



간호연구방법론

- 표본 수 산정과 연속자료 분석법 -

김 호(서울대학교 보건대학원 교수)

모집단은 연구자가 최종적으로 관심을 가지는 집단을 말하며 표본은 모집단에서 통계적 결정을 하기 위하여 모집단으로부터 대표성 있게 뽑은 집단을 뜻한다. 표본은 대표성 있게 모집단을 반영하여야 한다.

p-value는 두 집단의 평균이 같다고 가정했을 때 우리의 자료 혹은 더 차이가 나는 자료를 얻을 확률을 말한다.

작은 p-value는 위에서 말한 확률이 작은 것을 뜻하며, 이는 통계적으로 가능하지 않은 일이 일어났다, 두 집단의 평균이 같다는 가정에 문제가 있다, 두 집단의 평균은 같지 않다고 결론 내린다고 말할 수 있다.

작지 않은 p-value는 두 집단의 평균이 같다고 가정하면 우리의 자료를 관측할 확률이 작지 않다는 것을 뜻하며, 이는 두 집단의 평균이 같다는 가정에 문제가 없다고 말할 수 있다.

A는 얻은 자료를 뜻하고 B가 연구가설일 경우 귀무가설(-B)은 두집단에 차이가 없는 것이고(Ho) 대립가설(B)는 두 집단에 차이가 있는 것을 뜻한다(Ha). 또한 일종의 오류는 옳은 귀무가설을 기각할 확률(α)을 말하며 이종의 오류는 틀린 귀무가설을 기각할 확률(β)이다. Power=1- β 로 있는 차이를 발견할 확률을 말한다

검정방법 A(모수적인 방법, 검정력=90%)는 검정방법 B(비모수적인 방법, 검정력=70%)보다 더 큰 검정력을 주었다는 말은 실제로 차이가 있을 때 A 방법을 100번 실시해서 90번의 경우 차이가 있다고(귀무가설 기각) 결정하였고, B 방법을 실시하였을 때는 100번 중 70번 귀무가설을 기각한 것을 뜻한다. 이 경우 A가 더 좋은 방법이며 검정력은 표본의 크기에 의존한다.

1. 표본수 산정

전통적 통계적 가설의 유의성은 표본수에 크게 의존하므로 표본수(실험의 비용)와 통계적 유의성(실험의 효용성)의 균형을 맞추는 것이 요구된다. 최소의 비용으로 효과를 증명하는것이 중요하다.

단순임의 추출(simple random sampling)에서 N:모집단의 크기, n이 표본의 크기라면

$$n = \frac{Npq}{(N-1) + pq}$$

연구대상수의 계산 시 고려할 사항은 어떠한 연구설계 인



가 (연구설계의 형태), 처치 후 결과는 언제 나타나는가? (처치효과의 지연효과), 처치 결과는 이산형인가 연속형인가? (결과의 척도), 연구기간을 얼마나 길게 할 것인가? (예상되는 추적시간), 중도에 탈락하는 사람은 얼마나 될 것인가? (예상되는 추적손실), 연구자의 지시대로 치료를 잘 받는가? (예상되는 처치의 비 순응률), 할당비는 얼마로 할 것인가? (시험군:비교군), 시험 집단의 수는 몇 개로 할 것인가? 이다.

이분형 결과에 대한 비교 (비율을 비교하는 경우)

$$n = \frac{(Z_{\alpha}\sqrt{2pq} + Z_{\beta}\sqrt{(P_1Q_1 + P_2Q_2)})^2}{d^2}$$

P1, P2, α , β 만 알면 두 집단의 비율 비교시의 표본수표를 이용할 수 있다.

연속형 변수의 비교는 두 집단의 평균값이 다르다는 것 말하는 것이 목적이며 양쪽집단의 평균, 분산만 알면 계산이 가능하다.

$$n_c = \frac{2(Z_{\alpha} + Z_{\beta})^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

연구대상수를 최소화하는 방법은 연속변수를 사용하거나 변수를 보다 정밀하게 측정하고 짝 측정치를 사용한다. 또한 이형 할당법을 사용하고 보다 일반적인 결과를 사용하는 방법이 있다.

2. 연속형 변수 자료분석

모수적 방법과 비모수적 방법에서 중앙값(median)은 평균에 비하여 이상치에 대하여 둔감(robust)하다.

자료의 정규성 분포가정을 하면 평균과 분산을 통하여 모



집단의 성질을 완전히 파악할 수 있다(모수적 방법).

비모수적 방법은 자료의 정규성 분포가정을 하지 않는다. 자료의 평균과 분산이 아닌 순위를 이용한 방법을 사용한다. 자료의 분포가정(eg 정규성)이 만족되면 효율이 떨어진다.

• t-test(연속변수의 두 집단 평균비교)
관심변수가 연속일 때(정규분포를 따를 때) 두 집단 간의 평균의 차이를 보는 검정, 즉 두 개의 독립적인 집단 간의 차이를 말한다.

Paired(짝지은) t-test는 한 개체에서 짝 지은 관찰치들의 동질성을 볼 때 즉 처치전의 값과 후의 값을 비교할 때 사용한다.

표본수가 적은 경우에는 정규분포 가정을 확인하기가 곤란하므로 비모수적 방법을 사용하고 두개 이상의 집단 혹은 다른 변수로 보정을 할 때는 ANOVA(분산분석)를 사용한다.

• 회귀분석(Regression Analysis)
: 연속형 설명변수가 연속형 종속변수에 미치는 영향을 분석한다

단순회귀와 중회귀에서 회귀계수들의 의미차이는 단순회

귀는 X 1단위 증가시 Y 증가분의 기대치를 말하며 중회귀는 다른 X들이 일정한 값으로 남아 있을 때 관심 X가 1단위 증가시 Y의 기대치의 증가분을 뜻한다.

한편 $Y = \beta_0 + \beta_1 * X_1 + \epsilon$ 이라면

β_1 은 X_2 와는 아무런 관계없이 X_1, Y 의 그림에서 기울기이다.

만약 β_1 과 β_1^* 이 다르다면 X_1 의 효과를 보는데 있어서 X_2 를 고려하느냐 마느냐 하는것에 따라서 결론이 다르게 된다. 이러한 경우 X_2 를 혼란변수(confounder)라고 한다. 혼란변수를 고려하지 않은 모형에서의 결론은 올바른 결론 이라 할 수 없다. 그러므로 연구설계시부터 혼란변수로 작용할 수 있는 모든 변수를 고려해야 한다.

중회귀방정식을 위해 독립변수를 선정하는 방법은 변수 중에서 중요도에 따라 한 개씩 뽑아내는 Forward selection 과 안 좋은 것부터 뽑아내어 더 이상 없을 때 까지 뽑아내는 Backward elimination, 결과에 따라 변수를 넣고 빼고를 반복하면서 기준에 따라 변수를 조정하는 Stepwise selection 법이 있다.

* 주) 상기내용은 2009년 통계학술대회의 강의내용은 요약한 것임(요약: 이순희 교수)