의학학술지에서 볼 수 있는 통계 오류

건국의대 산부인과학교실

김수녕

학습 목표: 우리 나라 의학학술지에서 흔히 쓰는 통계를 설명하고 오류의 특징을 기술할 수 있어야 한다.

구체 목표:

1) 흔히 쓰이는 통계를 3가지 이상 나열할 수 있어야 한다.
2) 통계에서 오류를 일으켜 투고자나 심의자가 주의하여야 할 내용을 3 가지 이상 기술할 수 있어야 한다.
3) 통계 오류를 최소로 하는 방안을 마련할 수 있어야 한다.

1. 서론

의학논문에서 통계분석은 의학이 발달할수록 그 중요성이 증가하고 있으며 분석방법도 점차 다양화되고 있다. 근거중심의학의 시대에 의학학술지의 통계적 내용에 대한 평가는 필수적이라 할 수 있다. 통계분석 방법의 오류는 논문 결론의 오류를 뜻하는 매우 중대한 잘못임에도 불구하고 대다수 국내외 의학학술지에서 간과되고 있는 실정이다. British Medical Journal에서는 의학학술지에 통계분석의 중요성을 인식하여 1979년도부터 논문심사 시에 통계부분의 심사를 별도로 시행하는 논문심사 제도를 도입하였다. 우리나라에서는 소수의 의학학술지를 제외하고는 통계에 대한 전문적인 심사가 이루어지지 않고 있으며 이에 대한 체계적인 교육도 거의 전무한 실정이다. 본 연제에서 국내 의학학술지의 질적 향상을 위한 방법의 일환으로 국내 의학학술지에서 흔히 사용되는 통계적 방법에 대하여 투고자나 심사자가 반드시 알아야할 내용을 설명한 다음 국내 의학학술지에서 흔히 발견되는 통계적 오류와 그 원인을 분석하고 이에 대한 개선책을 논하고자 한다.

2. 통계학의 기초개념

2.1. 모집단(population)과 표본(sample)

모집단이란 조사 단위의 전체 집합을 뜻한다. 모집단은 크기에 따라 유한모집단과 무한모집단으로 나누어진다. 표본이란 모집단의 일부분을 뜻한다. 실제 통계 분석에서 모집단을 대상으로 자료를 수집하기란 거의 불가능하고 많은 시간과 막대한 경비를 소모하여야하므로 모집단에 관한 정보를 얻기 위하여 대부분 표본을 사용한다. 표본은 모집단을 대표할 수 있도록 임의로 추출하여야 한다. 이를 임의표본(random sample) 또는 확률표본이라 한다.

2.2. 모수(parameter)와 통계량(statistic)

모수는 모집단의 특성을 나타내는 양적인 측도를 뜻한다. (모평균, 모분산)
통계량은 표본의 특성을 나타내기 위하여 표본 자료에서 계산된 측도를 뜻한다. (표본평균, 표본분산)

2.3. 변수(variable)

모든 자료의 특성은 자료들 간에 차이가 있다. 자료의 특성을 나타내는 요소들을 변수 또는 변인이라 부른다. (연령, 체중, 혈압 등)

2.4. 자료(data)의 형태

자료의 형태는 연산을 할 수 있는가 없는가에 따라서 다음과 같이 나눌 수 있다.

2.4.1. 명목자료(nominal data)

측정대상을 분류하는 역할을 하는 자료이다. 연산은 할 수 없는 자료이다. (예, 성별)

2.4.2. 순위자료(ordinal data)

측정 대상의 크기를 순위로 나타낸 자료를 뜻한다. 측정 대상 간의 대소 관계, 높고 낮음 등의 순위는 알 수 있으나 양적인 비교는 할 수 없다. 명목자료와 마찬가지로 연산은 할 수 없는 자료이다. (예, 병기)

2.4.3. 구간자료(interval data)

구간자료는 등간자료라고도 불린다. 측정 대상이 가지는 속성에 따라 순위를 부여하며 순위 사이의 간격이 동일한 자료를 뜻한다. 구간 자료에서 0은 없다는 뜻이 아니며 임의의 기준에 의하여 정해진 것이다. (예, 온도) 구간 자료에서 덧셈, 뺄셈

은 의미가 있으나 비율을 나타내는 곱셈, 나눗셈의 계산은 의미가 없다.

2.4.4. 비율자료(ratio data)

비율 자료는 구간 자료의 특징을 모두 지니며 또한 측정하고자 하는 속성이 전혀 존재하지 않는 상태인 절대영점을 지니므로 측정치 사이의 비율 계산이 가능하다. (예, 체중, 키)

3. 통계분석 방법

통계학은 크게 기술통계학(descriptive statistics)과 추측통계학(statistical inference)으로 나눌 수 있다.

3.1. 기술통계학(descriptive statistics)

자료를 수집, 분류, 정리하거나 모집단이나 표본자료를 전시하는 것을 말한다. 기술통계에는 자료의 분포 상태를 나타내는 빈도분석과 분포의 특성을 나타내는 평균, 중앙값, 최빈값, 분산, 표준편차, 범위, 왜도, 첨도 등이 속한다.

3.2. 추측통계학(inferential statistics)

표본에 내포된 정보를 바탕으로 모집단의 특성을 추리하는 것을 말한다. 추측통계에는 적합도검정, 비교통계, 상관분석, 회귀분석 등이 속한다.

3.3. 통계적 가설검정

3.3.1. 통계적 가설

통계적 조사나 실험을 하는 경우 연구 대상인 모집단의 특성에 관한 가정을 통계적 가설이라 한다. [예] 항암제 A와 B의 약효는 A 약이 B 약보다 좋다.
통계적가설에는 귀무가설(null hypothesis)과 대립가설이 있다.

① 귀무가설(null hypothesis)
검정하고자 하는 가설을 말한다. 귀무(null)이란 “없다”는 뜻이며 이는 “차이가 없다”는 의미이다. 귀무가설은 영가설이라고도 하며 기호 H_0으로 표기한다. 귀무가설은 연구자의 주장과 반대되는 가설이다. 귀무가설은 부정하려고 세운 것이며 귀무가설을 기각(부정)하게 되면 대립(연구)가설을 채택하게 된다.

② 대립가설(alternate hypothesis)
귀무가설을 반대하는 가설을 말한다. 대립가설은 연구가설이라고도 하며 연구자의 예상 또는 주장에 대한 설명이다. 대립가설은 기호 H1으로 표기한다.

3.3.2. 가설검정

통계적 가설검정이란 통계적 가설을 검정하는 절차를 뜻한다.
가설검정의 절차는 6 단계로 나눌 수 있다.

① 가설(statistical hypothesis) 설정
② 통계적 검정법 선정
③ 유의수준(level of significance) 결정
④ 표본자료에 의한 검정통계량(test statistic) 산출
⑤ 기각역(rejection region) 설정
⑥ 통계적 의사결정

연구에서 귀무가설이란 연구자의 주장에 반대되는 가설이며 연구자는 귀무가설을 기각함으로써 연구자의 주장이 옳다는 결론을 얻게 된다. 실제 거의 모든 연구는 표본을 사용하므로 표본추출로 인해 귀무가설이 참인데도 부정하는 경우(제1종 과오)가 확률적으로 발생하게 된다. 제1종 과오를 범할 확률을 검정의 유의수준 혹은 α수준이라 한다. 다시 말하면 유의수준이란 참인 귀무가설을 기각하는 확률을 말한다. 일반적으로 의학통계에서는 가설의 검정시 유의수준으로 5%(0.05)를 사용한다.

① 기각역 설정
기각역이란 귀무가설이 기각되는 영역으로서 검정통계량의 값의 범위로 지정한다. 기각역의 위치는 대립가설에 의하여 결정된다. 기각역의 위치에 따라 양측검정(two-tailed test)과 단측검정(one-tailed test)으로 나누어 진다.

- 양측검정
  대립가설이 차이의 방향(크거나 작음)을 지적하고 있지 않다면 표본분포의 양쪽 끝을 기각역으로 사용하게 된다.

- 단측검정
  대립가설이 차이의 예측되는 방향(크거나 작음)을 지적한다면 표본분포의 한쪽 끝을 기각역으로 사용하게 된다. 단측검정은 기각역의 방향에 따라 두 가지로 나누어 진다.
  우측검정(right-tailed test, upper-tailed test) : 기각역이 오른쪽(큰 쪽)에 있는 경우

좌측검정(left-tailed test, lower-tailed test): 기각역이 왼쪽(작은 쪽)에 있는 경우

② 통계적 유의성(significance)

통계적으로 계산된 통계치의 확률(p)이 귀무가설을 부정하는 확률인 유의수준(α)보다 낮은 경우(귀무가설의 부정영역에 속하면) 통계적 유의성이 있다고 한다. 흔히 통계적 유의성은 확률(probability)을 나타내는 알파벳 소문자 p(p-value)로 표시한다. 유의수준(α)을 0.05로 설정하였다면 p가 0.05보다 작으면 통계적으로 유의하다고 한다.

3.4. 통계 검정법의 선정

통계 분석법은 모수 통계 검정법(parametric statistical analysis)과 비모수 통계 검정법(nonparametric statistical analysis)으로 나누어진다. 모수 통계 검정법은 모집단의 모수에 대한 강한 가정을 전제로 하며 비모수통계 검정법은 모집단에 대한 가정이 모수통계 검정법에 비하여 약하다. 비교통계에 관한 검정법에 대해서만 알아보기로 한다.

3.4.1. 모수통계 검정법

모집단의 모수에 대한 추측을 하는 통계적 검정법이다. 따라서 모집단의 분포 및 모수에 대한 전제조건이 필요하다.

① 모수통계 검정법의 종류

z 검정, Student t 검정, F 검정, 분산분석(ANOVA) 등

② 전제조건

- 모집단이 정규분포를 한다고 가정한다.
- 집단내의 분산은 같아야 한다.
- 변수의 측정치는 등간형이거나 비율형이어야 한다.

③ 장점

모집단에 대한 가정이 충족되는 경우 비모수 통계 분석에 비하여 검정력이 크다.

④ 단점

- 모수에 관한 조건을 전제로 하는데 이에 대한 검정은 일반적으로 통계적 분석 과정으로는 불가능하다.
- 등간형이나 비율형 자료에만 분석이 가능하다.

3.4.2. 비모수통계 검정법

① 비모수통계 검정법의 종류

카이자승(chi-square) 검정
콜모고로프-스머노프(Kolmogorov-Smirnov) 검정
런(runs) 검정
이항(binominal) 검정
피셔(Fisher) 검정
중앙값(Median) 검정
맨-휘트니(Mann-Whitney) U 검정
왈드-울포위츠 런 검정(Wald-Wolfowitz runs test)
모우지스(Moses) 검정
무작위검정(randomization test)
맥네머(McNemar) 검정
부호검정(sign test)
윌콕슨(Wilcoxon) 부호순위 검정
왈쉬(Walsh) 검정
코크란(Cochran) Q 검정
프리드만(Friedman) 검정
크루스칼-왈리스(Kruskal-Wallis) 검정

② 사용되는 경우

- 모수통계에서 필요로 하는 전제조건을 충족시키지 못하는 경우
- 순위로만 이루어진 자료
- 모수가 관여되지 않는 분석 검정

③ 장점

- 모집단에서의 분포의 특성을 알 수 없을 때 특히 모집단의 분포를 정규분포로 가정할 수 없을 때에도 사용할 수 있다.
- 자료의 형태가 모수적 통계 분석이 가능한 등간형, 비율형뿐만 아니라 명목형, 순위형 자료인 경우도 검정할 수 있다.
- 표본의 수가 작은 경우(6 이하)에 모집단 분포의 성격을 정확히 몰라도 사용할 수 있다.

④ 단점

- 측정자료의 순위를 위주로 통계적 검정을 하므로 통계분석시 측정치의 크기가 고려되지 않아 자료에 대한 유용한 정보를 충분히 활용하지 못한다.
- 자료의 수가 많은 경우 모수적 통계에 비하여 오히려 계산 절차가 복잡하다.

3.4.3. 적절한 통계 분석법의 선정

통계 분석법은 일반적으로 다음 4가지 사항에 의하여 선정된다.

① 분석목적

② 분석하고자 하는 변수의 수
③ 표본 집단의 수
④ 측정된 자료의 형태

비교통계 분석은 분석하고자 할 변수의 수, 표본 집단의 수, 변수의 형태에 따라 다음과 같이 요약할 수 있다.

① 2개의 표본

- 표본이 독립적일 때

자료의 형태	분 석 방 법
명목자료	카이자승(chi-square) 검정
	피셔(Fisher) 검정
순위자료	맨-휘트니(Mann-Whitney) U 검정
등간/비율자료	Student t 검정

- 표본이 관련되어 있을 때

자료의 형태	분 석 방 법
명목자료	맥네머(McNemar) 검정
순위자료	윌콕슨(Wilcoxon) 부호순위 검정
등간/비율자료	paired t 검정

② 3개 이상의 표본

- 표본이 독립적일 때

자료의 형태	분 석 방 법
명목자료	카이자승(chi-square) 검정
순위자료	중앙값(Median) 검정
	크루스칼-왈리스(Kruskal-Wallis)
등간/비율자료	분산분석(ANOVA)

• 표본이 관련되어 있을 때

자료의 형태	분 석 방 법
명목자료	코크란(Cochran) Q 검정
순위자료	프리드만(Friedman) 검정
등간/비율자료	분산분석(ANOVA)

어떠한 통계적 검정법을 선정하여 통계처리시 이에 대한 정확한 이해가 없으면 통계적 오류를 범할 수 있다.

다음은 χ2 검정법에 대한 전제조건들이다.

① 측정치들이 독립적이어야 한다.
② 자유도가 1일때는 (2x2 분할표)
  • 총 표본이 20이상이어야하며
  • 표본수가 40미만인 경우 기대빈도가 5이상이어야 한다.
  • 표본수가 40이상인 경우 기대빈도가 1이상이어야 한다.
③ 자유도가 2이상일 때는 (r x c 분할표, 자유도=(r−1)(c−1)≥2)
  범주의 20%이상에서 기대빈도가 5미만이어서는 안되며 1보다
  작은 기대빈도가 있어서도 안된다.

다음은 chi-square 분석의 전제조건이 만족되지 않는 자료에 대하여 잘못된 결론을 내린 예로 의학학술지에 기재된 내용을 일부 변경하였다.

[예] chi-square 분석을 잘못 적용한 예

난소암과 난소낭종을 대상으로 CA125의 양성 유무에 따라 다음과 같은 교차분할표를 작성하여 통계패키지인 SPSS/PC를 사용하여 통계처리한 결과는 다음과 같다.

	난소암	난소낭종
CA125 양성	6	1
CA125 음성	12	18

[SPSS/PC 통계 처리 결과]

Chi-Square	D.F.	Significance	Min E.F.	Cells with E.F.<5
3.09428	1	0.07857	3.405	2
4.74787	1	0.02933	(Before Yates Correction)	

[해설]
Min E.F. : 최소 기대빈도(Minimum Expected Frequency)
Cells with E.F.< 5 : 기대빈도가 5미만인 난이 2(2/4 = 50%).

위의 계산 결과에서 Significance 만으로 통계적 해석을 내리게 되면 Yates 연속성 수정 후의 χ2 (3.09428)에 대한 확률이 0.07857로 0.05보다 크므로 통계적 유의성이 없다. 즉 "CA125 양성유무는 난소암과 난소낭종의 구분과 무관하다."고 결론을 내린다. 2x2 교차분할표에서 chi-square 검정법의 전제조건에서 기대빈도가 5미만인 난이 없어야 하는데 예제에서는 2개의 난이 기대빈도가 5미만이므로 chi-square 검정법을 적용할 수 없다. 이 경우 Fisher's exact test를 적용하여야 하며 two-tail probability가 0.04224 < 0.05로 통계적 유의성이 있다고 결론을 내린다.

4. 통계분석 방법의 분류

의학연구의 종류를 원저, 증례, 종설, 기타 4가지로 분류할 수 있다. 원저인 논문에서 통계분석을 사용하였던 경우에 통계기법을 기술통계(descriptive statistics)와 추측통계(interferential statistics)로 나눈다. 기술통계는 자료의 특성을 요약하기 위한 분석방법으로 백분율, 대표값(평균, 중앙값, 최빈값), 산포도(표준편차, 범위)의 사용을 포함한다. 추측통계는 Emerson과 Colditz가 제시한 분류를 기준으로 하였다 (Table1).

Table 1. Classification of Statistical Methods

0. Z-test
1. Student's t-test
2. paired t-test
3. ANOVA
    (1) one way / two-way / multi-way
    (2) repeated measures
4. multiple comparisons
    (1) Bonferroni (2) Scheffe (3) Duncan (4) other
5. Pearson correlation analysis
6. simple linear regression
7. contingency tables
(1) chi-square test (2) Fisher's exact test
(3) McNemar's test (4) other
8. survival analysis
(1) life table method (2) Kaplan-Meier method
(3) log-rank test (4) Cox regression
(5) other
9. epidemiological statistics
(1) sensitivity/specificity, odds ratio, relative risk
    (2) other
10. nonparametric test
(1) Mann-Whitney U test (Wilcoxon rank sum test)
    (2) Wilcoxon signed rank test
    (3) Kruskal-Wallis test
    (4) test for trend
    (5) Spearman rank correlation
    (6) other
11. multivariate analysis
(1) multiple regression
    (2) logistic regression
    (3) other
12. others

5. 의학학술지 통계적 평가기준

국제의학학술지 편집인위원회에서 제시한 통계적 서술 통일양식에 근거하였으며 British Medical Journal에서 사용하였던 통계평가 점검표를 참고로 하여 새로운 점검표를 작성하였다(Table 2). 추측통계분석을 시행한 논문들을 대상으로 통계분석

방법의 적용 타당성, 통계학적 방법론, 통계분석 결과의 서술에 대한 평가를 한다.

Table 2. Check Lists in Assessing the Statistical Contents

연구종류

1) 원저 2) 증례 3) 종설 4) 기타

통계기법 종류

1) 추측통계 2) 기술통계 3) 사용안함 4) 기타

통계기법 적용

1. 기술통계

1) 대표값 (1) 적절 (2) 부적절 (3) 사용안함
2) 산포도 (1) 적절 (2) 부적절 (3) 사용안함

2. 추측통계

1) 방법
2) 적정성 (1) 적절 (2) 부적절
   통계적 가정
      (1) 정규성 (2) 등분산성 (3) 독립성
      (4) 표본크기 (5) 기대도수 (6) 기타
   통계적절차
      (1) 사후검정 (2) 기타

3) 방법론 (기법의 서술)
      (1) 상세 기술 (2) 유의수준 (신뢰구간) (3) 통계프로그램
      (4) 양측/단측검정 (5) 기타
4) 분석결과
      (1) p-값 (2) 신뢰구간 (3) 통계량, 자유도 (4) 해석 오류
      (5) 계산 오류 (6) 비교군 언급 (7) 연구가설과 일치
      (8) 한계점 (9) 인용문헌의 통계적서술 (10) 기타
5) 기타
      (1) 통계용어
      (2) 보조적 분석기법 (Table, Chart,…)
      (3) 통계학적 방법론에 관한 참고문헌

6. 흔한 통계분석 방법

Student's t-test, chi-square test, Fisher's exact test, one-way ANOVA test, 비모수통계분석 등이 많은 의학학술지에서 사용되었다.

7. 흔한 통계적 오류

7.1. 통계적 분석기법

기술통계기법에서 대표값과 산포도 사용이 부적절한 경우가 많았다.
추측통계기법 중 흔한 분석방법 선택의 오류는 Student's t-test, chi-square test, Fisher's exact test, one-way ANOVA test, 비모수통계분석 등 이었다. 논문에서 사용된 통계분석방법 이외에 추가적으로 요구되는 추측통계기법으로는 비모수통계분석이 많았으며 그 다음 다중비교분석, chi-square test, Fisher's exact test 등이다.
추측통계분석에 요구되는 가정에 대한 검정은 정규성, 등분산성, 기대도수, 표본크기, 자료의 독립성, 기타 회귀분석과 관련된 다중공선성, 잔차분석 등 이며 거의 대부분 의학학술지 논문에서 시행되지 않고 있다.

7.2. 통계학적 방법론

통계학적 방법론의 기술에서 가장 많이 기술하지 않는 항목은 양측/단측 검정이며 상세기술, 사용한 통계프로그램, 유의수준에 대한 서술 등이다.

7.3. 통계분석 결과의 서술

분석결과에 대한 서술에 대한 평가항목으로 인용문헌의 통계적 서술이 불충분한 논문이 가장 많으며 p-value의 부정확한 기술, 신뢰구간 제시, 통계량, 자유도에 대한 기술이 없거나 통계분석 결과에 대한 해석오류 등이다.

7.4. 기타

통계학 용어의 잘못된 표기, Table, Chart 등 보조기법 사용시 문제점, 통계학적 방법론에 관한 참고문헌이 필요한 논문에 참고문헌을 인용하지 않은 경우 등이다.

참고문헌

1. Gardner MJ, Altman DG, Jones DR, Machin D. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? Br Med J 1983;286:1485-8.
2. Shen J, Gou L, Tang J. Analysis of common on statistical errors in biomedical journals. Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi. 2007;21:541-3.
3. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. BMJ 1986;292:810-2.
4. Garcia-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Medical Research Methodology. 2004;4:1-5.
5. 박일규, 강정옥, 김신규, 금동극. 대한임상병리학회지에서 사용된 통계기법의 평가. 대한임상병리학회지. 1999;19:460-4.
6. 고흥, 곽일용, 김광우, 함병문, 최익현. 대한마취과학회지에 게재된 논문의 통계적 분석에 관한 고찰(1981~1990년). 대한마취과학회지. 1993;26:22-7.
7. 송현, 박계현, 김웅한, 전태국. 대한흉부외과학회지에 게재된 통계적 분석에 관한 고찰. 대한흉부외과학회지. 1994;27:732-7.