윤리교육

의생명과학연구에서 올바른 통계처리

서울대 보건대학원 보건학과

김 호

Ethical Problems of Statistical Methods in Biomedical Researches

Ho Kim

Graduate School of Public Health, Seoul National University

윤리적인 연구를 위해서는 통계적인 고려가 매우 중요하다. 통계적인 방법은 항상 오류를 포함하고 있으며 그 오류를 정확히 이해하고 해석하는 것이 올바른 연구를 진행하는데 필수적인 사항이다. 통계학에서는 일종의 오류와 이종의 오류를 구분하고 있는데 일종의 오류란 귀무가설이 올바름에도 불구하고 귀무가설을 기각하는 오류이다. 전형적인 임상시험에서는 약효가 없음에도 불구하고 약효가 있다고 판단하는 오류이다. 이종의 오류는 그 반대의 경우로 약효가 있음에도 불구하고 약효가 없다고 판단하는 경우이다. 검정력이란 이종의 오류의 반대의 확률, 즉 약효가 있는 경우 그 약효를 발견할 확률이다. 올바른 연구를 위해서는 통계분석을 포함한 모든 연구진행 방법이 사전에 고정하는 것이 대단히 중요하다. 연구대상수도 이러한 의미에서 미리 정하는 것이 당연하다. 연구대상수가 클수록 큰 검정력을 얻을 수 있으므로 비용이 허락하는 한에서 최대한의 연구대상수를 확보하려는 경향이 있는데 이러한 경향은 연구윤리 측면에서 대단히 잘못된 방법이다. 임상시험을 포함한 의생명 연구는 불확실한 처치를 인간을 대상으로 실행하는 것이기 때문에 그 효과가 공인되기 전까지는 최소한의 사람을 대상으로 연구가 진행되어야 한다. 이러한 의미에서 효과를 증명할 수 있는 최소한의 연구대상수를 미리 산출하여서 연구를 진행하는 것이 중요하다.

이 이외에도 실행하는 가설이 많은 경우 다중비교의 문제가 발생하므로 p-value의 기준을 더 엄격하게 유지해야만 연구전체의 일종의 오류를 원하는 수준으로 유지할 수 있게 된다. 연구에서 이상치를 자의적으로 판단하여서 이들을 제외하고 분석을 하는 것은 연구윤리에 많은 문제를 발생시킨다. 이러한 경우 이상치 판단의 근거가 사전에 정의되어 있어야하며 모든 연주가들이 동의하는 상태에서 분석에서의 제외가 결정되어야한다. 또한 이상치 제외전과 후의 분석이 모두 제시되어서 이상치 제외가 분석결과에 어떠한 영향을 끼쳤는지도 제3자가 판단할 수 있도록 하여야한다. 연구자료 분석에서 사용된 모든 통계적 모형에는 가정이 존재하게 된다. 자료의 분석 후에는 이러한 가정들이 문제없이 적용될 수 있는지에 대한 진단이 이루어져야 하고 진단이 어려운 경우에는 민감도 분석을 통하여 연구의 결과가 모형의 가정에 얼마나 의존하는 가를 평가하여야 한다.

윤리적인 연구를 진행하기 위해 연구자가 반드시 확보해야 할 것들이 여러 가지가 있겠지만 통계적인 고려는 그 중 가장 중요한 것 중의 하나이다. 이러한 고려가 충분하지 않은 연구에서의 결과는 윤리적인 문제와 더불어 과학적 정확성에도 많은 문제를 가지고 있음은 아무리 강조하여도 지나치지 않다.

의생명과학에서 올바른 통계처리: 연구에서 흔히 발견되는 통계 오류 & 기초 통계학 및 통계자료의 이해

서울대학교 보건대학원 김 호

❖ "세상에는 세가지 거짓말이 있다. 거짓말, 새빨간 거짓말, 그리고 통계.(There are three kinds of lies: lies, damned lies, and statistics)"
- 벤자민 디즈레일리(Benjamin Disraeli)

● "진정으로 교육받은 사람의 징표는 그가 통계에 의해 깊은 감동은 받았다는 것이다. (It is the mark of a truly intelligent person to be moved by statistics.)"
- 조지 버나드 쇼 (George Bernard Shaw)

❖ "너의 자료를 고문하라. 그러면 어떠한 결과라도 얻을 수 있을 것이다.(Torture numbers, and they'll confess to anything): ~Gregg Easterbrook

❖ "만약 당신의 한쪽 발이 얼음 속에 다른 발이 끓는 솥 속에 있다고 하면, 통계적으로는 당신은 아주 평안한 상태이다. (Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable.)": Bobby Bragan, 1963

❖ There are two kinds of statistics, the kind you look up and the kind you make up. ~Rex Stout, *Death of a Doxy*

통계의 오용의 종류

❖ 의도된 자료의 취사선택
❖ 의도된 설문
❖ 과장된 일반화
❖ 왜곡된 표본
❖ 추정된 오차에 대한 오류 혹은 오용
❖ 잘못된 원인결과
❖ 자료조작

기초통계

❖ 두 집단의 비교 (예: 신약 vs 위약(Placebo))

1) 비교하고자 하는 변수가 연속형인 경우 (예, 수축기 혈압)
2) 비교하고자 하는 변수가 이산형인 경우 (예, 병에 호전여부)

❖ T-test : 두 집단 간의 평균비교
❖ Chi-square test : 두 집단간의 비율 비교

기초통계

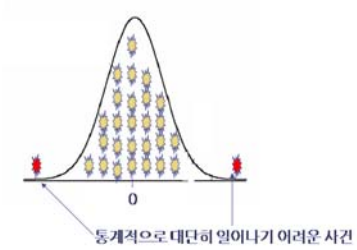
P-value (1)

❖ 연구목적 : 관심변수의 (모)평균이 두 집단에서 다르다.

❖ $\bar{Y}_1$ 첫 번째 집단에서의 표본 평균
❖ $\bar{Y}_2$ 두 번째 집단에서의 표본 평균
❖ 만약 두 집단에서의 모평균이 같다고 하면
❖ 두 표본 평균은 비슷할 것이다.
❖ 표본평균의 차이를 반복적으로 구해보면

기초통계

P-value (2)



기초통계

P-value (3)

- P-value = 두 집단의 평균이 같다고 가정했을 때 우리의 자료, 혹은 더 차이가 나는 자료를 얻을 확률

- 작은 p-value : 위의 확률이 작다
→ 통계적으로 가능하지 않은 일이 일어났다.
→ 두 집단의 평균이 같다는 가정에 문제가 있다.
→ 두 집단의 평균은 같지 않다고 결론 내린다.

기초통계

P-value (3)

- 작지 않은 p-value : 두 집단의 평균이 같다고 가정하면 우리의 자료를 관측할 확률이 작지 않다.
→ 두 집단의 평균이 같다는 가정에 문제가 없다.

양쪽검정, 한쪽검정

기초통계

- A(얻은 자료) -> B (연구가설)
- -B -> -A
- 귀무가설 (-B) : 두 집단에 차이가 없다. (Ho)
- 대립가설 (B) : 두 집단에 차이가 있다. (Ha)
- 일종의 오류 : 옳은 귀무가설을 기각할 확률
  = Pr (reject Ho | Ho is true)
  약효가 없는데도 불구하고 약효가 있다고 판단할 확률
- 이종의 오류 : 틀린 귀무가설을 받아들일 확률
  = Pr (Not reject Ho | Ha is true)
  약효가 있는데도 불구하고 약효가 없다고 판단할 확률
- Power = 1- 이종의 오류
  = 있는 차이를 발견할 확률

기초통계

대립가설 vs 귀무가설

대립가설 연구자가 주장하고자 하는 가설
직접 검정 불가
alternative hypothesis(H_1)

귀무가설 대립가설의 여사상
통계적 검정의 대상
null hypothesis(H_0)

기초통계

통계분석의 기본 개념

➢ 가설의 검정

- 귀무가설 & 대립가설
- 제1종 오류(α) & 제2종 오류(β)

➢ α, β 의 결정기준?
→ 절대적 기준은 없다!
→ 연구자가 주관적으로 결정

실제 검정결과	귀무가설이 옳 음	귀무가설이 틀 림
귀무가설 채 택	옳음	제 2 종 오류
귀무가설 기 각	제 1 종 오류	옳음

일반적으로
α = 0.01 or 0.05
β = 0.1 or 0.2

의도된 자료의 취사선택

- 어떤 계약회사에서 신약의 효과를 검정하기 위해서 유의수준 95%의 임상시험을 20번 했다고 가정

의도된 자료의 취사선택

- 어떤 계약회사에서 신약의 효과를 검정하기 위해서 유의수준 95%의 임상시험을 20번 했다고 가정
- 만약 약효가 전혀 없다고 가정했을 때 유의수준 95%의 의미는 ?

의도된 자료의 취사선택

- 어떤 계약회사에서 신약의 효과를 검정하기 위해서 유의수준 95%의 임상시험을 20번 했다고 가정
- 만약 약효가 전혀 없다고 가정했을 때 유의수준 95%의 의미는 ?
- 19 임상시험의 결과를 폐기하고 1 결과를 홍보용으로 사용한다면 ?

의도된 자료의 취사선택

- 약효가 전효 없는 물질로 200번 임상시험을 한다 해도 확률적으로 10번 정도는 약효가 관찰될 것이다.

의도된 자료의 취사선택

- 약효가 전효 없는 물질로 200번 임상시험을 한다 해도 확률적으로 10번 정도는 약효가 관찰될 것이다.
- 모든 임상시험 결과를 공표하지 않는 기관 (계약회사, 담배회사 등)

의도된 자료의 취사선택

- 변수의 수(종속변수: 효과를 나타냄, 설명변수)를 늘림 -> 우연한 발견을 할 확률을 높임
- 모든 자료를 공개하는 것이 바람직

의도된 설문

- 종부세 찬성 여부를 묻는데

1) 부동산세금이 외국에 비해서 아직도 적다 혹은 종부세 부과 대상의 70%가 다가구보유 세대다 라고 이야기하고

2) 작년에 비해서 300%까지 오른 곳이 많고 소득이 없는 일가구 소유 노인들의 예를 들면서 ….

과장된 일반화 (Overgeneralization)

- 고온과 사망률 증가 -> 기온과 사망률 증가
- 많은 경우 연구자 보다는 해석 과정에서 발생
- TV 인터뷰

왜곡된(biased) 표본

- 표본조사와 전수조사
- Target Population 과 sampling Population
- 표본오차 비표본오차 (예, 전화조사)

추정된 오차에 대한 오류 혹은 오용

- 천만명의 서울시민 중 천명을 대상으로 설문조사를 하였다.
- 대표성 ? 임의성 (randomness)
- 65 % +- 5%

추정된 오차에 대한 오류 혹은 오용

- 천만명의 서울시민 중 천명을 대상으로 설문조사를 하였다.
- 대표성 ? 임의성 (randomness)
- 65 % +- 5%
- 추정오차에 대한 설명 생략은 결과가 100% 정확하다는 잘못된 해석을 하게함

잘못된 원인결과

- A <-> B : 상관관계 관찰

잘못된 원인결과

- A <-> B : 상관관계 관찰
  - A -> B
  - B -> A
  - C -> A & C -> B
  - due purely to chance

잘못된 원인결과

- A <-> B : 상관관계 관찰
  - A -> B
  - B -> A
  - C -> A & C -> B
  - due purely to chance
- 해변에서 아이스크림 사먹은 사람수, 익사자수, (해변에 나온 사람수)

기초통계

표1. 질병상태와 노출상태에 따른 위험도 (예제1)

노출상태	질병상태		위험도
	유	무	
노출	81	29	81/(81+29)=0.7364
비노출	28	182	28/(28+182)=0.1333
상대위험도			0.7134/0.1333=5.52

결론 : 노출상태와 질병상태에는 연관이 있다.

기초통계

표2. 혼란변수 유무에 따른 위험도 (예제1)

남성

노출상태	질병상태		위험도
	유	무	
노출	1	9	0.100
비노출	20	180	0.100
상대위험도			1.00

여성

노출상태	질병상태		위험도
	유	무	
노출	80	20	0.800
비노출	8	2	0.800
상대위험도			1.00

결론 : 남녀 모두에서 노출상태와 질병상태에는 연관이 없다.

기초통계

요약하면

- 전체 집단에서는 질병과 노출에 연관 있다.
- 남자에서는 질병과 노출에 연관 없다.
- 여자에서는 질병과 노출에 연관 없다.
- ???

기초통계

표3. 질병상태와 노출상태에 따른 위험도 (예제2)

노출상태	질병상태		위험도
	유	무	
노출	240	420	0.3636
비노출	200	350	0.3636
상대위험도			1.0000

결론 : 노출상태와 질병상태에는 연관이 없다.

기초통계

표4. 혼란변수 유무에 따른 위험도 (예제2)

남성

노출상태	질병상태		위험도
	유	무	
노출	135	415	0.2455
비노출	5	45	0.1000
상대위험도			2.45

여성

노출상태	질병상태		위험도
	유	무	
노출	105	5	0.9545
비노출	195	305	0.3900
상대위험도			2.45

결론 : 남녀 모두에서 노출상태와 질병상태에는 연관이 있다.

기초통계

요약하면

- 전체 집단에서는 질병과 노출에 연관 없다. (RR=1.00)
- 남자에서는 질병과 노출에 연관 있다. (RR=2.45)
- 여자에서는 질병과 노출에 연관 있다. (RR=2.45)
- ???

기초통계

정리

- 질병상태와 노출여부는 성별에 의해 혼란(Confounding) 되고 있다
- 이러한 경우 올바른 자료의 분석을 위해서는 성별은 질병상태와 노출여부와 함께 반드시 고려해야 한다. (성별을 혼란변수라고 부른다.)
- 이와 마찬가지로 어떠한 분석을 할 때 가능한 혼란변수를 모두 고려해야만 올바른 분석결과를 얻을 수 있다.

기초통계

표5. Housing tenure by CHD(coronary heart disease) outcome after six years, SHHS (Scottish Heart Health Study) men

Housing Tenure	CHD ?		Risk
	Yes	No	
Rented	85	1821	0.0466
Owner-Occupied	77	2400	0.0311
Relative Risk			1.43

기초통계

문제점

- 집을 소유하지 못한 사람들은 빈곤한 사람들(more disadvantaged social group)이 많다.
- >> 집소유 형태는 생활양식(lifestyle)에 의해 혼란되고 있을 수 있다.
- >> 특히 흡연자의 비율은 57% 대 35% 로 세입자들이 높다. 그리고 흡연은 CHD에서 잘 알려진 위험요인이다

기초통계

표6. Housing tenure by CHD(coronary heart disease) outcome after six years, SHHS (Scottish Heart Health Study) men

Smokers

Housing Tenure	CHD		Risk
	Yse	No	
Rented	33	923	0.0345
Owner-occupied	48	1722	0.0271
RR			1.27

NonSmokers

Housing Tenure	CHD		Risk
	Yse	No	
Rented	52	898	0.0547
Owner-occupied	29	678	0.0410
RR			1.33

기초통계

결론

- 표5에서와 마찬가지로 가옥소유형태는 CHD에 위험인자로 작용한다.
- 하지만, 흡연을 고려한 후에는 상대위험비가 줄었다.
- 흡연자, 비흡연자 모두에서 상대위험비의 감소가 일어나므로 흡연은 혼란변수라고 볼 수 있다.
- * 상대위험비의 감소 폭은 크지 않으므로 혼란의 정도는 약하다고 결론 내린다.
- 원자료 : Mark Woodward (1999), Epidemiology-study design and data analysis

자료조작

- selective reporting에서부터 완전한 허구 자료의 보고 까지 다양하게 존재
- 연구가설에 부합하는 자료만 사용하는 경우
- Outlier 처리에 주의 : 연구윤리와 연결된 중요한 문제

References

- Christensen, R. and T. Reichert, 1976 "Unit Measure Violations in Pattern Recognition, Ambiguity and Irrelevancy," *Pattern Recognition*, vol. 4, pp. 239-245. Pergamon Press.
- Hooke, R., 1983, *How to tell the liars from the statisticians*: Marcel Dekker, Inc., New York, NY.
- Jaffe, A.J. and H.F. Spirer, 1987, *Misused Statistics*: Marcel Dekker, Inc., New York, NY.
- Campbell, S.K., 1974, *Flaws and Fallacies in Statistical Thinking*: Prentice Hall, Inc., Englewood Cliffs, NJ.
- Oldberg, T., "An Ethical Problem in the Statistics of Defect Detection Test Reliability," 2005, Speech to the Golden Gate Chapter of the American Society for Nondestructive Testing. Published on the Web by ndt.net at http://www.ndt.net/article/v10n05/oldberg/oldberg.htm.
- Oldberg, T. and R. Christensen, 1995, "Erratic Measure" in *NDE for the Energy Industry 1995*: The American Society of Mechanical Engineers, New York, NY. Republished on the Web by ndt.net at http://www.ndt.net/article/v04n05/oldberg/oldberg.htm.

기초통계

Multiple Comparisons

ex) 한 test 에서 유의수준이 α 인 test가 있다고 하자.

Let $H_{01}: \alpha_1 = 0$, $\Pr(\text{do not reject } H_{01} \mid H_{01} \text{ is true}) = 1-\alpha$

$H_{02}: \alpha_2 = 0$, $\Pr(\text{do not reject } H_{02} \mid H_{02} \text{ is true}) = 1-\alpha$

then $\Pr(\text{do not reject } H_0 \mid H_0)$ where $H_0 = H_{01}$ and H_{02}

$= \Pr(\text{do not reject } H_{01} \text{ and do not reject } H_{02} \mid H_0)$

$= (1-\alpha)(1-\alpha) = (1-\alpha)^2$

일반적으로 $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_k = 0$ 를

multiple comparison을 한다면 $(1-\alpha)^k \le (1-\alpha)$

$1 - 0.1855 = 0.8145 = (.95)^4 \le .95$

$\therefore$ overall α 는 0.05가 아니라 0.1855가 되므로 type I error가 Inflate 되었다.

기초통계

Multiple Comparisons

Bonferroni Correction : 만약 m개의 multiple comparison을 한다면 각각의 유의수준을 α/m 로 하면 전체의 유의수준을 α 에 가깝게 할 수 있다.

예) m이 4인 경우 $(1-\frac{0.05}{4})^4 \cong 0.95 = 1-0.05$

응용) 10개의 mean을 비교하는 경우

p값의 기준을 0.05로 하면 overall p값을 유지할 수 없으므로 각각의 경우 $\frac{0.05}{10} = 0.005$ 를 기준으로 test를 실시한다.

이를 "Bonferroni corrected p-value"라고 한다.

KEY MESSAGE: 다중비교의 문제가 있을 때는 individual test는 엄격하게 시행한다.

NEWS This Week

Science, 2/23/2003

PUBLIC HEALTH

AIDS Vaccine Trial Produces Disappointment and Confusion

	Total	Infected	Percentage Infected
All subjects	1679	98	5.8%
	3330	191	5.7
White and Hispanic	1508	81	5.4
	3003	179	6.0
Black, Asian, and Other (combined)	171	17	9.9
	327	12	3.7
Black	111	9	8.1
	203	4	2.0
Asian	20	2	10.0
	53	2	3.8
Other minorities	40	6	15.0
	71	6	8.5

PLACEBO
VACCINE

Black and white? Despite overall negative results, VaxGen's Donald Francis sees hope in the subgroup analyses.

Science, 3/7/2003

HIV/AIDS

Vaccine Results Lose Significance Under Scrutiny

subgroup analyses conducted.

Gurwith says VaxGen did nine substudies based on race. A Bonferroni correction would change the *P* value for the black subgroup to between 0.09 and 0.18. "So it wouldn't be significant," acknowledges Gurwith. He says the finding of significance in the group that combined blacks, Asians, and people of mixed race would remain, however. (Adjusted, *P* would be less than 0.04.) "This looks like a real result, and it makes some biological

Close look. Claimed protection in blacks may not be statistically significant.

기초통계

Get Motivated <예시1>

	Trt A	*Trt B*	
+	n_{11} 52	n_{12} 48	n_{1+}
-	n_{21} 48	n_{22} 52	n_{2+}
	n_{+1}	n_{+2}	n_{++}

$$\chi^2 = \frac{\left(n_{11} - n_{+1}n_{1+}/n_{++}\right)^2}{v_{11}}, \quad v_{11} = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n_{++}^2\left(n_{++}-1\right)}$$

$$\chi^2 = \frac{\left(52 - 100\times 100/200\right)^2}{\left(100\times 100\times 100\times 100\right)/\left(200^2\times 199\right)} = 0.32, \; p > 0.05$$

기초통계

- $n_{ij}^* = 100\times n_{ij}$ 라고 하고, χ^2 를 다시 계산하면

$$\chi^{*2} = 100^2/100\times\chi^2 = 32.00, \; p < 0.01$$

- 두 예에서 비율은 정확히 같음에도 불구하고 통계적 유의성은 상당히 다르다. ???
- 전통적 통계적 가설 검정의 유의성은 표본수에 크게 의존한다.
- 통계적 유의성이 없었던 경우라도 표본수를 크게 하면 유의성을 볼 수 있다.
- 표본수(실험의 비용)와 통계적 유의성(실험의 효용성)의 균형을 맞추는 것이 요구됨
- 최소의 비용으로 효과를 증명하고 싶다.

기초통계

통계학에서의 표본수 계산

- 표본조사의 경우
  - 목적 : 추정 (estimation)
  - 도구 : 표본오차
  - 예 : 여론조사
- 임상시험의 경우
  - 목적 : 검정 (testing)
  - 도구 : 제1종의 오류, 제2종의 오류
  - 예 : 임상시험

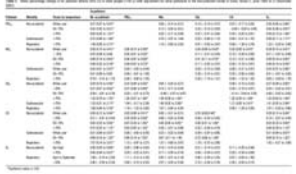
연구대상자를 충분히 확보하고 시작하는 연구가 비윤리적이다. !!
(이떻게 생각하세요?)

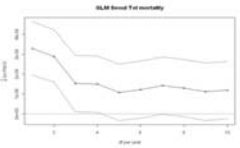
참고서적

1) Piantadosi, S. (1997). Clinical Trials. A methodological perspectives. John Wiley & Sons, New York.
2) Machin, D. et al. (1987). Sample size tables for clinical studies. Blackwell Sciences, London.
3) Schein-Chung Chow and Jen-pei Liu (1998). Design and analysis of clinical trials: Concepts and methodologies. Wiley
4) Schuster, J.J. (1993). Practical handbook of sample size guidelines for clinical trials. CRC Press, Florida.
5) 김 호 (2002). 적절한 연구대상수의 산출. 대한마취과학회지: 42(1). 1-10.
6) 박예경, 김호 (2007). 유전체 연관연구에서의 검정력 및 연구대상수 계산 고찰. 예방의학회지 40(2):114-121.
7) 김호 (2008) 튜토리얼 : R과 SAS를 이용한 표본수 및 검정력 계산 (Sample Size and Power Calculation Using R and SAS) 한국보건정보통계학회지 33(1): 11-20.

Sensitivity Analysis

- 모형에서 가정한 것들이 바뀌면서 연구의 결론이 이떻게 바뀌는지에 대한 분석,
- 어떠한 가정이 결론에 영향을 많이 미치는지에 대한 언급도 필요





ARTICLE

The Use of Probabilistic Decision Models in Technology Assessment

The Case of Total Hip Replacement

Andrew Briggs,[1] Mark Sculpher,[2] Jill Dawson,[3] Ray Fitzpatrick,[4] David Murray[5] and Henrik Malchau[6]

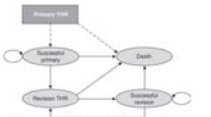


Fig. 1. Diagram showing the structure of the model. THR = total hip replacement.

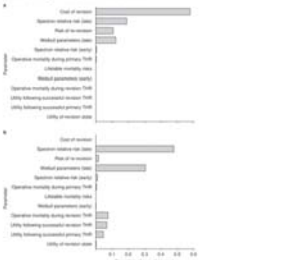


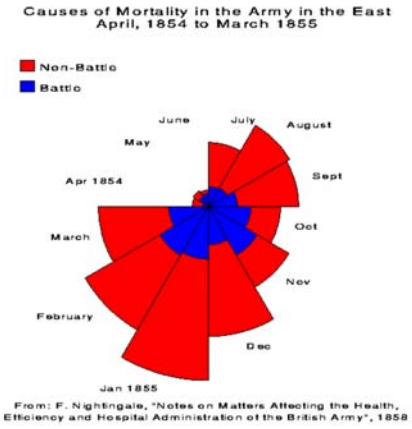
Fig. 3. Summary of individual parameter contributions to model sum of squares (analysis of covariance [ANCOVA]) for (a) incremental costs and (b) additional quality-adjusted life-years. THR = total hip replacement.

질문

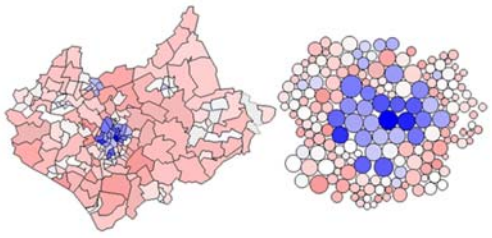
- 미국에서 수행한 연구를 매우 비슷한 상황에서 우리나라에서 반복하였다. 미국에서는 매우 유의한 결과를 얻었지만 우리나라에서는 유의한 결과를 얻는데 실패하였다. 이러한 경우 어떠한 해석이 가능한가 ?
- 한 연구자가 일종의 오류 5%를 기준으로 논문을 작성하였다. 10년간 100편의 논문을 발표하였다면 일종의 오류가 의미하는 바가 무엇인가 ?

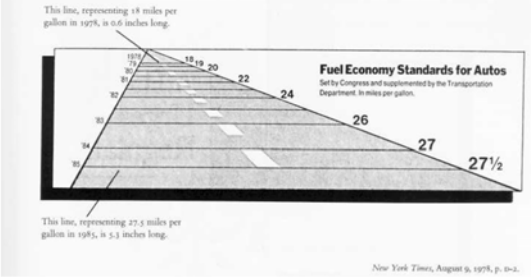
Statistical Graphics

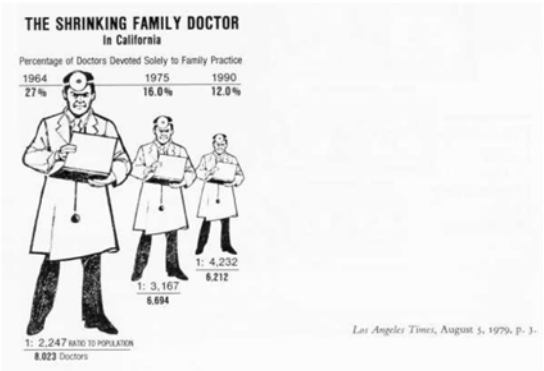
- 기술통계학(descriptive statistics)의 중요한 부분
- 자료의 직관적 이해에 도움을 줌
- http://www.math.yorku.ca/SCS/Gallery/



The 1854 London Cholera Epidemic







윤리적인 연구, 공부가 필요합니다.

감사합니다.

김호

hokim@snu.ac.kr
02) 880-2711
http: //plaza.snu.ac.kr/~hokim