Common Statistical Errors in Medical Research:

How to Report Statistics in Medical Journals

안형진, Ph.D.
고려대학교 의과대학
의학통계학교실

소개

- 통계학이란
  - 연구의 결론을 객관적으로 도출하기 위하여 자료(data)를 수집, 처리, 해석하는 학문

- 올바르지 않은 통계분석법으로 도출된 결과
  - 신뢰도 하락 → 논문의 질 하락
  - 사회적 비용의 증가
  - 예: 신약과 위약의 비교 임상시험
    - Real efficacy but not significant: 새로운 치료의 기회 박탈
    - No efficacy but significant: 잠재적 부작용에 노출, 후속연구의 진행으로 인한 사회적 비용증가

Introduction

- 왜 의과학자로서 통계학의 지식이 필요한가?
  - 전문통계학자와의 공동 연구 시 의사소통 (반대로 통계학자도 연구에 관한 기본지식 필요)
  - 수행 연구결과 발표 시 연구에 수행된 통계분석법을 이해하여야 연구결과의 신뢰성을 줄 수 있음.
  - 논문을 통한 최신의학방법 습득 시 논문의 이해도를 증가시킴
  - 좋은 연구와 그렇지 않은 연구를 구분할 수 있는 능력의 향상

연구설계

- 통계학의 중요한 역할
  - 궁극적으로 연구가 비교할만하고(comparable) 일반화할 만한 (generalizable)한지 증명
  - 연구계획 및 설계가 매우 중요

- 연구는 인과관계 추론에서 발생할 수 있는 여러 가지 편향 (bias)을 최소화할 수 있도록 설계해야 함.
  - 전향적 무작위 중재연구 (best, 일반화의 문제 있을 수 있음.)
  - 관찰연구 (confounding bias 통제 필요, 규모가 큰 연구가 가능, 일반화의 측면에서 좋을 수 있음)

- 잠재적 교란변수(potential confounders)를 미리 수집

연구의 크기

- 표본수 산정은 연구설계 시 필요

- 좋은 연구계획은 임상적 유의성과 통계적 유의성을 동시에 보일 수 있도록 충분히 큰 규모의 연구이어야 하나 임상적 유의성이 없음에도 통계적 유의성을 보일 정도의 너무 큰 연구는 지양하여야 함.

- 표본수 계산에 필요한 요소
  - 통계적 분석방법
  - 유의수준 (일반적으로 0.05)
  - 검정력 (일반적으로 0.8 이상)
  - 연구효과의 크기 (예: 군간 비율차이, 군간 평균차이)
  - 다른 관련 모수 (예: 표준편차 등)

- 준비연구 필요

통계분석 및 표현방법

- 자료의 분포적 특징
  - 기술통계 (그림 또는 요약숫자)
  - 연속형 자료: 평균±표준편차, 중앙값±사분위범위
  - 범주형 자료: 절대빈도 및 상대빈도 (%)
  - 이상값 및 특이값 검출
  - 특이값의 제외 시 그 이유를 논문에 기술

- 연구결과의 질은 얼마나 많은 통계분석방법을 사용하였느냐 또는 얼마나 어려운 통계분석방법을 사용하였느냐에 결정되는 것이 아니라 얼마나 적절한 방법을 사용하였느냐에 의존함.

가설 검정

- 가설
  - 연구의 목적과 관련된 모집단, 분포, 모수 등에 관한 어떤 주장이나 설명
- 귀무가설 (Null hypothesis)
  - 현재 믿어지고 있는 상태
  - 틀렸음을 보이고자 하는 것
  - 연구자가 보이고자 하는 주장(대립가설)을 증명할 수 없을 때 돌아가는 곳
  - H_0으로 표기
- 대립가설(Alternative hypothesis)
  - 연구가설(research hypothesis)이라고도 함
  - 연구를 통해 보이고자 하는 상황이나 새로운 주장
  - H_1 또는 H_A로 표기

가설 검정

	실제	
검정의 결과	H_0 참	H_0 거짓
H_0 기각 실패	OK	제 2종의 오류 Type II Error (β)
H_0 기각	제1종의 오류 Type I Error (α)	OK

- 제1종의 오류는 귀무가설이 참일 때 표본에 근거하여 검정한 결과 귀무가설을 기각할 때 발생한다.
- 제1종의 오류를 저지를 확률을 α로 표기한다.
- 제 2종의 오류는 귀무가설이 거짓일 때 표본에 근거하여 검정한 결과 귀무가설을 기각하지 못할 때 발생한다.
- 제 2종의 오류를 저지를 확률을 β로 표기한다.
- 확률 $(1-\beta)$를 검정력(power of the test)이라고 부른다.

검정방법

- α와 β 모두 최소화할 수 있는 검정법을 찾으면 가장 이상적이겠으나 α가 작아지면 β는 증가한다.
- 그래서 통계적 가설 검정에서는 α를 고정시키고 그에 따른 기각역(rejection region)을 구한다.
- 이제 귀무가설이 참이라고 가정한 상태에서 표본으로부터 검정통계량을 구하게 되고 이 검정통계량이 기각역에 있게 되면 귀무가설을 기각하고 기각역 밖에 있으면 귀무가설을 기각하지 못한다.
- 즉, 귀무가설이 참일 때 귀무가설을 기각할 확률이 α보다 작거나 같다.
- 일반적으로 α는 0.05를 사용하고 유의수준(level of significance)이라고 부른다.
- Note: 우리가 귀무가설을 기각하지 못한다고 해서 대립가설을 채택하는 것은 아니다. (법원에서 피고에게 무죄를 선고했다고 해서 꼭 그 피고가 죄가 없다고 할 수는 없다. 단지 그 피고가 유죄임을 보일 증거가 불충분해서 그럴 수도 있기 때문이다.) 즉, 자료가 대립가설을 증명하기에 불충분해서 귀무가설을 기각못했을 수도 있다.

통계적 유의성

- 만일 표본으로부터 얻은 증거들이 대립가설을 지지하고 그래서 귀무가설을 기각한다면, 검정결과는 (통계적으로) 유의하다고 말한다. (Statistically Significant)

- 만일 표본으로부터 얻은 증거들이 대립가설을 지지하지 않고 그래서 귀무가설을 기각하지 못한다면 검정결과는 (통계적으로) 유의하지 못하다고 말한다. (Not Statistically Significant)

p-value

- p-value는 귀무가설이 사실이라고 가정한 상황에서, 해당 관찰결과를, 또는 그 보다 더 극단적인 결과를 얻게될 가능성을 의미

- 만일 p-value가 크면, 귀무가설이 사실이라는 가정하에서, 이런 검정 통계량 값을 얻을 가능성이 높다. 그러므로 귀무가설의 타당성을 의심할 충분한 이유가 없다.

- 만일 p-value가 작으면, 귀무가설이 사실이라는 가정하에서 이런 검정 통계량 값을 얻을 가능성이 작다. 그러므로 귀무가설을 기각할 충분한 이유가 있다.

가설검정의 절차

1. 연구가설에 맞는 모수를 정한다.
2. 이 모수를 이용하여 귀무가설과 대립가설을 세운다.
3. 유의수준 α를 선택한다. 일반적으로 0.05를 사용한다.
4. 검정에 사용할 검정통계량을 지정한다.
5. 표본으로부터 검정통계량을 구한다.
6. p-값을 구한다.
7. p-값을 α와 비교하여 α보다 작으면 귀무가설을 기각하고 α보다 크면 귀무가설을 기각하지 못한다.
8. 검정결과를 바탕으로 결론을 도출한다. 이 때 결론은 연구가설과 관련된 말로 설명을 해야한다.

Note: 가설검정에서 기각역을 이용하는 방법도 있음.

95% 신뢰구간과의 관계

American Statistical Association's Statement on P-values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.

American Statistical Association's Statement on P-values

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- In some sense the p-value offers a first defense line against being fooled by randomness, separating signal from noise…The p-value is a very valuable tool, but it should be complemented – not replaced – by confidence intervals and effect size estimators (as is possible in the specific setting). Tal Galili (R-bloggers)

검정력 (Power)

- A power analysis is a way to find either: The effect size you'll be able to detect given a set sample size, OR
- The sample size you'll need to detect a specific effect size.

- •Doing a power analysis makes you **think critically** about your proposed study.

검정력에 영향을 주는 요인들

1. The design of the study and the type of measurements
   - Paired data? Two groups? Three groups? Continuous data? Nominal or ordinal data?
2. The variability of the data (e.g., standard deviation)
3. The significance level of the test
   - Usually 0.05
4. The effect size of interest (Clinical significance)
   - Ask yourself "What am I hoping to find?"
   - And "Would it be important if I found half that difference?"
5. The sample size
   - What can you afford to do?

표본수 계산에 필요한 요소들

1. The design of the study and the type of measurements
   - Paired data? Two groups? Three groups? Continuous data? Nominal or ordinal data?
2. The variability of the data (e.g., standard deviation)
3. The significance level of the test
   - Usually 0.05
4. The effect size of interest (Clinical significance)
   - Ask yourself "What am I hoping to find?"
   - And "Would it be important if I found half that difference?"
5. Adequate power
   - Usually at least 80% (or 90%)

통계분석 및 표현방법

- 적절한 통계방법의 선택 기준
  - 연구의 목적 (가설)
  - 연구설계방법
  - 교란변수 보정
  - 분석 변수의 수
  - 비교하고자 하는 군의 수
  - 자료의 종류 (연속형, 이분형, 범주형, 생존기간 등)

통계분석 및 표현방법

- 비록 적절한 통계적 방법을 사용하여 결과를 도출하였더라도 통계결과의 부적절한 표현으로 인하여 논문이 거절될 수 있음.

- 논문작성 시 주의해야 할 몇 가지 통계적 고려 사항
  - 방법(method)부분에 연구설계와 자료수집과정을 자세하게 기술
    - 통계분석방법은 방법의 이름만을 나열하는 것이 아니라 어떤 연구가설을 보이기 위하여 어떤 방법을 사용하였는지 자세히 기술
    - 통계적 유의수준 지정
    - 분석에 사용한 통계프로그램의 명시

통계분석 및 표현방법

- 결과(result)부분에 분석결과 제시
  - 표에 제시된 숫자를 다시 반복하는 것이 아니라 가능하면 표에 제시된 값들의 질적인 표현에 중점을 둠
  - 유의한 결과는 P-값과 함께 추정치(예: 오즈비)와 95% 신뢰구간을 함께 제시
  - P-값이 유의하지 않더라도 표에서 NS(not significant)로 제시하지 말고 원 P-값을 그대로 제시
  - 표, 그림, 본문에 사용되는 숫자의 소수점은 자료의 단위와 임상적으로 관련 있는 만큼 제시 (일반적으로 소수점 둘째 또는 셋째 자리)
  - 그림으로 결과를 표현하는 경우에는 x축과 y축의 단위를 포함한 변수 설명을 명확히 제시하고 범례(legend)를 구체적으로 표시
  - 표에서 약어를 사용하는 경우 주석에 풀어서 제시
  - 그림과 표의 핵심은 표나 그림만으로 저자가 표현하고 자 하는 바를 독자가 이해할 수 있도록 함.

통계분석 및 표현방법

- 결론(conclusion/discussion)부분에는 표본수, 비일반화, 관찰연구 등의 연구 제약점을 기술

  - 이 때 주의할 점은 제약점을 결론의 시작에 기술하는 것이 아니라 연구의 강점, 논문이 주는 시사점을 먼저 설명한 후 마지막에 연구의 제약점을 기술

Common Statistical Errors

- 연구설계
  - 연구 목적과 주 결과 변수가 불명확한 경우
  - 표본수를 밝히지 않는 경우
  - 연구에서 제외된 표본을 밝히지 않는 경우
  - RCT의 경우 표본수 계산에 관련된 사항이 없는 경우
  - RCT의 경우 불명확한 무작위배정 과정
  - 적절하지 않은 대조군의 사용

Common Statistical Errors

- 부적절한 자료분석
  - 적절하지 않은 통계 검정
    - 자료의 종류와 맞지 않는 통계 검정 방법 사용
    - 짝지은 자료에서 독립 표본 검정 방법 사용
    - 모수적 방법의 부적절한 사용
  - 제 1종의 오류 통제 어려움
    - 다중비교 방법 사용하지 않음
    - 적절하지 않은 사후-부집단 분석
  - 사용한 분석의 가정을 검토하지 않음
    - Outlier, Influential value 등
  - 관찰연구에서 교란변수를 보정하지 않음.

Common Statistical Errors

- 논문에서 부적절한 결과 표현
  - 분석에 사용한 모든 통계분석법은 명확하고 정확하게 서술해야 함.
  - 올바른 통계분석법의 이름을 사용해야 함.
  - 일반적이지 않은 통계분석법을 사용한 경우에는 명확한 설명이나 참고문헌을 제시하여야 함.

Common Statistical Errors

- 논문에서 부적절한 결과 표현
  - 부적절한 기술통계 제시
    - 기술통계에서 표준편차 대신 표준오차 제시
    - 자료의 분포가 치우친 경우에는 평균과 표준편차 대신 또는 더불어 중앙값과 사분위 범위 제시
  - 신뢰구간은 제시하지 않고 P-값만 제시
  - 부적절한 p-값의 표현
  - 너무 많은 소수점 이하 자리수
- 부적절한 결과 해석
  - 유의하지 않음을 차이가 없음이나 영향이 없음으로 해석하는 경우
  - 연구 자료 분석결과와 관련 없는 결론

Frequently used statistical methods
for independent continuous response

Situation	Statistical Methods
One sample	**1. Parametric: One sample t-test** **2. Non-parametric: Wilcoxon signed rank test**
Independent two-sample comparison	**1. Parametric: Independent two-sample t-test** **2. Non-parametric: Wilcoxon rank sum test (A.K.A. Mann-Whiteny U test)**
Three or more group Comparison	**1. Parametric: analysis of variance(ANOVA) with multiple comparison** **2. Non-parametric: Kruskal-Wallis test**
Relationship with continuous predictors	**1. One predictor: simple linear regression** **2. Multiple predictors: multiple linear regression**
Relationship with continuous and categorical predictors	**1. Analysis of covariance (ANCOVA) or** **2. General linear models (GLM)**

Frequently used statistical methods for correlated continuous response

Situation	Statistical Methods
Paired data	**1. Parametric: paired t-test** **2. Non-parametric: Wilcoxon signed rank test on differences**
Repeated measures	**1. Repeated measures ANOVA (RM-ANOVA) or** **2. Linear mixed model (LMM)**

Frequently used statistical methods for independent categorical response

Situation	Statistical Methods
Relationship with one categorical predictor	**1. Chi-square test with large sample** **2. Fisher's exact test with small sample** **3. Simple linear logistic regression**
Relationship with categorical and continuous predictor	**1. Multiple logistic regression for binary response** **2. Multinomial or ordinal logistic regression for categorical response (more than two levels)**

Frequently used statistical methods for correlated categorical response

Situation	Statistical Methods
Paired categorical data analysis	**1. McNemar's test**
Repeated categorical responses	**1. Marginal model using generalized estimating equation (GEE)** **2. Generalized linear mixed model (GLMM)**

Frequently used statistical methods
Other methods

Situation	Statistical Methods
Correlation between two continuous variables	**1. Pearson correlation coefficient or** **2. Spearman's correlation coefficient**
Survival data	**1. Kaplan-Meier survival curve** **2. Log-rank test** **3. Cox's proportional hazard model**
일치도 분석	**1. Sensitivity, Specificity** **2. Kappa measure** **3. Concordant correlation coefficient)** **4. ROC Analysis including AUROC curve** **5. Bland-Altman analysis**

출판이 거절되는 흔한 이유

- 주제가 임상적으로 중요하지 않음
- 고유한 연구가 아님
- 실제로 저자의 가설을 검증한 연구가 아님
- 연구설계의 문제
- 연구계획대로 하지 못한 연구
- 표본의 크기가 작은 연구
- 대조군이 없거나 선정에 문제가 있는 연구
- 부적절하거나 잘못된 통계분석
- 자료에 근거하지 않은 결론을 유도
- 이해관계의 상충의 의심
- 이해하기 힘들 정도로 글이 엉망인 경우
- 심사위원을 잘못 만난 경우 (?)

통계의 오용

- P-값이 0.05보다 작게 나올 때까지 갖가지 방법을 사용
- 교란변수를 통제하지 않고 분석하고 결론을 내림
- 사용한 통계분석의 가정을 확인하지 않음 (가장 많은 오용)
- 중도탈락자와 무응답자를 무시함.
- 인과관계와 연관관계의 혼용
- 분석결과가 좋지 않으면 몇 몇 값을 자료에서 제외 (특히, 특이값)
- 보이고자 하는 결과만 보임 (6개월을 예상한 연구에서 4개월째 유의한 결과를 보이면 연구를 중단하고 논문작성, 6개월의 결과가 좋지 않으면 임의로 연구를 6개월 더 연장)
- 특정 집단을 계속 나누어 유의한 결과를 보일 때까지 분석

결론

- 의학연구를 수행하고 타당한 결과를 도출하여 논문으로 출판하기 위해서는 먼저 명확하고 의미 있는 연구주제를 확립하고 이 연구주제에 맞는 올바른 연구설계를 하여야 함.

- 연구설계대로 자료를 수집하고 적절한 통계분석법을 이용하여 결과를 내고 논문에 명확하게 통계방법과 결과를 기술함.

- 이 때 기준은 같은 자료가 있다면 독자들이 같은 통계방법을 시행할 수 있을 정도로 명확하게 기술함.

- Curran-Everett와 Benos(2004)가 제시한 의학저널에 통계를 보고하는 10가지 가이드라인.

Guideline

1. If in doubt, consult a statistician when you plan your study.
2. Define and justify a critical significance level α appropriate to the goals of your study.
3. Identify your statistical methods, and cite them using textbooks or review papers.
4. Control for multiple comparisons.
5. Report variability using a standard deviation.
6. Report uncertainty about scientific importance using a confidence interval.
7. Report a precise P value.
8. Report a quantity so the number of digits is commensurate with scientific relevance.
9. In the Abstract, report a confidence interval and a precise P-value for each main result.
10. Interpret each main result by assessing the numerical bounds of the confidence interval and by considering the precise P-value.

관찰연구에서 선택편향 (교란편향)을 보정하는 방법

서론

- 편향 (Bias)
  - 연구결과와 실제현상 사이에 구조적인 차이 (systematic difference)가 존재하면 편향(bias)이 발생되었다고 한다
    - 조사자 편향 (observer bias) 및 평가편향 (assessment bias) – 임상시험(눈가림)
    - **교란편향 (confounding bias)**
    - **선택편향 (selection bias) – 관찰연구**
    - **할당편향 (allocation bias) – 시험연구에서 비 무작위배정**
    - 정보편향 (information bias) - 측정오차
    - 출판편향 (publication bias) – Systematic Review
    - 기억편향 (recall bias) – 사례-대조 연구
    - 건강인 진입효과 (healthy entrant effect) – 코호트 연구

서론

- 연구설계의 중요성
  - 잘못된 연구설계 → 부적절한 결론도출
  - 잘못된 연구설계 → 어떤 통계적 방법도 설계상의 문제를 해결할 수는 없다.
  - "… a poor design cannot be salvaged by good statistics."

- Good Work → Team Work
  - Get biostatisticians on board in the beginning of the study.

연구설계의 종류

- 시험연구 또는 중재연구 (Experimental, Interventional Study)
  - 외부요인(extraneous factor)들을 통제하면서 연구에 관련된 요인을 조작
  - Randomized Controlled Trials (RCT)은 실험연구의 한 예 (다른 위험요인들의 직접적 통제는 불가능).
  - 일반적으로 연구자가 결과에 영향을 미칠지도 모르는 다른 요인들을 통제할 수 있으므로 연구가설을 증명 (또는 인과관계의 추론)할 수 있는 가장 확실한 연구설계라 할 수 있다..
  - 윤리문제 등 여러 가지 제약으로 특히 사람을 대상으로 하는 연구에서는 불가능한 경우가 많다.

연구설계의 종류

- RCT의 장점
  - 엄격하고 근거중심
  - 객관적이고 독립적
  - 할당편향(교란편향)의 최소화
  - 알려지거나 알려지지 않은 기저변수의 보정
  - 인과관계를 조사할 가능성이 높음
- RCT의 단점
  - 대조군의 선택이 어려움
  - 비용과 시간
  - 윤리적인 문제
  - 일반화의 문제
  - 결과변수의 제한

연구설계의 종류

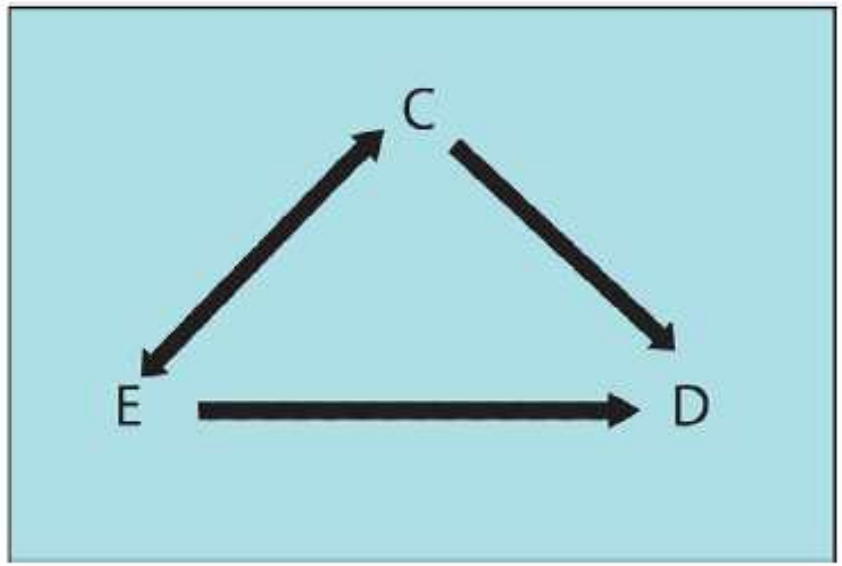
- 관찰연구 (Observational Study)
  - 실험에 미치는 요인을 통제할 수 없으며 단지 관찰만 가능한 연구
  - 예: Cohort Study, Case-Control Study, Cross-Sectional Study (표본조사)
  - 역학연구(epidemiological study): 일반 인구집단을 대상으로 관심요인과 질병과의 관계를 평가하는 연구이며 관찰연구에 속한다.

연구설계의 종류

- 관찰연구의 장점
  - 유연성
  - 일반적으로 윤리적
  - 상대적으로 큰 표본수
  - 일반화의 가능성

- 관찰연구의 단점
  - 대조군 선택의 어려움
  - 선택편향 (교란편향)의 발생
  - 인과관계를 밝히기 어려움
  - 군간 표본수의 불균형

교란편향 (Confounding Bias)



- 교란변수(C: confounder)에 의한 처치(E: Exposure)와 병(D: Disease)의 관계 도식화

교란편향을 보정하는 통계적 방법

- 짝짓기 (Matching)
  - 처치군과 대조군에서 교란변수의 분포를 동일하게 하는 개체를 선택하는 방법 (예: one-to-one matching)
  - 장점
    - 교란변수의 각 수준에서 처치군과 대조군의 수가 적절하게 보장된다면 이 방법은 교란편향을 최소화할 수 있는 효율적인 방법
    - Matching 후의 분석방법은 비교적 쉬움.
  - 단점
    - 교란변수의 수가 증가할 수록 짝지을 개체를 찾기가 어려움
    - 짝지은 자료만으로도 충분한 검정력을 갖도록 표본의 수가 커야함.
    - 측정되지 않은 교란변수의 통제는 불가능
    - 사례-대조 연구의 경우 Over-matching의 문제가 있을 수 있음.

교란편향을 보정하는 통계적 방법

- 층화 (Stratification)
  - 교란변수의 균일 층(homogeneous strata)내에서 처치와 결과를 평가함.
  - 예제

TABLE II Stratified Analysis of Nail Fixation Compared with Plate Fixation and the Effect on Development of Adult Respiratory Distress Syndrome or Multiple Organ Failure[17]

	Chest Injury		No Chest Injury		Total Cohort	
	Nail	Plate	Nail	Plate	Nail	Plate
Developed adult respiratory distress syndrome or multiple organ failure	5	2	4	1	9	3
Number at risk	117	104	118	114	235	218
Risk	0.043	0.019	0.033	0.009	0.038	0.014
Risk difference (95% confidence interval)	0.024 (−0.022, 0.069)*		0.025 (−0.012, 0.062)*		0.024 (−0.004, 0.054)	
Summary risk difference (95% confidence interval)	0.024 (−0.005, 0.053)†				P value = 0.11‡	

*Risk differences between strata are not significantly different; that is, no interaction (test for heterogeneity; p value = 0.96). †Given the absence of interaction, a pooled summary validly estimates the risk difference, adjusting for chest injury. ‡P-value testing the null hypothesis of no association between treatment method and outcome, adjusting for chest injury and assuming no interaction.

Saam Morshed et al. Analysis of Observational Studies: A Guide to Understanding Statistical Methods. J Bone Joint Surg Am. 2009;91 Suppl 3:50–60

교란편향을 보정하는 통계적 방법

- 층화
  - 장점
    - 간편함
    - Effect modification을 볼 수 있음
  - 단점
    - 여러 개의 교란변수를 통제하기 어려움
    - 연속형 교란변수의 통제가 어려움

교란편향을 보정하는 통계적 방법

- 다변수 회귀분석 (Multiple Regression)
  - 여러 개의 독립변수와 결과변수와의 함수적 관계를 모형화하여 교란변수를 통제함.

  - 예제: 결과변수의 형태에 따른 회귀모형

결과변수의 형태	회귀분석 모형	효과의 추정치
연속형	선형 회귀	평균의 차이, 기울기
이분형	로지스틱 회귀	오즈비 (odds ratio)
사건까지의 시간	콕스의 비례위험	위험비 (hazard ratio)
율 (rate)	포아송 회귀	율의 비 (rate ratio)

교란편향을 보정하는 통계적 방법

- 다변수 회귀분석
  - 장점
    - 여러 개의 교란변수를 효율적으로 통제할 수 있음.
    - 개개 인자의 효과를 비교적 쉽게 평가할 수 있음.
  - 단점
    - 모형의 선택에 따라 결과가 민감할 수 있음.
    - 잘못된 모형 선택 시 편향이 더욱 심해질 수 있음.
    - 모형의 가정에 민감할 수 있음
    - 다중공선성 등 수학적 문제 발생 가능성

교란편향을 보정하는 통계적 방법

- 성향점수 짝짓기 (Propensity Score Matching: PSM)
  - 처치군에 속할 확률에 따라 짝짓기
  - 짝지을 변수가 너무 많은 경우 일반 짝짓기 방법은 현실적으로 불가능
  - 성향점수에 근거하여 짝짓기

    성향점수 = Pr(Treatment | Confounders)
  - $0 \leq$ 성향점수 ≤ 1
  - 성향점수 이론에 의하면 성향점수가 주어진 경우에 편향이 없다면 교란변수가 주어진 경우에도 편향이 없음.

교란편향을 보정하는 통계적 방법

- 성향점수 짝짓기 (Propensity Score Matching: PSM)
  - 성향점수는 알려져 있지 않기 때문에 자료로부터 예측하여야 함.
  - 처치가 2개의 수준으로 구성된 경우에는 일반적으로 로지스틱 회귀분석이나 프로빗 회귀분석을 이용하여 성향점수를 예측함.

$$\ln\left(\frac{PS}{1-PS}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$$PS = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\right)}$$

교란편향을 보정하는 통계적 방법

- 최근접 이웃 매칭(nearest neighbor matching)은 가장 일반적인 매칭 방법으로 처치군에 대조군을 매칭하고 매칭이 되지 않은 대조군들은 제외하기 때문에, 거의 항상 처치군의 평균처치효과를 추정할 수 있는 방법
  - Greedy Matching
  - Optimal Matching

교란편향을 보정하는 통계적 방법

- 균형 확인 방법
- 표준화 차이
  - 연속형 변수

$$d = \frac{(\bar{x}_{\text{치료군}} - \bar{x}_{\text{비치료군}})}{\sqrt{\frac{s^2_{\text{치료군}} + s^2_{\text{비치료군}}}{2}}}$$

  - 범주형 변수

$$d = \frac{(\hat{p}_{\text{치료군}} - \hat{p}_{\text{비치료군}})}{\sqrt{\frac{\hat{p}_{\text{치료군}}(1-\hat{p}_{\text{치료군}}) + \hat{p}_{\text{비치료군}}(1-\hat{p}_{\text{비치료군}})}{2}}}$$

- 공변량의 균형을 파악하기 위해 표준화 차이를 이용하는 방법은 표본 수에 영향을 받지 않으며, 표준화 차이를 이용하여 성향점수 보정 후의 잔차 불균형을 일으키는 특정한 공변량을 구분할 수도 있음.
- 또한 측정단위에 의존하지 않으며, 경험적으로 표준화 차이가 0.1보다 큰 경우 해당 공변량이 불균형이라고 판단함. (Austin &Mamdani, 2006;Austin 등, 2007;Normand 등, 2001).

교란편향을 보정하는 통계적 방법

- PSM의 절차
  - 성향점수 모형을 구축함.
  - 모든 자료를 이용하여 성향점수를 예측함.
  - 짝짓기 알고리즘과 성향점수를 이용하여 처치군과 대조군을 짝지음.
  - 짝지은 자료에서 교란변수들이 군간 균형을 이루는 지 평가함. 만일 균형이 충분하지 않은 경우 성향점수 모형을 재구축함.
  - 짝지은 자료를 이용하여 평균처치효과를 추정함.
  - 이 이외에도 층화, 성향점수 회귀보정, 성향점수 가중치 방법도 있음.

교란편향을 보정하는 통계적 방법 예제

- **결혼여부와 비만 관련성의 성별차이 분석: 성향점수매칭방법**
  - **김다양, 이광수 (2015) 보건경제와 정책연구 제21권 제2호**
- 대상
  - 질병관리본부 주관으로 시행되고 있는 국민건강영양조사(Korea National Health and Nutrition Examination Survey) 제5기 3차년도 (2012) 자료를 이용
  - 2012년 국민건강영양조사 자료 중 법적으로 결혼이 가능한 연령 만 18세 이 상을 연구 대상자로 하였다. 그리고 결측값이 있는 대상자와 결혼상태가 이혼, 사별, 별거인 대상자는 제외하여 연구 대상자는 남자 2,184명, 여자 2,643명으로 총 4,827명
  - 본 강의에서는 여성만 고려함.
- 비만은 BMI 25이상은 비만 미만은 정상으로 분류, 결혼여부는 대상자에서 이혼, 사별, 별거를 제외한 기혼과 미혼으로 분류
  - 주 결과 변수: 비만여부
  - 단변수 분석(unadjusted analysis): Unadjusted odds ratio – 3.55 (2.82, 4.46), 유의함
  - 여성인 경우 기혼이 미혼보다 비만일 오즈(위험)는 약 3.55배이다.

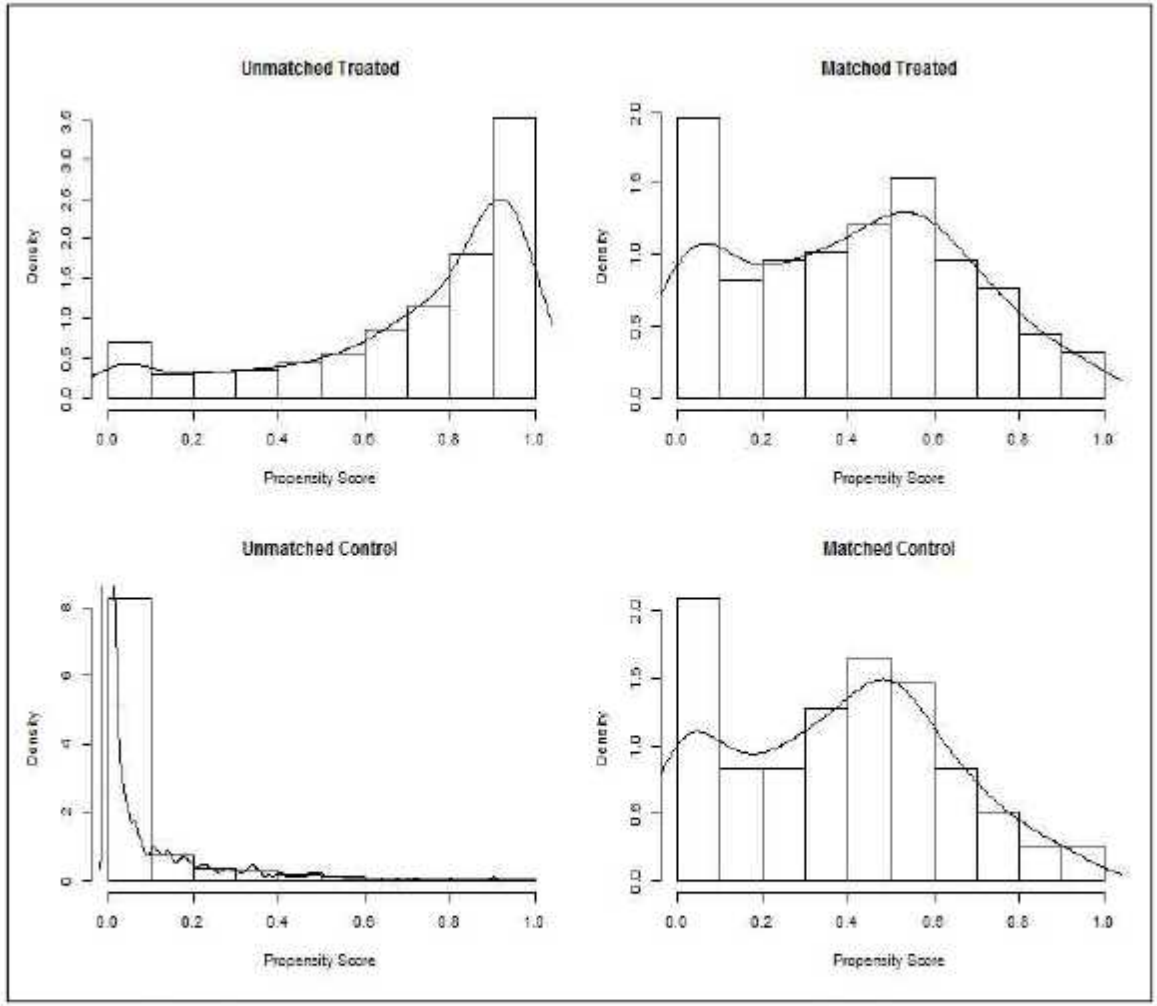
교란편향을 보정하는 통계적 방법 예제

- 성향점수 짝짓기 방법
- 성향점수는 8개의 잠재적 교란변수(교육수준, 가구소득, 1년 간 음주빈도, 현재 흡연여부, 스트레스 인지여부, 우울경험여부, 걷기실천여부, 연령)를 독립변수로 결혼여부를 종속변수로 하여 로지스틱 회귀모형으로 예측
  - 1:1 최근접 이웃방법을 이용하여 매칭된 157쌍을 얻어 표본수가 314명으로 자료의 약 12%를 사용하여 분석을 실시
  - 매칭된 자료를 이용하여 구한 Odds ratio가 1.23 (95% CI: 0.76, 1.98)으로 미혼이 기혼보다 우울증에 걸릴 Odds가 1.23배 높았으나 유의하지 않음.
  - 보정하지 않은 결과는 유의하였으나 매칭으로 보정한 결과는 유의하지 않았음. (OR 3.55 → 1.23)

<표 2> 성향점수 매칭 전후 차이분석(여성)

변수		매칭 전			매칭 후		
		기혼 (n=2,195)	미혼 (n=448)	t/x²	기혼 (n=157)	미혼 (n=157)	t/x²
교육수준	초등학교 이하	583(26.6)	2(0.4)	205.260**	8(5.1)	2(1.3)	7.203
	중학교	269(12.3)	19(4.2)		7(4.5)	3(1.9)	
	고등학교	732(33.3)	212(47.3)		57(36.3)	50(31.8)	
	대학교 이상	611(27.8)	215(48.0)		85(54.1)	102(65.0)	
가구소득	하	310(14.1)	33(7.4)	22.957**	8(5.1)	7(4.5)	2.113
	중하	589(26.8)	110(24.6)		50(31.8)	39(24.8)	
	중상	628(28.6)	129(28.80		52(33.1)	58(36.9)	
	상	668(30.4)	176(39.3)		47(29.9)	53(33.8)	
1년간 음주빈도	지난 1년간 전혀 마시지 않음	412(18.8)	48(10.7)	115.814**	34(21.7)	30(19.1)	2.703
	월1회 미만	564(25.7)	127(28.3)		38(24.2)	41(26.1)	
	월1회 정도	251(11.4)	73(16.3)		23(14.6)	23(14.6)	
	월2-4회	362(16.5)	130(29.0)		31(19.7)	39(24.8)	
	주2-3회	123(5.6)	42(9.4)		15(9.6)	14(8.9)	
	주4회 이상	41(1.9)	9(2.0)		3(1.9)	2(1.3)	
현재 흡연여부	평생 술을 마셔본 적이 없음	442(20.1)	19(4.2)	68.242**	13(8.3)	8(5.1)	0.880
	피움	59(2.7)	43(9.6)		13(8.3)	15(9.6)	
	가끔 피움	18(0.8)	13(2.9)		5(3.2)	3(1.9)	
	과거엔 피웠으나, 현재 피우지 않음	88(4.0)	27(6.0)		8(5.1)	10(6.4)	
	흡연한 적이 없음	2,030(92.5)	365(81.5)		131(83.4)	129(82.2)	
스트레스 인지여부	적게 느낌	1,629(74.2)	252(56.3)	58.516**	91(58.0)	96(61.1)	0.331
	많이 느낌	566(25.8)	196(43.8)		66(42.0)	61(38.9)	
우울 경험 여부	2주 연속 우울감 없음	1,880(85.6)	382(85.3)	0.044	131(83.4)	138(87.9)	1.271
	2주 연속 우울감	315(14.4)	66(14.7)		26(16.6)	1,912.1)	
걷기운동 여부	걷기1회 30분 이상, 주5일 이상 x	1,444(65.8)	223(49.8)	40.941**	96(61.1)	95(60.5)	0.013
	걷기1회 30분 이상, 주5일 이상	751(34.2)	225(50.2)		61(38.9)	62(39.5)	
연령		51.09(13.26)	26.2(7.5)	272.607**	34.44(10.18)	33.01(8.45)	0.831
비만여부	아니오	993(45.2)	334(74.6)	127.892**	104(66.2)	111(70.7)	0.723
	예	1,202(54.8)	114(25.4)		53(33.8)	46(29.3)	

*P< 0.05 **P< 0.01



(a) Before PSM (b) After PSM

[그림 3] 성향점수 매칭 전후 분포 비교(여성)
Treated : 미혼, Control : 기혼

<표 3> 성향점수 매칭 후 로지스틱 회귀분석

변수		남성(n=314)		여성(n=314)	
		Odds Ratio	95% CI	Odds Ratio	95% CI
연령		0.98	(0.94 - 1.01)	1.03	(0.99 - 1.06)
결혼여부 (0: 미혼 1: 기혼)		1.77*	(1.09 - 2.88)	1.06	(0.64 - 1.76)
가구소득	하(reference)				
	중하	1.01	(0.4 - 2.9)	1.72	(0.46 - 6.39)
	중상	1.52	(0.55 - 4.1)	1.58	(0.43 - 5.82)
	상	1.49	(0.55 - 3.96)	1.47	(0.39 - 5.6)
교육수준	초등학교 이하(reference)				
	중학교	7.29	(0.59 - 88.73)	0.28	(0.03 - 2.5)
	고등학교	3.72	(0.34 - 40.37)	0.25	(0.04 - 1.63)
	대학교 이상	5.57	(0.51 - 61.47)	0.16	(0.03 - 1.01)
1년간 음주빈도	최근 1년간 전혀 마시지 않았다(reference)				
	월1회 미만	1.26	(0.47 - 3.34)	1.00	(0.48 - 2.08)
	월1회 정도	0.48	(0.17 -1.41)	0.81	(0.34 - 1.94)
	월2-4회	1.13	(0.48 - 2.6)	0.92	(0.43 - 2.00)
	주2-3회 정도	1.25	(0.50 - 3.1)	0.6	(0.21 - 1.69)
	주4회 이상	1.26	(0.33 - 4.87)	1.06	(0.11 - 10.64)
	평생 음주한적 없음	1.36	(0.26 - 7.07)	0.98	(0.3 - 3.15)
현재 흡연여부	피움(reference)				
	가끔 피움	1.47	(0.28 - 7.9)	0.34	(0.03 - 3.45)
	과거엔 피웠으나, 현재 피우지 않음	0.69	(0.36 - 1.3)	1.07	(0.28 - 4.12)
	흡연한 적이 없음	0.75	(0.4 - 1.4)	1.09	(0.43 -2.79)
스트레스 인지여부		1.09	(0.6 - 1.89)	1.31	(0.77 - 2.25)
우울 경험 여부		0.97	(0.38 - 2.47)	1.14	(0.54 - 2.43)
걷기 운동		0.6*	(0.37 - 0.99)	0.76	(0.44 - 1.29)

교란편향을 보정하는 통계적 방법

- PSM 예제: 결혼여부와 비만과의 관련성
- 결론 및 제약점
  - 이 연구의 표본은 아마도 RCT보다 좀 더 실제 모집단을 대표할 수 있는 표본임.
  - PSM은 RCT가 아님
  - 단지 측정된 교란변수만 보정이 가능함.
    - 민감도 분석
  - 처치와 대조군 사이에 겹치는 부분이 많아야 함.
  - 큰 표본수가 필요함.
  - 단면연구라는 제약점 (인과성 해석의 제약)
- 성향점수를 이용한 다른 방법
  - 가중치 방법
  - 회귀 보정

결론

- 관찰연구에서는 선택편향 또는 교란편향을 최소화할 수 있는 방법을 선택해야 한다.
- 각 방법의 장,단점을 잘 이해하고 가장 적절한 방법을 선택하여 분석을 실시한다.
- 방법을 적용하는 경우 방법의 가정을 만족하는 지 꼭 확인하여야 한다.
- 결과에서 인과관계(causality)를 해석할 때는 연구설계, 편향의 최소화 등 다각도로 고려해야 하며 관찰연구에서의 제한점을 꼭 숙지하고 있어야 한다.
- 관찰연구가 종적(longitudinal)이면 중도탈락 등으로 인한 결측이 많이 발생할 수 있다. 결측은 또 다른 편향을 발생시킬 수 있으므로 결측을 고려한 분석법을 적용하여야 한다.