쉽게 배우는 통계입문

Introduction to Statistics

Present by Hobin Kwak

MetaCode

1 통계

통계는 데이터의 **수집**, **분석**, **추론**, **요약** 등의 방법론을 다룬다.
(The art and science of learning from data)

- Design (설계/계획)
- Description (요약)
  : 데이터를 요약 표현하기 위한 시각적(Graphical), 수치적(numerical) 방법
- Inference (추론)
  : 표본에 기반한 모집단에 대한 추론/예측

MetaCode

1 통계

모집단(Population): 통계학에서 관심/조사의 대상이 되는 개체의 전체 집합

모수(Parameter): 모집단에 대한 수치적 요약
- 고등학생의 1일 평균 온라인게임 플레이 시간
- 강아지보다 고양이를 좋아하는 성인의 비율

표본(Sample): 모집단을 적절히 대표하는 모집단의 일부

통계량(Statistic): 표본에 대한 수치적 요약
- 고등학생 1000명의 1일 평균 온라인게임 플레이 시간
- 강아지보다 고양이를 좋아하는 성인의 비율 (1000명)

sample statistic → population parameter!

2 자료의 종류

1. **범주형 자료 :** 속성의 범주화, 상대적 서열도 표현
    1. 명목형 자료 : 단순히 속성을 분류하기 위함 (혈액형)
    2. 순서형 자료 : 상대적인 크기 비교 (만족도, 최종학력)

2. **양적 자료 :** 자료자체가 숫자로 표현됨
    1. 이산형 자료 : 셀 수 **있음** (빈도수, 불량품의 수)
    2. 연속형 자료 : 셀 수 **없음** (길이, 시간)

3 통계량 - 중심

1. 최빈값 (mode)

- 발생빈도가 가장 높은 값
- 극단값에 영향을 받지 않음
- 주로 범주형 자료에 대한 대표값
- 2개 이상 존재 가능

사이즈	수량
S	5
M	**25**
L	10
XL	0

2. 중앙값 (median)

- 크기 순으로 정렬된 자료에서 가운데에 위치하는 값
- 관측값 변화에 민감하지 않음
- 극단값에 영향을 받지 않음

1 2 3 4 5 6 7 8 9

1 2 3 4 5 6 7 8 9 10

3 통계량 - 중심

3. 산술평균 (Arithmetic Mean)

- 모든 자료의 값을 더하여 자료의 수로 나누어 준 값
- 모든 값을 반영하므로 극단값에 영향을 받음

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

4. 가중평균 (Weighted Mean)

- 자료의 중요성이 각기 다를 경우 중요도에 따라 가중치를 부여한 평균

$$\bar{X} = \frac{w_1\bar{X}_1 + w_2\bar{X}_2 + \cdots + w_i\bar{X}_i}{w_1 + w_2 + \cdots + w_i} = \frac{\sum w_i\bar{X}_i}{\sum w_i}$$

통계량 - 중심

5. 기하평균 (Geometric Mean)

- 자료가 성장률, 증가율 등 앞 시점에 대한 비율로 나타난 경우 유용한 통계량
- 음수가 아닌 자료값 only
- 연간 물가 상승률

$$\text{기하평균 } (\mathbf{G}) = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad (x_i = \text{상승률})$$

Ex) 일일 주가 상승률 : 1% 3% 5% 10% : 1.0374….

3 통계량 - 중심 : 예제

1. 1반과 2반의 학생이 각각 30명, 50명이고
   평균성적은 각각 70점, 80점일 때, 두 반 전체의 평균 성적은?

2. 국어(3학점) A+, 영어(2학점) B, 컴퓨터(4학점)
   C+, 과학(2학점) B0이고, A+부터 C+까지 4.5~2.5의 평점을
   가질 때 전체 과목 평점의 평균은?

통계량 - 중심 : 예제

3. 5년간 물가 상승률이 각각 3%, 5%, 6%, 2%, 4%일 때,
   물가 상승률의 평균은?

4. 주어진 Data가 다음과 같을 때, 중앙값은?

Data: 8, 5, 6, 2, 9, 4, 3

3 통계량 - 산포

1. **분산 (Variance)**

   - 편차 제곱의 합을 자료의 수로 나눈 값

$$\sum_{i=1}^{n}(x_i-\overline{x})^2/(n-1)$$

2. **표준편차 (Standard Deviation)**

   - 분산을 제곱근한 값

$$\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2/(n-1)}$$

MetaCode

통계량 - 산포 : 예제

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) = \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) / (n-1)$$

3

통계량 - 산포 : 예제

2. 표본의 크기가 10일 때, data의 합은 20이고
   data의 제곱의 합은 75일 때, 표본분산의 값은?

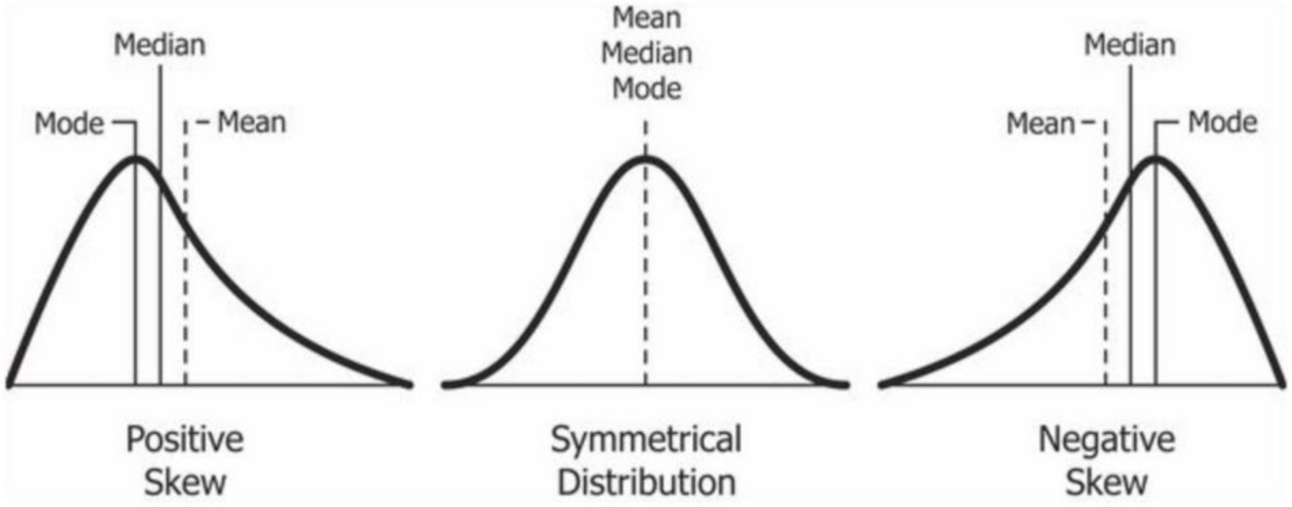
MetaCode

3

통계량 - 형태

1. 왜도 (Skewness)

- 분포의 비대칭도



2. 첨도 (Kurtosis)

- 뾰족한 정도
- 표준정규분포의 첨도는 3이 된다.

3 통계량 - 상관

1. **상관 (Correlation)**
   - 확률변수 X,Y의 변화가 서로 관계가 있을 때 상관관계가 있다고 함
   - 선형적 관련성을 파악함

2. **공분산 (Covariance)**

$$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

3. **상관계수 (Correlation Coefficient)**

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2]^{1/2}}$$

3 통계량 - 상관

3. 상관계수 (Correlation Coefficient)

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}$$

- 공분산을 두 변수의 표준편차의 곱으로 나눈 값
- $-1 < r < 1$
- 두 양적 변수 간의 선형적 연관성의 강도 측정
- 단위가 없음
- 절댓값이 1에 가까울 수록 연관성의 강도가 높다

3 통계량 - 상관 : 예제

1. 독서량(X)과 성적(Y)의 상관관계를 조사하고자 할 때,
   학생 5명을 뽑아 다음과 같이 조사하였다. 두 변수의 상관계수는?

독서량(X):	0	2	3	6	6
성적(Y):	10	50	45	70	60

MetaCode

4 확률과 확률변수 : 확률 정의

1. **표본공간(S) :** 랜덤한 현상의 모든 가능한 결과의 집합

2. **사건(event)** : 표본공간의 부분집합

    1. 합사상
    2. 곱사상
    3. 여사상
    4. 배반사상

3. **Flipping Coin Twice**

    1. 표본공간 S : {HH, HT, TH, TT}
    2. 사건 A : 동전을 두 번 던지는 시행에서 동전의 앞면이 1번만 A = {HT, TH}

MetaCode

4 확률과 확률변수

1. **확률의 고전적 정의**
   : 가능한 결과가 N가지이고, 각 결과가 나타날 가능성이 모두 같을 때, 사건 A에 속하는 결과가 m개라면 A의 확률

$$P(A) = \frac{m}{N}$$

2. **경험적 정의 (상대도수)**

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

MetaCode

3. 확률의 공리적 정의

: 표본공간 S에서의 임의의 사상 A에 대하여,

- $0 \le P(A) \le 1$
- $P(S) = 1$
- 서로 배반인 사상들에 대하여

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

이 때, P(A)를 사상 A의 확률이라고 함

MetaCode

1. 확률의 성질

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A^c) = 1 - P(A)$
- An이 서로 배반사상일 때

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

- $A \subset B$이면 $P(A) \leq P(B)$

4 확률과 확률변수 : 조건부확률

한 사건이 일어날 것을 전제로 다른 사건이 일어날 확률
(**변화된 표본공간**에서의 사건 발생 확률)

- B가 일어났을 때 A가 일어날 확률

- A가 일어났을 때 B가 일어날 확률

MetaCode

4 확률과 확률변수 : 예제

1. 한 보험회사의 고객은 타입 A,B로 분류된다. A고객이 사고가 날 확률은 30%, B고객이 사고가 날 확률은 10%이다.

   (a) 모집단의 20%가 타입A일 때, 새 고객이 사고가 날 확률은?

   (b) 새 고객이 사고가 날 때, 그 고객의 유형이 타입A일 확률은?

4 확률과 확률변수 : 독립과 종속

1. **독립사건 :** 한 사건의 발생이
   다른 사건의 발생 확률에 영향을 주지 않음

- 사건 A와 B가 독립이면 $P(A \cap B) = P(A)P(B)$

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

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

2. **종속사건 :** 한 사건의 발생이 다른 사건의 발생 확률에 영향을 줌

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

MetaCode

4 확률과 확률변수 : 베이즈 정리

사건 A1, … An이 표본공간 S의 분할이고 P(A) >0, P(B) >0 일 때,

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

- P(A_k)는 원인의 가능성 : 사전확률
- P(B|A_k)는 원인 A_k의 결과로서 B가 관측될 확률
- P(A_k|B)는 B가 관측된 후에 원인 A_k의 가능성 : 사후확률
- 사전확률을 사후확률로 전환할 수 있음

4 확률과 확률변수 : 확률변수

1. **확률변수**
   - 표본공간에서 정의된 실수값 함수
     - 실수가 아니면 확률분포함수 정의할 수 없음
   - 일정 확률을 가지고 발생하는 사건에 수치를 부여한 것
   - 변수가 어떤 값을 취하는지가 확률적으로 결정된다
     - 통계적 규칙성은 있다고 봄

2. **확률분포**
   - 확률변수의 값과, 확률을 대응시켜
     표, 그래프, 함수로 표현한 것

4 확률과 확률변수 : 이산/연속확률변수

1. **이산확률변수**

 - 이산표본공간에서 정의된 확률변수의 값이
  유한 혹은 countably infinite
 - 확률질량함수
  : 이산확률변수 X의 값 x1,⋯, xn의 각 확률을 대응

2. **연속확률변수**

 - 특정 구간 내의 모든 값을 취하는 확률변수
 - 확률변수의 값이 무한개이며 셀 수 없음
 - 확률밀도함수
  : 확률변수 X가 어떤 구간 [l, u]의 모든 값을 취하고
   이 구간에서의 함수 f(x)

$$(a) f(x) \geq 0, \int_l^u f(x)dx = 1$$

$$(b) P\{a \leq X \leq b\} = \int_a^b f(x)dx \quad (\text{단}, l \leq a < b \leq u)$$

MetaCode

4 확률과 확률변수 : 기대값

1. 기대값 (expected value)

- 확률변수의 모든 값의 평균
- 이산확률변수
  - 확률변수의 값이 x1,⋯ 이고 X=xi일 확률이 f(xi)일 때,

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$$

- 연속확률변수
  - 확률변수 X가 [l, u] 구간의 모든 값을 취하고
    X의 확률밀도함수가 f(x)일 때,

$$E(X) = \int_l^u x f(x) dx$$

4 확률과 확률변수 : 기대값의 성질

1. **기대값의 성질** (a, b는 상수이고 X, Y는 확률변수)

$$E(a) = a$$

$$E(aX) = a \cdot E(X)$$

$$E(X \pm b) = E(X) \pm b$$

$$E(aX \pm b) = a \cdot E(X) \pm b$$

$$E[c_1 g_1(X) + c_2 g_2(X)] = c_1 E[g_1(X)] + c_2 E[g_2(X)]$$

MetaCode

4 확률과 확률변수 : 분산과 표준편차

1. 분산

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

   1. 이산확률변수 $Var(X) = E[(X-\mu)^2] = \sum (x_i - \mu)^2 f(x_i)$
   2. 연속확률변수 $Var(X) = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)dx$

2. 표준편차

$$sd(X) = \sqrt{Var(X)} = \sqrt{E(X-\mu)^2}$$

MetaCode

4 확률과 확률변수 : 분산과 표준편차의 성질

1. 분산과 표준편차의 연산

$$Var(X \pm b) = Var(X)$$

$$\sigma(X \pm b) = \sigma(X)$$

$$Var(aX) = a^2 Var(X)$$

$$\sigma(aX) = a\sigma(X)$$

$$Var(aX \pm b) = a^2 Var(X)$$

$$\sigma(aX \pm b) = a\sigma(X)$$

MetaCode

4 확률과 확률변수 : 공분산과 상관계수

1. 공분산과 상관계수

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2]^{1/2}}$$

$$Cov(X, Y) = E[(X - \mu_1)(Y - \mu_2)]$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

MetaCode

4 확률과 확률변수 : 공분산과 상관계수

1. 공분산과 상관계수의 성질

$$Cov(aX + b, cY + d) = acCov(X,Y)$$

$$Corr(aX + b, cY + d) = \begin{cases} Corr(X,Y) & ac > 0 \\ -Corr(X,Y) & ac < 0 \end{cases}$$

$$-1 \le Corr(X,Y) \le 1$$

2. 두 확률변수 합의 분산

- Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)
- Var(X - Y) = Var(X) + Var(Y) - 2Cov(X,Y)

MetaCode

4

확률과 확률변수 : 예제

1. 한 동전을 두 번 던질 때 앞면이 나온 횟수를 X라 하고,
확률변수 $Y = (X+1)^2$ 라 할 때, 두 확률변수의 기댓값은?

2. $$Var(X) = \sigma^2 = E(X-\mu)^2 = E(X^2) - \mu^2$$

MetaCode

확률과 확률변수 : 예제

3. $Cov(X,Y) = E(XY) - E(X)E(Y)$

4 이산확률분포 : 이항분포

1. **베르누이 시행**

: 사상이 두 개뿐인 시행 (성공 or 실패)

- 각 시행에서 성공확률과 실패확률의 합은 1
- 각 시행은 서로 독립
- 베르누이 시행을 n번 독립 시행했을 때의 확률변수 x의 분포는 이항분포

x	0	1
f(x)	1 - p	p

- 이 때, 확률변수 X의 평균(기댓값) : p
- 확률변수 X의 분산 : p(1-p)

4 이산확률분포 : 이항분포

1. 이항확률분포

: 베르누이 시행을 반복하여
특정한 횟수의 성공/실패가 나타날 확률

2. 이상확률분포의 확률질량함수

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}={}_nC_x\,p^x(1-p)^{n-x} \qquad (0\le x\le n)$$

- n : 시행 횟수, x : 성공 횟수, p : 성공 확률

- 기댓값 : np
- 분산 : np(1-p)

MetaCode

4 이산확률분포 : 포아송분포

1. 포아송분포

: 단위시간, 단위공간 내 발생하는 사건의 횟수를
확률변수 X라고 할 때, X는 λ를 모수로 갖는 포아송분포 따름
: 발생빈도가 낮은 사건의 단위 당 발생 수

$$X \sim P(\lambda)$$

2. 포아송분포의 확률함수

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x=0,1,2\ldots \quad , 0 < \lambda < \infty$$

λ = 단위시간당 평균 발생 횟수

- 기댓값 : λ
- 분산 : λ

MetaCode

4 이산확률분포 : 예제

1. 확률변수 X가 베르누이분포를 따를 때,
   X의 분산이 p(1-p)임을 보여라

2. 주사위를 5번 던질 때, 4 이상의 눈이 두 번 나올 확률은?

MetaCode

4 이산확률분포 : 예제

3. 동전을 5번 던질 때, 앞면이 나온 횟수를 X라고 하자.
   이 때, X의 기댓값과 분산은?

4. 1000명의 보험가입자가 있을 때,
   한 해에 보험금을 청구할 확률이 1/2000이다.
   어떤 해에 보험금이 3회 청구될 확률은?

MetaCode

4 연속확률분포 : Uniform Distribution

1. Uniform Distribution

: 연속확률분포 중 가장 간단한 분포

2. 확률밀도함수

$$f(X) = \frac{1}{b-a} \quad a \le X \le b$$

- 기댓값 : (a+b)/2
- 분산 : $\frac{(b\_a)^2}{12}$

4 연속확률분포 : 정규분포

1. 정규분포 (가우스분포)

: 연속확률분포 중 가장 널리 사용
: 표본을 통한 통계적 추정 및 가설검정이론의 기본

2. 확률밀도함수

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

MetaCode

4 연속확률분포 : 정규분포의 특징

1. Bell Shaped : 평균을 중심으로 좌우 대칭의 종모양

2. 평균 = 중앙값 = 최빈값

3. 평균에 의해 분포의 위치가 결정

4. 표준편차에 의해 분포의 모양이 결정
   - 표준편차가 크면 평평한 곡선이 됨

5. 확률변수 X가 어느 구간에 속할 확률은
   그 구간과 분포함수로 이루어진 면적값

6. 이항분포와 포아송분포는 일정조건이 만족될 때 정규분포로 근사 가능
   - $np > 5$ and $n(1-p) > 5$
   - $\lambda > 5$

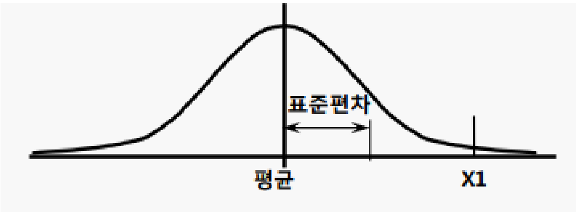
MetaCode

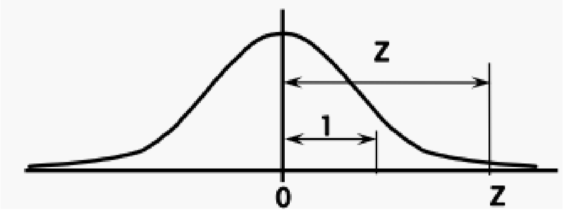
4 연속확률분포 : 표준정규분포

1. 표준정규분포

: 평균이 0이고 표준편차가 1인 정규분포
: Z분포로도 불림
: 정규분포를 따르는 X확률변수 X를 표준화

$Z = (X - \mu)/\sigma$ $Z \sim N(0,1)$





2. 표준정규분포의 확률밀도함수

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

MetaCode

4 연속확률분포 : 표본분포

1. **표본분포 (sampling distribution)**
   : 모집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때,
   그 모든 표본의 통계량의 확률분포

2. **표본평균의 평균과 표준편차**
   : $X_1, \cdots, X_n$이 모평균 μ, 모표준편차 σ인
   모집단으로부터의 확률표본 (i.i.d)일 때,

표본평균 : $\bar{X} = \frac{\sum X_i}{n}$

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}[E(X_1) + \cdots + E(X_n)] = \frac{1}{n}n\mu = \mu$$

$$\mathrm{Var}(\bar{\mathrm{X}}) = \frac{\sigma^2}{\mathrm{n}}$$

MetaCode

4 연속확률분포 : 중심극한정리

1. **중심극한정리**

: 평균이 μ, 표준편차 σ인 임의의 모집단으로부터
크기 n인 표본에서의 표본평균은 n이 크면 근사적으로
평균이 μ이고 분산이 $\frac{\sigma^2}{\mathrm{n}}$인 정규분포를 따름

: **모집단이 정규분포라면** 표본평균은 표본 개수와
상관없이 **항상 정규분포**를 따른다.

MetaCode

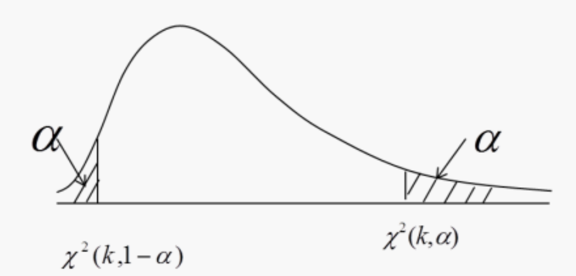
연속확률분포 : 카이제곱 분포

1. **카이제곱(χ^2) 분포**

: 표본분산과 관련된 분포
: 확률변수 $Z_1, \cdots, Z_k$가 각각 표준정규분포를 따르고 독립일 때 그들의 제곱합은 자유도 k인 카이제곱 분포 $\chi^2_{(k)}$를 따름

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2_{(k)}$$

: 표본분산을 알고 모분산을 추정할 때 사용하는 분포
(표본크기 클 수록 치우침이 적어짐)



4 연속확률분포 : 카이제곱 분포의 특징

1. 단봉분포
2. 오른쪽에 꼬리를 가짐
3. 항상 양수값을 가짐
4. 자유도가 커지면 정규분포에 가까워짐
5. 모분산 추정 및 검정에 활용
6. 적합성, 동질성, 독립성 검정 등에 사용

4 연속확률분포 : t분포

1. **t분포**

: X의 분포가 정규분포일 때, 표본평균의 분포에서
모집단의 표준편차를 모를 경우
모표준편차 대신 표본표준편차를 사용
: t분포는 자유도에 의해 모양이 결정됨
: Z ~ N(0,1), V ~ $\chi^2_{(k)}$ 이고 Z와 V는 서로 독립일 때,

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k)$$

: $X_1, \cdots, X_n$ ~ N(μ, σ^2)일 때,

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

MetaCode

4 연속확률분포 : t분포의 특성

1. T분포는 정규분포보다 넓게 퍼져 있고 꼬리부분이 더 평평함

2. Bell Shaped

3. 표본크기가 커질수록 분포가 중심부근에서 점점 더 뾰족해짐

   - 표본 크기가 30 이상이 되면 **정규분포에 근사**

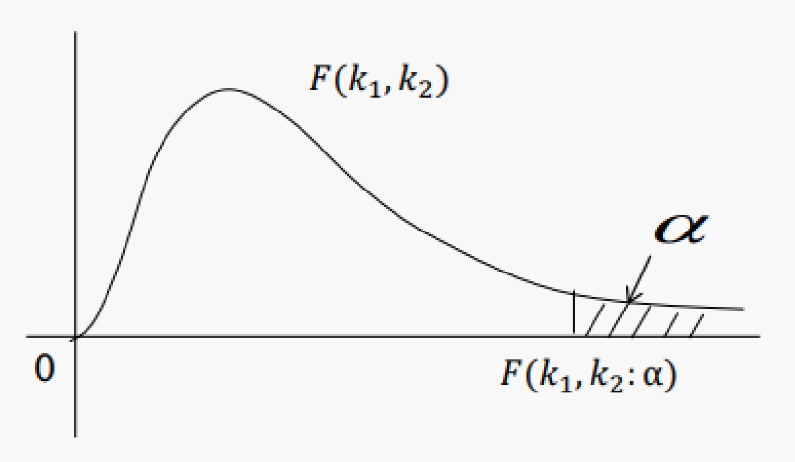
4. 주로 모평균 추정 혹은 모평균차이에 대한 추정 시
   모표준편차를 모를 때 t분포를 사용함

5. 표본 크기가 30 이상일 경우에는 표준정규분포, 미만일 때는 t분포

4 연속확률분포 : F-분포

1. **F분포**
   : F-분포는 두 정규모집단의 분산을 비교하는 추론에 사용
   : V1과 V2는 각각 자유도 k1, k2인 카이제곱분포를 따르는
   독립인 확률변수

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2) \quad F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$



MetaCode

4 연속확률분포 : 예제

1. X1, ⋯ Xn 이 모평균, 모분산 μ, σ^2인 모집단의 확률표본일 경우, 표본평균의 분산이 $\frac{\sigma^2}{n}$임을 보여라

2. 학생 100명의 성적 평균이 70점, 표준편차가 10점이다. 60~80점 사이의 성적을 받은 학생 수는? (단, 성적은 정규분포를 따르고, P(Z>1)=0.159)

4 연속확률분포 : 예제

3. X가 정규분포를 따를 때, $P(X<5)=0.5$, $P(X<10)=0.9$, $P(X<3)=0.3$ 이다. 이 때 X의 기댓값은?

4. 평균이 10, 표준편차가 0.4인 정규분포를 따르는 모집단에서 20의 표본을 임의로 추출한 경우 표본평균의 확률분포는?

4 연속확률분포 : 예제

5. $Z_i \sim N(0,1),\ i = 1, \dots, k$이고 각각 독립일 때, $\sum_{\{i=1\}} Z_i^2$의 분포는?

MetaCode

5 통계적 추정

1. **통계적 추정**

   : 표본의 통계량을 기초로 하여
   모집단의 모수를 추정하는 방법론

2. **통계적 추정의 종류**
   1. 점추정
      - 모수를 단일한 값으로 추측하는 방식
      - 신뢰도를 나타낼 수 없음
   2. 구간추정
      - 모수를 포함한다고 추정되는 구간을 구하는 방식
      - 신뢰도를 나타낼 수 있음

MetaCode

5 통계적 추정 : 기준

1. **불편성 (Unbiasedness)**
   : 모수의 추정량의 기댓값이 모수가 되는 성질

2. **유효성 (Efficiency)**
   : 추정량이 불편추정량이고 분산이
   다른 추정량에 비해 가장 작은 분산을 갖는 성질

3. **일치성 (Consistency)**
   : 표본 크기가 커질 수록 추정량이 모수에 수렴하는 성질

4. **충분성 (Sufficiency)**
   : 모수에 대해 가능한 많은 표본정보를 내포하는 성질

5 통계적 추정 : 점추정

1. **표준오차 (Standard Error)**
   : 통계량의 표준편차 $\sigma/\sqrt{\mathrm{n}}$
   : 표본크기가 클 수록 작아짐
   : 추정량의 표준편차가 작을 수록 좋음

2. **점 추정량**
   1. 모평균 : 표본평균
   2. 모분산 : 표본분산
   3. 모표준편차 : 표본표준편차
   4. 모비율 : 표본비율

MetaCode

5 통계적 추정 : 구간추정

1. **구간추정**

: 표본에서 얻어지는 정보를 이용하여 모수가 속할 것으로 기대되는 범위(신뢰구간)를 택하는 과정

: 통계적 추정은 일반적으로 신뢰구간의 추정을 활용

: 모수 θ에 대하여 P(a < θ < b) = 1 - α 일 때 구간 (a,b)을 모수 θ에 대한 100(1- α)% 신뢰구간이라고 한다.

2. **신뢰구간**

: 모수를 포함할 것으로 추정한 구간

3. **신뢰수준**

: 신뢰구간이 모수를 포함할 확률 (1 - α)   * α : 오차율

: 동일한 표본추출을 통해 구한 신뢰구간들 중 100 x (1- α) %는 모수를 포함

MetaCode

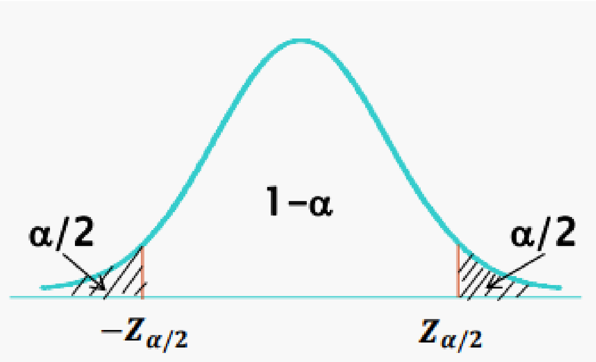
5 통계적 추정 : 모평균의 구간추정

1. 모분산을 아는 경우

가정) 모분산을 안다.

모집단의 평균이 μ, 분산이 σ^2인 정규분포

Z통계량을 사용



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

$$P\left\{\mu \in (\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})\right\} = 1-\alpha$$

- 90% 신뢰구간 : $Z_{0.05}$ = 1.64
- 95% 신뢰구간 : $Z_{0.025}$ = 1.96
- 99% 신뢰구간 : $Z_{0.005}$ = 2.57

5

통계적 추정 : 예제

예) 우리나라 대학생들의 월 평균 지출은 얼마일까.
100명을 랜덤 샘플링하여 조사한 결과 평균 30만원이고
모집단의 표준편차는 12만원이다. 90% 신뢰구간으로 모평균을
구간추정해보자.

풀이)

표준오차 : 12000

$$\bar{X} - Z_{0.05}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{0.05}\sigma_{\bar{X}}$$

$$\bar{X} - 1.64\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.64\sigma_{\bar{X}}$$

MetaCode

5 통계적 추정 : 모평균의 구간추정

1. **모분산을 모르는 경우**

가정) 모분산을 모른다.

모집단의 평균이 μ, 분산이 σ^2인 정규분포

t통계량을 사용

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

$$P\left\{-t(n-1,\frac{\alpha}{2}) \leq \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t(n-1,\frac{\alpha}{2})\right\} = 1-\alpha$$

- 표본 크기가 클 경우 Z통계량을 사용

5 통계적 추정 : 예제

예) 우리나라 대학생들의 월 평균 지출은 얼마일까.
16명을 랜덤 샘플링하여 조사한 결과 평균 30만원이고
표준편차는 10만원이다. 90% 신뢰구간으로 모평균을
구간추정해보자.

풀이)

$$\bar{X} - t_{0.05}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{0.05}\frac{S}{\sqrt{n}}$$

5

통계적 추정 : 예제

1. 20대 직장인의 1인당 월평균 지출액을 추정한다.
   10명을 랜덤 샘플링하고 표본 평균은 150만원이었다.
   모집단의 분포가 표준편차 10만원의 정규분포를 따를 때,
   모평균에 대한 95% 신뢰구간을 구하라.
   $(Z_{0.025} = 1.96, \quad Z_{0.05} = 1.645, \quad Z_{0.1} = 1.282)$

MetaCode

5 통계적 추정 : 예제

2. 고등학교 3학년 학생의 일평균 공부시간을 추정한다.
   500명을 랜덤 샘플링하고 조사 결과, 표본 평균은 4시간, 표준편차는 1시간이었다. 모집단의 일평균 공부시간에 대한 95% 신뢰구간은?
   $(Z_{0.025} = 1.96, \quad Z_{0.05} = 1.645, \quad Z_{0.1} = 1.282)$

MetaCode

통계적 추정 : 예제

3. 한 공장에서 생산한 돌고래 인형 5개를 추출해 무게를 측정하였다. 표본평균은 50g이었고, 표본표준편차는 10g이다. 이 공장의 돌고래 인형의 평균 무게 90% 신뢰구간은?
   ($Z_{0.025} = 1.96$, $Z_{0.05} = 1.645$, $Z_{0.1} = 1.282$, $\mathrm{t}_{5,0.05} = 2.015$, $t_{4,0.05} = 2.1318$)

5 통계검정 : 가설

1. 가설 검정

: 설정한 가설이 옳을 때 표본에서의 통계량과 통계량의 분포에서 이론적으로 얻는 특정 값을 비교하여 가설의 기각/채택 여부를 판정하는 방법

: 확률적 오차 범위를 넘어서면 가설을 기각한다.
: 유의수준(α) : 기각/채택 여부의 판단기준

2. 가설의 종류

: 귀무가설 (H0)
- 대립가설과 상반되는 가설로, 일반적인 사실을 귀무가설로 설정
- 효과가 없다, 차이가 없다 등의 내용

: 대립가설 (H1)
- 입증하고자 하는 가설
- 효과가 있다, 차이가 있다 등의 내용

MetaCode

5 통계검정 : 오류

1. 가설설정의 오류

- 제1종 오류 (α)
  : 귀무가설을 채택해야 했음에도 이를 기각할 오류
  : 표본으로부터 얻은 검정결과가 우연에 의해 잘못 판단되었을 가능성
  : α는 일반적으로 5%로 설정

- 제2종 오류 (β)
  : 귀무가설을 기각해야 했음에도 이를 채택할 오류
  : 실제로는 효과가 없는데 효과가 있다고 잘못 결론 내릴 가능성
  : β는 일반적으로 10%로 설정

MetaCode

통계적 검정 : 요소

1. **유의수준 (significance level)**
   - 제1종 오류를 범할 확률의 최대 허용한계

2. **유의확률 (p-value)**
   - 검정통계량 값에 대해 귀무가설을 기각할 수 있는 최소의 유의수준으로 귀무가설이 사실일 확률
   - $\alpha >$ p-value : 귀무가설 기각
   - $\alpha <$ p-value : 귀무가설 채택

3. **임계값 (critical value)**
   - 기각역과 채택역을 나누는 경계값
   - 기각역 : 귀무가설을 기각하게 되는 검정통계량의 관측값의 영역
   - 채택역 : 귀무가설을 채택하게 되는 검정통계량의 관측값의 영역
   - 검정통계량의 관측값이 기각역에 속하면 귀무가설 기각

5

통계검정 : 절차

1. 검정할 가설을 설정

2. 유의수준을 설정

3. 임계치를 결정하고 검정통계량과 임계치를 비교
   (혹은 유의수준과 유의확률 비교)

4. P-value값이 유의수준보다 작으면 귀무가설을 기각

MetaCode

5

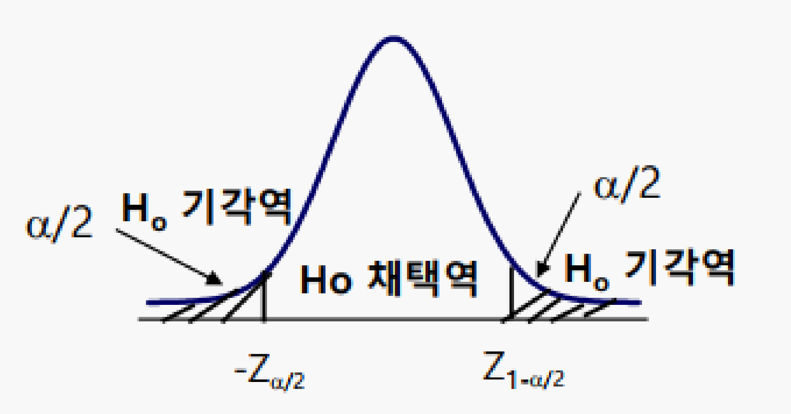
통계검정 : 양측검정과 단측검정

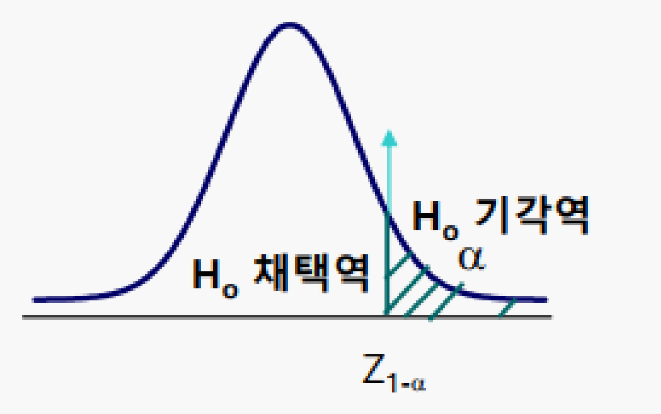
1. 양측검정 (Two-sided)

- 기각역이 각각 왼쪽과 오른쪽 두 부분으로 구성된 가설검정
- 양쪽 기각역의 합 = 유의수준

2. 단측검정 (One-sided)

- 기각역이 한쪽으로만 구성되는 가설검정
- 한쪽 기각역이 유의수준





5 통계검정 : 모평균 검정

1. 정규모집단의 경우

1. 모분산이 알려진 경우
   : Z 검정 통계량

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

2. 모분산을 모르는 경우
   : t 검정 통계량 (자유도 n-1)

$$t' = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

2. 표본 크기가 큰 임의의 모집단

1. 모분산이 알려진 경우
   : Z 검정 통계량

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

2. 모분산을 모르는 경우
   : Z 검정 통계량

$$Z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

5 통계검정 : 예제

1. 모 초등학교 4학년 학생의 평균 키가 150cm라고
   알려져 있는데 실제는 이와 다른지 검사하고자 학생 25명을
   추출하였고 표본 평균은 148cm가 나왔다. 평균 키는 정규분포를
   따르고 모 표준편차는 10cm로 알려져 있다.
   유의수준 5%로 검정해보자.

MetaCode

5 통계검정 : 예제

2. 고등학교 3학년 학생의 일평균 공부시간이 4시간, 표준편차가 1시간인 정규분포를 따를 때, 이를 검정하기 위해 30명의 학생을 랜덤 샘플링하여 표본평균이 3.5시간이다. 이를 통해 고3 학생의 일평균 공부시간이 4시간보다 작다고 할 수 있을까?

$(P(|Z| < 1.645) = 0.9, P(|Z| < 1.96) = 0.95)$

MetaCode