

(그림으로 이해하는) 닥터 배의 술술 보건의학통계

Part_2 중급 보건의학통계 맛보기
라 가 영



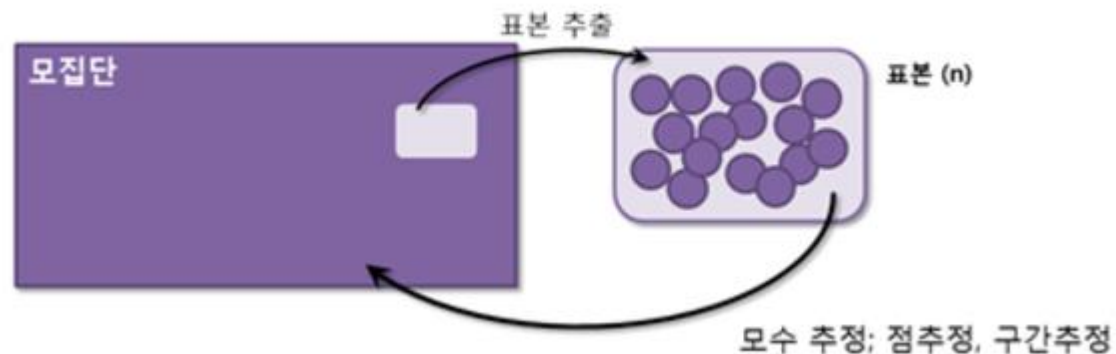
08.
이 여덟 가지만
더 알고
중급 통계에
들어가자

- 가. 95% 신뢰구간
- 나. 상대위험도와 교차비
- 다. 지수함수
- 라. 일반화 선형모형
- 마. 우도
- 바. 교란변수와 교호작용
- 사. 가변수의 설정
- 아. 단변수 분석과 다변수 분석

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 점 추정(point estimation) : 모집단으로부터 표본을 추출하고, 추출된 표본의 통계량(표본 평균, 표본 표준편차)을 통해 모수(모 평균, 모 표준편차)를 추정
- 구간추정(interval estimation) : 모평균이 존재할 구간을 확률적으로 추정
- 신뢰구간(confidence interval; CI) : 대표적인 구간추정. 이 구간 내에 실제로 모수가 존재할 것으로 예측 되는 구간. 90%, 95%, 99%... 보건의학 통계 분야에서 가장 널리 사용되고 있는 것은 95% 신뢰구간



08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 모평균에 대한 신뢰구간

- 연속형 자료의 경우 표본의 크기(표본 수)가 충분히 크다면 중심극한정리에 의해 표본 평균이 정규분포를 따른다고 가정할 수 있음
- 정규분포라는 가정 하에 표본 평균과 표본 표준편차로부터 모평균의 신뢰 구간을 구할 수 있음

$$\text{모평균의 95\% 신뢰구간} = \bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}}$$

(표본 평균 \bar{X} , 표본 표준편차 s , 표본의 크기 n)

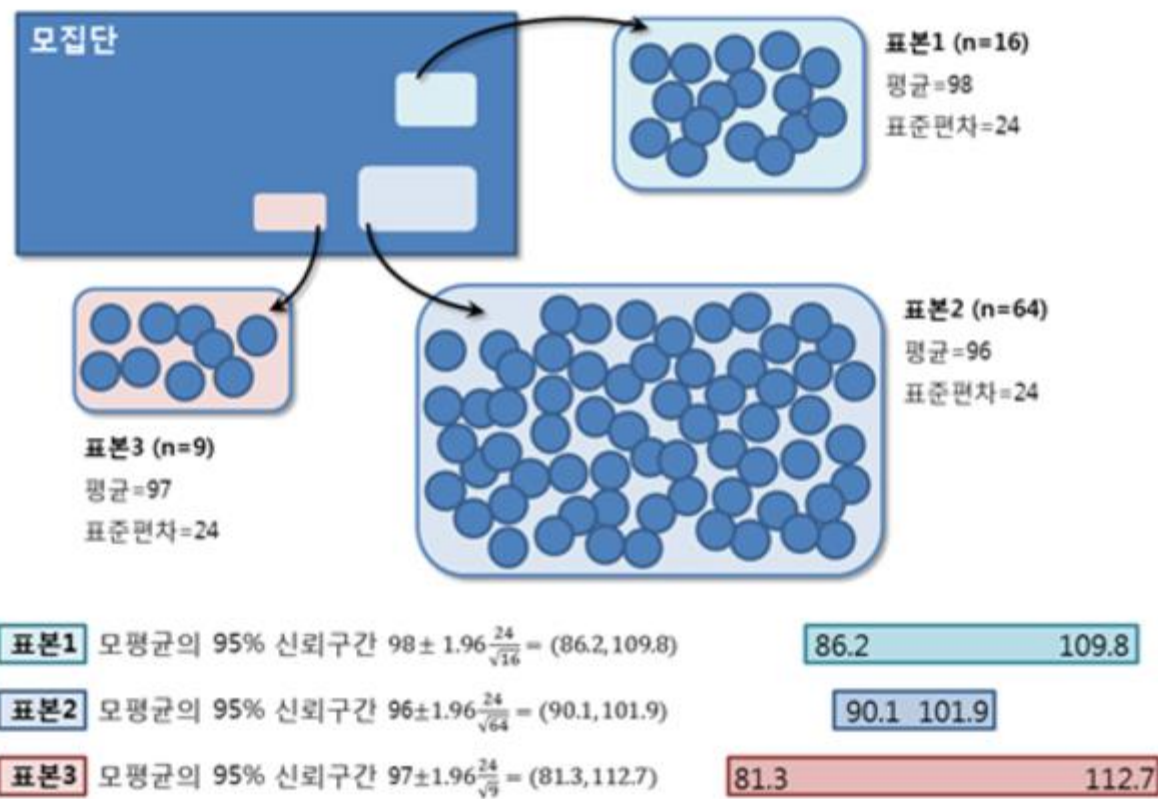
신뢰수준 C 에 따른 임계값 z^*

90% - 1.645, 95% - 1.960, 99% - 2.576

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 모평균에 대한 신뢰구간



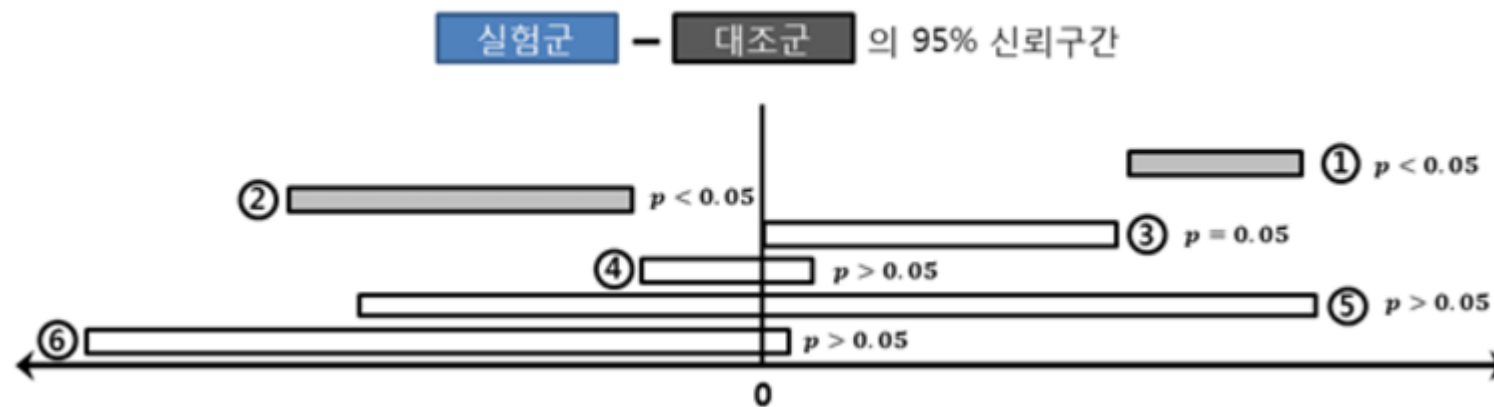
표본들의 평균과 표준편차가 비슷하다면, 95% 신뢰구간의 폭은 표본 수에 영향을 받음

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 모평균에 대한 신뢰구간

- 신뢰 구간은 T 검정에도 이용
- 4, 5, 6은 모두 95% 신뢰구간이 0을 포함하고 있으므로 두 군이 통계적으로 유의한 차이가 있다고 말할 수 없음($p > 0.05$)



08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 모비율에 대한 신뢰구간

- 비율로 측정되는 자료의 경우 표본수가 충분히 크다면 모비율의 신뢰구간을 추정 가능
- 모비율의 95% 신뢰구간은 실제 모비율이 이 구간 내에 있을 가능성이 95%라는 의미

$$\text{모비율의 95\% 신뢰구간} = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

(표본의 관심 사건의 비율 p , 표본의 크기 n)

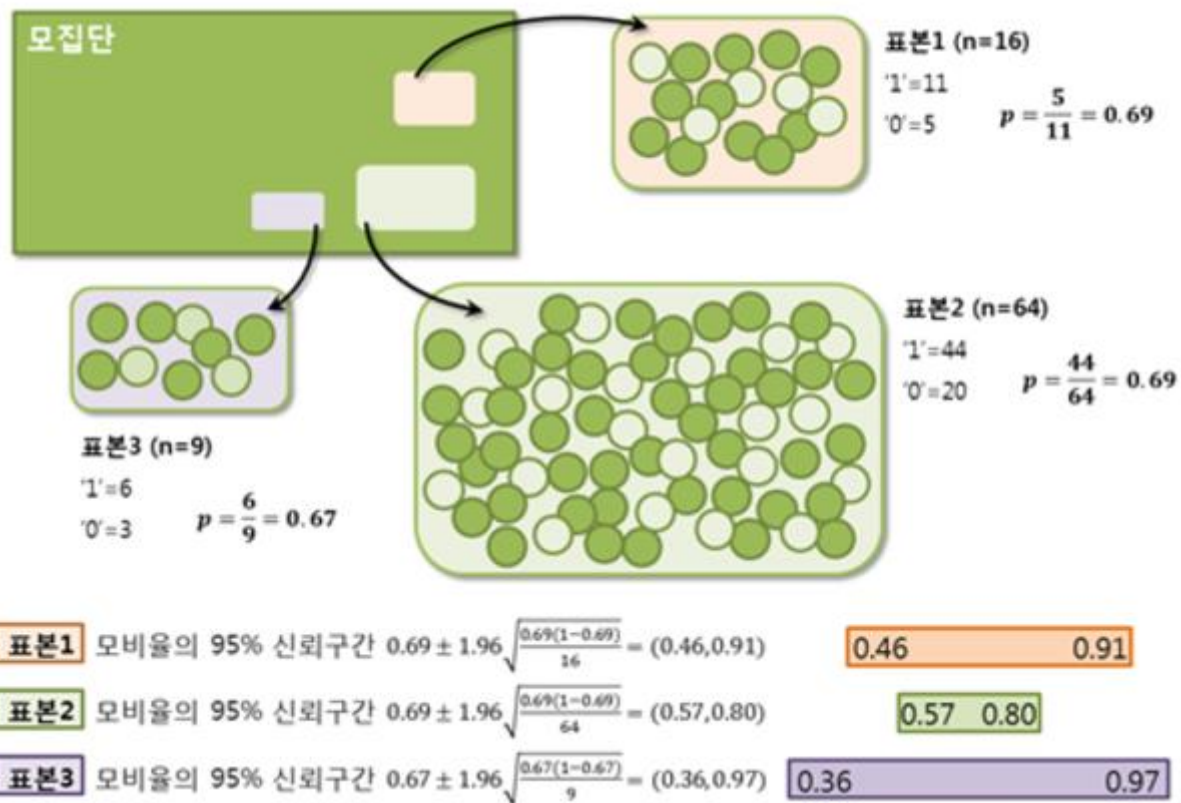
신뢰수준 C 에 따른 임계값 z^*

90% - 1.645, 95% - 1.960, 99% - 2.576

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 모바일에 대한 신뢰구간



모바일의 신뢰구간도 표본 크기에 영향을 많이 받음

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 비의 신뢰구간

- 비(ratio) : a / b 의 형태를 띠기 때문에 대개 0 이상으로 표현. 두 값이 동일한 경우 1, 비교대상보다 큰 경우 1 이상의 값
- 교차비(odds ratio) : 질병과 개별 위험인자 사이의 연관성의 정도를 추정

교차비 Odds ratio

- 질병이 있는 경우 위험인자 유무의 비와 질병이 없는 경우 위험인자 유무의 비의 비
- 환자-대조군 연구에서 주로 사용
- 통계분석에서 수학적인 장점이 있음

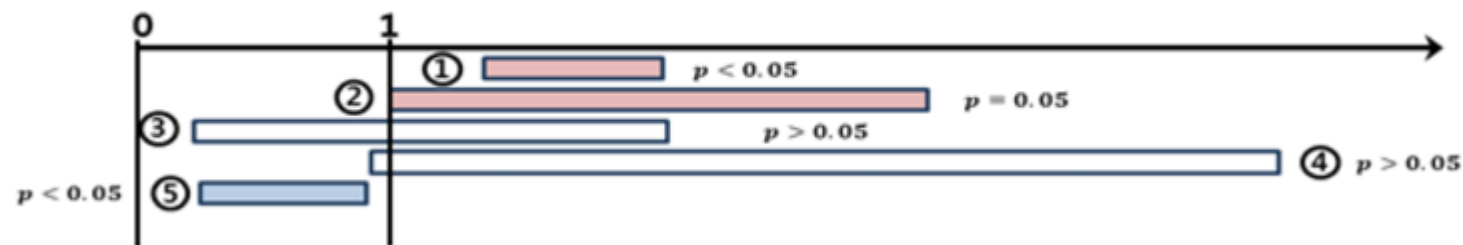
08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

가. 95% 신뢰구간

- 비의 신뢰구간

- 1의 경우 95% 신뢰구간이 1보다 큰 구간에 위치. 실제 교차비가 1이어서 질병과 위험인자 사이에 연관성이 전혀 없을 가능성은 5% 미만($p < 0.05$)
- 5의 경우 95% 신뢰구간이 0과 1사이에 완전하기 위치. 위험인자가 있는 경우 질병의 위험도가 통계적으로 유의하게 낮음($p < 0.05$)

교차비(Odds ratio)의 95% 신뢰구간



08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

나. 상대위험도와 교차비

- 코호트 연구

- 코호트(Cohort) : 특정 인구집단
- 코호트 연구(Cohort study) : 코호트를 연구대상으로 선정
 - > 여러 위험인자들(ex. 흡연, 음주, 혈압, 가족력 등)을 미리 조사
 - > 장기간 추적을 통해 실제 질병이 발생하는 것을 시간순서대로 관찰
 - > 질병의 발생과 위험인자와의 관계를 밝히는 연구
- 코호트 연구에서 위험인자 유무를 이미 알고 있기 때문에 위험인자가 있는 사람들 중 몇 %에서 질병이 발생하였고, 없는 자의 몇 %에서 질병이 발생하였는지 알 수 있음

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

나. 상대위험도와 교차비

- 코호트 연구

- 상대위험도(relative risk) : 위험인자가 있는 경우 질병이 발생하는 비율과 없는 경우 질병이 발생하는 비율의 비
- 상대위험도가 클수록 위험인자와 질병 간에 연관성이 큰 것으로 간주

코호트 연구 (전향적)

위험인자 유/무

질병발생 유/무

	질병 발생	질병 미발생	전체
위험인자 있음	A	B	A+B
위험인자 없음	C	D	C+D

$$\text{상대위험도 } Relative\ risk = \frac{A/(A+B)}{C/(C+D)}$$

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자




나. 상대위험도와 교차비

- 환자-대조군 연구

환자대조군 연구 (후향적)

위험인자 유/무

질병발생 유/무

	질병 발생	질병 미발생	
위험인자 있음	A	B	
위험인자 없음	C	D	

질병이 있을 때 위험인자의 $Odds = \frac{A}{C}$

질병이 없을 때 위험인자의 $Odds = \frac{B}{D}$

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

나. 상대위험도와 교차비

- 환자-대조군 연구

- 상대 위험도 - 대규모 코호트를 미리 설정하고 장기간에 걸쳐 추적관찰 하는 엄청난 시간과 비용, 노력
- 교차비 - 간단히 질병 발생의 위험인자를 사려 볼 수 있음

상대위험도 Relative risk

- 위험인자가 없는 경우에 비해 위험 인자가 있을 때 질병이 발생할 상대적 위험도
- 코호트 연구에서 사용
- 임상적 의미가 분명함

교차비 Odds ratio

- 질병이 있는 경우 위험인자 유무의 비와 질병이 없는 경우 위험인자 유무의 비의 비
- 환자-대조군 연구에서 주로 사용
- 통계분석에서 수학적 장점이 있음

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

나. 상대위험도와 교차비

- 상대위험도와 교차비의 관계

- 상대위험도는 위험인자가 없는 경우에 비해 있을 때 질병이 발생할 상대적 위험도로 그 의미가 분명
- 교차비는 질병이 있는 경우 위험인자 유/무 비와 질병이 없는 경우 위험인자 유/무 비로서 그 의미가 임상적으로 와 닿지 않음

	질병 발생	질병 미발생	전체
위험인자 있음	p_2	$1 - p_2$	1
위험인자 없음	p_1	$1 - p_1$	1

$$\text{교차비(odds ratio)} = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{p_2(1-p_1)}{p_1(1-p_2)} \cong \frac{p_2}{p_1} = \text{상대위험도(relative risk)}$$

만약 질환 발생이 드물어 p_1, p_2 가 매우 작다면

$$\frac{1-p_1}{1-p_2} \cong 1 \text{에 근사한다고 가정할 수 있다.}$$

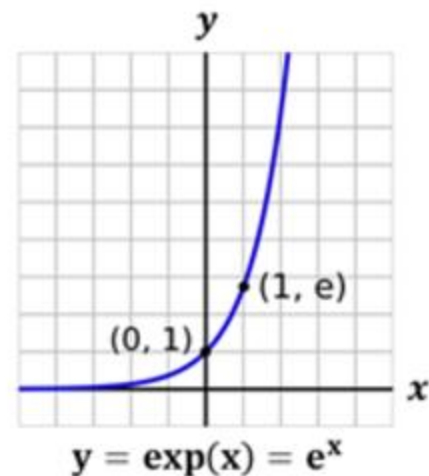
08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

다. 지수함수

$$- f(x) = \exp(x) = e^x$$

$$\exp(\beta) = e^\beta$$

- $\exp(0) = e^0 = 1$
- $\beta < 0 \rightarrow \exp(\beta) = e^\beta < 1$
- $0 < \beta \rightarrow 1 < \exp(\beta) = e^\beta$
- $\ln\{\exp(x)\} = \ln e^x = x$
- $\exp(\ln x) = e^{\ln x} = x$
- $\exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$
 $= e^{(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)} = e^{\beta_1 x_1} \times e^{\beta_2 x_2} \times \dots \times e^{\beta_k x_k}$
 $= \exp(\beta_1 x_1) \times \exp(\beta_2 x_2) \times \dots \times \exp(\beta_k x_k)$
- $\ln(x \times y) = \ln x + \ln y$



통계분석에서 자연로그(ln)를 취하는 경우가 많아 이에 상응하는 지수함수(exponential function)에 대한 이해가 필요

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

라. 일반화 선형모형

- 다중회귀분석

중회귀모형 : multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad i=1, 2, \dots, n$$

여기서, y_i 는 종속변수의 관측값

$x_{1i}, x_{2i}, \dots, x_{ki}$ 는 i 번째 관측값의 독립변수들의 값

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 는 회귀계수

ε_i 는 독립적으로 $N(0, \sigma^2)$ 를 따르는 오차항

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\Rightarrow y = X\beta + \varepsilon$$

반응변수를 설명하는데 있어 하나의 독립변수로 충분하지 않은 경우
반응변수의 변화를 설명하기 위하여 $k(k \geq 2)$ 개의 설명변수가 사용되어 질 때

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

라. 일반화 선형모형

- 일반선형모형

- **일반선형모형**(general linear model) : 다중회귀분석의 대표적인 다변수 분석법. 최소 제곱법으로 연속형 변수 사이의 회귀식을 추정. 종속변수와 독립변수의 선형 결합으로 모형화
- 일반선형모형의 4가지 기본 가정
 - ① 독립변수와 종속변수 사이의 선형성
 - ② 오차항의 정규성
 - ③ 독립성
 - ④ 등분산성

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

라. 일반화 선형모형

- **일반화선형모형**(generalized linear model) : 일반선형모형의 가정들이 적용될 수 없는 경우를 위해 확장한 것. 종속변수를 적절한 함수화로 변화시킨 $f(x)$ 와 독립변수를 선형 결합으로 모형화
- 임상 자료는 연속형이 아닌 경우가 많기 때문에 일반화선형모형을 사용
Ex) 질병 발생 유무(범주형), 질병 발생 건수(정수), 생존분석(시간의 개념이 변수에 합산)

일반선형모형

$$y = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

일반화 선형모형

$$f(x) = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

로지스틱 회귀분석

$$\ln \frac{p}{1-p} = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

Cox의 비례위험모형

$$\ln \frac{h_i(t)}{h_0(t)} = \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

포아송 회귀분석

$$\ln(\text{발생수}) = \ln(\text{전체 인구}) + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

마. 우도

- 확률 : 모수로부터 특정 현상이 관찰되는 것
- 우도(가능도, likelihood) : 확률의 반대 개념. 주어진 현상을 가지고 이 현상이 추출될 가능성을 가늠하는 것
- 최대우도법(maximum likelihood method) : 종속변수와 독립변수가 선형관계에 놓여 있지 않은 일반화 선형모델에서 사용. 우도가 가장 높은 모수를 거꾸로 추적하는 방법

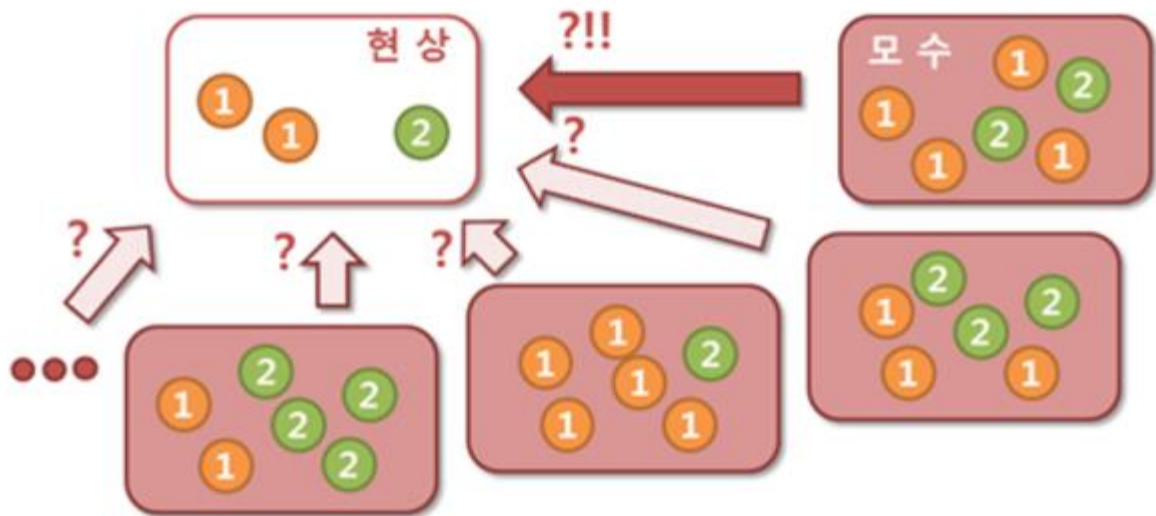
08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

마. 우도

A. 확률(probability): 모수로부터 다음과 같이 관찰될 확률은?



B. 우도(likelihood): 현상에 대해 가장 가능성이 높은(우도가 높은) 모수는?



최소제곱법 - 다중회귀분석에서 주어진 독립변수들로 종속변수를 가장 잘 예측하는 회귀식을 구할 때
최대우도법(최대편우도법) - 로지스틱 회귀분석이나 Cox의 비례위험모형에서 주어진 위험인자들로부터 종속변수를 예측하는 회귀식을 추정할 때

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

마. 우도

- 우도비 검정



뇌졸중 발생이라는 하나의 '현상'을 설명하기 위해 여러 위험인자들 중 어떤 변수들이 뇌졸중의 발생을 설명하기 위해 선택될 수 있을까?

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

마. 우도

- 우도비 검정



우도비 검정(likelihood ratio test) : 두 개의 우도의 비를 계산해서 두 모형의 우도가 유의한 차이가 있는지 비교하는 방법

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

바. 교란변수와 교호작용

- 교란변수

- **교란변수**(혼란변수, confounder) : 인과 관계를 교란시킬 수 있는 요소
- 통계적인 연관성을 입증하는 것 뿐 아니라 독립변수가 실제 위험요인인지 교란변수가 존재하지 않는지 파악하는 것이 중요



08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

바. 교란변수와 교호작용

- 교호작용

- **교호작용(interaction)** : 독립변수 사이에 상호작용에 있어서 두 효과의 합이 산술적으로 예상되는 결과가 나타나지 않는 것
- 음의 교호작용이 있어서 서로 위험요인이 상쇄되는 경우도 있을 수 있음

교호작용이 없는 경우

	정상체중	비만
정상혈압	x1	x2
고혈압	x2	x4

교호작용이 있는 경우

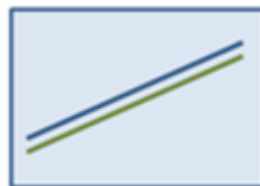
	정상체중	비만
정상혈압	x1	x2
고혈압	x2	x6

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

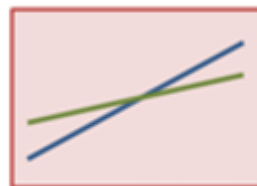
바. 교란변수와 교호작용

- 교호작용

- 두 군에 서로 다른 치료를 가한 뒤 결과변수를 여러 시점에서 반복적으로 측정
- 시간 요인을 x축, 결과변수를 y축



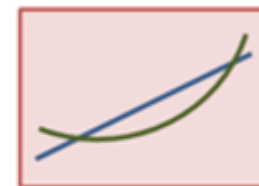
교호작용 없음



교호작용 있음



교호작용 있음



교호작용 있음

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

사. 가변수의 설정

- 2개의 범주로 나뉘는 경우 -> 0(무)과 1(유)로 코딩하여 해당변수의 회귀계수(b)를 추정
- 독립변수가 0(무)인 경우에 비해 1(유)인 경우 회귀계수(b)만큼 종속변수가 증가(혹은 감소)한다.
- 3개 이상의 범주로 나누어지는 경우 -> 가변수(dummy variable)를 설정
- 한 항목(순위 척도인 경우에는 가장 낮은 순위)을 참조항목(reference)으로 설정하고, 참조항목에 대한 다른 항목의 차이를 회귀계수(b)로 각각 추정

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

사. 가변수의 설정

자료 코딩					$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots$				
Reference		유방암 1기	0	0	0				
		유방암 2기	1	0	0				
		유방암 3기	0	1	0				
		유방암 4기	0	0	1				

결과 해석					Reference				
유방암 2기	의 y는		유방암 1기	에 비해 β_1 만큼 크고,					
유방암 3기	의 y는		유방암 1기	에 비해 β_2 만큼 크고,					
유방암 4기	의 y는		유방암 1기	에 비해 β_3 만큼 크다.					

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

아. 단변수 분석과 다변수 분석

- 단변수 분석(univariable analysis) : 두 변수 사이의 관계를 밝히는 것
- 다변수 분석(multivariable analysis) : 하나의 종속변수에 영향을 미치는 여러 위험인자들의 영향을 서로 보정하여 각 인자들이 실제로 미치는 영향의 정도를 하나의 모형으로 설명한 것. 2개 이상의 독립변수를 함께 분석
- 다변량 분석(multivariate analysis) : 2개 이상의 종속변수를 함께 분석

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

아. 단변수 분석과 다변수 분석

- **보정되지 않은 교차비**(unadjusted OR) : 영향을 줄 것으로 예상되는 모든 위험인자를 대상으로 1 대 1로 단변수 분석을 통해 얻은 교차비
- **보정된 상태의 교차비**(adjusted OR) : 단변수 분석에서 연관성이 의심되는 선별된 위험인자를 대상으로 로지스틱 회귀분석을 시행하여 얻은 교차비. 최대우도법을 통해 회귀모형을 추정
- 일반적으로 단변수 분석 뒤 다변수 분석을 시행

08. 이 여덟 가지만 더 알고 중급 통계에 들어가자

아. 단변수 분석과 다변수 분석

