2과목.빅데이터 탐색

(Ch_03. 통계기법의 이해 - SEC 01. 기술통계)

빅데이터 분석 기사(2과목. 빅데이터 탐색)

CHAPTER 1. 데이터 전처리

CHAPTER 2. 데이터 탐색

CHAPTER 3. 통계 기법 이해

통계기법의 이해

통계기법의 이해 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 기술통계

2. 추론통계

2. 통계기법의 이해 – 기술통계

기술통계

● 기술통계(Descriptive Statistics)란 데이터 분석의 초기 단계에서 데이터 분포의 특징을 파악하기 위해 사용되는 통계기법이다.

● 수집된 데이터의 전수조사가 어려운 경우 데이터의 특징을 담고 있는 표본 데이터를 추출하기 위해 기술통계 작업을 수행하며, 이러한 작업을 통해 데이터에 대한 정확한 이해가 가능하게 된다.

● 기술통계의 '기술'은 Technology가 아닌 Descriptive(기술하는, 서술하는)임을 기억하도록 한다.

01 데이터 요약

1) 대푯값

; 대푯값은 주어진 데이터를 대표할 수 있는 값으로 중위수, 평균값, 최빈수, 사분위수가 있다.

① 중위수(Median)

- ▶ 중위수는 모든 데이터를 오름차순으로 정렬했을 때 가장 중앙에 취한 데이터 값을 의미한다.
- ▶ 중위수는 이상치에 영향을 받지 않는다.
- ▶ 중위수의 개수가 짝수인 경우 중앙에 있는 두 개의 값의 평균을 중위수로 정한다.

$$d_{median} = \frac{n+1}{2}\text{번째 값 } (n : \text{데이터 개수})$$

2. 통계기법의 이해 – 기술통계

1) 대푯값

② 평균값(Average)

- ▶ 평균값은 주어진 데이터를 모두 더한 후 데이터의 개수만큼 나눈 값을 의미한다.
- ▶ 평균값은 모두 같은 가중치를 두며, 이상값에 민감하다.
- ▶ 평균의 종류에는 산술평균, 기하평균, 조화평균이 있다.

산술평균 (Arithmetic mean)	$A=\frac{1}{n}\sum_{i=1}^{n}a_i=\frac{a_1+a_2+\cdots+a_n}{n}$ a_i : i번째 데이터 n : 대상 데이터 수	대상 데이터 n개의 합의 평균
조화평균 (harmonic mean)	$H=\frac{n}{\frac{1}{a_1}+\frac{1}{a_2}+\cdots+\frac{1}{a_n}}$ a_i : i번째 데이터 n : 대상 데이터 수 두 수가 주어진 경우 조화평균 $H=\frac{2(a_1a_2)}{a_1+a_2}$ a_1 : 첫 번째 데이터 a_2 : 두 번째 데이터	• 역수의 산술평균의 역수 • 역수 차원에서 산술평균을 구하고, 다시 역수를 취해 원래의 차원의 값으로 돌아오는 것 • 주로 다른 두 속력의 평균을 구하는 데 사용된다.

기하평균 (Geometic mean)	$\left(\prod_{i=1}^{n}a_i\right)^{\frac{1}{n}}=\sqrt[n]{a_1a_2a_3\cdots a_n}$ a_i : i번째 데이터 n : 대상 데이터 수	• 대상 데이터 n개의 양수 곱의 n제곱근 • 물가상승률, 성장률 등의 평균값 연산에 활용 예 1, 2, 3의 기하평균을 구할 경우 $1\times2\times3=6$을 구한 뒤 3제곱근을 취한 값 $6^{\frac{1}{3}}=\sqrt[3]{6}$이 된다.

2. 통계기법의 이해 – 기술통계

1) 대푯값

② 평균값(Average)

▶ 평균은 대상 범위에 따라 모평균, 표본평균으로 나뉜다.

구분	수식	설명
모평균	$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ X_i : i번째 데이터 N : 모집단 데이터 수	모집단의 데이터가 N개일 때 X_1, X_2, X_3, …, X_n에 대한 평균
표본평균	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ X_i : i번째 데이터 n : 표본집단 데이터 수	표본의 데이터가 n개일 때 X_1, X_2, X_3, …, X_n에 대한 평균

③ 최빈수(Mode)

▶ 최빈수는 데이터 값 중에서 가장 빈도수가 높은 데이터를 의미한다. 즉, 가장 여러 번 관측된 데이터라고 할 수 있다.

④ 사분위수(Quartile)

▶ 사분위수는 모든 데이터를 순서대로 배열했을 때, 4등분한 지점에 있는 값을 의미한다.

2. 통계기법의 이해 – 기술통계

2) 산포도

; 산포도는 데이터의 흩어진 정도를 나타내는 값이다. 산포도를 나타내는 값으로 분산, 표준편차, 범위, IQR, 사분편차, 변동계수가 있다.

① 분산(Variance)

▶ 분산은 데이터가 평균으로부터 얼마나 떨어져 있는지 나타내는 값을 의미한다.

▶ 편차의 제곱의 평균값으로 관측값에서 평균을 뺀 값인 편차를 모두 더하면 0이 나오기 때문에 제곱해서 더한다.

▶ 단, 편차의 제곱이기 때문에 본래의 데이터보다 큰 값으로 표현된다.

구분	수식
모분산	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ (μ : 모평균, X_i : i번째 데이터, N : 데이터의 수)
표본분산	$s^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$ ($\overline{X}$: 표본평균, X_i : i번째 데이터, n : 데이터의 수)

2. 통계기법의 이해 – 기술통계

2) 산포도

② 표준편차(Standard Deviation)

▶ 표준편차는 분산에 양의 제곱근을 취한 값을 의미한다.

▶ 본래의 데이터와 동일한 단위로 데이터를 분석할 수 있어서 데이터가 커지는 분산의 단점을 보완할 수 있다.

구분	수식
모표준편차	$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$ (μ : 모평균, X_i : i번째 데이터, N : 데이터의 수)
표본 표준편차	$s=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$ ($\overline{X}$: 표본평균, X_i : i번째 데이터, n : 데이터의 수)

2. 통계기법의 이해 – 기술통계

2) 산포도

③ 범위(Range)

▶ 범위는 데이터의 최댓값과 최솟값의 차이를 의미한다.

$$R = X_{max} - X_{min} \ (X_{max} : \text{데이터 최댓값}, X_{min} : \text{데이터 최솟값})$$

④ IQR(사분범위, 사분위수범위)

▶ IQR(InterQuartile Range)은 제3사분위수와 제1사분위수의 차이값을 의미한다.

$$IQR = Q_3 - Q_1 \ (Q_3 : \text{제3사분위수}, Q_1 : \text{제1사분위수})$$

⑤ 사분편차(Quartile Deviation)

▶ 사분편차는 제3사분위수와 제1사분위수의 차인 IQR의 절반 값을 의미한다.

$$\text{사분편차} = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2} \ (Q_3 : \text{제3사분위수}, Q_1 : \text{제1사분위수})$$

2. 통계기법의 이해 – 기술통계

2) 산포도

⑥ 변동계수(CV, 변이계수, 상대표준편차)

- ▶ 변동계수(Coefficient of Variation)는 표준편차를 평균으로 나눈 값을 의미한다.
- ▶ 변동계수는 측정 단위가 다른 데이터의 산포도를 상대적으로 비교할 때 사용된다.

구분	수식
모집단	$CV=\frac{\sigma}{\mu}$ (σ : 모표준편차, μ : 모평균)
표본집단	$CV=\frac{s}{\overline{X}}$ (s : 표본표준편차, $\overline{X}$: 표본평균)

2. 통계기법의 이해 – 기술통계

3) 데이터 분포

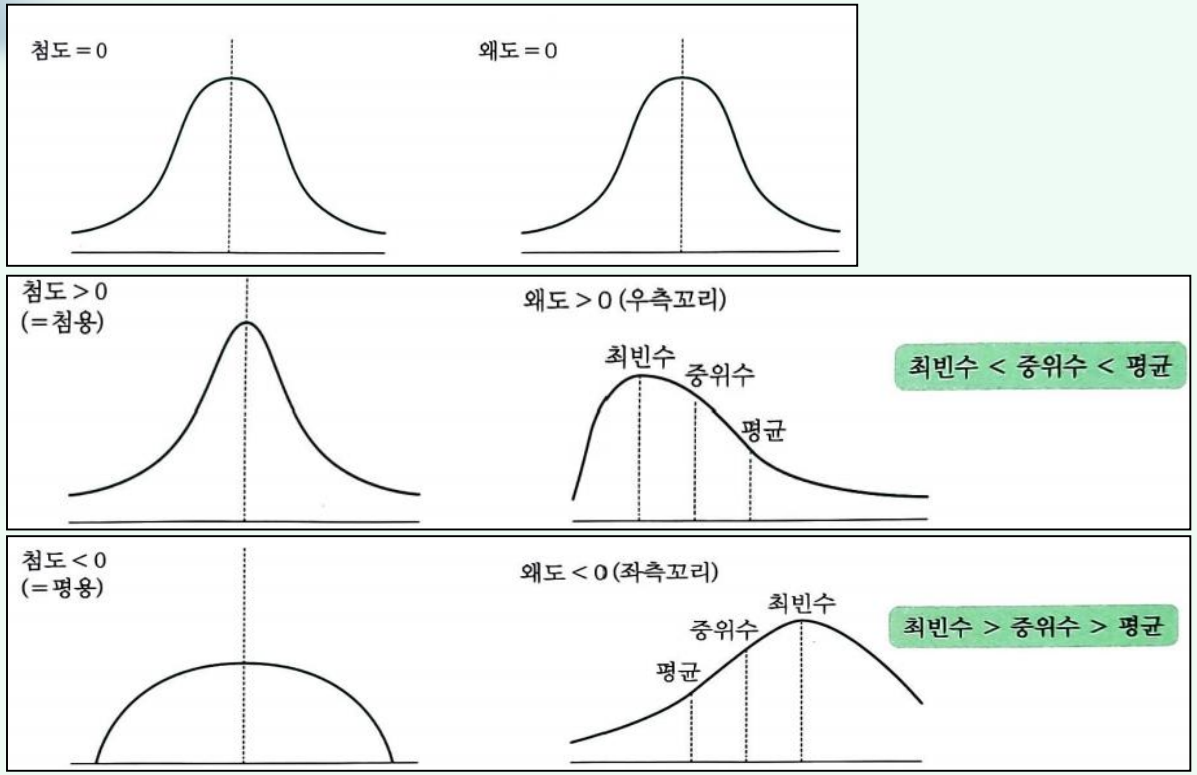
; 데이터 분포를 표현하는 통계량에는 첨도와 왜도가 있다.

① 첨도(Kurtosis)

▶ 데이터 분포의 뾰족한 정도를 나타내는 통계량이다.

② 왜도(Skewness)

▶ 데이터 분포의 기울어진 정도를 나타내는 통계량이다.



2. 통계기법의 이해 – 기술통계

4) 공분산(Covariance)

- 공분산은 2개의 변수 사이의 연관성을 나타내는 통계량을 의미한다.
- 공분산으로 상관관계의 상승 또는 하강 경향을 이해할 수 있으나 선형 관계의 강도를 나타내지는 못한다.
- 공분산의 종류에는 모공분산, 표본공분산이 있다.

모공분산	$Cov(X, Y)=\sigma_{XY}=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)$ (μ_X : X 모집단 평균, μ_Y : Y 모집단 평균)
표본공분산	$S_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$ ($\overline{X}$: X 표본집단의 평균, $\overline{Y}$: Y 표본집단의 평균)

- 공분산 해석 : 공분산은 다음과 같이 해석할 수 있다.

공분산 값	내용
($Cov>0$)	2개의 변수 중 하나의 값이 상승할 때 다른 하나의 값도 상승하는 경우 공분산은 양수가 된다.
($Cov<0$)	2개의 변수 중 하나의 값이 상승할 때 다른 하나의 값은 하강하는 경우 공분산은 음수가 된다.

2. 통계기법의 이해 – 기술통계

5) 상관관계(Correlation)

- 두 변수 사이에 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 방법으로, 상관관계로 인과관계는 알 수 없다.
- 공분산은 선형관계의 강도를 나타낼 수 없지만, 상관계수는 선형관계의 강도를 나타낼 수 있다.
- 상관계수는 -1 ~ 1 사이의 값을 가지며 1에 가까울수록 강한 양(+)의 상관관계를, -1에 가까울수록 강한 음(-)의 상관관계를 가진다.
- 상관관계 분석 방법에는 피어슨 상관계수, 스피어만 상관계수, 카이제곱 검정이 있다.

2. 통계기법의 이해 – 기술통계

개념 체크

01 다음 중 기술통계에 대한 설명으로 옳지 않은 것은?

① 기술통계는 데이터 분석 초기 단계에서 데이터 분포의 특징을 확인하기 위해 사용된다.

② 데이터의 전수조사가 어려운 경우 표본 데이터를 추출하여 분석한다.

③ 기술통계의 기술은 'Technology'를 의미한다.

④ 기술통계를 통해 데이터에 대한 정확한 이해가 가능하게 된다.

기술통계

▶ 기술통계(Descriptive Statistics)란 데이터 분석의 초기 단계에서 데이터 분포의 특징을 파악하기 위해 사용되는 통계 기법이다.

▶ 수집된 데이터의 전수조사가 어려운 경우 데이터의 특징을 담고 있는 표본 데이터를 추출하기 위해 기술통계 작업을 수행하며, 이러한 작업을 통해 데이터에 대한 정확한 이해가 가능하다.

▶ 기술통계의 '기술'은 Technology가 아니라

02 다음 중 평균값에 대한 설명으로 옳지 않은 것은?

① 평균값은 주어진 데이터를 모두 더한 뒤, 데이터의 개수만큼 나눈 값을 의미한다.

② 평균값은 이상치에 영향을 받지 않는다.

③ 평균값의 종류로 모평균과 표본평균이 있다.

④ 모평균의 수식은 $\mu = \frac{1}{N}\sum_{i=1}^{N} Xi$ 와 같다.

평균값(Average)

▶ 평균값은 주어진 데이터를 모두 더한 뒤, 데이터의 개수만큼 나눈 값을 의미한다.

▶ 평균값은 모두 같은 가중치를 두며, 이상값에 민감하다.

▶ 평균의 종류에는 산술평균, 기하평균, 조화평균이 있다.

▶ 평균의 대상 범위에 따라 모평균, 표본평균이 있다.

2. 통계기법의 이해 – 기술통계

03 다음과 같은 데이터에서 중위수는 얼마인가?

13, 5, 3, 2, 6, 10, 1, 20, 8, 11

① 5 ② 8

③ 6 ④ 7

주어진 데이터를 오름차순으로 정렬을 먼저 한다.

1, 2, 3, 5, 6, 8, 10, 11, 13, 20

다만 위와 같이 주어진 데이터의 개수가 10개(짝수)이므로

중앙의 두 수(6, 8)를 더한 값의 평균이 중위수가 된다.

(6 + 8) / 2 = 7, 7이 중위수이다.

04 다음 중 대푯값에 대한 설명이 바르지 않는 것은?

① 대푯값에는 평균값, 중위수, 사분위수, 범위가 있다.

② 표본평균의 수식은 $\bar{X} = \frac{1}{N}\sum_{i=1}^{n} Xi$ 와 같다.

③ 중위수 수식은 $\mathrm{d}_{\mathrm{median}}\ \frac{n+1}{2}$ 번째 값$(n$: 데이터 개수$)$와 같다.

④ 사분위수는 모든 데이터를 순서대로 배열했을 때, 4등분한 지점에 있는 값을 의미한다.

대푯값에는 평균값, 중위수, 최빈수, 사분위수가 있다.

범위(Range)는 산포도에 속한다.

2. 통계기법의 이해 – 기술통계

05 다음 중 산포도에 속하지 않는 것은?

① 분산

② 최빈수

③ 표준편차

④ IQR

산포도에는 분산, 표준편차, 범위, IQR, 사분편차, 변동계수가 있다. 최빈수는 대푯값에 속한다.

06 다음 중 분산에 대한 설명이 옳은 것은?

① 분산은 데이터가 커지는 단점을 보완할 수 있다.

② 분산의 수식은 $s = \sqrt{\frac{\sum_{i=1}^{N}(Xi - \bar{X})^2}{n-1}}$와 같다.

③ 분산은 데이터의 흩어진 정도를 나타낸다.

④ 분산은 표준편차에 양의 제곱근을 취한 것이다.

1번과 2번은 표준편차에 대한 설명이다. 표준편차는 분산에 양의 제곱근을 취한 값이다.

분산(Variance)

▶ 분산은 데이터가 평균으로부터 얼마나 떨어져 있는지를 나타내는 값을 의미한다.

▶ 편차의 제곱의 평균값으로 관측값에서 평균을 뺀 값인 편차를 모두 더하면 0이 나오기 때문에 제곱해서 더한다.

▶ 단, 편차의 제곱이기 때문에 본래의 데이터보다 큰 값으로 표현된다.

2. 통계기법의 이해 – 기술통계

07 다음 중 IQR에 대한 설명으로 옳은 것은?

① IQR은 사분범위, 사분위수범위와 같은 말이다.

② IQR은 제4사분위수와 제2사분위수의 차이값을 의미한다.

③ IQR은 시각화 도구인 막대그래프에서 확인할 수 있다.

④ 사분편차는 IQR X 2로 표현된다.

2번 같은 경우는 제3사분위수와 제1사분위수의 차이값을 의미하는 것이 바로 IQR이다.

3번 같은 경우는 시각화 도구인 상자수염그림(Box-Plot)에서 확인할 수 있다.

4번 같은 경우는 사분편차 IQR / 2이다.

▶ IQR(Inter Quartile Range)은 제3사분위수와 제1사분위수의 차이값을 의미한다.

IQR = Q3 - Q1

사분편차(Quartile Deviation)

▶ 사분편차는 IQR의 절반 값을 의미한다.

사분편차 = IQR / 2

08 다음 중 데이터 분포에 대한 설명으로 옳지 않은 것은?

① 데이터 분포는 첨도와 왜도로 나뉜다.

② 첨도는 데이터의 뾰족한 정도를 나타낸다.

③ 왜도는 데이터의 기울어진 정도를 나타낸다.

④ 왜도 > 0일 때, 데이터는 좌측 꼬리 모형을 갖는다.

왜도 > 0 일 때, 데이터는 우측 꼬리 모형을 갖는다.

2. 통계기법의 이해 – 기술통계

09 다음 중 왜도 < 0일 때, 올바른 배열은?

① 최빈수 < 중위수 < 평균

② 최빈수 > 중위수 > 평균

③ 최빈수 < 중위수 = 평균

④ 최빈수 = 중위수 < 평균

왜도 < 0 일 때, 최빈수 > 중위수 > 평균과 같은 분포를 갖는다.

1번 같은 왜도 > 0일 때에 대한 설명이다.

10 다음 중 공분산에 대한 설명으로 옳지 않은 것은?

① 공분산은 2개의 변수 사이의 연관성을 나타내는 통계량이다.

② 공분산으로 상승 또는 하강 관계를 이해할 수 있다.

③ 공분산으로 선형관계에 대한 강도를 나타낼 수 있다.

④ 공분산은 모공분산과 표본공분산으로 나뉜다.

공분산으로 상관관계의 상승 또는 하강 경향을 이해할 수 있으나 선형관계의 강도를 나타내지는 못한다.

공분산(Covariance)

- 공분산은 2개의 변수 사이에 연관성을 나타내는 통계량을 의미한다.
- 공분산으로 상관관계의 상승 또는 하강 경향을 이해할 수 있으나 선형 관계의 강도를 나타내지는 못한다.
- 공분산의 종류에는 모공분산, 표본공분산이 있다.

2. 통계기법의 이해 – 기술통계

11. 다음 중 상관관계에 대한 설명으로 틀린 것은?

① 두 변수 사이에 어떤 선형적 혹은 비선형적 관계가 있는지 분석하는 방법이다.

② 상관관계의 수치를 나타내는 상관계수는 -1 ~ 1의 범위 를 갖는다.

③ 상관계수가 1에 가까울수록 강한 양(+)의 상관관계를 갖는다고 할 수 있다.

④ 상관관계로 두 변수 사이의 인과관계를 알 수 있다.

상관관계는 두 변수 사이에 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 방법으로 상관관계로 인과관계는 알 수 없다.

상관관계(Correlation)

● 두 변수 사이에 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 방법으로 상관관계로 인과관계는 알 수 없다.

● 공분산은 선형관계의 강도를 나타낼 수 없지만, 상관계수는 선형관계의 강도를 나타낼 수 있다.

● 상관계수는 -1 ~ +1 사이의 값을 가지면 1에 가까울수록 강한 양(+)의 상관관계를, -1에 가까울수록 강한 음(-)의

2. 통계기법의 이해 – 기술통계

02 표본추출

1) 표본추출의 정의

; 표본추출(Sampling)은 모집단의 일부를 정해진 규칙에 따라 표본으로 추출하는 것을 의미한다.

2) 표본추출 기법

; 표본추출 기법에는 단순무작위추출, 계통추출, 층화추출, 군집추출이 있다.

① 단순무작위추출(Simple Random Sampling)

- ▶ 모집단에서 정해진 규칙 없이 표본을 추출하는 방식이다.
- ▶ 표본의 크기가 커질수록 정확도가 높아지고, 추정값의 분산이 작아진다.

② 계통추출(Systematic Sampling)

- ▶ 모집단을 일정한 간격 및 구간으로 추출하는 방식이다.

예) 10명에게 번호표를 나눠주고, 짝수 번호인 사람 선정

③ 층화추출(Stratified Sampling)

- ▶ 모집단을 여러 계층으로 나누고, 계층별로 무작위 추출하는 방식이다.
- ▶ 데이터 특징이 층 내에서는 동질하고, 층간에서는 이질한 특징이 있다.

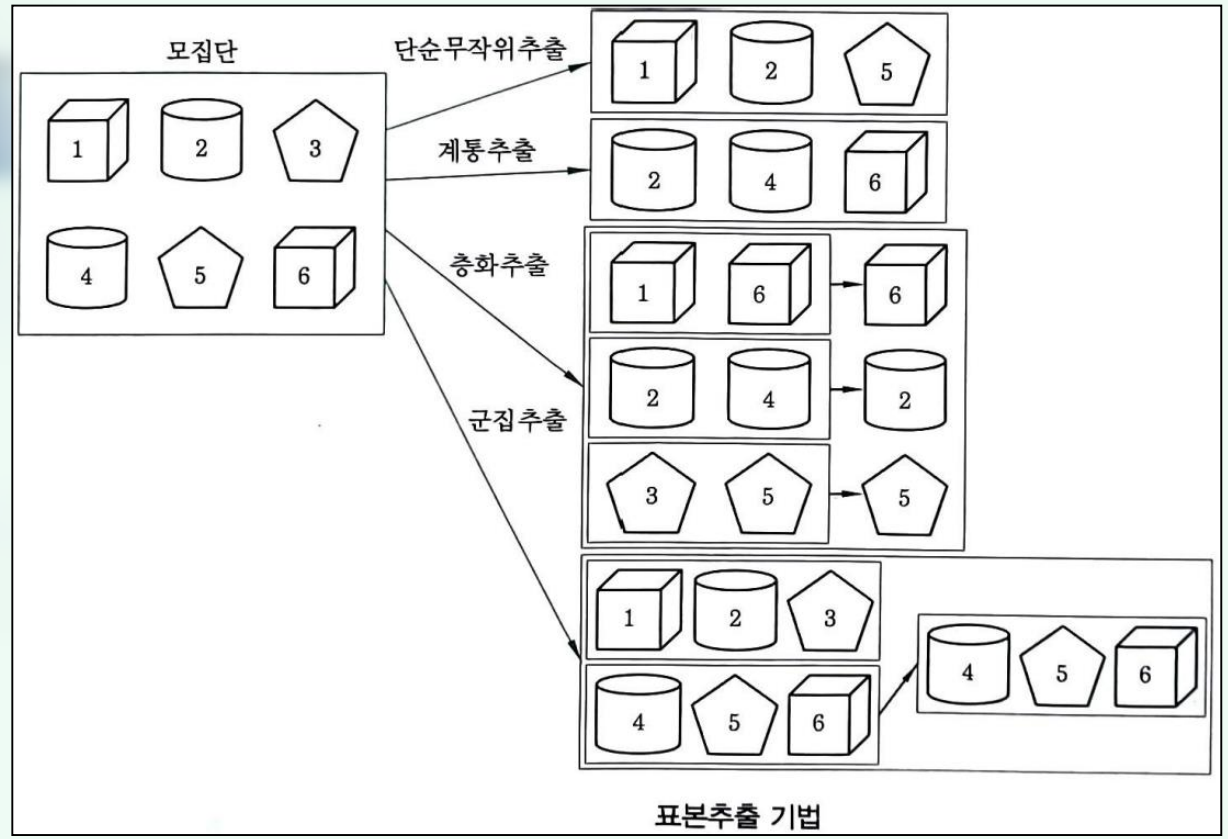
모집단(population) : 정보를 얻고자 하는 관심 대상의 전체 집합

2. 통계기법의 이해 – 기술통계

2) 표본추출 기법

④ 군집추출(Cluster Random Sampling)

- ▶ 모집단을 여러 군집으로 나누고 일부 군집의 전체를 추출하는 방식이다.
- ▶ 데이터 특징이 집단 내에서는 이질적이고, 집단 외에서는 동질한 특징이 있다.



표본추출 기법

2. 통계기법의 이해 – 기술통계

03 확률분포

1) 확률의 개념

; 확률(Probability)이란 어떠한 사건이 발생할 가능성을 의미하며, 0~1 범위의 수로 표현된다.

2) 조건부 확률

; 조건부 확률은 어떤 사건이 일어난다는 조건에서 다른 사건이 일어날 확률을 의미한다.

사건 A가 조건으로 일어날 때 사건 B가 발생할 확률

$$P(B|A)=\frac{P(A\cap B)}{P(A)}$$

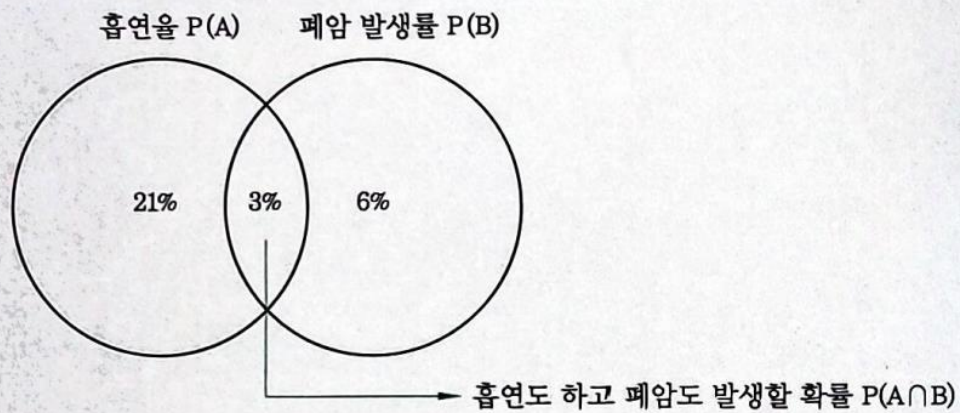
사건 B가 조건으로 일어날 때 사건 A가 발생할 확률

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

2. 통계기법의 이해 – 기술통계

2) 조건부 확률

예시



흡연할 때 폐암에 걸릴 확률 : $P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{3}{24}=12.5\%$

폐암에 걸렸을 때 흡연할 확률 : $P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{3}{9}=33.3\%$

2. 통계기법의 이해 – 기술통계

3) 베이즈 정리(Bayes' Theorem)

- 베이즈 정리는 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 설명하는 확률이론으로 B가 발생할 때, A가 발생할 확률을 의미한다.
- 어떤 사건이 서로 배반(俳反)하는 원인 둘에 의해 일어난다고 할 때 실제 사건이 일어났을 때 이것이 두 원인 중 하나일 확률을 구하는 정리이다.

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

예시) 한 회사에서 A공장은 부품을 50% 생산하고 불량률은 1%, B공장은 부품을 30% 생산하고 불량률 2%, C공장은 부품을 20% 생산하고 불량률이 3%이다. 부품을 선택했을 때 C공장에서 생산한 불량 부품일 확률을 구하시오.

A_1 : A공장, A_2 : B공장, A_3 : C공장, B : 불량률

$P(A_1)$: A공장 부품 생산율 50%, $P(B|A_1)$: A공장 불량률 1%
$P(A_2)$: B공장 부품 생산율 30%, $P(B|A_2)$: B공장 불량률 2%
$P(A_3)$: C공장 부품 생산율 20%, $P(B|A_3)$: C공장 불량률 3%

$P(A_3|B)$: 불량품이 C공장에서 생산될 확률

$$=\frac{P(A_3)P(B|A_3)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+P(A_3)P(B|A_3)}$$

$$=\frac{20\%\times 3\%}{(50\%\times 1\%)+(30\%\times 2\%)+(20\%\times 3\%)}=\frac{60}{50+60+60}=\frac{6}{17}$$

배반 : 두 개의 사건이 동시에 일어날 수 없는 경우

2. 통계기법의 이해 – 기술통계

4) 확률분포(Probability Distribution)

● 확률분포는 확률변수가 특정한 값을 가질 확률을 나타내는 분포이다. 확률변수의 종류에 따라 이산확률분포와 연속확률분포로 나뉜다.

① 이산확률분포(Discrete Probability Distribution)

▶ 이산확률분포는 이산확률변수 X가 갖는 확률분포를 나타낸다.

▶ 이산확률변수는 확률변수 X 가 0, 1, 2, 3, ... 과 같이 하나씩 셀 수 있는 값을 갖는다.

▶ 이산확률분포의 종류에는 푸아송 분포, 베르누이 분포, 이항 분포가 있다.

종류	설명
푸아송 분포	주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률분포 $P=\frac{\lambda^n e^{-\lambda}}{n!}$ (λ : 평균, n : 발생 횟수)
베르누이 분포	특정 실험의 결과가 성공 또는 실패로 두 가지 중 하나의 결과를 얻는 확률분포
이항 분포	n번 시행 중에 각 시행의 확률이 P일 때, k번 성공할 확률분포 $P=\binom{n}{k}P^k(1-P)^{n-k}$ (n : 시행 횟수, P : 특정 사건이 성공할 확률, k : 성공 횟수)

이산확률변수 : 셀 수 있는 확률변수

2. 통계기법의 이해 – 기술통계

4) 확률분포(Probability Distribution)

② 연속확률분포(Continuous Probability Distribution)

- ▶ 연속확률분포는 연속확률변수 X 가 갖는 확률분포를 나타낸다.
- ▶ 연속확률분포의 종류에는 정규분포, 표준정규분포(Z-분포), T -분포, x^2 분포(카이제곱 분포), F -분포, 지수 분포, 감마 분포가 있다.

㉮ 정규분포 : 분포 곡선이 평균값을 중심으로 좌우 대칭한 종 모양의 분포로 가우스 분포라고도 표현한다.

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

(σ^2 : 모분산, μ : 모평균, x : 확률변수, e : 자연상수(2.718…))

㉯ 표준정규분(Z -분포) : 정규분포에서 x를 Z로 정규화한 분포로 평균이 0이고, 분산이 1인 정규분포이다.

$$Z=\frac{\overline{X}-\mu}{\sigma}$$ (평균 0, 분산 1)

(σ : 모표준편차, μ : 모평균, $\overline{X}$: 표본평균)

연속확률변수 : 연속적인 구간 내의 실수값을 가진 확률변수

2. 통계기법의 이해 – 기술통계

4) 확률분포(Probability Distribution)

② 연속확률분포(Continuous Probability Distribution)

㉰ T-분포

- ▶ 모집단이 정규분포라는 정도만 알고 모표준편차(σ)는 모를 때, 모집단의 평균을 추정하기 위해 사용하는 분포
- ▶ 표본의 크기가 작은 경우 사용하고, 중심극한정리에 의해 Z-분포는 정규분포를 따른다.
- ▶ 두 집단의 평균이 동일한지 확인하고자 할 때 검정통계량으로 사용된다.

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n-1}}$$

(s : 표본표준편차, μ : 모평균, $\overline{X}$: 표본평균, n : 자유도(표본의 개수))

자유도(degree of freedom) : 정보의 수에서 추정된 매개변수를 뺀 것으로 정보의 수가 n일 경우 자유도는 n - 1의 값을 가짐.

2. 통계기법의 이해 – 기술통계

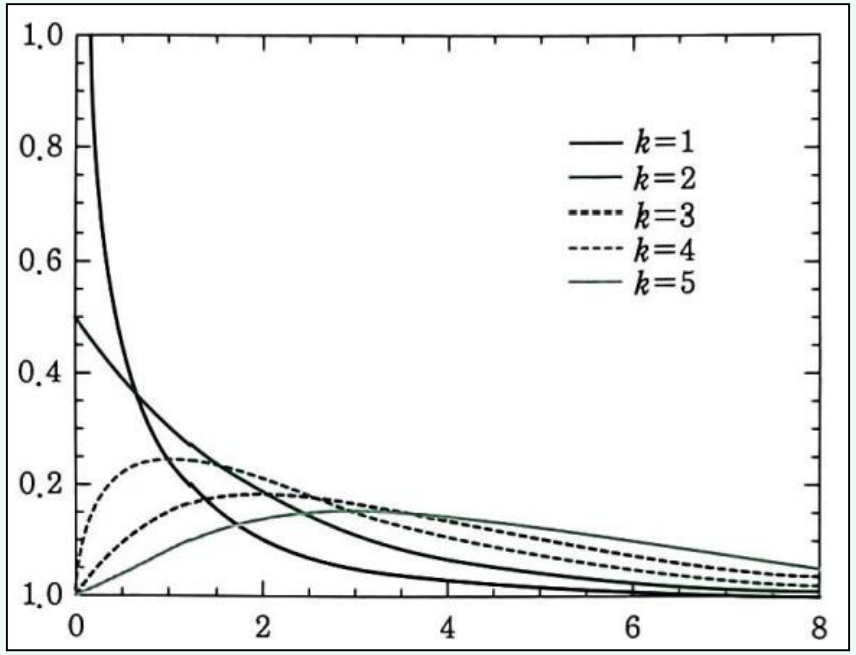
4) 확률분포(Probability Distribution)

② 연속확률분포(Continuous Probability Distribution)

㉮ x^2분포(카이제곱 분포)

- ▶ k(자유도)개의 서로 독립적인 표준정규확률변수를 각각 제곱한 다음 합해서 얻는 분포
- ▶ 카이제곱 분포는 신뢰구간이나 가설 검정 등의 모델에서 자주 활용된다.

$$\chi^2 = Z_1^2 + Z_1^2 + \cdots + Z_k^2$$



2. 통계기법의 이해 – 기술통계

4) 확률분포(Probability Distribution)

② 연속확률분포(Continuous Probability Distribution)

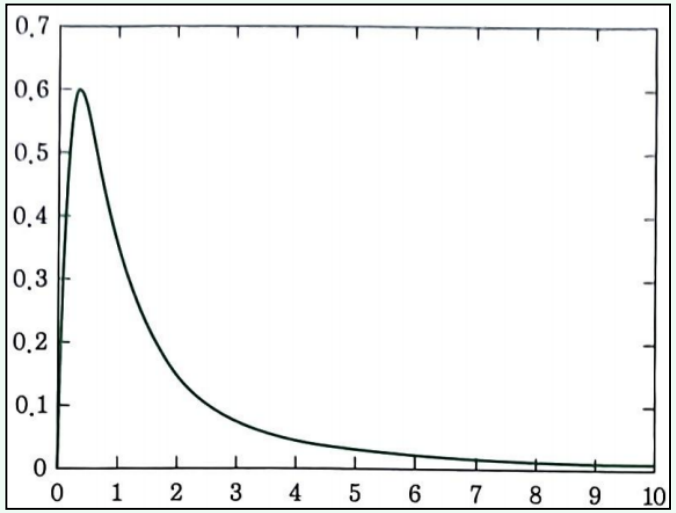
㉮ $F-$분포

▶ 독립적인 x^2분포(카이제곱 분포)가 있을 때, 두 확률변수의 비

▶ 모집단 분산이 서로 동일하다고 가정되는 두 모집단으로부터 표본 크기가 각각 n_1, n_2인 독립적인 2개의 표본을 추출했을 때, 2개의 표본분산 s_1^2, s_2^2율 $\left(\frac{s_1^2}{s_2^2}\right)$

$$F=\frac{s_1^2}{s_2^2}$$

(s_1^2 : 첫 번째 집단의 표본분산, s_2^2 : 두 번째 집단의 표본분산)



2. 통계기법의 이해 – 기술통계

5) 최대우도법(Maximum Likelihood Method)

- 우도(likelihood, 가능도)는 현재 얻은 데이터가 해당 분포로부터 나왔을 가능성을 의미한다.
- 최대우도법은 어떤 확률변수에서 표집한 값들을 토대로 그 확률변수의 모수를 구하는 방법이다.
- 즉, 우리가 알고 싶은 데이터(θ, 모수)가 있을 때, 다양한 관측치들을 통해서 그 데이터가 나올 수 있게 하는 가장 그럴듯한 값을 추정하는 통계 기법이다.
- 연산의 편의성을 위해 자연 로그를 취하고, 양 변을 미분하여 0이 되게 하는 감마(r)를 찾는다.

모수(Parameter) : 모집단 분포 특성(모평균, 모분산 등)을 규정짓는 척도로 관심의 대상이 되는 모집단의 대푯값이다.

2. 통계기법의 이해 – 기술통계

03 표본분포

표본분포(Sample Distribution)는 모집단에서 추출한 일정한 크기의 표본에 대한 분포상태를 의미한다.

1) 표본분포 용어

; 표본분포 용어에는 모집단, 모수, 통계량, 추정량 등이 있다.

① 모집단(Population) : 정보를 얻고자 하는 관심 대상의 전체 집합이다.

② 모수(Parameter) : 모집단 분포 특성(모평균, 모분산 등)을 규정짓는 척도로 관심의 대상이 되는 모집단의 대푯값이다.

③ 통계량(Statistic) : 표본의 몇몇 특징을 수치화한 값(평균, 표준오차). 통계량을 통해 모수를 추정하고, 무작위로 추출할 경우 각 표본에 따라 달라지는 확률변수이다. 표준 오차(standard error, SE)는 통계의 표본 분포의 표준 편차이다.

④ 추정량(Estimator) : 모수의 추정을 위해 구해진 통계량이다.

2. 통계기법의 이해 – 기술통계

2) 표본분포와 관련된 법칙

① 큰 수의 법칙(Law Large Number)

▶ 데이터를 많이 선택할수록(n이 커질수록) 표본평균의 분산은 0에 가까워진다.

② 중심극한정리(Central Limit Theorem)

▶ 데이터의 크기가 커지면 데이터의 표본분포는 최종적으로 정규분포의 형태를 따른다.

2. 통계기법의 이해 – 기술통계

개념 체크

01 다음 중 표본추출 기법에 속하지 않는 것은?

① 단순추출

② 계통추출

③ 층화추출

④ 군집추출

표본추출 기법에는 단순무작위추출 , 계통추출 , 층화추출 , 군집추출이 있다 .

단순무작위 추출

▶모집단에서 정해진 규칙 없이 표본을 추출하는 방식이다 .

▶표본의 크기가 커질수록 정확도 높아지고, 추정값의 분산이 작아진다 .

계통추출

▶ 모집단을 일정한 간격 및 구간으로 추출하는 방식이다 .

층화추출

▶ 모집단을 여러 계층으로 나누고, 계층별로 무작위 추출하는 방식이다 .

▶ 데이터 특징이 층 내에서는 동질하고, 층간에서는 이질한

02 다음에 설명하는 표본추출 기법은 어느 것인가?

모집단을 일정한 간격 및 구간으로 추출하는 방식이다.

예) 특정한 지역에 주민등록번호 끝 번호가 3인 사람을 추출

① 계통추출

② 계층추출

③ 군집추출

④ 층화추출

2. 통계기법의 이해 – 기술통계

03 다음 중 조건부 확률에 대한 설명으로 옳지 않은 것은?

① 조건부 확률은 어떤 사건이 일어난다는 조건에서 다른 사건이 일어날 확률을 의미한다.

② 사건 A가 조건으로 일어날 때 사건 B가 발생할 확률은

$P(B|A) = \frac{P(A \cap B)}{P(A)}$ 와 같다.

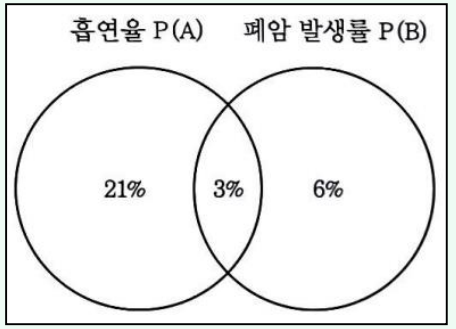
③ 조건부 확률은 -1~1의 범위를 갖는다.

④ 사건 B가 조건으로 일어날 때 사건 A가 발생할 확률은

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ 와 같다.

확률은 0~1범위의 값을 갖는다 .

04 다음과 같은 데이터가 주어졌을 때, 흡연할 때 폐암에 걸릴 조건부 확률은?



① 1/7 ② 1/10

③ 1/3 ④ 1/8

주어진 데이터에서 흡연할 때 폐암에 걸릴 확률은

$P(B|A) = \frac{P(A \cap B)}{P(A)}$ 이므로

= 3/24 = 1/8 = 12.5%가 된다 .

2. 통계기법의 이해 – 기술통계

05 다음 중 이산확률분포의 종류가 아닌 것은?

① 푸아송 분포

② 카이제곱 분포

③ 베르누이 분포

④ 이항 분포

카이제곱 분포는 이산확률분포가 아닌 연속확률분포에 속한다 .

이산확률분포

▶ 이산확률분포는 이산확률변수 X가 갖는 확률분포를 나타낸다 .

▶ 이산확률변수는 확률변수 X가 0,1,2,3…과 같이 하나씩 셀 수 있는 값을 갖는다 .

▶ 이산확률분포의 종류에는 푸아송 분포, 베르누이 분포, 이항 분포가 있다 .

06 다음 중 연속확률분포에 대한 설명으로 옳지 않은 것은?

① 정규분포는 분포 곡선이 평균값을 중심으로 좌우 대칭한 종 모양의 분포이다.

② 표준정규분포는 정규분포에서 x를 Z로 정규화한 분포로 평균이 0이고 분산이 1인 정규분포이다.

③ T-분포는 모표준편차를 알 때, 모집단의 평균을 추정 하기 위해 사용하는 분포이다.

④ F-분포는 독립적인 x^2 분포가 있을 때, 두 확률변수의 비를 나타낸다.

T-분포

▶모집단이 정규분포라는 정도만 알고 모표준편차는 모를 때, 모집단의 평균을 추정하기 위해 사용하는 분포

▶ 표본의 크기가 작은 경우 사용하고, 중심극한정리 에 의해 Z-분포는 정규분포를 따른다 .

▶ 두 집단의 평균이 동일한지 확인하고자 할 때 검정통계량으로 사용된다 .

2. 통계기법의 이해 – 기술통계

07 다음 중 표본분포에 대한 설명으로 틀린 것은? .

① 모수는 정보를 얻고자 하는 관심 대상의 전체 집합을 의미한다.

② 표본분포 용어에는 모집단, 모수, 통계량, 추정량 등이 있다.

③ 표본분포는 모집단에서 추출한 일정한 크기의 표본에 대한 분포 상태를 의미한다.

④ 표본분포와 관련된 법칙으로는 큰 수의 법칙, 중심극한정리가 있다.

1 번은 모집단에 대한 설명이다 .

2. 통계기법의 이해 – 기술통계 예상문제

예상문제

01 다음 중 왜도 < 0일 때, 올바른 배열은?

① 최빈수 < 중위수 < 평균

② 최빈수 > 중위수 > 평균

③ 최빈수 < 중위수 = 평균

④ 최빈수 = 중위수 < 평균

왜도 < 0 일 때는 왜도가 음수이다. 왜도가 음수일 때는 좌측꼬리를 가진다. 최빈수 > 중위수 > 평균과 같은 분포를 갖는다. 1번 같은 경우는 왜도 > 0 일 때에 대한 설명이다.

02 다음과 같은 특징을 갖는 표본추출 기법은?

모집단을 여러 군집으로 나누고 일부 군집의 전체를 추출하는 방식이다.

데이터 특징이 집단 내에서는 이질적이고, 집단 외에서는 동질한 특징이 있다.

① 군집추출 ② 계통추출

③ 층화추출 ④ 집단추출

표본추출 기법에는 단수무작위추출, 계통추출, 층화추출, 군집추출이 있다.

03 다음 중 평균값에 대한 설명으로 옳지 않은 것은?

① 평균값은 주어진 데이터를 모두 더한 뒤, 데이터의 개수만큼 나눈 값을 의미한다.

② 평균값은 이상치에 영향을 받지 않는다.

③ 평균값의 종류로 모평균과 표본평균이 있다.

④ 모평균의 수식은 $\mu = \frac{1}{N}\sum_{i=1}^{N} Xi$ 와 같다.

평균값(Average)

▶ 평균값은 주어진 데이터를 모두 더한 뒤, 데이터의 개수만큼 나눈 값을 의미한다.

▶ 평균값은 모두 같은 가중치를 두며, 이상치에 굉장히 민감하다.

▶ 평균의 종류에는 산술평균, 기하평균, 조화평균이 있다.

▶ 평균의 대상 범위에 따라서 모평균, 표본평균이 있다.

04 다음 중 데이터 분포가 오른쪽 꼬리를 갖는 왜도 형태일 경우 올바른 데이터 순서는?

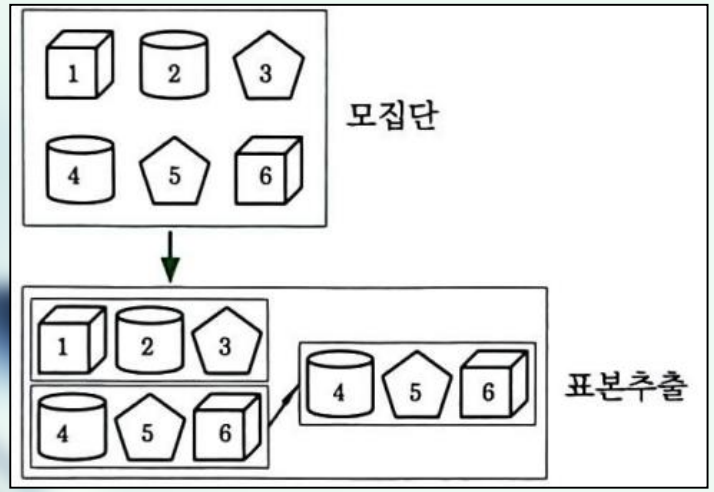
① 평균 < 최빈수 < 중위수

② 평균 < 중위수 < 최빈수

③ 최빈수 < 평균 < 중위수

2. 통계기법의 이해 – 기술통계 예상문제

05 다음 그림이 설명하는 표본추출 기법은?



① 단순무작위추출

② 층화추출

③ 군집추출

④ 계통추출

그림이 설명하는 표본추출 기법은 집단 내에서는 이질적이고, 집단 외에서는 동질한 특징을 갖는 군집추출이다.

※ 다음과 같은 모집단 데이터를 보고 물음에 답하시오. (06~08)

구분	A	B	C	D	E
남자	170cm	173cm	175cm	170cm	167cm
여자	161cm	162cm	161cm	165cm	166cm

07 주어진 데이터에서 여자 집단의 분산은 얼마인가?

① 4.2 ② 4.4

③ 4.5 ④ 4.7

주어진 데이터에서 여자 집단의 분산을 구하기 위해서는 먼저 평균을 구한다. 여자 집단의 평균은

161 + 162 + 161 + 165 + 166 = 815 / 5 = 163cm가 된다.

분산은 편자 제곱의 평균이므로 다음과 같이 연산된다.

$(161 - 163)^2 + (162 - 163)^2 + (161 - 163)^2 + (165 - 163)^2 + (166 - 163)^2 = 4 + 1 + 4 + 4 + 9 = 22 / 5 = 4.4$ 가 분산이 된다.

08 주어진 데이터에서 여자 집단의 표준편차는 얼마인가? (단, 소수점의 경우 소수점 4자리에서 반올림한다.)

① 2.025 ② 2.098

③ 2.074 ④ 2.121

표준편차는 분산에 양의 제곱근을 취한 값을 의미하기 때문에, 여자 집단의 표준편차는 4.4 루트(제곱근) = 2.098

이 표준편차가 된다.

2. 통계기법의 이해 – 기술통계 예상문제

09 다음 중 표본분포에 대한 설명으로 틀린 것은?

① 모수는 정보를 얻고자 하는 관심 대상의 전체 집합을 의미한다.

② 표본분포 용어에는 모집단, 모수, 통계량, 추정량 등이 있다.

③ 표본분포는 모집단에서 추출한 일정한 크기의 표본에 대한 분포 상태를 의미한다.

④ 표본분포와 관련된 법칙으로는 큰 수의 법칙, 중심극한정리가 있다.

1번은 모집단에 대한 설명이다.

표본분포 용어

① 모집단(Population) : 정보를 얻고자 하는 관심 대상의 전체 집합이다.

② 모수(Parameter) : 모집단 분포 특성(모평균, 모분산 등)을 규정짓는 척도로써 관심의 대상이 되는 모집단의 대푯값이다.

③ 통계량(Statistic) : 표본의 몇몇 특징을 수치화한 값(평균, 평균오차). 통계량을 통해 모수를 추정하고, 무작위로 추출할 경우 각 표본에 따라 달라지는 확률변수이다.

11 다음 중 상관관계에 대한 설명으로 틀린 것은?

① 두 변수 사이에 어떤 선형적 혹은 비선형적 관계가 있는지 분석하는 방법이다.

② 상관관계의 수치를 나타내는 상관계수는 -1 ~ 1의 범위 를 갖는다.

③ 상관계수가 1에 가까울수록 강한 양(+)의 상관관계를 갖는다고 할 수 있다.

④ 상관관계로 두 변수 사이의 인과관계를 알 수 있다.

상관관계는 두 변수 사이의 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 방법으로 상관관계로는 인과관계는 알 수가 없다.

상관관계(Correlation)

● 상관관계는 두 변수 사이의 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 방법으로 상관관계로는 인과관계는 알 수가 없다.

● 공분산은 선형관계의 강도를 나타낼 수 없지만, 상관계수는 선형관계의 강도를 나타낼 수 있다.

● 상관계수는 -1 ~ +1 사이의 값을 가지며 1에 가까울수록

2. 통계기법의 이해 – 기술통계 예상문제

13 한 회사에서 A공장은 부품을 50% 생산하고 불량률이 1%이고, B공장은 부품을 30% 생산하고 불량률이 2%이며, C공장은 부품을 20% 생산하고 불량률이 3%이다. 불량품이 발생했을 때 B공장에서 생산할 부품일 확률은 얼마인가?

① 5/17 ② 6/17

③ 2/3 ④ 1/5

다음과 같이 베이즈 정리로 연산을 한다.

A1 : A공장, A2 : B공장, A3 : C공장, B : 불량률

P(A1):A공장 부품 생산율 50%, P(B|A1):A공장 불량률:1%

P(A2):B공장 부품 생산율 30%, P(B|A2):B공장 불량률:2%

P(A3):C공장 부품 생산율 20%, P(B|A3):C공장 불량률:3%

P(A2|B):불량품이 B공장에서 생산될 확률

= P(A2)P(B|A2) / P(A1)P(B|A1) + P(A2)P(B|A2) + P(A3)P(B|A3)

= 30% * 2% / (50% * 1%) + (30% * 2%) + (20% * 3%)

= 60 / 50 + 60 + 60 = 60 / 170 = 6 / 17 이 된다.

14 지수 분포를 따르는 확률밀도함수에서 표본 3, 1, 2, 3, 4가 추출되었다. 최대우도 추정법을 이용해서 최대우도 추정치를 구하면 얼마인가? (단, 표본은 서로 독립적이

15 다음과 같은 <사례>에서 A농구팀 연봉의 대푯값을 산출하기 위해 가장 적절한 통계량은?

<사례>

A농구팀 일부 팀원의 연봉이 구단 전체 연봉의 60% 이상을 차지하고, 나머지 선수들은 일반적인 연봉 범위에 포함된다.

① 중위수 ② 평균

③ 최빈값 ④ 최댓값

일부 팀원의 연봉이 기준 이상으로 높은 경우 이는 이상치에 해당한다. 이상치에 영향을 받지 않는 통계량은 중위수이다.

중위수(Median)

● 중위수는 모든 데이터를 오름차순으로 정렬했을 때, 가장 중앙에 위치한 데이터 값을 의미한다.

● 중위수는 이상치에 대해서 영향을 전혀 받지 않는다.

● 중위수의 개수가 짝수인 경우 중앙에 있는 두 개의 값의 평균을 중위수로 정한다.

16. 다음 중 대푯값에 대한 설명과 특징으로 옳지 않은 것은?

2. 통계기법의 이해 – 기술통계 예상문제

17 다음 중 모집단과 표본의 통계량에 대한 설명으로 옳지 않은 것은?

① 모집단의 분포와 상관없이 표본의 수가 큰 표본평균의 분포는 정규분포를 따른다.

② 표본분포의 평균은 모집단의 평균과 동일하다.

③ 표본평균의 분산은 n의 크기와 관계없이 모평균의 분산 을 따른다.

④ 동일한 모집단의 표준편차에서 표본의 크기가 커지면 커질수록 표준오차는 줄어든다.

표본평균의 평균은 모평균과 동일하고, 표본평균의 분산은 모분산을 표본의 크기로 나눈 것이다. 따라서 모집단의 분산이 σ^2인 경우 표본분포의 분산은 $\frac{\sigma^2}{n}$이 된다.

18. 다음 중 이산확률분포의 종류가 아닌 것은?

① 푸아송 분포

② 카이제곱 분포

③ 베르누이 분포

④ 이항 분포

19 다음 중 IQR에 대한 설명으로 옳은 것은?

① IQR은 사분범위, 사분위수범위와 같은 말이다.

② IQR은 제4사분위수와 제2사분위수의 차이값을 의미한다.

③ IQR은 시각화 도구인 막대그래프에서 확인할 수 있다.

④ 사분편차는 IQR X 2로 표현된다.

2번 같은 경우는 제3사분위수와 제1사분위수의 차이값을 의미하는 것이 바로 IQR이다.

3번 같은 경우는 시각화 도구인 상자수염그림(Box-Plot)에서 확인할 수 있다.

4번 같은 경우는 사분편차(절반 값)는 IQR / 2를 한 것이다.

20. 다음과 같은 데이터에서 중위수는 얼마인가?

13, 5, 3, 2, 6, 10, 1, 20, 8, 11

① 5 ② 8

③ 6 ④ 7

주어진 데이터를 오름차순으로 정렬을 먼저 해야 한다.

1, 2, 3, 5, 6, 8, 10, 11, 13, 20 -> 10개로 짝수이다.

위와 같이 주어진 데이터의 개수가 10개(짝수)이므로 중앙의 두 수(6, 8)를 더해서 2로 나눈 값이 중위수가 된다.

감사합니다.