 이때 $\bar{y}$와 s^2이 서로 독립이라 하였으므로, Z와 U는 독립이라고 가정할 수 있다. 그러면

$$t = \frac{\bar{y}}{\sqrt{\frac{s^2}{n}}} = \frac{\frac{\sigma}{\sqrt{n}} Z}{\sqrt{\frac{\sigma^2 U}{n(n-1)}}} = \frac{Z}{\sqrt{U/(n-1)}} \sim t_{(n-1)}$$

이 t분포의 정의로부터 얻어진다.

Problem. 확률변수 Y는 아래와 같은 확률밀도함수를 가진 분포를 따른다고 한다.

$$f_Y(y) = \begin{cases} \frac{24}{y^4} & y > 2 \\ 0 & \text{o.w.} \end{cases}$$

$Y_1, Y_2, \cdots, Y_{48}$을 $f_Y(y)$에서 뽑은 표본이라고 가정했을 때, 중심극한정리를 이용하여 $P\left(\sum_{i=1}^{48} Y_i > 180\right)$의 근사값을 구하시오.

Solution.

$$P\left(\sum_{i=1}^{48} Y_i > 180\right) = P(\bar{Y} > 3.75)$$

이다. 또한 Y_i의 평균과 분산은

$$\mathbb{E}[Y_i] = \int_2^\infty y \frac{24}{y^4} dy = \left[-12\frac{1}{y^2}\right]_2^\infty = 3$$

$$\mathbb{E}[Y_i^2] = \int_2^\infty y^2 \frac{24}{y^4} = \left[-24\frac{1}{y}\right]_2^\infty = 12$$

$$\mathrm{Var}(Y_i) = \mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2 = 12 - 9 = 3$$

을 얻고, $Y_1, \cdots, Y_{48}$이 IID 확률변수열이므로 중심극한정리에 의하여

$$\frac{\bar{Y} - 3}{\sqrt{3}} \dot{\sim} N(0, 1)$$

이므로 표준정규분포를 따르는 확률변수 Z에 대하여

$$P(\bar{Y} > 3.75) = P\left(\frac{\bar{Y}-3}{\sqrt{3}} > \frac{3.75-3}{\sqrt{3}}\right)$$
$$\approx P(Z > \frac{\sqrt{3}}{4}) \approx 0.333$$

Problem. 확률변수 Z에 대해서 $Z_1, Z_2, \cdots, Z_n$을 확률밀도함수가 다음과 같은 분포에서 뽑은 표본이라고 가정하자. 단 $\theta > 0$이다.

$$f_Z(z) = \begin{cases} \frac{1}{2\theta} & 0 < z < 2\theta \\ 0 & \text{o.w.} \end{cases}$$

$Z_{(n)}$을 $\max(Z_1, Z_2, \cdots, Z_n)$으로 정의하였을 때 $\mathbb{E}[Z_{(n)}]$을 구하고, θ의 불편추정량을 도출하시오.

Solution. $Z_{(n)}$의 누적분포함수를 $F(z)$라고 하면, $z \le 0$에서는 $F(z) = 0$이고 $z \ge 2\theta$에서는 $F(z) = 1$이며 $0 < z < 1$인 경우

$$\begin{aligned} F(z) &= P(Z_{(n)} \le z) \\ &= P(Z_1 \le z, Z_2 \le z, \cdots, Z_n \le z) \\ &= \prod_{i=1}^{n} P(Z_i \le z) \\ &= \prod_{i=1}^{n} \left(\int_0^z \frac{1}{2\theta} dx\right) \\ &= \prod_{i=1}^{n} \frac{z}{2\theta} \\ &= z^n/(2\theta)^n \end{aligned}$$

을 얻는다. 따라서 이를 미분하면 그 확률밀도함수는

$$f_{Z_{(n)}}(z) = nz^{n-1}(2\theta)^{-n} I_{[0,2\theta]}(z)$$

으로 주어진다. 따라서 기대값은

$$\mathbb{E}[Z_{(n)}] = \int_0^{2\theta} z \times nz^{n-1}(2\theta)^{-n} dz = \frac{n}{(n+1)(2\theta)^n}[z^{n+1}]_0^{2\theta} = \frac{2n\theta}{n+1}$$

이다.

한편 우변을 θ로 만들기 위해 양변에 $\frac{n+1}{2n}$을 곱하면

$$\mathbb{E}\left[\frac{(n+1)Z_{(n)}}{2n}\right] = \theta$$

이다. 따라서 θ의 불편추정량으로 아래를 사용할 수 있다.

$$\hat{\theta} = \frac{(n+1)Z_{(n)}}{2n}$$

Problem. 아래와 같이 절편이 없는, 즉 원점을 통과하는 선형회귀모형을 적합하고자 한다.

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, 2, \cdots, n$$

이때 오차항 ϵ_i는 서로 독립이고 $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$이라고 가정한다. 또한 관측된 설명변수의 벡터를 $X = (x_1, \cdots, x_n)^T$로 쓰자. x_i와 ϵ_i는 서로 독립임을 가정한다. 오차의 제곱합을 최소로 하는 β의 추정치를 $\hat{\beta}$라 할 때, $\hat{\beta}$를 구하시오. 그리고 모델이 참이라고 가정했을 때 $\mathbb{E}[\hat{\beta}|X]$와 $\mathrm{Var}(\hat{\beta}|X)$를 구하시오.

Solution.

$$\hat{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

이므로, 양변을 β로 미분함으로써 normal equation

$$\sum_{i=1}^{n} -2x_i(y_i - \hat{\beta} x_i) = 0$$

을 얻고, 이를 풀어

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

을 얻는다. 모형이 참인 경우,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \middle| X \right] \\
&= \frac{\mathbb{E}[\sum_{i=1}^{n} x_i y_i | X]}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{\sum_{i=1}^{n} x_i \mathbb{E}[y_i|X]}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta
\end{aligned}$$

로 이는 불편추정량이며,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}|X) &= \mathrm{Var}(\hat{\beta} - \beta | X) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} - \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} \middle| X \right) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i (y_i - \beta x_i)}{\sum_{i=1}^{n} x_i^2} \middle| X \right) \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \mathrm{Var}(\sum_{i=1}^{n} x_i \epsilon_i | X) \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} x_i^2 \mathrm{Var}(\epsilon_i) \\
&= \frac{\mathrm{Var}(\epsilon_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

을 얻는다.

Problem. 어떤 전구의 남은 수명을 X라고 할 때, 확률변수 X는 모수가 θ인 지수분포를 따르고(즉, $X \sim \exp(\theta)$) 그 확률밀도함수를 $f_X(x|\theta) = \theta e^{-\theta x}$ $(x \geq 0, \theta > 0)$이라고 하자. 표본 $X_1, X_2, \cdots, X_n$에 대하여 θ의

최우추정량을 $\hat{\theta}_{MLE}$라고 하자. $\bar{X} \neq 0$일 때 $\hat{\theta}_{MLE} = \frac{1}{\bar{X}}$임을 증명하시오. 그리고 n이 3이고 $X_1 = 1, X_2 = 2, X_3 = 3$일 때 그 값을 구하시오.

Solution. 가능도함수는

$$L(\theta; x_i) = \prod_{i=1}^{n} (\theta e^{-\theta x}) I(\theta > 0) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) I(\theta > 0)$$

으로 주어지고, 로그가능도함수는

$$l(\theta; x_i) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

이다. 양번을 θ로 미분하여 일계조건을 구하면

$$\frac{d}{d\theta} l(\theta; x_i) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

을 얻으며, 이로부터

$$\hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$$

를 얻는다. 표본수 n이 3이고 X_1, X_2, X_3가 각각 1, 2, 3인 경우에는 $\bar{X} = 2$이므로

$$\hat{\theta}_{MLE} = \frac{1}{2}$$

을 얻는다.

Problem. 위 문제에서 어떤 전문가가 θ의 분포는 지수분포를 따라야 하고, θ의 평균이 $\frac{1}{3}$이라고 믿는다고 가정하자. 전문가의 믿음에 기초하여 θ의 사전확률분포를 도출하시오.

Solution. 이 전문가는 사전확률분포가

$$\pi(\theta) = 3e^{-3\theta} I(\theta \geq 0)$$

으로 주어진다고 믿고 있다. 이때 문제에서 제시된 지수분포의 정의에서는 그 평균이 모수의 역수임을 상기하자.

Problem. 위 문제에서 $X_1 = 1, X_2 = 2, X_3 = 3$일 때, θ의 사후확률분포를 구하시오.

Solution. θ의 사후확률분포를 $p(\theta|x_i)$라고 하면, 베이즈 법칙에서

$$\begin{aligned}
p(\theta|x_i) &\propto L(\theta; x_i) \times \pi(\theta) \\
&= \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) I(\theta > 0) \times 3e^{-3\theta} I(\theta \geq 0) \\
&= 3\theta^n \exp\left(-\theta \left(\sum_{i=1}^{n} x_i + 3\right)\right) I(\theta > 0) \\
&\propto \theta^{(n+1)-1} \frac{e^{-\theta/(\sum_{i=1}^{n} x_i + 3)^{-1}}}{(\sum_{i=1}^{n} x_i + 3)^{-(n+1)} \Gamma(n+1)} I(\theta > 0)
\end{aligned}$$

이므로 사후분포는

$$\theta|X \sim \Gamma(n+1, (\sum_{i=1}^{n} x_i + 3)^{-1})$$

으로 주어진다. ($n+1$은 shape parameter, $(\sum_{i=1}^{n} x_i + 3)^{-1}$은 scale parameter) 우리의 경우 $n = 3$이고 $\sum_{i=1}^{n} x_i = 6$이므로, 사후분포는 $\Gamma(4, 1/9)$를 따른다.

Problem. 앞선 문제에서 θ의 베이즈 추정량을 $\hat{\theta}_{Bayes}$라고 하자. $X_1 = 1, X_2 = 2, X_3 = 3$일 때 구한 사후확률분포에 기초하여 이를 구하고, $\hat{\theta}_{MLE}$와의 차이를 서술하시오.

Solution. 베이즈 추정량은 사후확률분포의 평균인

$$\hat{\theta}_{Bayes} = 4 \times 1/9 = \frac{4}{9}$$

으로 주어진다. 이는 앞서 구한 $\hat{\theta}_{MLE} = \frac{1}{2}$에 비하여 작다. 이는 표본으로만 구한 최우추정량은 1/2인 것으로부터 알 수 있듯 표본은 θ를 1/2로 추정하나, 베이즈 추정량의 경우 사전적으로 θ의 평균을 그보다 작은 1/3으로 생각하고 있다. θ의 베이즈 추정량 $\hat{\theta}_{Bayes}$는 사전확률분포로부터 사전정보를 얻어 추정량을 얻으므로, 이를 반영하여 표본만 보았을 때보다 더 작은 추정량을 내놓게 된다.

Problem. 어느 전기자동차 공장에서 생산하는 제품의 배터리 수명을 높이기 위해서 출하 시점까지의 노출온도(A)를 세 수준으로 하고, 노출시간(B)를 세 수준으로 설정한 뒤 랜덤한 순서로 실험하여 다음과 같은 배터리 수명 데이터를 얻었다.

		노출온도		
		A_1	A_2	A_3
노출시간	B_1	15	10	5
	B_2	6	13	4
	B_3	9	8	11

위 실험의 데이터 구조식과 조건을 기술하고, 각각의 가설을 설정하시오.

Solution. 데이터 구조식은 아래와 같이 쓸 수 있다.

$$x_{ij} = \mu + a_i + b_j + \epsilon_{ij}, \quad (i,j) \in \{1,2,3\}^2$$

이때

- x_{ij}: A_i와 B_j에서 얻은 측정값
- μ: 실험 전체의 모평균
- a_i: A_i가 주는 효과
- b_j: B_j가 주는 효과
- ϵ_{ij}: A_i와 B_j에서 얻은 측정값의 오차

이때 조건에서는 $\sum_{i=1}^{3} a_i = \sum_{j=1}^{3} b_j = 0$을 가정하며, ϵ_{ij}는 $N(0, \sigma_E^2)$을 따르는 IID 확률변수임을 가정한다. 이때 가설은 노출온도에 대해서는

$$H_0 : a_1 = a_2 = a_3 = 0, \quad H_1 : a_i \neq a_{i'} \text{ for some } i \neq i'$$

, 노출시간에 대해서는

$$H_0 : b_1 = b_2 = b_3 = 0, \quad H_1 : b_j \neq b_{j'} \text{ for some } j \neq j'$$

으로 설정하여 검정할 수 있다.

Problem. 위 문제 상황에서 총변동을 A의 변동, B의 변동, 오차변동으로 분해할 때 아래 비어 있는 세 괄호에 들어갈 수식을 채우고 각 항목의 의미를 설명하시오.

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{\bar{x}})^2 = \sum_i\sum_j(\quad)^2 + \sum_i\sum_j(\quad)^2 + \sum_i\sum_j(\quad)^2$$

Solution. 먼저 아래를 정의하자.

$$\bar{x}_{i\cdot} := \frac{1}{3}\sum_{j=1}^{3} x_{ij}, \quad \bar{x}_{\cdot j} := \frac{1}{3}\sum_{i=1}^{3} x_{ij}$$

그렇다면 아래처럼 분해할 수 있다.

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{\bar{x}})^2 &= \sum_{i=1}^{3}\sum_{j=1}^{3}((\bar{x}_{i\cdot} - \bar{\bar{x}}) + (\bar{x}_{\cdot j} - \bar{\bar{x}}) + (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}))^2 \\ &= \sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 + \sum_i\sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 + \sum_i\sum_j(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \\ &+ 2\sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})(\bar{x}_{\cdot j} - \bar{\bar{x}}) + 2\sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}) \\ &+ 2\sum_i\sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}))\end{aligned}$$

이때

$$\sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})(\bar{x}_{\cdot j} - \bar{\bar{x}}) = \sum_i(\bar{x}_{i\cdot} - \bar{\bar{x}})\sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}}) = 0 \times 0 = 0$$

$$\sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}) = \sum_i(\bar{x}_{i\cdot} - \bar{\bar{x}})\sum_j(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}) = \sum_i(\bar{x}_{i\cdot} - \bar{\bar{x}})(3\bar{x}_{i\cdot} - 3\bar{x}_{i\cdot} - 3\bar{\bar{x}} + 3\bar{\bar{x}}) = 0$$

$$\sum_i\sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}) = \sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})\sum_i(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}) = \sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})(3\bar{x}_{\cdot j} - 3\bar{\bar{x}} - 3\bar{x}_{\cdot j} + 3\bar{\bar{x}}) = 0$$

이므로,

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij} - \bar{\bar{x}})^2 = \sum_i\sum_j(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 + \sum_i\sum_j(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 + \sum_i\sum_j(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

을 얻는다. 첫째 항은 동일한 A_i를 받은 개체들의 평균에 대한 편차제곱합이므로 A에 의한 변동을 의미하며, 둘째 항은 동일한 B_j를 받은 개체들의 평균에 대한 편차제곱합이므로 B에 의한 변동을 의미한다. 마지막 항의 경우, x_{ij}에서 A_i에 의한 효과 $\bar{x}_{i\cdot}$과 B_j에 의한 효과 $\bar{x}_{\cdot j}$을 제외한 뒤 중복하여 빼주어 만든 처리에 무관한 오차항의 편차제곱합이므로, 오차변동을 의미한다.

Problem. 위 문제들을 바탕으로 아래의 분산분석표의 빈칸을 채우시오. 그리고 노출온도와 노출시간에 따른 특성치의 차이가 존재하는지 유의수준 5%로 검정하시오.

Solution.

인자	제곱합 (SS)	자유도 (df)	평균제곱합 (MS)	F비 (F ratio)	F(0.05)
A					
B					
오차					
총변동					

먼저 제곱합 열부터 채우자.

- 인자 A의 제곱합은 공식에 의하여

$$SS_A = \frac{(30^2 + 31^2 + 20^2)}{3} - \frac{81^2}{9} = \frac{74}{3} = 24.7$$

- 인자 B의 제곱합은 공식에 의하여

$$SS_B = \frac{(30^2 + 23^2 + 28^2)}{3} - \frac{81^2}{9} = \frac{26}{3} = 8.7$$

- 총변동은 공식에 의하여

$$SS_T = (15^2 + 10^2 + 5^2 + 6^2 + 13^2 + 4^2 + 9^2 + 8^2 + 11^2) - \frac{81^2}{9} = 108$$

- 오차변동은

$$SS_E = SS_T - SS_A - SS_B = 108 - 24.7 - 8.7 = \frac{224}{3} = 74.7$$

다음으로 자유도는, 인자 A와 B에 대해서는 요인의 개수 3에서 1을 빼 2씩이고, 총변동에 대해서는 $3\times3-1 = 8$이며, 오차는 8에서 $2+2$를 빼 4를 얻는다.
평균제곱합 열은,

- 인자 A의 평균제곱합은 제곱합을 자유도인 2로 나누어 12.3을 얻는다.
- 인자 B의 평균제곱합은 제곱합을 자유도인 2로 나누어 4.3을 얻는다.
- 오차의 평균제곱합은 제곱합을 자유도인 4로 나누어 18.7을 얻는다.

F비는 각 인자에 대하여 인자의 평균제곱합과 오차의 평균제곱합 사이의 비로 결정되며, 우리의 경우 각각

$$\frac{37}{56} \approx 0.7, \frac{13}{56} \approx 0.2$$

를 얻는다.
$F(0.05)$는 인자에 상관없이 자유도가 $2, 4$인 F분포의 95% quantile을 알면 된다. 두 셀 모두 6.9를 넣는다. 검정에서는 $F(0.05)$셀에 있는 값에 비하여 둘 모두 매우 작으므로, 요인 A, B에 따른 배터리 수명의 차이가 존재한다고 말하기 어렵다. 즉 귀무가설을 채택할 만한 충분한 근거가 없다.

Problem. 한 사회학자가 어떤 도시의 1인당 소득을 추정하고자 한다. 도시내 총 30개의 집락에서 추출된 20개 집락에 대해 조사한 결과는 아래와 같다.

집락(i)	거주자수(m_i)	집락당 총 소득(y_i)	집락(i)	거주자수(m_i)	집락당 총 소득(y_i)
1	2	930	11	2	300
2	12	50	12	3	700
3	3	970	13	9	870
4	11	910	14	12	910
5	5	70	15	6	790
6	8	160	16	4	130
7	10	660	17	6	930
8	7	790	18	5	440
9	9	880	19	6	240
10	3	490	20	6	770
				$\sum_{i=1}^{20} m_i = 129$	$\sum_{i=1}^{20} y_i = 11,990$

이 도시의 1인당 소득을 추정하고, 추정분산을 구하시오.

Solution. 집락추출에서 1인당 소득을 추정하는 경우, 아래의 비추정량을 사용한다.

$$\hat{\mu} = \frac{\hat{Y}_{HT}}{\hat{M}_{HT}} = \frac{\sum_{i=1}^{20} \pi_i^{-1} y_i}{\sum_{i=1}^{20} \pi_i^{-1} m_i}$$

이때 $\hat{Y}_{HT}, \hat{M}_{HT}$는 각각 모집단의 총소득과 도시의 총 거주자수에 대한 홀비츠-톰슨 추정량을 의미한다. 여기에서는 집락이 단순확률추출되었다고 생각할 수 있으므로 $\pi_i = \frac{20}{300}$으로 모두 같기에, 이에 따라 1인당 소득은

$$\hat{\mu} = \frac{\sum_{i=1}^{20} y_i}{\sum_{i=1}^{20} m_i} = \frac{11990}{129} = 92.94574$$

으로 추정된다. 한편 추정분산은 전체 도시의 거주자수를 모를 때

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{n_I \bar{m}^2} \left(1 - \frac{n_I}{N_I}\right) \frac{\sum_{i=1}^{n_I} (y_i - \hat{\mu} m_i)^2}{n_I - 1}$$

로 구할 수 있음이 알려져 있다. 여기에서 표본에 포함된 집락의 수 n_I는 20, 전체 집락의 수 N_I는 300이다. $\bar{m}$은 표본에 포함된 집락들에 대한 거주자수의 평균 $\frac{1}{20}\sum_{i=1}^{20} m_i = 6.45$이다. 계산을 통하여,

$$\widehat{\mathrm{Var}}(\hat{\mu}) = 203.40$$

임을 얻는다.

Problem. 위 표의 자료가 도시 거주자들의 소득에 대한 예비표본자료라고 가정하자. 추정오차의 한계를 50으로 할 때의 1인당 평균소득을 추정하려면, 향후 표본조사에서 표본의 크기를 얼마로 해야 하는지 구하시오.

Solution. 유의수준을 0.05라고 하자. 그렇다면 추정오차의 한계는

$$1.96 \times \sqrt{\widehat{\mathrm{Var}}(\hat{\mu})}$$

으로 주어지고, 이를 50으로 제한하려 하기에

$$\widehat{\mathrm{Var}}(\hat{\mu}) \leq \left(\frac{50}{1.96}\right)^2 = 650.77$$

을 만드는 n_I를 찾아야 한다. 이때 우리가 위 문제의 표를 예비표본자료로써 사용하고 있으므로, $\bar{m}$과 $\frac{\sum_{i=1}^{n_I}(y_i-\hat{\mu}m_i)^2}{n_I-1}$를 이들의 것으로 대체하여 사용한다면,

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{N_I - n_I}{n_I N_I} \times \frac{181328.5}{41.6025} = \frac{300 - n_I}{n_I} \times 14.52865 \leq 650.77$$

이므로

$$n_I \geq \frac{300 \times 14.52865}{650.77 + 14.52865} = 6.55$$

따라서 향후 표본조사에서 표본집락의 개수는 최소 7개로 하여야 한다.

Problem. 위 표의 자료가 "층1"의 표본을 나타낸다고 하자. 이때 이웃에 있는 보다 작은 도시를 "층2"로 선정하였다. "층2"에서는 총 200개의 집락에서 10개의 집락을 추출하였다. 추가 자료가 다음 표와 같이 주어졌을 때, 두 도시 전체의 1인당 평균소득을 추정하고 추정분산을 구하시오. 이 표본추출 방법론에 대해 설명하고, 집락추출과 비교하시오.

집락(j)	거주자수(m_j)	집락당 총 소득(y_j)
1	3	200
2	5	500
3	3	100
4	4	200
5	2	800
6	7	100
7	6	300
8	8	700
9	3	200
10	2	700
	$\sum_{j=1}^{10} m_j = 43$	$\sum_{j=1}^{10} y_j = 3,800$

Solution.
여전히 비추정과 동일한 아이디어를 사용하되, 일차표본포함확률 π_i를 조정하면 된다. "층1"의 집락들에게는 $\frac{20}{300}$, "층2"의 집락들에게는 $\frac{10}{200}$을 사용하는 것만 다르다. 따라서

$$\hat{\mu} = \frac{\hat{Y}_{HT}}{\hat{M}_{HT}} = \frac{\frac{300}{20} \times 11990 + \frac{200}{10} \times 3800}{\frac{300}{20} \times 129 + \frac{200}{10} \times 43} = 91.53846$$

을 1인당 평균소득의 추정량으로 얻는다.
한편 추정분산의 경우, 아래처럼 추정할 수 있음이 알려져 있다.

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{\hat{M}_{HT}^2} \sum_{i=1}^{2} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) \frac{\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}m_{ij})^2}{n_i - 1}$$

이때 N_i, n_i는 각각 "층i"에 존재하는 총 집락의 수와 표본집락의 수이며, y_{ij}와 m_{ij}는 "층i"의 j번째 집락에서 얻은 y_j와 m_j이다. 이때 주의할 것은 $\hat{\mu}$는 i에 의존하지 않고, 두 층 모두를 이용하여 구한 91.53846이라는 것이다. 따라서 계산하면

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{(91.53846)^2} \left(\frac{300^2}{20} \left(1 - \frac{20}{300}\right) \times 178982.4 + \frac{200^2}{10} \left(1 - \frac{10}{200}\right) \times 119379.7 \right) = 154.2966$$

을 얻는다.

이러한 층화집락추출법은 여러 층을 두고 해당 층 내에서 집락을 추출한다. 이에 따라 전체 집단에 대해 집락을 추출할 때에 비하여 도시별, 층별 추정량을 특정하게 수 있다는 장점이 있다. 특히 표본이 특정 도시에서만 뽑히는 문제를 층별 집락추출을 함으로써 해결하기 때문에, 표본의 대표성을 확보하고 추정량의 분산을 감소시키는 장점이 있다. 우리의 경우에도 “층2”를 사용하면 도시2에 대한 정보를 얻음으로써 표본의 크기를 증가시키는 동시에 층화를 수행하기에, 추정분산이 일단계 집락추출에서의 203.4보다 작은 154.3 가량이 됨을 알 수 있다.

Problem. 시계열 문제

Solution. 설명편 참고

2023년

Problem. 스피어만의 순위상관계수, 인자분석에서의 공통성, ARCH 모형에 대해 설명하시오.

Solution. 설명편 참고

Problem. Gauss-Markov 정리에 대해 설명하시오.

Solution. 선형모형

$$Y = X\beta + \epsilon$$

이 참이고

- $X \perp\!\!\!\perp \epsilon$
- $\epsilon \sim_{i.i.d} N(0, \sigma_\epsilon^2)$
- X has a full rank

이 만족된다면, 최소제곱추정량 $\hat{\beta}^{OLS} = (X^TX)^{-1}X^TY$는 BLUE(Best Linear Unbiased Estimator)가 된다는 정리이다.

Problem. 분산팽창인자(variance inflation factor)에 대해 설명하시오.

Solution. 회귀분석에서 j번째 설명변수에 대한 분산팽창인자는 아래와 같이 정의된다.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

이때 R_j^2는 j번째 설명변수를 반응변수로, 나머지 $k-1$개의 설명변수를 설명변수로 하여 회귀모형을 적합하였을 때의 결정계수이다. 분산팽창인자가 클수록 해당 j번째 변수가 반응변수에 줄 수 있는 선형적인 영향이 나머지 변수들의 선형결합으로 쉽게 표현될 수 있다는 것이므로, 다중공선성을 가질 확률이 크다는 것이며, 해당 변수의 포함으로 인해 추정량의 분산이 불안정해질 수 있음을 의미한다.

Problem. 다음과 같은 자료생성과정을 가정하자.

$$Y_i = \mu + u_i$$

여기서 u_i는 0.5의 확률로 1 또는 0인 확률변수이며 모수 μ는 1 또는 0임이 알려져 있다. 관측된 n개의 랜덤표본 $Y_1, \cdots, Y_n$을 이용해 가설

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu = 1$$

을 검정하고자 한다. 제1종의 오류율이 0이고 검정력이 0이 아닌 검정과 그 검정력을 구하시오.

Solution. 모수 μ가 0인 경우에는, Y_i는 0.5의 확률로 1 또는 0이다. 반면 모수 μ가 1인 경우에는, Y_i는 0.5의 확률로 1 또는 2이다. 즉 귀무가설 하에서 생각할 수 있는 검정통계량 중 하나는

$$TS = \sum_{i=1}^{n} I(Y_i = 2)$$

이다. 귀무가설 하에서는

$$P(TS = 0) = 1$$

이다. 따라서 이 검정통계량을 기반으로 기각역을 $TS > 0$으로 잡을 수 있다. 그렇다면 귀무가설 하에서는 $P(TS > 0) = 0$이므로 제1종의 오류율은 0이다. 반면 대립가설 하에서는 $P(TS > 0) = 1 - (1 - 0.5)^n = 1 - \frac{1}{2^n} > 0$이므로, 검정력은 $n > 1$에서 0이 아니다. 또한 검정력은 표본수 n에 대해

$$\gamma(n) = 1 - \frac{1}{2^n}$$

으로 써진다.

Problem. 어느 섬유공장에서 처리액의 농도가 섬유탄성도에 미치는 영향을 알아보기 위해 실험을 실시하였다. 처리액 농도(A)를 인자로 하여 4가지 수준(A_1: 5%, A_2: 10%, A_3: 15%, A_4: 20%)을 취하고 각 처리액 농도에서 3회씩, 총 13회의 실험을 랜덤한 순서로 시행하여 다음과 같은 데이터를 얻었다.

	처리액 농도 수준			
실험의 반복	A_1: 5%	A_2: 10%	A_3: 15%	A_4: 20%
	12.0	11.0	7.0	7.5
	12.0	9.0	9.0	8.5
	11.0	10.0	9.0	8.0

위 실험의 데이터 구조식과 조건을 기술하고 아래 분산분석표의 빈칸을 채우시오.

인자	제곱합 (SS)	자유도 (df)	평균제곱합 (MS)	F값
A				
E(오차)				
T(합)				

Solution.
데이터 구조식은

$$x_{ij} = \mu + a_i + \epsilon_{ij}$$

이며, 이때

- x_{ij}: i번째 인자수준 A_i에 대한 실험에서 j번째 반복으로부터 얻은 자료
- μ: x_{ij}의 모평균

- a_i: 인자수준 A_i의 효과
- ϵ_{ij}: x_{ij}의 오차항

이다. 또한 그 조건은 $\sum_{i=1}^{4} a_i = 0$과 $\epsilon_{ij} \sim_{i.i.d} N(0, \sigma_\epsilon^2)$으로 주어진다.
분산분석표에서 A에 의한 변동은

$$SS_A = 3\sum_{i=1}^{4}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 = 3((35/3 - 9.5)^2 + (10 - 9.5)^2 + (25/3 - 9.5)^2 + (8 - 9.5)^2) = \frac{77}{3} \approx 25.67$$

이며, 이때 $\bar{x}_{i\cdot}$은 i번째 인자수준에서 반복을 통해 얻은 평균값, $\bar{\bar{x}}$는 전체 평균이다. 전체 변동은

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{3}(x_{ij} - \bar{\bar{x}})^2 = \frac{73}{2} \approx 31.5$$

이다. 따라서 오차의 변동은

$$SS_E = SS_T - SS_A \approx 5.83$$

을 얻는다. 한편 자유도의 경우, 각각 $3, 8, 11$임을 쉽게 알 수 있다. 평균제곱합은 각 인자의 제곱합을 자유도로 나누어

$$MS_A \approx 8.56, \quad MS_E \approx 0.73$$

을 얻으며, F통계량은 둘의 비인 11.74로 보고된다.

Problem. 위 문제에서 유의수준 $\alpha = 0.05$에서 인자 A에 대한 가설검정을 실시하고, 3수준 A_3에서의 섬유 탄성도 모평균 $\mu(A_3)$에 대한 90% 신뢰구간을 구하시오.

Solution. 먼저 가설검정은 귀무가설과 대립가설이

$$H_0 : a_1 = a_2 = a_3 = a_4 = 0, \quad H_1 : a_i \neq a_{i'} \text{ for some } i \neq i'$$

으로 주어진다. 이 가설에 대한 가설검정은 우리가 앞서 구한 F통계량 11.74가 자유도가 3, 8인 F분포의 95% 퍼센타일 $F_(0.95, 3, 8) = 4.07$을 초과하는지 확인함으로써 가능하다. 우리의 F통계량이 기각역에 포함되므로, 모든 인자 수준에서 효과가 동일하다는 귀무가설을 기각할 수 있다.
그 다음으로 $\mu(A_3) = \mu + a_3$에 대한 90% 신뢰구간은

$$\bar{x}_{3\cdot} \pm t(8, 0.05)\sqrt{\frac{MS_E}{3}}$$

으로 주어지므로,

$$(7.42, 9.25)$$

가 된다.

Problem. 위 문제에서 오차의 분산 σ_ϵ^2의 95% 신뢰구간을 구하시오.

Solution. $\frac{SS_E}{\sigma_\epsilon^2}$는 자유도가 8인 카이제곱분포를 따른다. 따라서 그 95퍼센트 신뢰구간은

$$\left(\frac{SS_E}{\chi^2_{0.975}(8)}, \frac{SS_E}{\chi^2_{0.025}(8)}\right) = (0.33, 2.67)$$

로 주어진다.

Problem. 위 문제에서처럼 처리액 농도가 섬유탄성도에 미치는 영향을 알아보기 위한 또다른 실험으로 처리액 농도를 위와 같은 4가지 수준으로 구분하고 실험일(B)을 랜덤하게 4일 선택하여 총 16회의 실험을 실시하였다. 이 실험의 데이터 구조식 및 조건, 특징을 기술하고, 앞선 실험과의 차이점을 서술하시오.

Solution. 이 실험 디자인에서는 모수인자는 1개이며 실험일(B)는 관측하고자 하는 모수가 아니라 변량인자 1개로 바라보아야 한다. 즉 이는 난괴법 디자인으로 취급하여야 한다. 이 경우 데이터 구조식은

$$x_{ij} = \mu + a_i + b_j + \epsilon_{ij}$$

으로 주어지며, 각각은

- x_{ij}: i번째 인자수준 A_i에 대한 실험에서 j번째 실험일으로부터 얻은 자료
- μ: x_{ij}의 모평균
- a_i: 인자수준 A_i의 효과
- b_j: j번째 실험일 B_j에서의 효과
- ϵ_{ij}: x_{ij}의 오차항

이다. 또한 $\sum_{i=1}^{4} a_i = 0$이도록 설정하며, b_j는 어떠한 모수가 아닌 변량이므로 $b_j \sim_{i.i.d} N(0, \sigma_B^2)$이고 $\mathrm{Cov}(b_j, \epsilon_{ij} = 0)$임을 가정한다. 이때 ϵ_{ij}는 이전과 같이 $N(0, \sigma_\epsilon^2)$을 따르는 랜덤오차이다. 여기에서 특징은 실험일에 따른 변동을 고정된 값이 아니라 확률변수 b_j로 둔다는 것이다. 앞선 실험에서는 각 반복이 동일한 시점에 이루어진다고 가정하여 각 인자수준에서의 반복에 따른 관측값이 동일한 분포를 따르고 동질적이나, 여기에서는 실험일을 블록으로 하여 보기에 다른 블럭에 있는 관측값들이 동질적이라고 말하기 어렵다. 따라서 이원배치법에서처럼 추론하게 된다.

Problem. 한 도매업체는 신상품 수요 조사를 위하여 해당 상품의 업체별 평균 매출액을 추정하고자 한다. 해당 상품은 4개의 대형 체인($N_1 = 24, N_2 = 36, N_3 = 30, N_4 = 30, N = 120$)에서 거래되고 있기 때문에 각 체인을 층으로 하는 층화확률추출을 적용하기 위하여 아래와 같이 총 20개 업체에서의 매출액을 조사하였다.

A 체인	B 체인	C 체인	D 체인
94	91	108	92
90	99	96	110
102	93	100	94
110	105	93	91
	111	93	113
	101		
$n_1 = 4$	$n_2 = 6$	$n_3 = 5$	$n_4 = 5$
$s_1^2 = 78.67$	$s_2^2 = 55.6$	$s_3^2 = 39.5$	$s_4^2 = 112.5$

층화확률추출을 이용하여 평균 매출액 추정값 $\bar{y}_{st}$와 추정분산 $\widehat{\mathrm{Var}}(\bar{y}_s t)$을 구하시오.

Solution. 평균 매출액 추정값은

$$\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{4} N_i \bar{y}_i$$

으로 주어진다. 이때 $\bar{y}_i$는 i번째 체인에서의 표본평균을 의미한다. 따라서

$$\bar{y}_{st} = \frac{1}{120}(24 \times 99 + 36 \times 100 + 30 \times 98 + 30 \times 100) = 99.3$$

이다. 추정분산은

$$\begin{aligned}\widehat{\mathrm{Var}}(\bar{y}_{st}) &= \frac{1}{N^2}\sum_{i=1}^{4}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)\frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}{n_i-1}\\ &= \frac{1}{120^2}\sum_{i=1}^{4}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)s_i^2\\ &= \frac{1}{14400}(9440.4+10008+5925+16875)=2.93\end{aligned}$$

으로 주어진다.

Problem. 위 문제에 대하여, 조사 결과를 단순확률추출의 결과로 보는 경우의 추정값 $\bar{y}$와 추정분산 $\widehat{\mathrm{Var}}(\bar{y})$을 구하시오. 그리고 이를 바탕으로 단순확률추출과 층화확률추출을 비교하시오.

Solution. 추정값은

$$\bar{y}=\frac{1}{120}\sum_{i=1}^{20}\left(\frac{20}{120}\right)^{-1}y_i=99.3$$

으로 동일하다. 그러나 추정분산은

$$\widehat{\mathrm{Var}}(\bar{y})=\frac{1}{20^2}\times 20\left(1-\frac{20}{120}\right)\frac{\sum_{i=1}^{20}(y_i-\bar{y})^2}{20-1}=2.49$$

로 더 작다. 일반적으로는 지금과 같은 비례배정의 경우 층화확률추출의 분산이 작으나, 단순확률추출의 분산이 여기에서는 더 작다. 이는 표본에서 n_i들이 작아 s_i^2가 S_i^2를 잘 추정하지 못하는 것과 층 내의 큰 변동 때문이라고 생각된다.

Problem. $X_1, X_2, \cdots, X_n \sim_{i.i.d} N(\mu, \sigma^2)$이라고 할 때 μ, σ^2의 적률이용추정량, 최대우도추정량을 구하고, 각각이 불편성과 일치성을 가지는지 확인하시오.

Solution. 먼저 적률이용추정량을 구하자.

$$\mu=\mathbb{E}[X]$$

이므로,

$$\hat{\mu}_{MME}=\underset{n}{\mathbb{E}}[X]=\frac{1}{n}\sum_{i=1}^{n}X_i$$

이다. 그 다음

$$\sigma^2=\mathbb{E}[X^2]-(\mathbb{E}[X])^2$$

이므로,

$$\hat{\sigma}^2_{MME}=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

으로 주어진다. 한편

$$\mathbb{E}[\hat{\mu}_{MME}]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

이므로 $\hat{\mu}_{MME}$는 불편추정량이고

$$\mathbb{E}[\hat{\sigma}^2_{MME}]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\,\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}[\bar{X}^2] \\
&= \mathbb{E}[X_i^2] - \mathrm{Var}(\bar{X}) - (\mathbb{E}[\bar{X}])^2 \\
&= (\sigma^2 + \mu^2) - \frac{\sigma^2}{n} - \mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

으로 n이 유한할 때 $\hat{\sigma}^2_{MME}$은 불편추정량이 아니다.

한편 큰 수의 법칙에 의하여

$$\hat{\mu}_{MME} = \underset{n}{\mathbb{E}}[X] \xrightarrow{p} \mu$$

$$\hat{\sigma}^2_{MME} = \underset{n}{\mathbb{E}}[X^2] - (\underset{n}{\mathbb{E}}[X])^2 \xrightarrow{p} (\mu^2 + \sigma^2) - \mu^2 = \sigma^2$$

이다. 따라서 이는 일치추정량이다.

둘째로 최대우도추정량을 구하면, 가능도함수가

$$L(\mu, \sigma^2; x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/2\sigma^2)$$

이고 로그가능도함수가

$$l(\mu, \sigma^2; x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

으로 주어진다. 그렇다면

$$
\begin{aligned}
\frac{\partial}{\partial \mu} l(\mu, \sigma^2; x) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\
\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; x) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2
\end{aligned}
$$

이므로 이들에 대한 일계조건에서

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

을 얻는다. 이들의 불편성과 일치성은 이들이 MME와 같음에 따라 자연스럽게 확인될 수 있으며, $\hat{\mu}_{MLE}$는 불편추정량이면서 일치추정량이지만 $\hat{\sigma}^2_{MLE}$는 일치추정량이면서 편의추정량이다.

Problem. 시계열 문제

Solution. 설명편 참고

Problem. 다음과 같은 표준화된 다중선현회귀모형에 대하여, 아래의 물음에 답하시오.

$$Y = X\beta + \epsilon, \quad \mathrm{Var}(\epsilon) = \sigma^2 I, \quad 1^T Y = 0,\ 1^T X = \mathbf{0}$$

σ^2이 1일 때, 최소제곱추정량 $\hat{\beta}^{OLS}$의 분산을 구하시오. 단,

$$X^TX = \begin{pmatrix} 1 & 4/5 \\ 4/5 & 1 \end{pmatrix}$$

이다.

Solution.

$$\mathrm{Var}(\hat{\beta}^{OLS}) = \sigma^2(X^TX)^{-1} = \frac{25}{9}\begin{pmatrix} 1 & -4/5 \\ -4/5 & 1 \end{pmatrix} = \begin{pmatrix} 25/9 & -20/9 \\ -20/9 & 25/9 \end{pmatrix}$$

Problem. 위 문제에서 회귀계수 β를 추정하기 위하여 주어진 능형모수 $\lambda > 0$ 하에서 $S(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$를 최소화하는 추정량 $\hat{\beta}^{\lambda}$를 구하고자 한다. 능형추정량 $\hat{\beta}^{\lambda}$를 도출하시오.

Solution. $S(\beta)$를 다시 쓰면

$$\begin{aligned} S(\beta) &= (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^T(X^TX + \lambda I)\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^T(X^TX + \lambda I)\beta \end{aligned}$$

이므로, 양변을 β로 미분하면

$$\nabla S(\beta) = -2X^TY + 2(X^TX + \lambda I)\beta$$

을 얻는다. $X^TX + \lambda I$가 strictly positive definite matrix이므로 $S(\beta)$는 convex function이고, $S(\beta)$는 $\nabla S(\beta) = 0$인 점에서 minimize된다. 따라서 우리가 구하는 능형추정량은

$$\hat{\beta}^{\lambda} = (X^TX + \lambda I)^{-1}X^TY$$

으로 주어진다.

Problem. 능형추정량 $\hat{\beta}^{\lambda}$의 불편성에 대해 논하고, 불편추정량이 아닐 경우 모수에 대한 편의를 구하시오.

Solution.

$$\mathbb{E}[\hat{\beta}^{\lambda}] = \mathbb{E}[(X^TX + \lambda I)^{-1}X^TY] = \mathbb{E}[(X^TX + \lambda I)^{-1}X^T(X\beta + \epsilon)] = \mathbb{E}[(X^TX + \lambda I)^{-1}X^TX\beta]$$

이다. 그렇다면

$$\begin{aligned} \mathbb{E}[\hat{\beta}^{\lambda}] &= \mathbb{E}[(X^TX + \lambda I)^{-1}X^TX\beta] \\ &= \mathbb{E}[(X^TX + \lambda I)^{-1}(X^TX + \lambda I)\beta] - \lambda\,\mathbb{E}[(X^TX + \lambda I)^{-1}\beta] \\ &= (I - \lambda(X^TX + \lambda I)^{-1})\beta \end{aligned}$$

이며, $\lambda > 0$일 때 편의추정량이다. 한편 그 편의는

$$\mathrm{bias}(\hat{\beta}^{\lambda}) = \mathbb{E}[\hat{\beta}^{\lambda}] - \beta = (I - \lambda(X^TX + \lambda I)^{-1})\beta - \beta = -\lambda(X^TX + \lambda I)^{-1}\beta$$

으로 주어진다.

Problem. 능형추정량 $\hat{\beta}^{\lambda}$의 분산을 구하여 최소제곱추정량의 분산과 비교하고, 이를 통해 능형추정량이 가지는 성질에 대해 논하시오. 단 능형모수 λ는 1/5로 가정한다.

Solution. 능형추정량의 분산은 $\sigma^2 = 1$으로 가정했을 때

$$\mathrm{Var}(\hat{\beta}^{\lambda}) = (X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$$

이다. λ를 1/5로 가정하면

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}^{\lambda}) &= (X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1} \\
&= \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 4/5 \\ 4/5 & 1 \end{pmatrix} \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix}^{-1} \\
&= \frac{25}{16} \begin{pmatrix} 6/5 & -4/5 \\ -4/5 & 6/5 \end{pmatrix} \begin{pmatrix} 1 & 4/5 \\ 4/5 & 1 \end{pmatrix} \begin{pmatrix} 6/5 & -4/5 \\ -4/5 & 6/5 \end{pmatrix} \\
&= \frac{25}{16} \begin{pmatrix} 14/25 & 4/25 \\ 4/25 & 14/25 \end{pmatrix} \begin{pmatrix} 6/5 & -4/5 \\ -4/5 & 6/5 \end{pmatrix} \\
&= \frac{25}{16} \begin{pmatrix} 68/125 & -32/125 \\ -32/125 & 68/125 \end{pmatrix} \\
&= \begin{pmatrix} 17/20 & -2/5 \\ -2/5 & 17/20 \end{pmatrix}
\end{aligned}$$

이다. 그렇다면

$$\mathrm{Var}(\hat{\beta}^{OLS}) - \mathrm{Var}(\hat{\beta}^{\lambda}) = \begin{pmatrix} 437/180 & -82/45 \\ -82/45 & 437/180 \end{pmatrix} \succeq 0$$

이다. 즉 능형추정량의 분산이 더욱 감소하는 것을 확인해줄 수 있다. 능형추정량은 약간의 편의를 허용하지만, OLS에 비해 추정량의 분산을 감소시키는 성질이 있다. 이는 λI를 X^TX에 추가함으로써 설명변수 사이에 존재하는 다중공선성이 분산을 팽창시키는 것을 방해하기 때문이다.

Problem. 확률벡터 $X = (X_1, X_2)^T$가 다음과 같은 이변량 정규분포를 따른다.

$$X \sim N_2(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad |\rho| < 1, \sigma_1 > 0, \sigma_2 > 0$$

확률벡터 X의 확률밀도함수를 구하시오.

Solution.

$$\begin{aligned}
\mathrm{pdf}_X(x_1, x_2) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_1^2(x_1-\mu_1)^2 - 2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2) + \sigma_2^2(x_2-\mu_2)^2\right) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right)
\end{aligned}$$

Problem. $X_1 = x_1$일 때 위 문제에서 X_2의 조건부분포가

$$N(\mu_2 + \rho\sigma_2(x_1 - \mu_1)/\sigma_1, \sigma_2^2(1-\rho^2))$$

임을 보이시오.

Solution. $(X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1$의 분포는

$$(X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1 \sim N(0, \sigma^2)$$

이고

$$\sigma^2 = \begin{pmatrix} -\rho\sigma_2/\sigma_1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} -\rho\sigma_2/\sigma_1 \\ 1 \end{pmatrix} = \rho^2\sigma_2^2 - 2\rho^2\sigma_2^2 + \sigma_2^2 = \sigma_2^2(1-\rho^2)$$

이다. 그러면 $(X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1$와 X_1은 둘 다 정규분포를 따르면서

$$\mathrm{Cov}((X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1, X_1) = \rho\sigma_1\sigma_2 - \rho\sigma_2\sigma_1^2/\sigma_1 = 0$$

이므로 독립이다. 따라서

$$\begin{aligned} \mathrm{cdf}_{X_2|X_1=x_1}(x_2) &= P(X_2 \le x_2 | X_1 = x_1) \\ &= P((X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1 \le (x_2 - \mu_2) - \rho\sigma_2(x_1 - \mu_1)/\sigma_1 | X_1 = x_1) \\ &= P((X_2 - \mu_2) - \rho\sigma_2(X_1 - \mu_1)/\sigma_1 \le (x_2 - \mu_2) - \rho\sigma_2(x_1 - \mu_1)/\sigma_1) \\ &= \Phi\left(\frac{(x_2 - \mu_2) - \rho\sigma_2(x_1 - \mu_1)/\sigma_1}{\sigma_2\sqrt{1-\rho^2}}\right) \end{aligned}$$

이고 미분하면

$$\mathrm{pdf}_{X_2|X_1=x_1}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{(x_2 - (\mu_2 + \rho\sigma_2(x_1 - \mu_1)/\sigma_1))^2}{2\sigma_2^2\sqrt{1-\rho^2}}\right)$$

을 얻는다. 따라서 X_2의 조건부분포는

$$X_2 | X_1 = x_1 \sim N(\mu_2 + \rho\sigma_2(x_1 - \mu_1)/\sigma_1, \sigma_2^2(1-\rho^2))$$

이다.

Problem. 앞선 문제에서

$$\mu = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 16 & 12 \\ 12 & 25 \end{pmatrix}$$

를 가정할 때 $P(5 < X_2 < 12 | X_1 = 7)$을 계산하시오.

Solution. 앞에서 구한 것처럼

$$X_2 | X_1 = x_1 \sim N(1 + 0.6 \times 5(7-3)/4, 25(1 - 0.6^2)) = N(4, 16)$$

이므로,

$$P(5 < X_2 < 12 | X_1 = 7) = \Phi(2) - \Phi(1/4) = 0.5759$$

Problem. 만약

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

이라면 $\mathbb{E}[\max(X_1, X_2)]$의 값은?

Solution. X_1, X_2가 서로 독립인 표준정규분포이므로, $X_3 = \max(X_1, X_2)$의 CDF는

$$\begin{aligned} F_{X_3}(x) &= P(X_3 \leq x) \\ &= P(\max(X_1, X_2) \leq x) \\ &= P(X_1 \leq x, X_2 \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x) \\ &= \Phi^2(x) \end{aligned}$$

이고, pdf는 $2\phi(x)\Phi(x)$으로 주어진다.

$$\begin{aligned} \mathbb{E}[\max(X_1, X_2)] &= \mathbb{E}[X_3] \\ &= \int_{-\infty}^{\infty} 2x\phi(x)\Phi(x)dx \\ &= -2\int_{-\infty}^{\infty} \phi'(x)\Phi(x)dx \\ &= -2[\phi(x)\Phi(x)]_{-\infty}^{\infty} + 2\int_{-\infty}^{\infty} \phi(x)\phi(x)dx \\ &= 2\int_{-\infty}^{\infty} \frac{1}{2\pi}\exp(-x^2)dx \\ &= \frac{2 \times 2^{-1/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}2^{-1/2}}\exp(-x^2/(2 \times (2^{-1/2})^2))dx \\ &= \frac{\sqrt{2}}{\sqrt{2\pi}} = \frac{1}{\sqrt{\pi}} \end{aligned}$$

을 얻게 된다.

2022년

Problem. 다중선형회귀모형 $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1},\ \epsilon_{n\times 1} \sim N(0, \sigma^2 I_n)$에 대하여, β를 최소제곱법을 활용하여 추정할 경우 오차제곱합 SSE가 $Y^T M_X Y$임을 보이시오. (단, $M_X = I_n - X(X^TX)^{-1}X^T$)

Solution.

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - X_i^T(X^TX)^{-1}X^TY)^2 \\
&= (M_XY)^T(M_XY) \\
&= Y^T(M_X^TM_X)Y \\
&= Y^TM_XY
\end{aligned}$$

(M_X: symmetric and idempotent)

Problem. 아래의 정리를 적용하여 SSE/σ^2가 자유도 $(n-p)$의 중심 χ^2 분포를 따름을 보이고 MSE가 분산 σ^2의 불편추정량임을 밝히시오.

> 정리. 확률벡터 y가 평균벡터 μ이고, 분산-공분산행렬이 Σ인 정규분포를 따를 때$(y \sim N(\mu, \Sigma))$, $P\Sigma$가 멱등행렬(idempotent matrix)일 경우 y의 이차형식 $y'Py$는 자유도가 $P\Sigma$의 계수(rank)이고, 비중심모수가 $\lambda = \frac{1}{2}\mu' P\mu$인 비중심 χ^2분포를 따른다. (행렬 A가 $A \cdot A = A$일 경우 A를 멱등행렬이라고 한다.)

Solution.

$$\frac{\text{SSE}}{\sigma^2} = \left(\frac{Y}{\sigma}\right)^T M_X \left(\frac{Y}{\sigma}\right)$$

이고 $Y/\sigma \sim N(X\beta, I_n)$이다. 또한 M_X는 멱등행렬이다. 따라서 주어진 정리에 의하여

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(\text{rank}(M_X))$$

이다. 그런데 멱등행렬의 랭크는 트레이스와 동일하므로, 자유도는

$$\text{rank}(M_X) = \text{tr}(M_X) = \text{tr}(I_n - X(X^TX)^{-1}X^T) = n - p$$

이다. 따라서 $\mathrm{SSE}/\sigma^2 \sim \chi^2(n-p)$이다. 또한 이로부터

$$\mathbb{E}\left[\frac{\mathrm{SSE}}{\sigma^2}\right] = n - p$$

를 얻으므로, 정리하면

$$\mathbb{E}[\mathrm{MSE}] = \mathbb{E}\left[\frac{\mathrm{SSE}}{n-p}\right] = \sigma^2$$

이다. 따라서 MSE는 σ^2의 불편추정량이다.

Problem. $\beta = (\beta_1^T, \beta_2^T)^T$일 때, $p-r$차원 벡터 β_2에 대하여 $H_0 : \beta_2 = 0$임을 검정하고자 한다. 다음의 분산분석표를 활용하여 F 검정통계량을 계산하고 유의수준 $\alpha = 0.05$ 하에서 검정하시오.

. 분산분석표(ANOVA table)

	오차제곱합(SSE)	자유도(df)	평균제곱합(MSE)
$Reduced$모형	672	22	30.5
$Full$모형	480	20	24

주: $Full$은 제약을 부과하지 않은 모형, $Reduced$는 제약을 부과한 모형을 나타냄.

Solution.
F 검정통계량은

$$F = \frac{(\mathrm{SSE}_{\mathrm{Reduced}} - \mathrm{SSE}_{\mathrm{Full}})/(\mathrm{df}_{\mathrm{Reduced}} - \mathrm{df}_{\mathrm{Full}})}{\mathrm{SSE}_{\mathrm{Full}}/\mathrm{df}_{\mathrm{Full}}} = \frac{192/2}{480/20} = 4$$

으로 나타나며, 이는 귀무가설 하에서 $F(2,20)$을 따른다. 기각역은 $F_{0.05}(2,20) = 3.493$으로 4가 더 크기에, 유의수준 0.06에서 $\beta_2 = 0$이라는 귀무가설을 기각한다.

Problem. 주사위와 동전을 던지는 실험을 실시하고자 한다. 주사위를 120번 던진 실험의 결과가 아래와 같을 때 각 주사위 면이 나올 확률이 모두 동일한지 유의수준 $\alpha = 0.05$ 하에서 검정하시오.

주사위 눈	1	2	3	4	5	6
빈도수	18	23	16	21	18	24

Solution.
아래처럼 카이제곱검정을 수행할 수 있다. 검정통계량은

$$U = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{4+9+16+1+4+16}{20} = 2.5$$

로, 기각역과 비교할 때 $U < \chi^2_{0.05}(5) = 11.07$이므로 귀무가설을 기각하지 못한다. 따라서 각 주사위 면이 나올 확률이 모두 동일하다는 귀무가설을 기각할 만한 충분한 근거가 없다.

Problem. 동전 및 주사위의 각 면이 나올 확률이 공정하다고 가정하고 게임을 실시한다. 게임은 두 단계로 이루어지며 동전을 한 번 던지고 추가로 주사위를 여섯 번 던진다. 게임참여자는 아래의 조건 중 한 가지라도 만족할 경우 승리한다. 동 참여자가 승리할 확률을 구하시오.
조건:

- 동전이 앞면일 경우 던진 여섯 번의 주사위 중 5 이상이 네 번 이상 나타남
- 동전이 뒷면일 경우 던전 여섯 번의 주사위 중 5 미만이 다섯 번 이상 나타남

Solution. 구하는 확률은

$$p = \frac{1}{2} \times \sum_{i=4}^{6} \binom{6}{i} (1/3)^i (2/3)^{6-i} + \frac{1}{2} \times \sum_{i=5}^{6} \binom{6}{i} (2/3)^i (1/3)^{6-i} = 0.22565$$

이다.

Problem. 동전 던지기와 같이 각 시행에서 성공할 확률이 θ이고 실패할 확률이 $1-\theta$인 n번의 독립시행을 실행한다고 하자. x를 n번의 시행에서 성공한 횟수라고 할 때 x는 모수가 (n, θ)인 이항확률분포가 된다. $\lambda = n\theta$는 상수이면서 $n \to \infty, \theta \to 0$일 때 아래와 같은 이항확률분포의 극한분포가 포아송분포를 따름을 보이시오.

Solution. 아래처럼 증명할 수 있다.

$$\begin{aligned}
p(x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\
&= \frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} \\
&= \frac{n!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&\approx \frac{\lambda^x}{x!} e^{-\lambda} \times \frac{n!}{(n-x)!n^x} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} e^{-\lambda} \times \frac{1 \times (1 - 1/n) \times \cdots \times (1 - (x-1)/n)}{(1 - \lambda/n) \times (1 - \lambda/n) \times \cdots \times (1 - \lambda/n)}
\end{aligned}$$

고정된 x에서 뒤의 항은 $n \to \infty$임에 따라 1으로 수렴하므로,

$$p(x) \overset{n \to \infty}{\to} \frac{\lambda^x}{x!} e^{-\lambda}$$

이고 극한분포는 포아송분포가 된다.

Problem. $\{y_i\}_{i=1}^n$는 $\{Y_j\}_{j=1}^N$의 값을 갖는 모집단에서 추출한 단순확률표본이라고 하자. 모집단의 평균을 μ, 분산을 σ^2이라고 할 때,

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1}\right)$$

임을 보이시오.

Solution. 먼저, 아래를 얻는다.

$$\begin{aligned}
\mathrm{Var}(y_i) &= \sigma^2 \\
\mathrm{Cov}(y_i, y_j) &= \mathbb{E}[y_i y_j] - \mu^2 \\
&= \frac{1}{N-1} \mathbb{E}[y_i (N\mu - y_i)] - \mu^2 \\
&= \frac{N}{N-1} \mu^2 - \frac{1}{N-1} \mathbb{E}[y_i^2] - \mu^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N-1}\mu^2 - \frac{1}{N-1}(\sigma^2+\mu^2) \\
&= -\frac{1}{N-1}\sigma^2
\end{aligned}
$$

따라서

$$
\begin{aligned}
\mathrm{Var}(\bar{y}) &= \frac{1}{n^2}\,\mathrm{Var}(y_1+y_2+\cdots+y_n) \\
&= \frac{1}{n^2}\left(n\sigma^2 - \frac{n(n-1)}{N-1}\sigma^2\right) \\
&= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)
\end{aligned}
$$

Problem. $\dfrac{N-n}{N}\cdot\dfrac{s^2}{n}$이 $\mathrm{Var}(\bar{y})$에 대한 불편추정량임을 보이시오.

Solution.

$$
\begin{aligned}
\mathbb{E}\left[\frac{(N-n)s^2}{Nn}\right] &= \frac{N-n}{Nn}\,\mathbb{E}[s^2] \\
&= \frac{N-n}{Nn(n-1)}\,\mathbb{E}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\
&= \frac{N-n}{Nn(n-1)}\,\mathbb{E}\left[\sum_{i=1}^{n}(y_i-\mu)^2 - n(\bar{y}-\mu)^2\right] \\
&= \frac{N-n}{N(n-1)}(\mathbb{E}[(y_i-\mu)^2] - \mathbb{E}[(\bar{y}-\mu)^2]) \\
&= \frac{N-n}{N(n-1)}(\sigma^2 - \mathrm{Var}(\bar{y})) \\
&= \frac{N-n}{N(n-1)} \times \frac{N(n-1)}{n(N-1)}\sigma^2 \\
&= \frac{N-n}{n(N-1)}\sigma^2 = \mathrm{Var}(\bar{y})
\end{aligned}
$$

Problem. 한 지역의 가구당 월평균소비지출을 표본조사를 통해 파악하려고 한다. 총 조사대상 가구 $N = 3000$으로부터 $n = 300$인 단순임의표본을 추출하였다. 표본평균이 $\bar{y} = 320$이며 표본분산은 $s^2 = 1470$이다. 동 지역 가구의 월평균소비지출액 μ를 추정하고, 동 추정량 표준오차의 한계를 구하시오.

Solution.

$$
\hat{\mu} = \frac{\sum_{i=1}^{300}\pi_i^{-1}y_i}{\sum_{i=1}^{300}\pi_i^{-1}} = \frac{\sum_{i=1}^{300}y_i}{300} = \bar{y} = 320
$$

$$
2\times\widehat{\mathrm{s.e}}(\hat{\mu}) = 2\times\widehat{\mathrm{Var}}(\bar{y}) = 2\times\frac{3000-300}{3000}\times\frac{s^2}{300} = 8.82
$$

Problem. 위의 표본조사에서 단순임의추출 대신 전체 가구를 소득분위별로 나누어 층화확률표본을 추출하였다. 전에 표본에서 각 층별 표본수가 차지하는 비율이 일정하다고 할 때 전체 가구 월평균소비지출액의 추정평균오차를 구하시오.

$$
s_1^2 = 540,\ s_2^2 = 1350,\ s_3^2 = 1620,\ s_4^2 = 1350,\ s_5^2 = 540
$$

Solution.

$$s^2 = \frac{1}{25}\sum_{i=1}^{5}\widehat{\mathrm{Var}}(\bar{y}_i) = \frac{1}{25}\sum_{i=1}^{5}\frac{N_i - n_i}{N_i} \times \frac{s_i^2}{n_i}$$

층화추출을 분위별로 수행하고 전체 표본에서 그 비율이 일정하도록 하였으므로 $N_i = 600, n_i = 60$이며, 이에 따라

$$s^2 = \frac{3}{5000}\sum_{i=1}^{5} s_i^2 = \frac{81}{25}$$

이다. 따라서 구하는 추정표본오차는 $s = 9/5 = 1.8$이다.

Problem. 모수가 λ인 포아송분포를 따르는 서로 독립인 랜덤표본 $X_1, \cdots, X_n$에 대하여, $X_1, \cdots, X_n$이 단조가능도비를 가짐을 보이시오.

Solution. 가능도함수는

$$L(\lambda; x) = \prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} = e^{-n\lambda} \times \lambda^{\sum_{i=1}^n x_i} \Big/ \prod_{i=1}^{n} x_i!$$

이며, $\lambda_1 < \lambda_2$에 대하여

$$\frac{L(\lambda_1; x)}{L(\lambda_2; x)} = e^{n(\lambda_2 - \lambda_1)}\left(\frac{\lambda_1}{\lambda_2}\right)^{\sum_{i=1}^n x_i}$$

으로 각 x_i에 대하여 감소함수가 된다. 따라서 $X_1, \cdots, X_n$이 단조가능도비를 가진다.

Problem. $H_0 : \lambda \leq \lambda_0,\ H_1 : \lambda > \lambda_0$를 검정할 때 전역최강력검정을 구하시오. 단, 전역최강력검정의 정의를 포함하여 기술하시오.

Solution. 우리의 상황에서 어떠한 검정 ϕ^{UMP}이 유의수준 α에서 전역최강력검정이라는 것은,

- $\sup_{\lambda \leq \lambda_0} \mathbb{E}_\lambda[\phi^{\mathrm{UMP}}(X)] = \alpha$
- 모든 유의수준 α인 검정 ϕ와 $\lambda_1 > \lambda_0$에 대하여,

$$\mathop{\mathbb{E}}_{\lambda_1}[\phi^{\mathrm{UMP}}(X)] \geq \mathop{\mathbb{E}}_{\lambda_1}[\phi(X)]$$

임을 의미한다. 단조가능도비가 있는 경우,

$$\phi^*(x) = \begin{cases} 1 & \sum_{i=1}^n x_i > c \\ \gamma & \sum_{i=1}^n x_i = c \\ 0 & \sum_{i=1}^n x_i < c \end{cases}$$

형태의 검정을 유의수준 α로 설계하는 경우 유의수준 α에서의 전역최강력검정이다. 한편 독립적인 푸아송분포들의 합

$$\sum_{i=1}^{n} X_i$$

는 모수가 $n\lambda$인 푸아송분포를 따르므로, 모수가 $n\lambda_0$인 푸아송 분포에서

$$P\left(\sum_{i=1}^{n} X_i < c\right) \leq \alpha$$

가 되도록 하는 모수가 $n\lambda_0$인 포아송 분포 U의 $1-\alpha$ quantile $c_\alpha = \min_{c\in\mathbb{Z}}(P(U \geq c) \geq 1-\alpha)$을 구한 뒤

$$\phi^*(x) = \begin{cases} 1 & \sum_{i=1}^n x_i > c_\alpha \\ (\alpha - P(U \leq c_\alpha - 1))\,/P(U = c_\alpha) & \sum_{i=1}^n x_i = c_\alpha \\ 0 & \sum_{i=1}^n x_i < c_\alpha \end{cases}$$

으로 세우면 이것이 전역최강력검정이다.

다음과 같은 다중선형회귀모형

$$Y = X\beta + \epsilon$$

에 대하여, $\epsilon \sim N(0, \sigma^2 I_n)$인 경우 β의 최소제곱추정량의 분포를 구하고, 그 추정량이 BLUE임을 증명하시오.

Solution.

$$\hat{\beta} = (X^TX)^{-1}X^TY \sim N(\beta, \sigma^2(X^TX)^{-1})$$

임은 증명 없이 넘어가려 한다. 이 추정량이 BLUE임을 보이자. 어떠한 다른 선형불편추정량 $\tilde{\beta} = CY$를 고려하자. 이때 $C \in \mathbb{R}^{p\times n}$이다. 따라서 $CX = I_p$이며, 불편추정량의 정의에서

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[CY] = CX\beta$$

이므로, $CX\beta = \beta$이다. 한편 이 추정량의 분산은

$$\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(CY) = C\,\mathrm{Var}(Y)C^T = \sigma^2 CC^T$$

이며,

$$\begin{aligned} \mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}) &= \sigma^2 CC^T - \sigma^2((X^TX)^{-1}) \\ &= \sigma^2(CC^T - (X^TX)^{-1}) \\ &= \sigma^2(C - (X^TX)^{-1}X^T)(C - (X^TX)^{-1}X^T)^T \end{aligned}$$

으로 나타난다. 그러므로 모든 p차원 벡터 γ에 대하여

$$\gamma^T(\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}))\gamma = \sigma^2\gamma^T(C-(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T)^T\gamma = \sigma^2\|\gamma^T(C-(X^TX)^{-1}X^T)\|^2 \geq 0$$

이므로 $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}) \succeq 0$이고, 최소제곱추정량 $\hat{\beta}$가 BLUE이다.

Problem. 만약 다중선형회귀모형에서 $\epsilon \sim N(0, V),\ V = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$이라면 최소제곱추정량은 BLUE인가? 그렇지 않다면 BLUE를 구하시오.

Solution. 이 경우 OLS는 BLUE가 아니다. 이 경우 WLS

$$\hat{\beta} = (X^TVX)^{-1}X^TVY$$

가 BLUE임이 잘 알려져 있다.

Problem. 다양한 기타분야 통계학 문제들

Solution. 설명편 참고

Problem. 베르누이분포를 따르는 확률변수에 대한 베이즈통계 문제

Solution. 설명편 참고

Problem. $X_1, \cdots, X_n$은 다음의 동일한 확률밀도함수를 갖는 i.i.d. 확률변수이다.

$$f(x) = \begin{cases} \frac{1}{2\theta} & |x| < \theta \\ 0 & \text{o.w.} \end{cases}$$

θ의 최대우도추정량 $\hat{\theta}_n$을 구하여라.

Solution. 가능도함수는

$$L(\theta; x) = \prod_{i=1}^{n} \frac{1}{2\theta} I(|x_i| < \theta) = \frac{1}{2^n \theta^n} I(\theta \geq \max\{|x_1|, \cdots, |x_n|\})$$

이므로, 최대우도추정량은

$$\hat{\theta}_n = \max\{|X_1|, |X_2|, \cdots, |X_n|\}$$

이다.

Problem. 위 문제에서 $(\theta - \hat{\theta}_n)$의 누적분포함수를 구하고, $C_n(\theta - \hat{\theta}_n)$이 0이 아닌 분산을 갖는 극한분포로 수렴하게 하는 C_n과 그 극한분포를 구하시오.

Solution. 먼저 누적분포함수 $F_n(x)$를 구하자.

$$\begin{aligned} F_n(x) &= P((\theta - \hat{\theta}_n) \leq x) \\ &= P(\hat{\theta}_n \geq \theta - x) \\ &= 1 - P(\hat{\theta}_n < \theta - x) \\ &= 1 - P(\max_i |X_i| < \theta - x) \\ &= 1 - P(x - \theta < X_1 < \theta - x, \cdots, x - \theta < X_n < \theta - x) \\ &= 1 - \prod_{i=1}^{n} P(x - \theta < X_i < \theta - x) \\ &= \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - \left(\frac{\theta - x}{\theta}\right)^n & \text{if } 0 < x \leq \theta \\ 1 & \text{if } \theta < x \end{cases} \end{aligned}$$

한편 이로부터 확률밀도함수를 구하면,

$$f_n(x) = \frac{d}{dx}F_n(x) = \frac{n}{\theta}\left(\frac{\theta - x}{\theta}\right)^{n-1} I(0 < x < \theta)$$

이며 기대값은

$$\begin{aligned}
\mathbb{E}[\theta - \hat{\theta}_n] &= \int_0^\theta x\frac{n}{\theta}\left(\frac{\theta - x}{\theta}\right)^{n-1} dx \\
&= \frac{n}{\theta^n}\left(\int_0^\theta x(\theta - x)^{n-1}dx\right) \\
&= \frac{n}{\theta^n}\left(\theta\int_0^\theta (\theta - x)^{n-1}dx - \int_0^\theta (\theta - x)^n dx\right) \\
&= \frac{n}{\theta^n}\left(\frac{\theta^{n+1}}{n} - \frac{\theta^{n+1}}{n+1}\right) \\
&= \frac{1}{n+1}\theta
\end{aligned}$$

, 이차적률은

$$\begin{aligned}
\mathbb{E}[(\theta - \hat{\theta}_n)^2] &= \int_0^\theta x^2\frac{n}{\theta}\left(\frac{\theta - x}{\theta}\right)^{n-1} dx \\
&= \frac{n}{\theta^n}\int_0^\theta x^2(\theta - x)^{n-1}dx \\
&= \frac{n}{\theta^n}\left(\int_0^\theta (\theta - x)^{n+1}dx - 2\theta\int_0^\theta (\theta - x)^n dx + \theta^2\int_0^\theta (\theta - x)^{n-1}dx\right) \\
&= \frac{n}{\theta^n}\left(\frac{\theta^{n+2}}{n+2} - 2\frac{\theta^{n+2}}{n+1} + \frac{\theta^{n+2}}{n}\right) \\
&= \frac{2}{(n+1)(n+2)}\theta^2
\end{aligned}$$

이므로 그 분산은

$$\mathrm{Var}(\theta - \hat{\theta}^n) = \frac{n}{(n+1)^2(n+2)}\theta^2$$

이다. 따라서 C_n

$$\frac{(n+1)\sqrt{n+2}}{\sqrt{n}}$$

과 같은 $\Theta(n)$ 수준으로 설정하면 극한분포를 구할 수 있음을 예상할 수 있다. $C_n = n$으로 설정하면, $n(\theta - \hat{\theta}_n)$의 누적분포함수 $\tilde{F}_n(x)$는

$$\begin{aligned}
\tilde{F}_n(x) &= P(n(\theta - \hat{\theta}_n) \le x) \\
&= P((\theta - \hat{\theta}_n) \le x/n) \\
&= F_n(x/n) \\
&= \begin{cases} 0 & \text{if } x/n \le 0 \\ 1 - \left(\dfrac{\theta - x/n}{\theta}\right)^n & \text{if } 0 < x/n \le \theta \\ 1 & \text{if } \theta < x/n \end{cases}
\end{aligned}$$

을 만족하며 $0 < x/n < \theta$일 때

$$\lim_{n\to\infty} \tilde{F}_n(x) = \lim_{n\to\infty} 1 - \left(\frac{\theta - x/n}{\theta}\right)^n = \lim_{n\to\infty} 1 - \left(1 - \frac{x}{n\theta}\right)^n = 1 - e^{-x/\theta}$$

이므로 $n(\theta - \hat{\theta}_n)$는 평균이 θ인 지수분포를 극한분포로 가진다.

2021년

Problem. X와 Y의 결합확률분포가 다음과 같다. 각각의 주변확률분포와 $\mathbb{E}[X], \mathbb{E}[X^2], \mathrm{Var}(X), \mathbb{E}[Y], \mathbb{E}[X+Y]$를 구하고, 독립인지 판정하시오.

Y \ X	0	2	4
1	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{3}{30}$
2	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{3}{30}$
3	$\frac{2}{30}$	$\frac{8}{30}$	0

Solution. 주변확률분포는 각각의 확률밀도함수가

$$P(X=a)=\begin{cases}\frac{3}{10} & a=0\\ \frac{1}{2} & a=2\\ \frac{1}{5} & a=4\end{cases}$$

$$P(Y=b)=\begin{cases}\frac{4}{15} & b=1\\ \frac{2}{5} & b=2\\ \frac{1}{3} & b=3\end{cases}$$

으로 나타난다. 한편

$$\begin{aligned}\mathbb{E}[X]&=\frac{9}{5}\\ \mathbb{E}[X^2]&=\frac{26}{5}\\ \mathrm{Var}(X)&=\frac{49}{25}\\ \mathbb{E}[Y]&=\frac{5}{3}\\ \mathbb{E}[X+Y]&=\frac{52}{15}\end{aligned}$$

한편 $P(X=4,Y=3)=0$으로 $P(X=4)P(Y=3)=\frac{1}{15}$와 다르기에, X와 Y는 서로 독립이 아니다.

Problem. X와 X_n은 각각 하나의 확률변수와 확률변수열이다. 모든 $\epsilon > 0$에 대해

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

이면 X_n은 X에 확률수렴한다고 하고, $X_n \xrightarrow{p} X$로 표기한다. $X_n \xrightarrow{p} X$이고 $Y_n \xrightarrow{p} Y$일 때, 상수 a에 대해

$$X_n + Y_n \xrightarrow{p} X + Y$$
$$aX_n \xrightarrow{p} aX$$

임을 보이시오.

Solution. 모든 $\epsilon > 0$에 대하여,

$$\begin{aligned} 0 \leq \lim_{n\to\infty} P(|X_n + Y_n - X - Y| \geq \epsilon) &\leq \lim_{n\to\infty} P(|X_n - X| + |Y_n - Y| \geq \epsilon) \\ &\leq \lim_{n\to\infty} P(\max\{|X_n - X|, |Y_n - Y|\} \geq \epsilon/2) \\ &\leq \lim_{n\to\infty} P(|X_n - X| \geq \epsilon/2) + \lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon/2) = 0 \end{aligned}$$

이고 $a \neq 0$이면

$$0 \leq \lim_{n\to\infty} P(|aX_n - aX| \geq \epsilon) = \lim_{n\to\infty} P(|X_n - X| \geq \epsilon/|a|) = 0$$

, $a = 0$이면

$$\lim_{n\to\infty} P(|aX_n - aX| \geq \epsilon) = \lim_{n\to\infty} 0 = 0$$

이므로 확률수렴의 정의에 따라 증명된다.

Problem. 어떤 상수 a에 대해, $X_n \xrightarrow{p} a$이며, 실함수 g가 a에서 연속이다. $g(X_n) \xrightarrow{p} g(a)$임을 보이시오.

Solution. 상수 a에서 실함수 g가 연속이므로, 모든 $\epsilon > 0$에 대하여 상응하는 $\delta(\epsilon) > 0$이 있어 $|y - a| < \delta(\epsilon)$이면 $|g(y) - g(a)| < \epsilon$임을 보장할 수 있다. 그렇다면

$$\begin{aligned} 0 \leq \lim_{n\to\infty} P(|g(X_n) - g(a)| \geq \epsilon) &= \lim_{n\to\infty} [P(|g(X_n) - g(a)| \geq \epsilon, |X_n - a| \geq \delta(\epsilon)) \\ &\quad + P(|g(X_n) - g(a)| \geq \epsilon, |X_n - a| < \delta(\epsilon))] \\ &\leq \lim_{n\to\infty} [P(|X_n - a| \geq \delta(\epsilon)) + 0] = 0 \end{aligned}$$

이 $X_n \xrightarrow{p} a$임에 따라 성립한다. 따라서 $g(X_n) \xrightarrow{p} g(a)$.

Problem. X_n을 공통 평균 μ와 공통 분산 $\sigma^2 < \infty$을 갖는 i.i.d. 확률변수의 열이라고 하자. 표본평균 $\bar{X} = \sum_{i=1}^{n} X_i/n$은 공통 평균 μ로 확률수렴함을 보이시오. 또한 이로부터 $\bar{X}$가 μ의 어떤 추정량인지 논하시오.

Solution. 힌트에서 마코프 부등식을 제시하였으므로, 여기에서 파생된 체비셰프 부등식 역시 안다고 가정하자. 그렇다면 임의의 $\epsilon > 0$에 대하여,

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\epsilon^2}$$

임을 안다. 이때 $\mathbb{E}[\bar{X}]| = \mu, \mathrm{Var}(\bar{X}) = \sigma^2/n$이므로

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

이고, 모든 ϵ에 대하여,

$$0 \leq \lim_{n\to\infty} P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

이므로 $\bar{X}$는 μ로 확률수렴한다. 이에 따라, $\bar{X}$는 μ의 일치추정량이다.

Problem. 모집단 분포가 연속형이고 그 누적분포함수가 $F(x)$일 때, 랜덤표본 $X_1, \cdots, X_n$에 대한 순서통계량을 $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$이라고 하자. 함수 $h(y) = F^{-1}(1 - e^{-y})I_{(0,\infty)}$에 대하여 다음을 보이시오.

$$(X_{(r)})_{1\leq r\leq n} \overset{d}{\equiv} \left(h\left(\frac{1}{n}Z_1 + \cdots + \frac{1}{n-r+1}Z_r\right)\right)_{1\leq r\leq n}, \quad Z_r \sim_{i.i.d.} \mathrm{Exp}(1)$$

Solution. 랜덤표본 $X_1, \cdots, X_n$이 $F(x)$로부터 비롯된다면, $F(X_1), \cdots, F(X_n)$은 $U(0,1)$으로부터 비롯되는 랜덤표본이며, 이에 따라 각각을 $U_1, \cdots, U_n$, 그로부터의 순서통계량을 $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$이라 하면

$$(X_{(r)})_{1\leq r\leq n} \overset{d}{\equiv} (F^{-1}(U_{(r)}))$$

이 F가 단조증가함수임에 따라 성립한다. 한편 $U_1 = 1 - e^{-Y_1}, \cdots, U_n = 1 - e^{-Y_n}$으로 쓰면 $1 - e^{-y}$ 역시 단조증가함수이므로,

$$(X_{(r)})_{1\leq r\leq n} \overset{d}{\equiv} h(Y_{(r)})$$

임을 알 수 있다. 따라서 이제 $Y_{(r)}$에 대해 논의하면 된다. 이때 $1 - e^{-y}$는 Exp(1)의 누적분포함수이므로, $Y_1, \cdots, Y_n$은 Exp(1)로부터의 랜덤표본이다. 그렇다면

$$Z_1 = nY_{(1)},\ Z_2 = (n-1)(Y_{(2)} - Y_{(1)}),\ \cdots, Z_n = Y_{(n)} - Y_{(n-1)}$$

과 같이 정의하자. 지수분포에서 순서통계량의 확률밀도함수는

$$f_{Y_{(1)},\cdots,Y_{(n)}}(y_1, \cdots, y_n) = n!e^{-y_1-y_2-\cdots-y_n}I(y_1 < y_2 < \cdots < y_n)$$

이며

$$Y_{(1)} + \cdots + Y_{(n)} = Z_1 + \cdots + Z_n$$

이고 $Z \mapsto Y$ 변환행렬의 야코비안이 $1/n!$이므로, $Z_1, \cdots, Z_n$의 확률밀도함수는

$$f_{Z_1,\cdots,Z_n}(z_1, \cdots, z_n) = e^{-z_1-z_2-\cdots-z_n} = \prod_{i=1}^{n} e^{-z_i}$$

이다. 따라서 $Z_1, \cdots, Z_n$은 i.i.d. Exp(1)이며, 이에 따라 각 $Y_{(r)}$은

$$Y_{(r)} = \frac{1}{n}Z_1 + \cdots + \frac{1}{n-r+1}Z_r$$

이므로 이를 대입하여

$$(X_{(r)})_{1\leq r\leq n} \overset{d}{\equiv} \left(h\left(\frac{1}{n}Z_1 + \cdots + \frac{1}{n-r+1}Z_r\right)\right)_{1\leq r\leq n}, \quad Z_r \sim_{i.i.d.} \mathrm{Exp}(1)$$

를 얻는다.

Problem. $X_1, \cdots, X_n$을 크기가 n인 랜덤표본이라고 하자. 이때, 지니의 평균차는 그 식이 아래와 같다.

$$G_n = \frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}|X_i - X_j|}{\binom{n}{2}}$$

$n = 10$일 때, $G_{10} = \sum_{i=1}^{10} a_i X_{(i)}$가 성립하도록 $a_1, \cdots, a_{10}$을 구하시오.

Solution. G_n의 분자만 보면, 가장 큰 $X_{(10)}$은 총 9번 더해지고, $X_{(9)}$는 8번 더해지며 1번 빼지고, 이와 같은 방식으로 $X_{(r)}$은 $r-1$번 더해지고 $10-r$번 빼진다. 따라서 총합으로는 $2r-11$번 더해진다고 볼 수 있다. 따라서

$$a_r = \frac{2r-11}{\binom{n}{2}}$$

Problem. $X \sim N(\mu, \sigma^2)$일 때, $\mathbb{E}[|X-\mu|] = \sigma\sqrt{2/\pi}$임을 보이시오.

Solution.

$$\begin{aligned}
\mathbb{E}[|X-\mu|] &= \int_{-\infty}^{\infty} |x-\mu| \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \\
&= \sqrt{2/\pi}\int_{0}^{\infty} (x-\mu) e^{-(x-\mu)^2/2\sigma^2} dx \\
&= \sigma\sqrt{2/\pi}\left[e^{-(x-\mu)^2/2\sigma^2}\right]_0^{\infty} = \sigma\sqrt{2/\pi}
\end{aligned}$$

Problem. 위의 결과를 이용하여 표본을 $N(\mu, \sigma^2)$에서 추출했을 때, 지니의 평균차 G_n에 대해 $\mathbb{E}[G_n] = 2\sigma/\sqrt{\pi}$임을 보이시오.

Solution.

$$\begin{aligned}
\mathbb{E}[G_n] &= \mathbb{E}\left[\frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}|X_i - X_j|}{\binom{n}{2}}\right] \\
&= \mathbb{E}[|X_1 - X_2|] \\
&= \sqrt{2/\pi} \times \sqrt{2\sigma^2} = 2\sigma/\sqrt{\pi}
\end{aligned}$$

이때 $X_1 - X_2 \sim N(0, 2\sigma^2)$임을 이용한다.

Problem. 어떤 공장에서 네 종류의 절삭공구를 생산하고 있다. 제품 간에 절삭력의 차이가 있는지 알아보기 위해 각 종류별 절삭공구의 절삭력을 3회 반복하여 측정하였다. 그 결과는 아래 표와 같다. (매 회 새로운 절삭공구 사용)

	절삭공구 종류			
	㉠	㉡	㉢	㉣
1회	66	55	45	60
2회	56	57	40	56
3회	70	60	51	53

위 실험에 적절한 구조식과 그 제약조건을 기술하시오.

Solution. 편의상 ㄱ, ㄴ, ㄷ, ㄹ을 A1, A2, A3, A4라고 쓰자. 이는 반복이 있는 일원배치법과 동일하다. 따라서 적절한 구조식은

$$x_{ij} = \mu + a_i + \epsilon_{ij}$$

이때 μ는 전체 평균, a_i는 i번째 인자에 해당하는 효과 평균, ϵ_{ij}는 i번째 절삭도구의 j회 시행에서의 오차이며,

$$\sum_{i=1}^{4} a_i = 0$$

을 제약조건으로 한다. ϵ_{ij}는 $N(0, \sigma_\epsilon^2)$을 따르는 i.i.d. 오차이다.

Problem. 위 문제에서, 총제곱합을 요인제곱합과 오차제곱합으로 분해하시오.

Solution.

$$\begin{aligned} SS_T &= \sum_{i=1}^{4}\sum_{j=1}^{3}(x_{ij} - \bar{\bar{x}})^2 \\ &= 740.25 \\ SS_A &= \sum_{i=1}^{4} 3(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \\ &= 538.25 \\ SS_E &= \sum_{i=1}^{4}\sum_{j=1}^{4}(x_{ij} - \bar{x}_{i\cdot})^2 \\ &= 202 \end{aligned}$$

Problem. 아래의 분산분석표를 완성한 후 유의수준 $\alpha = 0.05$에서 요인(A)에 대한 검정을 하시오.

	제곱합(SS)	자유도($d.f.$)	평균제곱합(MS)	F값
요인(A)				
오차(E)				
총(T)				

Solution. 첫째 열은 앞서 구하였듯이 각각 538.25, 202, 740.25이다. 둘째 열은 각각 $3(=4-1)$, $8(=4(3-1))$, $11(=4\times 3-1)$이다. 셋째 열은 각각을 자유도로 나누어 위의 두 셀을 179.42, 25.25로 채운다. 마지막으로 F값은 요인 A에 대하여 MS_A/MS_E인 7.11을 얻는다. 한편 $F(3,8)$의 분포 하에서 95% 퍼센타일은 4.066으로, 우리의 검정통계량이 더 크기에 유의수준 $\alpha = 0.05$에서 $H_0 : a_1 = a_2 = a_3 = a_4 = 0$을 기각할 수 있다.

Problem. 절삭공구 A1, A3의 절삭력 모평균에 대한 95퍼센트 신뢰구간을 각각 구하시오. 또한 절삭공구 A1, A3의 절삭력 모평균차에 대한 95퍼센트 신뢰구간을 구하시오.

Solution. A1의 절삭력 모평균은 $\mu + a_1$이다. 그 95퍼센트 신뢰구간은

$$\bar{x}_{1\cdot} \pm t_{0.025}(8)\sqrt{MS_E/3} = 64 \pm 6.69 = (57.31, 70.69)$$

이다. A3의 절삭력 모평균은 $\mu + a_4$이다. 그 95퍼센트 신뢰구간은

$$\bar{x}_{3\cdot} \pm t_{0.025}(8)\sqrt{MS_E/3} = 45.33 \pm 6.69 = (48.74, 52.03)$$

이다. A1, A3의 절삭력 모평균차는 $a_1 - a_3$이다. 그 95퍼센트 신뢰구간은

$$(\bar{x}_{1\cdot} - \bar{x}_{3\cdot}) \pm t_{0.025}(8)\sqrt{2 \times MS_E/3} = 18.67 \pm 9.46 = (9.23, 28.13)$$

이다.

Problem. MA(2) 시계열의 가역성 문제

Solution. 설명편 참고

Problem. 다중선형회귀모형 $Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I_n)$에 대하여 일반적인 선형가설

$$H_0 : C\beta = m, \quad H_1 : C\beta \neq m$$

을 검정하기 위한 검정통계량을 구하시오. 이때 C는 rank가 q인 full column rank matrix이다.

Solution.

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$$

임이 잘 알려져 있다. 따라서 귀무가설 하에서, $C\beta - m$이므로

$$C\hat{\beta} - m \sim N(0, \sigma^2 C(X^TX)^{-1}C^T)$$

이다. 따라서 이를 이용하면 가설검정을 할 수 있다. 특히 카이제곱분포의 정의에 의하여

$$(C\hat{\beta} - m)^T(\sigma^2 C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta} - m) \sim \chi^2(\text{rank}(C(X^TX)^{-1}C^T)) = \chi^2(q)$$

이다. 이때 σ^2은 모르므로,

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$$

이며 이것이 $\hat{\beta}$와 독립적임을 감안하면

$$F = (C\hat{\beta} - m)^T(\hat{\sigma}^2 C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta} - m)/q \sim F(q, n-p)$$

을 검정통계량으로 사용할 수 있다.

Problem. 다중선형회귀분석에서 하나의 회귀계수 β_j에 대한 유의성검정을 할 때, t검정과 F검정의 결과가 일치함을 보이시오.

Solution. 이 경우 $C = e_j^T$이다. e_j는 j번째 원소만 1이고 나머지는 0인 벡터이다. 그렇다면 $q = 1$임을 감안할 때 앞에서 구한 F통계량은

$$F = \hat{\beta}_j^T\hat{\beta}_j^T/\hat{\sigma}^2(X^TX)^{-1}_{jj} = \left(\frac{\hat{\beta}_j}{\hat{\sigma}_{jj}}\right)^2 = t^2$$

으로 나타내어진다. 또한 F분포와 t분포의 성질에서 $F(1, n-p)$분포를 따르는 V는 $t(n-p)$분포를 따르는

U의 제곱으로 표현될 수 있다. 따라서 $F_\alpha(1,n-p)=t_\alpha^2(1,n-p)$이다. 따라서 t검정에서

$$|t|\geq t_\alpha(1,n-p)$$

이라면, 항상

$$F=|t|^2\geq t_\alpha^2(1,n-p)^2=F_\alpha(1,n-p)$$

이므로 F검정도 기각한다. 반대로 F검정에서

$$F\geq F_\alpha(1,n-p)$$

으로 기각한다면, t검정에서

$$|t|=\sqrt{F}\geq\sqrt{F_\alpha(1,n-p)}=t_\alpha^2(1,n-p)$$

으로 t검정도 기각한다. 따라서 두 검정은 완벽히 동일하다. 주의할 것은 단순히 기각역이 같을 뿐만 아니라, 동일한 표본에서 어떠한 검정 방법을 사용해도 결과가 같다는 것이다.

Problem. p차원 정규분포 $N_p(\mu_x,\Sigma)$에서 추출한 크기가 n인 랜덤표본 $X_1,\cdots,X_n$와 $N_p(\mu_y,\Sigma)$에서 추출한 크기가 m인 랜덤표본 $Y_1,\cdots,Y_m$이 있다. 합동공분산행렬 S는 아래와 같이 정의된다.

$$S=\frac{(n-1)S_x+(m-1)S_y}{n+m-2}$$

우도함수를 구하고, μ_x,μ_y,Σ의 최대우도추정량을 구한 뒤, $\mu_0=\mu_x=\mu_y$라는 제약을 가했을 때의 최대우도추정량과 비교하여라.

Solution. 우도함수는

$$\begin{aligned}L(\mu_x,\mu_y,\Sigma;x,y)&=\prod_{i=1}^n\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp(-(x_i-\mu_x)^T\Sigma^{-1}(x_i-\mu_x)/2)\\&\times\prod_{j=1}^m\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp(-(y_j-\mu_y)^T\Sigma^{-1}(y_j-\mu_y)/2)\\&=(2\pi)^{-(m+n)p/2}|\Sigma|^{-(m+n)/2}\\&\times\exp\left(-\frac{1}{2}\left(\sum_{i=1}^n(x_i-\mu_x)^T\Sigma^{-1}(x_i-\mu_x)+\sum_{j=1}^n(y_j-\mu_y)^T\Sigma^{-1}(y_j-\mu_y)\right)\right)\end{aligned}$$

으로 주어진다. 로그가능도함수는

$$\begin{aligned}l(\mu_x,\mu_y,\Sigma;x,y)&=-\frac{(m+n)p}{2}\log(2\pi)-\frac{m+n}{2}\log|\Sigma|\\&-\frac{1}{2}\left(\sum_{i=1}^n(x_i-\mu_x)^T\Sigma^{-1}(x_i-\mu_x)+\sum_{j=1}^n(y_j-\mu_y)^T\Sigma^{-1}(y_j-\mu_y)\right)\end{aligned}$$

으로 주어진다. 양변을 μ_x로 미분하면 Σ가 symmetric matrix이고 Σ^{-1}도임에 따라

$$\nabla_{\mu_x}l=-\sum_{i=1}^n\Sigma^{-1}(x_i-\mu_x)$$

이므로 첫째 일계조건은 $\mu_x = n^{-1}\sum_{i=1}^{n} x_i$로 얻는다. 동일한 이유로, μ_y로 미분하면

$$\nabla_{\mu_y} l = -\sum_{j=1}^{m} \Sigma^{-1}(y_j - \mu_y)$$

을 얻기에 둘째 일계조건은 $\mu_y = m^{-1}\sum_{j=1}^{m} y_j$로 얻는다. 한편 $\mu_x = \bar{x}, \mu_y = \bar{y}$로 가정하고 로그가능도함수를 Σ^{-1}에 대해 미분하면,

$$\begin{aligned}
\frac{\partial l}{\partial \Sigma^{-1}} &= \frac{m+n}{2} \times \frac{1}{|\Sigma^{-1}|} \times |\Sigma^{-1}|\Sigma \\
&\quad - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \mu_x)^T(x_i - \mu_x) + \sum_{j=1}^{n}(y_j - \mu_y)^T(y_j - \mu_y)\right) \\
&= \frac{m+n}{2}\Sigma - \frac{1}{2}((n-1)S_x + (m-1)S_y)
\end{aligned}$$

이고 일계조건에서

$$\Sigma = \frac{(n-1)S_x + (m-1)S_y}{m+n} = \frac{m+n-2}{m+n}S$$

를 얻는다. 따라서 제약이 없으면

$$\begin{aligned}
\hat{\mu}_x &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\hat{\mu}_y &= \frac{1}{m}\sum_{j=1}^{m} Y_j \\
\hat{\Sigma} &= \frac{m+n-2}{m+n}S
\end{aligned}$$

이다. 한편 제약이 있는 경우, 로그가능도함수는

$$\begin{aligned}
l(\mu_0, \Sigma; x, y) &= -\frac{(m+n)p}{2}\log(2\pi) - \frac{m+n}{2}\log|\Sigma| \\
&\quad - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^T\Sigma^{-1}(x_i - \mu_0) + \sum_{j=1}^{n}(y_j - \mu_0)^T\Sigma^{-1}(y_j - \mu_0)\right)
\end{aligned}$$

이므로

$$\nabla_{\mu_0} l = -\sum_{i=1}^{n}\Sigma^{-1}(x_i - \mu_0) - -\sum_{j=1}^{m}\Sigma^{-1}(y_j - \mu_0)$$

이다. 따라서 일계조건으로부터 $\mu_0 = \frac{1}{m+n}(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j)$을 얻는다. 동시에, Σ^{-1}으로 미분하면

$$\begin{aligned}
\frac{\partial l}{\partial \Sigma^{-1}} &= \frac{m+n}{2} \times \frac{1}{|\Sigma^{-1}|} \times |\Sigma^{-1}|\Sigma \\
&\quad - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^T(x_i - \mu_0) + \sum_{j=1}^{n}(y_j - \mu_0)^T(y_j - \mu_0)\right) \\
&= \frac{m+n}{2}\Sigma - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^T(x_i - \mu_0) + \sum_{j=1}^{n}(y_j - \mu_0)^T(y_j - \mu_0)\right)
\end{aligned}$$

이므로,

$$\hat{\mu}_0 = \frac{n\bar{X} + m\bar{Y}}{m+n}$$
$$\hat{\Sigma} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_0)^T(x_i - \hat{\mu}_0) + \sum_{j=1}^{n}(y_j - \hat{\mu}_0)^T(y_j - \hat{\mu}_0)}{m+n}$$

으로 주어진다. 즉 제약이 존재하는 경우 $m+n$개의 IID 표본이 있다고 가정하고 구한 최대우도추정량과 동일하다.

2019년

Problem. 변동계수(coefficient of variation)에 대해 간략히 설명하시오.

Solution. 변동계수는 표준편차를 평균으로 나눈 값

$$cv = \frac{\sigma}{\mu}$$

이다. 이는 데이터의 전체적인 스케일에 무관하게 그 변동이 평균의 어느 정도인지를 묘사한다.

Problem. 젠센(Jensen)의 부등식에 대해 간략히 설명하시오.

Solution. 통계학의 관점에서, Jensen의 부등식은 아래와 같다.
$X \in \mathbb{R}^d$가 $\mathbb{E}[X]$가 잘 정의되는 확률변수이고 $\phi : \mathbb{R}^d \to \mathbb{R}$이 convex이라면,

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$

이다.

Problem. 로지스틱(logistic) 분포에 대해 간략히 설명하시오.

Solution. 로지스틱 분포는 아래의 확률밀도함수를 가지는 연속확률분포이다.

$$f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$$

그 평균은 μ이며 분산은 s^2에 비례한다. 그 누적분포함수는 $x \to -\infty$이면 0, $x \to \infty$이면 1인 sigmoidal 형태이다.

Problem. 지수분포의 무기억 성질(memoryless property)에 대해 간략히 설명하시오.

Solution. 지수분포를 따르는 확률변수 X과 양의 실수 a, b에 대하여,

$$P(X \geq a) = P(X \geq a + b | X \geq b)$$

이 성립한다. 이를 무기억 성질이라고 부른다. 이는 X의 $X \geq b$라는 기억은 X가 그로부터 추가적으로 a만큼 더 커질 확률에 영향을 주지 못함을 의미한다.

Problem. Fisher의 정보에 대해 간략히 설명하시오.

Solution. Fisher의 정보는 score function의 variance

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X;\theta) \right)^2 \right]$$

로 정의된다. 이때 $f(X;\theta)$는 θ가 true parameter일 때의 X에서의 확률밀도함수이다.

Problem. X_1와 X_2가 연속형 확률변수이며 X_2의 분산이 유한하다고 하면,

$$\mathbb{E}[\mathbb{E}[X_2|X_1]] = \mathbb{E}[X_2]$$

임을 보이시오.

Solution.
모든 $A \in \sigma(X_1)$에 대하여, 조건부 기대값의 정의에 의하여

$$\mathbb{E}[\mathbb{E}[X_2|X_1]\mathbf{1}_A] = \mathbb{E}[X_2\mathbf{1}_A]$$

이다. $A = \mathbb{R}^2$으로 두면,

$$\mathbb{E}[\mathbb{E}[X_2|X_1]] = \mathbb{E}[X_2]$$

을 얻는다.

Problem. X_1와 X_2가 연속형 확률변수이며 X_2의 분산이 유한하다고 하면,

$$\mathrm{Var}(\mathbb{E}[X_2|X_1]) \leq \mathrm{Var}(X_2)$$

임을 보이시오.

Solution. 앞선 문제를 응용하면

$$\begin{aligned}
\mathrm{Var}(X_2) &= \mathbb{E}[(X_2 - \mathbb{E}[X_2])^2] \\
&= \mathbb{E}[(X_2 - \mathbb{E}[\mathbb{E}[X_2|X_1]])^2] \\
&= \mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1] + \mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])^2] \\
&= \mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])^2] \\
&+ 2\,\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])] \\
&+ \mathbb{E}[(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])^2] \\
&\geq \mathrm{Var}(\mathbb{E}[X_2|X_1])
\end{aligned}$$

이 성립한다. 이때 마지막 부등호는

$$\begin{aligned}
\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])^2] &\geq \mathbb{E}[0] = 0 \\
\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])] &= \mathbb{E}[\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])|X_1]] \\
&= \mathbb{E}[(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])\,\mathbb{E}[X_2 - \mathbb{E}[X_2|X_1]|X_1]] \\
&= \mathbb{E}[(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]])(\mathbb{E}[X_2|X_1] - \mathbb{E}[X_2|X_1])] = 0
\end{aligned}$$

임에 따라 성립한다.

Problem. 모집단을 특정 그룹 단위로 나누어 표본을 추출할 때, 그룹의 특성을 고려하여 집락표집(cluster sampling)과 층화표집(stratified sampling)을 비교하시오.

Solution. 먼저 그룹의 특성을 보면, 집락표집에서는 각 집락 내부에서는 이질적이지만 집락들끼리는 동질적이어 집락들이 하나의 조사 단위로서 작동한다. 따라서 모집단을 집락 여러 개로 나눈 후, 그 중 몇 개의 집락만 얻는 방식으로 표본을 추출한다. 반면 층화표집에서는 각 층 내에서는 동질적이지만 층들끼리는 이질적이어, 모집단을 잘 설명하는 표본을 얻기 위해서는 각 층마다 적절한 수의 표본을 얻어야 한다. 집락표집은 집락 안의 원소들이 이질적일수록, 층화표집은 층 안의 원소들이 동질적일수록 그 효율이 증가한다.

Problem. 다음은 10개의 표본을 추출하여 단순선형회귀분석을 실시한 결과이다. 결과에 근거하여 다음 질문들에 답하시오.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} N(0, \sigma^2),\ i = 1, 2, \cdots, 10$$

<모수추정 결과>

모수	추정값	표준오차(s.e.)
β_0	3.5	1.12
β_1	8.4	2.34

<분산분석표>

요인	제곱합(SS)	자유도(df)	평균제곱합(MS)	F-값
회귀(Regression)				㉠
오차(Error)	44.8	㉡		-
합계(Total)	㉢		-	-

모수추정 결과를 토대로 β_1의 95퍼센트 신뢰구간을 구하여라.

Solution. β_1은

$$\frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \sim t(10-2)$$

을 따르므로 95퍼센트 신뢰구간은

$$(\hat{\beta}_1 - \text{s.e.}(\hat{\beta}_1) \times t_{0.025}(8), \hat{\beta}_1 + \text{s.e.}(\hat{\beta}_1) \times t_{0.025}(8)) = (3.19, 13.61)$$

으로 주어진다.

Problem. 위 문제에서, 분산분석표에 있는 빈칸들을 채우고 결정계수를 구하시오.

Solution. $8.4 = \hat{\beta}_1 = S_{xy}/S_{xx}$이며, 오차의 자유도가 8이므로 평균제곱합은 5.6이며,

$$(2.34)^2 = \text{s.e.}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{S_{xx}} = \frac{5.6}{S_{xx}}$$

이므로

$$SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx} = 8.4^2 \times \frac{5.6}{(2.34)^2} = 72.2$$

이며, $SST = SSR + SSE = 117$을 얻는다. 한편 F값은

$$F = \frac{SSR/1}{SSE/8} = \frac{72.2}{5.6} = 12.9$$

로 나타난다. 한편 결정계수는

$$R^2 = \frac{SSR}{SST} = \frac{72.2}{117} = 0.62$$

로 나타난다.

Problem. 위 모형에서 $\beta_0 = 3.5$라는 정보가 추가적으로 주어졌다 하자. 이를 참고하여 새로운 모형을 제시하고 회귀계수의 최소제곱 추정량을 유도하시오. 또한, 기존 모형과 비교하여 새로운 모형이 가질 수 있는 특징을 간략히 서술하시오.

Solution.

새로운 모형에서는

$$y_i = 3.5 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} N(0, \sigma^2),\ i = 1, 2, \cdots, 10$$

로 모형이 나타나며, 여기에서는 사실상 $y_i - 3.5$를 새로운 데이터 z_i로 하고 상수항이 없는 회귀모형 $z_i = \beta_1 x_i + \epsilon_i$를 적합하는 것이나 마찬가지이다. 이 경우

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} z_i x_i}{\sum_{i=1}^{10} x_i^2} = \frac{\sum_{i=1}^{10} (y_i - 3.5) x_i}{\sum_{i=1}^{10} x_i^2}$$

을 얻는다.

$$\begin{aligned}\sum_{i=1}^{10} (y_i - 3.5) x_i &= \sum_{i=1}^{10} y_i x_i - 3.5 \sum_{i=1}^{10} x_i \\ &= S_{xy} + 10\bar{x}\bar{y} - 35\bar{x}\end{aligned}$$

이며 $S_{xy} = \hat{\beta}_1 S_{xx} = 8.4 \times \dfrac{5.6}{(2.34)^2} = 8.59$이다. 앞선 정보에서 $\bar{y} = 3.5 + 8.4\bar{x}$이므로 이를 대입하면 $10\bar{x}\bar{y} - 35\bar{x} = 84\bar{x}^2$이다. 그리고

$$84\bar{x}^2 = 8.4 \times 10\bar{x}^2 = 8.4 \times (\sum_{i=1}^{10} x_i^2 - S_{xx})$$

인데 분모에서

$$\sum_{i=1}^{10} x_i^2 = 10 \times \frac{S_{xx}}{\hat{\sigma}^2} \times \text{s.e.}(\hat{\beta}_0)^2 = \frac{10 \times (1.12)^2}{(2.34)^2} = 2.3$$

이므로, 분자는

$$S_{xy} + 84\bar{x}^2 = 8.59 + 8.4(2.3 - 5.6/(2.34)^2) = 19.32$$

이며 이를 분모로 나누면 새 추정량으로 여전히 8.4를 얻는다. 즉 추정량의 값 자체는 같다. 글그러나 새로운 추정량은 상수항 부분을 고정하였기에, 상수항을 추정해야 함에 따라 손실되는 효율성을 β_1을 추정하는 데 사용할 수 있다. 따라서 분산이 작아지고, 신뢰구간의 길이가 감소하는 장점이 있다. 기존에 $\hat{\beta}_1$의 분산이

$$\frac{\hat{\sigma}^2}{S_{xx}} = (2.34)^2 = 5.48$$

이었던 반면, 새로운 추정량의 분산은

$$\frac{\hat{\sigma}^2}{\sum_{i=1}^{10} x_i^2} = \frac{44.8/(10-1)}{2.3} = 2.16$$

이다. 따라서 우리는 더 효율적인 추정량을 얻는다.

Problem. 시계열 추정 문제

Solution. 설명편 참고

Problem. 다음 확률밀도함수

$$f_X(x) = (2 - 4|x - 0.5|)I_{(0,1)}(x)$$

에 대하여 $\mathbb{E}[e^{tX}]$를 구하고, $\mathbb{E}[X]$가 $\lim_{t\to 0} \dfrac{d\mathbb{E}[e^{tX}]}{dt}$가 같은지 확인한 뒤 적률생성함수가 존재하는지 설명하시오.

Solution.

$$\begin{aligned}
\mathbb{E}[e^{tX}] &= \int_0^1 e^{tx}(2 - 4|x - 0.5|)dx \\
&= \int_0^{0.5} e^{tx}(2 - 4(0.5 - x))dx + \int_{0.5}^1 e^{tx}(2 - 4(x - 0.5))dx \\
&= \int_0^{0.5} 4xe^{tx}dx + \int_{0.5}^1 4(1 - x)e^{tx}dx \\
&= 4\left[\left(\frac{x}{t} - \frac{1}{t^2}\right)e^{tx}\right]_0^{0.5} + 4\left[\left(-\frac{x}{t} + \frac{1}{t} + \frac{1}{t^2}\right)e^{tx}\right]_{0.5}^1 \\
&= \frac{2}{t}e^{0.5t} - \frac{4}{t^2}e^{0.5t} + \frac{4}{t^2} + \frac{4}{t^2}e^t - \frac{2}{t}e^{0.5t} - \frac{4}{t^2}e^{0.5t} \\
&= \frac{4}{t^2}(e^{0.5t} - 1)^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 x(2 - 4|x - 0.5|)dx \\
&= \int_0^{0.5} 4x^2 dx + \int_{0.5}^1 4x(1 - x)dx \\
&= \frac{1}{6} + \frac{3}{2} - \frac{7}{6} = \frac{1}{2}
\end{aligned}$$

$$\begin{aligned}
\lim_{t\to 0} \frac{d\mathbb{E}[e^{tX}]}{dt} &= \lim_{t\to 0} -\frac{8}{t^3}(e^{0.5t} - 1)^2 + \frac{4}{t^2}e^{0.5t}(e^{0.5t} - 1) \\
&= \lim_{v\to 0} \frac{e^v(e^v - 1)}{v^2} - \frac{(e^v - 1)^2}{v^3} \quad (v = 0.5t) \\
&= \lim_{v\to 0} \frac{e^v - 1}{v} \times \frac{e^v v - e^v + 1}{v^2} \\
&= \lim_{v\to 0} \frac{ve^v}{2v} = \frac{1}{2}
\end{aligned}$$

따라서 둘은 같다. 따라서 이를 이용하면 mgf를

$$M_X(t) = \begin{cases} \frac{4}{t^2}(e^{0.5t}-1)^2 & t \neq 0 \\ 1 & t = 0 \end{cases}$$

이게 정의하면 이는 $\mathbb{E}[e^{tX}]$와 almost everywhere에서 같으면서 moment를 generating하는 적률생성함수이다.

Problem. 확률표본 $X_1, \cdots, X_n$은 독립이고 다음의 확률밀도함수를 갖는 지수분포를 따른다고 한다.

$$f_X(x) = \alpha e^{-\alpha x},\ x > 0,\ \alpha > 0$$

$\beta > 0$에 대하여 $Y = \beta e^X$의 확률밀도함수와 n차 적률을 구하시오. 또한 이로부터 그 기대값과 분산을 구하시오.

Solution.

누적분포함수는 $y \geq \beta$에 대하여

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\beta e^X \leq y) \\ &= P(e^X \leq y/\beta) \\ &= P(X \leq \ln(y/\beta)) \\ &= 1 - e^{-\alpha \ln(y/\beta)} \\ &= 1 - (y/\beta)^{-\alpha} \end{aligned}$$

이며 $y < \beta$에 대해서는 $F_Y(y) = 0$이므로,

$$f_Y(y) = \alpha\beta^\alpha y^{-\alpha-1} I(y \geq \beta)$$

이 확률밀도함수이다.

n차 적률을 구하기 위해서는 아래처럼 할 수 있다. $n \geq 1$인 정수일 때, $n \neq \alpha$라면

$$\begin{aligned} \mathbb{E}[Y^n] &= \int_\beta^\infty y^n \alpha\beta^\alpha y^{-\alpha-1} dy \\ &= \alpha\beta^\alpha \int_\beta^\infty y^{n-\alpha-1} dy \\ &= \frac{\alpha\beta^\alpha}{n-\alpha}[y^{n-\alpha}]_\beta^\infty \\ &= \begin{cases} \infty & (\text{if } n > \alpha) \\ \dfrac{\alpha\beta^n}{\alpha - n} & (\text{if } n < \alpha) \end{cases} \end{aligned}$$

이며 $n = \alpha$인 경우에는 ∞이다. 따라서 Y의 n차 적률은

$$\mathbb{E}[Y^n] = \begin{cases} \infty & (\text{if } n \geq \alpha) \\ \dfrac{\alpha\beta^n}{\alpha - n} & (\text{if } n < \alpha) \end{cases}$$

이며, 잘 정의되려면 $n < \alpha$여야 한다. 이때 기대값과 분산은 $\alpha > 2$일 때만 잘 정의될 수 있고,

$$\mathbb{E}[Y] = \frac{\alpha\beta}{\alpha - 1}$$
$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \frac{\alpha\beta^2}{\alpha - 2} - \frac{\alpha^2\beta^2}{(\alpha-1)^2} = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$$

이다.

Problem. Y의 분포에서 모수 α, β의 최대가능도 추정량을 구하고, 그 분산을 대표본 근사로써 구하시오.

Solution. $y_1, \cdots, y_n$을 표본으로써 얻었다고 하자. 그렇다면

$$\begin{aligned} L(\alpha, \beta; y_i) &= \prod_{i=1}^{n} \alpha\beta^\alpha y_i^{-\alpha-1} I(y_i \geq \beta) \\ &= \alpha^n \beta^{n\alpha} \left(\prod_{i=1}^{n} y_i \right)^{-\alpha-1} I(\min\{y_i\} \geq \beta) \end{aligned}$$

이며 로그가능도함수는

$$l(\alpha, \beta; y_i) = n\log(\alpha) + n\alpha\log(\beta) - (\alpha+1)\sum_{i=1}^{n}\log(y_i) + \delta(\min\{y_i\} \geq \beta)$$

으로 주어진다. 로그가능도함수가 최대화되려면, 적어도 $\beta \leq \min\{y_i\}$여야 한다. 이 조건 하에서 일계조건은

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{n}{\alpha} + n\log(\beta) - \sum_{i=1}^{n}\log(y_i) \\ \frac{\partial l}{\partial \beta} &= \frac{n\alpha}{\beta} \end{aligned}$$

이므로, α가 고정일 때 β는 최대한으로 증가해야 하므로, $\hat{\beta} = \min\{y_i\}$이다. 해당 조건 하에서,

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n}\log(y_i) - n\log(\min\{y_i\})}$$

을 얻는다. 한편 (α, β)의 최대가능도추정량의 분산은 대표본 하에서 피셔의 정보량의 역행렬로 수렴한다. 따라서 피셔의 정보량을 구하자.

$$\begin{aligned} I(\alpha, \beta) &= \mathrm{Var}\left(\frac{\partial}{\partial\theta} \log f(Y; \theta) \right) \\ &= \mathrm{Var}\left(\begin{pmatrix} \frac{1}{\alpha} + \log(\beta) - \log(Y) \\ \frac{\alpha}{\beta} \end{pmatrix} \right) \\ &= \mathrm{Var}\left(\begin{pmatrix} -\log(Y) \\ 0 \end{pmatrix} \right) \\ &= \begin{pmatrix} \mathrm{Var}(\log(Y)) & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Var}(X) & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} \frac{1}{\alpha^2} & 0 \\ 0 & 0 \end{pmatrix}$$

따라서, 근사적으로 $(\hat{\alpha}, \hat{\beta})$의 분산은

$$\begin{pmatrix} \frac{\alpha^2}{n} & 0 \\ 0 & 0 \end{pmatrix}$$

이다. (추가 필요: 아마 $\sqrt{n}$ 대신 n을 곱하면 β 분산도 0이 아니게 될텐데, 이것까지 과연 해야 할지)

Problem. Neymann-Pearson의 정리를 기술하고, 이를 증명하시오.

Solution. 네이만 피어슨 보조정리는 가설

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1$$

의 검정함수 ϕ^*가 어떤 $\gamma \in [0,1]$과 $k \geq 0$에 대하여

$$\phi^*(x) = \begin{cases} 1 & \frac{f(x;\theta_1)}{f(x;\theta_0)} > k \\ \gamma & \frac{f(x;\theta_1)}{f(x;\theta_0)} = k \\ 0 & \frac{f(x;\theta_1)}{f(x;\theta_0)} < k \end{cases}$$

꼴이면서 그 크기가 $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$이면, 유의수준 α에서 최강력검정임을 의미한다.
일반적인 유의수준 α인 검정 ϕ를 생각하면, $\mathbb{E}_{\theta_0}[\phi(X)] \leq \alpha$여야 한다. 그렇다면 검정력 측면에서는

$$\begin{aligned} \mathbb{E}_{\theta_1}[\phi^*(X)] - \mathbb{E}_{\theta_1}[\phi(X)] &= \mathbb{E}_{\theta_1}[\phi^*(X) - \phi(X)] \\ &\geq \mathbb{E}_{\theta_1}[\phi^*(X) - \phi(X)] - k(\mathbb{E}_{\theta_0}[\phi^*(X)] - \mathbb{E}_{\theta_0}[\phi(X)]) \\ &= \int (\phi^*(x) - \phi(x))(f(x;\theta_1) - kf(x;\theta_0))dx \\ &\geq 0 \end{aligned}$$

이 성립한다. 이때 마지막 부등식은 $f(x;\theta_1) > kf(x;\theta_0)$인 경우 $\phi^*(x) - \phi(x) \geq 0$으로 integrand가 양수이고, $f(x;\theta_1) < kf(x;\theta_0)$인 경우 $\phi^*(x) - \phi(x) \leq 0$으로 여전히 integrand가 양수이기 때문이다. 따라서 MP test가 된다.

2013년

Problem. 판별함수, 비모수 검정 문제

Solution. 설명편 참고

Problem. 층화확률추출법의 개념을 설명하고 동 방법을 적용하여 전체 표본을 각 층에 배분하려고 할 때 고려해야 할 요인은 어떠한 것이 있는지 약술하시오.

Solution. 단순확률추출을 하는 경우 표본이 모집단을 잘 설명하지 못하는 형태로 추출될 수 있다. 이러한 문제는 추정량의 분산 증가로 이어진다. 모집단에 포함된 각 개체들을 설명할 수 있는 표본을 뽑기 위해서는, 특정 개체들이 뽑히지 않는 극단적인 경우를 줄이기 위하여 모집단을 층으로 나누고 각 층에서 표본 추출을 함으로써 최종적으로 얻은 표본이 모든 층의 원소를 포함하도록 할 수 있다. 이는 전체 분산 중 층 내에서의 분산의 비중이 작을 때 더욱 극대화된다. 따라서 전체 모집단을 각 층에 배분하려고 할 때, 층 내부의 개체들은 동질적이면서 층끼리는 이질적이도록 설계하여 분산을 최소화하고 효율적인 추정량을 얻을 수 있도록 해야 한다.

Problem. 다중회귀모형 $y = X\beta + \epsilon,\ \epsilon \sim N(0, I\sigma^2)$에 관한 다음 물음에 답하시오, β는 p차원 벡터이며, 관측값은 n개이다. 최소제곱법에 따른 최소제곱추정량 $\hat{\beta}$를 구하고, 그 분산-공분산행렬을 구하시오.

Solution.

$$\begin{cases}\text{minimize} \quad (y - X\beta)^T(y - X\beta)\end{cases}$$

β로 미분하면, normal equation은

$$2X^TX\beta - 2X^Ty = 0$$

이며, 따라서

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

분산-공분산행렬은

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((X^TX)^{-1}X^Ty) \\ &= (X^TX)^{-1}X^T\,\mathrm{Var}(y)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

Problem. 표준화된 두 개의 독립변수 (x_1, x_2)와 하나의 종속변수 (y)의 선형모형

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

에서 β의 최소제곱추정치 $\hat{\beta}$의 분산-공분산행렬을 (x_1, x_2)의 표본상관계수 r_{12}를 이용하여 나타내시오. 또한

이를 이용하여 두 독립변수가 높은 상관관계를 가질 때 발생할 수 있는 문제를 기술하시오.

Solution.

$$\begin{aligned}\sigma^2(X^TX)^{-1} &= \sigma^2\begin{pmatrix}\sum_{i=1}^n x_{1,i}^2 & \sum_{i=1}^n x_{1,i}x_{2,i}\\ \sum_{i=1}^n x_{1,i}x_{2,i} & \sum_{i=1}^n x_{2,i}^2\end{pmatrix}^{-1}\\ &= \sigma^2\begin{pmatrix}1 & r_{12}\\ r_{12} & 1\end{pmatrix}^{-1}\\ &= \frac{\sigma^2}{1-r_{12}^2}\begin{pmatrix}1 & -r_{12}\\ -r_{12} & 1\end{pmatrix}\end{aligned}$$

따라서 각 추정량 $\hat{\beta}_1, \hat{\beta}_2$의 분산은 $1-r_{12}^2$에 반비례한다. 따라서 r_{12}가 크다면, 그 분산이 커지며, 다중공선성의 문제로 인하여 회귀계수 추정량들이 비효율적이게 된다.

Problem. 서로 독립이고 크기가 n인 확률표본 $X_1, \cdots, X_n$을 아래와 같은 확률밀도함수 $f(x;\theta)$를 가지는 분포에서 추출하였다.

$$f(x;\theta) = \begin{cases}\theta x^{\theta-1} & \text{if } 0 < x \le 1,\ 0 < \theta < \infty\\ 0 & \text{o.w.}\end{cases}$$

$-\ln(X_1)$의 기댓값과 분산을 구하시오.

Solution. $Y_i = -\ln(X_i)$으로 정의하면,

$$\begin{aligned}P(Y_i \le y) &= P(-\ln(X_i) \le y)\\ &= P(e^{-y} \le X_i)\\ &= \int_{e^{-y}}^1 \theta x^{\theta-1}dx\\ &= [x^\theta]_{e^{-y}}^1\\ &= 1 - e^{-y\theta}\end{aligned}$$

으로 지수분포의 누적분포함수와 같다. 따라서 Y_i들은 모수가 기대값이 $1/\theta$인 지수분포로부터의 랜덤표본이다. 따라서 기대값과 분산은 각각 θ^{-1}, θ^{-2}이다.

Problem. 위 문제에서, $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$의 기대값과 분산을 구하시오.

Solution.

$$-\frac{1}{n}\ln(X_1X_2\cdots X_n) = \frac{1}{n}\sum_{i=1}^n(-\ln X_i) = \frac{1}{n}\sum_{i=1}^n Y_i$$

이므로, 그 기대값과 분산은 각각 $\theta^{-1}, n^{-1}\theta^{-2}$이다.

Problem. 위 문제에서, $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$이 θ^{-1}에 대한 완비충분통계량임을 보이시오.

Solution.

$$f(x;\theta) = \exp(\ln\theta + \theta\ln x - \ln x)$$

으로 지수족이고, 그 support는 θ에 무관하게 $0 < x < 1$이며, 모수공간이 열린집합이고, 앞선 문제에 의하여

$\mathrm{Var}(c(-\ln X)) = 0$인 nonzero c가 없으므로,

$$\sum_{i=1}^{n}(-\ln X_i) = -\ln(X_1X_2\cdots X_n)$$

은 완비충분통계량이고, 여기에 상수를 곱한 $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$ 역시도 θ에 대한 완비충분통계량이 된다. $\theta \mapsto \theta^{-1}$ 연산은 invertible이므로, 이는 θ^{-1}에 대한 완비충분통계량이기도 하다.

Problem. 위 문제에서, $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$이 θ^{-1}에 대한 최소분산불편추정량임을 보이시오.

Solution. 레만-쉐페 정리를 이용할 수 있다. 앞선 문제로부터 $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$이 θ^{-1}에 대한 불편추정량이므로, 라오-블랙웰화를 통해 얻은 추정량 역시 아래와 같이 자신과 같다.

$$-\frac{1}{n}\ln(X_1X_2\cdots X_n) = \mathbb{E}_{\theta^{-1}}\left[-\frac{1}{n}\ln(X_1X_2\cdots X_n)\middle| -\frac{1}{n}\ln(X_1X_2\cdots X_n)\right]$$

따라서 $-\frac{1}{n}\ln(X_1X_2\cdots X_n)$이 θ^{-1}는 UMVUE이다.

2012년

Problem. 제조업체의 설비투자액이 클수록 매출액이 큰지를 조사하기 위하여 같은 제품을 생산하는 10개 업체의 설비투자액과 매출액을 조사하였다. 아래와 같이 매출액을 종속변수로, 설비투자액을 설명변수로 하는 단순선형회귀모형을 설정하고 회귀모형에 대해 적합결여검정을 실시하고자 한다. 다음 물음에 답하시오.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \cdots, 10, \ (\epsilon_i \sim_{i.i.d.} N(0, \sigma^2))$$

<설비투자액과 매출액 자료>

설비투자액(X)	4	6	6	8	8	9	9	10	12	12
매 출 액(Y)	39	42	45	47	50	50	52	55	57	60

요 인	자유도(DF)	제곱합(SS)	평균제곱합(MS)
회귀(regression)	1	378.5	378.5
오차(error)	8	17.6	2.2
적합결여	①	③	⑤
순오차	②	④	3.875
합계	9	396.1	

적합결여검정을 위한 분산분석표를 채우고, 그 귀무가설을 제시한 뒤 유의수준 5%에서 검정하시오.

Solution. 오차 중 적합결여에 의한 오차는 중복된 X값이 있는 6, 8, 9, 12에서 나타나며 이에 따라 적합결여의 자유도는 $6 - 2 = 4$, 순오차의 자유도는 $10 - 6 = 4$이다. 한편 순오차의 평균제곱합이 3.875이므로 제곱합은 여기에 4를 곱한 15.5이고, 적합결여의 제곱합은 17.6에서 15.5를 뺀 2.1이며, 그 평균제곱합은 2.1을 4로 나눈 0.525이다.
적합결여검정의 귀무가설은

$$H_0 : \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$$

으로 나타나며, F통계량은 $0.525/3.875 = 0.1355$로 $F(4, 4)$분포의 95% quantile보다 작은 것이 분명하므로, 귀무가설을 기각할 만한 충분한 근거가 없다.

Problem. 전구 생산시간 문제

Solution. 설명편 참고

Problem. 대한이, 민국이, 한은이 세 사람은 1970년-2010년 중 우리나라의 국내총생산(Y_t)과 수출(X_t) 자료를 이용하여 각자 알맞은 회귀모형을 적합하고자 한다. 먼저, 대한이는 모형 I을 적합시켜 다음의 ANOVA 테이블 및 추정결과를 얻었다. 표의 빈칸을 채우시오.

모형 I : $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \quad (\epsilon_t \sim N(0,\sigma^2)$이고 서로 독립)

<ANOVA 테이블>

	DF	SS	MS	F-value
X_t	1	504.1	504.1	①
residuals	39	5,235.9	134.3	

<추정결과 I>

	Estimate	Std.Error	t-value	Pr(>|t|)
β_0	14.777	3.686	②	0.0003
β_1	0.296	0.153	③	④

Solution. 1번은 MS까리 나누어 F값 504.1/134.3 = 3.754로 나타난다. 한편 추정결과 테이블에서는, t값은 estimate를 std.error로 나눈 4.009와 1.935로 나타나며, β_1은 자유도가 39인 t분포를 따르기에 해당 분포 하에서 1.935의 위치를 확인함으로써 약 0.0603을 얻는다.

위 상황에서, 민국이는 변수를 로그변환하여 분석을 실시해야 한다고 주장하면서 모형 II를 적합하여 아래의 결과를 얻었다. 이 결과로부터 모형 II가 대한이가 이용한 모형 I에 비해 우월하다고 판단할 수 잇는지에 대한 의견을 제시하시오.

모형 I : $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \quad (\epsilon_t \sim N(0,\sigma^2)$이고 서로 독립)

<ANOVA 테이블>

	DF	SS	MS	F-value
X_t	1	504.1	504.1	①
residuals	39	5,235.9	134.3	

<추정결과 I>

	Estimate	Std.Error	t-value	Pr(>|t|)
β_0	14.777	3.686	②	0.0003
β_1	0.296	0.153	③	④

Solution. Y_t를 $\ln Y_t$로 변형시켜 스케일이 달라졌기 때문에 완전한 비교는 어렵지만, 이 모형이 더욱 우월해 보인다. 회귀계수의 표준오차가 매우 감소한 것을 볼 수 있는데, 이는 추정량의 효율성이 증가하였음을 의미하기 때문이다.

Problem. 한은이는 각 변수를 로그변환한 후 1차 차분하여 모형 III을 적합하였다. 세 사람이 적합시킨 모형에 대하여 각각 3개의 통계량을 얻었다. 이 결과로부터 어느 한 모형이 다른 모형에 비하여 우월하다고 판단할 수 있는가? 그 이유를 설명하시오.

모형 III: $\Delta \ln Y_t = \alpha_0 + \alpha_1 \Delta \ln X_t + \epsilon_t \quad (\epsilon_t \sim N(0,\sigma^2)$이고 서로 독립)

Solution.

<모형별 주요 통계량>

	AIC통계량	DW통계량	Jarque Bera
모형 I	321.19	0.504	0.151
모형 II	107.64	0.734	< 0.001
모형 III	219.25	1.530	< 0.001

(단, Jarque Bera는 p-값을 나타냄)

먼저 AIC 통계량을 보았을 때, AIC 통계량은 로그가능도가 커질수록 작아짐을 고려하면 모형I, 모형III에 비해 모형II가 더욱 우월해 보인다(선형성 가정). 반면 DW 통계량을 보면, DW 통계량이 2에 가까울수록 잔차의 자기상관성이 없음을 고려할 때 모형I과 모형II는 잔차의 자기상관성이 커 보인다. 즉 시계열적 관계를 제대로 묘사하지 못하고 있어, 모형III이 우월하다고 보인다(독립성 가정). 마지막으로 Jarque Bera의 검정 결과를 보면, 모형 II와 모형 III에서 얻은 잔차는 정규분포를 따른다고 말하기 어려운 반면 모형 I은 이들에 비해 그나마 잔차가 정규분포에 가까웠다(정규성 가정). 즉 각각의 모형이 모두 장단점을 가지고 있어, 모형들 간의 우월성을 판단하기 어렵다.

Problem. 크기 n인 확률표본 $X_1, \cdots, X_n$을 확률밀도함수가 $f(x)$인 균등분포에서 추출하였을 때, n번째 순서통계량 $Y = X_{(n)}$의 확률밀도함수 $g(y)$는 아래와 같다.

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}, \quad 0 < \theta < \infty, \qquad g(y) = \begin{cases} n\frac{y^{n-1}}{\theta^n} & \text{if } 0 \le y \le \theta \\ 0 & \text{otherwise} \end{cases}, \quad 0 < \theta < \infty$$

확률변수 Y가 θ의 불편추정량인지 판단하고 그 근거를 제시하시오. 그 다음 Y가 θ의 일치추정량이자 충분통계량임을 보이시오. 그 다음 이를 Y가 θ의 완비충분통계량임을 가정한 뒤 이를 이용하여 θ의 UMVUE를 구하시오.

Solution.

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^\theta ny\frac{y^{n-1}}{\theta^n}dy \\ &= \frac{n}{n+1}\left[\frac{y^{n+1}}{\theta^n}\right]_0^\theta = \frac{n}{n+1}\theta \end{aligned}$$

이므로 θ의 불편추정량은 아니다. 한편 모든 $\epsilon > 0$에 대하여

$$\begin{aligned} P(|Y-\theta| \ge \epsilon) &= 1 - P(\theta - \epsilon < Y < \theta + \epsilon) \\ &= 1 - \int_{(\theta-\epsilon)\wedge 0}^{\theta} n\frac{y^{n-1}}{\theta^n}dy \\ &= \min\left\{\frac{(\theta-\epsilon)^n}{\theta^n}, 0\right\} \to 0 \quad \text{as } n \to \infty \end{aligned}$$

이므로, $Y \xrightarrow{p} \theta$이며 Y는 θ의 일치추정량이다. 한편 표본으로부터의 가능도함수가

$$L(\theta; x) = \frac{1}{\theta^n} I_{(\theta \ge y)}$$

으로 주어지기에, 분해정리에 의하여 Y는 θ의 충분통계량이다. Y가 θ의 완비충분통계량임을 가정하면, 레만-쉐페 정리에 의하여 θ의 불편추정량 $\frac{n+1}{n}Y$의 Y에 대한 조건부 기대값 $\frac{n+1}{n}Y$가 θ의 UMVUE가 된다.

2011년

Problem. 확률변수 X, Y는 서로 독립이고 표준정규분포 $N(0,1)$을 따른다. 확률변수 $Z = \dfrac{Y}{X}$를 정의할 때, 확률변수 Z의 확률밀도함수를 구하고, 이에 대해 중심극한정리를 적용할 수 없음을 보이시오.

Solution. 아래처럼 누적분포함수를 구할 수 있다.

$$\begin{aligned}
P(Z \le z) &= P(Y/X \le z) \\
&= P(Y/X \le z, X > 0) + P(Y/X \le z, X < 0) \\
&= P(Y \le zX, X > 0) + P(Y \ge zX, X < 0) \\
&= \int_0^\infty \int_{-\infty}^{zx} \frac{1}{2\pi} e^{-x^2/2 - y^2/2} dy dx + \int_{-\infty}^0 \int_{zx}^\infty \frac{1}{2\pi} e^{-x^2/2 - y^2/2} dy dx \\
&= \int_0^\infty \frac{1}{2\pi} e^{-x^2/2} \int_{-\infty}^{zx} e^{-y^2/2} dy dx + \int_{-\infty}^0 \frac{1}{2\pi} e^{-x^2/2} \int_{zx}^\infty e^{-y^2/2} dy dx \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Phi(zx) dx + \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Phi(-zx) dx \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Phi(zx) dx + \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Phi(zx) dx \\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-x^2/2} \Phi(zx) dx
\end{aligned}$$

확률밀도함수는 양변을 z로 미분하여

$$f_Z(z) = \sqrt{\frac{2}{\pi}} \int_0^\infty x e^{-x^2/2} \phi(zx) dx = \frac{1}{\pi} \int_0^\infty x e^{-x^2/2} e^{-z^2 x^2/2} dx = \frac{1}{\pi(1+z^2)}$$

을 얻는다. 즉 코시분포와 같다. 코시분포는 대표적으로 평균과 분산을 가지지 않는 분포로, 중심극한정리에서는 이차적률까지 존재함을 가정하므로 중심극한정리를 적용할 수 없다.

Problem. 화학약품을 합성하는 공정에서 반응온도의 수준에 따라 생산량이 차이가 있는지를 검정하기 위한 실험을 실시하여 아래의 데이터를 얻었다. 분산분석표의 빈칸을 채우고, 생산량이 반응온도의 변화에 따라 유의한 영향을 받는가에 대한 F검정을 실시하시오.

<반응온도에 따른 생산량>

T_1 = 80℃	T_2 = 90℃	T_3 = 100℃	T_4 = 110℃
90.1	89.8	91.6	91.3
90.0	90.5	91.4	90.0
89.5	90.8	91.1	90.6

<분산분석표>

요인	제곱합(SS)	자유도(df)	평균제곱합	F통계량
반응온도(T)	3.52	(B1)	(C)	(E)
잔차(Error)	(A)	(B2)	(D)	
합계	5.23	(B3)		

Solution. 먼저 잔차제곱합 A는 합계제곱합에서 반응온도에 대한 제곱합을 빼어 1.71을 얻는다. 일원배치법 하에서 B1, B2, B3의 자유도는 각각 $4-1=3, 12-4=8, 12-1=11$으로 주어진다. 평균제곱합 C, D는 제곱합을 각각 자유도로 나누어 1.17, 0.21로 나타나며, F통계량은 그 비인 5.57로 나타난다. 한편 이는 자유도가 3, 8인 F분포 하에서의 95퍼센트 퀀타일인 4.07보다 크므로, 귀무가설을 기각할 수 있다. 즉 생산량이 반응온도의 변화에 따라 유의한 영향을 받는다.

Problem. 한국대학교 1학년을 대상으로 정규수업시간을 제외한 주당 학습시간에 대해 표본조사를 실시하였다. 전체 1500명의 학생 중 200명을 단순확률추출하여 조사한 결과 135명의 학생이 응답하였으며, 조사결과를 요약한 내용은 다음과 같다.

<주당 학습시간에 대한 조사결과>

	학생수(N_i)	표본수	응답수(n_i)	$\sum_j^{n_i} y_{ij}$	$\sum_j^{n_i} y_{ij}^2$
남학생	600	70	31	249	2,939.5
여학생	900	130	104	1,287	20,515.5
계	1,500	200	135	1,536	23,455.0

단, $i=1$이면 남학생, $i=2$이면 여학생

조사결과에서 무응답을 무시하고 표본수 135인 단순확률표본이라고 간주했을 때, 표본평균과 표본평균의 추정분산을 계산하시오.

Solution. 단순확률표본으로 간주하는 경우

$$\bar{y} = \frac{\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}}{n_1 + n_j} = \frac{1536}{135} = 11.378$$

을 얻으며, 추정분산은

$$\hat{V}(\bar{y}) = \left(1 - \frac{135}{1500}\right) \frac{S^2}{135} = \frac{1}{135}\left(1 - \frac{135}{1500}\right) \frac{23455 - 135 \times (11.378)^2}{135 - 1} = 0.301$$

으로 얻는다.

Problem. 위 문제에서, 성별에 따른 응답률 격차가 너무 크다고 판단하여 사후층화를 통해 무응답조정을 하고자 한다. 사후층화 보정 후 표본평균 $\bar{y}_{ps}$과 표본평균의 추정분산 $\hat{V}(\bar{y}_{ps})$를 계산하고, 앞 문항의 결과와 비교하여 차이점을 약술하시오.

Solution. 사후층화 시

$$\bar{y}_{ps} = \frac{600}{1500} \times \frac{249}{31} + \frac{900}{1500} \times \frac{1287}{104} = 10.638$$

을 표본평균으로 얻는다. 그 추정분산은

$$\begin{aligned}\hat{V}(\bar{y})_{ps} &= \left(1 - \frac{135}{1500}\right)\left(\frac{600}{1500} \times \frac{2939.5 - 31 \times (8.032^2)}{135 \times 30} + \frac{900}{1500} \times \frac{20515.5 - 104 \times (12.375^2)}{135 \times 103}\right) \\ &\quad + \frac{1}{135}\left(\frac{900}{1500} \times \frac{2939.5 - 31 \times (8.032^2)}{135 \times 30} + \frac{600}{1500} \times \frac{20515.5 - 104 \times (12.375^2)}{135 \times 103}\right) \\ &= 0.2667\end{aligned}$$

이며 앞 문항에 비하여 분산이 줄어들어 효율적인 추정량이 되었음을 알 수 있다. 이는 층화를 통하여 무응답조정을 수행하였기 때문이다.

Problem. 확률변수 X가 아래 분포를 따르는 하나의 확률표본이라고 할 때, $g(\theta) = \dfrac{1}{\theta^2}$의 최소분산불편추정량을 구하려 한다.

$$f(x|\theta) = \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}},\ -\infty < x < \infty,\ \theta > 0$$

$T = \log(1 + e^{-X})$이 θ에 대한 완비충분통계량임을 보이시오.

Solution. X의 확률밀도함수는 아래의 지수족처럼 쓸 수도 있다.

$$f(x|\theta) = \exp\left(\theta \log(1+e^{-x}) + \log\theta - x - \log(1+e^{-x})\right)$$

여기에서 분포의 토대 $\mathbb{R}$이 θ에 불변하며, 모수공간 $(0, \theta)$가 $\mathbb{R}$의 열린 집합이며, $\mathrm{Var}(c\log(1+e^{-X})) = 0$인 $c \neq 0$이 없으므로 $\log(1+e^{-X})$는 θ에 대한 완비충분통계량이 된다.

Problem. 앞 문항에서 구한 완비충분통계량에 대해 $\mathbb{E}[T] = \theta^{-1}, \mathrm{Var}(T) = \theta^{-2}$임을 보이시오.

Solution. T의 분포를 구하자. $t \geq 0$에 대하여

$$\begin{aligned}P(T \leq t) &= P(\log(1+e^{-X}) \leq t) \\ &= P(1 + e^{-X} \leq e^t) \\ &= P(e^{-X} \leq e^t - 1) \\ &= P((e^t-1)^{-1} \leq e^X) \\ &= P(-\ln(e^t - 1) \leq X) \\ &= \int_{-\ln(e^t-1)}^{\infty} \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}} dx \\ &= \int_1^{e^t} \frac{\theta}{v^{\theta+1}} dv \quad (v = 1 + e^{-x}) \\ &= [-v^{-\theta}]_1^{e^t} = 1 - e^{-t\theta}\end{aligned}$$

이므로 이는 평균이 $1/\theta$인 지수분포를 따른다. 따라서 평균이 θ^{-1}, 분산이 θ^{-2}이다.

Problem. 위 문항을 이용하여 $g(\theta)$의 UMVUE를 구하시오.

Solution. 위 문제로부터

$$\mathrm{Var}(T) = \mathbb{E}[T^2] - (E[T])^2 = \mathbb{E}[T^2] - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

임을 알기에

$$\mathbb{E}[0.5T^2] = g(\theta)$$

이다. 따라서 $0.5T^2$은 $g(\theta)$의 불편추정량이고 T가 CSS이므로 레만 쉐페 정리에 의하여 UMVUE은 $0.5T^2$이다.

Problem. 아래는 한국시리즈 5차전 입장관중 200명을 단순확률추출하여 이전 경기까지 얼마나 많은 경기를 관전했는지 조사한 결과이다.

<조사결과>

관전경기 횟수	응답인원(명)
0	33
1	67
2	66
3	15
4	19

위의 조사결과가 아래와 같은 분포를 따르는지 검정하고자 한다.

$$p_0 = (1-\theta)^4,\ p_1 = 4\theta(1-\theta)^3,\ p_2 = 6\theta^2(1-\theta)^2,\ p_3 = 4\theta^3(1-\theta),\ p_4 = \theta^4,\ (0<\theta<1)$$

(단, p_i 는 경기를 i번 관전할 확률. 예를 들면 p_3 는 전체 네 경기 중 세 경기를 관전할 확률)

θ의 최대우도추정량을 구하고, 이를 바탕으로 주어진 가설을 검정하기 위한 χ^2 검정통계량을 구한 뒤 적합도 검정을 실시하시오.

Solution. 먼저 최대우도추정량을 구하자. 가능도함수는

$$\begin{aligned} L(\theta; x) &= ((1-\theta)^4)^{33} \times (4\theta(1-\theta)^3)^{67} \times (6\theta^2(1-\theta)^2)^{66} \\ &\quad \times (4\theta^3(1-\theta))^{15} \times (\theta^4)^{19} \\ &= 4^{82} \times 6^{66} \times \theta^{320} \times (1-\theta)^{480} \end{aligned}$$

으로 주어지므로, 로그가능도함수는

$$l(\theta; x) = 82\log 4 + 66\log 6 + 320\log\theta + 480\log(1-\theta)$$

이며 일계조건은

$$\frac{320}{\theta} - \frac{480}{1-\theta} = 0$$

으로 주어진다. 따라서 $\theta = 0.4$이다. $\hat{\theta} = 0.4$를 최대우도추정량으로 할 때, 예상되는 응답비율은 0번부터 각각

$$\frac{81}{625}, \frac{216}{625}, \frac{216}{625}, \frac{96}{625}, \frac{16}{625}$$

이며 200을 곱하면

$$25.92, 69.12, 69.12, 30.72, 5.12$$

이다. 이를 바탕으로 적합도 검정을 실시하면 χ^2 검정통계량은 47.81이고, 이는 자유도가 4인 카이제곱분포에서의 95퍼센트 퀀타일보다 분명히 더 크다. 따라서 귀무가설을 기각한다.

2010년

Problem. 베이즈 추정량 문제

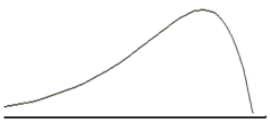
Solution. 설명편 참고

Problem.

☐ 다음 중 <u>옳지 않은</u> 것은?

A. 추정량(estimator)의 기대값이 모수와 같으면 이러한 추정량을 불편추정량(unbiased estimator)이라 한다.

B. 분포의 모양이 아래와 같이 왼쪽 꼬리가 긴(skewed to the left) 경우 평균에 대한 3차 적률(moment)은 양수(陽數)가 된다.



C. 1종 오류를 범할 확률을 유의수준(significance level)이라 한다.

D. 산술평균을 A, 기하평균을 G, 조화평균을 H라 할 때 A≥G≥H이다.

E. 추정량이 모수에 확률수렴(convergence in probability)하면 이러한 추정량을 일치추정량(consistent estimator)이라 한다.

Solution.

A. 그렇다.

B. 그렇지 않다. 왼쪽 꼬리가 긴 경우 평균에 대한 3차 적률은 음수가 된다.

C. 그렇다.

D. 그렇다.

E. 그렇다.

Problem. PCA 문제

Solution. 설명편 참고

Problem. 우리나라의 2000.Q1-2009.Q2의 GDP자료 y를 이용하여 계절성이 있는지를 검정하려고 한다. GDP 자료는 로그변환($ly = \log(y)$)하여 사용하였고 1, 2, 3, 4분기에 대한 계절더미변수 $s1, s2, s3, s4$를 생성하였다. 각 계절더미변수는 해당 분기에는 1의 값을 가지고 여타 분기에는 0의 값을 지닌다. 아래 회귀분석모형 I을 적합하여 다음과 같은 결과를 얻었다.

회귀분석모형 Ⅰ : $ly_t = \beta_0 + \beta_1 t + \beta_2 s2_t + \beta_3 s3_t + \beta_4 s4_t + \epsilon_t$

회귀분석모형 추정 결과

Dependent Variable: ly

Number of Observations Read 38
Number of Observations Used 38

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.55578	0.13894	333.46	<.0001
Error	33	0.01375	0.00041667		
Corrected Total	37	0.56953			

Root MSE	0.02041	R-Square	0.9759
Dependent Mean	12.24953	Adj R-Sq	0.9729
Coeff Var	0.16664		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	11.99182	0.00864	1387.90	<.0001	5701.94146	802.62535
t	1	0.01040	0.00030228	34.40	<.0001	0.50475	0.49298
s2	1	0.06400	0.00913	7.01	<.0001	0.00109	0.02046
s3	1	0.05895	0.00938	6.29	<.0001	0.00069661	0.01646
s4	1	0.10201	0.00938	10.87	<.0001	0.04924	0.04924

계절더미변수들이 모두 유의한지를 검정하는 귀무가설을 쓰고 이를 검정하시오.

Solution.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

이를 검정하기 위한 F통계량은

$$F = \frac{(SSE(RM) - SSE(FM))/3}{SSE(FM)/33} = \frac{0.00109 + 0.00069661 + 0.04924}{3 \times 0.00041667} = 40.821$$

으로, 자유도가 3, 33인 F분포 상에서 굉장히 큼이 분명하다. 따라서 귀무가설을 기각한다.

Problem. $s1, s2, s3$을 이용하여 아래의 회귀분석모형 II를 적합하였을 때 s1의 계수 γ_2 및 그 분산을 구하시오.

회귀분석모형 Ⅱ : $ly_t = \gamma_0 + \gamma_1 t + \gamma_2 s1_t + \gamma_3 s2_t + \gamma_4 s3_t + \eta_t$

Solution. 위 문제에서, 회귀분석모형 I의 적합에 사용되는 model matrix X를 고려하자. 그렇다면 X는

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 38 & 1 & 0 & 0 \end{pmatrix}$$

과 같으며, 1열은 1, 2열은 1-38이 나열되어 있으며, 3, 4, 5열은 각각 0100, 0010, 0001가 반복되는 형태를 가진다. 이로부터 우리가 얻는 회귀모형 하에서 $\hat{\beta}$는

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

이다. 이때 $Y = (ly_1, ly_2, \cdots, ly_{38})^T$이다. 둘째로 회귀분석모형 II의 적합에 사용되는 model matrix는

$$\tilde{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 1 & 3 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 38 & 0 & 1 & 0 \end{pmatrix}$$

으로, X와 1, 2열은 같으나 3, 4, 5열이 각각 1000, 0100, 0010이 반복되는 형태이다. 한편 계절성에 대한 고려로부터,

$$\tilde{X} = X \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 \end{pmatrix} := XK$$

임을 알 수 있으며,

$$\hat{\gamma} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^TY = (K^TX^TXK)^{-1}K^TX^TY = K^{-1}(X^TX)^{-1}X^TY = K^{-1}\hat{\beta}$$

를 얻는다. 한편

$$K^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

이므로, $\hat{\gamma}_2 = -\hat{\beta}_4 = -0.10201$이며, 분산은 $(0.00938)^2 = 0.000088$이다.

Problem. 위 문제에서, 누군가가 계절요인들의 크기가 시간에 비례하여 달라질 수도 있다는 의문을 제기하였다. 이를 확인하기 위해 적절한 회귀분석 모형식을 서술하시오.

Solution.

$$ly_t = \delta_0 + \delta_1 t + \delta_2 s2_t + \delta_3 s3_t + \delta_4 s4_t + \delta_5 ts2_t + \delta_6 ts3_t + \delta_7 ts4_t + \zeta_t$$

등을 사용할 수 있다.

Problem. A시청의 세무담당자는 도시내 식료품 소매상들의 평균 매출액이 어느 정도인지 알고 싶어졌다. 총 800개의 소매상을 모두 조사할 수 없어서 세무담당자는 표본조사를 실시하기로 하였다. 세무담당자는 표본설계 과정에서 도시내 식료품 소매상이 소규모, 중규모, 대규모로 구분되고 각각 400, 320, 80개라는 것을 알게 되었다. 표본추출방법은 층화추출, 표본크기는 40으로 결정하고 비례배분으로 표본을 각 층에 배분한 후 표본조사를 실시하였다. 다음의 조사결과를 이용하여, 각 층의 표본크기, 평균매출액과 총매출액의 값과 분산을 계산/추정하시오.

표본조사 결과

	층(h)		
	소규모($h=1$)	중규모($h=2$)	대규모($h=3$)
$\sum_{i=1}^{n_h} y_{hi}$	200	240	120
$\sum_{i=1}^{n_h} y_{hi}^2$	2,038	3,675	3,660

단, y_{hi}는 층 h에서의 i번째 소매상의 매출액

Solution. 비례배분을 하였으므로 각 층별 표본크기는 20, 16, 4로 나타난다. 평균 매출액은 비례배분하였으므로

$$\bar{y}_{st} = \frac{200+240+120}{20+16+4} = 14$$

이며, 분산은

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{3} \frac{N_h^2}{n_h N^2}\left(1-\frac{n_h}{N_h}\right)s_h^2 = 0.119$$

로 나타난다. 총매출액은

$$\hat{Y}_{st} = 14 \times 800 = 11200$$

로 추정되며, 그 분산은

$$\hat{V}(\hat{Y}_{st}) = \sum_{h=1}^{3} \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)s_h^2 = 76000$$

으로 추정된다.

Problem. 악력 계산 문제

Solution. 설명편 참고

Problem. 서로 독립인 확률변수 $X_1, \cdots, X_n$이 성공확률 p인 베르누이 분포를 따른다. $S = \sum_{i=1}^{n} X_i$에 대하여 S^2/n^2이 p^2의 최대우도추정량이지만 불편추정량은 아님을 보여라.

Solution. 가능도함수는

$$L(p;x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = (1-p)^n \times \left(\frac{p}{1-p}\right)^{\sum_{i=1}^{n} x_i}$$

으로 주어지며, $s = \sum_{i=1}^{n} x_i$라 할 때 로그가능도함수는

$$l(p;x) = s\log(p) + (n-s)\log(1-p)$$

이며 일계조건으로부터

$$\hat{p} = \frac{S}{n}$$

을 얻고, 양변을 제곱하면 p^2의 최대우도추정량이 S^2/n^2임을 알 수 있다. 그러나

$$\begin{aligned}\mathbb{E}[S^2/n^2] &= \frac{1}{n^2}\mathbb{E}[S^2] \\ &= \frac{1}{n^2}(\mathrm{Var}(S) + \mathbb{E}[S]^2)\end{aligned}$$

$$= \frac{1}{n^2}(np(1-p) + n^2p^2)$$
$$= p^2 + \frac{p(1-p)}{n}$$

으로 $p \in (0,1)$인 경우 0이 아닌 편의 $p(1-p)/n$이 존재한다. 따라서 p^2의 불편추정량은 아니다.

Problem. 위 문제에서, 아래의 잭나이프 추정량 J가 p^2의 불편추정량임을 보이시오.

$$J = n\frac{S^2}{n^2} - \frac{n-1}{n}\left(S\frac{(S-1)^2}{(n-1)^2} + (n-S)\frac{S^2}{(n-1)^2}\right)$$

Solution.

$$\begin{aligned}
\mathbb{E}[J] &= \mathbb{E}\left[n\frac{S^2}{n^2} - \frac{n-1}{n}\left(S\frac{(S-1)^2}{(n-1)^2} + (n-S)\frac{S^2}{(n-1)^2}\right)\right] \\
&= n\,\mathbb{E}[S^2/n^2] - \frac{1}{n(n-1)}\,\mathbb{E}\left[S(S-1)^2 + (n-S)S^2\right] \\
&= np^2 + p(1-p) - \frac{1}{n(n-1)}\,\mathbb{E}[-2S^2 + S + nS^2] \\
&= np^2 + p(1-p) - \frac{(n-2)(n^2p^2 + np(1-p)) + np}{n(n-1)} \\
&= \frac{1}{n-1}(np^2 + p(1-p) - p) = p^2
\end{aligned}$$

Problem. 위 문제에서 주어진 잭나이프 추정량은 p^2의 UMVUE인가? 만일 그렇다면 그 근거를 제시하고 그렇지 않다면 최소분산 불편추정량을 제시하시오.

Solution. 베르누이 분포의 확률밀도함수가

$$f(x) = p^x(1-p)^{1-x} = \exp\left(x\log p + (1-x)\log(1-p)\right) = \exp((\log p/(1-p))x + \log(1-p)$$

이고 분포의 토대가 $\{0,1\}$으로 $p \in (0,1)$에서 불변이고 $(0,1)$은 열린 집합이며, $\mathrm{Var}(cX) = 0$인 c가 존재하지 않으므로

$$S = \sum_{i=1}^{n} X_i$$

는 CSS이다. J는 p^2의 불편추정량이면서, S에 조건부 기대값을 취해도 자신과 같으므로, 레만-쉐페 정리에 의하여 UMVUE이다.

Problem. 행렬 X가 다음과 같이 주어져 있다. X의 각 열을 $C_1, \cdots, C_5$로 명명하기로 한다.

$$X = \begin{bmatrix} 1 & 1 & 2 & 8 & 9 \\ 1 & 1 & 2 & 8 & 9 \\ 1 & 1 & 2 & 8 & 9 \\ 1 & 1 & 1 & 7 & 7 \\ 1 & 1 & 1 & 7 & 7 \\ 1 & 1 & 1 & 7 & 7 \\ 1 & -1 & 1 & 1 & 7 \\ 1 & -1 & 1 & 1 & 7 \\ 1 & -1 & 1 & 1 & 7 \\ 1 & -1 & 1 & 1 & 7 \end{bmatrix},\ C_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ C_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix},\ C_3 = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ C_4 = \begin{bmatrix} 8 \\ 8 \\ 8 \\ 7 \\ 7 \\ 7 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ C_5 = \begin{bmatrix} 9 \\ 9 \\ 9 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \end{bmatrix}$$

위 행렬에서 C_4, C_5는 C_1, C_2, C_3의 선형결합으로 표시할 수 있다. 아래 식의 빈칸을 채우고, X^TX의 계수를 구하시오.

$$C_4 = 3C_1 + (\ ①\ )C_2 + (\ ②\ )C_3$$
$$C_5 = (\ ③\ )C_1 + (\ ④\ )C_2 + 2C_3$$

Solution. 단순 계산을 통하여,

$$C_4 = 3C_1 + 3C_2 + C_3$$
$$C_5 = 5C_1 + 0C_2 + 2C_3$$

임을 안다. 한편 C_1, C_2, C_3는 선형독립이므로, X는 column rank로 3을 가진다. 따라서 X^TX는 rank로 3을 가진다.

Problem. 위 문제에서, $Y = aC_4 + bC_5$라고 하자. Y를 C_1, C_2, C_3에 대하여 회귀 적합한다고 할 경우 세 변수에 대한 회귀계수를 a, b에 대해 쓰고, 오차항 벡터를 구하시오.

Solution.

$$\begin{aligned} Y &= aC_4 + bC_5 \\ &= a(3C_1 + 3C_2 + C_3) + b(5C_1 + 0C_2 + 2C_3) \\ &= (3a + 5b)C_1 + 3aC_2 + (a + 2b)C_3 \end{aligned}$$

으로 나타내지고, 오차항벡터는 영벡터이다.

Problem. 위 문항에서 주어진 Y를 C_1, C_2, C_3, C_4, C_5에 대하여 회귀 적합한다고 할 경우 회귀분석과 관련하여 함축하고 있는 의미를 간략히 기술하시오.

Solution. 이들 모두를 사용하는 경우,

$$\begin{aligned} Y &= aC_4 + bC_5 \\ &= (3a + 5b)C_1 + 3aC_2 + (a + 2b)C_3 \end{aligned}$$

의 두 표현이 모두 가능하며 최적의 회귀계수이다. 즉 X가 full column rank가 아닌 경우, Y의 회귀식을 표현할 수 있는 유일한 방법이 존재하지 않고, identifiable하지 않게 됨을 의미한다.

2009년

Problem.

□ 변수 X와 Y간의 두 회귀식이 $Y=aX+b$, $X=a_1Y+b_1$ 이고, r을 두 변수 간의 상관계수라고 할 때 다음 중 **옳은** 것은?

A. $r=a\times a_1$ B. $a=a_1$ C. $r=\sqrt{\frac{a}{b}\times\frac{a_1}{b_1}}$

D. $a_1=0$이면 $r=0$ E. $r^2=a\times a_1$

Solution.

$$\begin{aligned} r &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ &= \frac{\mathrm{Cov}(a_1Y+b_1, aX+b)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ &= a_1 a \frac{\mathrm{Cov}(Y,X)}{\sqrt{\mathrm{Var}(Y)\,\mathrm{Var}(X)}} = a_1 a r \end{aligned}$$

이므로, $a_1=0$이면 $r=0$이다. 즉 D가 참이다.

Problem. 어떤 도시의 유권자를 대상으로 A 후보의 공약에 대한 찬반을 조사하였다. 임의로 900명을 추출하여 찬성과 반대를 조사한 결과 찬성이라고 대답한 사람이 378명이었을 때, 이 도시의 유권자 전체에 대한 찬성 비율의 95퍼센트 신뢰구간은?

Solution. 정규근사 플러그인 신뢰구간은

$$378/900 \pm 1.96 \times \sqrt{378\times 522/900^3} = (0.39, 0.45)$$

이다.

Problem.

□ 다음 확률의 성질 중 **옳지 않은** 것은 ? (단, $P(A)>0,\ P(B)>0,\ P(C)>0$)

A. $P(A|B)\geq P(A)$이면, $P(B|A)\geq P(B)$이다.

B. $P(A\cap C|B)=P(A|B)P(C|B)$이면, $P(A|B\cap C)=P(A|B)$이다.

C. $P(A\cup B|C)+P(A\cap B|C)=P(A|C)+P(B|C)$이다.

D. $P(A\cap B)=P(A)P(B),\ P(A\cap C)=P(A)P(C),\ P(B\cap C)=P(B)P(C)$이면, A, B, C는 서로 독립(mutually independent)이다.

E. $P(A)=P(A|B)P(B)+P(A|B^c)P(B^c)$이다.

Solution.

A.

$$\begin{aligned}P(B|A)&=\frac{P(A\cap B)}{P(A)}\\&=\frac{P(A|B)P(B)}{P(A)}\\&\geq\frac{P(A)P(B)}{P(A)}=P(B)\end{aligned}$$

B.

$$\begin{aligned}P(A|B\cap C)&=\frac{P(A\cap B\cap C)}{P(B\cap C)}\\&=\frac{P(A\cap C|B)P(B)}{P(B\cap C)}\\&=\frac{P(A|B)P(C|B)P(B)}{P(B\cap C)}\\&=P(A|B)\end{aligned}$$

C.

$$\begin{aligned}&P(A\cup B|C)+P(A\cap B|C)\\&\qquad=P((A-B)\sqcup B|C)+P(A\cap B|C)\\&\qquad=P(A-B|C)+P(A\cap B|C)+P(B|C)\\&\qquad=P((A-B)\sqcup(A\cap B)|C)+P(B|C)\\&\qquad=P(A|C)+P(B|C)\end{aligned}$$

D. 반례가 있다. 주사위 두 개를 던져 나오는 눈을 바탕으로 A를 첫 주사위가 3이 나오는 사건, B를 둘째 주사위가 4가 나오는 사건, C를 두 주사위의 합이 7이 나오는 사건이라 하면 각각은 pairwise independent하지만,

$$\frac{1}{36}=P(A\cap B\cap C)\neq P(A)P(B)P(C)=\frac{1}{216}$$

이다.

E.

$$P(A)=P(A\cap B)+P(A\cap B^C)$$

$$= P(A|B)P(B) + P(A|B^C)P(B^C)$$

Problem. 계절조정 관련 문제

Solution. 설명편 참고

Problem. X_1, X_2, X_3가 아래와 같은 확률밀도함수를 갖는 서로 독립인 확률변수라고 할 때, $Y = X_{(1)}$의 확률밀도함수 $g(y)$를 구하여라.

$$f(x) = \begin{cases} e^{-x} & 0 < x < \infty \\ 0 & \text{o.w.} \end{cases}$$

Solution. Y의 정의상 이는 $0 < Y < \infty$의 값을 가진다. 따라서 $0 < y < \infty$에 대하여

$$\begin{aligned} \int_0^z g(y)dy &= P(Y \le z) \\ &= 1 - P(Y > z) \\ &= 1 - P(X_1 > z, X_2 > z, X_3 > z) \\ &= 1 - \left(\int_z^\infty e^{-x} dx \right)^3 \\ &= 1 - e^{-3z} \end{aligned}$$

이며, 양변을 z로 미분하여

$$g(y) = 3e^{-3y} I(0 < y < \infty)$$

를 얻는다.

Problem. 주성분분석 관련 문제

Solution. 설명편 참고

Problem. ACF 관련 문제

Solution. 설명편 참고

Problem. A시의 인구가 3만 명, B시의 인구가 2만 명, C시의 인구가 4만명이다. 이 세 도시로부터 총 400명의 표본을 층화추출하려고 한다. 비례배분시 각 층별 표본크기를 결정하시오.

Solution. 간단한 계산을 통하여 133명, 89명, 178명을 배정해야 함을 알 수 있다.

Problem. 위 문제의 상황에서, 각 층의 표준편차가 14, 12, 20이라고 한다. 최적배분을 통하여 각 층별 표본크기를 결정하시오.

Solution. 층내비용이 동일하다고 가정하면, 최적표본수는

$$n_h^* \propto N_h S_h$$

이도록 설정한다. 즉 가중치는 각각 42, 24, 80이 되며, 이를 바탕으로 배정하면 115명, 66명, 219명이다.

Problem. 새로운 벼 품종 A_1, A_2, A_3, A_4의 비교 실험을 위하여, 논을 토양이 다를 것이라 예상되는 3개의 블록 B_1, B_2, B_3으로 층별하고 각 블록을 4개의 구획으로 나누어, 시험하려는 품종 A_1, A_2, A_3, A_4를 각 블록마다 각 구획에 랜덤하게 배치하여 쌀의 수확량을 비교하였다.

$$
\begin{aligned}
x_{ij} &= \mu + a_i + b_j + e_{ij} \\
e_{ij} &\sim N(0, \sigma_E^2) \text{이고 서로 독립} \\
b_j &\sim N(0, \sigma_B^2) \text{이고 서로 독립} \\
Cov(e_{ij}, b_j) &= 0 \\
i &= 1,2,3,4, \ j = 1,2,3
\end{aligned}
$$

$$
S_A = \sum_i \sum_j (\bar{x}_{i\cdot} - \bar{\bar{x}})^2 = 300, \ S_B = \sum_i \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2 = 300
$$

$$
S_E = \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 = 100, \ S_T = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 = 700 \text{ 이라고 하자.}
$$

인자 A가 유의한지 검정하기 위한 검정통계량을 구하고 유의수준 5퍼센트에서 유의성을 검정하시오.

Solution. 인자 A에 의한 변동의 자유도는 $4-1=3$이며, 그 변동은 300이다. 오차항에 대해서는 그 자유도가 $(4-1)(3-1)=6$이고, 그 변동은 100이다. 따라서 F통계량은

$$
F = \frac{300/3}{100/6} = 6
$$

이며, 자유도가 3, 6인 F통계량의 95퍼센타일이 4.76이므로 유의수준 5퍼센트에서 인자의 효과는 유의하다.

Problem. 위 문제에서, 변량인자 B의 산포 σ_B^2의 추정값을 구하시오.

Solution.

$$
\hat{\sigma}_B^2 = \frac{V_B - V_E}{4} = \frac{300/2 - 100/6}{4} = 33.33
$$

으로 쉽게 구할 수 있다.

Problem. $U_1, \cdots, U_n$은 $(0,1)$ 구간에서 균등분포를 갖는 서로 독립인 확률변수이고 $Y_n = \left(\prod_{i=1}^n U_i\right)^{-1/n}$이라고 할 때, $\sqrt{n}(Y_n - e)$의 극한분포를 구하려고 한다. $X_i = -\log U_i$의 평균과 분산을 구하고, 그로부터 $\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1])$의 극한분포를 구한 뒤 δ 방법을 이용하여 구하고자 $\sqrt{n}(Y_n - e)$의 극한분포가 평균이 0이고 분산이 e^2인 정규분포임을 보이시오.

Solution.

$$
\begin{aligned}
P(X_i \le x) &= P(-\log U_i \le x) \\
&= P(\log U_i \ge -x) \\
&= P(U_i \ge e^{-x}) = 1 - e^{-x}
\end{aligned}
$$

이 $x > 0$에 대해 성립한다. 따라서 X_i는 평균이 1인 지수분포를 따르며, 평균과 분산이 각각 1이다. 또한 중심극한정리에 의하여,

$$
\sqrt{n}(\bar{X}_n - 1) \xrightarrow{d} N(0,1)
$$

임을 안다. 한편

$$Y_n = \left(\prod_{i=1}^{n} U_i\right)^{-1/n} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log U_i\right) = \exp(\bar{X}_n)$$

이므로, $\exp(1) = e$임과

$$\left.\frac{d}{dx}e^x\right|_{x=1} = e$$

임을 고려하면 델타 방법에 의하여

$$\sqrt{n}(Y_n - e) \xrightarrow{d} N(0, e^2)$$

임을 쉽게 안다.

Problem. $X_1, \cdots, X_n$이 $N(\mu, \sigma^2)$을 따르는 서로 독립인 확률변수이고, μ는 미지이며, $\sigma^2 > 0$은 알려져 있다고 가정하자. μ의 완비충분통계량을 구하고, 이로써 $t \neq 0$일 때 $e^{t\mu}$의 유일한 UMVUE를 구하시오. 또한 해당 추정량의 크래머-라오 하한을 구하시오.

Solution.
확률밀도함수를 다시 쓰면

$$f(x;\mu) = \exp\left(-\mu x/\sigma^2 + \mu^2/2\sigma^2 - \log(\sqrt{2\pi}\sigma) + x^2/2\sigma^2\right)$$

형태로 지수족이고 분포의 토대 $\mathbb{R}$은 μ에 불변이며, μ의 모수공간은 $\mathbb{R}$으로 열린집합이고, $\mathrm{Var}(cx/\sigma^2) = 0$인 상수 c는 존재하지 않는다. 따라서 μ의 완비충분통계량으로

$$T = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i$$

를 사용할 수 있다. 또한 σ^2이 이미 알려진 양의 상수이므로, $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ 역시도 CSS가 된다. 한편 정규분포의 mgf로부터

$$\mathbb{E}[e^{tX}] = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

임을 알며, $\bar{X}$는 평균이 μ, 분산이 σ^2/n인 정규분포를 따르기에

$$\mathbb{E}[e^{t\bar{X}}] = e^{t\mu + \frac{\sigma^2}{2n}t^2}$$

이고, σ^2과 n은 알고 있기에

$$\mathbb{E}\left[e^{t\bar{X} - \frac{\sigma^2}{2n}t^2}\right] = e^{t\mu}$$

이다. 따라서 $e^{t\bar{X} - \frac{\sigma^2}{2n}t^2}$는 $e^{t\mu}$의 불편추정량이며, 레만-쉐페 정리에 의하여 유일한 UMVUE가 된다($\bar{X}$에 대한 조건부 기대값이 자기 자신이다.). 한편 그 크래머-라오 하한은 이 추정량이 불편추정량임에 따라 정보량 $I(\mu)$에 대하여 $(nI(\mu))^{-1}$와 동일하다.

$$\begin{aligned} I(\mu) &= \mathrm{Var}\left(\frac{\partial}{\partial\mu}\log f(x;\mu)\right) \\ &= \mathrm{Var}\left(-X/\sigma^2\right) \\ &= \frac{1}{\sigma^2} \end{aligned}$$

이기에, 하한은 σ^2/n으로 주어진다.

Problem. 검정 관련 문제

Solution. 설명편 참고

2008년

Problem. θ_1과 θ_2는 서로 독립이고, $\mathrm{Var}(\theta_1) = \sigma_1^2$, $\mathrm{Var}(\theta_2) = \sigma_2^2$을 만족한다. $\theta_3 = a\theta_1 + (1-a)\theta_2$라고 할 때, $\mathrm{Var}(\theta_3)$을 최소로 하는 상수 a값은 얼마인가?

Solution.

$$\begin{aligned}\mathrm{Var}(\theta_3) &= \mathrm{Var}(a\theta_1 + (1-a)\theta_2) \\ &= a^2\sigma_1^2 + (1-a)^2\sigma_2^2\end{aligned}$$

이므로, 양변을 a로 미분하면 일계조건이

$$2a\sigma_1^2 - 2(1-a)\sigma_2^2 = 0$$

으로 주어진다. 따라서

$$a = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

이다.

Problem. 다중공선성을 진단하는 방법으로 적절한 것은?
보기: DFFITS, White검정, 분산팽창인자, 쿡 통계량, Mallows C_p

Solution. DFFITS, 쿡 통계량, Mallows C_p는 influential point에 대해 논하며, White 검정은 이분산성에 대해 논한다. 분산팽창인자만이 다중공선성을 진단하는 방법이다.

Problem. 절편이 포함된 단순선형회귀모형에서 b를 x에 대한 y의 회귀계수, r을 x와 y의 상관계수라고 할 때 다음 중 옳지 않은 것은?

A. $b > 0$ 이면 $r > 0$ 이다.

B. $b < 0$ 이면 $r < 0$ 이다.

C. $r = 0$ 이면 $b = 0$ 이다.

D. $b = 1$ 이면 $r = 1$ 이다.

E. b'을 y에 대한 x의 회귀계수라고 할 때 $r^2 = b \cdot b'$이 성립한다.

Solution.

$$b = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \sqrt{\frac{S_{yy}}{S_{xx}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

이므로 b, r의 부호는 같다. 따라서 A, B는 참이다. 동일하게 C 역시도 참이다. 한편

$$r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = \frac{S_{xy}}{S_{xx}} \times \frac{S_{xy}}{S_{yy}} = b \cdot b'$$

도 성립한다. 그러나 D는 일반적으로 참이 아니다.

Problem. 다음은 표본설계방법에 대한 설명이다. 옳은 것은?

A. 층화임의표집은 층간분산이 작을수록 유리하다.

B. 집락표집에서는 집락내는 이질적, 집락간은 동질적이어야 집락표집의 효과가 있다.

C. 층화임의표집에 대한 네이만 배분법을 따르면, 층내분산이 작은 층에서 표본을 더 많이 선택해야 한다.

D. 순서모집단에 계통표집을 적용하는 것은 바람직하지 못하다.

E. 비용측면에서 표본조사계획을 고려한다면 층화임의표집이 집락표집보다 적절하다.

Solution.

A. 아니다. 층화임의표집은 층간분산이 클수록 유리하다.

B. 그렇다.

C. 아니다. 층내분산이 큰 층에서 표본을 더 많이 선택해야 한다.

D. 아니다. 계통표집은 순서모집단보다는 주기성을 가지는 모집단에 적용될 때 약점을 가진다.

E. 아니다. 일반적으로 집락표집이 층화표집보다 더 적은 비용이 필요하다.

Problem.

☐ 카이제곱분포에 대한 다음 설명 중 옳지 <u>않는</u> 것은 ?

A. 서로 독립인 확률변수 $X_i (i=1,2,....,n)$ 들이 각각 자유도 k_i 인 카이제곱분포를 따르면 $Y=\sum_{i=1}^{n} X_i$ 는 자유도가 $\sum_{i=1}^{n} k_i$ 인 카이제곱분포를 따른다.

B. 카이제곱분포는 감마분포의 특수한 경우이다.

C. 확률변수 Z가 $N(0,1)$을 따를 때, $X=Z^2$은 $\chi^2(1)$ 분포를 가진다.

D. $X \sim \chi^2(n)$ 이면, X의 적률생성함수 $M_X(t)=(1-2t)^{-n/2}$ 이다. 단, $t<\frac{1}{2}$

E. $X \sim \chi^2(n)$ 이면, $E(X)=n$, $Var(X)=n^2$ 이다.

Solution. 나머지는 다 맞는 설명이지만, E는 틀렸다. $\mathrm{Var}(X) = 2n$이다.

Problem. 라오-크래머의 정보부등식에 대해 약술하시오.

Solution.

$$X_i \sim_{i.i.d.} f(x;\theta),\ \theta \in \Theta \subseteq \mathbb{R}^k,\ i = 1, 2, \cdots, n$$

이며 $\eta = \eta(\theta) \in \mathbb{R}^d$일 때, 적절한 조건 하에서 η의 추정량 $\hat{\eta} = \hat{\eta}(X_1, \cdots, X_n)$에 대해 아래가 성립함을 의미한다.

$$\mathrm{Var}(\hat{\eta}_n) \succeq \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\eta}_n]\right)^T [nI(\theta)]^{-1} \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\hat{\eta}_n]\right)$$

Problem. 능형회귀에 대해 약술하시오.

Solution. 선형회귀의 진행에 있어 X가 full column rank를 가지지 않으면 $\hat{\beta}$가 identifiable하지 않다. 따라서 적절한 수준의 양수 $\lambda > 0$에 대하여, X^TX에 λI만큼을 더해 이를 invertible하게 만들어 능형회귀추정량

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I)^{-1}X^TY$$

를 얻을 수 있다. 이는 회귀계수의 $l2$-norm을 벌점항으로 넣는 벌점화회귀로 취급할 수도 있다.

Problem. 시계열 문제

Solution. 설명편 참고

Problem. $X_1, \cdots, X_n$은 $f(x) = e^{-(x-\theta)},\ x \geq \theta$를 따르는 확률분포로부터 추출한 랜덤표본이다. $X_{(1)}$이 θ의 충분통계량임을 보이고, $X_{(1)} + a$가 θ의 불편추정량이 되기 위한 a의 값을 구하여라.

Solution. 확률밀도함수가

$$f(x_1, \cdots, x_n;\theta) = \prod_{i=1}^{n} e^{-(x_i-\theta)}I(x_i \geq \theta) = e^{-\sum_{i=1}^n x_i + n\theta}I(\min\{x_i\} \geq \theta) = e^{-\sum_{i=1}^n x_i} \times e^{n\theta}I(\min\{x_i\} \geq \theta)$$

으로, 분해정리에 의하여 $X_{(1)}$이 θ의 충분통계량이 된다. 한편 $Y = X_{(1)}$의 누적분포함수는 $y \geq \theta$에 대하여

$$\begin{aligned} P(Y \leq y) &= 1 - P(Y > y) \\ &= 1 - P(X_1 > y, \cdots, X_n > y) \\ &= 1 - (e^{-(y-\theta)})^n \\ &= 1 - e^{-n(y-\theta)} \end{aligned}$$

으로 주어지므로, $Y - \theta$는 평균이 $\dfrac{1}{n}$인 지수분포를 따른다. 따라서 $a = \dfrac{1}{n}$이면 $X_{(1)} + n^{-1}$이 θ의 불편추정량이 된다.

Problem. $X_1, \cdots, X_n$은 다음의 분포를 따르는 독립적 확률변수이다.

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1, \theta > 0 \\ 0 & \text{o.w.} \end{cases}$$

θ에 대한 최대우도추정량을 구하시오.

Solution. 가능도함수는

$$L(\theta; x) = \prod_{i=1}^{n} \theta x_i^{\theta-1} I(\theta > 0)$$

이며, 로그가능도함수는

$$l(\theta; x) = (\theta - 1) \sum_{i=1}^{n} \log(x_i) + n \log \theta - \delta(\theta \leq 0)$$

으로 나타난다. 따라서 일계조건은

$$\frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0$$

으로 나타나며, 이로부터 최대우도추정량은

$$\hat{\theta} = \left(-\frac{1}{n} \sum_{i=1}^{n} \log(X_i) \right)^{-1}$$

임을 얻는다.

Problem. $Y = -\sum_{i=1}^{n} \ln X_i$라고 할 때, 기대값 $\mathbb{E}[Y]$를 구하시오.

Solution.

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[-\sum_{i=1}^{n} \ln X_i \right] \\
&= -n\, \mathbb{E}[\ln X_1] \\
&= -n \int_0^1 \ln x \times \theta x^{\theta-1} dx \\
&= -n[\ln x \times x^{\theta}]_0^1 + \int_0^1 n x^{\theta-1} dx \\
&= \frac{n}{\theta}
\end{aligned}$$

Problem. 위 문제에서 귀무가설과 대립가설이 아래와 같이 주어졌다.

$$H_0 : \theta = 1 \text{ v.s. } H_1 : \theta = 2$$

유의수준이 α인 최강력검정의 기각역을 제시하시오.

Solution.

$$\frac{f(x; 2)}{f(x; 1)} = 2^n \prod_{i=1}^{n} x_i$$

이므로, 원하는 최강력검정의 기각역은

$$X_1 X_2 \cdots X_n > k$$

형태로 주어진다. 이제 유의수준을 α로 하는 k를 찾으면 된다. 양변에 로그를 씌운 뒤 -1을 곱하면, 기각역은

$Y < -\ln(k)$처럼 쓸 수도 있다. 그런데 $-\ln(X_i)$의 경우 $a > 0$에 대하여

$$P(-\ln(X_i) \le a) = P(X_i \ge e^{-a}) = \int_{e^{-a}}^{1} \theta x^{\theta-1} = 1 - e^{-a\theta}$$

이므로 $-\ln(X_i)$는 평균이 θ^{-1}인 지수분포를 따르며, 그 합인 Y는 모수가 n, θ^{-1}인 감마분포를 따른다. 따라서

$$-\ln k = \Gamma_{0.05}(n, \theta^{-1})$$

으로 결정되며, 이에 따라 기각역은

$$X_1 X_2 \cdots X_n > \exp(-\Gamma_{0.05}(n, \theta^{-1}))$$

으로 주어진다.

Problem. 세 종류의 산업 A, B, C에 종사하는 국내 기업인들을 대상으로 향후 국내경기전망에 대한 설문을 실시하여 다음과 같은 결과를 얻었다.

산업 \ 전망	긍 정	부 정	계
A	10	40	50
B	30	20	50
C	20	30	50
계	60	90	150

산업에 따라 기업인들의 향후 경기전망에 대한 판단이 다른지를 검정하기 위한 귀무가설과 대립가설을 설정하고, 유의수준 5퍼센트에서 검정하시오.

Solution. 동질성 검정을 수행할 수 있다. 산업 A, B, C의 긍정평가비율을 각각 p_A, p_B, p_C,라고 한다면, 귀무가설과 대립가설은

$$H_0 : p_A = p_B = p_C, \quad \text{v.s.} \quad H_1 : \text{not } H_0$$

으로 주어지고 카이제곱 검정통계량은

$$\frac{(10-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(40-30)^2}{30} + \frac{(20-30)^2}{30} + \frac{(30-30)^2}{30} = 16.667$$

으로, $\chi^2(2)$를 따르는 분포에서의 95퍼센트 퍼센타일보다 훨씬 크다. 따라서 귀무가설을 기각할 수 있다.

Problem. 위 문제에서, 각 산업별로 향후 경기를 긍정적으로 전망한 비율의 95퍼센트 신뢰구간을 구하시오.

Solution. 단순히 정규근사 플러그인 신뢰구간을 구하면 될 듯하다. A의 경우 $0.2 \pm 1.96 \times \sqrt{0.2 \times 0.8/50} = (0.09, 0.31)$이며, B의 경우 $0.6 \pm 1.96 \times \sqrt{0.6 \times 0.4/50} = (0.46, 0.74)$을 얻는다. C의 경우 동일한 방식으로 $(0.26, 0.54)$를 얻는다.

Problem. 시계열 문제

Solution. 설명편 참고

2007년

Problem.

□ 다음은 여러 분포에 대한 설명이다. 옳지 **않은** 것은 ?

A. T가 자유도 n인 t분포를 가졌으면 T^2은 $F(n,1)$분포를 갖는다.

B. 포아송(λ) 분포로부터 랜덤표본 $X_1, X_2, \cdots, X_n$을 얻었을 때 표본의 크기 n이 커짐에 따라 표본평균 $\overline{X_n}$는 모평균 λ로 수렴한다.

C. $F \sim F(n,m)$ 이면 $\frac{1}{F} \sim F(m,n)$을 따른다.

D. $X \sim \chi^2(n)$ 이면 $E(X)=n$ 이고 $Var(X)=2n$ 이다.

E. 서로 독립인 확률변수 $X_i (i=1,2,....,n)$들이 각각 자유도 k_i인 카이제곱분포를 따르면 $Y=\sum_{i=1}^{n} X_i$는 자유도가 $\sum_{i=1}^{n} k_i$인 카이제곱분포를 따른다.

Solution. A가 거짓이다. T^2은 $F(1,n)$ 분포를 갖는다.

Problem. 두 확률변수 X와 Y의 결합확률밀도함수 $f(x,y)$가 다음과 같이 주어질 때, $\mathbb{E}[XY]$의 값은?

Y \ X	0	1	2	3
0	1/24	3/24	2/24	1/24
1	3/24	8/24	2/24	0
2	1/24	2/24	1/24	0

Solution. 직접 계산해보면,

$$\mathbb{E}[XY] = \frac{8+4+4+4}{24} = \frac{5}{6}$$

을 얻는다.

Problem. 다음 모형 중 선형으로 변환할 수 없는 것은?

A. $Y = \beta_0 X^{-\beta_1} \epsilon^{\epsilon}$

B. $Y = e^{\beta_0 + \beta_1 X + \epsilon}$

C. $Y = \frac{e^{\beta_0 + \beta_1 X + \epsilon}}{1 + e^{\beta_0 + \beta_1 X + \epsilon}}$

D. $Y = \beta_0 + \beta_1 \frac{1}{X} + \epsilon$

E. $Y = \beta_0 \beta_1^X + \epsilon$

Solution.

A. $\ln Y = \ln \beta_0 - \beta_1 \ln X + \epsilon$으로 선형변환된다.

B. $\ln Y = \beta_0 + \beta_1 X + \epsilon$으로 선형변환된다.

C. $\ln \frac{Y}{1-Y} = \beta_0 + \beta_1 X + \epsilon$으로 션형변환된다.

D. 이미 선형 형태이다.

E. 안된다.

Problem.

□ 다음은 표본조사에서의 오차에 대한 설명이다. 옳지 **않은** 것은 ?

A. 표본오차는 표본의 크기가 커짐에 따라 일반적으로 증가한다.

B. 비표본오차는 전수조사와 표본조사에서 모두 발생한다.

C. 무응답오차는 비표본오차의 일종이다.

D. 비표본오차는 일반적으로 표본조사보다 전수조사에서 더 크게 발생한다.

E. 전수조사에서는 표본오차가 발생하지 않는다.

Solution. A. 표본오차는 표본의 크기가 커짐에 따라 일반적으로 감소한다.

Problem. 오차항의 이분산성과 자기상관성 존재 여부를 검정하는 방법을 순서대로 나열하시오.

Solution. White 검정, Durbin-h 검정.

Problem. $f(x) = k\frac{1}{\sqrt{x(1-x)}}$, $0 < x < 1$가 확률밀도함수가 되기 위한 k값은?

Solution.

$$\begin{aligned}
\int_0^1 \frac{1}{\sqrt{x(1-x)}} dx &= \int_0^1 \frac{1}{\sqrt{-(x-0.5)^2 + 0.25}} dx \\
&= \int_{-0.5}^{0.5} \frac{1}{\sqrt{0.25 - v^2}} dv \quad (v = x - 0.5) \\
&= \int_{-\pi/2}^{\pi/2} \frac{1}{\sqrt{0.25 - 0.25\sin^2\theta}} \times 0.5\cos\theta d\theta \quad (v = 0.5\sin\theta)
\end{aligned}$$

$$= \pi$$

이므로, $k = \frac{1}{\pi}$여야 한다.

Problem. 시계열 관련 문제 두 개

Solution. 설명편 참고

Problem. 정규성 검정 방법들을 쓰시오.

Solution. 대표적으로 자크-베라 검정, 콜모고로프-스미르노프 검정이 있다. 이외에도 샤피로-윌크 검정, 앤더슨-달링 검정, D'Agostino's K-squared test 등이 있다.

Problem. 가우스-마코프 정리를 간략하게 설명하고 이를 증명하시오.

Solution. 선형모형

$$y = X\beta + \epsilon$$

에서, 이 선형모형이 참이며, ϵ이 등분산성과 독립성을 만족한다면, 최소제곱추정량 $\hat{\beta}$가 β의 가장 효율적인 선형불편추정량, 즉 BLUE라는 정리이다. 증명은 앞서서 수행하였으므로 생략한다.

Problem. 희망초등학교 6학년 학생 200명 중 남자는 112명, 여자는 88명이다. 이중 40명을 대상으로 1년 동안 읽은 책의 평균권수 μ를 조사하는데 층화표집을 사용하여 남자는 22명, 여자는 18명을 추출하였다. 조사결과 각 층에 대한 자료는 다음과 같이 구해졌다.

층	N_h	n_h	$\overline{y_h}$	s_h^2
남학생	112	22	30	60
여학생	88	18	45	40
계	200	40		

6학년 학생 전체에 대하여 1년 동안 읽은 책의 평균 권수의 추정량을 구하고, 그 분산을 구함으로써 95퍼센트 신뢰구간을 구하시오.

Solution.

$$\bar{y}_{st} = \frac{112}{200} \times 30 + \frac{88}{200} \times 45 = 36.6$$

이다. 분산의 추정량은

$$\hat{V}(\bar{y}_{st}) = \frac{112^2}{22 \times 200^2}(1 - 22/112) \times 60 + \frac{88^2}{18 \times 200^2}(1 - 18/88) \times 40 = 1.029$$

로 나타난다. 이를 통한 95퍼센트 신뢰구간은 $36.6 \pm 1.96 \times 1.029 = (34.6, 38.6)$이다.

Problem. $Y_1, \cdots, Y_n$이 구간 $[0, \theta]$에서 균등분포를 따르는 랜덤표본이라고 하자 ($\theta > 0$). 다음의 가설 $H_0 : \theta = 1$ v.s. $H_1 : \theta \neq 1$에 대해 기각역은 다음과 같다.

$$RR = \{Y_{(n)} \leq k \text{ or } Y_{(1)} > 1\}$$

α가 유의수준을 나타낼 때 k를 α와 n으로 나타내시오.

Solution.

$$\begin{aligned}\alpha &= \mathbb{E}_1[I(RR)] \\ &= \mathbb{E}_1[I(Y_{(n)} \le k)] \\ &= P_1(Y_{(n)} \le k) \\ &= P_1(Y_1 \le k, \cdots, Y_n \le k) \\ &= k^n\end{aligned}$$

이므로, $k = \alpha^{1/n}$이다.

Problem. 위 검정의 검정력함수를 구하고, $n = 2, \alpha = 0.05$일 때 $\theta = 0.5$에서의 검정력을 구하여라. 또한 $\alpha = 0.05$일 때 $\theta = 2$에서의 검정력이 0.9 이상이 되기 위해서 필요한 최소한의 표본크기를 구하여라.

Solution. 먼저 $\theta < 1$이라면,

$$\begin{aligned}\gamma(\theta) &= \mathbb{E}_\theta[I(RR)] \\ &= \mathbb{E}_\theta[I(Y_{(n)} \le \alpha^{1/n})] \\ &= (\alpha^{1/n} \wedge \theta)^n / \theta^n \\ &= 1 \wedge \frac{\alpha}{\theta^n}\end{aligned}$$

이 검정력함수이다. 한편 $\theta > 1$이라면,

$$\begin{aligned}\gamma(\theta) &= \mathbb{E}_\theta[I(RR)] \\ &= P_\theta(Y_{(n)} \le \alpha^{1/n}) + P_\theta(Y_{(n)} > 1) \\ &= \frac{\alpha}{\theta^n} + 1 - \frac{1}{\theta^n} \\ &= 1 - \frac{(1-\alpha)}{\theta^n}\end{aligned}$$

이다. 따라서 검정력함수는

$$g(\theta) = \begin{cases} 1 & \theta < \alpha^{1/n} \\ \dfrac{\alpha}{\theta^n} & \alpha^{1/n} \le \theta < 1 \\ 1 - \dfrac{(1-\alpha)}{\theta^n} & 1 < \theta \end{cases}$$

으로 주어진다. 주어진 값들을 대입하면 $\alpha^{1/n} = 0.224$이고 θ는 이보다 크면서 1보다는 작으르모, 검정력은 $0.05/0.5^2 = 0.2$로 주어진다. 한편 검정력이 0.9 이상이 되기 위해서는,

$$1 - \frac{0.95}{2^n} \ge 0.9$$

여야 한다. 따라서 $2^n \ge 0.95/0.1 = 9.5$이기에, $n \ge 4$여야 한다.

Problem. 시계열 자기상관 및 조건부분산 문제

Solution. 설명편 참고

Problem. 500만원을 가진 주식투자자 홍길동 씨는 A, B, C의 세가지 주식 중 하나를 매입하여 1년 후에 매각할 계획을 갖고 있다. 각 주식의 현재 시장 가격은 10000원으로 동일하다. 한편 1년 후 주가의 시나리오는 아래 세 가지가 있다.

	시나리오1	시나리오2	시나리오3
A주식	12,000원	9,000원	6,000원
B주식	7,000원	15,000원	9,000원
C주식	10,000원	11,000원	8,000원

maximax, maximin, minimax 기준에 의하여 홍길동씨는 각각 어느 주식을 매입하게 되는가?

Solution. maximax 기준을 따르는 경우, 최대한의 이득을 얻을 가능성이 있는 B주식을 매입할 것이다. maximin 기준을 따르는 경우, 최소액이 가장 큰 C주식을 택할 것이다. minimax 기준을 이용하는 경우, maximum risk를 최소화하는 C주식을 택한다.

Problem. 시계열 분해 문제

Solution. 설명편 참고