의학연구통계

울산의과대학 이 무 송

본 원고는 'Essentials of Medical Statistics. Kirkwood BR. Blackwell Scientific Publications Oxford 1988 1장-10장'을 요약하였음

가. 기본적인 이해

1. 통계학이란 무엇인가?

- ***자료를 수집, 요약, 제시 및 해석***하며, 자료로써 가설(hypothesis)을 검정하는 학문
- 의학 연구에서 통계학의 중요성
- 광범위하고 공식적으로 자료를 재구성하는 방법을 제시
- 의학 분야에서 정량적인 측정이 점증
- 대부분의 생물학적 현상에는 필연적으로 내적인 변이가 있기 때문
- **변이(variability)**를 감안한 자료 해석은 통계학의 기본적 개념
- 특정한 스트레스를 많이 받는 직업과 관련된 질병 양상을 연구하는데 있어, 일반 인구의 정상치보다 높게 관찰된 평균 혈압이 단순히 ***우연에 기인한 변이 폭인지*** 아니면 실제로 그 직업이 가지는 건강상의 위해를 반영하는지 판단하기 위하여 통계학적 방법이 필요
- 백신 접종군에서 질병에 걸리지 않은 사람의 분율이 백신 비접종군보다 낮다면?
- 실제 백신에 효과가 있는 것인가?
- 이 결과가 단지 우연에 의해서 나타났다고 보아야 하는가?
- 백신을 맞은 집단을 선정하는 과정에서 오류가 있었던 것은 아닌가?
- 통계적 분석은 이상의 가능성 중 첫 번째 두 가지 가능성을 구별하는데 이용
- 적절한 연구 설계로써 세 번 째 가능성을 배제

2. 모집단과 표본 (Population and Sample)

- 대상 인구집단 전수를 조사한 경우를 제외하고, 얻어진 모든 자료는 표본에서 얻어진 것
- 표본은 그보다 대규모의 집단, 즉 모집단에서 추출
- 연구자가 관심이 있는 것은 표본 자체가 아니라 표본이 모집단에 대하여 제시하고 있는 사실
- 우연에 의하여, ***표본이 다를 경우 다른 결과***가 얻어짐
- 표본을 이용하여 모집단에 대한 추론을 하고자 할 때 이 점을 반드시 감안

표본 변이(sampling variation)

- 모집단이나 그 경계는 정확히 정의하기 어려운 경우가 많고, 모집단을 적절히 대표할 수 있는 표본의 추출에는 상당한 주의가 요구
- **대상 모집단(target population)**

3. 자료의 정의 (Defining the data)

- 연구자가 얻은 기초 자료(raw data)
- 각 개인들에 대하여 이루어진 관찰치(observations)
- 각 개인의 숫자: 표본 수(sample size)
- 각 개인에 대하여 측정된 값(혈압 등)이나 기록된 값(나이, 성별 등): **변수(variable)**
- 통계적 분석 방법: 변수의 형태에 따라 결정
- ***변수의 형태***: 정성적(또는 범주적) / 정량적(또는 수치적)
- 정성적 변수(qualitative variable)
- 이분성(binary) 변수
- 정량적 변수(quantitative variable)
- 숫자로서 값이 한정(discrete)되어 있거나 연속적(continuous)

4. 자료의 분석과 결과의 제시

- **일반적인 세 가지 원칙**
  - 단지 복잡한 분석 방법을 그 자체만으로 적용하는 것은 절대로 피해야 한다.
  - 처음에는 기본적인 요약 지표를 구하고 도시적인 기법으로 자료를 살펴보아야 한다.
  - 분석에 사용되는 기법은 자료의 성격에 비추어 꼭 필요한 기법 중 가장 간단한 것으로 선택하여야 한다.
  - 통계적인 추론은 일반적인 상식과 반드시 합치되는 방향으로 진행되어야 한다.
  - 도시적 기법(graphical techniques)을 폭넓게 사용하여야 한다.
  - 변수간의 관계, 경향성, 또한 집단간의 차이
  - 그림 (diagram)이나 표는 반드시 쉽게 알아볼 수 있도록 만들어야 하며 자체로 설명이 가능하여야 한다.
  - 축의 척도를 부적절하게 사용하여 자료를 잘못 해석하게 되는 예

나. 빈도, 빈도 분포, 히스토그램

- ***분석의 첫 단계는 자료를 요약하는 일***
- 정확히 표현되고 그 자체로 자료를 설명할 수 있는 그림으로 도시

1. 빈도: 정성적 자료

- 각 범주별로 해당되는 관찰치의 숫자: 빈도(frequency)
- 상대 빈도(relative frequency): 전체 개체의 숫자 중에서 그 범주가 차지하는 분율
- 빈도 및 상대빈도는 흔히 막대 도표(bar diagram)나 파이 도표(pie chart)로 표현이 가능
- 막대 도표에서 막대의 길이는 빈도에 비례
- 파이 도표에서는 잘려진 파이의 면적이 빈도와 비례

2. 빈도 분포: 정량적 자료

- 빈도 분포: 특정 값 내지는 범위 안에 들어 있는 관찰치의 개수를 정리한 표
- 연속 변수인 경우 특정 범위를 지정
- 관찰치의 총 개수, 최소값, 최대값을 확인
- 자료를 범주화할 것인지, 범주화한다면 ***범주를 자르는 경계점***을 어떻게 정할지를 결정
  - 관찰치의 총 수에 따라 5-20개까지의 범주를 사용
  - ***범주의 시작점은 가능하면 떨어지는 값(round number)으로 지정***
  - 각 범주의 넓이는 동일한 것이 권장
  - ***각 범주의 경계 정도에 해당하는 값이 어느 범주에 포함되는 지를 명확하게 결정***
  - 예를 들어 범주를 8-9, 9-10 같이 표시하는 경우 9라는 측정치가 어느 범주에 포함되는지 혼동
- 표의 형식을 정한 후에는 각 범주별 관찰치의 숫자를 파악

3. 히스토그램

- 빈도 분포를 흔히 히스토그램으로 도시
- 빈도나 분율을 사용하여 도시
- 각 범주의 폭이 다를 경우 주의
- 사각형의 높이는 해당 범주의 관찰치 개수를 구간 폭으로 나눈 값
- 즉 히스토그램에 쓰이는 막대의 면적이 해당 범주의 빈도와 비례

4. 빈도 다각형(fequency polygon)

- 빈도 분포를 도시
- 동일 그림에 두 개 이상의 빈도 분포를 도시하여 이를 비교할 때 유용
- 히스토그램을 가상적으로(아니면 연필로 흐리게) 그려놓고, 각 막대 상단의 중간점을 선으로 연결

5. 모집단의 빈도 분포

- 자료로부터 일반적인 결론을 도출하는데 얼마나 신뢰를 가질 수 있는지는 '얼마나 많은 사람에 대하여 측정치가 얻어졌느냐'에 따라 결정
  - 표본이 클수록, 구간을 세밀하게 범주화 할 수 있을수록
  - 히스토그램 내지는 빈도 다각형의 모양이 완만한 형태를 가지게 되어
  - 좀더 모집단의 분포와 비슷한 모양

6. 빈도 분포의 형태

- 흔한 형태
  - 분포의 중앙에 해당되는 부분에 빈도가 가장 많고
  - 양 극단(분포의 상위 꼬리 내지는 하위 꼬리)의 빈도가 가장 적음

- 중앙에 대하여 대칭적인(symmetrical) 형태
  - 종 모양의 분포
- 비대칭적이고 비뚤어진(skewed) 형태
  - 상위 꼬리가 하위 꼬리보다 긴 형태
    - 양의 방향으로[오른쪽으로] 비뚤어짐: positively[right] skewed
  - 하위 꼬리가 상위 꼬리보다 긴 형태
    - 음의 방향으로[왼쪽으로] 비뚤어짐: negatively[left] skewed
- 최빈값이 두 개인(bimodal) 분포
  - 자료가 두 개의 다른 분포로 구성되어 있는 경우
- 역 J자 형 분포 / 균일한 분포(uniform distribution)

다. 평균, 표준편차, 표준오차

- 정량적 변수의 경우 단지 두 가지의 값만으로도 요약이 가능
  - 평균적인 값을 나타내는 지표 / 변수 값의 산포도를 나타내는 지표

1. 평균, 중앙값, 최빈값

- 산술 평균(arithmetic mean)

$$\text{평균(Mean)},\ \bar{x} = \frac{\sum x}{n}$$

x: 변수 값 / Σ: '합' / n: 관찰치의 개수

- 중앙값(median)
  - 변수의 분포를 반으로 나누는 지점에 해당하는 값
  - 관찰치를 오름차순으로 정리한 경우, 가운데 관찰치에 해당
  - 관찰치의 개수가 짝수인 경우 중앙에 있는 두 개 관찰치의 평균

$$Median = \frac{(n+1)}{2} \text{ th value of ordered observations}$$

- 최빈값(mode)
  - 해당 값의 관찰치가 가장 많은 지점
- 평균은 개개의 관찰치를 감안할 뿐 아니라 수학적, 통계학적 기법에 적용하기 쉬어 선호
- 한 개나 두 개의 관찰치가 극단적으로 높거나 낮게 나오는 경우, 중앙값이 유용
- 분포가 대칭적이고 정점이 하나인 경우 평균, 중앙값과 최빈값은 일반적으로 동일
- 분포가 양의 방향으로 비뚤어진 경우 기하 평균(geometric mean)이 산술 평균보다 유용

cf. inter-quartile range

2. 변이의 지표

- 범위(range): 최대 값과 최저 값의 차이
- 분산(variance): 평균과 각 관찰치 간 차이를 제곱하여 평균을 구한 것
  - 거리의 제곱에 n으로 나누기보다는 (n-1)로 나눔
  - n-1로 나누는 것이 모집단의 분산을 좀 더 정확히 반영

$$Variance, s^2 = \frac{\sum(x-\bar{x})^2}{(n-1)}$$

1) 자유도

- 분산 계산시의 분모인 n-1: 분산의 자유도(degrees of freedom)를 나타내는 숫자
- 각 관찰치의 평균에서의 거리(deviation)인 $x-\bar{x}$ 중 독립적인 값을 가질 수 있는 것은 n개가 아니라 n-1개
- 마지막 하나는 나머지 거리로부터 계산이 되는데, 모두 더하면 0이 된다는 성질이 있기 때문

2) 표준편차(standard deviation)

- 분산의 제곱근

$$s.d., s = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} \quad 또는 \quad s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}}$$

- 흔히 계산기에 σ_{n-1} 로 표시

3) 해석

- 일반적으로 관찰치 중 70%는 평균으로부터 상하 1 표준편차 사이의 구간에 포함
  - 약 95%는 평균에서 상하 2*표준편차의 구간에 포함
  - 이론적인 빈도 분포인 정규분포에 근거하여 얻어진 결과

4) 변이계수(coefficient of variation)

$$c.v. = \frac{s}{x} * 100\%$$

- 표본 평균에 대비한 표준편차의 분율
- 관찰치 값의 상대적인 변이 폭을 보는데 유용 / 관찰치의 단위와 무관

3. 빈도 분포에서 평균과 표준편차의 계산

- 원래 자료는 없고 빈도 분포만 있는 때
  - 각 범주의 중간 값을 이용하여 평균이나 표준편차를 계산

4. 측정 단위의 변화

- 일정한 크기의 상수를 관찰치에 더하거나 빼면 그 크기만큼 평균이 달라지지만 표준편차는 불변
- 상수를 곱하거나 나누는 경우에는 그 크기만큼 평균과 표준편차가 변화

5. 표본 변이와 표준오차

- ***표본은 그 자체로 의미를 갖는 것이 아니라 표본이 대표하는 모집단에 대한 정보를 제공***
- ***표본 평균 $\bar{x}$과 표준편차 s는 모집단의 평균 μ과 표준편차 σ를 추정하는데 활용***
  - ***표본평균이 정확히 모집단 평균과 동일할 가능성은 거의 없음***
  - 표본이 다르면 값이 달라지며, 이는 표본 변이에 기인한다.
- 표본 수가 같은 독립적인 표본을 여러 번 추출하여 각 표본의 평균을 구한 경우 다음과 같이 ***표본평균들의 분포***를 생각할 수 있음
  - 분포의 평균 = 모집단의 평균
  - 분포의 표준편차 = $\frac{\sigma}{\sqrt{n}}$
    - 이를 ***표본 평균의 표준오차***(standard error of the sample mean)
    - ***모집단 평균이 표본 평균에 의하여 얼마나 정확히 추정될수있는지 반영하는 지표***
  - 표준오차의 크기는 모집단 내에서의 변이도 뿐 아니라 표본의 크기에 따라 결정
    - 표본이 클수록 표준오차는 작아짐
  - 모집단의 표준편차인 σ를 알 수 있는 경우는 거의 없기 때문에 표본의 표준편차인 s를 이용하여 표준오차를 추정

$$s.e. = \frac{s}{\sqrt{n}}$$

< 예: 250명 비행기 조종사의 혈압 측정치 >

- 모집단의 평균 μ = 78.2 mmHg / 모집단 표준편차 σ = 9.4 mmHg
- 각 측정치를 250개의 조그만 딱지에 적어서 주머니 안에 집어넣었다. 각 학생들은 주머니를 잘 흔든 후, 10개의 딱지를 골라내어, 그 값을 기록, 그 평균 $\bar{x}$를 구하고 딱지는 다시 주머니에 넣었다. 이런 식으로 30개의 각기 다른 표본이 추출되어, 30개의 각각 다른 표본평균이 구해졌는데 이들 각각은 동일한 모집단 평균을 추정하는데 활용된다.
- 이 표본 평균들의 평균은 78.23 mmHg으로 모집단 평균과 거의 같은 값이다.

- 표본 평균의 표준편차는 3.01 mmHg로 이론적인 표준오차 값인

  $\frac{\sigma}{\sqrt{n}} = \frac{9.4}{\sqrt{10}} = 2.97 mmHg$ 과 거의 같은 값이다.

- 이러한 연습을 반복하였는데 이번에는 표본 수를 20개로 하였다.
  - 표본 평균의 평균은 78.14 mmHg로 모집단 평균과 거의 동일
  - 표준편차는 2.07 mmHg으로 이론적인 값인 $\frac{9.4}{\sqrt{20}} = 2.10\ mmHg$ 과 일치

1) 해석

- (반복적 표본 추출에 따라 얻어지는) ***표본 평균의 약 95%는 모집단 평균을 중심으로 2*표준오차 범위 안에 포함***
- ***위 사실을 이용, 표본 평균과 그 표준오차를 알면 모집단 평균(그 값을 모르는)이 있을 법한 범위를 알아낼 수 있다.***
  - 신뢰구간(confidence interval)

2) 유한모집단에 대한 보정

- 표본이 유한 규모의 모집단에서 추출된 경우
  - 유한모집단에 대한 보정(finite population correction)

$$s.e = \frac{\sigma}{\sqrt{n}} * \sqrt{(1 - \frac{n}{N})}$$

N: 모집단의 개체 수 / n/N: 표본추출 분율(sampling fraction)

  - 유한모집단 보정을 무시한 경우는 표준오차의 값이 과대 평가
  - 표본 추출 분율이 10% 미만인 경우 무시하여도 별 차이가 없음

라. 정규 분포 (Normal Distribution)

- 정규분포곡선
  - 평균을 중심으로 대칭적인 종-모양(bell-shape)
  - 표준편차가 작은 경우 종의 높이가 높아지는 대신 폭이 좁아지며, 큰 경우에는 반대로 높이가 낮아지는 동시에 폭이 넓어지게 된다.
  - 변수 아래 부분(area under the curve)의 면적은 항상 1.00(100%)
- 변수를 전환(transformation), 예를 들어 변수에 로그를 취하는 등, 하면 그 분포가 정규분포에 가까워지는 경우
- 많은 변수를 구체적으로 기술하는데 유용할 뿐 아니라. 통계 분석 기법에서 핵심적인위치

2. 표준정규분포(standard normal distribution)

- 변수가 정규분포를 따르는 경우 측정 단위가 변하여도 정규 분포
- 정규분포를 나타내는 변수의 단위를 적절히 변화함으로써 평균이 0이고 표준편차가 1인 표준정규분포로 전환
  - 각 관찰치에 해당 변수의 평균을 빼준 다음 표준편차로 나누어 줌

$$SND,\ z=\frac{x-\mu}{\sigma}$$

, 단 변수명은 x, x의 평균과 표준편차가 μ, σ

  - *z: 표준정규 편차(standard normal deviate; SND)*
- 표준정규분포표
  - 빈도분포곡선으로 둘러싸인 부분의 면적(area under the curve)
  - 퍼센티지 포인트(percentage points)

3. 정규 분포에서 곡선 아래의 면적에 대한 표

- 인구 집단 중에서 특정 범위 안에 포함된 사람의 분율이 어느 정도인지 결정하는데 유용
- 분포의 상위 꼬리 부분의 면적
- 분포의 하위 꼬리 부분의 면적
- 두 개의 값 안에 포함된 분포의 면적
- 꼬리 부분의 면적으로부터 해당되는 측정치의 역 추정

4. 정규 분포의 퍼센티지 포인트

- 표준정규편차(SND): 특정 변수 값이 평균으로부터 몇 표준편차만큼 떨어져 있는지
- 정확히 분포의 95%를 포함하고 있는 표준정규편차 z(즉 −z와 z사이 면적이 95%)는 1.96
  - 1.96은 정규 분포의 5% 퍼센티지 포인트(percentage point)
  - 2.58는 1% 퍼센티지 포인트
- 양 극단을 포함한(two-sided) 퍼센티지 포인트로서 양 극단의 면적을 포함
  - 일부 교과서에서는 일단의(one-sided) 퍼센티지 포인트를 사용
  - 일단의 a% 퍼센티지 포인트는 양단을 이용한 2a% 퍼센티지 포인트와 동일
  - 1.96은 일단의 2.5% 퍼센티지 포인트인 동시에 양단의 5% 퍼센티지 포인트

마.단일 평균의 신뢰 구간

1. 서론

- 표본평균과 그 표준오차를 이용하여 모집단의 평균이 위치한 값의 범위를 파악하는 방법

2. 표본 수가 큰 경우: 정규 분포 이용

- 반복적 표본 추출에서 얻어지는 표본평균의 분포에서 약 95%의 표본평균은 모집단 평균에서 2×표준편차 사이의 범위에 존재
  - 표본평균의 분포가 정규 분포임을 가정
  - 표본평균 분포의 평균은 모집단의 평균 μ이고 표준편차는 표본 평균의 표준오차인

    $\frac{\sigma}{\sqrt{n}}$

    - 표본수가 큰 경우에, 예를 들어 표본 수가 60을 넘는 경우
    - 표본 평균의 분포는 거의 대부분 정규 분포를 따르며
    - 표본 수가 큰 경우의 표준편차인 s가 모집단 표준편차 σ의 믿을만한 추정치
- 특정한 표본 평균이 모집단 평균에서 1.96 표준오차 거리 안에 있으리라는 사실은 적어도 95% 확률로 정확하게 알 수 있음
- 표본평균이 모집단 평균에서부터 1.96 표준오차 거리에 있을 확률이 95%이기 때문에, 역으로 $\bar{x}-1.96\ s.e$ 과 $\bar{x}+1.96\ s.e$ 사이 구간이 모집단 평균을 포함하고 있을 확률 역시 95%
  - $\bar{x}-1.96\ s.e$ 과 $\bar{x}+1.96\ s.e$ 사이 구간은 모집단 평균이 있을만한 범위를 대표
  - 모집단 평균의 95% 신뢰구간 (confidence interval, c.i.)

$$Large-sample\ 95\%\ c.i.=\bar{x}\pm(1.96*\frac{s}{\sqrt{n}})$$

- 99% 신뢰구간은 $\bar{x}\pm(2.58\times s.e)$

$$Large-sample\ 95\%\ c.i.=\bar{x}\pm(z'*\frac{s}{\sqrt{n}})$$

3. 표본수가 적은 경우

- 표본 수가 크지 않을 때 두 가지 차이
  - 표본의 표준편차, s, 자체가 표본변이의 영향을 받기 때문에 모집단 표본편차 σ의 믿을만한 추정치가 아닐 수 있음
    - s의 표본변이로 인해 정규분포를 사용하여 신뢰구간을 추정하는데 타당성이 없어지는 경우
    - 정규 분포 대신에 t 분포를 사용하여야
      - 엄격하게 말하면, t 분포를 사용하는 것도 모집단이 정규 분포를 따를 때만 타당
        - 그러나 모집단이 극단적으로 비정규적 분포를 보이는 경우를 제외하면 t 분포를 사용하여 큰 문제는 없음: robustness
  - 모집단의 분포가 정규분포를 따르지 않는 경우 표본평균의 분포 자체가 정규분포를 따르지 않을 수도 있음
    - 표본 수가 아주 작으면서 (예를 들어 15 미만), 동시에 모집단의 분포가 정규 분포와 큰 차이를 보이는 경우에만 고려

cf. 중심극한정리(central limit theorem)

변수가 모집단에서 정규 분포를 따르지 않아도 표본평균은 정규분포를 따름

표본수 15 이상인 경우

1) t 분포를 이용한 신뢰구간 추정

- 정규분포를 이용하여 신뢰구간을 구하는 이론적 배경은 $(\bar{x}-\mu)/(\frac{\sigma}{\sqrt{n}})$ 이 표준정규분포를 따르며, 동시에 표본수가 큰 경우 σ 대신 s를 이용할 수 있다는 점
  - 엄밀히 말하면 $(\bar{x}-\mu)/(\frac{s}{\sqrt{n}})$ 은 표준정규분포가 아니라 (n−1)의 자유도를 갖는 t분포
    - 종 모양의 대칭적 분포로 평균이 0이지만, 산포도가 정규분포보다 좀 더 크기 때문에 꼬리가 약간 길
- t 분포의 모양: 표본표준편차인 s의 자유도인 n−1에 따라 결정
  - 자유도가 작을수록 분포의 퍼진 정도가 큼

$$Small\text{-}sample\ c.i. = \bar{x} \pm (t' * \frac{s}{\sqrt{n}})$$

- 자유도가 작은 경우 퍼센티지 포인트는 정규분포에 비하여 상당히 큰 값
  - 이 경우 표본의 표준편차인 s가 모집단 값인 σ를 적절히 반영하지 못하는 지표이므로 이러한 불확실성을 감안하다보면 신뢰구간도 σ가 믿을 만하게 추정된 경우보다 넓어지게 됨
- 자유도가 큰 경우에는 t 분포가 거의 정규분포와 동일한데, 표본표준편차인 s가 모집단 값인 σ의 믿을 만한 추정치가 되기 때문

2) 모집단 분포가 극단적인 비정규 분포일 때

- 측정치의 척도를 전환시켜 전환된 변수의 분포는 정규 분포를 따르도록
- 비모수적인 방법(non-parametric method)으로 신뢰 구간을 구하는 방법

4. 가능한 방법의 요약

Table 5.1 Recommended procedures for constructing a confidence interval

(a) Population standard deviation σ unknown

Sample size	Population distribution	
	Approximately normal	Severely non-normal*
60 or more	$\bar{x} \pm (z' * \frac{s}{\sqrt{n}})$	$\bar{x} \pm (z' * \frac{s}{\sqrt{n}})$
Less than 60	$\bar{x} \pm (t' * \frac{s}{\sqrt{n}})$	Non-parametric

(b) Population standard deviation σ known

Sample size	Population distribution	
	Approximately normal	Severely non-normal*
15 or more	$\bar{x} \pm (z' * \frac{\sigma}{\sqrt{n}})$	$\bar{x} \pm (z' * \frac{\sigma}{\sqrt{n}})$
Less than 15	$\bar{x} \pm (z' * \frac{\sigma}{\sqrt{n}})$	Non-parametric

* It is preferable to transform the scale of measurement to make the distribution more normal.

바. 단일 평균의 유의성 검정

1. 서론

- 표본평균치가 모집단 평균으로 우리가 상정하고 있는(hypothesized) 값과 합치되는지
  - 유의성 검정 (significance testing)

2. 짝지어진 자료의 t 검정

- 각 개인에 대하여 측정된 변수의 짝간(pairwise) 차이가 평균적으로 0인지를 검정
  - 예: 수면제의 효과를 파악하기 위하여 10명 환자의 수면 시간을 약을 준 날 밤과 위약을 준 날 밤, 이틀에 걸쳐 측정
  - 수면제를 복용한 날의 수면 시간이 증가한 것은 두 가지로 해석
    - 우연에 기인한 결과: 약과 위약의 영향이 동일하다는 전제
    - 관찰된 수면 시간의 증가가 실제 약효에 기인

- 두 가지 가능성 중 보다 적절한 설명이 무엇인지를 결정하기 위하여 유의성 검정

- 유의성 검정은 첫 번 째 해석, 즉 우연에 기인한 것이라는 해석이 맞다고 가정하는데서 출발
  - 귀무가설(null hypothesis), 즉 약이나 위약 간에 수면 시간에 실제 차이가 없다는 내용, 을 규정
    - 수면 시간 차이의 실제 평균(μ)이 0이다.
  - 귀무가설이 옳다고 할 때 관찰치인 $\bar{x}$가 우연히 얻어질 가능성이 어느 정도인지 계산
    - 현재 관찰된 수치(수면 시간 차이의 평균)보다 같거나 더 큰 정도의 값이 얻어질 확률
    - 연구 결과의 유의 수준(significance level)
    - 유의성 검정은 쉽게 얘기하면, 그 확률이 매우 작을 경우에만 첫 번째 가능성(즉 우연에 기인한 결과일 가능성)을 배제, 즉 귀무가설을 기각하여 실제 효과가 있다는 두 번째 가능성을 받아들이겠다는 것
- t 분포를 이용, 연구 결과의 유의 수준을 계산
  - 수면시간의 차이가 정규분포를 따른다면 $(\bar{x}-\mu)/(\frac{s}{\sqrt{n}})$ 은 자유도 9인 t 분포를 따름
  - 귀무가설은 수면 시간 차이의 평균이 0, 즉 모집단 평균인 μ가 0이라는 것이므로

    짝지어진 자료의 $t=\frac{\bar{x}}{s/\sqrt{n}}$, $d.f.=n-1$

  - t의 값이 클수록, 결과가 우연에 의하여 나타났을 가능성이 적어짐
    - 관찰된 t 값인 3.18은 2.82와 3.25 사이에 있으므로, 확률은 2%와 1% 사이
      - 수면 시간의 차이는 2% 수준에서 유의(significant at the 2% level)
      - 이 정도의 큰 차이가 우연에 의해서 나타날 확률은 2%가 안 되기 때문
      - $P<0.02$로 표시
    - 이 확률이 매우 작기 때문에 귀무가설은 그럴듯하지 않다(implausible)고 생각하여 그 약이 수면 시간에 영향을 준다고 해석
- 유의수준이 작을수록, 결과는 좀더 유의: 귀무가설이 틀렸다는 증거가 좀더 강화되기 때문
  - 일반적으로 확률이 5% 미만인 경우: 귀무가설이 틀렸다는 타당한(reasonable) 증거
  - 1% 미만인 경우: 강력한(strong) 증거
  - 0.1% 미만인 경우: 매우 강력한 증거(very strong)
- 5%가 넘는 경우: 유의하지 않다고 하며, 그 정도 결과라면 귀무가설과 부합된다고 생각하는 것이 타당하기 때문
  - ***귀무가설이 반드시 옳다는 것은 아니며, 이를 반박할 적절한 증거가 없을 뿐 통계적으로 시사적인 소견, 중간 정도의 유의성 : statistially suggestive, marginally significant***
- 일단 약이 실제 효과가 있다는 사실을 파악한 후 그 약에 의하여 증가하는 수면 시간이 어느 정도인지 제시
  - 예를 들어 95% 신뢰 구간으로 제시

3. 신뢰 구간과 유의성 검정의 관계

- 연구 결과가 특정 유의 수준에서 가설에서 정한 값과 유의한 차이가 있다면
  - 그 유의 수준에 해당하는 신뢰 구간은 귀무가설에서 정한 값을 포함하지 않음
- 반면 연구 결과가 유의하지 않다면 신뢰구간에는 그 값이 포함

4. 단측(one-sided)과 양측(two-sided) 유의성 검정

- 양측 검정: 가설에서 벗어나는 방향이 양, 음 양측으로 가능
  - 귀무가설도 약과 위약 간에 차이가 없다는 것이고 귀무가설이 기각될 때 받아들이는 대립가설(alternative hypothesis)도 두 가지 모두 가능하고 또한 관심을 가지는 것
  - 양쪽 극단의 확률이 유의 수준의 계산에 포함
- 단측 검정: 약이 위약보다 낫거나 최소한 동일한 정도의 수면 시간을 보장한다는 전제
  - 이 경우 약을 복용한 사람의 평균 수면시간이 위약 복용시보다 적다는 결과는 그 감소된 수면 시간이 아무리 크다 하더라도 실제 효과가 아니라, 우연에 의한 것으로 생각
- 대부분의 경우 양측 검정을 사용하는 것이 좋음
  - 단측 검정은 유의한 결과를 얻을 가능성이 높아지기 때문에 끌리게 되긴 하지만, 같은 이유로 귀무가설을 기각하고 실제 효과가 있다고 잘못 결론 내릴 가능성 또한 증가
  - 단측 검정은 그 사용에 충분한 이유가 있어야 하며, 자료를 수집하기 전에 단측 검정 수행 여부를 결정하여야 하고, 연구 결과를 제시할 때도 이 점을 명확히 함

5. 단일 표본의 t 검정

- 짝지어진 자료의 t 검정
  - 표본 평균이 특정한 값 μ (항상 0이어야 할 필요는 없다.)와 다른 지를 검정하기 위한 단일 표본 t 검정의 특수한 예
- 단일 표본 t 검정의 일반적인 식

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}},\ d.f.=n-1$$

6. 정규 검정(Normal test)

- 표본수가 큰 경우(60을 넘는 경우)의 평균치 검정
- 모집단 표준편차를 아는 경우의 평균치 검정

표본 수가 큰 경우

$$z=\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

σ를 아는 경우

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

7. 유의성 검정에서의 오류의 형태

- 연구 결과가 우연에 의하여 나타난 것인지 실제 효과나 차이에 의해 나타난 것인지 판단하는 유의성 검정에서는, 두 가지 가능성 중 무엇이 진실인지 증명해 내지 못하는, 두 종류의 오류가 발생
- 귀무가설은 실제로 옳음에도 불구하고 기각될 수 있으며, 반대로 실제로 틀림에도 기각하지 못하는 경우
  - 제 1종 오류(type Ⅰ error) / 제 2종 오류(type Ⅱ error)
- 유의성 검정의 유의 수준
  - 귀무가설이 옳은 경우에, 얻어진 연구 결과 내지는 그보다 극단적인 연구 결과가 얻어질 확률
  - 제 1종 오류를 범할 확률, 즉 귀무가설을 기각할 확률은 따라서 그 검정의 유의 수준과 동일
  - 예를 들어 5% 유의 수준에서 유의한 연구 결과가 단지 표본 변이만으로도 나타날 확률은 5%이므로, 그 결과로부터 귀무가설을 기각한다면, 오류의 확률이 5%
- 제 2종 오류: 실제 귀무가설이 틀림에도 불구하고 기각하지 못하는 경우
  - 실제 모집단 평균(귀무가설에서 상정한 μ가 아닌 다른 값 μ′)을 중심으로 한 표본 평균의 분포와, μ를 중심으로 한 가상적 표본 평균 분포 간에 겹치는 부분이 있기 때문에 발생
  - 겹치는 부분은 실제 표본평균의 분포 중에서 귀무가설을 그대로 받아들이는 지역(귀무가설에 부합되는 것으로 판단하는 지점, 즉 5% 유의 수준)과 겹치는 부분 (b%)
  - 유의 수준을 낮게 설정한 경우 제 1종 오류의 가능성은 감소하지만, 빗금친 부분의 면적은 증가하여 제 2종 오류가 커진다. 역도 마찬가지로 성립
  - 제 2종 오류를 범하지 않을 가능성(100−b%): 검정력(power of the test)
    - 표본 수가 커질수록 검정력이 커지는데, 분포 곡선이 높이는 높아지고 폭은 좁아져서 빗금친 부분의 면적이 감소

사. 두 평균치의 비교

1. 두 평균간 차이의 표본분포

- 두 독립된 표본의 평균 차이인 $\overline{x_1}-\overline{x_2}$ 은 각 평균의 분포가 정규 분포를 따르는 경우 정규 분포를 따름
  - 이 분포의 평균은 두 모집단 평균의 차이인 $\mu_1-\mu_2$
  - 유의성 검정에서의 귀무가설은 두 평균이 동일하다는 것으로 $\mu_1-\mu_2=0$
  - 표준오차는 각 평균의 표준오차를 이용하여 계산

$$s.e.=\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}$$

이는 표본의 표준편차인 s_1, s_2로부터 추정

2. 정규 검정

● 두 표본 수가 모두 크거나 모집단의 표준편차(σ)를 아는 드문 경우

표본 수가 큰 경우

$$z=\frac{x_1-x_2}{\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}}$$

또는

σ들을 아는 경우

$$z=\frac{x_1-x_2}{\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}}$$

● 평균 차이의 신뢰구간

표본 수가 큰 경우

$$c.i.=(\overline{x_1}-\overline{x_2})\pm(z'*s.e.)$$

$$s.e.=\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}$$

또는

σ 들을 아는 경우

$$c.i.=(\overline{x_1}-\overline{x_2})\pm(z'*s.e.)$$

$$s.e.=\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}$$

3. t 검정: 표본 수가 적고, 표준편차가 같은 경우

● 모집단의 분포는 정규분포를 따라야 한다. 단, 그 가정이 단일 표본의 경우에서와 같이 반드시 엄밀하게 지켜져야 하는 것은 아님

● 두 평균 차이의 표준오차

$$s.e.=\sqrt{(\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2})} \text{ 또는 } \sigma\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$$

σ: 공통된 표준편차

● 두 개의 표본에서 구해진 σ의 추정치 s_1, s_2을 이용하여 모집단 표준편차의 추정치 s를 추정: s의 자유도는 $(n_1-1)+(n_2-1)=n_1+n_2-2$

$$s=\sqrt{[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}]}$$

● 표본수가 많은 표본에서 얻어진 추정치인 경우 신뢰도가 높아지며, 가중치를 많이 부여

● 두 평균 간 차이의 표준오차

$$s.e.=s\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$$

● t 값의 계산

$$t=\frac{\overline{x_1}-\overline{x_2}}{s\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}},\ d.f.=n_1+n_2-2$$

● 평균 차이의 신뢰구간(confidence interval)

$$c.i.=(\overline{x_1}-\overline{x_2})\pm(t'*s.e.),\ s.e.=s\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$$

4. 표본수가 적고 표준편차가 각기 다를 때

● 두 집단의 모집단 표준편차가 다를 경우
  ● 변수의 척도를 변화시켜 t 검정이 사용 가능하게 조정
  ● 예를 들어 표준편차가 그 절대값에 있어 평균치에 비례하는 경향이 있다면 각 관찰치에 로그를 취하여 해결이 가능
  ● 대안으로 비모수적 방법, Fisher-Behrens 내지는 Welch 검정

아. 세 개 이상 평균치의 비교—분산분석

1. 서론

● 집단이 세 개 이상인 경우 구성 집단 개개의 평균을 비교
  ● t 검정을 여러 번 시행?
  ● 번거로울 뿐 아니라, 유의성 검정을 여러 번 시행할 경우 엉뚱하게 유의한 결과
    ● 실제적으로는 차이가 없는 경우라도, 20번 검정하면 1번 정도에서는 '5% 유의 수준에서 유의하다'는 분석 결과
● 일원 분산분석
  ● 비교하고자 하는 집단이 하나의 요인(factor)만으로 정의되는 경우
  ● 예를 들어 사회경제적 수준에 따른 비교, 인종간의 비교 등
● 이원 분산분석(two-way analysis of variance): 집단이 두 가지 요인으로 구분되는 경우
● 분산분석에 포함되는 요인
  ● 그 값에 따라 구분되는 집단간의 평균을 비교하는 경우
  ● 그 요인이 측정치의 변이원(source of variation)으로서 반드시 고려하여야 할 경우
● 하나 이상의 변수에 의하여 구분되는 자료를 분석하기 위해서는 분산분석과 밀접하게 연관된 기법으로서 다변량 회귀분석(multiple regression)을 사용
  ● 회귀분석이 보다 일반화된 방법
  ● 동일한 분석결과가 얻어지지만, 회귀분석이 좀더 일반화된, 즉 광범위한상황에서 사용

2. 일원 분산 분석(one-way analysis of variance)

- 자료가 전체적으로 갖는 변화도 중 집단간 평균 차이에 기인한 부분을 동일 집단내에서의 개인 차이에 기인한 부분과 비교
- 각 개인이 어느 집단에 소속되었는지를 무시하고, 전체 관찰치의 분산을 계산
  - 분산은 표준편차의 제곱으로, 각 관찰치를 전체 평균(overall mean)에서 뺀 값을 제곱하여 모두 더해준 다음, 자유도로 나눈 값
- 일원 분산분석에서는 이 제곱합(sum of squares, SS)을 두 개의 구성 요소로 분리

  (i) 집단간 평균 차이에 기인한 제곱합

  (ii) 집단내에서 각 관찰치간 차이에 기인한 제곱합: 잔차 제곱합 (residual sum of squares)
- 전체 자유도도 비슷하게 분할
- 자유도당 변이의 정도: 평균 제곱합(mean square, MS)
- 집단간 차이를 보기 위한 유의성 검정은 군간 및 군내 평균 제곱합에 의하여 수행
  - 군간 평균의 차이가 우연에 의해서도 나올 수 있는 정도라면 군간 평균간의 변이의 양이나, 각 군내 개인간의 변이의 양이나 거의 같은 정도
  - 군간에 차이가 있다면, 군간 변이가 상대적으로 클 것
- 평균 제곱합을 F 검정을 이용하여 비교: 분산 비 검정(variance-ratio test)

$$F = \frac{Between\text{-}groups\ MS}{Within\text{-}groups\ MS},\ d.f. = d.f._{Between\text{-}groups},\ d.f._{Within\text{-}groups} = (k-1,\ N-k)$$

N: 총 관찰치의 숫자, k: 비교 집단의 숫자

- 집단간 차이가 없는 경우 F는 약 1이 되며, 증가할수록 그 차이가 있음을 시사
- 집단간 차이가 단순히 우연 때문이라는 귀무가설하에서, 그 비(ratio)는 F 분포
  - 자유도는 분자의 자유도 k−1, 분모의 자유도 (N−k)

1) 가정

- 자료가 정규 분포를 따라야 함
- 각 집단이 추출된 모집단의 표준편차가 모두 동일하여야 함
- 정규분포에서 약간 벗어나는 정도는 무방하나, 표준편차가 다른 경우 결과에 심각한 오류

2) 두 표본의 t 검정과의 관계

- 두 표본 t 검정의 확장

3. 이원 분산분석(two-way analysis of variance)

- 자료가 두 가지 변수로 분류되는 경우
  - 예를 들어 연령 및 성별로 집단을 구분하는 경우
  - 각 집단마다 관찰치의 숫자가 동일한 경우: 균형 설계(balanced design)
    - 한 군당 2개 이상의 관찰치가 있는 경우: 반복이 있는 경우
    - 한 군당 1개의 관찰치만 있는 경우: 반복이 없는 경우
  - 다른 경우: 비균형 설계(unbalanced design)

4. 반복이 있는 균형 설계(Balanced design with replication)

- 총 3종 (strain)의 실험용 쥐에 대하여, 종별로 각각 다섯 마리의 남성 및, 다섯 마리 여성 쥐 성장 호르몬으로 처치한 실험 결과
  - 연구 목적은 각 종별로 처치에 대한 반응이 동일한지, 성별 차이가 있는지를 파악
  - 측정한 반응 지표는 처치 7일 후의 체중 증가
- 이원 분산 분석에 의하여 총 제곱합을 4개의 구성 요소로 분할

(i) 종간의 차이에 기인한 제곱합: 종의 주된 효과 / 자유도: 종의 숫자에 1을 뺀 값 = 2

(ii) 성별 차이에 의한 제곱합: 성별의 주된 효과 / 자유도: 성별이 2가지이므로 2−1=1

(iii) 종과 성별의 상호 작용(interaction)에 의한 제곱합

상호 작용은 종간의 차이가 성별에 따라 동일하지 않다는 의미

또는 성별 차이가 종별로 다르다는 의미

자유도: 두 변수의 자유도를 곱한 값 2×1=2

(iv) 잔차 제곱합: 각 종-성별 군내에서 개인별 차이에 기인하는 값

자유도: 24, 종의 숫자(3), 성별의 종류(2가지)를 곱하고 여기에 각 군당 관찰치−1(4)을 곱한 값

- 주된 효과와 상호 작용은 F 검정을 이용하여 유의성을 검정

5. 반복이 없는 균형 설계

- 재태 기간을 추정하는 5가지의 방법을 10명의 산모에 적용
  - 각 군당(즉 각 추정 방법 당) 하나씩의 관찰치밖에 없기 때문에 잔차 제곱합이 계산되지 않음
- 상호작용이 우연에 기인한 부분인 것으로 가정하여 주된 효과의 영향을 평가하는 F 값을 계산할 때 상호 작용의 평균 제곱합을 잔차 제곱합으로 간주

1) 제곱합을 좀더 세분할 필요성

- 예제에서 관찰된 유의한 차이를 좀 더 자세히 검토

quickening 방법에 의한 재태 기간은 평균적으로 다른 방법에 비하여 상당히 높은 경향

- 추정 방법의 주된 효과에 기인한 제곱합을 다음과 같이 세분

(i) quickening 방법으로 추정한 재태 기간과 다른 방법으로 추정한 재태 기간간의 차이에 기인한 제곱합으로 자유도 1

(ii) 다른 4 가지 방법간의 차이에 기인한 제곱합으로 자유도 3

- 각각의 구성 요소는 통상적인 방법으로 F 검정
- 주의할 점: 추정 방법에 기인한 제곱합은 이외에도 다른 방식으로 세분화할 수 있다.
  - 해당 제곱합의 자유도만큼 세분화 할 수 있으며 이 경우 4개로 세분화가 가능
  - 연구자가 차이를 파악하고자 하는 부분이 무엇인가에 따라 세분화 방법이 결정
    - 자료 분석 전에 충분한 근거에 의하여 결정
    - 세분화 작업은 선형 대비(linear contrast) 방법에 의하여 수행

2) 단일 표본 t 검정과의 관계

- 반복이 없는 균형 설계의 이원 분산분석은 단일 표본의 짝지어진 t 검정의 확장
- 한 개인에 대하여 3번 이상 측정된 값들을 비교하는 방법
- 이 예제에서 각 산모에 대하여 5 가지 다른 방법에 의하여 재태 기간이 추정되었으므로 분석할 변수가 개인당 다섯 개
  - 두 개의 변수만 측정하였을 경우는 두 가지 방법의 결과가 동일하며, 구해진 F 값은 t 값의 제곱

6. 비균형 설계

- 구충 감염과 혈색소치를 조사한 자료
  - 두 요인, 즉 성별과 구충 감염 정도에 따라 분류
  - 각 성별로, 혈색소치는 구충 감염 정도가 심할수록 감소하는 경향
  - 각 감염 정도별로, 여성의 평균 혈색소치가 남자보다 낮은 경향
  - 비균형 설계로, 각 군마다 대상자 수가 각기 다르므로 성별 및 구충 감염 정도에 따른 효과의 분리가 어려워지며, 그 결과로 연구 자료의 해석이 어려움
- 총 제곱합은 두 요인에 기인한 독립적인 제곱합으로 분리될 수 없음
  - 분산분석표를 구하는 과정이 약간 다름
  - 우선 성별에 따른 차이에 기인하는 제곱합을 계산
  - 구충 감염 정도에 기인한 추가적인 제곱합을 계산
    - 이 제곱합은 구충 감염 군 간의 성별 차이를 보정한 후, '혈색소 치와 구충 감염 정도와의 관련성'을 평가하는데 활용
- 분석하는 방법으로서 이전과 달리 먼저 구충 감염 정도의 제곱합을 계산한 후, 이를 보정한 성별의 제곱합을 구하는 방법도 가능
  - 이 때도 구충 감염 정도의 제곱합에는 성별의 차이에 의한 영향이 일부 반영
  - 성별에 따라 구충 감염 정도가 다른 것이 이미 보정된 후, 성별이 혈색소 치에 미치는 영향을 평가
- 비균형 설계의 경우 두 가지 분석을 모두 시행하는 것이 바람직하지만, 이 예제에서는 논리적으로 볼 때 성별의 영향을 먼저 파악하는 것이 적절

7. 고정과 무작위 효과(Fixed and random effects)

- 변수나 요인의 두 가지 형태
  - 고정 효과와 무작위 효과
- 성별, 연령과 같은 변수는 고정 효과로서 해당 변수의 각 수준(levels)은 모두 특정한 고정된 값을 갖는다. 예를 들어 성별은 항상 남자 아니면 여자이다. 그러나 무작위 효과의 각 관찰치들은 그 자체로는 관심의 대상이 되지 않는 성격을 가지며, 각 수준은 단지 그 변수가 가질 수 있는 변이원을 대표하는 수준(levels)으로 구성된 하나의 표본이라 할 수 있다. 예를 들어 가정에서 만든 경구 수액제의 나트륨과 자당(sucrose) 농도 변이를 보는 연구를 한다고 하자. 10명에게 각각 8개의 수액제를 만들라고 하였다면, 이 때 10명이라는 사람은 단지 각기 다른 사람에 의하여 만들어진 수액간의 변이를 대표한다는 점에서만 의미를 가질 수 있다. 따라서 여기서 사람이라는 변수는 무작위 변수이다. 이 예에서 검정하고자 하는 것은 개인이라는 변수의 유의성 뿐만 아니라, '동일한 사람이 만든 수액에서, 또는 다른 사람이 만든 수액간의 농도 변이의 크기가 어느 정도인가' 하는 것이다. 이러한 것들은 변이의 구성 요소(components of variation)라 한다.
- 일원 설계나, 반복 없는 이원 설계에서는 고정 효과나 무작위 효과나 유의성 검정 방법이 동일
- 반복이 있는 이원 설계나 그 이상의 설계에서는 동일하지 않음
  - 두 효과가 모두 고정 효과인 경우, 각각의 평균 제곱합을 잔차의 평균 제곱과 비교
  - 모두 무작위 효과인 경우에는 효과의 평균 제곱합을 잔차와 비교하는 것이 아니고 상호 작용의 평균 제곱합과 비교
  - 하나는 고정, 하나는 무작위 효과인 경우에는 무작위 효과는 잔차와 비교, 고정 효과의 평균 제곱합은 상호 작용과 비교

자.상관 및 선형 회귀 분석

1. 서론

- 2개 이상의 연속변수간의 상호 관계를 파악
- 관련성의 밀접성을 평가하는 지표가 상관계수인 반면, 선형회귀분석은 그 관련성을 표현하는데 가장 적합한 직선의 방정식을 도출하며 동시에 한 변수로부터 다른 한 변수의 값을 예측(prediction) 할 수 있도록 하는 것이 주된 목적

2. 상관분석

- 산포 도표(scatter diagram)
  - 관련성을 상관계수(correlation coefficient) r로 정량화

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2\sum(y-\bar{y})^2]}}$$

- 상관 계수
  - 항상 -1에서 1까지의 숫자, 변수간의 관련성을 전혀 없는 경우 0
- 유의성 검정
  - t 검정을 이용, r이 유의하게 0과 다른지, 즉 관찰된 상관계수가 우연에 기인하였을 가능성을 검정

$$t = r\sqrt{[\frac{n-2}{1-r^2}]},\ d.f. = n-2$$

  - 유의수준은 상관계수의 값뿐 아니라 관찰치의 숫자에 따라서도 결정
    - 표본수가 많은 경우에는 상관관계가 약함에도 불구하고 통계적으로 유의한 결과가 나오게 되며, 반면 표본 수가 매우 적은 경우 강한 상관 관계가 있어도 유의하지 않게 검정될 수 있다.

3. 선형 회귀 분석

- x 변수의 증가에 따른 y 변수의 증가(또는 감소) 양상을 나타내는 직선의 식을 도출
  - 두 변수 중 어느 것을 y로 정할 것인지가 중요
- 회귀 직선의 식

$$y = a + bx$$

a: 절편(intercept), b: 직선의 기울기(slope)

  - a, b의 값은 직선에서부터 각 점에의 수직 거리를 제곱하고 이를 모두 더한 값(sum of the squared vertical distances)이 가장 작아지도록 하여 추정
    : 방법을 최소제곱에 의한 추정(least squares fit)
  - 경사도 b: 회귀 계수(regression coefficient)
    - 그 부호는 상관 계수와 동일하며, 상관 관계가 전혀 없으면 b도 0

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \qquad \text{그리고} \qquad a = \bar{y} - b\bar{x}$$

  - 추정된 a, b의 값은 모집단에서 x와 y와의 선형적 관련성을 평가하는 회귀 직선의 절편과 경사도 참값 (모집단 값)에 대한 표본 추정치
    - 표본 변이가 발생하며 그 정밀성(precision)은 그 표준오차로서 정량화

$$s.e.(a) = s\sqrt{[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}]} \text{ and } s.e.(b) = \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$$

$$\text{단, } s = \sqrt{[\frac{\sum(y-\bar{y})^2 - b^2\sum(x-\bar{x})^2}{(n-2)}]}$$

s: 직선으로부터 각 점들이 가지는 표준 편차 / s의 자유도: n−2

1) 유의성 검정

- t 검정을 이용, b가 유의하게 특정 값, β와 다르다는 검정

$$t=\frac{b-\beta}{s.e.(b)},\ d.f.=n-2$$

- 특히 b가 유의하게 0이 다른지를 검정하기 위하여 활용
  - 그 결과는 r=0에 대한 t 검정과 동일

2) 예측(Prediction)

- 특정 x값(x′이라 하자.)에서의 y값을 예측하는데 사용

$$y'=a+bx'$$

- 그 예측치는

$$s.e.(y')=s\sqrt{[1+\frac{1}{n}+\frac{(x'-\bar{x})^2}{\sum(x-\bar{x})^2}]}$$

- 예측치의 표준 오차는
- y′의 표준 오차는 x′가 평균에 가까울수록 그 값이 작아진다.

3) 가정

- 모든 x 값에 대하여 y는 정규 분포
- 직선 상에서 관찰치들이 보이는 산포도는 회귀직선 전 구간에 걸쳐 동일

차.다변량 회귀분석

1. 서론

- 독립 변수가 여러 개 있을 때, 종속 변수의 의존성을 파악하여야 하는 경우
  - 독립 변수간의 상호 관련성이 있을 수 있다는 점을 감안하여 변수들의 종합적인 영향(종속 변수에 미치는) (joint influence)을 평가

2. 단변수 회귀분석에서의 분산분석적 접근 방법

- 산포 도표로 두 변수의 관계를 파악
- a, b의 추정
  - 회귀 직선에서 관찰치까지의 편차를 제곱하여 합한 값이 가장 작아지도록 하여 추정
- 회귀 직선을 기준으로 계산한 편차 제곱합(sum of squared deviations)이 최소화되었을 때, 그 값이 잔차 제곱합(residual sum of squares)
  - 자유도=표본수-2

- 전체 변이 중에서 잔차 제곱합을 제외한 부분은 회귀 식에 의하여 설명되는 제곱합(sum of squares explained by the regression)
  - 자유도=1
- 변수간 관련성이 없다면, 회귀에 의한 평균 제곱합은 잔차에 의한 평균 제곱합의 값이나 거의 같을 것이고, 관련성이 있다면 전자의 값이 더 클 것이다. 따라서 F 검정으로 변수간의 관련성을 평가

$$F=\frac{regression\ mean\ square}{residual\ mean\ square},\ d.f.=(1,\ n-2)$$

3. 상관계수와 분산분석표의 관계

- 상관 계수의 제곱 r^2는 회귀에 의한 제곱합을 총 제곱합으로 나눈 값
  - 총 변이 중에 회귀에 의하여 설명되는 부분

4. 독립변수가 두 개 있는 경우의 다변량 회귀분석

$$y=a+b_1x_1+b_2x_2$$

- 예) 출생 체중$=a+b_1$산모의 키$+b_2$임신기간
  임신 기간이 어떠하든지 간에, 출생 체중은 산모의 키와 선형적 관련성이 있으며, 또한 산모의 키가 어떠하든지 간에 출생 체중은 임신 기간과 선형적 관련성이 있다.
  b_1, b_2: 부분 회귀계수(partial correlation coefficient),
  이에 상응하는 상관성이 부분상관성(partial correlations)
  - F 검정: 회귀부분의 자유도가 2
  - 회귀에 의하여 설명되는 변이: R^2
  - $R=\sqrt{0.2026}=0.45$: 다변량 상관 계수(multiple correlation coefficient)
    - 항상 양수인데, 독립 변수가 두 개 이상인 경우 상관 계수의 방향성을 부여할 수 없기 때문
- 변수에 의한 회귀로 설명되는 제곱합은 단변수 회귀 분석에서 추정된 '임신 기간으로 설명되는 제곱합'에 임신 기간을 감안한 후 산모의 키에 기인한 추가적인 제곱합으로 구성

5. 독립변수가 세 개 이상인 다변량 회귀분석

- 독립변수의 숫자는 적절한 정도로 줄이는 것이 권장
- 회귀 식에 포함될 독립 변수의 선택

1) 단계를 높이는 회귀분석(step-up regression)

- 각 독립변수에 대하여 단변수 회귀분석
- 종속변수 변이의 설명력이 가장 큰 변수를 골라서 첫 번째 변수로 한 후, 남은 독립변수를 각각 선정하여 독립 변수가 2개인 회귀 분석을 수행
- 변수가 두 개인 회귀식 중 변이의 설명력이 가장 큰 것을 선택
- 이상의 과정을 반복하여 각 단계마다 변수를 하나씩 추가
- 남은 변수 중 어느 것을 추가하여도 추가 변수에 따른 설명력의 증가가 유의하지 않거나, 분석 전에 정한 회귀 식의 최대 독립 변수 수에 도달할 경우 중지

2) 단계를 낮추는 방법(step-down regression)

- 독립변수를 포함한 다변량 회귀분석
- 한 번에 변수 하나씩 제외
  - 각 단계에서 떨어뜨리는 변수는 설명력에 미치는 영향이 가장 낮은 변수로 선택
- 이 과정은 회귀 식에 남아 있는 모든 변수가 통계적으로 유의하거나, 그 숫자가 너무 많은 경우에는 사전에 정한 최대 변수 수에 달할 때까지 계속

3) 최적 조합 회귀 분석 (optimal combination regression)

- 위의 계단식 방법(stepwise method)은, 분석 종료시 남아 있는 변수의 숫자가 같을지라도, 남아 있는 변수의 내용이 다를 수 있다. 따라서 어느 방법도 주어진 변수의 숫자에서 최적의 회귀식을 도출한다고 자신 있게 말할 수는 없다. 따라서 변수가 하나일 때, 두개일 때, 세 개일 때 등등 각각의 변수 숫자에 따라, 최적의 변수 조합을 선정하는 방법이 선호된다. 그러나 예를 들어 변수 수가 두 개일 때 최적인 두 개의 변수 중에는 하나일 때 최적인 변수가 포함되지 않는 경우도 발생한다는 점을 기억하여야 한다.

6. 값이 비연속적인(discrete) 독립변수의 다변량 회귀 분석

- 비연속변수(factor)는 가변수(dummy variable)를 정의하여 회귀식에 포함
- 값 (level)이 3개 이상인 경우, 예를 들어 연령군과 같이, 에는 가변수를 여러 개 사용
  - 각 가변수는 변수의 값 (level)간의 차이를 지칭
  - 한 요인이 k개의 값을 가지고 있다면 가변수는 k−1개 정의하여야 하며, 해당 요인의 자유도는 k−1

7. 비직선적인 독립변수일 때의 다변량 회귀 분석

- 변수를 몇 개의 아군으로 재정의하여 각 아군당 하나씩의 값을 가지는 요인(factor)으로서 회귀식에 포함
  - 분석 초기 단계에서는 독립변수를 연속적인 형태와 요인화한 형태로 모두 분석
    - 두 가지 방법에서 얻어진 제곱합의 차이를 이용하여, 해당 변수와 종속변수 간의

관계가 비선형적인지를 검토

- 대부분의 경우 연속 변수를 표본 수에 따라 3~5개 정도로 나누면, 관련성의 비선형 여부를 파악하는데 충분

- 독립 변수를 적절히 전환
- 연관성에 대한 적절한 수학적 기술을 찾아내는 것으로 예를 들어 관계가 정방형 (quadratic) 이라면, 회귀 식에 변수 자체 (x) 외에도 그 제곱 (x^2)을 포함

8. 다변량 회귀분석과 분산분석의 관계

- 독립변수가 모두 비연속적인 다변량 회귀 분석은 요인이 여러 개 있는 분산 분석과 동일
- 공분산분석(analysis of covariance)
  - 군간 차이를 볼 때 비연속 변수 외에도 연속 변수가 포함된 경우 사용

9. 다변량 분석(multivariate analysis)

- 두 개 이상의 종속 변수가 동시에 어떻게 변화하는지를 파악하는 기법
- 주성분 분석(principal component analysis)
  - 성분(component)이라 명명하는 변수의 몇 가지 조합들이 전체적인 변이를 적절히 설명하도록 하여 자료를 단순화하는 기법
- 판별 분석(discriminant analysis)
  - 각 집단을 가장 잘 구별해 낼 수 있는 변수들의 조합(판별 함수, discriminant function)을 찾아내는 기법
  - 판별 함수를 이용하면, 연구 대상 이외의 새로운 개체에 대하여 어느 집단에 소속될 가능성이 높은지 예측
- 요인 분석(factor analysis)
  - 심리학적 검사 방법을 분석하는데 흔히 사용
  - 여러 가지 검사 항목에 대하여 보이는 반응이 내재하고 있는 요인, 예를 들어 감정(emotion)이나 이성적 추론 등, 에 의하여 영향을 받는지 알아 낼 수 있다.
- 군집 분석(cluster analysis)
  - 여러 가지 변수의 측정치를 이용하여 각 개인들이 자연스럽게 체계적인 집단을 이루게 되는지를 파악하는 기법