

경영경제 통계학 노트: 과제 답안

1장: 통계학과 자료의 이해

1.

명목척도	성별(남/여), 직업군(사무직/생산직/전문직), 주민등록번호, 선수의 등 번호, 좋아하는 색상(빨강/파랑/노랑), 흡연여부(Yes/No), 학번, 신입/경력직, 우편번호, 지역구분(서울/경기)
서열척도	사회계층(상/중/하), 마라톤 순위, 브랜드 선호순위(1/2/3), 학년(1/2/3/4), 학교성적(A/B/C/D), 기업규모(대/중/소)
등간척도	체온, 기온, IQ, <u>주가지수*</u>
비율척도	연령, 월소득, 체중, 수영 시간, 높이뛰기 거리, 어제 운동시간, 매출액, 직원수, 연봉, 결근일수, 불량품수, 방문고객수, 교통사고 건수

* 지수(指數): 물가나 임금 등 변화하는 상황을 알기 쉽도록 보이기 위해 어느 해의 수량을 기준 100으로 하고, 상대적인 다른 해의 수량을 비율로 나타낸 수치

주가지수: 특정시점을 기준으로 시가총액을 비교해서 지수를 산출하는 방식

(참고) 우리나라의 종합주가지수(KOSPI) 산출방법: 한국거래소의 유가증권 시장에 상장된 회사들의 시가총액을 기준시점에 비교해서 나타낸 시장지표. 1980년 1월 4일을 기준시점으로 비교시점의 시가총액을 지수화 ($KOSPI = (\text{비교시점의 시가총액} / \text{기준시점의 시가총액}) \times 100$)

3. (1) 서열; (2) 범주형; (3) 할 수 없다

2장: 자료의 시각화

1.

변수의 수	자료유형	시각적 표현 방법
단변량 (univariate)	질적자료	도수분포표, 막대그래프, 파이차트
	양적자료	계급구간 이용 도수분포표, 히스토그램, 줄기-잎 그림, 상자-수염 그림, 누적백분율곡선

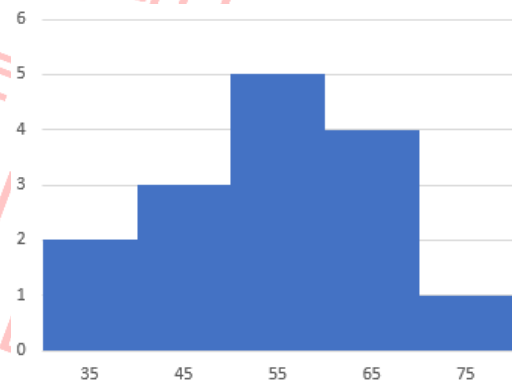
2. [생략]

3.

순서(저-고)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
점수	22	25	34	36	38	41	45	46	47	49	54	55	57	59	67
편차	-23	-20	-11	-9	-7	-4	0	1	2	4	9	10	12	14	22

(1)

구간	빈도	상대빈도	누적상대빈도
(20, 30]	2	0.13	0.13
(30, 40]	3	0.20	0.33
(40, 50]	5	0.33	0.67
(50, 60]	4	0.27	0.93
(60, 70]	1	0.07	1.00
계	15	1.00	-



(2)

2	25
3	468
4	15679
5	4579
6	7

3장: 자료의 수치요약

1. [생략]

2.

(1) 평균과 중앙값, 최빈값

- 평균: 34
- 중앙값: 30
- 최빈값: 30

(2) 범위, 사분위범위, 분산, 표준편차, 변동계수

- 범위: $80 - 10 = 70$
- 사분위범위(IQR): 위치의 수식을 이용한 결과와 EXCEL 함수를 이용한 결과가 상이할 수 있으니 주의하세요!!

① 위치의 수식 이용: $50 - 30 = 20$

- 1사분위(Q1): 30 수식 이용: $L_{25} = (11+1)(25/100) = 3$

- 3사분위(Q3): 50 수식 이용: $L_{75} = (11+1)(75/100) = 9$

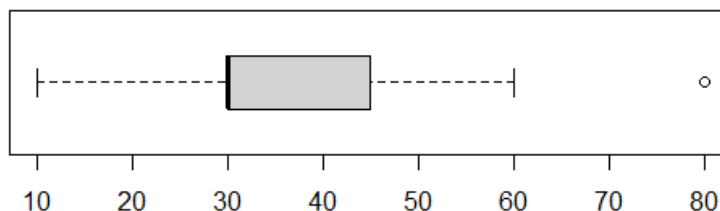
② EXCEL 함수 이용: $50 - 30 = 20$ (0, 1 제외) 또는 $45 - 30 = 15$ (0, 1 포함)

순서		1	2	3	4	5	6	7	8	9	10	11	
점수		10	20	30	30	30	30	40	40	50	60	80	
0, 1 제외	0			.25			.5			.75			1
0, 1 포함		0		.25			.5		.75			1	

- 0, 1 제외: QUARTILE.EXC(array, 1) 또는 PERCENTILE.EXC(array, .25) → Q1=30, Q3=50
- 0, 1 포함: QUARTILE.INC(array, 1) 또는 PERCENTILE.INC(array, .25) → Q1=30, Q3=45
- 기존의 함수 QUARTILE과 PERCENTILE은 0, 1 포함(INC)으로 계산됨.
- 소프트웨어의 함수의 계산 방법에 따라 상이한 결과가 산출될 수 있으니, Q1과 Q3를 개념적으로 이해함에 초점을 두기 바랍니다. (데이터가 크면 차이는 무시할 수 있을 정도로 미미하게 될 것입니다.)

- 분산: 376.36
- 표준편차: $\sqrt{376.36} = 19.40$
- 변동계수(CV): $19.40/34 = 0.57$

(3) 상자-수염그림



3.

(1) 평균과 중앙값, 최빈값

- 평균: 45

- 중앙값: 46
- 최빈값: N/A(Not available)

(2) 범위, 사분위범위, 분산, 표준편차, 변동계수

- 범위: $67-22=45$
- 사분위범위(IQR): 위치의 수식을 이용한 결과와 EXCEL 함수를 이용한 결과가 상이할 수 있으니 주의하세요!!

① 위치의 수식 이용: $55-36=19$

- 1사분위(Q1): 36 수식 이용: $L25=(15+1)(25/100)=4$

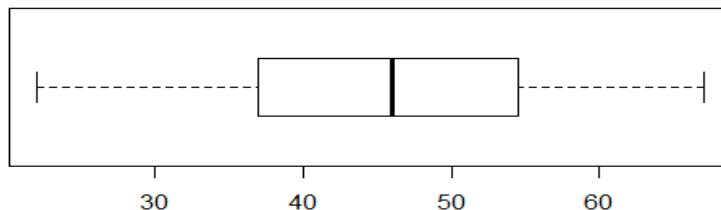
- 3사분위(Q3): 55 수식 이용: $L75=(15+1)(75/100)=12$

② EXCEL 함수 이용: $55-36=19$ (0, 1 제외) 또는 $54.5-37=17.5$ (0, 1 포함)

순서		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
점수		22	25	34	36	38	41	45	46	47	49	54	55	57	59	67	
0, 1 제외	0				.25				.5				.75				1
0, 1 포함		0			.25				.5			.75				1	

- 0, 1 제외: QUARTILE.EXC(array, 1) 또는 PERCENTILE.EXC(array, .25) -> Q1=36, Q3=55
- 0, 1 포함: QUARTILE.INC(array, 1) 또는 PERCENTILE.INC(array, .25) -> Q1=37, Q3=54.5
- 기존의 함수 QUARTILE과 PERCENTILE은 0, 1 포함(INC)으로 계산됨.
- 소프트웨어의 함수의 계산 방법에 따라 상이한 결과가 산출될 수 있으니, Q1과 Q3를 개념적으로 이해함에 초점을 두기 바랍니다. (데이터가 크면 차이는 무시할 수 있을 정도로 미미하게 될 것입니다.)
- 분산: $2222/14=158.71$
- 표준편차: $\sqrt{158.71}=12.60$
- 변동계수(CV): $12.60/45=0.28$

(3) 상자-수염그림



4. 3 표준편차 이내(± 3 SD)에 포함되는 데이터(관측치)의 규모(확률)를 경험법칙과 체비셰프의 부등식을 활용하여 각각 및 비교하여 설명하시오.

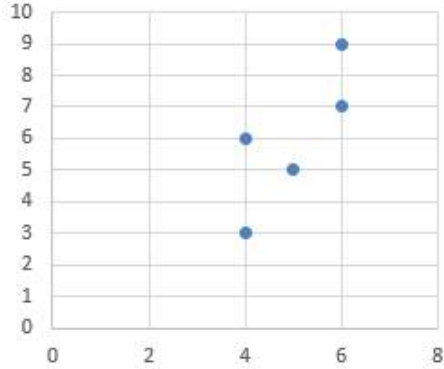
3표준편차 이내에 포함되는 데이터의 규모(확률):

- 경험법칙에 따르면 “모든 관측치의 약 99.7%가 평균으로부터 3표준편차 이내에 속한다”이고,
- 체비셰프 부등식을 이용하면 “K=3일 경우 모든 관측치의 적어도 $8/9(=0.88888...)$, 즉, 약 88.9%는 평균으로부터 3표준편차 이내에 속한다”라고 할 수 있음

종합하면 3표준편차 이내에는 어떤 형태의 분포이던지 약 88.9% 이상은 포함되며, 특히 종 모양 분포의 경우에는 약 99.7%가 포함됨

4장: 자료의 연관성 측도

- (1) 수학성적(X)과 과학성적(Y)의 자료유형: 양적변수(비율척도), 적절한 측도: 상관계수, 산점도
(2) 공분산: 1.75, 상관계수: 0.78
(3) 수학점수와 과학점수는 강한 양의 상관관계를 보인다.



- (1) 시점(COVID-19 전후) 변수와 식품유형 변수의 자료유형: 질적변수(명목척도), 적절한 측도: 분할표, 이차원 막대그래프
(2) 모든 셀 해석: COVID-19 이전에는 판매된 식품 182건 중 신선식품, 냉동식품, 반조리식품이 97, 27, 58건 판매되었고, COVID-19 이후에는 판매된 418건 중 신선식품, 냉동식품, 반조리식품이 203, 73, 142건 판매되었다.

	신선식품	냉동식품	반조리식품	계
COVID-19 이전	97	27	58	182
COVID-19 이후	203	73	142	418

- (3) 행 합계 기준 표준화: COVID-19 이전에는 신선식품, 냉동식품, 반조리식품이 (53%, 15%, 32%)의 비율로 판매되었고, COVID-19 이후에는 신선식품, 냉동식품, 반조리식품이 (49%, 17%, 34%)의 비율로 판매되었다.

	신선식품	냉동식품	반조리식품	계
COVID-19 이전	0.53	0.15	0.32	1.00
COVID-19 이후	0.49	0.17	0.34	1.00

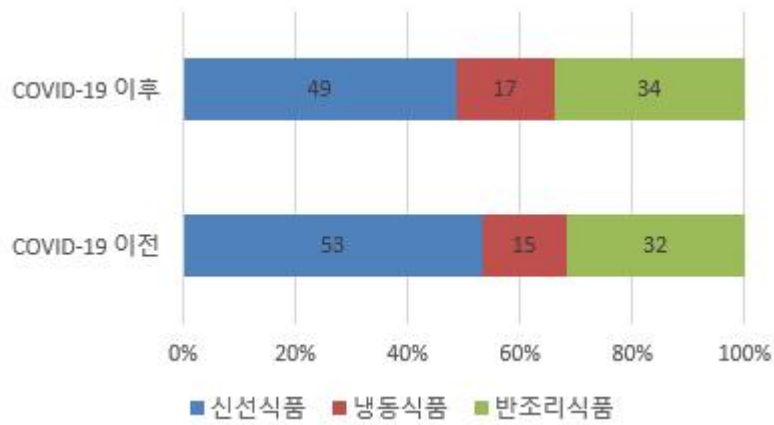
- (4) 열 합계 기준 표준화: 신선식품 판매량 중 COVID-19 이전 32%, 이후 58% 판매되었고, 냉동식품 판매량 중 COVID-19 이전 27%, 이후 73% 판매되었으며, 반조리식품 판매량 중 COVID-19 이전 29%, 이후 71% 판매되었다.

	신선식품	냉동식품	반조리식품
COVID-19 이전	0.32	0.27	0.29
COVID-19 이후	0.68	0.73	0.71
계	1.00	1.00	1.00

- (5) 전체 합계 기준 표준화: COVID-19 전후 판매된 600건의 식품은 COVID-19 이전 신선식품 16%, COVID-19 이전 냉동식품 5%, COVID-19 이전 반조리식품 10%, COVID-19 이후 신선식품 34%, COVID-19 이후 냉동식품 12%, COVID-19 이후 반조리식품 24%로 구성된다.

	신선식품	냉동식품	반조리식품	계
COVID-19 이전	0.16	0.05	0.10	0.30
COVID-19 이후	0.34	0.12	0.24	0.70
계	0.50	0.17	0.33	1.00

- (6) (2)~(5) 중 가장 적절한 방법: 행 합계 기준 표준화
(7) COVID-19 이후에 신선식품의 판매는 감소하고, 냉동식품과 반조리식품의 판매량이 증가하였다.

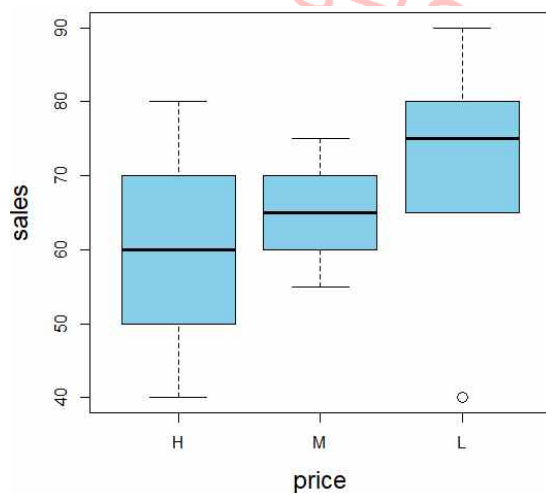


3. A백화점에서 각 지점의 가격수준을 높음, 중간, 낮음 3등급으로 분류하고, 가격수준별로 5개 지점씩 임의로 추출하여 매출액을 조사하였다(매출액 단위: 백만 원).

높음	중간	낮음
80	70	90
70	75	40
60	65	65
40	60	75
50	55	80

- (1) 가격수준(높음, 중간, 낮음): 질적변수(서열척도), 매출액: 양적변수(비율척도), 적절한 측도: 그룹별 기술통계량 또는 분할표, 상자수염그림

- (2) 가격수준별로 매출액의 분포 시각화:

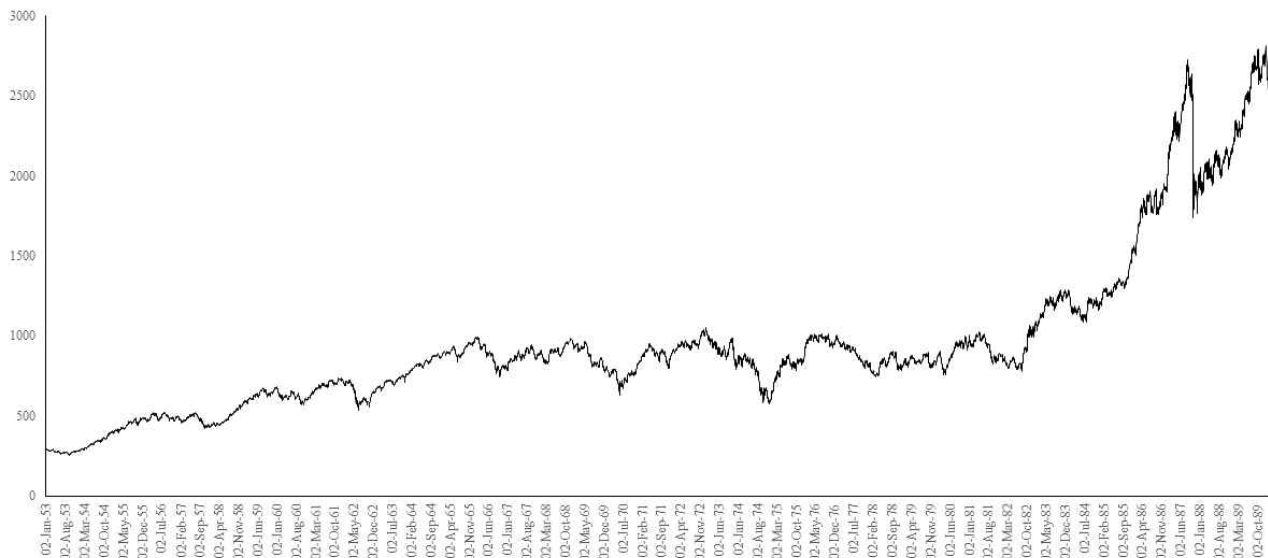


- (3) 가격수준별로 매출액의 평균: 높음 60, 중간 65, 낮음 70, 표준편차: 높음 15.81, 중간 7.91, 낮음 19.04 (가격수준이 낮을수록 매출액의 평균은 높아지고, 낮은 경우에 변동성도 가장 높다. 단, 중간 가격의 지점에서의 변동성이 가장 낮다.)

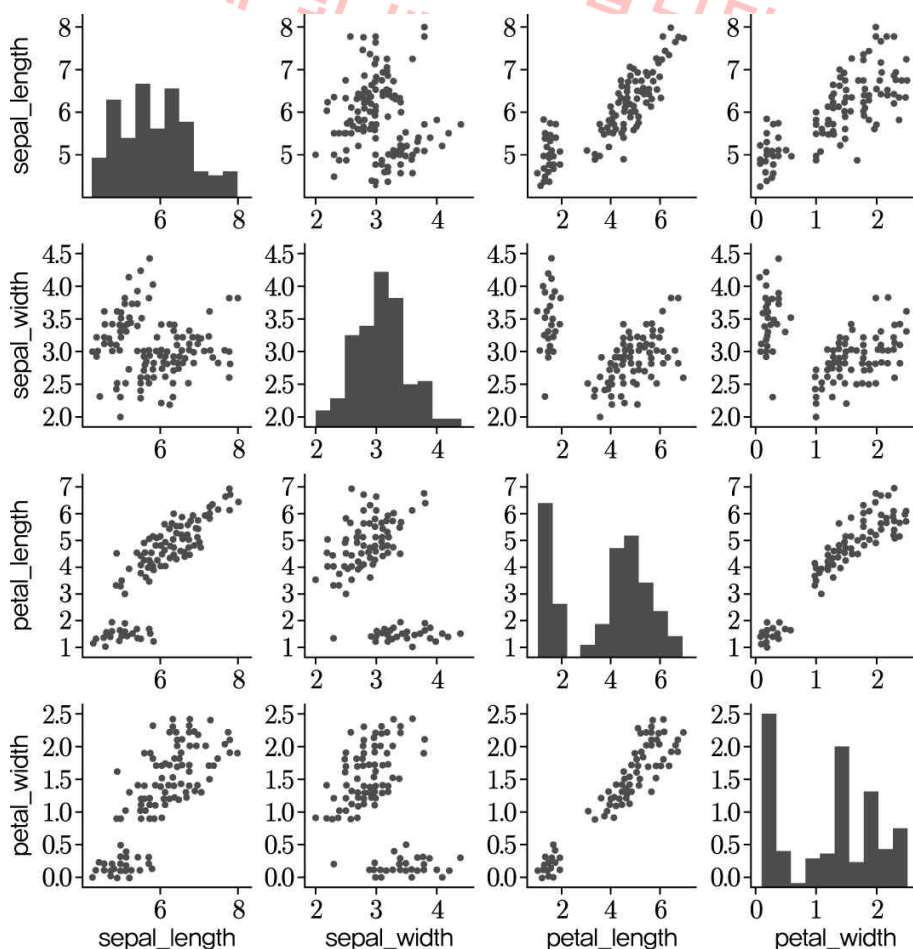
- (4) 가격수준이 중간 또는 높은 지점에서는 40%의 지점에서 매출액이 높은 반면, 가격수준이 낮은 지점에서는 60%의 지점에서 매출액이 높다.

		가격수준		
		높음	중간	낮음
매출액	높음	2	2	3
	낮음	3	3	2

4. 다우 종합지수 시계열 도표



5. [그림 4-4]는 150송이의 붓꽃(iris)에 관하여 꽃받침 길이(sepal_length), 꽃받침 넓이(sepal_width), 꽃잎 길이(sepal_length), 그리고 꽃잎 넓이(sepal_width)를 측정된 자료의 상관도표를 나타낸 것이다. 4개의 양적변수 각각을 짝지어 총 12개($=_4C_2$)의 산점도와 네 개의 히스토그램을 보여주고 있다.



대각선 부분에는 개별 변수의 분포를 파악하는 히스토그램을 넣는다. 대칭이므로 상삼각행렬(upper triangular matrix) 부분은 아래쪽과 중복되므로 경우에 따라 생략하거나 분포를 나타내는 타원형 등 다른 유형의 그래프 또는 상관계수 등을 표현하기도 한다.

5장: 확률이론

1. 야구시합에서 A, B팀이 있다. 전적 상으로 A팀이 이길 확률 0.6, B팀이 이길 확률 0.4이다. 야구전문가는 B팀이 이긴다고 전망했고, 야구전문가의 전망이 맞을 확률이 0.8, 틀릴 확률이 0.2라고 한다. B팀이 이길 확률은?

$P(\text{B팀이 이김} | \text{B팀이 이긴다는 전망})=?$

사전확률: $P(A)=\text{A팀이 이길 확률}=0.6$

$P(B)=\text{B팀이 이길 확률}=0.4$

우도확률: $P(B|B)=\text{(B가 이긴 상황) 전문가가 B팀이 이긴다고 했을 때 확률}=0.8$

$P(B|A)=\text{(B가 진 상황, 즉 A가 이긴 상황) 전문가가 B팀이 이긴다고 했을 확률}=0.2$

*주의: 우도확률을 고려함에 있어서 시간적인 전후관계가 아닌 두 사건의 관계에 대한 일반적인 정보를 이용한다는 점에서 접근하길 바랍니다.

Bayes' 이론을 적용하면,

사후확률 $P(B|B+)=\text{전문가가 B팀이 이긴다고 했을 때 B팀이 이길 확률} = (P(B) \cap P(B+)) / P(B+)$

$$= \{P(B) \cdot P(B|B)\} / \{P(B) \cdot P(B|B) + P(A) \cdot P(B|A)\}$$

$$= \frac{0.8 \cdot 0.4}{0.8 \cdot 0.4 + 0.6 \cdot 0.2} = \frac{0.32}{0.44} = \frac{8}{11} = 0.73$$

즉, 우도확률을 모를 때는 B팀이 이길 확률이 0.4이지만, 야구전문가의 예측 적중 확률정보를 활용하면 B팀의 확률은 0.73으로 올라갑니다.

2. MBA 프로그램을 지원하는 K지원자가 이 MBA프로그램에 입학하기 위해서는 650점 이상의 GMAT 점수를 취득해야 한다. 그런데 현재 650점을 취득할 확률은 10%이다. 그래서 \$500의 비용이 드는 학원수강을 고려중이며, 만일 650점 이상의 GMAT점수를 취득할 확률이 두 배 증가한다면 그는 기꺼이 학원을 수강할 것이다. 그는 어떻게 해야 하는가?

[MBA학생들을 대상으로 실시한 서베이 결과는 다음과 같다.]

- 650점 이상 GMAT점수를 취득한 학생: 52%가 수강
- 650점 미만 GMAT점수를 취득한 학생: 23%가 수강

사전확률: 650점을 취득할 확률 $P(S)=0.1$, 취득 못할 확률 $P(F)=0.9$

학원수강 사건: +

우도확률: $P(+|S)=0.52$, $P(+|F)=0.23$

$$\text{사후확률: } P(S|+)=\frac{0.52 \cdot 0.1}{0.52 \cdot 0.1 + 0.23 \cdot 0.9} = \frac{0.052}{0.052 + 0.207} = \frac{0.052}{0.259} = 0.201$$

사전확률보다 2배 이상 증가하였으므로 학원을 수강할 것이다.

3. 아스피린이 2병 있고, 페니실린이 1병 있다. 이들 약을 먹고 약물중독이 발생하였을 경우, 페니실린 약을 먹고 중독이 되었을 확률은? (아스피린을 먹고 약물중독일 확률은 10%, 페니실린은 80% 확률이라고 가정)

사전확률: 아스피린을 먹었을 경우 $P(A1)=2/3$, 페니실린을 먹었을 경우 $P(A2)=1/3$

우도확률: 아스피린을 먹고 중독될 확률 $P(B | A1)=0.1$, 페니실린을 먹고 중독될 확률 $P(B | A2)=0.8$

사후확률: 약을 먹고 중독되었을 때 그 약이 페니실린일 확률= $P(A_2 | B) = \frac{\frac{1}{3} * 0.8}{\frac{1}{3} * 0.8 + \frac{2}{3} * 0.1} = \frac{\frac{8}{30}}{\frac{10}{30}} = 0.8$

4. 사건의 정의: 기계(A, B, C) & 불량품(Defect, D), $P(A|D)=?$

사전확률: $P(A)=0.35$, $P(B)=0.2$, $P(C)=0.45$

우도확률: $P(D|A)=0.01$, $P(D|B)=0.013$, $P(D|C)=0.02$

사후확률: $P(A|D)=35/(35+26+90)=35/151=0.23$

5. 사건의 정의: 암 건강검진 결과: 양성(+), 실제 암에 걸림: C (정상: N) -> 알아야할 확률: $P(C|+)$

1) 사전확률: $P(C)=0.01$, $P(N)=0.99$

2) 우도확률: $P(+|C)=0.90$, $P(+|N)=0.10$

3) 사후확률: $P(C|+)=0.009/(0.009+0.099)=0.009/0.108=0.0833$

6. 사건의 정의: 스포츠카 위치(A1, A2, A3) & 3번 문을 여는 사건(B3), $P(A1|B3)$ 와 $P(A2|B3)=?$

사전확률: $P(A1)=P(A2)=P(A3)=1/3$

우도확률: $P(B3|A1)=1/2$, $P(B3|A2)=1$, $P(B3|A3)=0$

사후확률: $P(A1|B3)=(1/3 * 1/2)/(1/3 * 1/2 + 1/3 * 1 + 1/3 * 0)=1/3$

$P(A2|B3)=(1/3 * 1)/(1/3 * 1/2 + 1/3 * 1 + 1/3 * 0)=2/3$

“선택을 바꾼다”

6장: 이론적 확률분포

1. K전자에서 생산된 휴대폰은 10%가 불량(defective)이라고 한다. 무작위로 선택된 10대의 휴대폰 중 1대 혹은 2대가 불량일 확률을 (1) 이항분포와 (2) 포아송분포를 가정하여 각각 계산하고($e=2.718$), 이 상황에서 어느 분포가 더 적절한지 설명하시오.

(1) 이항분포:

$$P(1 \leq X \leq 2) = {}_{10}C_1(0.1)^1(0.9)^9 + {}_{10}C_2(0.1)^2(0.9)^8 = 0.3874 + 0.1937 = 0.5811$$

(2) 포아송분포:

$$\mu = np = (10)(0.1) = 1$$

$$P(X=x) = \frac{\mu^x e^{-\mu}}{x!}, \quad p(1 \leq X \leq 2) = \frac{1^1 e^{-1}}{1!} + \frac{1^2 e^{-1}}{2!} = 0.3679 + 0.1839 = 0.5518$$

(3) 적절한 분포: 이항분포가 더 적절함 ($\because$ 10대를 검사하므로 불량품의 확률변수 값에 상한(upper limit) 10 존재)

2. K투자종목의 수익률이 평균이 10%이고 표준편차가 5%인 정규분포를 따른다. 손실이 발생할 확률을 구하시오. 또한, 표준편차가 10%일 때 손실이 발생할 확률을 구하시오.

확률변수 X : K투자종목의 수익률

(1) 표준편차가 5%인 경우 손실이 발생할 확률:

$$X \sim N(10, 5^2)$$

$$P(X \leq 0) = P(Z \leq \frac{0-10}{5}) = P(Z \leq -2) = 0.0228$$

(2) 표준편차가 10%일 때 손실이 발생할 확률:

$$X \sim N(10, 10^2)$$

$$P(X \leq 0) = P(Z \leq \frac{0-10}{10}) = P(Z \leq -1) = 0.1587$$

3. 통계학 시험에서 65~75점 사이를 받은 학생은 B를 받는다고 한다. 만약, 전 수강생의 평균이 70점이고 표준편차가 10점이며, B학점을 받은 학생의 수가 60명이라면 통계학을 수강하고 있는 학생은 모두 몇 명인가? (시험성적은 정규분포를 하고 있다고 가정)

통계학 점수 $X \sim N(70, 10^2)$

$$P(65 \leq X \leq 75) = P(\frac{65-70}{10} \leq Z \leq \frac{75-70}{10}) = P(-0.5 \leq Z \leq 0.5) = 2P(0 \leq Z \leq 0.5) = 2*0.1915 = 0.383$$

$$\text{총 수강학생수} = 60/0.383 = 156.7 \rightarrow \text{약 } 157 \text{명}$$

4. 300명 여성들의 키에 대한 자료는 평균 160cm, 표준편차 5cm의 정규분포를 한다고 한다.

확률변수 X : 여성의 키

(1) 162cm보다 크고 172cm보다 작은 여성은 몇 명인가?

$$X \sim N(160, 5^2)$$

$$P(162 < X < 172) = P\left(\frac{162-160}{5} < \frac{X-\mu}{\sigma} < \frac{172-160}{5}\right) = P(0.4 < Z < 2.4) = 0.4918 - 0.1554 = 0.3364$$

300명 중 33.64%이므로, $100.92 \rightarrow 100$ 명

(2) 몇 cm 이하의 여성들이 30%에 해당하는가?

$$P(Z < -0.52) = 0.30 \rightarrow \frac{X-\mu}{\sigma} = \frac{X-160}{5} = -0.52$$

$$X = 157.4 \text{ (cm)}$$

5. 100명이 탑승할 수 있는 비행기만을 운행하는 이 노선의 예약부도율은 10%이다. 오늘 노선은 100% 예약이 완료되었다. 오늘 출발하기로 예약한 고객 100명 중 미탑승자가 5명 이하일 확률을 다음 단계로 구하시오.

확률변수 X: 미탑승자 수

(1) 이항분포의 확률 계산:

$$X \sim B(100, 0.1)$$

$${}_{100}C_0 (0.1)^0 (0.9)^{100} + {}_{100}C_1 (0.1)^1 (0.9)^{99} + {}_{100}C_2 (0.1)^2 (0.9)^{98} + {}_{100}C_3 (0.1)^3 (0.9)^{97} + {}_{100}C_4 (0.1)^4 (0.9)^{96} + {}_{100}C_5 (0.1)^5 (0.9)^{95} \\ = 0.05757...$$

$$= \text{BINOM.DIST}(5, 100, 0.1, \text{TRUE}) = 0.0576$$

(2) 이항분포의 정규분포 근사값 계산:

$$n=100, p=0.1$$

$np=10 > 5$, $n(1-p)=90 > 5$ 만족하여 정규분포 근사 가능

$$\mu=np=10, \sigma^2=np(1-p)=9, \sigma=3 \rightarrow X \sim N(10, 3^2)$$

$$P(X \leq 5) = P\left(Z \leq \frac{(5+0.5)-10}{3}\right) = P(Z < -1.5) = 0.0668$$

(3) 위 두 결과를 비교하시오.

$0.0668 - 0.0576 = 0.0092$ 의 미미한 차이로 이항분포의 정규분포 근사 가능

7장: 표집분포와 추정

1. 모수(parameter)와 통계량(statistics)을 표현하는 부호와 발음을 적으시오.

	모수		통계량	
	부호	발음	부호	발음
크기	N	capital N	n	
평균	μ ()	mu	$\bar{x}$	x bar
비율	π	pi	p	
분산	σ^2	sigma square	s^2	
표준편차	σ	sigma	s	

2. 항공편에서 빈자리들은 항공사가 수입원을 놓치는 원인이 된다. 어느 항공사가 지난 한 해 동안 항공편 당 공석의 평균치를 추정하고자 한다. 무작위로 225기 항공편들의 기록을 추출하여 각 표본 항공편의 공석 수를 기록하였더니, 표본평균과 편차는 각각 $\bar{x}=11.6$ 과 $s=4.1$ 이었다(단위는 자리). 지난해 항공편 당 평균 공석 수의 90% 신뢰구간을 추정하시오.

$$\begin{aligned} \text{모평균에 대한 90\% 신뢰구간: } \bar{X} \pm Z_{\frac{\alpha}{2}} \sigma_{\bar{X}} &= \bar{X} \pm Z_{0.05} \left(\frac{\sigma}{\sqrt{n}} \right) \approx \bar{X} \pm 1.645 \left(\frac{s}{\sqrt{n}} \right) \\ &= 11.6 \pm 1.645 \left(\frac{4.1}{\sqrt{225}} \right) = 11.6 \pm 0.45 = (11.15, 12.05) \end{aligned}$$

(11.15, 12.05) 구간이 지난 해 항공편 당 평균 공석 수를 포함한다고 90% 확신할 수 있다.

3. 환경오염 감시단체의 비판에 직면한 어느 자동차 제조사는 새로운 엔진이 대기오염 기준을 충족시키는지 알아보기 위해 엔진시험을 하고자 한다. 이러한 형태의 모든 엔진은 평균 배출탄소량이 반드시 20 PPM 미만이어야 한다. 10대의 엔진들이 시험용으로 생산되었으며 이 표본 엔진들로부터 배출탄소량의 평균 및 표준편차는 각각 $\bar{x}=17.1$ 과 $s=3$ 이었다(단위는 PPM). 이 새로운 엔진의 평균 배출탄소량의 99% 신뢰구간을 추정하시오.

$$t_{1-\frac{\alpha}{2}, n-1} = t_{0.995, 9} = 3.25$$

$$99\% \text{ 신뢰구간: } \bar{X} \pm t_{0.005} \left(\frac{s}{\sqrt{n}} \right) = 17.1 \pm 3.25 \left(\frac{3.0}{\sqrt{10}} \right) = (14.02, 20.18)$$

4. 어떤 지역에서 고구마에 대한 평균가격을 추정하려고 한다. 지역 슈퍼마켓 20곳을 무작위로 표본추출하고 각 슈퍼마켓에서 고구마를 구매하여 구매가격의 평균과 표준편차를 산출하였다. 표본의 평균이 26,840원, 표준편차가 2,630원으로 나타났다. 이 지역 고구마의 평균가격에 대한 95% 신뢰구간을 구하시오.

(단위: 천원) $\bar{X}=26.84$, 표준편차가 $S= 2.63$

$$t_{n-1, \alpha/2} = t_{19, 0.025} = 2.093$$

$$\bar{X} \pm t_{19, 0.025} \left(\frac{s}{\sqrt{n}} \right) = 26.84 \pm 2.093 \left(\frac{2.63}{\sqrt{20}} \right) = 26.84 \pm 1.23 = (25.61, 28.07)$$

5. 새로 개발된 인적자원개발(HRD) 프로그램에 대한 선호도 조사를 위해 36명의 학생을 표본추출하여 조사한 결과, 20명이 찬성하였다. 선호도 모비율 π 에 대한 95% 신뢰구간을 구하시오.

$$p = \frac{x}{n} = \frac{20}{36} = 0.56$$

$$95\% \text{ 신뢰구간: } p \pm z \sqrt{\frac{p(1-p)}{n}} = 0.56 \pm 1.96 \sqrt{\frac{0.56(1-0.56)}{36}}$$

$$0.40 < \pi < 0.72$$

(0.40, 0.72) 구간이 새로 개발된 인적자원개발(HRD) 프로그램에 대한 선호도 모비율을 포함한다고 95% 확신할 수 있다.

6. 유력한 대선 후보 K에 대한 대학생들의 지지율을 파악하기 위해 무작위로 100명의 대학생을 조사한 결과 36명이 지지함을 확인하였다. 대학생들의 K 후보 지지율에 대한 95% 신뢰구간을 구하시오.

$$p = \frac{x}{n} = \frac{36}{100} = 0.36$$

$$95\% \text{ 신뢰구간: } p \pm z \sqrt{\frac{p(1-p)}{n}} = 0.36 \pm 1.96 \sqrt{\frac{0.36(1-0.36)}{100}}$$

$$0.27 < \pi < 0.45$$

(0.27, 0.45) 구간이 대학생들의 K 후보 지지율을 포함한다고 95% 확신할 수 있다.

7. 화장품 회사에서 100cc 용기에 액체를 채우는 기계의 일관성을 확인하고자 25개 샘플을 무작위로 표본추출하여 액체의 양을 측정한 결과, 분산으로 1.2가 산출되었다. 모분산 σ^2 에 대한 95% 신뢰구간을 구하시오.

우선 자유도 $n-1=24$ 인 χ^2 분포의 임계치 $\chi_{0.025, 24}^2 = 12.40$ 와 $\chi_{0.975, 24}^2 = 39.36$ 을 구한다.

$$95\% \text{ 신뢰구간: } \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \text{에 대입하면, } \frac{24(1.2)}{39.36} \leq \sigma^2 \leq \frac{24(1.2)}{12.40}$$

$$0.73 < \sigma^2 < 2.32$$

(0.73, 2.32) 구간이 액체 양의 모분산을 포함한다고 95% 확신할 수 있다. 분산을 포함한다고 95% 확신할 수 있다.

8. 어떤 여론조사기관에서 대통령이 직무수행을 얼마나 잘 수행하는가를 조사하고자 한다. 만일 조사된 추정량에 관하여 95% 신뢰수준에서 최대 오차범위가 3%라고 한다면 표본 크기를 몇으로 하여야 할지 계산하시오.

이 여론조사기관에서 수용가능한 최대오차범위(maximum margin of error)가 3%이므로 다음을 알 수 있다.

$$Z_{.975} \cdot \sqrt{\frac{p(1-p)}{n}} = 1.96 \cdot \sqrt{\frac{.5(1-.5)}{n}} \approx 0.03$$

$p=0.5$ 인 까닭은 (다른 어떤 값을 가질 때보다) 0.5일 때 $p(1-p)$ 가 최대가 되기 때문이고, 문제에서 요구한 95% 신뢰수준에 따라 $Z_{.975}=1.96$ 을 표준오차 s_p 에 곱한다. 위 식을 계산하면 n 은 적어도 1060명 정도가 되어야 최대오차범위가 3%를 맞출 수 있다. 수많은 여론조사에서 표본크기가 1000명 안팎인 이유가 바로 여기에 있다. 지지율의 오차범위를 더 좁히고 싶다면 물론 표본크기를 늘려야 한다.

8장: 단일집단 가설검정

1. (1) 1.645, (2) 0.8413, (3) -1.8 (-1.812), (4) 0.01, (5) 8.547, (6) 0.025

2. 진통제의 효과 발현 시간

① 가설: $H_0: \mu = 30$, $H_a: \mu < 30$

② 유의수준: $\alpha = 0.05$

③ 검정통계량: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{28.5 - 30}{5/\sqrt{50}} = -2.12$

④ 임계치: $Z_{0.05} = 1.645$, 기각역: $z \leq -z_{0.05} = -1.645$

⑤ 유의성 판단: 검정통계량의 값 $-2.12 < -1.645$ 이므로 유의수준 $\alpha = 0.05$ 에서 H_0 기각할 수 있음.

또는 p-value=0.017 < 0.05이므로 H_0 기각할 수 있음.

[결론] 관측결과로부터 유의수준 5%에서 새로운 진통제의 진통효과가 더 빠르다고 볼 수 있다.

3. 기기의 강도

① 가설: $H_0: \mu = 83.2$, $H_a: \mu \neq 83.2$

② 유의수준: $\alpha = 0.05$

③ 검정통계량: $t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{85.43 - 83.2}{\sqrt{6.445/9}} = 2.635$

④ 임계치: $t_{(8, 0.975)} = 2.306$, 기각역: $t_0 < t_{(8, 0.025)} = -2.306$ or $t_0 > t_{(8, 0.975)} = 2.306$

⑤ 유의성 판단: 검정통계량의 값 $2.635 > 2.306$ 이므로 유의수준 $\alpha = 0.05$ 에서 H_0 기각할 수 있음.

또는 p-value=0.015*2=0.03 < 0.05이므로 H_0 기각할 수 있음.

[결론] 기존 의료기기와 새로운 공법에 의한 의료기기는 강도에 차이가 있다.

4. 초등학생들이 스스로 의복을 구입하는 비율

① 가설: $H_0: \pi \leq 0.2$, $H_a: \pi > 0.2$

② 유의수준: $\alpha = 0.01$

③ 검정통계량: $Z_0 = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.3 - 0.2}{\sqrt{0.2 * 0.8 / 100}} = 2.50$

④ 임계치: $Z_{0.99} = 2.326$, 기각역: $Z_0 > 2.326$

⑤ 유의성 판단: 검정통계량 $2.50 > 2.326$, 즉, 검정통계량이 기각역에 포함되므로 H_0 기각할 수 있음.

또는 p-value=0.006*2=0.012 < 0.05이므로 H_0 기각할 수 있음.

[결론] 초등학생들이 스스로 의복을 구입하는 비율이 과거보다 높다.

5. 커피 자판기의 용량

① 가설: $H_0: \mu = 150$, $H_a: \mu \neq 150$

② 유의수준: $\alpha = 0.05$

③ 검정통계량: $t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{152 - 150}{6.46/\sqrt{8}} = 0.88$

④ 임계치: $t_{(7, 0.975)} = 2.36$, 기각역: $t_0 < t_{(7, 0.025)} = -2.36$ or $t_0 > t_{(7, 0.975)} = 2.36$

⑤ 유의성 판단: 검정통계량 0.8 이 기각역에 속하지 않으므로 유의수준 $\alpha = 0.05$ 에서 H_0 기각할 수 없음.

또는 $p\text{-value} = 0.205 * 2 = 0.41 > 0.05$ 이므로 H_0 기각할 수 없음.

[결론] 커피 자판기의 용량이 150ml와 다르다고 할 만한 충분한 근거가 없다.

해당 강의 자료는 교강사만
이용할 수 있습니다.
무단 복제 및 배포는 저작권법 위
반 행위오니 관리에 유의
하시기 바랍니다.

9장: 두집단 가설검정

1. F분포 확률계산

- (1) $P(F(2, 3) > F_0) = 0.05, F_0 = 9.55$
- (2) $P(F(3, 2) > F_0) = 0.05, F_0 = 19.16$
- (3) $P(F(2, 3) < F_0) = 0.05, F_0 = 0.052; F_0 = 1/F(0.95, 3, 2) = 1/19.16 = 0.052$
- (4) $P(F(9, 7) < F_0) = 0.1, F_0 = 0.399; F_0 = 1/F(0.9, 7, 9) = 1/2.505 = 0.399$

2. (예제 9-1)의 온라인 콘텐츠 교육성과 문제의 등분산/이분산 판단

- ① 가설: $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_a: \sigma_1^2 \neq \sigma_2^2$
 - ② 유의수준: $\alpha = 0.05$
 - ③ 검정통계량: $F_0 = \frac{S_1^2}{S_2^2} = \frac{25}{16} = 1.563$
 - ④ 임계치: $F(29, 29, 0.975) = 0.476, F(29, 29, 0.025) = 2.101$
기각역: $F < 0.476$ 또는 $F > 2.101$
(p-value = $2 * P(F_{29, 29} > 1.56) = 2 * 0.118 = 0.235$)
 - ⑤ 유의성 판단: $0.476 < F_0 = 1.563 < 2.101$ (p-value = $0.235 > 0.05$) $\rightarrow$ 5% 유의수준에서 귀무가설 기각할 수 없음.
- [결론] 전통적 콘텐츠와 게임형 콘텐츠의 교육성과 분산이 다르다고 할 충분한 근거가 없다. 따라서 등분산 검정을 실시해야 한다.

3. A, B 두 브랜드 배터리의 평균수명 비교

- (1) 두 배터리의 평균수명이 같은지 검정
 - ① 가설: $H_0: \mu_1 = \mu_2, H_a: \mu_1 \neq \mu_2$
 - ② 유의수준: $\alpha = 0.05$
 - ③ 검정통계량: $Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{810 - 850}{\sqrt{\frac{10000}{40} + \frac{2500}{50}}} = \frac{-40}{\sqrt{300}} = -2.31$
 - ④ 임계치: $Z_{0.975} = 1.96$, 기각역: $|Z| > 1.96$
 - ⑤ 유의성 판단: $T = -2.31 < -1.96$ 이므로 귀무가설 기각
- [결론] 유의수준 0.05에서 두 회사 배터리의 평균수명이 다르다.

p-value에 의한 방법

- ④ p-value = $P(|Z| \geq 2.31) = 2 \times 0.0104 = 0.0208$
 - ⑤ p-value = $0.0208 < \alpha = 0.05$ 이므로 귀무가설 기각
- [결론] 유의수준 0.05에서 두 회사 배터리의 평균수명이 다르다.

(2) 두 배터리의 평균수명 차이가 50시간인지 검정

- ① 가설 $H_0: \mu_1 - \mu_2 = 50, H_a: \mu_1 - \mu_2 \neq 50$
- ② 유의수준 $\alpha = 0.05$

$$\textcircled{3} \text{ 검정통계량 } Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(810 - 850) - (0 - 50)}{\sqrt{\frac{10000}{40} + \frac{2500}{50}}} = \frac{10}{\sqrt{300}} = 0.577$$

④ 임계치: $Z_{0.975} = 1.96$, 기각역: $|Z| > 1.96$

⑤ 유의성 판단: $-1.96 < T = 0.577 < 1.96$ 이므로 귀무가설을 기각할 수 없음

[결론] 유의수준 0.05에서 두 회사 배터리의 수명 차이가 50시간이 아니라고 할 만한 충분한 근거가 없다. (50시간임)

4. 수능점수의 지역별 편차

(1) 분산의 동질성 비교(양측검정)

① 가설: $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_a: \sigma_1^2 \neq \sigma_2^2$

② 유의수준: $\alpha = 0.05$

$$\textcircled{3} \text{ 검정통계량: } F_0 = \frac{S_1^2}{S_2^2} = \frac{8.2^2}{7.6^2} = 1.164$$

④ 임계치: $F(89, 99, 0.975) = 0.664$, $F(89, 99, 0.025) = 1.499$

기각역: $T < 0.664$ 또는 $T > 1.499$

$$(p\text{-value} = 2 * P(F_{89,99} > 1.164) = 2 * 0.230 = 0.461)$$

⑤ 유의성 판단: $0.664 < F_0 = 1.164 < 1.499$ ($p\text{-value} = 0.461 > 0.05$) $\rightarrow$ 5% 유의수준에서 귀무가설 기각 못함.

[결론] A지역과 B지역 학생들의 성적 분산이 다르다고 할 충분한 근거가 없다.

(2) 모평균 비교

① 가설 $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$

② 유의수준 $\alpha = 0.05$

$$\textcircled{3} S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{89 \times 67.24 + 99 \times 57.76}{90 + 100 - 2} = 62.25$$

$$\text{검정통계량 } T_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{76.4 - 81.2}{\sqrt{62.25(\frac{1}{90} + \frac{1}{100})}} = \frac{-4.8}{1.314} = -3.65$$

④ 임계치: $t(188, 0.025) = 1.973$, 기각역: $T < -1.973$ 또는 $T > 1.973$

$$\text{또는 } (p\text{-value} = 2 * P(|t_{188}| > 3.65) = 2 * 0.00017 = 0.00034)$$

⑤ 유의성 판단: $T_0 = -3.65 < -1.973$ ($p\text{-value} < 0.05$) $\rightarrow$ 5% 유의수준에서 귀무가설 기각.

[결론] A지역 학생과 B지역 학생의 수능 평균점수에 차이가 있다.

5. 갑작스러운 전염병의 발생으로 인해 개발한 신약의 효과

(1) 분산의 동질성 비교(오른쪽꼬리 검정: 다양한 예시를 위해 단측검정 수행)

① 가설: $H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq 1$ 이고, $H_a: \frac{\sigma_1^2}{\sigma_2^2} > 1$

② 유의수준: $\alpha = 0.05$

③ 검정통계량: $F_0 = \frac{s_1^2}{s_2^2} = 1.98$

④ p-value = F.DIST.RT(1.98, 29, 29) = 0.035

⑤ 유의성 판단: 검정통계량 $F_0 < 0.05$ 이므로 귀무가설 기각.

[결론] 분산이 다르다. → 이분산 채택

(2) 모평균 비교

① 가설 $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

② 유의수준 $\alpha = 0.01$

③ 검정통계량: $t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{11.4 - 12.3}{\sqrt{\frac{18.93}{30} + \frac{9.57}{30}}} = -0.92$

보정자유도 $\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}} = \frac{0.9025}{\frac{0.3982}{29} + \frac{0.1018}{29}} = \frac{0.9025}{0.0172} = 52.35$

④ 임계치: $t(52.35, 0.975) = 2.007$, 기각역: $|T| > 2.007$ <Excel 참조> =T.INV(0.975, 52.35)

⑤ 유의성 판단: $-2.007 < t_0 = -0.92 < 2.007 \rightarrow H_0$ 기각할 수 없음.

[결론] 신약과 위약의 치료효과가 나타나는 시간이 다르다고 할 만한 충분한 근거가 없다.

6. 공장 12개소에서 사고방지 프로그램을 운영하면서 프로그램의 효과

[12개의 공장에서 프로그램 전후에 측정한 것이므로 쌍체비교]

$\bar{d} = 3$, $s_d = \sqrt{\frac{25 + 4 + 64 + 9 + 49 + 16 + 36 + 25 + 25 + 25 + 0 + 64}{11}} = 5.576$

① 가설: $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 > 0$

② 유의수준: 0.05

③ 검정통계량: $t_0 = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{3}{\frac{5.576}{\sqrt{12}}} = 1.86$

④ 기각역: $t > t_{11, 0.95} = 1.796$ (p-value = 0.045)

=T.INV(0.95, 11) 및 =1-T.DIST(1.86, 11, TRUE) 참조

⑤ 유의성 판단: $t_0 = 1.86 > 1.796$ (또는 p-value < 0.05)이므로 H_0 기각할 수 있음.

[결론] 사고로 인한 손실작업시간이 프로그램의 운영 후 감소하였다고 할 수 있다.

7. 독감을 예방하는 신약이의 성별 약효 차이

$$p_1 = \frac{102}{200} = 0.51, \quad p_2 = \frac{38}{100} = 0.38, \quad \bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{102 + 38}{200 + 100} = 0.47$$

① 가설: $H_0: p_1 - p_2 = 0$ vs $H_1: p_1 - p_2 \neq 0$

② 유의수준: 0.05

$$\text{③ 검정통계량: } T = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.13}{\sqrt{0.47 \times 0.53 \times 0.015}} = 2.13$$

④ 기각역: $|Z| > Z_{0.025} = 1.96$

⑤ 유의성 판단: $Z = 2.13 > 1.96$ 이므로 H_0 기각할 수 있음.

[결론] 성별로 효과가 동일하지 않다고 판단된다.

8. P기업과 Q기업이 주식수익률의 변동성(분산)에 차이가 있는지 검정

① 가설: $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

② 유의수준: 0.01

$$\text{③ 검정통계량: } F = \frac{s_1^2}{s_2^2} = \frac{9.4}{6.8} = 1.38$$

④ 기각역: $F < F_{0.005, 40, 30} = 0.42$ 또는 $F > F_{0.995, 40, 30} = 2.52$

임계치: $=F.INV(0.005, 40, 30)$, $=F.INV(0.995, 40, 30)$ 참조

⑤ 유의성 판단: $0.42 < F = 1.38 < 2.52$ 이므로 H_0 기각할 수 없음.

[결론] 두 회사 주식수익률의 변동성에 유의한 차이가 없다.

[참고] 유의수준 10%에서는 $F < F_{0.05, 40, 30} = 0.57$ 또는 $F > F_{0.95, 40, 30} = 1.79$

10장: 분산분석

1. A 백화점 각 지점의 가격수준별 매출액 비교

- (1) 총제곱합(SS_{Total}): 2950
- (2) 가격 수준별 그룹 간 제곱합(SS_{Between}): 250
- (3) 가격 수준별 그룹 간 제곱평균(MS_{Between}): 125
- (4) 검정통계량과 비교할 임계치 F의 분모 자유도: 12
- (5) 가격 수준별로 매출액에 차이가 있다고 판단할 수 있는가? 없다.

2. K 공장의 A, B, C 기계별 산출량

(1) 분산분석표

변동요인	자유도	제곱합	제곱평균	F
그룹간	2	72	36	5.76
그룹내	8	50	6.25	
총변동	10	122		

(2) 기계별 생산량에 차이가 있는지 가설검정

- ① 가설: $H_0: \mu_A = \mu_B = \mu_C$ (세 그룹의 모평균은 모두 같다) vs. $H_a: \mu_i = \mu_j$ for any i, j (세 그룹의 모평균이 모두 같지는 않다. 적어도 한 그룹의 모평균은 같지 않다.)
- ② 유의수준: $\alpha = 0.05$
- ③ 검정통계량: $F_0 = 5.76 \sim F_{2,8}$
- ④ 임계치와 기각역: $F > F_{2,8,0.95} = 4.46$
- ⑤ 기각여부 판단: 귀무가설을 기각할 수 있다.
- (3) 결론: 세 기계별 생산량의 모평균이 모두 같지는 않다.

3. 브랜드별 휴대폰 사용자 만족도

(1) 분산분석표

변동요인	자유도	제곱합	제곱평균	F
그룹간	3	605.47	201.82	9.14
그룹내	13	287	22.08	
총변동	16	892.47		

(2) 기종별 만족도에 차이가 있는지 가설검정

- ① 가설설정: $H_0: \mu_A = \mu_B = \mu_C = \mu_D$ vs. $H_a: \mu_i = \mu_j$ for any i, j
- ② 유의수준: $\alpha = 0.05$
- ③ 검정통계량: $F_0 = 9.14 \sim F_{3,13}$
- ④ 임계치와 기각역: $F > F_{3,13,0.95} = 3.41$
- ⑤ 기각여부 판단: 귀무가설을 기각할 수 있다.
- (3) 결론: 브랜드별 사용자 만족도의 모평균이 모두 같지는 않다.

4. 평가 인증 여부에 따라 보육교직원 현원의 분산

가설은 다음과 같다.

$$H_0: \frac{\sigma_Y^2}{\sigma_N^2} \leq 1$$

$$H_a: \frac{\sigma_Y^2}{\sigma_N^2} > 1$$

데이터를 먼저 평가인증 여부에 따라 정렬(sort)한다. 그 다음 엑셀에서 [데이터]-[데이터 분석]을 통해 “기술 통계량”을 평가인증 여부 별로 확인해보자. F 값은 다음과 같다.

$$\frac{S_Y^2}{S_N^2} = \frac{29.607}{17.779} = 1.665$$

두 자유도는 각각 121과 83이므로, 엑셀에서 1-F.DIST(1.665, 121, 83, TRUE)로 유의확률(p -value)을 계산한다. 그 값은 0.00695 정도로 0.01보다 다소 작으므로 귀무가설은 1% 유의수준에서 기각될 수 있다. 다시 말해, 평가인증을 받은 어린이집은 (그렇지 않은 어린이집보다) 보육교직원현원의 분산이 통계적으로 유의하게 크다.

해당 강의 자료는 교강사만
이용할 수 있습니다.
무단 복제 및 배포는 저작권법 위
반 행위오니 관리에 유의
하시기 바랍니다.

11장: 카이제곱분석

1. 게임에 사용하고 있는 주사위가 공정한 주사위인지(즉, 이산균등분포를 하는지) 검정(n=120)

[기대값]

숫자	1	2	3	4	5	6
관측빈도	20	20	20	20	20	20

- ① 가설설정: $H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = \frac{1}{6}$
 H_a : 적어도 하나는 $\frac{1}{6}$ 이 아니다.
- ② 유의수준 설정: $\alpha = 0.05$
- ③ 검정통계량 계산: $\chi_0^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i} = 7.80 \sim \chi_5^2$
- ④ 유의성 검정: 검정통계량 7.80이 임계치 11.07보다 크지 않으므로 귀무가설 기각 못함.
- ⑤ 해석 및 결론: 공정한 주사위가 아니라고 할 충분한 근거가 없다.

2. 신용카드 회사의 고객의 신용도에 따른 등급 평가

[기대값]

신용등급	High	Middle	Low	계
관측고객 수	25	50	25	100

- ① 가설설정: $H_0: \pi_1 = .25, \pi_2 = .50, \pi_3 = .25$
 H_a : 적어도 한 등급은 π_j 와 다르다.
- ② 유의수준 설정: $\alpha = 0.01$
- ③ 검정통계량 계산: $\chi_0^2 = \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} = 4.32 \sim \chi_2^2$
- ④ 유의성 검정: 검정통계량 4.32가 임계치 9.21보다 크지 않으므로 귀무가설 기각 못함.
- ⑤ 해석 및 결론: 목표로 하는 등급 비율에 대한 기준이 적절하다.

3. 휴대폰 브랜드 S와 K의 시장점유율의 지역별 차이

[기대값]

	S 브랜드	K 브랜드	계
지역 A	48	48	96
지역 B	52.5	52.5	105
지역 C	45	45	89

- ① 가설설정: H_0 : 각 지역별 브랜드 비율이 같다. H_a : 각 지역별 브랜드 비율이 모두 같지는 않다.
- ② 유의수준 설정: $\alpha = 0.05$
- ③ 검정통계량 계산: $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 3.62 \sim \chi_{3(2-1)}^2$
- ④ 유의성 검정: 검정통계량 3.62가 임계치 7.81보다 크지 않으므로 귀무가설 기각 못함.
- ⑤ 해석 및 결론: 각 지역별 브랜드 비율이 다르다고 할 만한 충분한 근거가 없다.

4. COVID-19 이전과 이후의 식품유형별 판매 비율

[기대값]

	신선식품	냉동식품	반조리식품	계
COVID-19 이후	80.69	30.33	70.98	182
COVID-19 이전	185.31	69.67	163.02	418

- ① 가설설정: H_0 : COVID-19 이전과 이후의 식품유형별 판매 비율이 같다. H_a : 모두 같지는 않다.
 ② 유의수준 설정: $\alpha = 0.05$
 ③ 검정통계량 계산: $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 8.67 \sim \chi^2_{2(3-1)}$
 ④ 유의성 검정: 검정통계량 8.67이 임계치 9.49보다 크지 않으므로 귀무가설 기각 못함.
 ⑤ 해석 및 결론: COVID-19 이전과 이후의 식품유형별 판매 비율이 다르다고 할 만한 충분한 근거가 없다.

5. 자동차 색상의 선호도에 대한 성별 차이
[기대값]

색상	화이트	레드	차콜	블랙	계
남성	31.20	20.16	25.92	18.72	96
여성	33.80	21.84	28.08	20.28	104
계	65	42	54	39	200

- ① 가설설정: H_0 : 자동차 색상 선호도와 성별은 서로 독립이다(관련이 없다), H_a : 독립이 아니다(관련이 있다).
 ② 유의수준 설정: $\alpha = 0.05$
 ③ 검정통계량 계산: $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 19.85 \sim \chi^2_3$
 ④ 유의성 검정: 검정통계량 19.85가 임계치 7.81보다 크므로 귀무가설 기각.
 ⑤ 해석 및 결론: 성별 자동차 색상 선호도에 차이가 있다(독립이 아니다, 관련이 있다).

6. '숙의 기반의 주민참여 방식'을 도입에 대한 연령군별로 만족도 차이
[기대값]

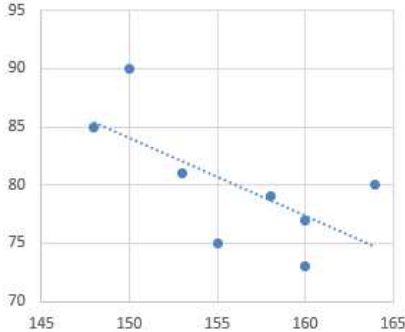
	만족	보통	불만족	계
청년층	45.22	58.52	48.26	152
장년층	33.02	42.735	35.24	111
노년층	40.76	52.745	43.50	137
계	119	154	127	400

- ① 가설설정: H_0 : 연령과 만족도는 서로 독립이다(관련이 없다), H_a : 독립이 아니다(관련이 있다).
 ② 유의수준 설정: $\alpha = 0.05$
 ③ 검정통계량 계산: $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 5.95 \sim \chi^2_4$
 ④ 유의성 검정: 검정통계량 5.95가 임계치 9.49보다 크지 않으므로 귀무가설 기각 못함.
 ⑤ 해석 및 결론: '숙의 기반의 주민참여 방식'을 도입에 대한 만족도는 연령군별로 차이가 없다(독립이다).

12장: 상관분석과 단순회귀분석

1. 청소년 대상 피트니스센터에서 8명의 회원의 신체검사: 키 $\bar{X}=156$, 가슴둘레 $\bar{Y}=80$

(1) 키와 가슴둘레의 관계를 파악하는 산점도



→ 키와 가슴둘레의 관계는 음(-)의 상관관계를 보임

2) 상관계수: $r = \frac{S_{XY}}{S_X S_Y} = \frac{-140}{\sqrt{210} \times \sqrt{210}} = -0.67$

3) 상관계수의 유의성 검정

① 가설: $H_0: \rho=0$ (키와 가슴둘레는 유의한 상관성이 없다), $H_a: \rho \neq 0$ (상관성이 있다)

② 유의수준: $\alpha=0.05$

③ 검정통계량: $t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.67}{\sqrt{\frac{1-(-0.67)^2}{8-2}}} = -2.19$

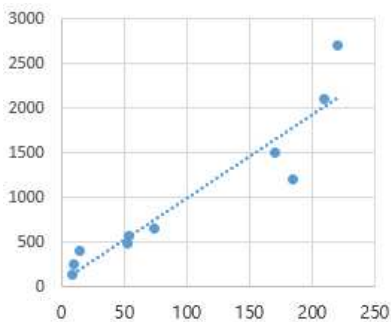
④ 임계치: $t_{(6,0.025)} = 2.45$, 기각역: $t < -2.45$ 또는 $t > 2.45$

⑤ $-2.45 < t_0 = -2.19 < 2.45 \rightarrow H_0$ 기각할 수 없음

[결론] 유의수준 0.05에서 키와 가슴둘레 사이에 유의한 상관성이 있다고 할 수 없다.

2. 기업의 광고비가 매출액에 미치는 영향

(1) 광고비를 독립변수(X), 매출액을 종속변수(Y)로 하는 산점도



→ 광고비와 매출액은 양(+)의 상관관계를 보임

(2) 최소제곱법에 의한 회귀식

$\bar{X}=110$, $\bar{Y}=900$, $S_{XY} = 286340$, $S_{XX} = 30758$

$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{286340}{30758} = 9.31$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 900 - 9.31 \times 110 = -124.04$

$\hat{Y}_i = -124.04 + 9.31X_i$

(3) 회귀계수의 유의성 검정

① 가설: $H_0: \beta_1=0$ (광고비는 매출액에 영향을 미치지 않는다), $H_a: \beta_1 \neq 0$ (영향을 미친다)

② 유의수준: $\alpha=0.05$

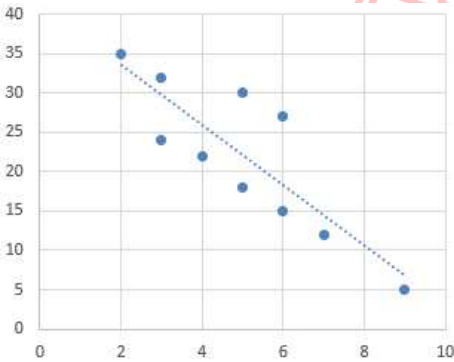
③ 검정통계량: $S_{b_1} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2 / (n-k-1)}{\sum(X_i - \bar{X})^2}} = 2.51, t_0 = \frac{b_1}{S_{b_1}} = \frac{9.31}{2.51} = 3.71$

④ 유의성 검정: p-value=0.006 < 0.05이므로 H_0 기각

⑤ 해석 및 결론: 기업의 광고비가 매출액에 유의한 영향을 미친다.

3. 중고차의 사용기간이 가격에 미치는 영향

(1) 산점도



→ 중고차의 사용기간과 가격은 음(-)의 상관관계를 보임

(2) 회귀모형 추정: $\hat{Y}_i = 41.13 - 3.83X_i$

(3) 회귀계수의 유의성 검정

① 가설: $H_0: \beta_1=0$ (중고차의 사용기간은 가격에 영향을 미치지 않는다), $H_a: \beta_1 \neq 0$ (영향을 미친다)

② 유의수준: $\alpha=0.05$

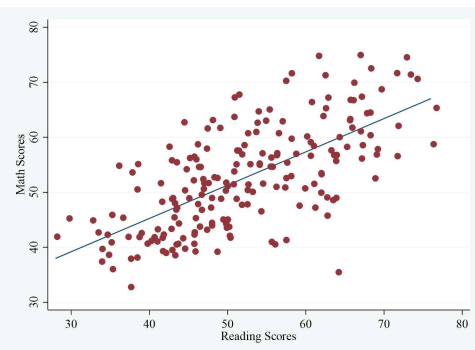
③ 검정통계량: $S_{b_1} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2 / (n-k-1)}{\sum(X_i - \bar{X})^2}} = 0.85, t_0 = \frac{b_1}{S_{b_1}} = \frac{-3.83}{0.85} = -4.50$

④ 유의성 검정: p-value=0.002 < 0.05이므로 H_0 기각

⑤ 해석 및 결론: 중고차의 사용기간은 가격에 영향을 미친다.

4. 읽기점수가 수학점수에 미치는 영향

(1) 산점도



→ 읽기점수와 수학점수는 양(+)의 상관관계를 보임

(2) 회귀모형 추정: $\hat{Y}_i = 21.038 + 0.605X_i$

(3) 회귀계수의 유의성 검정

① 가설: $H_0: \beta_1=0$ (읽기점수는 수학점수에 영향을 미치지 않는다), $H_a: \beta_1 \neq 0$ (영향을 미친다)

② 유의수준: $\alpha=0.001$

③ 검정통계량: $S_{b_1} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2 / (n - k - 1)}{\sum(X_i - \bar{X})^2}} = 0.0487$, $t_0 = \frac{b_1}{S_{b_1}} = \frac{0.605}{0.0487} = 12.44$

④ 유의성 검정: p-value < 0.001이므로 H_0 기각

⑤ 해석 및 결론: 읽기점수는 수학점수에 영향을 미친다.

해당 강의 자료는 교강사만
이용할 수 있습니다.
무단 복제 및 배포는 저작권법 위
반 행위오니 관리에 유의
하시기 바랍니다.