학술지 편집인이 알아야 하는 의학통계:
How to Report Statistics in Medical Journals

안형진
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.

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.