2.3 통계 기법 이해

개념 1

● 기술 통계

기술 통계학이란 수집된 자료를 정리하여 그림이나 표로 예약하거나 자료의 수치를 요약한 대푯값과 데이터 분포의 형태와 변동의 크기를 구하는 방법을 말한다.

예시) 최솟값,최댓값,중위수 등으로 데이터의 특성을 유추

개념 2

● 추론 통계

수집한 데이터를 바탕으로 추론 및 예측하는 통계 기법을 의미한다.

개념 3

● 표본 추출

전수조사란 관측하고자 하는 데이터의 모든 범위를 조사하는 방법이다. 전수조사는 많은 시간과 비용이 들기 때문에, 데이터 일부분만 추출하는 것을 표본추출이라고 한다.

모수 : 관심을 갖고 있는 모집단의 대푯값

통계량 : 표본추출된 데이터를 바탕으로 모수를 추정하는데 사용되는 것

* 모분산

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

* 표본분산

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

개념 4

● 확률 표본추출 방법

1. 단순 무작위 표본 추출

통계 조사의 기본으로 모집단으로부터 표본을 균등한 확률로 추출하는 것

2. 체계 표본 추출

모집단 관측치로부터 시간, 순서 및 공간의 동일한 구간을 정해서 무작위로 하나의 단위를 추출하고 그 이후 k번째 간격마다 하나씩 단위를 추축하는 방법

3. 층화 표본 추출

모집단을 겹치지 않는 여러 개의 층으로 분할한 후 각 층에서 표본을 단순 무작위 추출법에 따라 추출하는 방법이다.

4. 군집 표본 추출

모집단을 어떤 기준에 따라 서로 인접한 기본단위인 군집을 형성하고 그 군집중 하나를 추출하는 방법이다.

개념 5

● 비확률 표본추출 방법

1. 편의 표본 추출 방법

마음대로 표본추출 하는 방법

2. 판단 표본 추출 방법

조사자의 주고나적 판단으로 조사에 필요한 대상만을 조사하는 방법

3. 누적 표본 추출 방법

표본 조사 대상이 접근이 어려운 경우 사전에 알고 있는 대상을 조사하고 다른 표본의 대상도 조사하여 누적추출하는 방법

4. 할당 표본 추출

특정한 기준에 따라 여러 그룹으로 구분하여 그룹별로 필요한 대상을 추출하는 방법

개념 6

● 확률 개념

1. 이론적 확률

$P(A) = \frac{n(A)}{n(S)} = \frac{\text{사건 A의 빈도}}{\text{표본공간 S의빈도}}$

2. 객관적 확률

$lim_{n\to\infty}\frac{r_n}{n} = p$

3. 주관적 확률

관찰자의 주관적 견해로 표현되는 확률을 의미한다.

개념 7

● 표본공간과 확률의 기본성질

1. 표본공간

어떤 무작위 실험을 했을 때, 측정 가능한 모든 결과값들의 집합

2. 확률의 기본 성질

2.1 $0 \leq P(A) \leq 1$

2.2 $P(S) = 1$

2.3 $P(\varnothing) = 0$

3. 전사건

반드시 일어나는 사건

4. 공사건

절대 일어날 수 없는 사건

개념 8

● 확률분포의 개념

1. 확률 변수

확률변수란 결과를 예측할 수 없는 실험에서 나타날 수 있는 확률적 결과를 수치로 표현하는 함수를 의미한다.

2. 확률 분포

확률 변수의 모든 값과 그에 대응하는 확률이 어떻게 분포하는지를 그래프로 나타내는 것을 확률분포라고 정의한다.

3. 확률분포함수

확률 변수를 일직선상 공간에 표현한 함수다.

확률분포함수는 확률질량함수, 누적분포함수, 확률밀도함수로 나누어진다.

4. 확률질량함수

이산형 확률변수의 확률분포이다.

5. 누적분포함수

$F(x) = P(X \leq x)$

6. 확률밀도함수

연속형 확률변수의 확률분포이다.

**참고

누적분포함수의 특징

1. 단조증가

2. 오른쪽 연속

3. 최대 극한값 1, 최소 극한값 0

개념 9

● 연속형 확률분포

연속형 확률분포에는 정규분포, 균등분포, 감마분포, 베타분포, 지수분포, T-분포, F-분포, 카이제곱 분포 등이 있다.

개념 10

● 이산형 확률분포

이산형 확률분포에는 베르누이분포, 이항분포, 다항분포, 초기하분포, 포아송분포 등이 있다.

개념 11

● 표본분포

모집단으로부터 추출한 표본들의 분포

개념 12

● 중심극한정리

모집단의 분포와 관계 없이, 표본들의 평균은 표본의크기(n)이 무한에 가까워짐에 따라 정규분포에 근사한다는 이론이다.

개념 13

● 점추정

점추정이란, 모집단의 모수를 하나의 통계량값으로 추정하는 것을 말한다.

개념 14

● 추정량의 특징

1. 불편성

2. 효율성

3. 일치성

4. 충분성

개념 15

● 구간추정

구간을 통해서 모집단의 모수를 추정하는 방법을 말한다. 이러한 구간은 신뢰구간이라고 부르고, 신뢰하한 신뢰상한으로 구성되어 있다.

개념 16

● 신뢰구간의 계산

1. 모분산 σ^2이 알려져 있는 경우

$X_1, ..., X_n$:관찰한 Data

$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

2. 모분산 σ^2을 모르고 $n \geq 30$일 경우

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$\bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

3. 모분산 σ^2을 모르고 소표본 일때

$\bar{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

개념 17

● 모비율에 대한 추정

$\hat{p} = \frac{X}{n}$ 일때, 다음을 이용해서 신뢰구간을 구할 수 있다.

$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

개념 18

● 가설 검정

통계적 가설검정은 모집단의 특성에 대한 주장 또는 가설을 세우고 표본에서 얻은 정보를 이용해 가설이 옳은지를 판단하는 과정이다.

개념 19

● 가설검정 용어들

1. 귀무가설

대립가설과 상반되며, 일반적인 통념 및 주장

2. 대립가설

검정을 통해 입증하고 싶은 주장

3. 검정통계량

가설검정에 사용되는 통계량

4. 유의수준

제1종오류 확률

5. 기각역

귀무가설을 기각하는 검정통계량의 집합

6. 유의확률

귀무가설을 지지하는 정도를 나타내는 확률

개념 20

● 제1종 오류와 제2종 오류

1. 제1종 오류

귀무가설이 참일 때 귀무가설을 기각하는 오류

2. 제2종 오류

대립가설이 참일 때, 귀무가설을 채택하는 오류

개념 21

● 검정력

대립가설이 참일 때, 귀무가설을 기각하고 대립가설을 채택할 확률

개념 22

● 모평균 검정

관찰한 데이터 $X_1, ...X_n \sim^{iid} \mathrm{N}(\mu, \sigma^2)$ 라고 하자.

1. 모분산 σ^2을 아는 경우

case 1) $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{\frac{\alpha}{2}}$, $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{\frac{\alpha}{2}}$ 이다.

case 2) $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{\alpha}$ 이다.

case 3) $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{\alpha}$ 이다.

2. 모분산 σ^2을 모르는 경우

case 1) $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{\frac{\alpha}{2}}$, $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\frac{\alpha}{2}}$ 이다.

case 2) $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{\alpha}$ 이다.

case 3) $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

이때 기각역은 $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\alpha}$ 이다.

개념 23

● 단일 모비율 검정

1. $H_0 : p = p_0$ vs $H_1 : p \neq p_0$

기각역 : $\frac{\hat{p}-p}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_{\alpha/over2}$, $\frac{\hat{p}-p}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/over2}$

2. $H_0 : p = p_0$ vs $H_1 : p < p_0$

$\frac{\hat{p}-p}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_{\alpha/over2}$

3. $H_0 : p = p_0$ vs $H_1 : p > p_0$

$\frac{\hat{p}-p}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/over2}$

개념 24

● 독립표본 t 검정

서로 관계없는 두 그룹에서의 평균이 같은지 다른지 검정하는 것

개념 25

● 대응표본 t 검정

동일한 대상에 대하여 두 가지 관측치가 있는 경우 이를 비교할때 사용되는 것

개념 26

● 정규성 검사

1. QQ plot와 같은 것으로 관찰한 데이터가 정규분포를 따르는지 확인 할 수 있다.

2. 샤피로-윌크 정규성 검정으로 모집단이 정규분포를 따르는지 검정할 수 있다.

* 기출 정리

왜도 > 0 ⇒ 최빈수 < 중위수 < 평균

조화평균은 속도를 평균낼때, 사용하기에 적합하다.

$\chi^2(n)$, $n \geq 3$ 이면 단봉형 태를 갖는다.
자유도 n이 작을수록 원점으로 치우치는 비대칭적인 모양을 갖는다.

군집추출 ⇒ 집단 내 이질적, 집단 간 동질적 이다.

유의수준 : 제 I 종오류를 범할 최대 확률.

대응 표본에 대한 비모수 검정 : 윌콕슨의 부호검정

두 표본에 대한 비모수 검정 : 윌콕슨의 순위합 검정.

분산분석 : 크루스칼 왈리스 검정.

비모수 통계 : 모집단 분포에 대한 가정의 불만족으로 인한 오류의 가능성이 크다.
모수 방법에 비해 계산량이 ↓ 하다.