고급통계적 방법론

2장: Frequentist Inference

Lecturer: 장원철 *2023* 가을학기

컴퓨터 시대가 도래하기전 전자계산기의 시대가 있었으며 (컴퓨터공학과의 예전이름은 전자계산기공학과였다!) 빅데이터 이전에 스몰데이터가 있었다. 스몰데이터는 주로 제한된 실험 환경하에서 개별 과학자의 정성들인 노력으로 얻어진 수백개 이내의 자료들로서 아주 효율적인 분석이 요구된다. Pearson, Fisher, Neyman, Hotelling 과 같은 통계학자들은 19세기 초부터 이러한 자료분석을 위한 통계이론을 개발하였으며 이런 이론에 기반한 분석방법들은 탁상용 기계계산기를 이용해서 실행이 가능하였다. 이 이론들은 빈도주의 추론(frequentist inference) 라고 불리었으며 20세기 통계학 분야의 주요이론으로 자리잡았다. 이 장에서는 빈도주의에 대해 간단히 소개한 후 어떻게 데이터 분석에 적용되는지 살펴본다.

그림 2.1은 스탠포드 대학 신장내과 연구실에 211명의 신장 환자의 GFR(Glomerular Filtration Rate: 사구체 여과율)를 측정한 자료를 보여주고 있다. 사구체 여과율은 신장이 1분 동안에 깨끗하게 걸러주는 혈액의 양을 의미하며 정상 범위는 분당 $90 \sim 102$ ml 정도이다. 나이가 증가하면 GFR은 일반적으로 감소하며 45 ml이하일 경우 신장기능에 심각한 문제가 있다고 간주한다. 이 자료의 경우 환자들의 사구체 여과율 평균은 54.25, 표준오차의 추정치($=s/\sqrt{n}$)는 0.95이다.

이런 경우 주로 데이터의 대푯값을 54.25 ± 0.95와 같은 방식으로 표기하는데 여기서 ± 0.95는 표본평균의 불확실성을 나타내는 빈도주의 추론의 방식이다. 따라서 대푯값인 54.25의 경우 4라는 숫자 자체도 정확하지 않을

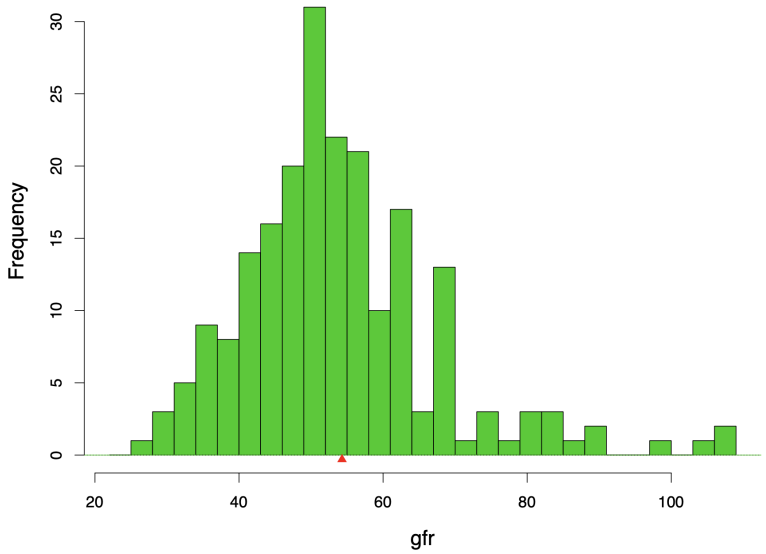

그림 2.1 211명의 신장환자의 GFR(Glomerular Filtration Rate: 사구체 여과율)자료. 평균은 54.25이고 표준편차는 0.95

수 있다는 걸 내포한다는 걸 명심하자.

일반적 통계적 추론은 관측 자료가 특정확률모형에서 생성되었다는 가정을 한다. 위의 예제의 경우 $n = 211$ 개의 gfr 관측치를 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$와 같이 하나의 벡터로 나타낼 수 있다. 이 자료의 생성과정을 확률모형을 통해 설명해보자. 먼저 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$가 확률분포 F부터 생성된 n의 서로 독립인 확률변수로 구성된 벡터라고 한다면 다음과 같이 표현할 수 있다.

$$F \to \mathbf{X}$$

여기서 F는 가능한 gfr 점수들의 분포를 나타내는 함수라고 간주할 수 있다. 실제로 확률변수 X가 x라는 값을 가지는 관측치가 될 경우 우리는 이러한 관측치를 바탕으로 F에 대한 추론을 하고자 한다. 만약 우리가 알고 싶은 것이 X의 평균이라고 한다면 다음과 같이 표현할 수 있다.

$$\theta = \mathbb{E}_F(\mathbf{X}).$$

물론 가장 자연스럽게 생각할 수 있는 θ의 추정치는 표본평균이다. 즉 $\widehat{\theta} = \bar{x}$. 만약 표본크기 n이 무한히 크다면 (예를 들면 $n = 10^{10}$) $\widehat{\theta}$는 θ의 값과 거의 동일할 것이다. 하지만 일반적으로 표본의 크기는 그 정도로 크지 않기 때문에 θ의 추정치에는 어느정도 불확실성이 내포되어 있고 이러한 불확실성이 어느 정도 있는냐는 추론의 문제라고 할 수 있다. $\widehat{\theta}$은 알려진 알고리즘에 의해서 데이터 $\mathbf{x}$로 부터 계산할 수 있으며 다음과 같이 $\mathbf{x}$의 함수로 표현할 수 있다.

$$\widehat{\theta} = t(\mathbf{x}).$$

위의 예제의 경우 $t(\mathbf{x}) = \bar{x} = \sum_{i=1}^{n} x_i/n$라고 할 수 있으며 추정량 (estimator) $\widehat{\Theta} = t(\mathbf{X})$에 실제 데이터 값을 넣어서 계산한 추정치 (estimate)로 간주할 수 있다.

여기서 빈도주의 추론의 첫번째 정의를 소개하자. 추정치의 정확도는 추정량의 확률적 정확도를 나타낸다. 이 정의는 마치 동의어 반복과 같은 느낌이 들지만 추정치는 숫자를 나타내고 추정량은 확률변수라는 것에 주목한다면 수긍이 가는 정의이다. 즉 추정량의 경우 다양한 값을 가질 수 있다는 점을 고려한다면 그 값들의 분포가 결국 정확도의 측도가 될 수 있다.

편의 (bias)와 분산 (variance)는 빈도주의 추론에서 자주 등장하는 중요한 개념이다. μ를 $\widehat{\Theta} = t(\mathbf{X})$의 기댓값이라고 하자.

$$\mu = \mathbb{E}_F(\widehat{\Theta})$$

이 경우 추정치 $\widehat{\theta}$의 bias와 variance는 다음과 같이 주어진다.

$$\text{bias} = \mathbb{E}_F(\widehat{\Theta}) - \theta = \mu - \theta, \quad \text{var} = \mathbb{E}_F\left[(\widehat{\Theta} - \mu)^2\right].$$

빈도주의는 종종 향후 시행에서 발생하는 무한개의 순열을 이용하여 정의된다. 예를 들자면 가상데이터 $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \ldots$이 있다고 상상한다면 여기에 대응되는 $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}, \ldots$을 만들수 있고 이 값들을 바탕으로 계산된 분산을 이용하여 $\widehat{\theta}$의 표준오차를 계산할 수 있을 것이다.

2.1 Frequentism in Practice

빈도주의 추론을 실제 상황에서 어떻게 사용하지 알아보자.

1. The plug-in principle. 표본평균의 표준오차는 다음과 같이 주어진다.

$$\text{se} = [\text{var}_F(X)/n]^{1/2}$$

여기서 주어진 관측치 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$을 이용하여 X의 분산을 표본분산을 이용하여 추정할 수 있다.

$$\widehat{\text{var}}_F = \sum (x_i - \bar{x})^2/(n-1)$$

위의 표본분산 공식을 표준오차 공식에 대입하면

$$\widehat{\text{se}} = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2/(n(n-1))\right]^{1/2}.$$

Plug-in principle은 말 그대로 평균의 불확실성을 나타내는 표준오차를 데이터에서 직접 추정하는 것을 의미한다.

2. Taylor-series expansion. 만약 통계량이 복잡한 형태 (예를 들자면 표본표준편차)을 띠고 있다면 delta method를 사용할 수 있다. 만약 알고자 하는 통계량이 $s(\widehat{\theta})$와 같이 우리가 잘 알고 있는 $\widehat{\theta}$의 함수로 표현될 수 있다면 테일러 전개를 통하여 주어진 통계량을 다음과 같이 표현할 수 있다.

$$\text{var}[s(\widehat{\theta})] \approx \text{var}[s(\theta) + s'(\theta)(\widehat{\theta} - \theta)] = [s'(\theta)]^2\text{var}(\widehat{\theta})$$

여기에 다시 위의 plug-in 방법을 사용한다면 최종 추정치는 $[s'(\widehat{\theta})]^2\widehat{\text{var}}(\widehat{\theta})$이 된다.

3. Parametric families and maximum likelihood theory.

이 내용은 4장에서 자세히 다루도록 한다.

4. Simulation and the bootstrap. 빈도주의의 가장 중요한 가정은 미래 시행을 통한 무한개의 자료를 관측한다는 걸 점이다. 만약 우리가 실제로 무한개의 자료를 생성할 수 있다면 빈도주의 추론의 문제는 아주 간단하게 해결될 것이다. 즉 F에서 무한개의 자료를 생성해서 거기에 대응하는 무한개의 통계량 $\widehat{\Theta}$를 생성한다면 우리가 궁금해하는 추론의 모든 문제는 사라진다. 하지만 현실은 F를 모르기 때문에 이 방법은 가능하지 않다. Bootstrap의 기본 아이디어는 $\widehat{F}$에서 자료를 생성해서 위의 절차를 되풀이하는 것이다. 여기서 $\widehat{F}$는 다양한 추정치를 사용할 수 있지만 일반적으로 nonparametric MLE인 emprical cdf를 주로 사용하고 이런 경우를 nonparametric bootstrap이라고 부른다. 앞의 세가지 방법들은 통계량들이 "smooth"한 함수인 경우에 사용될 수 있지만 bootstrap의 경우 이러한 제약이 없다.

표 2.1은 bootstrap을 사용하여 gfr의 위치모수를 다양한 통계량으로 추정한 것이다. 여기서 winsorized mean은 절삭평균(trimmed mean)을 의미하며 중앙값과 더불어 smooth하지 않는 (즉 미분가능하지 않는

표 2.1: qfr 데이터에 대한 3가지 대푯값 추정치와 추정치의 표준오차; 중앙값과 절삭평균의 표준오차는 bootstrap($B = 1000$)을 사용하여 구하였음.

	Estimate	Standard error.
mean	54.25	.95
25% WInsorized mean	52.61	.78
median	52.24	.87

점이 존재) 통계량이다. 이 2개의 통계량은 M-estimator의 일종으로 사실 점근분포 (asymptotic distribution)을 구하는 것은 어렵지 않다. 여기서는 bootstrap을 사용하여 표준오차를 구하였다.

5. Pivotal statistics. Pivotal statistics은 통계량의 분포가 F에 의존하지 않는 통계량을 의미한다. 이 경우 $\widehat{\Theta} = t(\mathbf{X})$의 이론적 분포를 바로 $\widehat{\theta}$에 적용할 수 있다. 우리에게 친숙한 2표본 t-검정을 통해서 pivotal statisitcs에 대해 알아보자.

2개의 모집단으로 부터 다음과 같은 자료를 관측한다고 가정하자.

$$\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1}) \quad \mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2}).$$

여기서 귀무가설은 두 모집단이 같은 분포를 가진다는 것 ($F_1 = F_2 = F$)이며 대립가설은 두번째 집단이 상대적으로 값들이 크다 (백혈병 환자 자료에서 AML 환자들의 유전자 발현정도가 ALL 환자보다 높았던 경우를 생각해보자)고 생각하자. 문제를 단순화 하기위해 우리는 두개의 모집단이 모두 정규분포를 따르고 같은 모분산 값을 가진다고 하자. 즉

$$X_{1i} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \quad i = 1, 2, \ldots, n_1,$$

그리고

$$X_{2i} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2), \quad i = 1, 2, \ldots, n_2.$$

이 경우 귀무가설은 결국 $H_0 : \mu_1 = \mu_2$이며 가장 쉽게 생각할 수 있는 검정통계량은 $\widehat{\theta} = \bar{x}_2 - \bar{x}_1$이다. 귀무가설하에서

$$\widehat{\theta} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

여기서 plug-in 방법을 사용하기 위해 σ^2의 다음과 같은 추정치를 사용할 수 있다.

$$\widehat{\sigma}^2 = \left[\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{21} - \bar{x}_2)^2\right] \Big/ (n_1 + n_2 - 2)$$

하지만 Student (William Gosset)이 다음과 같은 검정통계량을 사용할 경우 검정통계량의 분포는 F(정확히 이야기 하자면 σ)에 의존하지 않는다는 것을 증명하였다.

$$t = \frac{\bar{x}_2. - \bar{x}_1}{\widehat{\text{sd}}}, \quad \text{where } \widehat{\text{sd}} = \widehat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}$$

귀무가설하에서 위의 검정통계량은 자유도가 $n_1 + n_2 - 2$인 t-분포를 따른다. 따라서

$$\Pr_{H_0}(-1.99 \leq t \leq 1.99) = 0.95$$

이 경우 $\theta = \mu_1 - \mu_2$의 95% 신뢰구간은 $\bar{x}_2 - \bar{x}_1 \pm 1.99 \cdot \widehat{\text{sd}}$로 주어진다.,

2.2 Frequentist Optimality

빈도주의 추론이 인기있는 이유중 하나로 상대적으로 간단한 모형가정을 들 수 있다. 즉 데이터를 생성하는 확률모형 F와 추정량 $t(\mathbf{X})$를 계산하는 알고리즘이 우리가 필요한 전부이다. 하지만 gfr 예제에서 보듯이 우리는 여러개의 알고리즘(추정치)을 제시할 수 있다. gfr 예제의 경우 위치모수를 추정하고 싶다면 3가지 알고리즘(평균, 절삭평균, 중앙값)중 가장 적합한 알고리즘은 어느 것일까? 만약 가장 적은 표준오차를 기준으로 한다면 25% 절삭평균이 우리의 선택된 알고리즘이 될 수 있을 것이다.

그렇다면 빈도주의 추론에서 어떤 기준에서 최선의 알고리즘을 선택할까? 이 질문에 대한 대답으로 1920-1935년 사이에 개발된 빈도주의 최적성(optimality)에 관한 두가지 중요한 결과에 대해 알아보자. 첫번째는 모수적 모형에서 제안된 Fisher의 최대우도추정량 (MLE)와 Fisher's information bound이다. 이 결과는 MLE가 (점근) 표준오차를 최소화 하는 추정량이라는 것을 증명하고 있고 점근표준오차를 최적화의 기준으로 한다면 MLE가 최고의 추정량이라고 할 수 있다.

두번째는 가설검정에 관한 중요한 결과로 Neyman-Pearson lemma이다. NP lemma는 우리가 단 두가지 가능한 확률밀도함수만 선택할 경우를 가정한 후 데이터가 귀무가설하에서는 f_0에서, 대립가설하에서는 f_1에서 생성된다고 가정한다. 검정 룰 $t(x)$는 주어진 관측치를 바탕으로 데이터가 귀무가설하에서 생성이 되었다고 생각되면 0, 그렇지 않으면 1을 값을 가진다고 하자. 이 경우 두가지 가능한 오류가 있는데 실제로 귀무가설하에서 데이터가 생성되었는데 대립가설에서 생성되었다고 판단하는 경우를 제 1종의 오류, 그 반대의 경우를 제 2종의 오류라고 부르면 다음과 같이 표현할 수 있다.

$$\alpha = \Pr_{f_0}(t(\mathbf{x}) = 1),$$
$$\beta = \Pr_{f_1}(t(\mathbf{x}) = 0)$$

사실 2 종류의 오류을 동시에 최소화하는 것은 불가능하다. 하지만 주어진 여러가지 검정 룰중 다른 검정 룰과 비교하여 두가지 오류 모두에 대해서 항상 작은 값을 가지는 검정 룰을 찾는다면 최적화된 검정 룰이라고 할 수 있을 것이다.

이러한 검정 룰의 찾기위해 먼저 우도비 (likelihood ratio)는 다음과 같이 정의할 수 있다.

$$L(\mathbf{x}) = f_1(\mathbf{x}) / f_0(\mathbf{x})$$

이 경우 검정 룰을 다음과 같이 정의하자.

$$t_c(\mathbf{x}) = \begin{cases} 1, & \text{if } \log L(\mathbf{x}) \geq c \\ 0, & \text{if } \log L(\mathbf{x}) < c. \end{cases} \tag{1}$$

위의 검정 룰은 cutoff c가 주어지면 정확히 (유일하게) 정의된다. NP lemma가 얘기하는 것은 $t_c(\mathbf{x})$가 앞에서와 언급한 것과 같은 기준으로 최적이라는 것이다.

그림 2.2는 (α_c, β_c)를 cutoff c의 함수로 표현한 것이다. 여기서 표본크기는 10이며 f_0는 표준정규분포이고 f_1은 평균이 0.5이고 분산은 1인 정규분포이다. NP lemma가 얘기하는 것은 t_c와 다른 형태의 검정 룰의 경우 (α, β)의 값을 계산한다면 이 곡선위의 어디엔가 위치한다는 것을 의미한다.

$$\alpha_c < \alpha \text{ and } \beta_c < \beta$$

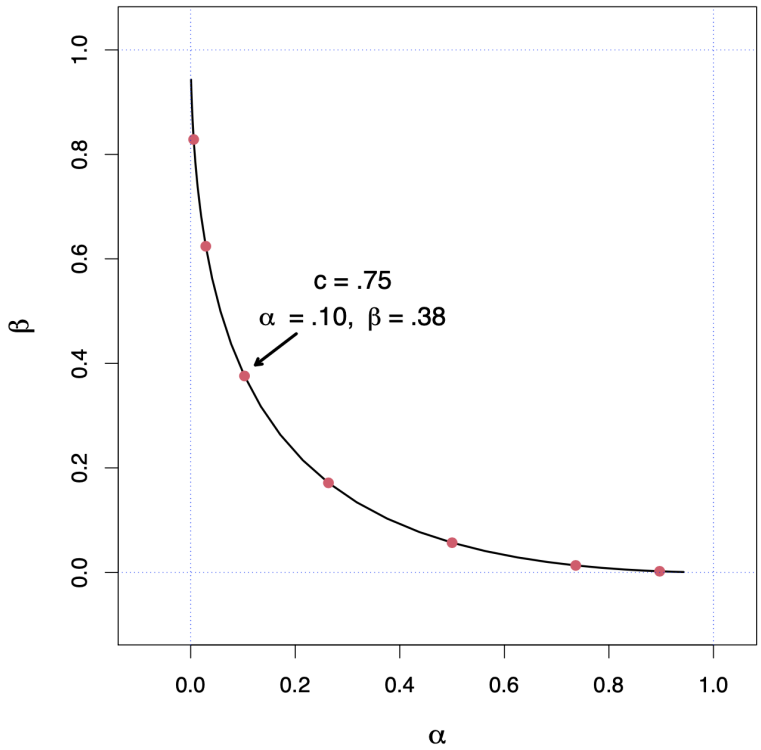

그림 2.2 표본크기가 10이며 $f_0 \sim N(0,1)$, $f_1 \sim N(0.5,1)$인 경우 Neyman-Pearson alpha 곡선. 빨간점은 cutoff 점 $(0.8, 0.6, 0.4 \ldots, -0.4)$를 나타낸다.