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

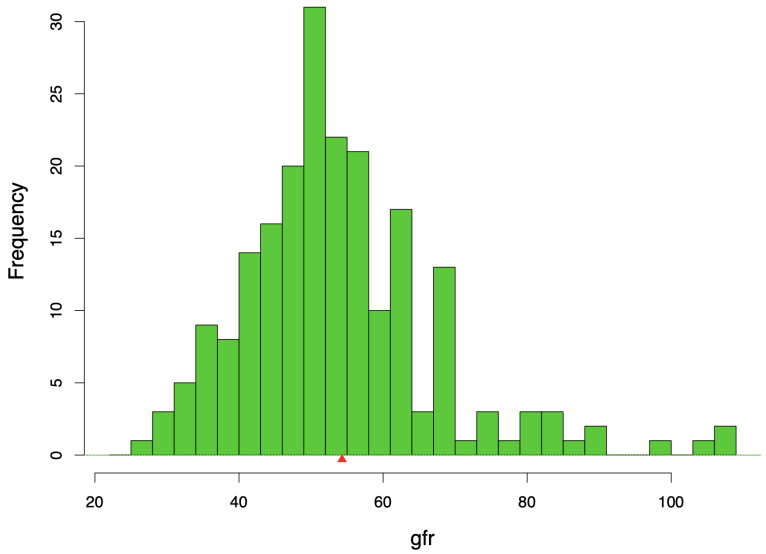

그림 2.1 211명의 신장환자의 GFR(Glomerular Filtration Rate: 사구체 여과율)자료. 평균은 54.25이고 표준편차는 0.95

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. 즉 만약 우리가 가상의 데이터 $\mathbf{X}^{(1)}.\mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \ldots$

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

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

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

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

1. *The plug-in principle.*
2. *Taylor-series expansion.*
3. *Parametric families and maximum likelihood theory.*
4. *Simulation and the bootstrap.*
5. *Pivotal statistics.*

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

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

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

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

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

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

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

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

$$\bar{x}_2 - \bar{x}_1 \pm 1.99 \cdot \widehat{\mathrm{sd}}$$

2.2 Frequentist Optimality

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

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

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

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

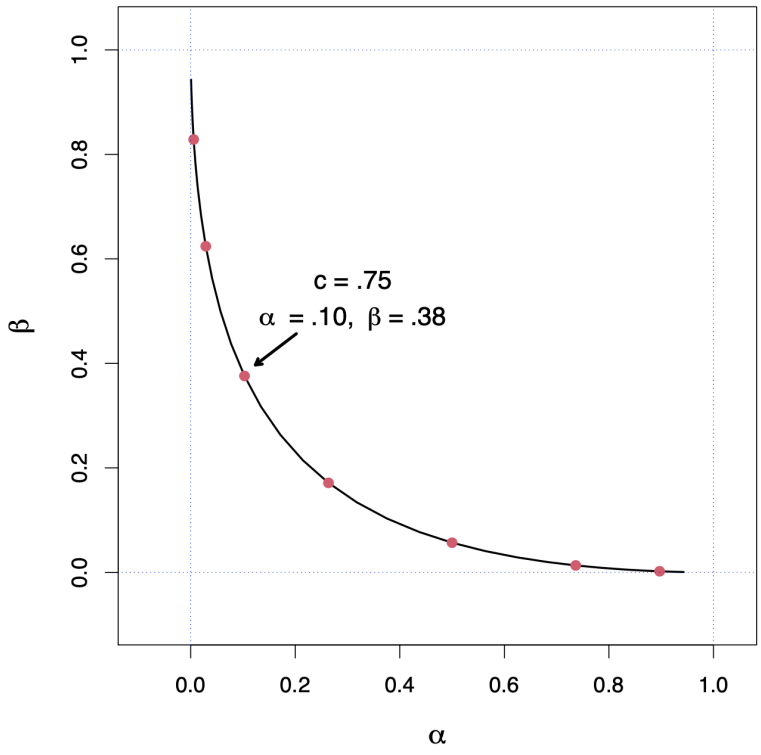

그림 2.2 표본크기가 10이며 $f_0 \sim N(0,1)$, $f_1 \sim N(0.5,1)$인 경우 Neyman-Pearson alpha 곡선. 빨간점은 cutoff 점 $(0.8, 0.6, 0.4 \ldots, -0.4)$를 나타낸다.