8장 통계적 추론 개요

8.1 표준편차 σ가 5인 정규모집단의 모평균 μ에 대한 구간추정을 위해 25개의 확률표본을 구하여 95% 구간을 구했더니 [22.54, 26.46]이었다고 할 때 아래의 물음에 대해 '맞음', '틀림', '모름'으로 답하여라.

⑴ 모평균 μ가 구간 [22.54, 26.46]에 있을 확률은 0.95이다. 틀림

⑵ 표본평균 $\overline{X}$가 구간 [22.54, 26.46]에 있을 확률은 0.95이다. 틀림

⑶ 표본평균 $\overline{x}$가 구간 [22.54, 26.46]에 있을 확률은 0.95이다. 틀림

⑷ 모평균 μ가 구간 [22.54, 26.46] 안에 있다. 모름

⑸ 표본평균 $\overline{X}$가 구간 [22.54, 26.46] 안에 있다. 모름

⑹ 표본평균 $\overline{x}$가 구간 [22.54, 26.46] 안에 있다. 맞음

⑺ 표본의 95%가 구간 [22.54, 26.46] 안에 있다. 모름

⑻ 신뢰수준을 99%로 하면 신뢰구간의 폭은 넓어진다. 맞음

⑼ 새로운 신뢰구간은 [22.54, 26.46] 안에 있다. 모름

⑽ 새로운 신뢰구간의 폭은 26.46 − 22.54 = 3.92 보다 좁다. 맞음

8.2 다음의 문제들에 대한 귀무가설과 대립가설을 세우고 적절한 형태의 검정통계량을 선택하고 비정상의 기준, 즉 기각역의 형태를 기술하여라.

(1) 어떤 제품을 생산하는 과정에서 품질의 변동, 즉 분산이 10보다 크면 문제가 발생하고 있다고 판정한다고 하자. 생산과정에서 문제가 발생하고 있는지를 알아보기 위해 무작위로 n개의 표본을 선택하여 조사한다.

σ^2: 품질의 분산, S^2: 표본분산

가설: $H_0 : \sigma^2 \le 10$ vs $H_1 : \sigma^2 > 10$

검정통계량: S^2 기각역의 형태: $S^2 \ge k$

(2) 어떤 지역의 셋째 아이의 남아출생비율이 둘째 아이의 남아 출생비율

보다 높다고 한다. 이를 알아보기 위해 이 지역에서 아이가 셋 이상인 가구 중 n 곳을 무작위로 방문하여 둘째와 셋째의 남아비율을 조사한다.

θ_i: i번째 남아출생비율, P_i : i번째 남아출생표본비율

가설: $H_0 : \theta_2 \geq \theta_3$ vs $H_1 : \theta_2 < \theta_3$

검정통계량: $P_2 - P_3$ 기각역의 형태: $P_2 - P_3 \leq k$

⑶ 조류는 좁은 공간에서 가두어 키우면 서로 잡아먹는 상호포식 현상이 종종 발생하는데 어떤 양계업자가 폐사된 닭 중 상호포식에 의해 폐사되는 비율이 4% 미만인지를 알아보고자 한다. 이를 위해 폐사된 닭 중 n 마리를 무작위로 추출하여 상호포식에 의한 비율을 조사한다.

θ: 상호포식비율, P: 상호포식표본비율

가설: $H_0 : \theta \geq 0.04$ vs $H_1 : \theta < 0.04$

검정통계량: P　　　　기각역의 형태: $P \leq k$

⑷ 흡연자의 소득이 비흡연자의 소득보다 높은지를 알아보고자 한다. 이를 위해 흡연자 중 m명을, 비흡연자 중 n명을 무작위로 선택하여 조사한다.

μ_1 : 흡연자의 평균소득, μ_2 :비흡연자의 평균소득

$\overline{X}$: 흡연자 표본평균소득, $\overline{Y}$: 비흡연자 표본평균소득

가설: $H_0 : \mu_1 \leq \mu_2$ vs $H_1 : \mu_1 > \mu_2$

검정통계량: $\overline{X} - \overline{Y}$　　기각역의 형태: $\overline{X} - \overline{Y} \geq k$

⑸ 여주 아울렛과 파주 아울렛의 가격에 차이가 있는지를 알아보고자 한다. 이를 위해 여러 개의 제품을 무작위로 선택하여 각각의 제품에

대해 여주와 파주 아울렛의 판매가격을 조사한다.

μ_1 : 여주의 평균가격, μ_2 :파주의 평균가격

$\overline{D}$: 각 상품의 여주와 파주 가격의 차의 표본평균

가설: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$

검정통계량: $\overline{D}$ 기각역의 형태: $|\overline{D}| \geq k$

8.3 분산 σ^2이 1인 정규모집단의 평균 μ에 대한 통계적 추론을 위해 16개의 확률표본을 얻었다고 하자. 다음과 같은 가설에 대해

$$H_0 : \mu = 0, \quad H_1 : \mu = 1$$

표본평균 $\overline{X}$가 검정통계량일 때 다음 물음에 답하여라. $\overline{X} \sim N(\mu, 1/16)$

(1) $\overline{X}$의 표준화된 검정통계량이 $4\overline{X}$인 것을 보여라

표준화된 검정통계량 ⇨ $\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{\overline{X}-0}{1/\sqrt{16}}=4\overline{X}$

⑵ 유의수준 5%에서 $\overline{X}$에 대한 기각역을 구하여라.

$H_1:\mu=1>\mu_0=0$ 이므로 기각역은 $4\overline{X}\geq z_{0.05}=1.645 \quad \Rightarrow \quad \overline{X}\geq 0.41125$

⑶(2)에서 구한 기각역을 적용했을 때 제2종 오류의 확률을 계산하여라.

$$\beta=P_{H_1}(\overline{X}<0.41125|H_1)=P_{H_1}\left(\frac{\overline{X}-1}{1/\sqrt{16}}<\frac{0.41125-1}{1/\sqrt{16}}\right)$$

$$=P(Z<-2.355)=0.0093$$

⑷ 표본평균이 $\overline{x}=0.7$이었다면 단측검정에 의한 유의확률을 구하여라.

유의확률: $P(\overline{X}\geq 0.7)=P(4\overline{X}\geq 2.8)=P(Z\geq 2.8)=0.0026$

8.4 기존의 방법에 의한 독감백신 평균배양시간은 30시간이었다고 하자. 이 배양시간을 단축하고자 새로운 방법을 개발하였다. 새로 개발된 세포배양법에 의한 독감백신 배양시간을 알아보고자 독립적으로 25번 반복실험을 하였다. 만약 배양시간이 정규분포를 따르고 표준편차는 기존의 방법에 의한 표준편차인 4시간 정도라고 가정하자. 만약 평균배양시간이 28.56시간으로 나왔다면,

(1) 이 세포배양법에 의한 독감백신 평균배양시간의 95%신뢰구간을 구하여라.

$$28.56 \pm 1.96\times4/5 = 28.56 \pm 1.568 = [26.992,\ 30.128]$$

새로운 방법의 평균배양시간이 기존 독감백신 평균배양시간인 30시간보다 단축되었는지를 알아보고자 한다.

⑵ 가설을 정하고 적절한 검정통계량과 해당 확률분포를 유도하여라.

μ : 새로운 방법의 평균배양시간, $\overline{X}$:25번 반복실험 결과의 평균배양시간

가설: $H_0 : \mu \geq 30$ vs $H_1 : \mu < 30$

검정통계량: $\overline{X}$ ⇨ 귀무가설 하에서 $\overline{X} \sim N(30, 4^2/25)$

$$\Rightarrow \; Z_0 = \frac{\overline{X} - 30}{4/5} \sim N(0,1)$$

⑶ 유의수준 5%에서 기각역을 구하고 25번의 실험결과로 나온 평균배양시간 28.56시간으로 결론을 도출하여라.

기각역 $Z_0 \leq k$ ⇨ $k = -z_{0.05} = -1.645$

$z_0 = \dfrac{28.56 - 30}{4/5} = -1.8 < -1.645$ 이므로 5% 유의수준에서 귀무가설 기각

⇨ 새로운 방법은 기존의 방법보다 배양시간을 단축시킨다고 할 수 있음

⑷ 동일한 자료에 대해 유의확률을 계산하고 1% 유의수준에서 결론을 도출하여라.

유의확률: $P(Z \le -1.8) = 0.0359 > 0.01$ 이므로 1% 유의수준에서는 귀무가설 기각시키지 못함 ⇨ 단축시킨다고 볼 수 없음

⑸ ⑶에서 구한 임계값으로 기각여부를 결정한다면 $H_1 : \mu = 28$ 일 때 제2종 오류 확률을 구하여라.

$$P_{H_1}\left(\frac{\overline{X}-30}{4/5} > -1.645\right) = P_{H_1}(\overline{X} > 28.684)$$
$$= P_{H_1}\left(\frac{\overline{X}-28}{4/5} > \frac{28.684-28}{4/5}\right) = P(Z > 0.855) = 0.1963$$

8.5 어떤 제과점에서 판매하던 기존 파이의 평균 칼로리는 245kcal이었다고 하자. 이 제과점에서 칼로리를 낮춘 다이어트 파이를 만들었다고 하자. 다이어트 파이는 기존 파이와 비교해 맛과 질감 등에서는 차이가 나지 않지만 칼로리만 낮추었다고 한다. 이를 알아보고자 16개의 파이를 조사하였다. 이 다이어트 파이의 칼로리 분포는 표준편차가 8인 정규분포라고 가정한다. $\mu_0 = 245,\ \sigma = 8,\ n = 16$

⑴ 16개의 평균 칼로리가 242kcal이었다고 하면, 다이어트 파이의 평균 칼로리에 대한 95%신뢰구간을 구하여라.

242 ± 1.96×8/4 = [238.08, 245.92]

다이어트 파이의 평균 칼로리가 기존 파이의 평균 칼로리보다 낮은지를 알아보고자 한다.

⑵ 가설을 설정하고 검정통계량과 해당 확률분포를 유도하여라.

μ : 다이어트 파이의 평균 칼로리, $\overline{X}$: 16개 파이의 표본 평균칼로리

가설: $H_0 : \mu \geq 245$ vs $H_1 : \mu < 245$

검정통계량: $\overline{X}$

귀무가설 하에서 $\overline{X} \sim N(245, 8^2/16)$ $\Rightarrow$ $Z_0 = \dfrac{\overline{X} - 245}{8/4} \sim N(0,1)$

⑶ 만약 16개의 평균 칼로리가 241.5kcal보다 낮을 때 다이어트 파이의 평균 칼로리가 기존 것보다 낮다고 판정한다면 유의수준을 구하여라.

$$P_{H_0}(\overline{X} \leq 241.5) = P\left(Z \leq \frac{241.5 - 245}{2}\right) = P(Z \leq -1.75) = 0.0401$$

⑷ 16개의 평균 칼로리가 242kcal이었다고 할 때, 유의수준 5%에서 기각역을 구하고 가설검정의 결론을 도출하여라.

5% 유의수준에서의 기각역 : $Z_0 \le k$ ⇨ $k = -z_{0.05} = -1.645$

$\Rightarrow\ z_0 = \dfrac{242-245}{2} = -1.5 > -1.645$ 이므로 귀무가설 기각시키지 못함.

다이어트 파이의 평균칼로리는 기존 파이의 평균칼로리보다 낮다고 보고 어려움

⑸ 16개의 평균 칼로리가 242kcal이었다고 할 때, 유의확률을 계산하고 5%와 10% 유의수준에서 결론을 도출하여라.

유의확률: $P(Z_0 \le -1.5) = 0.0668$ ⇒ 유의수준 5%에서는 귀무가설 기각시키지 못하지만 10%에서는 귀무가설 기각시킬 수 있음

9장. 단일모집단에 대한 통계적 추론

9.1 통계학 공부가 얼마나 스트레스를 주는지를 알아보기 위해 무작위로 선택된 6명의 학생을 대상으로 기초통계학을 수강하기 전보다 후의 스트레스 증가량에 측정한 결과 아래와 같은 자료를 얻었다. 이 증가량의 분포가 정규분포를 따른다고 가정할 때 물음에 답하여라.

4	-2	6	3	0	6

⑴ 스트레스 증가량의 표본평균과 표본표준편차를 계산하여라.

$$\overline{x} = 17/6 = 2.833\,,\ \ s^2 = (101 - 6 \times 2.833^2)/5 = 10.567$$

$$\Rightarrow\ \ s = \sqrt{10.567} = 3.251$$

⑵ 기초통계학이 학생의 학업스트레스를 높이는 요인이라고 할 수 있는

지를 5% 유의수준으로 검정하여라. μ: 평균 스트레스 증가량

◦ 가설: $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$

◦ 검정통계량: $T = \dfrac{\overline{X}}{S/\sqrt{n}} \sim t_{n-1}$, $t = \dfrac{2.833 - 0}{3.251/\sqrt{6}} = 2.135$

◦ $t = 2.135 > t_{0.05,5} = 2.015$ 이므로 5%유의수준에서 귀무가설 기각 할 수 있다. ⇒ 기초통계학이 스트레스를 증가시키는 요인이라고 할 수 있음

⑶ 스트레스 증가량의 표준편차에 대한 95% 신뢰구간을 구하여라.

$$\left[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}}}\right] \Rightarrow \left[\sqrt{\frac{5 \times 10.567}{12.83}}, \sqrt{\frac{5 \times 10.567}{0.83}}\right]$$

= [2.029, 7.973]

9.4 다음 값은 어떤 사이트에 올라온 숙대입구역 근처에 있는 원룸 중

보증금이 1000만 원인 월세 자료(단위:만 원)이다. 정규성 검정결과 정규분포를 따른다고 할 수 있다고 하면 물음에 답하여라.

70, 90, 80, 55, 100, 110, 110, 110, 100,
80, 80, 95, 68, 80, 95, 55, 85, 80

⑴ 이 자료를 이용하여 표본평균과 표본분산을 계산하여라.

$\overline{x}=85.722, \quad s^2=292.918$

⑵ 이 지역 원룸의 평균월세에 대한 95%신뢰구간을 구하여라.

$t_{0.025,17}=2.11$

$85.722 \pm 2.11\sqrt{\dfrac{292.918}{18}} \Rightarrow [77.211,\ 94.233]$

⑶ 월세의 표준편차에 대한 95%신뢰구간을 구하여라.

$$\left[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}},\sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}}}\right] \Rightarrow \left[\sqrt{\frac{17\times 292.918}{30.19}},\sqrt{\frac{17\times 292.918}{7.56}}\right]$$

= [12.843, 25.665]

(4) 부동산 전문가에 의하면 이 지역 원룸의 실제 평균월세는 70만원이라고 할 때 호가가 70만원보다 높다고 할 수 있는지 5%유의수준에서 검정하여라.

μ: 평균호가

◦ 가설: $H_0: \mu \le 70$ vs $H_1: \mu > 70$

◦ $t = \dfrac{85.722-70}{\sqrt{292.918/18}} = 3.897 > t_{0.05,17} = 1.74$이므로 5% 유의수준에서 귀무가설 기각 할 수 있다.

=> 호가가 실제 월세보다 높다고 할 수 있음

9.6 어떤 동물학자가 아프리카 코끼리의 평균수명에 대해 알아보기 위해

15마리의 야생코끼리에 대한 자료를 수집하였다. 이들 코끼리의 평균수명은 27.8년이고 표준편차는 5.5년이라고 하자.

$n = 15$, $\bar{x} = 27.8$, $s = 5.5$

⑴ 아프리카 코끼리의 실제 평균수명에 대해 95%신뢰구간을 구하여라.

$27.8 \pm 2.145 \times 5.5 / \sqrt{15}$ = [24.754, 30.846]

⑵ 이 신뢰구간추정이 적절하기 위해 필요한 가정에 대해 설명하여라.

=> 코끼리 수명은 $N(\mu, \sigma^2)$ 따르고 15마리 코끼리는 모집단으로부터 무작위로 선정

⑶ 인도 코끼리의 수명은 평균 60년이라고 한다. 아프리카 코끼리는 인

도 코끼리보다 수명이 짧은지 1% 유의수준에서 가설검정하여라.

μ: 아프리카 코끼리 평균수명

◦ 가설: $H_0: \mu \geq 60$ vs $H_1: \mu < 60$

◦ $t = \dfrac{27.8 - 60}{5.5/\sqrt{15}} = -23.24 < -2.624 = -t_{0.01,14}$ 이므로 1% 유의수준에서 귀무가설 기각 할 수 있다.

⇒ 아프리카 코끼리 수명은 인도코끼리 수명보다 짧다고 할 수 있음

9.7 어떤 연구에 의하면 우리나라 40대의 고혈압 유병율이 24.8%라고 한다. 이를 확인하기 위해 40대 성인 200명의 최근 의료기록을 무작위로 뽑아 유병률을 조사한 결과 68명이 고혈압에 걸린 것으로 나타났다.

유병율: $\theta = 0.248$, 표본유병율: $p = 68/200 = 0.34$,

$n = 200$ ⇨ 정규근사가능

⑴ 만약 이 연구의 주장이 옳다면 200명 중 고혈압에 걸린 사람은 평균 몇 명 정도인가?

200×0.248 = 49.6(명)

⑵ 조사 결과를 활용하여 우리나라 40대의 고혈압 유병률에 대한 95% 신뢰구간을 구하여라.

$0.34 \pm 1.96\sqrt{0.34 \times 0.66/200} =$[0.2743, 0.4057]

⑶ 조사 결과를 활용하여 우리나라 40대의 고혈압 유병률이 24.8%라는 주장에 대해 유의수준 5%에서 가설검정하여라.

가설: $H_0 : \theta = 0.248$ vs $H_1 : \theta \neq 0.248$

◦ $|z| = \left| \dfrac{0.34 - 0.248}{\sqrt{0.248 \times 0.752/200}} \right| = 3.068 > 1.96 = z_{0.025}$이므로 5% 유의수준에서 귀무가설 기각

⇒ 고혈압 유병율이 24.8%라는 주장은 문제가 있음

⑷ 위의 조사에서 사용된 표본이 우리나라 40대 고혈압 유병률을 분석하는 데 적절하다고 볼 수 있는가? 문제가 있다면 어떤 문제가 있으며 어떻게 표본을 선정해야 하는지 설명하여라.

=> 의료기록은 병원에 방문한 사람, 즉 환자들을 많이 포함되어 있을 가능성이 높아 40대 전체를 대표하는 표본으로 적절하지 않을 수 있음. 단순 치료를 위한 의료기록보다는 건강검진 자료와 같은 자료에서 표본을 선정하는 것이 적절함

9.8 어떤 공장에서 생산된 정밀 부품의 불량률 θ를 알아보려고 한다.

⑴ 95% 신뢰수준에서 한계오차를 $\pm 1.5\%p$까지 허용하고 싶을 경우, 몇 개의 표본을 추출해야 하는가?

$$n \geq 0.5(1-0.5)\left(\frac{1.96}{0.015}\right)^2 = 4268.444$$

⇒ 4269개 이상 뽑아야 한다.

⑵ 불량률 θ에 대해 추론하기 위해 200개의 부품을 무작위로 추출하였다. 비록 비복원추출이지만 이 공장에서는 매일 수만 개의 부품을 생산하고 있어 복원추출과 차이가 거의 없다고 가정할 수 있다고 하자.

① 만약 θ가 0.3이라면 불량제품의 표본비율이 0.27과 0.33사이에 있을 확률을 근사적으로 구하여라.

◦ $\theta = 0.3 \Rightarrow P \sim N\left(0.3, \frac{0.3 \times 0.7}{200}\right), \quad \frac{0.3 \times 0.7}{200} = 0.00105$

$$\begin{aligned} P(0.27 \le P \le 0.33) &= P\left(\frac{0.27 - 0.3}{\sqrt{0.3 \times 0.7/200}} \le Z \le \frac{0.33 - 0.3}{\sqrt{0.3 \times 0.7/200}}\right) \\ &= P(-0.93 \le Z \le 0.93) = 0.6476 \end{aligned}$$

② 200개의 표본 중 174개가 정상으로 관찰되었다고 할 때 θ에 대한 90%신뢰구간을 구하여라.

정상제품비율: $174/200 = 0.87$ ⇨ 불량제품비율 $p = 0.13$

$$0.13 \pm 1.645\sqrt{\frac{0.13 \times 0.87}{200}} = 0.13 \pm 0.039 = [0.091,\ 0.169]$$

③ 이 공장에서는 불량률이 20% 미만이라고 주장한다.

ⓐ 귀무가설과 대립가설을 θ로 설정하여라.

$H_0 : \theta \geq 0.2$ vs $H_1 : \theta < 0.2$

ⓑ 200개의 표본 중 175개 이상이 정상이면 이 회사의 주장을 채택한다면 제1종 오류를 범할 확률을 구하여라.

175개 이상 정상 ⇨ 25개 미만 불량

$$\alpha = P_{H_0}(X < 25) = P_{H_0}(X \leq 24.5) \approx P\left(Z \leq \frac{24.5 - 40}{\sqrt{32}}\right)$$
$$= P(Z \leq -2.74) = 0.0031$$

또는 $\alpha = P_{H_0}(P < 0.125) = \left(Z \le \dfrac{0.125 - 0.2}{\sqrt{0.2 \times 0.8/200}}\right)$

$$= P(Z \le -2.65) = 0.004$$

(c) 200개의 표본 중 179개가 정상으로 관찰되었을 때 5%유의수준에서 이 공장의 주장을 검정하여라.

179개 정상 ⇨ 21개 불량 ⇒ $p = 21/200 = 0.105$

$z = \dfrac{0.105 - 0.2}{\sqrt{02 \times 0.8/200}} = -3.359 < -1.645$ 이므로 5% 유의수준에서 귀무가설 기각 할 수 있다. ⇒ 불량률이 20% 미만이라고 할 수 있음

(d) (c)에 대한 p-값: $P(Z \le -3.359) = 0.0004$

10. 두 모집단의 비교

10.1 진통제 효과를 비교하는 실험에서 9명의 환자에게 A진통제를, 12명의 환자에게 B진통제를 복용시킨 후 무통시간을 측정하였다. 두 진통제의 무통시간은 동일한 분산을 가지는 정규분포를 따른다고 할 때 다음의 질문에 답하여라.

통계량	A진통제	B진통제
표본평균	5.8	5.0
표본표준편차	1.3	1.5

(1) 합동표준편차 S_p가 1,419인 것을 보여라.

$$\sqrt{\frac{8\times 1.3^2+11\times 1.5^2}{9+12-2}}=1.419$$

⑵ A진통제의 평균무통시간에 대해 95%신뢰구간을 구하여라.

$\overline{x} \pm t_{0.025,8}\, s/\sqrt{m} = 5.8 \pm 2.306 \times 1.3/\sqrt{9}$ =[4.801, 6.799]

⑶ A진통제의 무통시간 분산이 진통제의 무통시간 분산보다 작은지를 5%유의수준에서 검정하여라.

가설: $H_0 : \sigma_A^2 \geq \sigma_B^2$ vs $H_1 : \sigma_A^2 < \sigma_B^2$

검정통계량의 값: $f = 1.3^2/1.5^2 = 0.751 > F_{0.95,8,11} = 0.302$이므로

5% 유의수준에서 H_0 기각할 수 없음

⇒ A의 분산이 B의 분산보다 작다고 할 수 없음

⑷ A진통제가 B진통제보다 진통효과가 있는지를 5%유의수준에서 검정하여라. μ: 평균진통효과시간

- 가설: $H_0: \mu_A \le \mu_B$ vs $H_1: \mu_A > \mu_B$
- 검정통계량의 값: $t = \dfrac{5.8-5}{1.419\sqrt{1/9+1/12}} = 1.2785 < t_{0.05,19} = 1.729$ 이므로 5% 유의수준에서 H_0 기각할 수 없음 ⇒ A의 진통효과가 B의 진통효과보다 있다고 할 수 없음

10.2 다음 자료는 14년 6월 4일 지방선거에 당선한 17개 지역 광역단체장의 소속정당과 득표율을 정리한 것이다.

지역	충북	강원	인천	대전	경기	부산	충남	대구	서울
정당	새민련	새민련	새누리	새민련	새누리	새누리	새민련	새누리	새민련
지지율	49.75	49.76	49.95	50.07	50.43	50.65	52.21	55.95	56.12
지역	세종	광주	경남	제주	울산	전북	경북	전남	
정당	새민련	새민련	새누리	새누리	새누리	새민련	새누리	새민련	
지지율	57.78	57.85	58.85	59.97	65.42	69.23	77.73	77.96	

유권자수 또는 투표자수 등 다른 정보가 없다는 가정하에 이들 득표율만 이용하여 다음 물음에 답하여라.

(1) 각 정당의 평균득표율에 대한 95% 구간추정을 하여라.

새민련: $\overline{x} = \dfrac{520.73}{9} = 57.859, \quad s_x^2 = \dfrac{1}{8}(30889.17 - 9 \times 57.859^2) = 95.024$

$\Rightarrow \ \overline{x} \pm t_{0.025,8}\sqrt{s_x^2/m} = 57.859 \pm 2.306\sqrt{95.024/9} = [50.366,\ 65.352]$

새누리: $\bar{y} = \frac{468.95}{8} = 58.619$, $s_y^2 = \frac{1}{7}(28115.465 - 8 \times 58.619^2) = 89.4575$

$\Rightarrow \ \bar{y} \pm t_{0.025,7}\sqrt{s_y^2/n} = 58.619 \pm 2.365\sqrt{89.4575/8} =$ [50.712, 66.526]

⑵ 두 정당의 평균 득표율에 차이가 있는지를 5% 유의수준에서 검정하여라.

가설: $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$

- $s_y^2/s_x^2 = 0.9413$으로 분산이 같다고 할 수 있으므로 합동분산추정치는 $s_p^2 = \frac{8 \times s_x^2 + 7 \times s_y^2}{9 + 8 - 2} = 92.433$

- $|t| = \left| \dfrac{58.619 - 57.859}{\sqrt{92.433(1/9 + 1/8)}} \right| = 0.163 < t_{0.025, 15} = 2.131$이므로 5% 유의 수준에서 H_0 기각할 수 없음

  ⇒ 두 당의 평균득표율에 차이가 있다고 보기 어려움

10.4 아래 자료는 Gosset이 Student라는 필명으로 발표한 논문에서 언급한 자료이다. 이 자료는 수면제를 복용하지 않은 10명의 환자에 대해 수면시간을 측정하고 Ⓐ D. hyposcyamine hydrobromide와 Ⓑ L. hyposcyamine hydrobromide 처치를 한 후의 수면제 복용 전에 비해 늘어난 수면 시간을 기록하였다.

환자	Ⓐ	Ⓑ	Ⓑ−Ⓐ	환자	Ⓐ	Ⓑ	Ⓑ−Ⓐ
1	0.7	1.9	1.2	6	3.4	4.4	1.0
2	−1.6	0.8	2.4	7	3.7	5.5	1.8
3	−0.2	1.1	1.3	8	0.8	1.6	0.8
4	−1.2	0.1	1.3	9	0.0	4.6	4.6
5	−1.0	−0.1	0.9	10	2.0	3.4	1.4

(1) 처치 Ⓐ를 했을 때 수면시간이 얼마나 늘었는지 95%신뢰구간을 구하여라.

$m=10$, $\overline{x}=0.66$, $s_x^2=3.4516$, $s_x=1.8578$, $t_{0.025,9}=2.262$

- $\overline{x}\pm t_{\alpha/2,m-1}s_x/\sqrt{n}=0.66\pm 2.262\times 1.8578/\sqrt{10}=[-0.669, 1.989]$

(2) 처치 Ⓑ를 했을 때 수면시간이 늘었다고 할 수 있는지 5%유의수준에서 검정하여라.

가설: $H_0 : \mu_B \le 0$ vs $H_1 : \mu_B > 0$

- $n = 10$, $\overline{y} = 2.33$, $s_y^2 = 4.009$, $s_y = 2.0022$

- 검정통계량의 값: $t = \dfrac{2.33}{2.0022/\sqrt{10}} = 3.6799 > t_{0.05,9} = 1.833$ 이므로 5% 유의수준에서 H_0 기각할 수 있음 ⇒ 처치 Ⓑ를 했을 때 수면시간이 늘었다고 할 수 있음

⑶ 처치 Ⓑ가 처치 Ⓐ보다 수면시간을 늘려주는 효과가 있는지를 5%유의수준에서 검정하여라.

$\delta = \mu_y - \mu_x$ 라고 하면 $H_0 : \delta \le 0$ vs $H_1 : \delta > 0$

- $\overline{d} = 1.67$, $s_d = 1.1304$

- 검정통계량의 값: $t = \dfrac{1.67}{1.1304/\sqrt{10}} = 4.672 > t_{0.05,9} = 1.833$이므로

  5% 유의수준에서 H_0기각 할 수 있다.

  ⇒ 처치 Ⓑ가 유의하게 처치 Ⓐ보다 수면시간을 늘려주는 효과가 있다고 할 수 있음

⑷ 만약 9번째 환자의 Ⓑ - Ⓐ값을 17.6으로 잘못 계산했다고 하자.

① Ⓑ가 처치 Ⓐ보다 수면시간을 늘려주는 효과가 있느지를 5%유의수준에서 검정하여라.

가설 : $H_0 : \delta \le 0$ vs $H_1 : \delta > 0$

- $\bar{d} = 2.97,\ s_d = 5.1616$

- 검정통계량: $t = \dfrac{2.97}{5.1616/\sqrt{10}} = 1.8196 < t_{0.05,9} = 1.833$이므로 5% 유의수준에서 H_0기각 못함

⇒ 처치 Ⓑ가 유의하게 처치 Ⓐ보다 수면시간을 늘려주는 효과가 있다고 할 수 없음

② (3)의 결과와 비교하여 왜 이런 결과가 나왔는지 원인을 알아보아라.

차가 모두 양수임에도 불구하고 유의하지 않은 것은 17.6이라는 극단적인 이상치가 자료에 포함되었기 때문. 이 이상치는 검정통계량의 분자인 평균보다 분모에 있는 표준편차의 값이 상대적으로 커져 유의하지 않게 만듬

⑸ 위의 추론을 하는데 있어 필요한 가정을 자료가 만족하는지 확인하여라.

$Q_1 = 1.05,\ Q_3 = 1.7,\ IQR = 0.975,$ 상한= 2.675, 하한=0.075

상자그림 결과 9번째 자료가 4.6일 때도 이상치로 판정되었고 자료의 크기도 크지 않아 비모수적인 방법을 적용하는 것이 적절함

10.6 다음 자료는 기초통계학을 수강하는 학생들 중 6명을 무작위로 뽑아 중간시험과 기말시험의 점수를 표시한 것이다.

번호	1	2	3	4	5	6
중간점수	65	72	93	87	77	55
기말점수	69	70	99	91	88	62
기말-중간	4	-2	6	4	11	7

기말시험 점수가 중간시험 점수보다 높은지를 5%유의수준에서 검정하여라.

가설 : $H_0 : \delta \leq 0$ vs $H_1 : \delta > 0$

- $\bar{d} = 5,\ s_d = 4.2895$

- 검정통계량의 값: $t = \dfrac{5}{4.2895/\sqrt{6}} = 2.855 > t_{0.05,5} = 2.015$이므로

5% 유의수준에서 H_0기각 할 수 있다.

⇒ 기말시험 점수가 중간시험 점수보다 높다고 할 수 있다.

10.8 다음 자료는 A지역에서 1862명을 B지역에서 188명을 임의로 뽑아 태어난 계절을 조사한 결과이다.

계절		봄	여름	가을	겨울
지역	A	480	454	446	482
	B	46	53	49	40

(1) A지역의 자료만 이용하여 다음 물음에 답하여라.

① 봄과 겨울의 출생비율이 여름과 가을의 출생비율보다 높다고 할 수 있는지를 5%유의수준에서 검정하여라.

봄·겨울: 962명, 여름·가을: 900

θ: 봄·겨울 출생비율

- 가설: $H_0 : \theta \le 0.5$ vs $H_1 : \theta > 0.5$

- $n = 1862,\ p = 0.5166$

⇨ 검정통계량의 값: $z = \dfrac{0.5166 - 0.5}{\sqrt{\dfrac{0.5 \times 0.5}{1862}}} = 1.437 < 1.645 = z_{0.05}$이므로 5% 유의수준에서 H_0 기각못함 ⇒ 봄과 겨울의 출생비율이 여름과 가을의 출생비율보다 높다고 할 수 없음

② 봄의 출생비율에 대해 95%신뢰구간을 구하여라.

봄: 480 ⇨ $p = 480/1862 = 0.2578$

- $p \pm z_{0.025}\sqrt{\dfrac{p(1-p)}{n}} = 0.2578 \pm 1.96\sqrt{\dfrac{0.2578(1-0.2578)}{1862}}$ =[0.2379, 0.2777]

⑵ A지역의 여름 출생비율과 B지역의 여름 출생비율의 차에 대한 95% 신뢰구간을 구하여라.

θ_1 : A 지역 여름출생비율, θ_2 : B지역 여름출생비율

- $p_1 = 454/1862 = 0.2458$, $p_2 = 53/188 = 0.2819$
- $p_1 - p_2 \pm z_{0.025}\sqrt{\dfrac{p_1(1-p_2)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ =[-0.1053, 0.0291]

⑶ A지역과 B지역의 겨울 출생비율이 동일한지를 5% 유의수준에서 검정하여라.

θ_1 : A 지역 겨울출생비율, θ_2 : B지역 겨울출생비율

- 가설: $H_0 : \theta_1 = \theta_2$ vs $H_1 : \theta_1 \neq \theta_2$
- $n_1 = 1862, p_1 = 482/1862 = 0.2589,$

  $n_2 = 188, p_2 = 40/188 = 0.2128,$ $p = 522/2050 = 0.2546$
- 검정통계량의 값: $z = \dfrac{0.2589 - 0.2128}{\sqrt{0.2546(1-0.2546)\left(\dfrac{1}{1862} + \dfrac{1}{188}\right)}} = 1.3828$

  ⇨ $|z| < 1.96 = z_{0.025}$이므로 5% 유의수준에서 H_0 기각못함

⇒ A 지역 겨울출생비율과 B지역 겨울 출생비율에 유의한 차이가 없음

10.10 [예제10.9]에서 아스피린을 복용한 사람 중 뇌졸중에 걸린 비율을 θ_1, 위약을 복용한 사람 중 뇌졸중에 걸린 비율을 θ_2라고 할 때 다음 물음에 답하여라.

그룹	심장마비	뇌졸중	전체
아스피린	139	119	11,037
위약	239	98	11,034
전체	378	217	22,071

(1) 각 모수에 대한 95% 신뢰구간을 구하여라.

$$p_1 = \frac{119}{11037} = 0.0108,\ \ p_2 = \frac{98}{11034} = 0.0089$$

- θ_1: $p_1 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{11037}}$ = 0.0108 ± 0.0019=[0.0089, 0.0127]

- θ_2: $p_2 \pm 1.96\sqrt{\frac{p_2(1-p_2)}{11034}}$ =0.0089 ± 0.0018=[0.0071, 0.0107]

⑵ $\theta_1 - \theta_2$에 대한 95% 신뢰구간을 구하여라.

$$0.0108 - 0.0089 \pm 1.96\sqrt{\frac{0.0108 \times 0.9892}{11037} + \frac{0.0089 \times 0.9911}{11034}}$$

⇒ 0.0019 ± 0.0026 = (-0.0006, 0.0045)

(3) 아스피린을 복용한 사람들이 위약을 복용한 사람들보다 뇌졸중에 걸릴 확률이 높은지 5% 유의수준에서 검정하여라.

가설: $H_0 : \theta_1 \le \theta_2$ vs $H_1 : \theta_1 > \theta_2$

$p = (119 + 98)/(11037 + 11034) = 0.0098$

$$z = \frac{0.0108 - 0.0089}{\sqrt{0.0098 \times 0.9902}\sqrt{\frac{1}{11037} + \frac{1}{11034}}} = \frac{0.0019}{0.0013} = 1.462$$

기각역이 $z \geq 1.645$ 인데 검정통계치 1.462 < 1.645보다 작으므로 귀무가설을 기각할 수 없다.

$\Rightarrow$ 아스피린을 복용한 사람들이 위약을 복용한 사람들보다 뇌졸중에 걸릴 확률이 높다고 할 수 없다.