○ 자료분석

○ 자료의 종류와 분석 목적에 따라 분석방법을 선택

- 자료의 종류
    - 범주형자료: 명목자료, 순서자료 (ex) 남:0, 여:1 (ex) 설문지 → 구분을 위함
    - **수치자료**: 이산자료, 연속자료 → 값 자체가 의미있음
- 분석목적
    - **비교**: t-검정, 분산분석, 동질성검정, … → 두 집단의 비율이 같은지
    - 관계: 상관분석, 회귀분석, …

상관분석 → 두 변수가 관계있는지

회귀분석 → 설명변수와 반응변수의 함수(인과) 관계를 추정하고 추정된 함수를 이용하여 미래에 대해 값을 예측하는 분석

⇒ 자료의 형태에 맞는 최적의 분석방법으로 정보의 손실이 없도록

○ 비교실험에서 고려해야 할 사항

- 주요용어
  - 반응변수(response variable, 종속변수): 연구대상이 되는 변수
  - **요인(factor)**: 반응변수에 영향을 주는 변수로 질적인 변수
  - 처리(treatment): 실험단위에 적용되는 특정한 실험조건(요인의 특정값)
    - 수준(level): 어떤 한 요인이 가지는 실험조건
  - 효과(effect): 처리에 따른 반응변수의 평균차이
  - 대조(control): 처리 효과에 대한 비교 준거

예) 네 종류의 (A, B, C, D) 비료에 따른 농작물의 평균수확량에 차이가 있는지 알아보기 위한 실험

- 비료의 종류 : 요인
- A, B, C, D : 처리 혹은 수준 → 처리의 수 4
- 측정된 수확량 : 반응변수
- 수확량의 차이 : 효과 (처리효과)

(요인의 종류 = 처리
요인에 따른 반응변수의 차이 = 효과

- ◉ 담금질 용액(기름, 소금물, 혼합용액)에 따른 알루미늄 합금의 강도
  - ○ 반응변수: 알루미늄 합금의 강도
  - ○ 요인: 담금질 용액
  - ○ 처리(수준): 기름, 소금물, 혼합용액
    - 처리(수준)의 수=3

- 고려사항
  - ○ 기타 반응변수에 영향을 주는 요인에 대한 검토
  - ○ 각 처리별 반복(replication) 회수
  - ○ 처리배치(treatment allocation): 실험순서 등
  - ○ 통계분석방법 ⇨ 실험설계와 연계

○ 실험연구의 단계와 분산분석

① 연구문제의 인지 및 기술(recognition and statement of the research problem)

② 반응변수, 요인 및 수준의 선택(choice of response variable, factors, and levels)

③ 실험의 설계(designing the experiment)

④ 실험의 수행(performing the experiment)

⑤ 통계적 데이터 분석(statistical data analysis)

⑥ 결론 및 앞으로의 연구과제(conclusions and recommendations)

- 실험연구
  - 계획단계부터 통계전문가의 상담
  - 실험계획은 되도록 단순화 => 분석의 효율성 실제차이가 있어도 통계적으로 유의하지 않을수 있음
  - 실제적 유의성 (practical significance)과 통계적 유의성 (statistical significance)

  - effect size (약효의 크기, 효과 크기) : 시험약과 대조약의 치료율 차이

  ↳ 통계적으로 유의하게 차이가 있다면 얼마나?

귀무가설

Statistical significance (i.e. reject the null hypothesis) means that differences in group means are not likely due to sampling error.

○ 실험계획의 기본원리

- **반복화(replication)**
  - ○ 통계적 오차를 제어하고 추정가능하게 하는 역할
    - \- 실험은 조절할 수 없는 외부요인에 영향을 받음
    - \- 처리별 평균에서 여러 잡음요인의 상쇄효과 기대
- **확률화(randomization, 랜덤화, 임의화)**
  - ○ 실험의 객관성을 보장
  - ○ 실험에서 고려되지 않고 있는 다른 요인들의 영향을 상쇄
    - \- 통제할 수 없는 외적요인을 확률적으로 비슷하게 만듬
- **블록화(blocking)**
  - ○ 동일적인 실험단위로 묶어 실험의 정밀도를 향상

○ 비교연구의 유형

- 단일그룹 사후관측법:

  처리 ⇒ 관측

  - 대조그룹 없이 처리그룹만 사후 측정
  - 충분한 과거자료가 있어 대조그룹을 새로 측정할 필요가 없는 경우

- 처리-대조 사후관리법:

  처리 ⇒ 관측, 대조 ⇒ 관측

  - 처리그룹과 대조그룹을 모두 사후 측정
  - 처리와 대조 두 그룹의 확률화
    - 처리 전 두 그룹의 비슷한 성질을 가져야 함
    - 처리 전 두 그룹 간에 차이가 있는 경우 분석 시 보정항 추가

- 처리-대조그룹 사전·사후관측법:

  관측 ⇒ 처리 ⇒ 관측, 관측 ⇒ 대조 ⇒ 관측

  - 처리그룹과 대조그룹을 모두 사전, 사후 측정
  - 사전 관측값과 사후 관측값의 차이를 통계적으로 분석

    ↳ 같은샘플내에서 차이가있는지

- 코호트연구(cohort study)
  - 코호트(cohort): 동일한 특성을 가진 개체들의 집단(처리그룹, 대조그룹)
  - 전향적연구(prospective study): 시간순서로 실험이 이루어지는 연구 (follow up)

시간이 오래걸림
비용, 노력↑

    ⇎ 후향적연구(retrospective study) : 결과를 얻은 후 분류가 이루어지는 경우

    - 예) 폐암발생 여부 자료를 얻은 후 흡연 여부 확인

- 사례-대조연구(case-control study)
  - 처리배치가 확률화 되지 않은 실험에서는 처리 대신 사례(case)라는 용어 사용
  - 윤리적으로 문제로 실험이 불가능한 경우 아 후향적연구
    - 예) 약물복용여부에 따른 비행발생여부

- 편향(bias)
  - 선택편향(selection bias): 결과가 실험자에 유리하도록 실험개체를 선택하는 편향
  - 반응편향(response bias): 실험개체들이 처리에 보인 스스로의 효과로 인해 발생하는 편향
    - 어떤 약을 먹고 있는지 환자가 알고 있는 경우 → 환자에게 알려주면 안됨
  - 관측편향(observation bias): 처리결과의 측정 시 처리그룹에 유리하게 자료가 관측되는 편향
    - 어떤 약을 먹고 있는지 의사가 알고 있는 경우 → 의사에게 알려주면 안됨 자의적 판단 X

- 해결방법
  - 선택평향 ⇨ 임의배치(random allocation)
  - 반응편향, 관측편향 ⇨ 이중눈가림(double blinding, 이중맹검)

의사, 환자 모두에게 비밀

9/4

■ 통계적 가설 검정(statistical hypothesis testing)

- 모집단의 모수 또는 특성에 대한 주장을 설정하고 이것의 옳고 그름을 표본으로부터 얻어진 정보를 이용하여 확률적으로 판정하는 과정

전수조사를 할 수 없으니까

- 가설(hypothesis)
  ① 귀무가설(H_0) : 검정의 대상이 되는 가설
  ② 대립가설(H_1) : 표본으로부터 얻은 강력한 증거에 의해 입증하고자 하는 가설

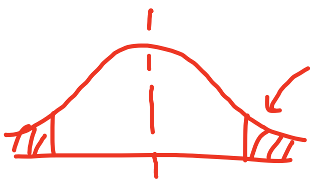
◉ 새로 개발된 항암제는 기존의 항암제보다 우수하다
  - 대립가설 : 기존 항암제보다 5년 생존율이 높다.
  - 귀무가설 : 5년 생존율에서 차이가 없다.

A→B
~B → ~A

[정상적인] 표본 ⇒ H_1 참
(대우) H_0 참 ⇒ **[비정상적인]** 표본

- 검정통계량(test statistics)
  - 귀무가설을 기각시킬 것인가, 채택할 것인가를 결정하기 위해 사용되는 통계량
  - 귀무가설 하에서 이 통계량의 확률분포를 이용하여 기각역(reject region)과 채택역(acceptance region)을 결정
  - 임계값은 대립가설의 형태(단측 또는 양측)와 **유의수준**(α)에 의해 결정
  - **유의확률(p-값)**을 이용하기도 함

H_0가 참일때 H_0를 기각할 확률 (오류)
ex) $\alpha=0.05$



기각역 : H_0 기각
H_1 채택
→ 차이가 유의하다

○ 두 모평균의 비교

independent identically distributed

- 가정: $Y_{11},\ldots,Y_{1m} \overset{iid}{\sim} N(\mu_1,\sigma^2),\ \ Y_{21},\ldots,Y_{2n} \overset{iid}{\sim} N(\mu_2,\sigma^2)$ ⇒ 등분산!!

- 중심축량(pivotal quantity): $$\frac{\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2},$$

자유도 : $(m-1)+(n-1)$

the functions and its value can depend on the parameters of the model, but its distribution must not.
(모수를 알아야하지만 분포자체는 모수에 의존하면 안됨)
값은

- $$S_p^2 = \frac{\sum(Y_{1i}-\overline{Y}_1)^2+\sum(Y_{2i}-\overline{Y}_2)^2}{m+n-2}$$

→ 합동표본분산 (pooled sample variance)
S_p^2 은 공통분산인 σ^2 의 추정량

- 가설: $H_0\ :\ \mu_1=\mu_2$ vs $H_1\ :\ \begin{cases} a.\ \ \mu_1>\mu_2 \\ b.\ \ \mu_1<\mu_2 \\ c.\ \ \mu_1\neq\mu_2 \end{cases}$

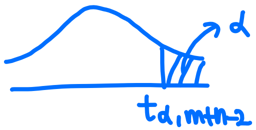
→ H_0 하에 중심축량 $=0$ 으로 두고 계산! $(\mu_1-\mu_2=0)$
중심축량은 분포가 모수에 의존하지 않음

- 검정통계량: $$\frac{\overline{Y}_1-\overline{Y}_2}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2}$$

σ

○ 유의수준을 α라고 하면, 기각역은 $\begin{cases} a.\ \ t_0 > t_{\alpha, m+n-2} & \mu_1 > \mu_2 \\ b.\ \ t_0 < -t_{\alpha, m+n-2} & \mu_1 < \mu_2 \\ c.\ \ \left| t_0 \right| > t_{\alpha/2, m+n-2} & \mu_1 \neq \mu_2 \end{cases}$

(+) t분포는 정규분포보다 꼬리가 두꺼움



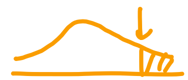
9/6

○ 여러 모집단의 평균비교

- 모든 쌍에 대해 t-검정

$P(A^c) = 1 - P(A)$
$P(A^c | B) = 1 - P(A|B)$
$\alpha = P(H_0 \text{ 기각} | H_0 \text{ 사실}) \Rightarrow 1-\alpha = P(H_0 \text{ 채택} | H_0 \text{ 사실})$

◉ 세 모집단 평균의 비교

- 가설 $H_{01}\ : \mu_1 = \mu_2$, $H_{02}\ : \mu_1 = \mu_3$, $H_{03}\ : \mu_2 = \mu_3$에 대해 검정을 실시
- 각각의 검정에 대해 유의수준을 α로 정함 : H_0가 사실일때 H_0를 기각할확률 $= \alpha$ 

  ⇨ $P(H_{0i} \text{ 채택} | H_{0i} \text{ 사실}) = 1 - \alpha$ 

- 각각의 가설검정에서 H_{0i} 를 모두 채택한 경우 H_0 채택한다고 하면?

  $P(H_0 \text{ 채택} | H_0 \text{ 사실}) = P(H_{01} \text{ 채택} \cap H_{02} \text{ 채택} \cap H_{03} \text{ 채택} | H_0 \text{ 사실})$

- Q: 유의수준 $P(H_0 \text{ 기각} | H_0 \text{ 사실})$은?

  ↳ $P(H_{01} \text{ 기각} \cap H_{02} \text{ 기각} \cap H_{03} \text{ 기각} | H_0 \text{ 사실})$

합집합의 확률은 각각의 확률의 합을 뛰어넘지 못함

- **Boole's inequality** : $P(A_1 \cup A_2 \cup A_3) \le P(A_1) + P(A_2) + P(A_3)$ $\quad P(\bigcup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} P(A_i)$

유도

⇨ **Bonferroni's inequality** :

$$P(A_1 \cap A_2 \cap A_3) \ge P(A_1) + P(A_2) + P(A_3) - 2$$

- $P(H_0 \text{ 채택} \mid H_0 \text{ 사실}) \ge (1-\alpha) + (1-\alpha) + (1-\alpha) - 2 = \boxed{1 - 3\alpha}$
- 각각의 검정에 대해 유의수준을 α로 한 경우, 실제 유의수준은

$$P(H_0 \text{ 기각} \mid H_0 \text{ 사실}) \le 1 - (1 - 3\alpha) = 3\alpha$$

→ n개의 모평균을 비교할때 유의수준을 $n\alpha$로, 원래의 유의수준보다 크게 잡음

p-value < α 인경우 H_0 기각
⇒ α를 줄일수록 기각이어려움!
즉 개별가설검정시 기각이 어려워짐 ($\frac{5\%}{3}$)
∴ type1 error를 15% (5% × 3)로 설정한다

* Bool's inequality → Bonferroni's inequality 유도

$A_i \to A_i^c$

$P(\bigcup_{i=1}^{n} A_i^c) \le \sum_{i=1}^{n} P(A_i^c)$

$P((\bigcap_{i=1}^{n} A_i)^c) \le \sum_{i=1}^{n} (1 - P(A_i))$

$1 - P(\bigcap_{i=1}^{n} A_i) \le n - \sum_{i=1}^{n} P(A_i)$

$P(\bigcap_{i=1}^{n} A_i) \ge \sum_{i=1}^{n} P(A_i) - (n-1)$

- 분산분석(analysis of variances, ANOVA)
  - 가정: $Y_{11},\dots,Y_{1m} \overset{\text{iid}}{\sim} N(\mu_1,\sigma^2)$, $Y_{21},\dots,Y_{2n} \overset{\text{iid}}{\sim} N(\mu_2,\sigma^2)$
  - 가설: $H_0 \ : \ \mu_1 = \mu_2$ VS $H_1 \ : \ \mu_1 \neq \mu_2$ ⇦ 양측검정
  - 검정통계량: $T = \dfrac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2}$

위에서 한 내용

- $T^2 \sim F_{1,m+n-2}$

$$T^2 = \frac{(\overline{Y}_1-\overline{Y}_2)^2}{S_p^2(1/m+1/n)} = \frac{\frac{mn}{m+n}(\overline{Y}_1-\overline{Y}_2)^2}{S_p^2} \sim F_{1,m+n-2}$$

1: 2-1

$$= \frac{m(\overline{Y}_1-\overline{Y})^2+n(\overline{Y}_2-\overline{Y})^2}{S_p^2}$$

확장 $\Rightarrow F = \dfrac{\left\{\sum_{i=1}^{p} n_i(\overline{Y}_i-\overline{Y})^2\right\}/(p-1)}{\sum\sum(Y_{ij}-\overline{Y}_i)^2/(N-p)} \sim F_{p-1,N-p}$

p는 집단의수, $N = \sum_{i=1}^{p} n_i$

○ 분자 $= \dfrac{mn}{m+n}(\overline{Y}_1 - \overline{Y}_2)^2 = \dfrac{(m+n)mn}{(m+n)^2}(\overline{Y}_1 - \overline{Y}_2)^2$

$$= \frac{m^2 n}{(m+n)^2}(\overline{Y}_1 - \overline{Y}_2)^2 + \frac{mn^2}{(m+n)^2}(\overline{Y}_1 - \overline{Y}_2)^2$$

$$= n\Big[\underbrace{\frac{m}{m+n}(\overline{Y}_2 - \overline{Y}_1)}_{②}\Big]^2 + m\Big[\underbrace{\frac{n}{m+n}(\overline{Y}_1 - \overline{Y}_2)}_{①}\Big]^2$$

$$= n(\overline{Y}_2 - \overline{Y})^2 + m(\overline{Y}_1 - \overline{Y})^2$$

- $\overline{Y} = \left(\sum_{i=1}^{m} Y_{1i} + \sum_{i=1}^{n} Y_{2i}\right)/(m+n)$

↳ $\overline{Y}_1$와 $\overline{Y}_2$는 종속관계
즉, 자유도가 m+n형태지만 1

각 집단과의 편차의 제곱을
가중치로 더한 값
⇒ 두 집단의 중심과의 거리라고 할 수 있음
즉, 일종의 분산

① $\dfrac{n}{m+n}(\overline{Y}_1 - \overline{Y}_2) = \dfrac{nm}{m(m+n)}(\overline{Y}_1 - \overline{Y}_2)$

$= \dfrac{1}{m(m+n)}\left(n\sum_{i=1}^{m} Y_{1i} - m\sum_{i=1}^{n} Y_{2i}\right)$

$= \dfrac{1}{m(m+n)}\left((n+m)\sum_{i=1}^{m} Y_{1i} - m\left(\sum_{i=1}^{m} Y_{1i} + \sum_{i=1}^{n} Y_{2i}\right)\right)$

$= \overline{Y}_1 - \overline{Y}$

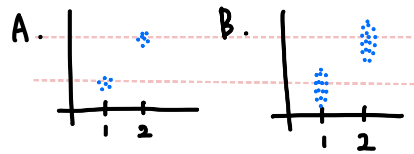
② $\dfrac{m}{m+n}(\overline{Y}_2 - \overline{Y}_1) = \dfrac{mn}{n(m+n)}(\overline{Y}_2 - \overline{Y}_1)$

$= \dfrac{1}{n(m+n)}\left(m\sum_{i=1}^{n} Y_{2i} - n\sum_{i=1}^{m} Y_{1i}\right)$

$= \dfrac{1}{n(m+n)}\left((n+m)\sum_{i=1}^{n} Y_{2i} - n\left(\sum_{i=1}^{m} Y_{1i} + \sum_{i=1}^{n} Y_{2i}\right)\right)$

$= \overline{Y}_2 - \overline{Y}$

(f) A와 B의 퍼진정도가 다를경우



A의 두집단의 평균차가 B의것보다 큼
(두 집단이 뚜렷하게 구분됨)

○ 결론

집단과전체의비교

$$\frac{\text{집단간의 변동}}{\text{집단내의변동}} = T^2 = \frac{m(\overline{Y}_1 - \overline{Y})^2 + n(\overline{Y}_2 - \overline{Y})^2}{\dfrac{\sum_{j=1}^{m}(Y_{1j} - \overline{Y}_1)^2 + \sum_{j=1}^{n}(Y_{2j} - \overline{Y}_2)^2}{m+n-2}} \sim F_{1,m+n-2}$$

→ σ^2을 추정하기위한 합동표본분산 S_p^2

↳ 각 집단내에 퍼져있는 정도

○ p개의 그룹 평균비교에 일반식 :

$$F = \frac{\sum_{i=1}^{p} n_i(\overline{Y}_i - \overline{Y})^2/(p-1)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/\sum_{i=1}^{p}(n_i - 1)} \sim F_{p-1,N-p}$$

$p-1$ → p개의 가중치 1개가 자유롭지 못함

$N-p$ → $\sum_{i=1}^{p}(n_i-1) = N-p$

- $N = \sum_{i=1}^{p} n_i$