Notes

2020년 8월 5일 수요일 오후 5:03

제 2장 - 자료의 기술 , PDF page 19

자료의 종류

- 질적 데이터 Qualitative Data
  - 명목형 nominal
  - 순서형 Ordinal
- 양적데이터 Quantitative Data
  - 이산형 discrete
  - 연속형 Continuous

평균 (mean, average)

- 모평균 (Population mean): $\frac{1}{N}\sum_{i=0}^{N} X_i = \mu$
- 표본평균 (Sample mean): $\frac{1}{n}\sum_{i=0}^{n} X_i = \bar{X}$

중앙값 median

- if n is an odd number, $\tilde{x} = X_{(\frac{n+1}{2})}$
- if n is an odd number, $\tilde{x} = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2})}}{2}$

분산 (Variance): 자료의 변동의 평균

- 모분산 Population Variance: $\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sigma^2$
- 표본분산 Sample Variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2$

표준편차 Standard Variation

- 모표준편차 Population Standard Deviation $= \sqrt{\sigma^2} = \sigma$
- 표본표준편차 Sample Standard Deviation $= \sqrt{S^2} = S$

자유도 Degrees of Freedom

: 변동을 계산하는데 이용되는 독립된 정보의 수

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2 = \frac{1}{N}\left\{\sum_{i=0}^{N}x_i^2 - \frac{\left(\sum_{i=0}^{N}x_i\right)^2}{N}\right\}$$

proof.

$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2$, let $P = \sum_{i=0}^{N}(x_i-\mu)^2$

$$\begin{aligned} P &= \sum_{i=0}^{N}(x_i-\mu)^2 \\ &= \sum_{i=0}^{N} x_i^2 - 2x_i\mu + \mu^2 \\ &= \sum_{i=0}^{N}x_i^2 - 2\mu\sum_{i=0}^{N}x_i + \sum_{i=0}^{N}\mu^2 , \quad \sum_{i=0}^{N}x_i = N\mu, \ \sum_{i=0}^{N}\mu^2 = N\mu^2 \\ &= \sum_{i=0}^{N}x_i^2 - 2N\mu^2 + N\mu^2 \\ &= \sum_{i=0}^{N}x_i^2 - N\mu^2 , \quad \mu = \frac{1}{N}\sum_{i=0}^{N}x_i \\ &= \sum_{i=0}^{N}x_i^2 - N\left(\frac{1}{N^2}\right)\left(\sum_{i=0}^{N}x_i\right)^2 \\ &= \sum_{i=0}^{N}x_i^2 - \frac{\left(\sum_{i=0}^{N}x_i\right)^2}{N} = P \end{aligned}$$

$$\therefore \sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2 = \frac{1}{N}\left\{\sum_{i=0}^{N}x_i^2 - \frac{\left(\sum_{i=0}^{N}x_i\right)^2}{N}\right\}$$

변동계수 Coefficient of Variation

- 서로 다른 평균과 표준편차를 가진 자료들을 비교할 수 있는 기준 . Ex) 단위가 다른 자료들

$CV = \frac{S}{\bar{x}} \times 100$, in percentage

왜도 Skewness 와 첨도 Kurtosis - PDF Page 32

왜도 Skewness - 자료의 대칭 정도를 나타내는 측도

- 왜도 계수 : $S_k = \frac{E[X-\mu]^3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left\{\frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2\right\}^{\frac{3}{2}}} = \frac{N}{(N-1)(N-2)}\frac{\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3}$

첨도 Kurtosis - 자료의 중앙 집중도를 나타내는 측도

- $K_u = \frac{E[X-\mu]^4}{\sigma^4} - 3 = \frac{\frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^4}{\left\{\frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2\right\}^2} - 3 = \frac{N(N+1)}{(N-1)(N-2)(N-3)}\frac{\sum_{i=0}^{N}(x_i-\mu)^3}{\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$

스터지의 규칙 Sturge's law

- 도수분포표를 그릴 때 필요한 구간의 수 k를 구하는 법
  - $k \geq \log_2 n$, $n = \#$ of obs.

제 3장 - 확률 , PDF page 47

표본공간 Sample Space : 어떤 실험에서 발생가능한 모든 단일사상들의 집합

확률의 공리 Axioms of Probability

정의 3.6

임의의 사상 A 에 대하여

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. 서로 배반인 사상열 A_i , $i = 1, 2, \ldots$ 에 대하여

   $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$

배반 사상 - 동시에 발생할 수 없는 사상

상호 배반 Mutually Exclusive - 어떠한 임의의 두 개의 사상도 교집합이 ∅ 일 때.

3.4 사상의 독립

정리 3.2 : 사상 A, B 가 독립이기 위한 필요충분조건

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(AB) = P(A)P(B)$

" 독립이면 상호배반이 아니다. " ⟷ " 상호배반이면 독립이 아니다.

정리 3.3 - 여러 사상들의 상호 독립

n 개의 사상들이 상호 독립이기 위해서,

$P(A_1 \cap A_2 \cap \cdots A_n) = P(A_1)\cdot P(A_2)\cdots P(A_n) = P\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} P(A_i)$

정리 3.4 확률의 합법칙

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

정리 3.5 - 서로 배반사상이라면,

$P(A \cup B) = P(A) + P(B)$, $A \cap B = \emptyset$

정리 3.6 확률의 곱법칙

$P(A \cap B) = P(B)P(A|B)$ or $P(A)P(B|A)$

만약 A와 B가 독립이라면,

$P(A \cap B) = P(A)P(B)$

정의 3.8 - 분할 Partition

사상 $B_1, B_2, \ldots, B_n$ 들이 표본공간 S에 대하여 다음의 조건을 만족한다면, S에 대한 분할 이라고 한다.

① $B_i \cap B_j \neq \emptyset$ for all i and j , $i \neq j$

② $\bigcup_{i=1}^{n} B_i = S$

정리 3.8 - 전확률의 정리 Theorem of Total Probabilities

상호 배반 사상인 B_n 이 있을 때, S 에 속하는 임의의 사상 A에 대해 다음이 성립한다.

$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$

정리 3.9 - 베이즈 정리

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

제 4장 - 확률변수 - PDF page 83

정의 4.1 - 확률변수

- 표본공간 내에 있는 각 원소를 하나의 실수값에 대응시키는 함수

정의 4.2 - 이산형 표본공간과 연속형 표본공간

discrete sample space & Continuous sample space

확률변수 X가 이산형 이라면, X의 확률분포는 확률질량 함수 Probability Mass Function 이라고 하고, 연속형이라면 확률 밀도 함수 Probability Density function 이라고 한다.

정의 4.4 - 확률 질량함수 - 다음 조건을 만족하는 이산형 확률변수 X의 확률질량함수 라고 한다.

① $p(x_i) > 0$, $i = 1, 2, \ldots$

② $p(x) = 0$, $x \neq x_i$, $i = 1, 2, \ldots$

③ $\sum p(x_i) = 1$

정의 4.5 - 누적분포함수

$F(x) = P(X \leq x) = \sum_{i: x_i \leq x} p(x_i)$, $-\infty < x < \infty$

정리 4.1 - 누적분포함수의 특성

① $F(-\infty) = 0$, $F(\infty) = 1$

② $F(x)$ 는 단조함수 이면서 비감소 함수이다.

③ $F(x)$ 는 우측으로 부터

정리 4.4

X가 이산형 확률분포 $p(x)$를 가지는 이산형 확률변수 또는 X가 연속형 확률분포 $f(x)$를 가지는 연속형 확률변수라고 할 때, 확률변수 $g(X)$

$$\mu_{g(x)} = E(g(x)) = \begin{cases} \sum_x g(x)p(x) \\ \int_{-\infty}^{\infty} g(x)f(x)\,dx \end{cases}$$

구간 선정에 신경쓰자 !!

정의 4.10

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \begin{cases} \sum_x (x-\mu)^2 p(x) \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \end{cases}$$

정리 4.5 - 분산

$Var(X) = E(X^2) - [E(X)]^2$

정리 4.6

$$Var[g(x)] = E[g(x) - \mu_{g(x)}]^2 = \begin{cases} \sum_x \{g(x) - E(g(x))\}^2 p(x) \\ \int_{-\infty}^{\infty} \{g(x) - E(g(x))\}^2 f(x)\,dx \end{cases}$$

정리 4.7 - 마코프 부등식 (Markov Inequality)

X는 확률변수이고, 임의의 함수 $U(x) \geq 0$ 와 임의의 상수 $c > 0$ 에 대해 다음의 부등식이 성립한다.

$$P(U(x) \geq c) \leq \frac{E[U(x)]}{c}$$

정리 4.8 - 체비셰프 부등식(Chebyshev' s Inequality)

확률변수 X의 평균과 분산이 각각 μ 와 σ^2 을 지닐 때, $k > 0$ 인 상수에 대해 다음의 부등식이 성립한다.

$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

$P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$

$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4}$

정리 4.9

임의의 상수 a, b 에 대하여

① $E(aX + b) = aE(X) + b$ 또는 $\mu_{aX+b} = a\mu_X + b$

② $Var(aX + b) = a^2 Var(X)$ 또는 $\sigma^2_{aX+b} = a^2\sigma^2_X$

③ $\sigma_{aX+b} = |a|\sigma_X$, $\sigma_{aX} = |a|\sigma_X$, $\sigma_{X+b} = \sigma_X$

정의 4.11 - 적률 Moments

X 가 기대값이 존재하는 확률변수라면 X의 r 차 적률(Moments)은 μ_r' 라고 표현하고 다음과 같이 구한다.

$\mu_r' = E(X^r)$

정의 4.12 - 중심적률 Central Moments

X 가 확률변수라면 a 에 대한 X 의 r 차 중심적률 (Central Moments)은 $E[(X-a)^r]$ 로 정의된다.

만약 $a = \mu_X$ 라면 μ_X 에 대한 X의 r차 중심적률은 μ_r 이라고 표현하고 다음과 같이 구한다.

$\mu_r = E[(X - \mu_X)^r]$

- X의 평균은 1차 적률, 분산은 2차 중심적률이다.
- 확률분포가 대칭이라면 X의 홀수차 중심적률은 항상 '0' 이다.

정의 4.15 계승 적률

확률변수 X에 대해 X의 r차 계승적률 (factorial moment) 은 다음과 같이 정의된다.

$E[X(X-1)\cdots(X-r+1)]$

정의 4.16 - 적률생성함수 - moment generating function

임의의 구간 $-h < t < h$, $h > 0$ 에서 모든 t에 대해 기대값이 존재한다면 e^{tx} 의 기대값을 X의 적률생성함수 라고 한다.

$$m(t) = E[e^{tx}] = \begin{cases} \sum_x e^{tx} p(x) \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \end{cases}$$

$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{\infty} x^r e^{xt} f_X(x)\,dx$

$\lim_{t \to 0}$, $\frac{d^r}{dt^r} m(0) = E[X^r] = \mu_r'$

정의 4.17 - 계승적률생성 함수

확률변수 X의 기대값이 존재한다면 $E(t^X)$를 계승적률생성함수 - factorial moment generating function 이라고 한다.

제 5 장 확률분포 , PDF 129

PDF 129

정의 5.1 - 이산형 균일분포

$P(x;a,b) = \frac{1}{b-a+1}$, $x = a, \dots, b$

정리 5.1 - 이산형 균일분포를 따르는 확률변수의 평균, 분산, 적률생성함수

$E(X) = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)(b-a+2)}{12}$,

$m(t) = E(e^{tX}) = \sum_{j=a}^{b} e^{jt} \frac{1}{b-a+1}$

$* \sum_{x=a}^{b} x = \frac{(b-a+1)(a+b)}{2}$

PDF 131

베르누이 실행 : 실험의 결과가 성공, 실패 두가지 결과만을 가지는 시행.

베르누이 분포 Bernoulli Distribution

$p(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$

정리 3.2 - 베르누이 분포를 따르는 확률변수의평균,분산, 적률생성함수

$E(X) = p$, $Var(X) = p(1-p)$, $m(t) = pe^t + q$

이항분포 - 한 번의 시행보다는 여러번의 시행 중에서 몇 번 성공할 것인가에 더 관심을 갖는 경우

조건 :
① 결과는 상호 배타적인 두 사건으로 구분.
② 매 시행마다 일정한 성공 확률
③ 각 시행은 서로 독립적

정의 5.3 - 이항분포

성공확률이 p 이고, 실패 확률이 $1-p$ 인 실험을 n 번 실행

$p(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 1, 2, 3, \dots, n$

정리 5.3 - 이항분포를 따르는 확률변수 x의 평균, 분산, 적률생성함수

$E(X) = np$, $Var(X) = np(1-p)$, $m(t) = \{(1-p) + pe^t\}^n$

이항분포의 특징
① $p = 0.5$ 이면 이항분포는 좌우대칭
② $p < 0.5$ 이면 skewed to the right
③ $p > 0.5$ 이면 skewed to the left
④ p 의 값에 상관없이 n 이 커질수록 정규곡선에 접근한다
* n 이 클 경우 정규분포를 이용하여 근사적으로 구할 수 있다.

PDF 137

다항분포 - Multinomial Distribution

- 베르누이 시행의 확장. 가능한 결과가 셋 이상

다항계수 - Multinomial Coefficient

$\binom{n}{x_1, x_2, \dots, x_k} = \binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-x_2-\cdots-x_{k-2}}{x_{k-1}} = \frac{n!}{x_1! x_2! \cdots x_k!}$

다항분포의 확률밀도함수, 평균, 분산

$p(x_1, x_2, \dots, x_k; n, p_1, \dots, p_k) = \binom{n}{x_1, \dots, x_k} p_1^{x_1} \cdots p_k^{x_k}$, $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$

$E(X_1) = np_1, \dots, E(X_k) = np_k$

$Var(X_1) = np_1(1-p_1), \dots, Var(X_k) = np_k(1-p_k)$

PDF 132

초기하분포 Hypergeometric Distribution

성공과 실패로 나뉘고 n 번 시행을 비복원 추출

$p(x; N, M, n) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

(actual success, success case, failure case, size, sample case, size)

초기하분포의 평균 분산

$E(X) = n \cdot \frac{M}{N}$, $Var(X) = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}$

모집단의 크기가 클 수록 이항분포와 닮아간다.

PDF 145

음이항 분포 Negative Binomial Distribution r번 성공할 때까지 총 실패횟수

조건 :
① 성공과 실패만으로 구성된 베르누이 시행이 독립적으로 반복된다
② 각 시행의 성공확률은 P로 동일하다.
③ 성공이 r 번 발생할 때까지 시행을 계속한다.

$p(x;p) = \binom{r+x-1}{x} p^r (1-p)^x = \binom{-r}{x} p^r (1-p)^x$, $x = 0, 1, 2, \dots$

$E(X) = \frac{r(1-p)}{p}$

$Var(X) = \frac{r(1-p)}{p^2}$

$m(t) = \left[\frac{p}{1-(1-p)e^t}\right]^r$

기하분포 Geometric Distribution 1번 성공할 때까지 총 실패 횟수

$P(x;p) = p(1-p)^x$

$E(X) = \frac{1-p}{p}$

$Var(X) = \frac{1-p}{p^2}$

$m(t) = \frac{p}{1-(1-p)e^t}$

정리 5.8

- 확률변수 X가 모수가 p인 기하분포를 따르면 다음이 성립한다.

$p(X \geq i+j \mid X \geq i) = P(X \geq j)$

PDF 151

포아송 분포 Poisson Distribution 단위시간이나 단위면적, 또는 단위공간 에서 발생한 사건의 수

λ는 단위시간 또는 단위공간에서 발생하는 사건의 평균

* 포아송분포는 e^x의 extension (교재 p142 설명 참조)

n 이 적당히 크고, np 가 적당히 작으면 $b(n,p) \approx pois(np)$ 로 근사할 수 있다.
$(n \to \infty)$ $(p \to 0)$

$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, 3, \dots$

$E(X) = \lambda$, $Var(X) = \lambda$, $m(t) = e^{\lambda(e^t-1)}$

PDF 156

연속형 균일분포

$f(x;a,b) = \frac{1}{b-a}$

$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases}$

$E(X) = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$, $m(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$

PDF 158

정의 5.10 - 정규분포

평균 μ와 표준편차 σ를 가지는 정규확률변수 X의 확률밀도함수는 다음과 같다.

$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

정규분포 $N(\mu, \sigma^2)$을 따르는 확률변수 X의 평균, 분산, 적률생성함수

$E(X) = \mu$, $Var(X) = \sigma^2$, $m(t) = e^{\mu t + \sigma^2 t^2/2}$

정의 5.11 - 표준 정규분포

평균이 0 이고, 분산이 1

$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} = \phi(z)$

$F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt$

$* \ P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$

$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$

제 6장 - 이변량 확률변수 PDF 189

정의 6.1- 결합확률질량함수

X, Y가 동일한 표본공간 내에서 정의된 이산형 확률변수라고 한다.

$p(x,y) = P(X=x, Y=y)$

X, Y 각각 $a \leq x \leq b$, $c \leq Y \leq d$ 라고 할 때, 결합확률 $P[a \leq X \leq b, c \leq Y \leq d]$은

$P[a \leq X \leq b \ \ c \leq Y \leq d] = \sum_{x=a}^{b}\sum_{y=c}^{d} p(x,y)$

정의 6.2- 주변확률질량함수 Marginal Probability Mass Function

$P(X=x) = \sum_{y} p(x,y)$, $P(Y=y) = \sum_{x} p(x,y)$

정의 6.3 - 결합확률밀도함수 Joint Probability Density Function

X, Y 가 연속형일 때 결합확률밀도함수는 다음을 만족한다
① $f(x,y) \geq 0$
② $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

정의 6.4 - 주변확률밀도 함수

정의 6.5. 결합누적분포함수

$F(x,y) = \sum_{w_1 \leq x}\sum_{w_2 \leq y} f(w_1, w_2)$

$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(w_1, w_2)\,dw_2\,dw_1$

PDF 196

정리 6.1 - 결합누적분포함수 F(x, y) 는 다음이 성립한다.

① 모든 y에 대해서 $F(-\infty, y) = \lim_{x \to -\infty} F(x,y) = 0$,
모든 x에 대해서 $F(x, -\infty) = \lim_{y \to -\infty} F(x,y) = 0$,
$\lim_{x \to \infty, y \to \infty} F(x,y) = F(\infty, \infty) = 1$

② 만약 $x_1 < x_2$, $y_1 < y_2$ 이면 ,
$P[x_1 < X \leq x_2 ; y_1 < Y \leq y_2]$
$= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0$

정의 6.6 - 독립 stochastically independent, statistically independent

$p(x,y) = P_X(x) P_Y(y)$, X, Y가 이산형일 때

$f(x,y) = f_X(x) f_Y(y)$, X, Y가 연속형일 때

정의 6.7 - 조건부 확률밀도함수

$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$, $-\infty < y < \infty$

정의 6.8 - Y의 조건부 누적분포함수

$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(w|x)\,dw$

PDF 204

정의 6.9 - 기대값

X와 Y의 함수인 $h(X,Y)$의 기대값은

$E[h(X,Y)] = \sum_{x}\sum_{y} h(x,y)p(x,y)$, X, Y 가 이산형

$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)\,dx\,dy$, X, Y가 연속형

정의 6.10 - 공분산 Covariance

$Cov(X,Y) = \sigma_{X,Y}^2 = E[(X-\mu_X)(Y-\mu_Y)]$

$= E(XY) - \mu_X\mu_Y$, $\mu_X = E(X)$, $\mu_Y = E(Y)$

정의 6.11-상관계수

$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

정리 6.2 - 코시-슈바르츠 부등식 (Cauchy - Schwarz Inequality)

If $E(X^2) < \infty$, $E(Y^2) < \infty$, then

$[E(XY)]^2 \leq E(X^2)E(Y^2)$

정리 6.4 - X, Y가 서로 독립일 경우

① $E(XY) = E(X)E(Y)$
② $Cov(X,Y) = 0$, $\rho_{X,Y} = 0$
③ $Var(X \pm Y) = Var(X) + Var(Y)$
④ $g_1(X)$와 $g_2(Y)$ 함수가 존재할 때,
$E[g_1(X) \cdot g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)]$

정의 6.12 - 조건부 기대값

$E[g(X,Y) \mid X = x] = \sum_{y} g(x,y) P_{Y|X}(y|x)$, 이산형

$E[g(X,Y) \mid X = x] = \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x)\,dy$

정의 6.13 - 조건부 분산

$Var(Y \mid X = x) = E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2$

정리 6.5

① $E[g(Y)] = E[E(g(Y)|X)]$
② $Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$

정리 6.6

① $E[g_1(Y) + g_2(Y) \mid X = x] = E[g_1(Y) \mid X = x] + E[g_2(Y) \mid X = x]$
② $E[g_1(Y)g_2(X) \mid X = x] = g_2(x)E[g_1(Y) \mid X = x]$

PDF 216

6.4 - 두 확률변수의 선형결합

정리 6.7 - 두 확률변수 X, Y와 임의의 상수 a, b에 대해 다음이 성립한다.

① $E(aX + bY) = aE(X) + bE(Y)$
② $Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2ab\,Cov(X,Y)$

정리 6.8 확률변수 X, Y, W, T 와 임의의 상수 a, b, c, d에 대해 다음이 성립한다.

① $Cov(aX+b, cY+d) = ac\,Cov(X,Y)$
② $Cov(X+Y, Z+W)$
$= Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W)$
③ a와 c의 부호가 같다면,
$\rho_{aX+b, cY+d} = \rho_{X,Y}$
a와 c의 부호가 다르다면
$\rho_{aX+b, cY+d} = -\rho_{X,Y}$

PDF 221

6.5 - 이변량 정규분포

1. $f(x,y)$는 결합 확률밀도함수이다.
2. $X \sim N(\mu_X, \sigma_X^2)$이고, $Y \sim N(\mu_Y, \sigma_Y^2)$ 이다.
3. ρ는 X와 Y의 상관계수이다.

이변량 정규확률밀도함수 Bivariate Normal Probability Density Function

$\Rightarrow f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-Q/2}$,

$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$

$\to X, Y$가 독립이라면

$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$

* 이는 일변량 정규확률밀도함수 $f(x), f(y)$ 의 곱과 동일하다

$f(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

PDF 233

제 7장 - 표본 분포 Sample Distribution

정의 7.1 - 무작위추출

n개의 관측값이 서로 독립이고, 각 관측값의 분포는 모집단의 분포와 같다.
각 관측값은 동일한 모평균 μ와 모분산 σ^2을 지닌다.

정의 7.2- 확률표본

각 관측값이 상호 독립적이며, 동일한 분포를 따를 때, 크기 n인 확률표본이라고 한다.
(independent, identically distributed, random sample)

PDF 237

7.2 표본평균의 표본분포

정리 1.1

확률표본이 평균 μ, 분산 σ^2인 모집단으로부터 무작위 추출되었다면, 표본평균 $\bar{X}$의 평균과 분산은,

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

정의 7.3 - 표준 오차 Standard Error

$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ 또는 $s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

표본평균의 표본분포 (the sampling distribution of the sample mean)
- 확률 법칙 Probability Rule
- 모의실험 Simulation Experiment

PDF 250

7.3 - 중심극한 정리와 대수의 법칙

정리 1.7 - 중심극한정리

모집단의 분포와 상관없이 n이 커짐에 따라 $N(\mu, \sigma^2/n)$인 정규분포를 따르게 된다.

정리 7.3 - 대수의 약법칙 weak law of large numbers

$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$

* 체비셰프 부등식을 사용하여 증명

$P(|\bar{X}_n - \mu| < \epsilon) = P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$

$\to P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$ 잘 기억하기

$\therefore \lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$ or $\lim_{n \to \infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$

PDF 261

7.4 - 정규분포에서 추출된 표본분포

평균이 μ이고, 분산이 σ^2/n 인 정규분포의 적률생성 함수

$m(t) = e^{\mu t + \frac{(\sigma t)^2}{2n}}$

정의 7.5

$f_X(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2 - 1} e^{-\frac{1}{2}x}$

X의 확률밀도 함수가 다음과 같다면 X는 자유도가 k인 카이제곱분포를 따른다고 한다.

* 카이제곱분포는 $r = \frac{k}{2}$, $\lambda = \frac{1}{2}$ 를 갖는 감마분포의 특별한 형태이다.

정리 7.7

확률변수 X_i가 평균 μ_i, 분산 σ_i^2인 정규분포를 따르고 상호독립일 때,

$U = \sum_{i=1}^{k}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ 는 자유도가 k인 카이제곱분포를 따른다.

* U의 적률생성함수

$\prod_{i=1}^{k} E[e^{tZ_i^2}] = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}$, $t < \frac{1}{2}$

정리 7.9

확률표본이 표준정규분포로부터 왔다면,
① $\bar{Z}$는 평균이 0 이고, 분산이 $\frac{1}{n}$을 갖는 정규분포를 따른다.
② $\bar{Z}$와 $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$는 독립이다.
③ $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$는 자유도가 $n-1$ 인 카이제곱분포를 따른다.

정리 7.10

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 가 평균이 μ, 분산이 σ^2인 정규분포로부터 추출된 확률표본의 표본분산이면 $U = \frac{(n-1)S^2}{\sigma^2}$ 는 자유도가 $n-1$ 인 카이제곱분포를 따른다.

정리 7.11

서로 독립인 두 확률변수 U, V가 각각 자유도가 m, n인 카이제곱분포를 따른다면 $X = \frac{U/m}{V/n}$ 는 자유도가 m, n인 F분포를 따른다.

정리 7.13

확률변수 Z는 표준정규분포를 따르고 U는 자유도가 k인 카이제곱분포를 따르며, Z와 U가 독립이면

$X = \frac{Z}{\sqrt{U/k}}$ 는 자유도가 k인 t분포를 따른다.

* 자유도가 커짐에 따라 표준정규분포에 근접해간다.
("자유도가 k인 t분포")^2 == "자유도가 1과 k인 F분포"

PDF 268

7.5 표본비율의 표본분포

정리 7.15 - 표본비율의평균, 분산, 표준오차

$E(\hat{p}) = p$

$Var(\hat{p}) = \frac{p(1-p)}{n}$

$\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

PDF 281

제 8장 - 점추정

추정량 Estimator - 모수를 추정하기 위해 사용되는 통계량
추정값 Estimate - 추정량이 관찰되어 얻어진 값

점추정량 Point Estimator - 하나의 값으로 모수를 추정하는 통계량
점추정값 Point Estimate - 점추정량으로 얻어진 값

좋은 추정값을 판단하는 기준 (필수불가결은 아님)
- 불편성
- 최소분산
- 일치성

정의 8.3 - 불편추정량 Unbiased Estimator

$E(\hat{\theta}) = \theta$

* 편의추정량 : $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$

정의 8.5 - 상대효율 Relative Efficieny

$\hat{\theta}_1, \hat{\theta}_2$ 를 모수 θ 에 대한 불편추정량이라할 때,
$\hat{\theta}_2$에 대한 $\hat{\theta}_1$의 상대효율은 다음과 같다.

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

* If $RE(\hat{\theta}_1, \hat{\theta}_2) > 1$, $\hat{\theta}_1$이 더 효율적인 추정량

If $RE(\hat{\theta}_1, \hat{\theta}_2) < 1$, $\hat{\theta}_2$이 더 효율적인 추정량.

추정오차 : $|\hat{\theta} - \theta|$

정의 8.6 - 평균제곱오차 mean - squared error

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \\ &= Var(\hat{\theta}) + bias(\hat{\theta})^2 \end{aligned}$$

* 편의는 크고, 분산이 작은 추정량과 분산이 크고 편의가 작은 추정량 중 어떤 추정량이 좋은 추정량인지 판단

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$$

PDF 293

8.1.3 일치성 Consistency

일치성 - 추정량의 분포가 표본크기를 무한히 증가시켰을 때 한없이 집중되는 것을 말한다 ($\hat{\theta} \to 0$, $bias \to 0$)

일치 추정량 - Consistent Estimator - 아래 식을 만족하면 $\hat{\theta}_n$은 θ의 일치 추정량

$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$

평균제곱오차 일치 추정량 MSE Consistent Estimator

$\lim_{n \to \infty} MSE(\hat{\theta}_n) = 0$

PDF 298

8.2.2 - 최대우도방법 Maximum Likelihood Method

정의 8.11 우도함수 Likelihood Function

우도함수 $L(\theta; x_1, \ldots, x_n)$은 $X_1, \ldots, X_n$을 통해 얻어지는 결합확률밀도 함수로 정의되며 모수 θ의 함수이다.

$$\underbrace{L(\theta; x_1, \ldots, x_n)}_{=L(\theta)} = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$$

$L(\theta)$를 최대화 시키는 θ의 값을 모수 θ의 추정값으로 선택하는 방법을 최대우도방법이라 한다.

1. 우도함수 $L(\theta)$ 구하기
2. 우도함수에 자연로그 취하기
3. 로그우도함수 미분하기
4. 1차 미분된 로그우도함수가 0이 되는 지점구하기
   = 최대값

최대우도 방법의 특징

일반적인 정칙조건 Regularity Condition

① $\hat{\theta}$는 θ의 근사 불편 추정량
② $\hat{\theta}$는 θ의 일치 추정량이다.
③ 근사적으로 $\hat{\theta} \sim N(\theta, I_\theta^{-1})$ 이다. $I_\theta = E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]$
④ 임의의 함수 $g(\cdot)$ 에 대하여 $\eta = g(\theta)$ 이면 $\hat{\eta} = g(\hat{\theta})$

제 9장 - 구간추정

* 8장에서 배운 점추정량은 모수 θ에 대해 불편성, 효율성, 일치성을 만족하면 바람직한 추정량 estimator 가 되지만, 한 점의 형태이기 때문에 분산을 갖는다. ($\frac{\sigma^2}{n}$)
이러한 점을 보완하여 생긴 개념이 구간추정.

정의 9.1 - 신뢰구간

$P(|\hat{\theta} - \theta| \le c) = 1 - \alpha$
$P(\hat{\theta} - c \le \theta \le \hat{\theta} + c) = 1 - \alpha$

여기서 $0 \le \alpha \le 1$, 구간 $(\hat{\theta} - c, \hat{\theta} + c)$를 모수 θ에 대한 $100(1-\alpha)\%$ 신뢰구간 이라고 한다.

모수의 신뢰한계 = 표본의 추정량 ± 표본오차 (Sampling Error)
표본오차 = 신뢰계수 X 추정량의 표준오차 (Standard Error)
신뢰계수 = 신뢰도의 범위를 나타내는 상수

* 95 % 신뢰구간의 의미는 100번 표본을 추출하여 신뢰구간을 계산했을 때 5번 정도는 신뢰구간이 모평균 μ를 포함하지 않는 경우가 발생한다는 것

정리 9.1

확률표본 $X_1, \ldots, X_n$은 $N(\mu, \sigma^2)$에서 추출되었거나 또는 모집단의 분포가 알려져 있지 않더라도 표본크기 n이 충분히 큰 경우 알려진 모분산 σ^2에 대해 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간과 신뢰한계는 다음과 같다.

$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \ \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$ 신뢰구간

$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 신뢰한계

* 90 % 신뢰구간일 때 z-value = ± 1.65
95 % 신뢰구간일 때 z-value = ± 1.96
99 % 신뢰구간일 때 z-value = ± 2.58

PDF 315

9.2.2 모분산을 모르는 경우

대부분의 경우 모분산을 모르기 때문에, 이럴 경우 표본분산 (S^2) 을 사용한다.

정리 9.2 - t -분포

크기 n의 확률표본의 표본평균의 표준화계수 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ 가 표준정규분포를 따르면 $\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$ 는 자유도가 $n-1$인 카이제곱분포를 따르며,

두 확률변수는 서로 독립이므로 확률변수 $\frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(1/\sigma^2)\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}} = \frac{\sqrt{n(n-1)}\,(\bar{X} - \mu)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$ 은 자유도가 $n-1$인 t분포를 따른다.

즉, 확률변수 $\frac{\bar{X} - \mu}{S/\sqrt{n}}$은 자유도가 $n-1$인 t분포를 따른다.

* t분포의 특징

1. 표준정규분포보다 중앙밀도가 낮다.
2. 모집단이 정규분포를 따르지 않고, 표본크기가 작으며 ($n < 30$), 모분산이 알려져 있지 않은 경우 t분포를 사용할 수 없다.

* 자유도 Degree of Freedom

Def : 자유롭게 값을 가질 수 있는 관측치의 개수

Ex: t분포의 자유도

- S^2을 계산하는데 사용된 정보의 양
- S^2의 분모
- " n - 제약식의 개수 "
- " $n-1$ "

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

표준정규분포에서는 얘가 σ

PDF 319

정리 9.3 - 모집단이 정규분포를 따르면 표본크기가 $n \le 30$ 일 때 모평균의 신뢰구간을 구하는 방법

$\left(\bar{X} - t_{(\alpha/2, n-1)} \frac{S}{\sqrt{n}} \ \ \bar{X} + t_{(\alpha/2, n-1)} \frac{S}{\sqrt{n}} \right)$ 신뢰구간

$\bar{X} \pm t_{(\alpha/2, n-1)} \frac{S}{\sqrt{n}}$ 신뢰한계

P 321

9.3 모평균 차의 구간추정

정리 9.4

n_1과 n_2 둘 다 독립적으로 추출된 표본일 때, $\bar{X} - \bar{Y}$의 분포는 다음과 같다.

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

정리 9.5 - 2개의 독립적인 모분산을 모두 아는 경우

$$\left[(\bar{X} - \bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \ (\bar{X} - \bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$ 신뢰구간

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$ 신뢰한계

* 만약 2개의 모분산이 동일하다면

$$\left[(\bar{X} - \bar{Y}) - z_{\alpha/2}\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \ (\bar{X} - \bar{Y}) + z_{\alpha/2}\,\sigma\sqrt{\frac{1}{n_1} - \frac{1}{n_2}} \right]$$

2개의 모분산 모두 모를 경우

1. 두 모집단의 모분산을 동일하다는 것을 가정

분산은 합동분산 (pooled variance) S_p^2 을 사용한다

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2}{(n_1 - 1) - (n_2 - 1)} = \frac{(n-1)S_1^2 + (n_2 - 1)S_2^2}{n + n_2 - 2}$$

2. 두 모집단의 모분산이 동일하지 않다는 것을 가정하는 경우

정리 9.6

if ...

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$

- $\left[(\bar{X} - \bar{Y}) - t_{(\alpha/2, n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \ (\bar{X} - \bar{Y}) + t_{(\alpha/2, n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$ 신뢰구간

$(\bar{X} - \bar{Y}) \pm t_{(\alpha/2, n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 신뢰한계

if ...

2. $\sigma_1^2 \ne \sigma_2^2$, then $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(v)}$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

- $\left[(\bar{X} - \bar{Y}) - t_{(\alpha/2, v)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \ (\bar{X} - \bar{Y}) + t_{(\alpha/2, v)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$

여기서 v는 웰치의 조정자유도 (Welch's adjusted DF) 또는 새터스웨이트의 자유도 (Satterthwaite's DF) 라고 한다.

PDF 331

9.3.2 대응표본

정리 9.8

[illegible]

을 의미한다.

제 10장 - 가설검정

PDF 355

10.1- 가설검정의 이론

10.1.1 통계적 가설과 검정통계량

Null Hypothesis / Alternative Hypothesis
귀무가설과 대립가설

- 가설검정을 할 때 관심있는 것은 대립가설이다.
- "귀무가설 채택" ≠ "귀무가설이 옳다."
  - 귀무가설을 기각하지 못한다만 말할 수 있다.

정의 10.3- 검정통계량 test statistic

가설검정에 사용되는 모든 통계량

EX: μ, σ^2

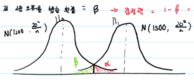
"제 1종 오류" = "Type I Error" = "False Negative" : 실제론 참인데 거짓이라고 주장
"제 2종 오류" = "Type II Error" = "False Positive" : 실제론 거짓인데 참이라고 주장

$\alpha = P($제 1종 오류$) = P(H_0$를 기각 $\mid H_0$이 사실$)$

$\beta = P($제 2종 오류$) = P(H_0$를 채택 $\mid H_1$이 사실$)$

제 1종 오류를 범할 확률 = α ⇝ 유의수준 - Level of significance

제 2종 오류를 범할 확률 = β ⇝ 검정력 = $1-\beta$ = Power of Test



두 오류의 확률 α와 β의 관계

① 주어진 표본크기 n 하에서는 α와 β는 동시에 줄일 수 없다. α와 β는 서로 반비례한다.

② α와 β를 동시에 줄이려면 n을 증가시키면 된다.

10.1.3 - 검정력

α와 β는 반비례하기 때문에, 옳은 결정을 할 확률인 $1-\alpha$와 $1-\beta$을 동시에 크게 만들 수 없다.
* 그래서 유의수준 α를 고정시키고 β를 줄여 $1-\beta$를 최대화시킨다.

1. α를 고정하는 이유 : 허용할 수 있는 최대의 α를 고정시켜 이 이상의 오류를 허용하지 않기 위해
2. β를 줄이는 이유: 귀무가설이 참이고, 채택하게 된다면 연구자에게 의미있는 성과가 아니다.

PDF 360

10.1.4 - 임계값 Critical Value

주어진 유의수준 α 하에서 귀무가설 H_0의 채택 또는 기각 여부를 판정하여 주는 기준값

정의 10.7 - 채택역과 기각역

① 채택역 : H_0을 받아들이는 영역. Acceptance Region

② 기각역 : H_0을 받아들이지 않는 영역. Rejection Region

10.1.6- p- 값의 이용

H_0가 참일 때 표본에서 얻어진 결과가 귀무가설을 기각하게 하는 확률.

PDF 364

10.1. 7 - 구간추정과 가설검정의 관계

신뢰구간 = 채택역

PDF 365

10.2 - 단일모평균의 검정

1. 모분산을 알 때.

$$-z_{\alpha/2} \le \sqrt{n}(\bar{x}-\mu_0)/\sigma \le z_{\alpha/2}$$

11 장 - 범주형 자료분석

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt$$

* $P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$

$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$

$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

정리 5.13

X가 $\mu = np$, $\sigma^2 = npq$, $q = 1-p$ 인 이항분포를 따른다면 $n \to \infty$ 일 때,

$Z = \frac{X-np}{\sqrt{npq}}$ 의 극한 분포는 표준정규분포가 된다.

정리 5.14

X가 $\mu = \lambda$, $\sigma^2 = \lambda$ 인 포아송분포를 따른다면 λ가 커짐에 따라

$Z = \frac{X-\lambda}{\sqrt{\lambda}}$ 의 분포는 표준정규분포에 근사해간다.

PD- 171

정의 5.17 - 지수분포

$f(x;\lambda) = \lambda e^{-\lambda x}$, or $f(x;\frac{1}{\theta}) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, $x>0$

λ : 단위시간에 발생하는 평균사건 수

θ : 사건들 사이에 걸리는 평균시간

정의 5.13- 감마분포

$$f(x; r, \lambda) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x}$$

or

$$f(x; r, \tfrac{1}{\theta}) = \frac{1}{\theta \Gamma(r)} \left(\frac{x}{\theta}\right)^{r-1} e^{-\frac{x}{\theta}}$$

r : 형태 (Shape) 모수로, 사건 발생 횟수이다.

λ : 척도 (Scale) 모수로, 단위시간에 발생한 평균사건 횟수이다.

정리 5.15 - 감마함수 Gamma Function

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\,dx, \quad r>0$$

* 감마함수 특징 :

a) $\Gamma(1) = 1$

b) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

c) $\Gamma(r+1) = r\Gamma(r)$

d) $\Gamma(r+1) = r!$

정리 5.16 - 지수분포와 감마분포의 평균. 분산, 적률생성함수

지수분포 : $E(X) = \frac{1}{\lambda} = \theta$, $Var(X) = \frac{1}{\lambda^2} = \theta^2$, $m(t) = \frac{\lambda}{\lambda - t}$, $t < \lambda$

* $F(x) = 1 - e^{-\lambda x}$

감마분포 : $E(X) = \frac{r}{\lambda} = r\theta$, $Var(X) = \frac{r}{\lambda^2} = r\theta^2$, $m(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$, $t < \lambda$

PDF 177

정의 5.14- 카이제곱분포

자유도가 v인 카이제곱분포는 다음과 같다.

$$f(x; v) = \frac{1}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}$$

$E(X) = v$, $Var(X) = 2v$

$m(t) = \left(\frac{1}{1-2t}\right)^{\frac{v}{2}}$, $t < 2$

베타 분포 Beta Distribution

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$E(X) = \frac{\alpha}{\alpha+\beta}$, $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

여기서 v는 웰치의 조정자유도 또는 새터스 웨이트의 자유도라고 한다.
Welch's adjusted DF Satterthwaite's DF

PDF 331

9.3.2 대응표본

정리 9.8

n 개의 차이 $D_1, \ldots, D_n$이 정규모집단으로부터의 추출된 확률표본이라면 모집단의 차이 Δ에 대한 $100(1-\alpha)\%$ 신뢰구간과 신뢰한계는 다음과 같다.

$$\left[\bar{D} - t_{(\alpha/2,\, n-1)} \frac{S_D}{\sqrt{n}},\ \bar{D} + t_{(\alpha/2,\, n-1)} \frac{S_D}{\sqrt{n}} \right]$$

$$\Delta = \bar{D} \pm t_{(\alpha/2,\, n-1)} \frac{S_D}{\sqrt{n}}$$

정리 9.9

표본비율 $\hat{p}$의 표본분포는 근사적으로 정규분포를 따른다.

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

표본 비율 $\hat{p}$를 표준화하면 다음과 같이 표준정규분포를 따른다.

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

정리 9.10

표본이 충분히 클 경우 모비율 p에 대한 $100(1-\alpha)\%$ 신뢰구간과 신뢰한계는 다음과 같다.

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$ 신뢰구간

$$p = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ 신뢰한계

PDF 342

9.6 모분산의 구간추정

정리 9.13

평균 μ, 분산 σ^2인 정규분포를 따르는 모집단으로부터 크기가 n인 확률표본을 반복하여 추출하여 각 표본에 대한 분산 S^2을 계산하면 확률변수 $\frac{(n-1)S^2}{\sigma^2}$은 자유도가 $(n-1)$인 카이제곱분포를 따른다.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

카이제곱분포의 신뢰구간

$$\Rightarrow P\left(\chi^2_{(1-\alpha/2,\, n-1)} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{(\alpha/2,\, n-1)}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S^2}{\chi^2_{(\alpha/2,\, n-1)}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{(1-\alpha/2,\, n-1)}}\right) = 1-\alpha$$

$$\Rightarrow \left[\frac{(n-1)S^2}{\chi^2_{(\alpha/2,\, n-1)}},\ \frac{(n-1)^2 S^2}{\chi^2_{(1-\alpha/2,\, n-1)}} \right]$$

* 여기서 $P(X^2 \ge \chi_{(\alpha/2,\, n-1)}) = P(X^2 \le \chi_{(1-\alpha/2,\, n-1)}) = \alpha/2$ 를 만족한다.

PDF 344

9.7 - 허용오차 및 표본크기의 결정

정리 9.15 - 모수를 θ, 모수를 추정하기 위한 추정량을 $\hat{\theta}$이라고 하면 허용오차의 종류는 다음과 같다.

① 절대오차 : $|\hat{\theta} - \theta|$

② 상대오차 : $\frac{|\hat{\theta} - \theta|}{\theta}$

③ 상대표준오차 : $\frac{\sigma(\hat{\theta})}{\theta}$

단일모집단일 때 표본오차구하는 방법

	모평균 추정		모비율 추정
	σ^2를 알 때	σ^2를 모를 때	
표본오차 (허용오차)	$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$e = t_{(\alpha/2,\, n-1)}\frac{s}{\sqrt{n}}$	$e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
표본크기	$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$	$n = \left(\frac{t_{(\alpha/2,\, n-1)} S}{e}\right)^2$	$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 p(1-p)$

비율 문제에서 모비율 p의 정보가 없을 때는 $p = 0.5$로 두고 시작한다.
$P = \{p \mid 0 \le p \le 1\}$ 이므로 $p = 0.5$ 일 때 $p(1-p)$가 최대값을 갖는다.

pf)

$f(p) = p(1-p) = p - p^2$

$f'(p) = 1 - 2p$

$f'(0.5) = 1 - 1 = 0$