4단계가 아니라면 대면수업 대면수업기간으로? 15분전? R을 좀 알아야함

◯ 자료분석

◦ 자료의 종류와 분석 목적에 따라 분석방법을 선택

↳ 방법론 하나하나 모두 조건이 있음

↓

자료형, 분석목적에 따른 사용법, …

- 자료의 종류 남/여 방향성이, 데이터의 특성 살리는 방향
  - ◦ 범주형자료: 명목자료, 순서자료
  - ◦ **수치자료**: 이산자료, 연속자료
    - ↳ 값 자체가 중요
- 분석목적
  - ◦ **비교**: t-검정, 분산분석, 동질성검정, … (확장, 집단내의 데이터가 같은지)
  - ◦ 관계: 상관분석, 회귀분석, … (선형관계)
    - ↳ 설명변수와 반응변수의 함수(인과)
      : 관계를 추정하고 추정된 함수를 이용하여 미래에 대한 값을 예측하는 분석
      설명, 반응변수 순서 중요.

t: 두집단사이차이

분산분석: 여러 집단끼리의 처리효과

선형회귀분석 (회귀분석중 가장 쉽고 기본, 심플)

◯ 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

예제) 네 종류(A,B,C,D)의 비료에 따른 농작물의 평균수확량에 차이가있는지 알아보기위한 실험

- 요인: 비료의종류
- 처리(수준): A,B,C,D
- 반응변수: 수확량

◉ 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

만약 비교의 영향이 있는데 우연히 by chance 비슷한 값이 있다면? (통제가 안된 경우)

- 고려사항
  - ○ 기타 반응변수에 영향을 주는 요인에 대한 검토
  - ○ 각 처리별 반복(replication) 회수 A 5번 B 5번, .. 아 A 10번 B 10번, ...
  - ○ 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) 관심있는것

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 → 좀더 효과적
  - 실험계획은 되도록 단순화 => 분석의 효율성
  - 실제적 유의성 (practical significance)과 통계적 유의성 (statistical significance)
    - effect size (약효의 크기, 효과 크기) : 시험약과 대조약의 치료율 차이

고정타면 얼마나 차이가 있는지? 추론
↳ effect size

비교A,B,C,D 두학급포함
→차이가 있을때
exactly 똑같지않음.
(sampling error)
true의 일부를 보는것이기때문.

⇒ 진짜차이가 있는경우
(통계적으로 유의한차이)

◯ 실험계획의 기본원리

- **반복화(replication)** sampling error는 여러번 반복할수록↓
  - 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

ex) 땅의 비옥도에 따라 수확량에 영향이 가므로
비옥도로 땅을 블록화

◯ 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

단점: 폐암발병률 자체가 낮음 → data 확보가 어려울수 있음.
따라서 follow up하는 수가 매우 커야함 → 비효율적
오랫동안 관찰해야할 수도 있음.

- 코호트연구(cohort study)
  - 코호트(cohort): 동일한 특성을 가진 개체들의 집단(처리그룹, 대조그룹)
  - 전향적연구(prospective study): 시간순서로 실험이 이루어지는 연구 (follow up)
    ⇔ 후향적연구(retrospective study) : 결과를 얻은 후 분류가 이루어지는 경우
    - 예) 폐암발생 여부 자료를 얻은 후 흡연 여부 확인

- 사례-대조연구(case-control study) : 대표적인 후향적연구
  - 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 ex. 신약실험에서 신약쪽에 건강한사람을 더넣는경우
  - 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 ex. 신약먹고감기걸렸는데 경미한 컨디션 난것로 기록
    - 어떤 약을 먹고 있는지 의사가 알고 있는 경우

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

처리약과똑같게보이도록 만든 대조약사용

결과측정하는 의사들에게 실험개체가 받고있는 처리가무엇인지 절대로알려주지IX

추론 : 모집단이 뭔지 알아맞추기
검정 : 모집단에 대한 주장 맞다/틀리다 (문장 有)

■ 통계적 가설 검정(statistical hypothesis testing)

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

표본은 모집단의 일부이기때문

- 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. → 팔수있는거
- ○ 귀무가설 : 5년 생존율에서 차이가 없다.

A → B 를보이기힘들때 ~B → ~A을보인다.

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

H0이 참일때내가가진표본이 비정상적임을보임

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

α

H0 기각	H0 채택
검정통계량 ∈ 기각역	검정통계량 ∈ 채택역
P값 < α	P값 > α

분산분석은 t검정의 확장
↳ m개 집단

○ 두 모평균의 비교

independent identically distributed
: 서로독립이고 같은 분포를따름
+) 모집단끼리도 독립

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

The function and its value can depend on the parameters of the model but its distribution must not.

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

알알려진 함수. 즉 값 자체는 모수에 의존

합동표본분산 (pooled sample variance)
S_p^2을 공통분산인 σ^2의 추정량 (등분산 가정)

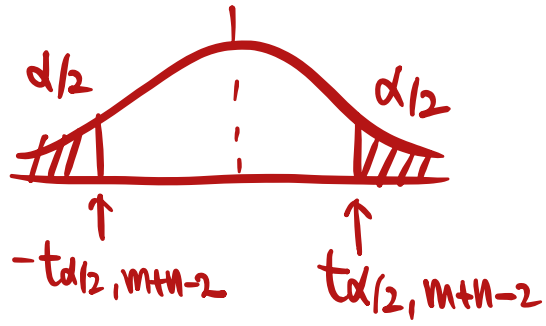
$$S_p^2 = \frac{\sum(Y_{1i}-\overline{Y_1})^2+\sum(Y_{2i}-\overline{Y_2})^2}{m+n-2}$$

분포는자유도에 의존(이미 정해진값) 즉 분포는모수에의존X

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

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

모수가 사라지므로 중심축량이 검정통계량의 좋은 후보가된다.



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

$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1-\mu_2, \left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2\right)$

$E(\bar{Y}_1-\bar{Y}_2) = E(\bar{Y}_1) - E(\bar{Y}_2) = E\left(\frac{\sum_{j=1}^{m} Y_{1j}}{m}\right) - E\left(\frac{\sum_{j=1}^{n} Y_{2j}}{n}\right)$

$= \frac{m\mu_1}{m} - \frac{n\mu_2}{n} = \mu_1-\mu_2$

$Var(\bar{Y}_1-\bar{Y}_2) = Var(\bar{Y}_1) + Var(\bar{Y}_2)$ → 독립이라 $Cov=0$

$= V\left(\frac{\sum_{j=1}^{m} Y_{1j}}{m}\right) + V\left(\frac{\sum_{j=1}^{n} Y_{2j}}{n}\right)$

$= \frac{1}{m^2}\cdot m\sigma^2 + \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{m}+\frac{\sigma^2}{n}$

α를 먼저 minimize → β minimize
∵ 현재사실 H_0을 부정하는 경우 impact가 더 크다.

결론 \ 실제	H_0 사실	H_1 사실
H_0 사실	0	β
H_1 사실	α	0

제 1종의 오류

○ 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

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

- 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? (제1종오류 α)

1-β : power(검정력)

어떤 검정이 좋다: 제1종오류는 α이하로 minimize하면서 동시에 β를 최소화 (즉 power를 maximize)

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

유도가능

A_i 대신 A_i^c 를넣는다

- Boole's inequality : $P(A_1 \cup A_2 \cup A_3) \le P(A_1) + P(A_2) + P(A_3)$
  - ⇨ Bonferroni's inequality : $P(A_1 \cap A_2 \cap A_3) \ge P(A_1) + P(A_2) + P(A_3) - 2$

$P(H_0 \text{ 채택} \mid H_0 \text{ 사실})$

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

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

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

↳ $1 - P(H_0 \text{채택} \mid H_0 \text{사실})$

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

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

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

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

각각의 검정에 대해유의수준을 5%로한경우

실제유의수준은 $3\alpha = 0.15$ (15%) 보다 작거나 같으며

5%라고할수없음.

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

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

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

$\frac{m+n}{mn}$

$\times \frac{m+n}{m+n}$

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

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

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

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

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

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

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

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

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

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

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

$= \frac{1}{(m+n)m}\left((n+m)\Sigma Y_{1i} - m(\Sigma Y_{1i} + \Sigma Y_{2i})\right)$

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

분자 : Σ (가중치·편차²)/자유도
→ 전체평균을 중심으로 각각의 평균이 얼마나 퍼져있는지 기대한 분산

○ 결론

일종의 편차

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

$\overline{Y}_1 - \overline{Y}$ 을알면 $\overline{Y}_2 - \overline{Y}$ 을알수있음
$\overline{Y}_2 - \overline{Y}$ 을알면 $\overline{Y}_1 - \overline{Y}$ 을알수있음
→ 자유도는 1

자유도 m-1
∵ $\sum_{j=1}^{m}(Y_{1j} - \overline{Y}_1) = 0$

자유도 n-1

↳ $T^2 = \frac{\text{집단간의 변화}}{\text{집단내의 변화}}$

즉 T^2 가클수록
H_1 support (특징에부합)
두집단이같지 않다
아래에서계속

○ p 개의 그룹 평균비교에 일반식 :

∵ $\sum_{i=1}^{p} n_i(\overline{Y}_i - \overline{Y}) = 0$

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

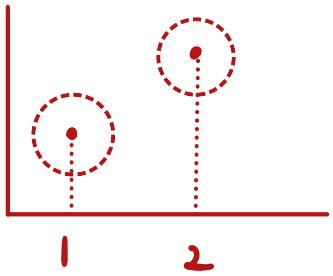
각각의 자유도의 합

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

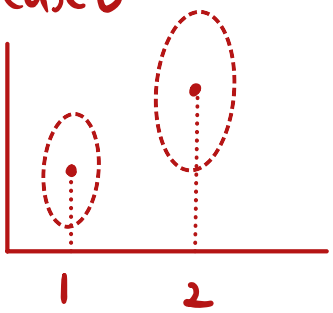
→ H_0 : 모든그룹의 평균이 같다
H_1 : 하나라도 같지 않다

어느쪽의 편차의 차이가 더 클까?

① case A



① case B



→ A의 편차의 차이가 더 크다.

집단내의 변동보다 집단간의 변동이 상대적으로 크기 때문

(즉 두 그룹이 훨씬 separate 돼있음)