기초통계학

기통1 : 기술통계학
기통2 : 추론통계학

수업진도

주차	강의내용	비고
1	강의소개 및 복습 (4장-7장)	
2	제8장 통계추론개요	
3	제9장 단일모집단에 대한 통계적추론	과제1: 8장,9장 관련연습문제
4	제10장 두모집단의 비교 -모평균비교	
5	제10장 두 모집단의 비교 - 모분산비교 -모비율비교	과제2: 10장관련 연습문제
6	제8장부터 제10장까지 복습	
7	과제풀이 및 질의 응답	
8	중간고사	비대면실시
9	제11장 분산분석	
10	제12장 회귀분석1	
11	제12장 회귀분석2	과제3: 11,12장 관련 연습문제
12	제13장 범주형자료분석	과제4: 13장 관련 연습문제
13	제11장에서 13장까지 복습	
14	과제풀이 및 질의응답	
15	기말고사	비대면실시

기초통계학 II 강의계획서

담당교수 : 최 승 경

e-mail : skchoi@sookmyung.ac.kr

교재 : 통계학 -기본개념과 원리- 여인권지음, 자유아카데미

평가계획 : 상대평가 특별기준 (50% => A) 6/22기준 25명 → A 12명

출석 10 과제(4회) 20 중간고사 35 기말고사 35

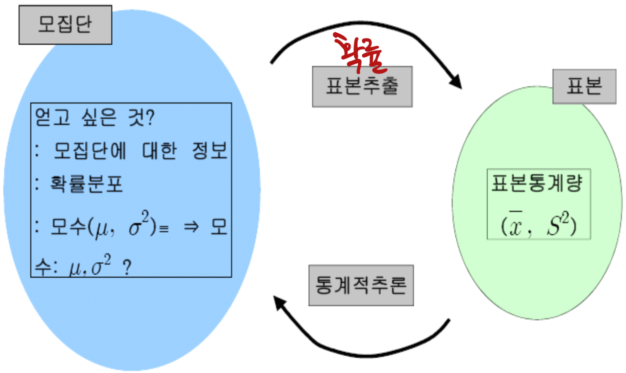
- 통계학(Statistics) : 일반화를위한 객관적인 증거 제공

관심 또는 연구 대상이 되는 집단(모집단)의 특성을 파악하기 위해, 모집단에서 일부(표본)를 대상으로 자료를 수집, 정리, 요약, 분석하여 표본의 특성을 파악하고 이것을 이용하여 모집단의 특성에 대해 추론하는 원리와 방법을 제공하는 학문

=> 통계로부터 정보를 뽑아내는 과학적인 방법

- **모집단(Population)** : 얻고자 하는 정보와 관련 있는 모든 개체로부터 얻을 수 있는 모든 관측값들의 집합.
  우리가 관심을 갖는 정보의 대상이 되는 전체 집단
- **표본(Sample)** : 모집단을 대표할 수 있도록 선택된 모집단의 일부(실제 조사 대상자)

확률표본추출: 표본이 나올 확률을 구할 수 있음
(결과값으로 모집단에 대한 추론할 때 정확도에 대한 수치 필요)
ex) 무작위(random)



N: 모집단의 크기, μ: 모평균, σ^2: 모분산,
n: 표본의 크기, $\bar{x}$: 표본평균, S^2: 표본분산

※ 통계량(statistic) :

-모수를 추정하기 위한 표본의 특성값

-통계량의 값은 표본에 따라 값이 변한다 → 확률변수 (확률분포: 표집분포, 표본분포)

※ 모수(parameter) :

-모집단의 성질을 대표하는 수치 μ, p, σ^2 등

-확률밀도함수(모집단의 모양)의 특성을 결정하지만 일반적으로 모르는 값

⊕ 분포에서 중요한 역할을 하는 수치
ex) 이항분포: n, p에 따라 달라짐

표본공간 ← 확률변수X → 수치

모든실험결과 ex) 동전2개 | ex) 앞면의수 | ex) 0 1 2

- **확률변수**(random variable) :

  각각의 근원사건들에 실수 값을 대응시키는 함수

- **확률분포**(probability distribution):

  확률변수가 갖는 값들과 그에 대응하는 확률값을 나타내는 것으로 나열된 표 또는 수식으로 표현 : 확률변수 X에 대한 모든정보

ex)

X	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

요약

▶ **기대값**(expected value)

요 $$\mu = E(X) = \sum_i x_i \cdot f(x_i)$$

▶ **분산**(Variance), **표준편차**(Standard Deviation)

요 $$\sigma^2 = Var(X) = E\left[(X-\mu)^2\right] = \sum_i (x_i-\mu)^2 \cdot f(x_i)$$

$$= \sum_i x_i^2 \cdot f(x_i) - \mu^2 = E[X^2]-\mu^2 = E[X^2]-[E(X)]^2$$

– 표준편차: $sd(X) = \sigma = +\sqrt{Var(X)}$

예제) 한 대리점에서 과거 300일 동안의 자동차 판매 자료

X : 하루 동안 판매한 자동차 수

X의 값	도수	$f(x)$	$F(x)$
0	54	0.18	0.18
1	117	0.39	0.57
2	72	0.24	0.81
3	42	0.14	0.95
4	12	0.04	0.99
5	3	0.01	1.00
합계	300	1.00	

$$E(X) = \sum_{i=1}^{6} x_i f(x_i) = 0 + 0.39 + 0.48 + 0.42 + 0.16 + 0.05 = 1.5$$

$$Var(X) = E(X^2) - [E(X)]^2 = 3.5 - 1.5^2 = 1.25,\ \ sd(X) = 1.118$$

$$E(X^2) = 0^2 \cdot 0.18 + 1^2 \cdot 0.39 + 2^2 \cdot 0.24 + 3^2 \cdot 0.14 + 4^2 \cdot 0.04 + 5^2 \cdot 0.01 = 3.5$$

이산확률변수

베르누이시행
① 성공/실패 두가지
② 각 시행이 독립
③ 성공확률이 매 시행 p로 동일

- 이항분포(binomial distribution): $X \sim Bin(n,p)$

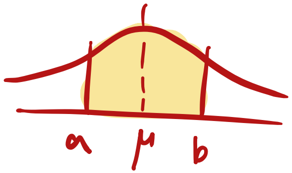
  확률질량함수 : $f(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$

  $$E(X) = np \quad , \quad Var(X) = np(1-p)$$

연속확률변수

- 정규분포(normal distribution) $X \sim N(\mu, \ \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad , \quad -\infty < x < \infty$$



→ $P(a \le X \le b) = \int_a^b f(x)dx$

▶ 표준정규분포 (Standard Normal Distribution)

$$Z \sim N(0,1)$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}z^2)$$: 표준정규분포의 확률밀도함수

직접적분하기 어려우므로 부록표 이용

< 정규분포일 때 확률계산>

X가 정규분포를 따르면 Z도 정규분포를 따른다

$$X \sim N(\mu,\ \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$:표준화

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right)$$

▪ 표집분포(sampling distribution) : 통계량의 확률분포

모집단

$N(\mu, \sigma^2)$

확률표본 (iid: 서로독립, 같은분포)

$\{X_1 \cdots X_n\}$

모집단의 분포와같은 확률변수

(1) 표본평균의 분포

◈ 모집단이 정규분포인 경우

정규분포 $N(\mu, \sigma^2)$을 따르는 모집단에서 크기가 n인 표본의 표본평균 $(\overline{X})$의 분포는 다음과 같다.

$\frac{X_1 + \cdots + X_n}{n}$

$E(\overline{X}) = \frac{1}{n} E(X_1 + \cdots + X_n) = \frac{1}{n} \cdot n \cdot \mu = \mu$

$Var(\overline{X}) = \frac{1}{n^2} Var(X_1 + \cdots + X_n)$

$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$

∵서로독립

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

◈ 모집단이 정규분포가 아닌 경우

대표본

평균이 μ이고 분산이 σ^2인 임의의 모집단에서 표본의 크기 n이 충분히 크면(30이상), 표본평균 $\overline{X}$는 근사적으로 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다.

즉 n이 충분히 클 때 다음이 성립한다.

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

⇨ 중심극한정리 (central limit theorem ; CLT)

모집단 (0또는1)

$p = 1$의수 $=$ 총합

$\{X_1 \cdots X_n\}$

$1 \to p$, $0 \to 1-p$ → 베르누이시행 n번

(2) 표본비율의 분포

* p : 모비율(population proportion),

$\hat{p}$: 표본비율(sample proportion)

표본의구성에따라달라짐

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$$

$X \sim Bin(n,p)$

표본비율의 평균 : $\mu_{\hat{p}} = E(\hat{p}) = E\left[\frac{X}{n}\right] = \frac{1}{n} np = p$

표본비율의 분산 : $\sigma^2_{\hat{p}} = Var(\hat{p}) = Var\left[\frac{X}{n}\right] = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$

모집단이정규분포가아님

$\hat{p}$ 이 어떤 분포인지 알수 없다.

$\hat{p} = \frac{\Sigma x_i}{n}$

← $\bar{X}$의 식과 같다

<표본크기가 큰 경우>

≥ 30

표본크기 n이 충분히 큰 경우, 표본비율 $\hat{p}$ 는 근사적으로 평균이 p이고 표준편차가 $\sqrt{p(1-p)/n}$ 인 정규분포를 따른다. (중심극한정리)

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

표준화

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

표본크기가 작은 경우는 아직 다루지 않음

예제) 다음 물음에 답하시오.

(1) 어느 대학교 입학생 1000명의 평균 수능성적은 315점이고, 분산은 400점이다. 크기 50의 표본을 추출하였을 때 표본평균의 기대값과 분산을 구하라.

$$E(\overline{X}) = \mu = 315 \ , \qquad Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{400}{50} = 8$$

$N=1000$
$\mu=315$
$\sigma^2=400$

$n=50$

어떤 확률변수에 대한 확률을 구하려면 그 분포를 알아야 함.

(2) 평균이 82이고 분산이 144인 모집단에 대하여 생각

(i) 크기 64인 표본이 임의추출되었다고 할 때 표본평균이 80.8에서 83.2사이에 있을 확률 → 중심극한정리

$$\overline{X} \sim N(82, \frac{144}{64}) = N(82, 1.5^2)$$

$$P[80.8 \le \overline{X} \le 83.2] = P\left[\frac{80.8-82}{1.5} \le Z \le \frac{83.2-82}{1.5}\right] = P[-0.8 \le Z \le 0.8]$$
$$= 0.7881 - 0.2199 = 0.5762$$

(ii) 크기 100일 때 표본평균이 80.8에서 83.2사이에 있을 확률

$$\overline{X} \sim N(82, \frac{144}{100}) = N(82, 1.2^2)$$

$$P[80.8 \le \overline{X} \le 83.2] = P\left[\frac{80.8-82}{1.2} \le Z \le \frac{83.2-82}{1.2}\right] = P[-1 \le Z \le 1]$$
$$= 0.8413 - 0.1587 = 0.6826$$

n이 커질수록 평균에 몰릴 확률↑

n 이 충분히 크므로 중심극한정리

(3) 어느 대학의 사회계열 학생들을 대상으로 하는 통계학입문 강의는 70명씩 10개반으로 나누어 진행되고 있다. 사회계열 학생들은 각자 수강하려는 모든 과목들의 시간표를 고려 하여 다른 교양과목이나 전공과목들과 시간이 겹치지 않도록 10개의 반 중 어느 하나를 수강해야 한다. 어느 특정 통계학입문 반의 여학생수가 과반이상이 될 확률을 구하여라. 사회계열 전체 여학생의 비율은 0.54이다.

$$p = 0.54, \quad \hat{p} \sim N\left(0.54, \frac{0.54 \cdot 0.46}{70} = 0.00355\right)$$

$$P(\hat{p} \geq 0.5) = P\left(Z \geq \frac{0.5 - 0.54}{\sqrt{0.00355}}\right) = P(Z \geq -0.67) = 0.7486$$

$= P(Z \leq 0.67)$