◯ 자료분석

◦ 자료의 종류와 분석 목적에 따라 분석방법을 선택

- 자료의 종류
  - ◦ 범주형자료: 명목자료, 순서자료
  - ◦ **수치자료**: 이산자료, 연속자료

- 분석목적
  - ◦ **비교**: t-검정, 분산분석, 동질성검정, …
  - ◦ 관계: 상관분석, 회귀분석, …

◯ 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

◉ 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
  - ○ 실제적 유의성 (practical significance)과 통계적 유의성 (statistical significance)
    - effect size (약효의 크기, 효과 크기) : 시험약과 대조약의 치료율 차이

◯ 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

- 코호트연구(cohort study)
  - 코호트(cohort): 동일한 특성을 가진 개체들의 집단(처리그룹, 대조그룹)
  - 전향적연구(prospective study): 시간순서로 실험이 이루어지는 연구
    ⇔ 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
    - 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
  - 관측편향(observation bias): 처리결과의 측정 시 처리그룹에 유리하게 자료가 관측되는 편향
    - 어떤 약을 먹고 있는지 의사가 알고 있는 경우

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

■ 통계적 가설 검정(statistical hypothesis testing)

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

- 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.

[정상적인] 표본 $\Rightarrow$ H_1 참

(대우) H_0 참 $\Rightarrow$ **[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**에 의해 결정
  - **유의확률(p-값)**을 이용하기도 함

◯ 두 모평균의 비교

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

◦ 중심축량(pivotal quantity): $\dfrac{\overline{Y_1}-\overline{Y_2}-(\mu_1-\mu_2)}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2}$,

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

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

◦ 검정통계량: $\dfrac{\overline{Y_1}-\overline{Y_2}}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2}$

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

○ 여러 모집단의 평균비교

- 모든 쌍에 대해 t-검정

◉ 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

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

- 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?

◦ Boole's inequality : $P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(A_2) + P(A_3)$

⇨ Bonferroni's inequality :

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

◦ $P(H_0 \text{ 채택} | H_0 \text{ 사실}) \geq (1-\alpha) + (1-\alpha) + (1-\alpha) - 2 = 1 - 3\alpha$

◦ 각각의 검정에 대해 유의수준을 α로 한 경우, 실제 유의수준은

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

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

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

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

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

$= n\left[\frac{m}{m+n}(\overline{Y}_2 - \overline{Y}_1)\right]^2 + m\left[\frac{n}{m+n}(\overline{Y}_1 - \overline{Y}_2)\right]^2$

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

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

○ 결론

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

○ p개의 그룹 평균비교에 일반식 :

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

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