데이터사이언스와 통계적사고 (2017년도 1학기)

< 목차 >

1. 통계학의 기본개념

가. 통계학과 과학적 접근 : 통계학은 데이터를 수량적으로 분석하는 학문으로서 지식을 얻는 과학적·합리적 방법

1) 기술통계 : 수집된 자료를 간단하고 편리하게 서술하는 통계로서 얻어진 자료를 분석하여 그 자료를 구성하는 대상들의 속성을 **설명·묘사**하려는 목적으로 사용되는 통계

2) 추리통계 : 얻은 자료를 가지고 우리가 관심을 가지는 집단의 속성을 추정하는 통계로서 **표본에서 얻은 자료의 결과를 모집단에 일반화**하려는 목적의 통계

나. 통계의 기본 용어와 개념

1) 모집단(Population) : 연구자가 연구의 대상으로서 관심을 갖는 집단 전체[1)]

2) 표본(Sample) : 연구 대상으로 하는 모집단의 일부, 표본은 모집단의 성격을 잘 대표해야한다.

3) 모수치(Parameter) : 모집단을 나타내는 수치

4) 통계량(statistics): 표본의 몇몇 특성을 수치화한 것 (예,)

5) 추정치(Estimator) : 모수치를 추정하기 위하여 모집단에서 추출된 표본의 속성[2)] (예, $\overline{X}$, s^2, s_{XY}, r)

다. 변수 : 변하는 특성 모두 ↔ 상수

1) 인과관계에 따른 변수의 분류[3)]

가) 독립변수(Independent Variable) : 원인이 되는 변수

나) 종속변수(Dependent Variable) : 결과가 되는 변수

다) 외재변수

(1) 매개변수(Intervening Variable) : 독립변수와 종속변수 사이에 끼어들어 영향을 미치는 변수[4)]로서 모든 실험에는 매개변수가 개입하는데 심각한 것은 통제[5)]를 해주어야 한다.

(2) 혼재변수/제3변수(Confounding Variable/The Third Variable) : 뒤에 숨어서 독립변수와 종속변수에 똑같이 영향을 미치는 것[6)]

(3) 실험이 우연에 의해 잘못되는 것은 용서가 가능하지만, 외재변수를 간과하는데서 오는 잘못은 연구자의 수치로 남는다.

2) 속성에 따른 변수의 분류

---

1) 모집단을 가지고 연구하고 싶지만 항상 그렇게 할 수는 없다. 그리고 모집단을 대상으로 조사를 하면 형편없는 결과를 얻게 된다. 예를 들어 대한민국 중학생 학력검사를 하기 위해 전수조사를 하면 엄청나게 많은 비용이 들고, 많은 데이터를 다룰 때는 반드시 중요한 실수가 발생한다. 따라서 모집단의 성질을 알기 위해 표본을 사용하는 것이다.

2) 표본을 가지고 추정치를 얻고 그것으로 모수치를 추정한다(=모집단의 속성을 안다)`

3) 독립변수와 종속변수 사이에는 인과관계가 있고, 과학은 두 변수 사이의 인과관계를 규명하는 것이다.

4) 예 : 비료의 종류 → 쌀 생산; 토양의 질 개입(매개변수), 교수법 → 시험; 소음(매개변수)

5) 빼거나 똑같이 만들어주는 것

6) 예 :

① “기온↑“ → 아이스크림↑ → 익사 ↑

② ”화재규모↑“ → 출동 소방관 수 ↑ → 피해액 ↑

③ “인구 2배” → 백주소비량 2배 → 위암 2배

가) 질적 변수 : 분류를 위하여 용어로 정의되는 변수
(1) 비서열 질적 변수 : 서열 X → 성별, 인종. 이름. 지지정당
(2) 서열 질적 변수 : 서열 O → 학력, 계급, 직급, 등급, 의견
나) 양적 변수 : 양의 크기를 나타내기 위해 수량으로 표시
(1) 연속변수 : 모든 수치 → 체중, 신장
(2) 이산변수 : 특정 수치 → 차량 대수, IQ 점수, 일수, 나이, 안타 수

라. 측정과 척도
1) 측정 : 관찰을 통해 사물의 특성을 수량화·범주화
2) 척도 : 사물의 속성을 구체화하기 위한 측정의 단위
가) 명명척도 : 사물을 구분하기 위해 이름 부여 → 출석부 번호, 축구선수 등번호
나) 서열척도 : 등간성[7] X, 사물의 등위를 나타내기 위해 사용 → 상영영화 인기순위
다) 등간척도 : 등간성[8] O, 임의영점 + 임의단위[9] → 온도, 학업성취 점수
라) 비율척도 : 등간성 O, 절대영점 + 임의단위[10] → 몸무게, 키

2. 확률
가. 확률의 의미
1) 오차의 특성을 이해하는 방식
가) 확률 = 불확실성에 대한 수량화
나) 오차는 통제불능 △ 정규분포
2) 반증의 과학
가) 오차가 존재하는 한 A → B의 엄밀한 인과관계 증명 X
나) A → B는 “우연에 의한 것” 기각 → “인과 관계에 의한 것” 더 합리적
3) 과학이라는 시스템 속에서의 확률의 의미 : 정확한 확률에 의지하는 것은 장기적 관점에서 매우 강력한 수단

나. 기본용어의 정의
1) 시행 : 임의로 발생하는 어떤 결과를 얻기 위한 과정 → 가능한 모든 결과 파악 가능
2) 표본공간과 표본점
가) 표본공간 : 한 시행에서 가능한 모든 결과의 집합
나) 표본점 : 시행의 결과로 가능한 각각의 결과
다) 표본공간 = ∑표본점
3) 사건 : 표본공간의 부분집합
4) 근원사건 : 단일한 표본점으로 구성된 사건
5) 확률 : 한 사건을 구성하는 근원사건의 확률의 합

---
7) 척도 단위 사이에 등간성 존재 X
8) 등간성이란 등급의 차이만큼 몇 배 차이난다고 할 수 있는 것을 의미한다.
9) 덧셈법칙 O 곱셈법칙 X
10) 덧셈법칙 O 곱셈법칙 O

가) $0 \leq P(E_i) \leq 1$

나) $\sum P(E_i) = 1$

다. 확률의 여러 정의

1) 고전적 정의(a priori 확률)

가) 전제조건 : 어떤 시행에서 각각의 근원사건이 발생할 확률이 모두 같음

나) 정의 : m개의 근원사건이 발생하는 어떤 시행에서 임의의 사건 E가 k개의 근원사건으로 구성된다면, 사건 E의 확률은 :

$P(E) = \frac{k}{m}$ (m : 한 시행의 모든 근원사건의 수, k : 임의의 한 사건의 근원사건의 수)

다) 문제점

(1) 근원사건의 수가 무한한 경우 확률이 정의될 수 없음

(2) 각각의 근원사건이 발생할 가능성이 모두 같다고 믿을 수 있는 경우에만 적용

(3) 전제조건에서 확률을 정의하기 위해 확률을 도입 → 동어반복의 모순

2) 상대도수에 의한 정의(a posteriori 확률)

가) 통계학은 귀납법을 믿음[11] → a posteriori 확률은 경험적 정의

나) 정의 : n이 충분히 클 때, 어떤 사건의 확률은 그 사건이 나타난 상대도수로서, 확률이란 상대도수의 극한값

$P(A) = \lim_{n\to\infty}\left(\frac{f(A)}{n}\right)$ (n: 사례 수, f(A) : 사건 A가 포함된 사례 수)

다) 문제점

(1) 얼마나 많이 반복해야 충분한 반복인지 명확하지 않음

(2) 극한값이라는 것도 단지 어떤 값에 한없이 가까워지는 것을 뜻하므로 얼마만큼 가까워졌는지 알기 위해서는 결국 확률의 개념을 도입해야하므로 모순 발생 → 연역적 정의 X

라. 대수의 법칙(큰 수의 법칙)[12]

1) 의미 : **독립표본**들의 **평균**은 **모집단 평균**에 **수렴**한다.

2) 정의 : 시행횟수 n, 관찰된 횟수 k에 대하여, n이 충분히 클 때 $\frac{k}{n}$는 확률 p에 수렴

3) 오해

가) 오차의 개념 오해 : 오차가 확률오차가 아닌 체계오차일 때 성립 X

나) 도박사의 오류 : 사건의 결과가 평균을 유지하는 방향으로 나타난다고 생각 → 각 사건은 독립적이므로 전체 확률을 유지해주기 위해 움직이지 않는다.

다) 대수의 법칙이 적용되지 않는 곳까지 적용 : 사건들이 독립적이지 않으면 적용 X

마. 확률의 계산

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

2) 곱셉법칙

가) 배반사건 : $A \cap B = \Phi$

---

11) 연역의 명제들은 모두 귀납법에서 오는데, 귀납에서 온 결론은 연역의 대전제로 사용하는 것은 순환 논증의 오류이다.

12) 대수의 법칙 : $\bar{X}$, 중심극한정리 : $\bar{X}$의 분포

나) 독립사건

(1) $P(B|A) = P(B)$

(2) $P(A|B) = P(A)$

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

(4) 복원추출

다) 종속사건

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

(2) $P(A \cap B) \neq P(A) \cdot P(B)$

(3) 비복원추출

라) 배반사건과 독립사건의 구분

(1) 배반 → 종속[13]

3. 확률분포

가. 기본적 용어의 정리

1) 확률변수 : 어떤 확률적 현상의 결과를 수량화한 측정치[14]로서 이산 확률변수와 연속 확률변수가 있다.

2) 이산 확률변수와 연속 확률변수

가) 이산 확률변수 : 셀 수 있는 횟수의 값 (예 : 동전을 던져서 앞면이 나온 횟수)

나) 연속 확률변수 : 구간 내에 무한히 많은 값 (예 : 키, 몸무게)

3) 확률변수와 확률분포

가) 확률분포 : 확률변수의 가능한 모든 값에 대하여 확률을 부여한 결과의 분포

나) 이산 확률변수의 기댓값 : $E(X) = \Sigma x P(x)$

다) 연속 확률분포의 구간과 막대그래프 : 확률변수가 해당 구간에 존재할 확률을 대응시킨 것으로서 확률은 전체 넓이에서 차지하는 해당 막대그래프의 넓이의 비

4) 이산 확률변수의 분산

가) X값과 평균의 차이의 제곱의 기대치

나) 수식

(1) $\sigma^2 = E[(X-\mu)^2] = \Sigma (x-\mu)^2 P(x)$

(2) $\sigma^2 = E[X^2] - \mu^2 = \Sigma x^2 P(x) - \mu^2$

5) 확률 질량함수와 확률 밀도함수

가) 확률 질량함수 : 이산 확률변수에서 **특정 확률변수에 대한 확률**을 나타내는 함수

나) 확률 밀도함수 : 연속 확률변수에서 **특정 확률변수가 그 값을 가질 상대적 가능성**을 나타내는 함수[15]

---

13) 종속 ↛ 배반; 반례 A={1,2,3}, B={2,3,5}

14) 무작위 추출이나 무작위 실행에 의해 얻어진 수치

15) 확률 밀도함수의 정의역은 연속적인 실수이기 때문에 특정한 x값에 대한 치역을 특정할 수 없다. 따라서 해당 구간에 포함될 확률로서 다룰 수밖에 없는 것이다.

다) 확률 질량함수의 Y는 **확률을 나타내지만**, 확률 밀도함수의 Y값은 확률과 밀접한 관계가 있는 **상대적 비율**이지만 **확률은 아니다**.

나. 정규분포

1) 정규분포의 특성

가) 종 형태의 곡선 모양을 가지며, 곡선 아래의 면적은 1이다.

나) 연속변수로서 곡선의 모양은 오직 μ와 σ에 의해서 결정된다.

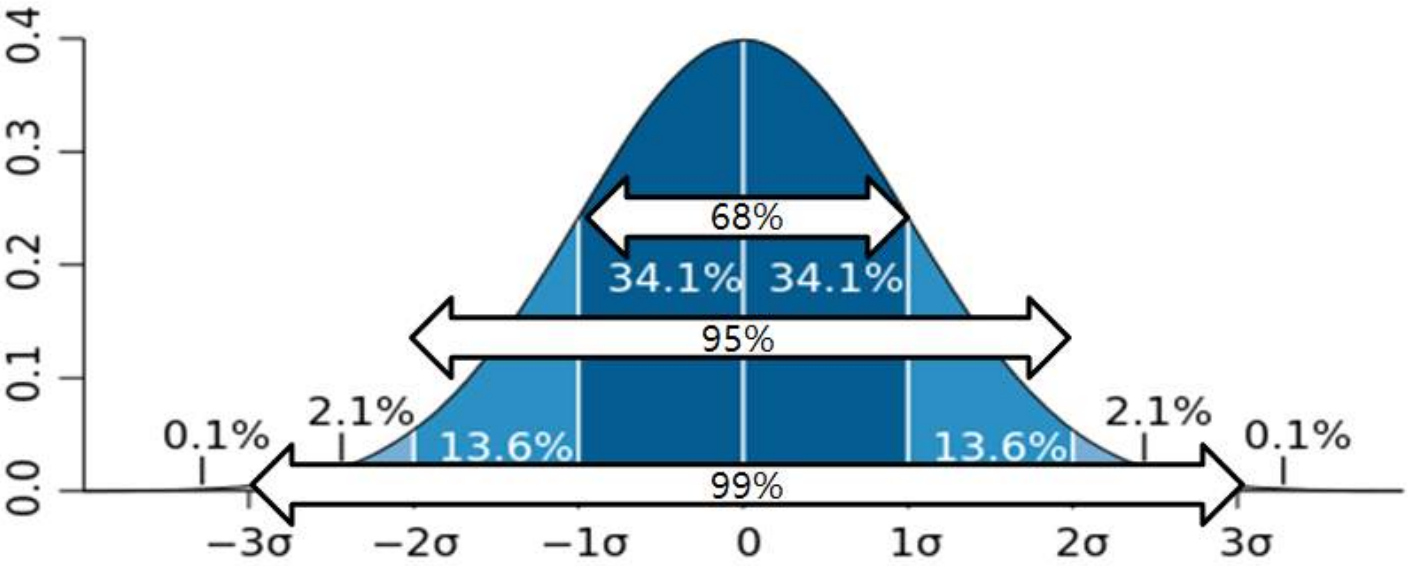
다) 평균값에 좌우대칭[16]이며 평균에서 최대값을 갖는다.

라) 곡선의 모양은 단봉분포이며, 평균=중앙값=최빈값이다.

마) x가 μ에서 멀어질수록 밀도함수 f(x)의 값은 0에 가까워진다.

바) 표준편차인 σ가 클수록 넓게 퍼지는 모양을 갖는다.

사) kσ에 따른 확률



2) 정규분포의 중요성

가) 오차의 분포가 정규분포에 근사하는 경우가 많다.

나) 이항분포를 비롯한 몇몇 확률분포도 정규분포에 근사할 수 있다.

다) 중심극한정리의 결과가 정규분포로 표시된다.

3) 정규분포의 밀도함수

가) X와 $\overline{X}$의 신뢰구간

(1) X

(가) $P[(\mu - k\sigma) < X < (\mu + k\sigma)]$

(나) $P[(X - k\sigma) < \mu < (X + k\sigma)]$

(2) $\overline{X}$

(가) $P[(\overline{X} - ks_{\overline{X}}) < \mu < (\overline{X} + ks_{\overline{X}})]$

16) x=a에 대칭일 때, f(a-x)=f(a+x), f(x)=f(2a-x)4

(나) $s_{\overline{X}} = \frac{s}{\sqrt{n}}$ [17]

(다) $P[(\overline{X} - k\frac{s}{\sqrt{n}}) < \mu < (\overline{X} + k\frac{s}{\sqrt{n}})]$ [18]

4) 표준정규분포 : z ~ N(0,1)

가) X ~ N(μ,σ)에 대해, $z = \frac{(X-\mu)}{\sigma}$이고 z는 표준정규분포를 이룬다.

나) 정규분포를 이루는 모든 확률변수 X값은 표준정규분포 상의 z값으로 전환 가능

4. 대푯값과 분산도

가. 평균, 중앙값, 최빈값(중심경향[19] 값)

1) 평균

가) 의미 : 변수의 기댓값

나) 표본평균 : $\overline{x} = E[X] = \frac{\Sigma x}{n}$, 모집단평균 : $\mu = E[X] = \frac{\Sigma x}{N}$

다) 평균은 특이값(outlier)에 민감하지만 편차에 가장 덜 영향을 받는다.

2) 중앙값

가) 의미 : 순서통계량 중 가장 가운데 위치한 값으로서 상위 또는 하위 50%에 속하는 값

3) 최빈값

가) 의미 : 가장 많은 관측빈도를 나타내는 값

나) 표본에 따른 안정성이 비교적 ↓

나. 대푯값의 비교

1) 평균과 중앙값의 비교

가) 평균은 특이값에 민감

나) 중앙값 = 특이값에 robust

2) 평균, 중앙값, 최빈값의 비교

3) 대푯값의 선택

가) 대푯값으로서 가장 안정된 측정치는 평균이지만 편포의 영향을 크게 받는다.

나) 편포를 이루는 경우 중앙값이 목적에 더 잘 부합 → 가구당 연간소득의 대푯값으로 적절[20]

다) 국세청은 평균에, 기성복 제조업자는 최빈값에 관심을 가질 것 → 어느 대푯값을 선택할지는 그것을 사용하는 목적과 문맥에 따라 달라지는 것

다. 표준편차, 분산, 표준점수

1) 편차 : 관측값의 평균의 차

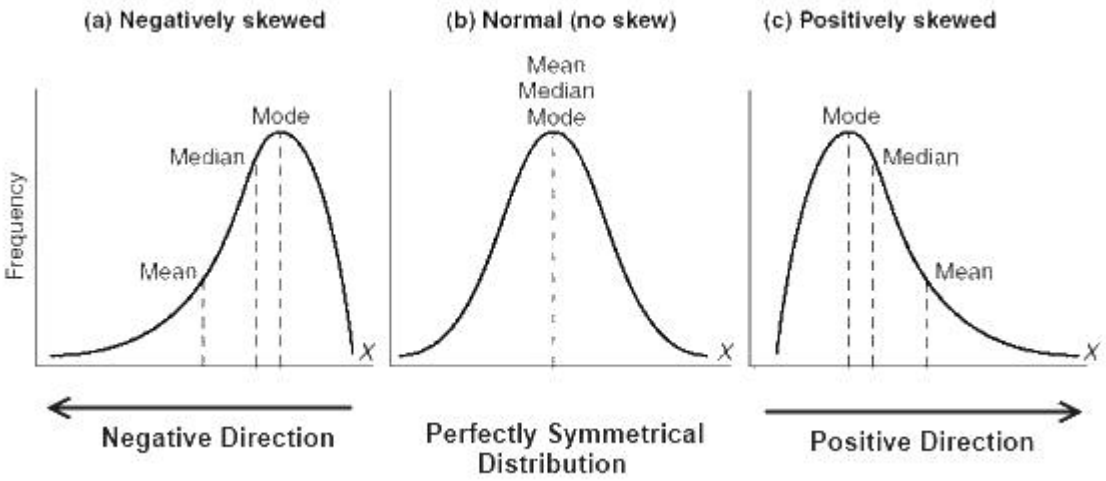
---

17) $s_{\overline{X}}$=$\overline{X}$의 표준편차, s=X의 표준편차

18) n이 커질수록 신뢰구간이 작아져서 정확한 결과를 얻게 된다.

19) 자료들이 어떤 특정한 값으로 몰리는 현상

20) 소수 고소득자의 소득에 영향을 받지 않기 때문



가) 하나의 관측치가 평균으로부터 얼마나 떨어져 있는지를 나타냄

나) 편차의 합은 0[21)]

2) 표준편차

가) 모집단의 표준편차 : $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{\frac{SS}{N}}$, 표본의 표준편차 : $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

나) 표준편차의 자유도와 불편추정치(n-1로 나누는 이유)

(1) 표본으로부터 모표준편차(σ)를 추정하는 경우 모평균(μ)을 모르는 상태이므로, 표본평균($\bar{x}$)을 사용하게 되는데 이로 인해 편의[22)]가 발생한다. → 편의에 의해 N으로 나누게 되면 모표준편차(σ)가 과소평가된다.[23)] → 불편추정치를 얻기 위해 자유도 개념[24)]을 도입하여 n-1로 나누는 것

(2) 자유도 도입 시 사례수가 1일 때 발생하는 불합리 해소 가능 : N을 사용하면 표준편차가 0이 되고, n-1을 사용하면 분모가 0이 되어 불능이 되는데 편차 자체가 존재하지 않는 상태에서는 표준편차가 0인 것보다는 불능인 것이 더 합리적이다.

3) 분산

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

나) 모집단 분산 : $\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$, 표본 분산 : $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{SS}{n-1}$

5. 표본평균의 분포와 중심극한정리

가. 표본평균의 분포

1) 표본평균($\bar{x}$)은 확률변수이기 때문에 확률분포를 가진다. → $\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$

2) $\mu_{\overline{X}} = \mu,\ \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$[25)]

---

21) $\Sigma(X_i - \overline{X}) = \Sigma X_i - n\overline{X} = \Sigma X_i - n \cdot \frac{\Sigma X_i}{n} = 0$

22) 편의 = 모평균 - 표본평균 → 오차

23) N이 n-1보다 크고 분모가 커지니까 모표준편차 추정치가 작아지는 것이 당연한 것

24) 표본평균을 사용할 경우, 평균에 맞추기 위해 사례 n개 중 하나는 정해지게 되어 자유도 상실

3) 표준오차

가) 표본평균의 표준편차

나) 측정치가 정확한 값에 얼마나 가까이 있는지를 나타냄

다) 크기 n인 표본에 대해 반복 측정한 결과($\overline{X}$)의 분산도

나. 중심극한정리

1) N이 커질수록 $\overline{X}$의 분포는 정규분포에 가까워지고, N이 충분히 큰 경우 모집단의 분포와 관계없이 $\overline{X}$는 정규분포한다.

2) 대수의 법칙 : $n \to \infty, \overline{X} \to \mu$, 중심극한정리 : n이 충분히 크면, $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$[26]

3) 중심극한정리에 의해 가능한 추론(정규분포가 가정이 되므로)

가) 표준오차단위에 따른 확률

나) 모집단 분포와 관계없이 모집단 추론 가능

다. 표본의 크기와 표본의 대표성 : 표본 ↑ → 모수에 대한 추정치 모수 근사

6. 추정과 신뢰구간

가. 의의 : 추정치가 모수치와 일치할 확률이 낮기 때문에 구간을 정하고 그 사이에 값이 존재할 확률을 생각

나. z분포에 의한 모집단의 평균 추정

1) 조건 : 모집단의 분산을 알거나 표본의 크기가 큰 경우

2) $\overline{X} \pm z \cdot s_{\overline{X}}$

다. t분포에 의한 모집단의 평균 추정

1) 조건 : 모집단의 분산을 모르며 표본의 크기가 작은 경우[27]

2) 어떤 정규분포로부터 표본이 추출된다면, t통계량은 z 통계량과 아주 흡사한 표분분포를 갖는다는 것

3) t분포는 z분포보다 변동 ↑, t분포의 변동의 정도는 n에 따라 결정

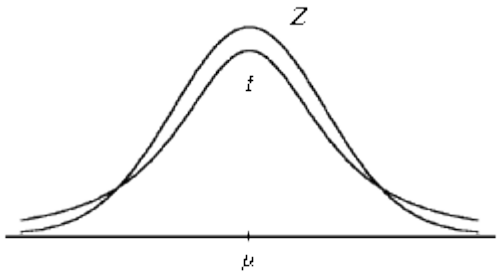
---

25) 모표준편차 ↑ → 표본평균표준편차↑, 표본수 ↑ → 표본평균의 모평균 반영 ↑ → 표본평균의 표준편차 ↓

26) $\overline{X}$의 표준점수 z가 표준정규분포한다.

27) ① 모집단이 근사한 정규분포를 이룬다면 $\overline{X}$의 분포도 근사한 정규분포를 이루는 점 이용

② $t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ 이용, z 통계량에서 모분산 대신에 표본표준편차(s)를 사용한 것



[그림 5-3] t-분포와 표준화정규분포

4) $\overline{X} \pm t\frac{s}{\sqrt{n}}$

라. 독립적인 두 모집단의 평균의 차이에 대한 추정

1) 모분산이 알려져 있거나 표본크기가 큰 경우

가) z 검정

나) $\overline{X_1} - \overline{X_2}$의 표분분포의 특성

(1) 표본크기 ↑ 근사 정규분포

(2) 표분분포의 평균 : $\mu_1 - \mu_2$

(3) 표분분포의 표준편차(표준오차) : $\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(4) 신뢰구간 : $(\overline{X_1} - \overline{X_2}) \pm z \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ (등분산가정 : $(\overline{X_1} - \overline{X_2}) \pm z\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$)

2) 모분산이 알려져 있지 않고 표본크기가 작은 경우

가) t 검정

나) 신뢰구간 :$(\overline{X_1} - \overline{X_2}) \pm t \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ (등분산가정 : $(\overline{X_1} - \overline{X_2}) \pm tS_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, $S_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$)

마. 두 평균 차이의 표본분포

1) 두 모집단이 독립이 아닌 경우 : $V(\overline{X_1} - \overline{X_2}) = V(\overline{X_1}) + V(\overline{X_2}) - 2Cov(\overline{X_1}, \overline{X_2})$

2) 두 모집단이 독립인 경우(등분산가정) : $\sigma^2_{\overline{X_1} - \overline{X_2}} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

바. 단측신뢰구간과 양측신뢰구간 : 모수가 어떤 값보다 최소한 같거나 크다는 가설을 설정할 필요가 있을 때 단측검정 시행[1)]

7. 가설검정의 이해

가. 가설의 의미 : 진위를 확인하기 위해 설정한 잠정적 판단

나. 가설의 조건

1) 명확성(조작가능성) : 공부를 잘하는 학생 X → 상위 50%에 속하는 학생

2) 반증가능성 : 참/거짓이 판별 가능해야함

다. 가설의 종류

1) 유의도 : 우연히 차이가 발생할 확률로서 기각역 또는 α라고 한다.

2) 영가설 : 두 모집단 간의 차이가 없으며 관찰된 차이는 순전히 우연에 의한 것이다.
→ 차이가 날 확률이 유의도보다 낮으면 영가설 기각
3) 대립가설 : 두 모집단 간의 차이는 우연에 의한 것이 아니다.
→ 차이가 날 확률이 유의도보다 높으면 영가설 기각 실패
4) 거증의 책임은 대립가설에 있으며, 영가설을 수용[28]한다는 표현은 쓰지 않는다.

라. 가설검정의 제한점

1) 잘못된 결과의 해석
가) 유일한 올바른 해석 : 가설을 기각했다 → 정말로 차이가 없는데 우연히 이런 차이를 나타냈을 확률이 5%이하이다.

2) 제한점
가) Fishing : 전체 연구에서 차이가 없다는 것만 뽑아 올림 → 효과 과소 평가
나) Publishing : 전체 연구에서 차이가 있다는 것만 발표 → 효과 과대 평가

8. 내·외적 타당도

가. 내적타당도

1) 의미 : 실험의 결과를 해석할 수 있게 하는 최소한의 조건으로서 다른 설명의 가능성을 배제하는 것
2) 내적타당도 저해 요소
가) 역사 : 실험 진행 중 외부적 사건으로 인해 결과 영향
(1) 예
(가) 정신교육 → (김정남 독살사건) → 반공의식 : 교육 때문인가? 사건 때문인가?
(나) 비료종류 → (산불) → 생산량 : 비료 때문인가? 산불 때문인가?
나) 성숙 : 피실험자의 내부적 변화
(1) 예
(가) IQ검사가 지루해서 마지막에 대충 풀어버림
(나) 어린아이 교육 → 지능, 1~2달동안 지능이 발달해버림
다) 검사 : 사전검사로 인해 연습효과가 생김
라) 도구 : 검사도구의 신뢰도
마) 선발 : 실험에 참가하는 집단을 어떻게 선발하였는가?
(1) 예 : 공부를 잘하는 학생들만 선발하면 X → 무작위추출
바) 탈락 : 피실험자 집단이 중간에 떨어져나감
(1) 우연적인 것 : 병원입원
(2) 체계적인 것 : 공부 못하는 학생들이 낙제되는 것은 체계적인 것으로서 결과에 영향
사) 통계적 회귀 현상 : 극단치 집합을 대상으로 하면 두 번째 결과는 평균으로 회귀한다.

나. 외적타당도

28) 영가설을 입증하는 것이 목표가 아니기 때문이다.

1) 의미 : 일반화 가능성

2) 외적타당도 저해 요소

가) 플라시보 효과 : (환자) 이 약을 **먹으면** 병이 나을거야

나) 자기성취적 예언 : (의사) 이 약을 **먹이면** 병이 나을거야

다) 존 헨리 효과 : "뭐? 나와 비교를 한다고?" → 불쾌 → 더 열심히

라) 호손 효과 : "와! 나를 관찰을 한다고?" → 신남 → 더 열심히

3) 연구의 두 갈래

가) 횡단적 연구 : 한 방에 하는 연구로서 실험적 통제에 의한 연구이다. ANOVA와 T-Test를 자주 쓴다.

나) 종단적 연구 : 표본을 장기적으로 추적하여 조사하기 때문에 문제가 생기지 않는다. 탈락한 사람들은 제외되기 때문에 문제가 생기지 않고 통제되지 않은 회귀적 상관연구를 주로 한다.

9. 가설검정의 예

가. z검정과 t검정

1) z검정

가) z검정을 위한 네 가지 조건

(1) 변수 = 양적변수

(2) 모집단의 분산을 알거나 n>30

(3) 모집단의 분포 = 정규분포

(4) 등분산성 가정 충족

나) 단일표본 z검정

(1) 한 모집단의 속성을 알기 위하여 한 표본의 통계치와 특정 수치 비교

(2) $z = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}}$

다) 두 독립표본 z검정

(1) 독립적인 두 표본을 가지고 두 모집단의 유사성 검토

(2) $z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

2) t검정

가) t검정의 조건

(1) 변수 = 양적변수

(2) 모집단의 분산을 모르는 경우

(3) 모집단의 분포 = 정규분포

(4) 등분산성 가정 충족

(5) 데이터가 서로 독립적(매우 중요)

나) 단일표본 t검정

(1) 모집단의 분산을 알지 못할 때 모집단에서 추출된 표본의 평균과 특정 수치 비교

(2) $t = \frac{\overline{X} - \mu_X}{s_{\overline{X}}}$

다) 두 독립표본 t검정

(1) $s_E = \sigma_{\overline{X_1} - \overline{X_2}} = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, $S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ (등분산 가정 충족 시)

나. 표본평균에 의한 모집단 평균 추정

1) 모분산을 알거나 표본의 크기가 큰 경우 : z 검정

2) 모분산을 모르며 표본의 크기가 작은 경우 : t 검정

다. 두 집단의 평균 차의 표분분포

1) 두 확률변수가 독립일 때

2) $\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\sigma_{\overline{X_1}}^2 + \sigma_{\overline{X_2}}^2}$

3) $V(\overline{X_1} - \overline{X_2}) = V(\overline{X_1}) + V(\overline{X_2})$[29]

라. 평균 차이의 표준편차 → t

1) 모집단이 정규분포 → 평균의 차의 표집분포 정규분포

2) $\mu_{\overline{X_1} - \overline{X_2}} = \mu_1 - \mu_2$

3) 표집평균의 표준오차 : $\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\sigma_{\overline{X_1}}^2 + \sigma_{\overline{X_2}}^2}$

4) $t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left(\frac{SS_1 + SS_2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

10. t검정의 통계학

가. t검정을 실시하기 위한 조건 검토

1) 진실험설계 : 무작위추출과 무작위할당 + 통제집단

2) 독립변수 : 명명척도 → 종속변수 : 구간척도 또는 비율척도

3) 비교하려는 두 집단의 모집단이 모두 정규분포

4) 표본간 독립성

나. 자료 분석방법의 재구성

1) $t = \frac{\text{두 집단의 차이}}{\text{오차의 표준편차}}$

2) 분자 : $(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)$

3) 분모(등분산가정) : $\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)s^{2_1} + (n_2-1)s^{2_2}}{n_1+n_2-2}}$

4) t 값이 기각역 내에 있으면(t 크리티컬보다 크면) → 영가설 기각[30]

---

29) $P(A \cap B) = P(A)P(B) \rightarrow E[AB] = E[A]E[B]$

다. 가설검정의 오류 가능성과 검정력

1) 1종 오류와 2종 오류

영가설	기각	기각 X
참	1종 오류(α)	O
거짓	검정력(1-β)	2종 오류(β)

2) 결과의 잘못된 해석

가) 상황 : 유의수준 0.05인 t검정에서 얻어진 평균치를 t점수로 전환한 값이 t크리티컬보다 커서 영가설을 기각한 경우

나) 유일한 옳은 해석 : 두 모평균이 같은데 우연히 이런 차이가 나타날 확률은 5%보다 작다.

11. 상관분석

가. 확률변수로서의 수학적 개념을 강조한 상관관계

1) 결합분포 : P(A∩B)

2) 조건부분포 : P(A | B)

3) 주변분포 : P(A), P(B)

4) 공분산

가) 의미 : 두 변수가 동시에 변하는 정도[31]

나) $Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E(XY) - E(X)E(Y)$

다) $s_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{N-1} = \frac{\sum X_i Y_i}{n-1} - \overline{XY}$

5) 상관계수 : 표준화된 공분산

가) $r = \frac{Cov(X,Y)}{S_X S_Y}$ (표본집단)[32]

나) 성질

(1) $-1 \le \rho \le 1$

(2) 지역독립성 : 두 변수에 일정한 값을 규칙적으로 더하거나 빼도 상관계수는 변하지 X

(3) 척도독립성 : 같은 값을 각기 두 변수에 곱하거나 나누어 주어도 상관계수는 변하지 X

(4) 상관계수 = 0 ↛ 통계적 독립 △ 독립 → 공분산 0 → 상관계수 0

나. 예측력을 강조한 상관계수의 설명 방법

1) $\frac{\text{회귀선이 예측한 변동량}}{\text{전체 변동량}} = \frac{\sum(\hat{Y}-\overline{Y})^2}{\sum(Y-\overline{Y})^2} = r^2$

다. 상관연구의 기본 가정과 주의점

1) 선형성 : 두 변수가 직선형이 아니라면 밀접한 관계가 있어도 상관관계 과소평가

2) 범주의 제한 : 자료가 절단되어 있으면 안된다.

3) 지수로서의 상관계수 : r^2은 한 변수의 분산이 다른 변수의 분산을 몇 % 설명하는지 나타내는 수치로 해석 가능

---

30) 우연이 작용할 확률이 유의값보다 작게되므로 영가설 기각

31) 분산은 자기 자신과의 공분산

32) 표본이므로 공분산 n-1로 나눠야함

4) 특이값 : 특이값은 상관변수에 큰 영향을 미치기 때문에 주의하여 살펴보아야 한다.

5) 상관관계 O → 인과관계 X[33)]

가) 매개변수

나) 제3변수

다) 우연

라) 인과관계를 입증하기 위해서는 실험을 해야한다.

6) 등분산성 : 등분산성 가정 X → 상관관계 ↓

12. 측정도구의 신뢰도

가. 고전 검사이론의 주요 가정

1) 점수 = 진점수 + 오차

2) 진점수와 오차는 서로 독립

3) 오차는 정규분포

4) 오차 점수는 서로 독립

나. X = T + E

1) $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$

다. 신뢰도 계수의 정의

1) 신뢰도 $= \frac{\text{진점수분산}}{\text{관찰된점수분산}}$

2) 신뢰도=상관계수인 이유(p. 198-199)

3) 측정의 표준오차 : $\sigma_E = \sigma_X\sqrt{1-r}$ (p. 199-200)

13. 단순회귀분석

가. 단순회귀모형

1) Y = α + βX (α, β는 모수치로서 관찰 X)

2) Y = a + bX 직선방정식으로 α, β 추정

나. 회귀선의 추정

1) $\sum\{Y_i - (a+bX)\}^2$를 최소화하는 값(최소제곱법)

2) a = $\overline{Y}$ - b$\overline{X}$, b = r($\frac{S_Y}{S_X}$)

다. 성질

1) $X = \overline{X}, Y = \overline{Y}$일 때 회귀선은 반드시 $(\overline{X}, \overline{Y})$를 지난다.

2) (X,Y)가 표준점수라면 b=r

3) $\frac{\text{회귀선이 예측한 변동량}}{\text{전체 변동량}} = \frac{\sum(\hat{Y}-\overline{Y})^2}{\sum(Y-\overline{Y})^2} = r^2$

라. 예측오차(회귀선에 의해 설명이 안되는 오차) : $e_i = Y_i - \hat{Y}$

33) 인과관계가 있으면 반드시 상관관계가 있다.

마. 분산(한 X값에 대한 Y의 분산) : $\sigma^2_{X|Y} = \frac{\sum(e_i - \bar{e})^2}{N}$

바. 표준오차 : $\sqrt{\sigma^2_{X|Y}}$

사. 결정계수 : $\frac{\text{회귀선이 예측한 변동량}}{\text{전체 변동량}} = \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2} = r^2 = \frac{MSS}{TSS}$

14. 다중회귀분석

가. 중회귀분석 : 한 변수의 값을 두 개 이상의 변수로 설명하는 방법[34]. 단순회귀모형보다 더 정확한 예측치를 줄 수 있다.

나. 중회귀모형 : $\hat{Y}= a+b_1X_1+b_2X_2$[35]

다. 다중공선성 : 독립변수들 간에 높은 선형관계(상관관계)가 존재하는 것으로서 모형 추정치가 불안정해지는 문제

1) 진단법[36]

가) 수정된 R^2

(1) 다중공선성 문제가 발생하는 것을 막을 수는 없지만, SPSS에서는 수정된 R^2를 활용하여 짐작한다.

(2) 수정된 R^2 : 독립변수의 수가 증가함에 따라 커지는 증가함수이기 때문에, 그 개수가 증가하면 결정계수도 커지게 된다. 이러한 단점을 보강하기 위해 사용

(3) $R^2_{adj}=1-\dfrac{SS_E/df_{SS_E}}{SS_T/df_{SS_T}}\ (df_{SS_E}=n-k-1, df_{SS_T}=n-1)$

(4) 특징

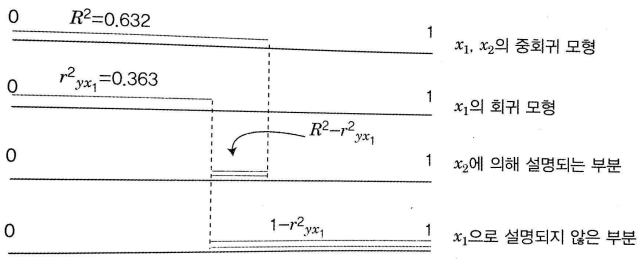
(가) $R^{2_{adj}}< R^2$

(나) k(독립변수 개수) ↓ + n(표본 수) ↑ → $R^{2_{adj}}$는 R^2에 가까워짐

나) 편상관

(1) 다른 예측변수의 영향을 제거한 특정 예측변수와 응답변수의 상관관계

(2) 편상관의 도식



(3) $r^2_{yx_2|x_1}=\dfrac{R^2-r^2_{yx_1}}{1-r^2_{yx_1}}$

2) 해결법

가) 상관관계가 높은 독립변수중 하나 혹은 일부를 제거한다.

---

34) 예측변수/설명변수 → 응답변수

35) Y의 예측치 = 상수 + 계수1*예측변수1 + 계수2*예측변수2

36) 참고

가) R^2값은 높아 회귀식의 설명력은 높지만 독립변수의 P-value 값이 커서 개별 인자들이 유의하지 않음 → 독립변수들 간에 높은 상관관계가 있다고 의심

나) 독립변수들 간의 상관관계를 구함

다) 분산팽창요인(VIF) >10 → 다중공선성 문제 O

나) 변수를 변형시키거나 새로운 관측치를 이용한다.

다) 자료를 수집하는 현장의 상황을 보아 상관관계의 이유를 파악하여 해결한다.

3) SPSS 결과분석

가) 입력/제거된 변수

입력/제거된 변수[a]

모형	입력된 변수	제거된 변수	방법
1	SES, events[b]		입력

a. 종속변수: mental

b. 요청된 모든 변수가 입력되었습니다.

(1) 독립변수 : SES, events 종속변수 : mental

나) 모형 요약

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.582[a]	.339	.303	4.556

a. 예측자: (상수), SES, events

(1) R^2와 R^2_{adj}와 큰 차이가 나지 않기 때문에 추가 검사를 도입할 필요는 없다.

(2) 추정값의 표준오차 : X에 대해 정규분포하는 Data 값이 회귀선으로부터 얼마나 떨어져 있는가 하는 표준거리로서 값이 적을수록 모형이 반응을 더 잘 예측한다.

다) ANOVA(회귀선의 유의도에 대한 검정)

ANOVA[a]

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	394.238	2	197.119	9.495	.000[b]
	잔차	768.162	37	20.761		
	전체	1162.400	39			

a. 종속변수: mental

b. 예측자: (상수), SES, events

(1) H_0 : 모든 회귀계수=0에 대한 F검정을 실시한 것으로서 F통계량과 p값을 계산

(2) 각 요소의 의미

ANOVA[a]

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	SSM	k	MSM(SSM/k)	MSM/MSE	p값
	잔차	SSE	n-k-1	MSE(SSE/n-k-1)		
	전체	SST	n-1			

라) 계수(개별 회귀계수들의 유의도에 대한 검정)

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	28.230	2.174		12.984	.000
	events	.103	.032	.428	3.177	.003
	SES	-.097	.029	-.451	-3.351	.002

a. 종속변수: mental

마) B : 회귀계수

바) 표준화 계수 : 어떤 변수가 어떤 방향으로 가장 영향력이 있는지 비교 가능

사) t와 유의확률 : 개별 회귀계수에 대한 검정[37]

라. 데이터의 구성

1) Model : $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 = 28.230 + 0.103 x_1 - 0.097 x_2$

Prediction equation: $\hat{y} = 28.23 + 0.103 x_1 - 0.097 x_2$

2) 가정 : 독립성, 정규성, 등분산성 (homoscedasticity)

가) 독립성 : 표본의 데이터는 서로 독립적이어야 한다.

나) 정규성 : 모집단의 조건부 확률분포(X에 대한 Y의 확률분포)가 정규분포해야한다.

다) 등분산성 : 모집단의 조건부 확률분포의 분산이 같다.

마. 회귀 모델의 유의도 검증

1) 가설검정

가) H_0 : $\beta_1 = \beta_2 = \cdots = 0$ or 모집단 다중상관계수[38] = 0

H_1 : 적어도 하나의 $\beta_i \neq 0$

나) $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)} = \dfrac{SS_M/k}{SS_E/n-k-1} = \dfrac{MS_M}{MS_E} (k = \text{예측변수의 수}, n = \text{전체 사례수})$

다) F값을 구하고 유의수준에 해당하는 F_{crit}값과 비교하여 귀무가설 기각 or 기각실패

라) **각 회귀계수의 t검정**이 유의하면 **귀무가설에 대한 F검정**도 일반적으로 유의[39]

2) 검정통계치의 분모 : 오차(잔차)의 제곱의 평균=분산[40]

바. 용어

1) 다중공선성 : 예측변수간 상관으로 모형추정치가 불안정해지는 문제

---

37) H_0 : $\beta_1 = 0$ → 선형관계가 존재하지 않는다. $t = \dfrac{B-0}{\text{표준오차}}$

38) 회귀선에서의 상관계수(r) = $\sqrt{\dfrac{\sum(\hat{Y}-\overline{Y})^2}{\sum(Y-\overline{Y})^2}} \rightarrow r^2 = \dfrac{X\text{들에 의해 설명되는 분산}}{Y\text{의 분산}}$

39) 모든 회귀계수들이 t값에 유의하지 않지만 F값은 유의한 경우 → 변수들의 전체 집합이 설명력을 가짐 △ 다중공선성 문제

40) z 검정에서는 표준오차

2) 상관 : 두 변수 간의 선형 관계

3) 편상관 : 다른 변수의 영향을 배제한 두 변수의 연합관계

4) 중다상관 : 여러 독립변수와 한 종속변수 사이의 상관

5) 수정결정계수 : 독립변수의 설명력을 자유도로 수정한 값

6) 분산팽창요인(VIF) : 한 독립변수와 나머지 독립변수의 중다상관으로 예측한 계수의 안정성 지수[41]

7) 더미변수 : 0과 1로써 어떤 관측치가 있음과 없음을 나타내는 인위적 변수

사. 연습문제

1) 다음은 4개의 예측변수에 따른 회사원의 급여액을 예측한 회귀분석 결과이다. 이 중 Gender는 유목변수이고 나머지는 양적변수이다. 빈 곳에 필요한 정보를 채우고 다음의 물음에 답하시오. (남자 = 1)

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F
Regression	23665352	1) 4	5916338	22.98
Residuals	2) 22656124.63	3) 88	4) 257455.96	
Total	46321476.63	92		
Coefficients				
Variable	Coefficient	s.e.	t	Sig.
Constant	3526.4	327.7	5) 10.76	0.000
Gender	6) 722.11	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
n = 93	$R^2 =$ 7) 0.511		$R^2_{adj} =$ 8) 0.489	

1) 23665352/5916338 = 4

3) 잔차제곱합 자유도 = n-k-1 = 93-4-1 = 88

4) 2) / 3)

2) 3) * 4)

5) 3526.4/327.7

6) 117.8 * 6.13

7) 1-SSM/SST = 1-23665352/46321476.63

8) 1-(23665352/4)/(46321476.63/92)

2. 교육 = 12년, 성별 = 남자(=1), 경험 = 10, 입사 개월 수 = 15 인 직원의 급여액 예측치는?

$\hat{Y} = 3526.4 + 722.11x_1 + 90.02x_2 + 1.2690x_3 + 23.406x_4$에 숫자 대입

41) VIF가 크면 표준오차 변동 ↑ → 계수 신뢰 X

$$VIF_{X_j} = \frac{1}{1 - R^2_{X_j}}$$

3. 다음에 제시된, 모형별 회귀분석 결과표를 보고, 이에 관한 설명 중 잘못된 곳을 지적하시오.

	Model		
Estimates	$E(y)=\alpha+\beta x_i$	$E(y)=\alpha+\beta_1 x_i+\beta_2 x_2$	$E(y)=\alpha+\beta_1 x_i+\beta_2 x_2+\beta_3 x_3$
β_1	0.450	0.400	0.340
β_2		0.003	0.002
β_3			0.002
R^2	0.25	0.34	0.38

a. y와 x_1 사이에는 정적 상관관계가 있다.

→ T, 계수가 양수이기 때문

b. 다른 변수의 영향을 제외한다면, x_1의 값 한 단위 당 y 값은 0.45 변화한다.

→ 계수에 대한 일반적인 설명

c. x_1 이 가장 강력한 partial effect를 미치고 있다.

→ partial effect는 두 변수의 영향이 겹치는 부분인데 주어진 자료로는 변수들 간 관계를 알 수 없다. 계수가 크다고 하여 알 수 있는 것이 아니다.

d. r^2_{yx3} 값은 0.40이다.

→ 변수들 간 상관관계는 알 수 없다.

15. 분산분석

가. 분산분석의 이해

1) t-검정 : 평균 검정 + 평균 차이 검정

2) 분산분석 : 평균비교

3) 분산분석의 의미 : 이 분석에 의한 유의도 검증이 두 종류의 분산[42]을 비교하는 것

4) 분산분석은 집단의 수가 둘 이상인 경우 사용할 수 있는 방법으로 t-검정을 일반화[43]

5) F 검정을 하는 이유 : t-검정을 여러번 실시하는 경우 1종 오류 가능성(통합 기각역)이 높아진다.[44]

6) 검정력을 높이는 방법

방법	효과
n ↑	F값 ↑(검정력 ↑)
α ↑	
처치효과 ↑	
표준오차 ↓	

나. F 분포

1) 모집단에서 무작위로 추출한 F값과 그 확률 나타낸 분포

2) F분포는 두 개의 자유도를 가지고 있음

---

42) $\frac{\text{두 집단 간 평균의 차이의 분산}}{\text{집단 내 개체들의 분산}}$

43) 집단 수가 2인 경우 $F=t^2$

44) $\alpha_{FW}=1-(1-\alpha)^c$

가) 분자(numerator)의 사례수가 n_1인 경우 자유도는 n_1-1

나) 분모(denominator)의 사례수가 n_2인 경우 자유도는 n_2-1

3) F분포는 항상 양수 : 두 분산의 비

4) F분포의 중앙값은 1에 가까우며 항상 정적편포를 이룬다.

다. 분산분석과 회귀분석의 관계

1) 더미변수를 이용하면 분산분석과 회귀분석이 같은 것

2) 일원설계 : $\hat{Y}=\hat{\mu}_3+(\hat{\mu}_1-\hat{\mu}_3)z_1+(\hat{\mu}_2-\hat{\mu}_3)z_2$를 계산한 뒤 $F=\frac{\hat{Y}-\overline{Y}\text{의 제곱합}}{Y-\hat{Y}\text{의 제곱합}}$

3) 이원설계 : 상호작용을 가정하지 않은 경우와 가정한 경우로 구분

가) 상호작용을 가정하지 않은 경우 : 각각 요인의 주효과에 대한 검정

나) 상호작용을 가정한 경우 : 상호작용에 대한 검정 or 오차로 편입된 상호작용 추출

4) 공분산분석 : 상호작용 가정한 회귀모델을 이용하여 회귀선 기울기 동일성 검정

가) 공변수와 독립변수의 상호작용이 없어야 회귀선의 기울기가 같은 것으로 드러남

라. 용어

1) 일원분산분석

가) 독립변수 = 요인(factor)

나) 수준, 처리 = 요인이 갖는 값

다) 분산분석의 종류

(1) 요인 1개, 종속변수 1개 → 일원배치 분산분석

(2) 요인 2개, 종속변수 1개 → 이원배치 분산분석

라) 더미변수 : 0과 1로서 유목변수의 존재유무를 나타내는 회귀분석 변수

마) 다중비교 : 통합분산분석 이후 각 집단 평균을 비교하는 방법

2) 이원분산분석

가) 요인설계 : 각 변수의 수준이 다른 변수의 수준과 관련되어 있는 실험설계

나) Factors : 실험설계의 변수를 요인이라고도 함

다) 수준 : 요인의 특정한 값

라) Cell : 한 변수와 특정 수준의 조합

마) 메타분석 : 같은 주제의 연구결과를 종합한 것

16. 일원배치분산분석

가. 개념

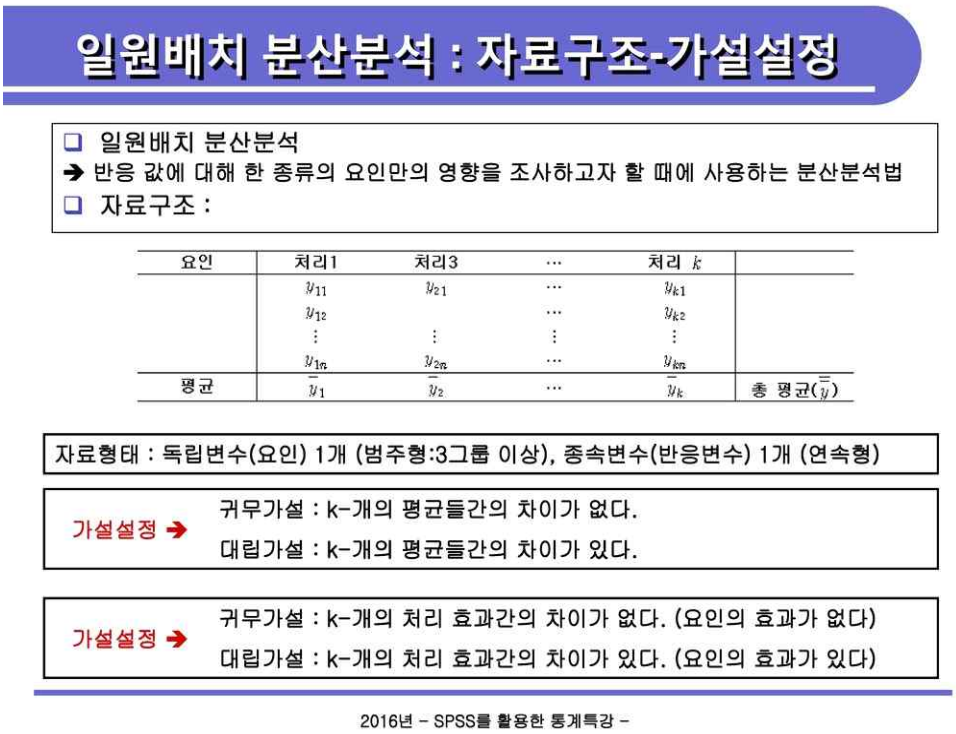
1) 일원설계는 일원배치법, 완전임의배치법 등의 용어와 같은 뜻으로 사용

2) 어떤 관심있는 하나의 인자(factor)의 영향을 조사하기 위해 쓰이는 가장 단순한 실험계획법

3) 조건 : 전집단의 무선표집·무작위할당 + 통제집단(비교집단)의 존재

4) 두 집단 이상이 평균차를 모집단의 평균차이 때문인 것으로 볼 수 있는지 통계적으로 확인

나. 자료구조



다. 모형 : $X=\mu+\alpha+\epsilon,\ \epsilon\sim N(0,\sigma^2)$

$$SST=SSA+SSE,\ \sum_i\sum_j(y_{ij}-\bar{y})^2=\sum_i n_i(\bar{y}_i-\bar{y})^2+\sum_i\sum_j(y_{ij}-\bar{y}_i)^2$$

라. 가정 : 독립성, 정규성, 등분산성

1) 독립성 : 집단들은 서로 독립이고, 집단 내의 개체들도 서로 독립이어야 한다.[45)]

2) 정규성 : 모집단 또는 표본평균의 분포가 정규분포를 따라야 함[46)]

3) 등분산성 : 모집단의 분산이 모두 같다.

마. 모델의 유의도 검증

1) 가설검정

가) H_0 : $\mu_1=\mu_2=\cdots=\mu_k$ or $all\ \alpha_i=0$

H_1 : 적어도 하나의 μ가 다르다 or $|all\ \alpha_i=0$

나) $F=\dfrac{\text{집단간평균의차이}}{\text{집단내점수의차이}}=\dfrac{\text{집단간분산}}{\text{집단내분산}}=\dfrac{SS_B/k-1}{SS_W/n-k}=\dfrac{MS_B}{MS_W}=\dfrac{\text{처치효과}+\text{오차}}{\text{오차}}$

$(k=\text{집단의수},\ n=\text{전체사례수})$

2) 분산분석표

45) 정규성과 정분산성은 현실적으로 지켜지지 않아도 분석이 민감하지 않지만, 독립성 위배는 1종오류 가능성을 크게 높인다.
46) n>30 → 중심극한정리에 의해 자동으로 만족

변동요인	제곱합(SS)	자유도(df)	평균제곱합(MS)	F	p
집단 간	SSA	k-1	SSA/k-1	MSA/MSE	
집단 내	SSE	n-k	SSE/n-k		
전체	SST	n-1			

3) 검정통계치의 분모 : 집단 내 분산(오차분산)

4) F값을 구하고 유의수준에 해당하는 F_{crit}값과 비교하여 귀무가설 기각 or 기각실패

5) 귀무가설을 기각하는 경우 다른 집단과 평균이 다른 집단을 찾기 위해 추가적인 분석 (사후검정, 다중비교) 실시 → 통합 1종 오류 경계

바. 연습문제

1) 다음 문장의 진위를 판단하여 T, F로 표시하시오 (1-5).

1. One-way ANOVA에서 $df_B + df_W = N - 1$의 등식이 성립한다.

→ T : k-1+n-k = n-1

2. 3개 이상의 집단에 대해 여러 차례의 t-test를 실시하면 Type 2 Error가 증가하게 된다.

→ F : Type 1 Error가 증가하게 되는 것

3. $s_B^2 + s_W^2 = s_T^2$

→ F. 분모의 자유도가 서로 다르기 때문에 성립 X

4. $SS_B + SS_W = SS_T$

→ T : SST = SSA + SSE

5. 비교하는 집단수가 2일 때 F 검정이 t 검정보다 더 검정력이 크다.

·→ F : 비교하는 집단 수가 2이면 F 검정과 t 검정은 같다.[47)]

2) 맞는 것을 보기에서 고르시오 (6-10).

6. 다음 중 F 분포에 관한 설명으로 맞지 않는 것은?

a. F는 두 변수의 분산의 비율에 관한 통계량이다. T

b. F 값은 음수가 될 수 없다. T

47) 집단 수가 2이면 $F = t^2$

c. F 분포는 부적편포를 나타낸다. F : F분포는 정적편포[48)]

d. F 분포는 두 가지 df 값에 따라 고유한 값을 나타낸다. T

7. 일원분산분석의 설명으로 맞지 않는 것은?

a. F 검정은 각 집단평균 중 유의미한 차이가 있는 짝이 있는지에 관한 전반적인 비교결과를 알려준다. T

b. 1종 오류 가능성 α 값을 유지하기 위해 t 검정 대신 사용된다. T

c. 독립변인이 분산 아닌 평균에만 영향을 미친다는 가정을 갖는다. T

d. 일원분산분석의 대립가설은 등가설의 형태를 취한다. F : 부등가설

8. F 값에 관한 설명으로 옳지 않은 것은?

a. 만일 $F<1$ 이라면 이것은 H_0 가 참인 것을 의미한다.

T : $F=\frac{\text{처치효과}+\text{오차}}{\text{오차}}$, F<1이면 처치효과가 없고, 따라서 영가설을 기각할 수 없게 된다. 대립가설이 참이라면 F>1

b. 집단 수 = 2 일 때 $F=t^2$ 이 된다. T

c. 독립변수의 효과가 커질수록 s_B^2 의 값은 커진다. T

d. F 값이 크다는 것은 처치효과의 크기가 크다는 것을 의미한다.

→ F : 사례 수가 많거나, 표준오차가 작아도 F값은 클 수 있다.

9. 다음 중 ANOVA의 가정이 아닌 것은?

a. 각 집단의 사례수는 같다.

b. 각 집단의 모집단은 정규분포를 이룬다.

c. 각 집단의 모집단 분산은 서로 같다.

d. 각 집단은 무작위 표본이며 표본의 데이터는 서로 독립적이다.

10. 분산분석의 검정력에 관한 설명 중 맞지 않는 것은?

a. 표본의 사례수가 커지면 검정력도 커진다. T : $F=(SS_B/k-1)/(SS_E/n-k)$

b. 유의수준 α 값이 작아지면 검정력은 커진다. F : α ↓ → β ↑ → 1-β ↓

c. 독립변수의 실제효과가 크면 검정력도 크다. T : 처치효과 ↑ → F값 ↑

d. 표본의 편차가 커지면 검정력은 작아진다. T : 표준오차 ↑ → F값 ↓

11. $s_B^2=37.9$, $s_W^2=44.5$ 일 때 F 값은?

48) F는 어려운 개념이다 → 어려운 시험은 정적편포를 이룬다 → F 분포는 정적편포

37.9/44.5=0.852

12. 어떤 과학자가 암세포 억제제의 효과를 실험하기 위해 동일한 종류의 암세포 24개를 배양하여 4개 집단에 6개씩 무선 할당하여 각 집단별로 한 집단은 통제집단으로 하고 나머지 3집단에 대하여 3 종류의 억제제를 주사하여 일주일 후 형성된 암세포의 수를 비교하여 일원분산분석으로 통계 처리하여 다음과 같은 결과를 얻었다.

	제곱합	df	평균제곱	F	유의확률
집단간	2590.458	3	863.468	39.205	.000
집단내	440.500	20	22.025		
합계	3030.958	23			

1) 이 실험의 영가설은? $\mu_1 = \mu_2 = \mu_3 = \mu_4$
2) 대립가설은? 적어도 한 집단의 모평균이 다르다.
3) SS_B 값은? 2590.458
4) SS_W 값은? 440.500
5) SS_T 값은? 3030.958
6) 결론은 무엇인가? 억제제의 효과가 있다.

13. 다음은 세 집단 분산분석을 위한 데이터이다. 일원분산분석의 모형은

$X = \mu + \alpha + \epsilon$

로 나타난다. 이 표의 6번 사례 점수(7점)를 모형에 의하여 분할한다면? 괄호 안을 채우시오.

case	group	score
1	1	9
2	1	8
3	1	13
4	1	10
	집단평균	10.00
5	2	5
6	2	7
7	2	8
8	2	4
	집단평균	6.00
9	3	5
1	3	3
11	3	6
12	3	6
	집단평균	5.00
	전체평균	7.00

$X = \mu + \alpha + \epsilon = \mu + (\overline{Y}_i - \mu) + (Y_i - \overline{Y}_i)$49)

7 = 7 + (6-7) + (7-6)

→ 점수는 다음과 같이 설명

전체 평균이 갖는 효과 + 자기가 속한 집단의 효과(집단 간 차이) + 오차(집단 내 차이)

각각 모두 더한 다음 자유도로 나누면 곧 분산이 된다.

14. 다음 분산분석 결과표의 빈 곳을 메우시오.

Source	SS	df	MS	F
Between	184.133	2	92.067	1.960
Within	563.600	12	46.967	
Total	747.733	14		

15. 일원분산분석에서 어떤 경우에 F 값이 0 또는 무한대가 될 수 있을까?

$$F = \frac{\text{집단 간 평균의 차이}}{\text{집단 내 점수의 차이}} = \frac{\text{집단 간 분산}}{\text{집단 내 분산}} = \frac{SS_B/k-1}{SS_W/n-k} = \frac{MS_B}{MS_W} = \frac{\text{처치효과} + \text{오차}}{\text{오차}}$$

$(k = \text{집단의 수}, n = \text{전체 사례수})$

F = 0 : 집단 간 평균이 같을 때

F → 무한대 : 집단 내 데이터가 같을 때

16. 각 집단의 표본평균, 표준편차, 사례 수를 알고 있을 때 분산분석을 통한 F 값을 구할 수 있을까? 아니면 다른 정보가 더 필요할까?

충분하다(SST=SSA+SSE를 떠올리며 분산분석표를 그려볼 것)

답

1. T 2. F 3. F 4. T 5. F 6. c 7. d 8. d 9. a 10. b

11. 0.852

12.

1) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
2) 적어도 한 집단의 모평균은 다르다.
3) 2590.458
4) 440.500
5) 3030.958
6) 억제제의 효과가 있다.

49) 항상 빼는 순서는 개별자료-모평균

13. 7 = 7 + (6 − 7) + (7 − 6)
14.
    (a) 563.600
    (b) 12
    (c) 92.607
    (d) 46.967
    (e) 1.960
15. 집단내의 데이터가 모두 같은 값일 때, 집단 평균이 모두 같을 때
16. 충분하다.

17. 이원배치분산분석
  가. 이원설계의 개념(장점)
    1) 연구의 돌파구(상호작용 효과의 검증)
      가) 상호작용 : 한 독립변수의 효과가 다른 독립변수의 수준에 따라 달리 나타나는 현상
      나) 실험결과 누적 → 통합된 결론 X → 연구의 전제·가정 재검토 → 상호작용 고려 → 연구의 돌파구
    2) 경제성 : 일원배치 분산분석의 경우보다 같은 수의 피험자로 더 많은 결과를 얻을 수 있음
    3) 실험적 통제 : 이원설계를 통해 집단의 동질성을 높여 검정력을 높일 수 있음
      가) 일원설계를 하게 되면 독립변수와 종속변수 사이에 끼어들어 독립변수의 순수한 효과를 판단하기 어렵게 만드는 외재변수를 통제하기에 한계가 있다.
      나) 외재변수가 존재함으로서 집단 내 분산이 증가하고 이로 인해 검정력이 떨어지는 효과가 있는데, 집단 내 특정 수준으로(예 : 아이큐 120-140) 관측대상을 제한함으로서 오차 분산을 줄이는 실험적 통제를 활용하여 검정력을 높이는 방법이 존재한다.
      다) 그러나 이 방법은 외적 타당도 문제를 발생시키므로, 차라리 그 외재변수를 또 하나의 독립변수로 도입하는 이원설계를 도입하여 오차 분산을 통제하자는 것
    4) 일반화 가능성 : 실험적 통제 대상을 관찰대상으로 하여 결과의 일반화 가능성 높임
      가) 내적타당도[50]를 확보하기 위해 외재변수를 통제하게 되면 통제된 환경에만 적용되는 결과를 얻게 되어 일반화 가능성이 떨어지게 된다.
      나) 이원설계는 이러한 변수를 통제하는 대신에 독립변수에 포함시키기 때문에 보다

50) 집단 간 종속변수의 차이가 순전히 독립변수의 효과에 의한 것으로 해석할 수 있는 조건을 의미

현실에 가까운 실험 환경을 만들어 줄 수 있고, 일반화 가능성 역시 올라가게 된다.

나. 자료구조

이원배치 분산분석 : 자료구조-가설설정

❑ 이원배치 분산분석
➔ 반응 값에 대해 두 종류의 요인의 영향을 조사하고자 할 때에 사용하는 분산분석법
❑ 자료구조 :

요인	B1	B2	…	Bq	
A1	y_{11}	y_{12}	…	y_{1q}	$\overline{y_{1.}}$
A2	y_{21}		…	y_{2q}	$\overline{y_{2.}}$
⋮	⋮	⋮	⋮	⋮	⋮
Ap	y_{p1}	y_{p2}	…	y_{pq}	$\overline{y_{p.}}$
평균	$\overline{y}_{.1}$	$\overline{y}_{.2}$	…	$\overline{y}_{.q}$	총 평균($\overline{\overline{y}}$)

자료형태 : 독립변수(요인) 2개 (범주형:3그룹 이상), 종속변수(반응변수) 1개 (연속형)

가설설정 ➔
귀무가설 : 요인(A)의 평균들간의 차이가 없다. (요인A의 효과가 없다)
대립가설 : 요인(A)의 평균들간의 차이가 있다. (요인A의 효과가 있다)

가설설정 ➔
귀무가설 : 요인(B)의 평균들간의 차이가 없다. (요인B의 효과가 없다)
대립가설 : 요인(B)의 평균들간의 차이가 있다. (요인B의 효과가 있다)

2016년 – SPSS를 활용한 통계특강 –

반복-이원배치 분산분석 : 자료구조

❑ 반복-이원배치 분산분석
➔ 반응 값에 대해 두 종류의 요인의 영향을 조사하고자 할 때에 사용하는 분산분석법
❑ 자료구조 :

요인	B1	B2	…	Bq	
A1	y_{111} ⋮ y_{11r}	y_{121} ⋮ y_{12r}	…	y_{1q1} ⋮ y_{1qr}	$\overline{y_{1..}}$
	$\overline{y_{11.}}$	$\overline{y_{12.}}$	…	$\overline{y_{1q.}}$	
A2	y_{211} ⋮ y_{21r}	y_{221} ⋮ y_{22r}	…	y_{1q1} ⋮ y_{2qr}	$\overline{y_{2..}}$
	$\overline{y_{21.}}$	$\overline{y_{22.}}$	…	$\overline{y_{2q.}}$	
⋮	⋮	⋮	⋮	⋮	⋮
Ap	y_{p11} ⋮ y_{p1r}	y_{p21} ⋮ y_{p2r}	…	y_{pq1} ⋮ y_{pqr}	$\overline{y_{p..}}$
	$\overline{y_{p1.}}$	$\overline{y_{p2.}}$	…	$\overline{y_{pq.}}$	
평균	$\overline{y}_{.1.}$	$\overline{y}_{.2.}$	…	$\overline{y}_{.q.}$	총 평균($\overline{\overline{y}}$)

2016년 – SPSS를 활용한 통계특강 –

다. 모형 : $X=\mu+\alpha+\beta+\alpha\beta+\epsilon,\ \epsilon\sim N(0,\sigma^2)$

$SST=SSA+SSB+SSAB+SSE$

$$\sum_i\sum_j\sum_k(y_{ijk}-\overline{y})^2=bn\sum_i(\overline{y}_{i.}-\overline{y})^2+an\sum_j(\overline{y}_{j.}-\overline{y})^2+n\sum_i\sum_j(\overline{y}_{ij}-\overline{y}_{i.}-\overline{y}_{j.}+\overline{y})^2+\sum_i\sum_j\sum_k(y_{ijk}-\overline{y}_{ij})^2$$

$a=$ 요인a 집단 수, $b=$ 요인b 집단수, $n=$ 전체 사례수

라. 가정 : 독립성, 정규성, 등분산성

1) 독립성 : 집단들은 서로 독립이고, 집단 내의 개체들도 서로 독립이어야 한다.[51)]

2) 정규성 : 모집단 또는 표본평균의 분포가 정규분포를 따라야 함[52)]

3) 등분산성 : 모집단의 분산이 모두 같다.

마. 모델의 유의도 검증

1) 가설검정

가) 가설

(1) H_0 : $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

H_1 : 적어도 하나의 $\alpha_i \neq 0$

(2) H_0 : $\beta_1 = \beta_2 = \cdots = \beta_b = 0$

H_1 : 적어도 하나의 $\beta_j \neq 0$

(3) H_0 : $(\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$

H_1 : 적어도 하나의 $(\alpha\beta)_{ij} \neq 0$

나) 분산분석표

분산원	제곱합(SS)	자유도(df)	평균제곱합(MS)	F	p
A	SSA	a-1	SSA/a-1	MSA/MSE	
B	SSB	b-1	SSE/b-1	MSB/MSE	
AB	SSAB	(a-1)(b-1)	SSAB/(a-1)(b-1)	MSAB/MSE	
E	SSE	n-ab	SSE/n-ab		
전체	SST	n-1			

다) 검정통계치의 분모 : 집단 내 분산(오차분산)

라) 각 영가설에 대한 F값을 구하고 유의수준에 해당하는 F_{crit}값과 비교하여 귀무가설 기각 or 기각실패 → 상호작용의 존재 여부에 따라서 해석이 달라짐

마) 상호작용

(1) 상호작용이 없을 때의 해석 : 주효과 그대로 해석

(2) 상호작용이 있을 때의 해석

(가) 순항상호작용 : 순서가 바뀌지 않아 주효과가 존재한다는 일반적 해석 가능

(나) 역순항상호작용 : 순서가 바뀌어 한 요인의 효과를 일반적으로 단정할 수 없고 요인별로 다른 설명

(3) 상호작용이 있을 때 독립변수의 효과를 계속해서 탐구하고자 할 때

(가) 단일요인 효과분석 : 요인설계를 분할하여 일련의 일원분산분석으로 환원[53)]

(나) 상호작용효과 분석 : 요인설계를 보다 작은 요인설계로 나누는 것

---

51) 정규성과 정분산성은 현실적으로 지켜지지 않아도 분석이 민감하지 않지만, 독립성 위배는 1종오류 가능성을 크게 높인다.

52) n>30 → 중심극한정리에 의해 자동으로 만족

53) 두 개의 일원설계로 나누거나, 셀을 다 풀어서 하나의 일원분산분석을 실시할 수 있음. 요컨대, 상호작용이 발견되면 분석은 일단 중지하고 후속 연구를 다시 생각해보는 것이 정수순

바. 다중공선성과 상호작용

1) 다중공선성 : 독립변수 사이에 상관관계가 존재(예: 평수, 침대 수, 방 수 → 집값)

2) 상호작용 : 독립변수 사이에는 상관관계가 없으나 종속변수와 관련되는 양상이 서로 독립적이지 않음(예[54]: 교통수단, 시간대 → 통학시간)

사. 연습문제

1. (Interaction) effect occurs when the effect of one factor is not the (same) at all levels of the other factor.

2. 다음 식을 언어적으로 표현한다면?

$$SS_T = SS_A + SS_B + SS_{AB} + SS_W$$

총 변동량 = A효과 변동량 + B효과 변동량 + 상호작용 변동량 + 오차 변동량

54) 독립변수인 교통수단과 시간대는 서로 독립적이지만 요인의 수준에 따라 종속변수인 통학시간이 다르게 나타난다.

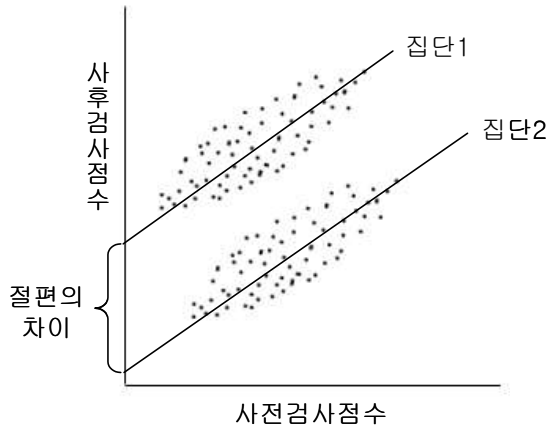
3. 다음 Table을 완성하시오. 단 각 셀에는 5명의 사례가 들어있음.
a=3, b=4, n=12*5=60

Source	SS	df	MS	F
A	1082.358	2	2164.716	0.014
B	2158.364	3	6475.092	0.041
A X B	785.228	6	4711.368	0.030
Within-cells	3290.875	48	157962.000	
Total		59		

4. 한 연구자가 나이가 들면 기억력이 감퇴하는지 이원 변량분석 (3 X 2 요인설계)으로 검증하고자 했다. 나이에 따른 세 수준, 30대, 40대, 50대 집단, 과제의 난이도에 따른 어려움, 쉬움의 두 수준으로 구성된 설계이다. 이 실험에 사용될 영가설은 무엇인가?
연령에 따른 기억력 감퇴 차이가 없다.
난이도에 따른 기억력 감퇴 차이가 없다.
나이와 난이도의 상호작용이 없다.

8. Two-way ANOVA에서는 3개의 F 값을 얻게 된다. T

18. 공분산분석
가. 개념
1) 공분산분석의 개념 모형



2) 변수 설명
가) 독립변수 : 프로그램 효과(두 직선)
나) 종속변수 ; 사후검사 점수
다) 공변수 : 사전검사 점수
3) 분산분석이 Y값의 평균을 비교하는 것이라면 공분산분석은 두 집단 점수의 회귀선이

만들어내는 Y절편의 차이가 통계적으로 의미 있는지를 검증하는 작업

4) 공분산분석은 일반 분산분석에서는 집단 내 분산에 속했을 공분산이 설명하는 부분을 제외하고 남은 부분의 분산분석이므로 오차항을 줄이는 효과를 낸다.

나. 모형 : $Y=\mu+\alpha_i+\beta(X_{ij}-\overline{X})+\epsilon_{ij}$

점수 = 모평균 + 트리트먼트 + 회귀(공변수)효과 + 오차

다. 가정 : 독립성, 정규성, 등분산성 + 추가 가정

1) 독립성 : 집단들은 서로 독립이고, 집단 내의 개체들도 서로 독립이어야 한다.[55]

2) 정규성 : 모집단 또는 표본평균의 분포가 정규분포를 따라야 함[56]

3) 등분산성 : 모집단의 분산이 모두 같다.

4) 선형성 : 종속변수와 공분산 사이의 직선형 관계

5) Equal slope : 회귀선의 기울기가 같음

6) 외삽 금지 : 비교하려는 두 집단의 공변수 범위가 서로 겹쳐야 함

라. 모델의 유의도 검증

1) 가설검정

가) 전제조건의 검정: 상호작용을 허용한 회귀모형의 회귀계수의 동일성 검정

(1) H_0 : all $\beta_1=\beta_2=....=\beta_j$

H_1 : at least one β is different

(2) 자유도: (k - 1), (N - 2k)

(3) $F=\frac{MS_{B.reg.}}{MS_{W.reg.}}$

나) 가설검정 2: 효과의 검정

(1) H_0 : all α_i=0;

H_1 : at least one $\alpha_i\neq 0$

(2) 자유도: $df_T=k-1$ $df_E=df_W=N-k-1$

(3) $F=\frac{MS_{B(adj.)}}{MS_{W(adj.)}}$

다) 검정통계치의 분모: 집단 내 분산 $MS_E=MS_{W(adj.)}$

라) 분산분석표

변량원	수정된 제곱합	자유도	수정된 제곱평균	F
집단 간	$SS_{B(adj.)}$	$k-1$	$SS_{B(adj.)}/df$	$\frac{MS_{B(adj)}}{MS_{W(adj)}}$
집단 내	$SS_{W(adj.)}$	$n-k-1$	$SS_{W(adj.)}/df$	

55) 정규성과 정분산성은 현실적으로 지켜지지 않아도 분석이 민감하지 않지만, 독립성 위배는 1종오류 가능성을 크게 높인다.
56) n>30 → 중심극한정리에 의해 자동으로 만족

마. 용어

1) 통계적 오차 통제 : 종속변수에서 공분산으로 설명되는 부분을 제외함으로서 검정력을 높임
2) 회귀선의 기울기의 동일성 검정 : 상호작용을 허용한 회귀분석이나 2원 분산분석으로 상호작용 효과 검정

바. 예상문제

1) 공분산분석을 해서 검정력이 증가하는 경우와 떨어지는 경우를 예를 들어 설명해보시오.

   어떤 학습 프로그램의 효과(독립변수)를 알아보기 위하여 사전검사(공변수)와 사후검사(종속변수)를 실시하는 실험을 생각해볼 수 있다. 사전검사와 사후검사의 상관관계가 클 경우 자료들이 회귀선을 중심으로 더 모일 것이다. 이는 집단 내 분산이 줄어드는 것을 의미하고 따라서 검정력이 증가한다. 공분산분석을 실시할 때, 사전검사와 사후검사의 상관관계가 강하지 않을 경우 검정력도 떨어지고 자유도만 잃는다.

2) 왜 공분산분석을 하면 검정력이 늘어나나?

   일반 분산분석을 했다면 집단 내 분산에 속했을 공분산이 설명하는 내용을 제외하고 남은 부분에 대한 분산분석이기 때문에 오차항이 줄어든다.
