

보건통계학 중간고사

222DBG04 빅데이터분석학 이지은

1. 중심극한정리(Central Limit Theorem)에 대하여 설명하시오. [16 점]

ans) 모집단이 정규분포를 따르지 않더라도 모집단에서 무작위로 추출된 표본의 크기가 30 이상일 경우, 표본은 평균 μ , 분산 σ^2 인 정규분포를 따른다.

2. 여학생 1000 명의 키(height)는 평균 160cm, 표준편차 8cm 인 정규분포를 따른다고 한다 . 키가 170cm 이상인 여학생의 수는 몇 명인가? (풀이 과정을 함께 제시하시오) [16 점]

ans)

$$z = (170 - 160) / 8 = 1.25$$

z분포의 정규분포표를 참고하면

$$p(Z \leq 1.25) = 0.8944$$

$$1 - 0.8944 = 0.1056$$

이때 $n=1000$ 을 곱해주면 약 105명이다

3. P 값(P value)의 정의를 기술하고, 두 군의 평균 수축기혈압을 비교한 결과 P 값이 0.03 이었다면 $P=0.03$ 의 의미를 해석하시오. [16 점]

ans)

p값은 귀무가설이 참이라 가정할 때, 통계값이 실제값 이상일 확률이다. 즉 귀무가설을 기각하는 확률이다. 이때 p값이 유의수준인 0.05보다 작기때문에 귀무가설을 기각한다. 즉 제안하는 가설인 두 군의 평균에는 차이가 있다는 것이 유의하다.

4. 1 종 오류(Type I error)와 2 종 오류(Type II error)에 대하여 각각 설명하시오. [16 점]

ans) 1종오류는 귀무가설이 참인데도 귀무가설을 기각하고 대립가설을 채택하는 오류이다. 2종오류는 귀무가설이 거짓인데도 대립가설을 기각하고 귀무가설을 채택하는 오류이다.

5. 연구대상자 200 명에 대하여 A 라는 항고혈압제를 복용하기 전과 후의 수축기 혈압을 각각 측정한 결과가 다음과 같다.

	대상자 1	대상자 2	대상자 3	대상자 4	대상자 5	...	대상자 200
복용전 수축기혈압 (mmHg)	100	120	120	140	80	...	150
복용후 수축기혈압 (mmHg)	108	120	150	135	120	...	140

항고혈압제 A 의 혈압 강하 효과를 확인하기 위하여 항고혈압제 복용전 수축기 혈압의 평균과 복용후 수축기혈압의 평균을 비교하고자 한다. 귀무가설(null hypothesis, H_0)과 대립가설(alternative hypothesis, H_1)을 각각 설정하고, 어떠한 통계검정방법을 사용할 것인지 제시하시오. [16 점]

ans) 항고혈압제 A의 복용 전 평균과 복용 후 평균의 차이가 없다는 것이 귀무가설이다. 대립가설은 복용전 평균과 복용후 평균의 차이가 있다는 것이다. 이때 두 집단간의 평균 차이를 검증하기 위해서는 t검정을 사용 해야 한다.

6. SAS 실습문제

2013 년 국민건강영양조사 자료를 이용하여 ‘혈중 카드뮴 농도와 갑상선기능저하의 관련성 (Association between blood cadmium level and hypothyroid status)’에 대해 분석하고자 한다.

(1) 2013 년 국민건강영양조사 원시자료의 ‘기본 DB’를 온라인에서 다운로드 받아(knhanes.kdca.go.kr), 다음의 대상자 포함기준과 제외기준을 만족하면서 다음의 12 개 변수만 포함하는 데이터셋을 생성하여 ‘work.midterm’이라는 이름으로 저장하는 SAS code 를 작성하시오. [10 점]

- 대상자 포함기준
 - o 혈중 카드뮴 농도 측정을 한 대상자만 포함 [힌트: `WHERE HE_Cd ne .;`]
- 대상자 제외기준 (다음의 제외기준에 속하는 대상자는 제외) [힌트: 조건을 만족하는 관측치를 삭제 `IF 조건 THEN DELETE;`]
 - o 만나이 65 세 초과
 - o 임신 중인 여성 [힌트: `LW_mp` 변수 이용]
 - o 갑상선암을 의사에게 진단받은 경우 [힌트: `DC7_dg` 변수 이용]
 - o 갑상선질환으로 현재 치료를 받고 있는 경우 [힌트: `TH_tx` 변수 이용]
 - o 요중요오드 농도 변수가 결측치인 경우 [힌트: `HE_UIod=.`]
 - o 요중크레아티닌 농도 변수가 결측치인 경우 [힌트: `HE_UCREA=.`]
- 새로 생성할 변수
 - o **UICre** [단위 $\mu\text{g/g}$] $\text{UICre} = \text{요중요오드} / \text{요중크레아티닌농도}$
- 포함할 변수 (다음의 변수만 포함)
 - o age
 - o sex
 - o HE_BMI
 - o HE_Cd
 - o HE_TSH
 - o HE_ft4
 - o HE_TPOAb
 - o HE_UIod
 - o HE_UCREA
 - o UICre
 - o sm_presnt
 - o BD1

[SAS code]

```
libname mid '/home/u63345097/data for midterm';
```

```
data work.midterm;
set MID.HN13_ALL(encoding = asciiany);
WHERE HE_Cd ne .;
IF age>65 THEN DELETE;
IF LW_mp=1 THEN DELETE;
IF DC7_dg=1 THEN DELETE;
IF TH_tx=1 THEN DELETE;
IF HE_UIod=. THEN DELETE;
IF HE_UCREA=.=. THEN DELETE;
```

```
keep age sex HE_BMI HE_Cd HE_TSH HE_ft4 HE_TPOAb HE_UIod HE_UCREA UICre sm_presnt BD1;
UICre=HE_UIod/HE_UCREA;
```

```
run;
```

(2) 위에서 작성한 work.midterm 데이터셋을 이용하여 다음의 Table 1을 완성하고, 필요한 SAS code를 작성하십시오. [10 점]
 (회색 글씨는 표시 형식을 보여주는 예이므로 삭제하고, 실제 숫자를 확인후 검은색 글씨로 채워 넣으시오)

	Men	Women
n	1053	896
Age, years	Mean \pm SD	38.9 \pm 16.0
Body Mass Index, kg/m ²	Mean \pm SD	22.7 \pm 3.5
Current smoker, n (%)	n (%)	85 (9.5)
Ever drinker, n (%)	n (%)	692 (80.7)
UI/Cre, μ g/g	Median [25th, 75th percentile]	230.3 [101.5, 616.3]
Blood cadmium, μ g/L	Mean \pm SD	0.89 \pm 0.45
TSH, μ IU/mL	Mean \pm SD	2.61 \pm 1.93
ftT4, ng/dL	Mean \pm SD	0.89 \pm 0.45
TPOAb, IU/mL	Median [25th, 75th percentile]	7.10 [5.80, 9.95]

Values are expressed as mean \pm standard deviation or median (25th, 75th percentile) or n (%).

BMI, body mass index; UI/Cre, urine iodine-to-creatinine ratio; TSH, thyroid-stimulating hormone; ftT4, free thyroxine; TPOAb, antithyroid peroxidase antibody.

[SAS code]

```
data work.midterm;
set work.midterm;
IF sex=1 THEN sex_char='MALE';
else if sex=2 THEN sex_char='FEMALE';
run;

proc freq data=work.midterm; /*현재 흡연 여부*/
table sex_char*sm_presnt;
run;
proc freq data=work.midterm; /*음주 여부 0이면 경험 없음, 1이면 있음 */
table sex_char*BD1;
run;
proc means data=work.midterm MEAN STD MEDIAN P25 P75;
class sex_char;
var age HE_BMI UICre HE_Cd HE_TSH HE_ft4 HE_TPOAb;
run;
```