

# 추정 이론

# 들어가기

- 통계적 추론(statistical inference) 에서 가장 중요하게 여겨지는 2가지의 분야
  - 추정이론(estimation theory)
  - 검정이론(testing theory)
- 추정이론
  - 연구나 관찰의 대상인 확률변수  $X$ 의 확률분포를  $P_\theta(x)$ ,  $\theta \in \Omega$ 로 표현하자. 미지의 모수  $\theta$ 를 제외하고는 확률분포의 유형이 결정되어 있다고 가정하자. 모수모형(parametric model)이라고 불리는 이런 경우에 모수  $\theta$ 는  $\theta = (\theta_1, \dots, \theta_k)$ 로 표현되는 벡터일 수 있고,  $\Omega$ 는 모수가 가질 수 있는 값의 집합, 즉 모수공간을 뜻한다.

# 들어가기

- 확률분포  $P_\theta(x)$ ,  $\theta \in \Omega$ 로부터 크기가  $n$ 인 표본  $X_1, \dots, X_n$ 을 얻었다고 하자.  $\mathbb{X} = (X_1, \dots, X_n)$ 이다.  
우리가 관심이 있는 문제는 확률분포  $P_\theta(x)$ ,  $\theta \in \Omega$ 를 특징짓는 모수  $\theta$  또는  $g(\theta)$ 의 추정이다.
- $g(\theta)$ 에 대한 추정방법
  - 점추정(point estimation) : 표본에 근거한 통계량  $T(\mathbb{X}) = T(X_1, \dots, X_n)$ 을 사용하여 하나의 값으로  $g(\theta)$ 를 추정하는 방법
  - 구간추정(interval estimation) : 두 개의 통계량  $T_1(\mathbb{X}_1) = T_1(X_1, \dots, X_n)$ 과  $T_2(\mathbb{X}_2) = T_2(X_1, \dots, X_n)$ 을 사용하여 구간  $[T_1(\mathbb{X}), T_2(\mathbb{X})]$ 에  $g(\theta)$ 가 포함될 확률을 고려하는 추정법
- 추정량의 선택기준으로는 평균제곱오차(mean square error)의 최소화를 사용

# 추정량과 추정값

- 미지의 모수를 포함하지 않는 랜덤포본  $X_1, \dots, X_n$ 의 함수를 통계량  $T$ 라고 하자.
  - **추정량(estimator)** : 모수  $\theta$ 의 함수  $g(\theta)$ 를 추정하기 위해 사용되는 통계량  $T(\mathbb{X}) = T(X_1, \dots, X_n)$ 을  $g(\theta)$ 의 추정량이라고 한다.
  - **추정치(estimate)** : 주어진 표본값  $X_1 = x_1, \dots, X_n = x_n$ 을 대입해서 구해진 추정량의 특정값,  $T(\mathbb{x}) = T(x_1, \dots, x_n)$ 을 추정값이라고 한다.
- 모수  $\theta$ 와 그의 추정량을 구별하기 편리하도록, 많은 경우에 추정량을  $\hat{\theta}_n$  또는  $\hat{\theta}$ 으로 표기하도록 한다.  $\hat{\theta}_n$ 에서 첨자  $n$ 은 추정량을 구하는 데 사용된 표본의 크기가  $n$ 임을 강조하기 위해 쓴 것이다.

# 추정량과 추정값 예제

- $X_1, \dots, X_n$ 을  $N(\mu, \sigma^2)$ 으로부터 얻은 랜덤표본이라고 하자. 모평균  $\mu$ 와 모분산  $\sigma^2$ 을 추정량 및 추정값을 구해보자.

- ✓ 모평균  $\mu$ 의 추정량 : 표본평균  $\hat{\mu} = T_1(\mathbb{X}) = \frac{\sum_{i=1}^n X_i}{n}$

- ✓ 모평균  $\mu$ 의 추정값 : 관찰된 표본이 주어졌을 때 그들의 값  $T_1(x_1, \dots, x_n) = \frac{\sum_{i=1}^n x_i}{n}$

- ✓ 모분산  $\sigma^2$ 의 추정량 : 표본분산  $\hat{\sigma}^2 = T_2(\mathbb{X}) = \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{n-1}$

- ✓ 모분산  $\sigma^2$ 의 추정값 : 관찰된 표본이 주어졌을 때  $T_2(x_1, \dots, x_n) = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)^2}{n-1}$

# 최대가능도 추정법의 Motivation

- 성공률이  $p$ 인 베르누이 분포에서 크기가 10인 랜덤표본을 뽑는다고 가정해보자. 이 분포의 성공률은 0.1이거나 0.9 둘 중에 하나라는 것을 알고 있으며, 관측된 표본을 통해 성공률을 추정하려고 한다고 가정하자.
  - 만약 10개의 관측치의 합이 8이었다면(1의 값이 8번 나왔다면) 성공률은 0.1보다는 0.9로 추정하는 것을 선호할 것이다.
  - 왜 0.9를 선호하게 되었는가? 0.9가 참일 경우 합이 8인 관측치를 얻을 확률이 0.1이 참일 경우보다 높기 때문이다.
  - 추정량을 선택할 때, 관측된 자료를 얻을 확률이 가장 높을 추정량을 구한다.

# 가능도 함수

- 주어진(관찰된) 자료에 대하여, **해당 자료가 얻어질 가능성을 모수에 대한 함수**로 나타냄
- 확률변수  $X_1, \dots, X_n$ 의 결합 확률밀도함수가  $f(x_1, \dots, x_n; \theta)$ 라고 하자. 결합 확률밀도함수  $f(x_1, \dots, x_n; \theta)$ 는 고정된 모수  $\theta$ 에 대하여  $(x_1, \dots, x_n)$ 의 함수로 사용된다. 그러나 반대로  $f(x_1, \dots, x_n; \theta)$ 를 관측치  $X_1 = x_1, \dots, X_n = x_n$ 이 주어졌을 때 모수  $\theta (\in \Omega)$ 의 함수로 생각해 볼 수도 있다. 즉  **$L(\theta) = L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$** 로 표기하고 이를  $X_1, \dots, X_n$ 의 **가능도함수**(likelihood function)라고 한다.

# 가능도 함수

- 가능도함수  $L(\theta)$ 는 주어진 자료  $(x_1, x_2, \dots, x_n)$ 에 대하여,  $(x_1, x_2, \dots, x_n)$ 이 얻어질 가능성을 모수  $\theta$ 에 대한 함수로 나타낸 것이다.

- $X_1, \dots, X_n$ 이 서로 독립적이고,  $X_i$ 가 확률밀도함수  $f_i(x_i; \theta)$ 를 가질 때의 가능도함수 :

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n \mathbf{f(x_i; \theta)} = f_1(x_1; \theta) f_2(x_2; \theta) \dots f_n(x_n; \theta)$$



# 최대가능도 추정량

- 랜덤표본의 가능도함수  $L(\theta; x_1, \dots, x_n)$ 를 최대화하는  $\theta$ 의 값을  $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n) \in \Omega$ 라고 할 때,  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ 를 모수  $\theta$ 의 **최대가능도 추정량**(maximum likelihood estimator)이라 한다.
- 최대가능도 추정량의 의미는 "실제로 관측된" 자료가 얻어질 확률을 가장 높게 만드는 (즉, 주어진 관측값을 가장 잘 설명하는)  $\theta$ 의 값을 모수  $\theta$ 의 추정량으로 삼는 것이다.

# 최대가능도 추정량 예제

- $X_1, X_2, \dots, X_5$ 가 서로 독립인 베르누이( $p$ ) 확률변수라고 하자. 이때  $Y = \sum_{i=1}^5 X_i$ 는  $B(5, p)$  분포를 따르며,  $Y$ 의 확률밀도함수는 몇 가지  $p$ 에 대해서 다음과 같이 주어진다.

$Y \backslash p$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.59050	0.32768	0.16807	0.07776	0.03125	0.01024	0.00243	0.00032	0.00005	0.0
1	0.32800	0.40960	0.36015	0.25920	0.15625	0.07680	0.02835	0.00640	0.00045	0.0
2	0.07290	0.20480	0.30870	0.34560	0.31250	0.23040	0.13230	0.05120	0.00810	0.0
3	0.00810	0.05120	0.13230	0.23040	0.31250	0.34560	0.30870	0.20480	0.07290	0.0
4	0.00045	0.00640	0.02835	0.07680	0.15625	0.25920	0.36015	0.40960	0.32800	0.0
5	0.00005	0.00032	0.00243	0.01024	0.03125	0.07776	0.16807	0.32768	0.59050	1.0

$Y$ 의 값이 3일 경우,  
 $p = 0.6$ 일 때 가능도함수의 값이 제일 크다.

관측값  $Y$ 의 값이 크면, 성공확률이 높은 쪽으로 추정  
 관측값  $Y$ 의 값이 작으면, 성공확률이 작은 쪽으로 추정

# 로그 가능도함수

- 가능도함수를 최대화하는 문제는 로그 가능도함수를 최대화하는 문제와 같다.
  - $\log L(\theta; x_1, \dots, x_n) = \log \prod_{i=1}^n f_i(x_i; \theta) = \sum_{i=1}^n \log f_i(x_i; \theta)$ 를 최대화
  - $\frac{d}{d\theta} \log L(\theta; x_1, \dots, x_n) = 0$ 의 해를 구하는 문제로 귀착된다.

# 최대가능도 추정법 예제 1

- $X_1, X_2, \dots, X_n$ 을  $EXP(\theta)$ 에서 추출된 랜덤표본이라고 할 때,  $\theta$ 의 최대가능도 추정량을 구해보자.

- ✓ 가능도 함수 :

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta) = \left(\frac{1}{\theta}\right)^n \exp\left(-\sum_{i=1}^n \frac{x_i}{\theta}\right)$$

- ✓ 로그 가능도 함수 :

$$\log L(\theta; x_1, x_2, \dots, x_n) = -n \log \theta - \sum_{i=1}^n \frac{x_i}{\theta}$$

- ✓ 로그 가능도 함수를  $\theta$ 에 대해 미분 :

$$\frac{d}{d\theta} \log L(\theta; x_1, x_2, \dots, x_n) = -\frac{n}{\theta} + \sum_{i=1}^n \frac{x_i}{\theta^2}$$

- ✓ 미분값을 0으로 만드는  $\theta$ 의 값은  $\bar{x}_n$ 이고, 이 값에서 로그가능도 함수의 이차 미분값은 0보다 작으므로, 이 값에서 로그가능도함수가 최대가 된다. 따라서  $\theta$ 의 최대가능도 추정량은  $\bar{x}_n$ 이다.

# 최대가능도 추정법 예제 2

- $X_1, X_2, \dots, X_n$ 을 포아송 분포( $\lambda$ )로부터 구한 랜덤표본이라고 할 때,  $\lambda$ 의 최대가능도 추정량을 구하여라.

- ✓ 가능도 함수 :

$$L(\lambda; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \lambda) = \frac{\exp(-n\lambda) \lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

- ✓ 로그 가능도 함수 :

$$\log L(\lambda; x_1, x_2, \dots, x_n) = -n\lambda + \sum_{i=1}^n x_i \log \lambda - \log \left( \prod_{i=1}^n x_i! \right)$$

- ✓ 로그 가능도 함수를  $\theta$ 에 대해 미분 :

$$\frac{d}{d\theta} \log L(\lambda; x_1, x_2, \dots, x_n) = -n + \sum_{i=1}^n \frac{x_i}{\lambda}$$

- ✓ 미분값을 0으로 만드는  $\theta$ 의 값은  $\bar{x}_n > 0$ 이고, 이 값에서 로그가능도 함수의 이차 미분값은 0보다 작으므로, 이 값에서 로그가능도함수가 최대가 된다. 따라서  $\theta$ 의 최대가능도 추정량은  $\bar{x}_n$ 이다.

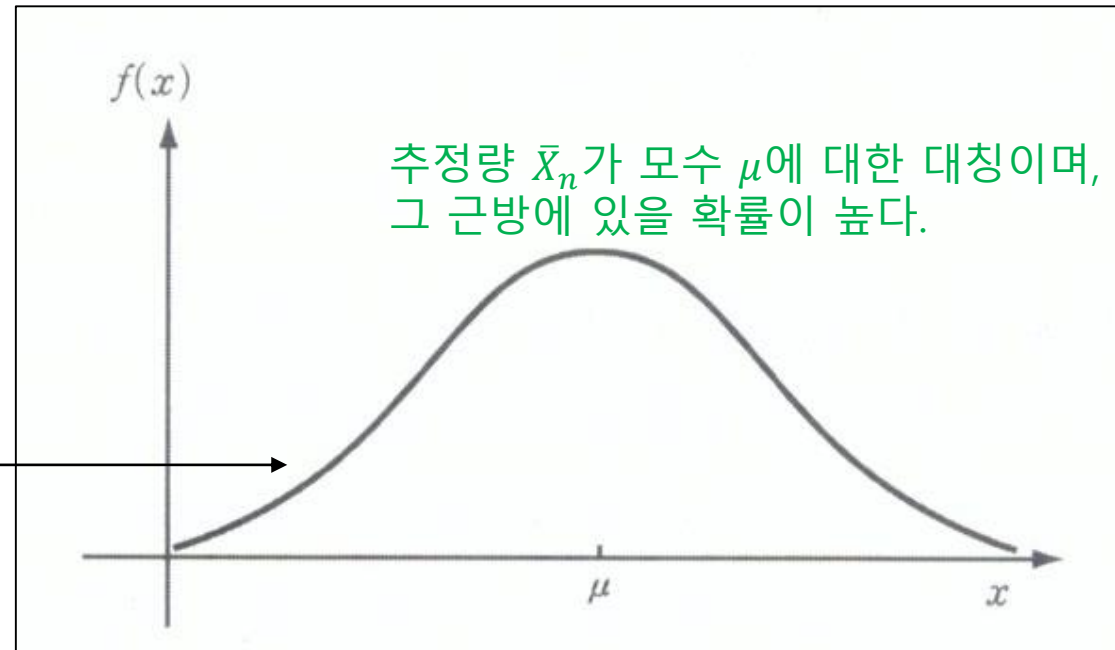
# 최대가능도 추정량의 불변성

- 불변성 원리 (invariance property)
  - $\theta$  를 추정하는 어떤 추정량이  $T(X)$  일 때,  $\theta$  의 함수  $g(\theta)$  의 추정량이  $g(T(X))$  가 되는 성질
- 최대가능도 추정량의 불변성
  - $X_1, \dots, X_n$ 을 확률밀도함수  $f(x; \theta)$  ,  $\theta \in \Omega$  를 갖는 분포에서 얻은 랜덤표본이라고 하자.  $\hat{\theta}_n$  이 모수  $\theta$  의 최대가능도 추정량이면,  $\theta$  의 함수인  $g(\theta)$  에 대하여,  $g(\hat{\theta}_n)$  이  $g(\theta)$  의 최대가능도 추정량이 된다.

# 추정량의 확률분포

- 표본의 함수인 추정량은 확률변수이다. 확률변수는 상응하는 확률분포를 가지므로 추정량도 그의 확률분포가 있다.
- 예를 들어,  $X_1, X_2, \dots, X_n$ 을  $N(\mu, \sigma^2)$ 으로부터 얻은 랜덤표본이라고 하자. 모평균  $\mu$ 를 표본평균  $\bar{X}_n$ 로 추정할 때 추정량  $\bar{X}_n$ 의 확률분포는 모수(모평균)에 중심을 둔  $N(\mu, \sigma^2/n)$  분포를 따른다.

추정량  $\bar{X}_n$ 의 확률밀도함수



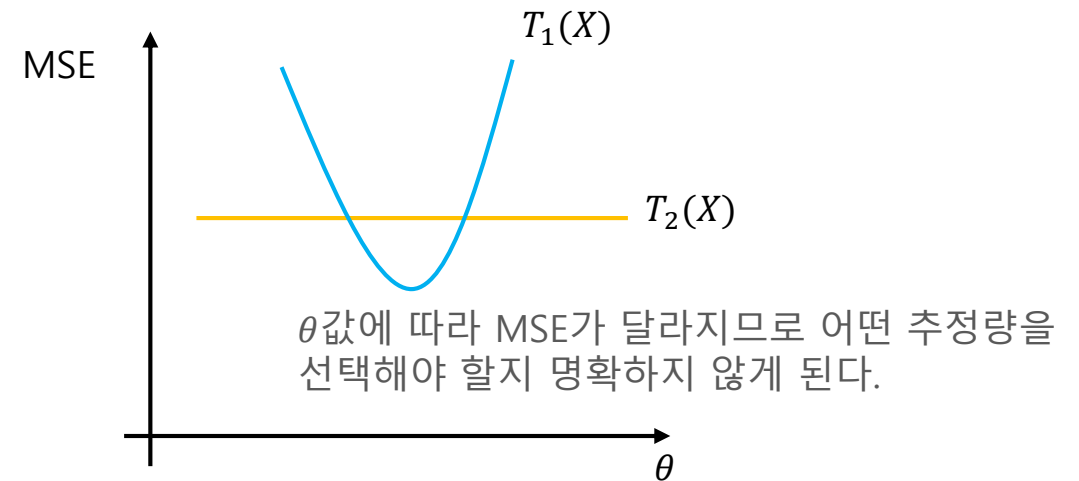
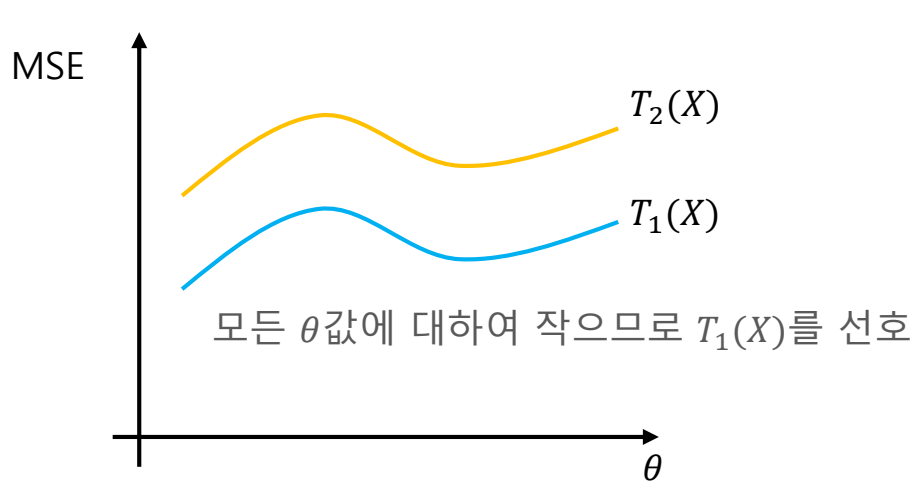
# 추정의 기준

- 우리가 원하는 바는 추정량  $T(\mathbb{X}) = T(X_1, \dots, X_n)$  이 추정하고자 하는 모수 또는 모수의 함수  $g(\theta)$  에 "가까운" 것이다.
- 추정량은 표본에 따라 값이 달라지므로 언제나  $g(\theta)$  와 동일한 값을 가질 수는 없다. 대신 추정량 이  $g(\theta)$  에 "가까운" 값을 제공한다면 그것은 훌륭한 추정량이라고 할 것이다.
- "가까운"의 기준? **손실함수** (loss function)
  - 절대오차 (absolute error) :  $|T(X) - g(\theta)|$
  - 제곱오차 (squared error) :  $(T(X) - g(\theta))^2$
  - $w(\theta)$  로 가중값을 준 가중제곱오차 (weighted squared error) :  $(T(X) - g(\theta))^2 w(\theta)$
- 이 손실함수들은 확률변수이다. 이들을 최소화하는 추정량  $T(X)$ 를 찾아내는 것은 불가능하며, 손실 함수의 기대값을 기준으로 삼는다.



# 평균제곱오차

- 일반적으로 제곱오차의 기댓값인 평균제곱오차 (Mean Squared Error , MSE) 를 추정의 기준으로 많이 사용한다.
  - $MSE = E(T(X) - g(\theta))^2$
- 평균제곱오차는 모수  $\theta$ 에 대한 함수이기 때문에 이를 최소화하는 추정량을 찾는 것이 항상 가능하지는 않다. 또한 모든  $\theta \in \Omega$ 에 대하여 균일하게 최소화하는 추정량  $T(X)$ 는 일반적으로 존재하지 않는다.



두 추정량의 평균제곱오차 MSE 비교

# 비편향추정량

- $\theta$ 를 추정하기 위한 어떤 추정량이  $T(\mathbb{X}) = \theta_0$  (어떤 정해진 상수) 이라고 하자. 즉 표본에서 어떤 값을 관측하든 추정값을  $\theta_0$  으로 하겠다는 의미이다. 만약 미지의 모수값이 실제로  $\theta_0$  이라면, 이 추정량의 평균제곱오차는 0 이 된다. 그러나 이러한, 표본과 무관한 추정방법은  $\theta$ 가  $\theta_0$ 로부터 멀리 떨어져 있을수록 평균제곱오차를 크게 만들 것이다. 이러한 추정량들은 고려대상에서 제외하는 것이 합리적이다.
- 이런 편협된 부적절한 추정량들을 제외하는 하나의 방법으로는 비편향추정량 (inbiased estimator) 들의 집합을 고려하고, 그 안에서 평균제곱오차를 최소화는 추정방법을 찾는 것이다.
- $T(\mathbb{X})$  를  $g(\theta)$  의 추정량이라고 할 때,  $E[T(\mathbb{X})] - g(\theta)$  를  $T(\mathbb{X})$  의 편향 (bias) 이라고 하며,  $E[T(\mathbb{X})] = g(\theta)$  (즉 편향 = 0) 이면  $T(\mathbb{X})$  를  $g(\theta)$  의 **비편향추정량**이라고 한다.

# 비편향추정량 예제

- $X_1, X_2, \dots, X_{10}$ 을  $N(\mu, \sigma_0^2)$ 으로부터 얻은 랜덤표본이라 하자.  $\sigma_0^2$ 은 알려진 값이라고 하자.

✓  $T_1(X) = \bar{X}_{10}$

$$E[T_1(X)] = E\left[\frac{1}{10} \sum_{i=1}^{10} X_i\right] = \mu$$

✓  $T_2(X) = \frac{X_1 + X_2}{2}$

$$E[T_2(X)] = E\left[\frac{X_1 + X_2}{2}\right] = \mu$$

- ✓  $T_1(X)$  ,  $T_2(X)$ 는 모평균  $\mu$ 의 비편향추정량이다.

# 평균제곱오차의 성질

- 모수의 함수  $g(\theta)$  의 추정량  $T(\mathbb{X})$  의 평균제곱오차(MSE) 는  $MSE = Var(T(\mathbb{X})) + (bias)^2$  을 만족한다.
- 평균제곱오차를 기준으로 추정할 때에는 분산과 편향의 두 요소가 존재함을 알 수 있다. 즉 평균제곱오차를 작게 하는 추정량은 편향의 절대값과 분산, 둘 다 작은 값을 가져야 한다.
- 비편향추정량들만의 집합을 고려할 때는 평균제곱오차를 최소화하는 추정량을 구하기 위해 분산을 최소화하는 추정량을 구하면 된다. 이 경우는 분산의 비 (variance ratio) 를 이용하면 편리하다.
- $T_1(\mathbb{X})$  와  $T_2(\mathbb{X})$  가  $g(\theta)$  의 비편향추정량일 때, 그 분산의 비, 즉  $\frac{Var[T_2(\mathbb{X})]}{Var[T_1(\mathbb{X})]}$  를 사용하고, 분산의 비를 추정량  $T_1(\mathbb{X})$  의  $T_2(\mathbb{X})$  에 대한 상대효율 (relative efficiency) 이라고 한다.

# 최소분산 비편향추정량

- 확률변수  $X_1, \dots, X_n$  의 결합 확률밀도함수가  $f(x_1, \dots, x_n; \theta)$  라고 하자. 함수  $g(\theta)$  의 추정량  $T^*(\mathbb{X})$  가 다음의 조건을 만족시키면 이를  $g(\theta)$  의 **최소분산 비편향추정량**(Minimum Variance Unbiased Estimator, MVUE)이라고 한다.
  - $E[T^*(\mathbb{X})] = g(\theta)$
  - $Var(T^*(\mathbb{X})) \leq Var(T(\mathbb{X}))$
- $g(\theta)$ 의 최소분산 비편향추정량을 구하는 데에 사용되는 두 가지 방법
  - 분산의 하한값인 "크라머-라오 하한값"을 갖는지 확인하는 방법
  - 완비 충분통계량을 활용하여 "라오-블랙웰의 정리"와 "레만-쉐페의 정리"를 사용하는 방법

# 구간추정은 왜 하는가?

- 모수의 함수에 대한 최적추정량들은 점추정문제여서 통계적 정확도(statistical accuracy)를 표현하지 못한다는 결점이 있다.
  - 포아송( $\lambda$ )의 경우 모수  $\lambda$ 의 추정량으로서 표본평균  $\bar{X}_n$ 은 최적성 등의 좋은 성질을 가지고 있으나, 정확성에 대한 아무런 정보를 포함하고 있지 않다.
- 실제 관측된 자료로부터 계산된 추정량(예를 들어, 표본평균)의 값이 정확히 모수와 같을 것이라 생각할 수 없다. 그러면, 이 값은 모수로부터 어느 정도 떨어져 있을까?
- 구간추정에서는 랜덤표본  $X_1, \dots, X_n$ 에 근거하여 모수에 대한 신뢰구간(=확률구간, random interval)을 정의하고 그 구간 내에 모수가 포함될 확률로써 통계적 정확도를 표현

# 신뢰구간

- 랜덤표본  $X_1, \dots, X_n$ 의 확률밀도함수가  $f(x; \theta)$ ,  $\theta \in \Omega$ 라고 하자. 확률구간  $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ 이,  $0 < \alpha < 1$ 에 대하여  $P[L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n)] = 1 - \alpha$ 를 만족시키면 이를 **모수  $\theta$ 의  $100(1 - \alpha)\%$  신뢰구간(confidence interval)**이라 부른다. 또한  $L(X_1, \dots, X_n)$ 과  $U(X_1, \dots, X_n)$ 을 각각 신뢰구간 하한 (lower limit) 과 상한 (upper limit) 이라 말한다.
  - 수많은 랜덤표본을 뽑아서 각 표본에 대해 같은 방법으로 신뢰구간을 구하면, 구해진 많은 신뢰구간들 가운데  $100(1 - \alpha)\%$  만큼의 신뢰구간들이 모수를 포함
  - $1 - \alpha$ 는 신뢰계수(confidence coefficient) 또는 신뢰도(confidence level)라고 한다.

# 추측변량

- 랜덤표본  $X_1, \dots, X_n$ 의 분포가 확률밀도함수  $f(x; \theta)$ ,  $\theta \in \Omega$ 를 따른다 하자. 이때 표본과 모수  $\theta$ 의 함수인 **확률변량(확률변수)**  $T(X_1, \dots, X_n; \theta)$ 의 분포가 모수  $\theta$ 에 의존하지 않으면 이를 **추측변량**이라고 한다.
  - ✓ 예를 들어  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$  랜덤표본에서,
    - $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{n-1}$ 일 때,  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S}$ 의 분포는  $t(n-1)$ 로 모수  $(\mu, \sigma^2)$ 에 의존하지 않으므로 추측변량이다.



# 추측변량을 이용한 신뢰구간

- 추측변량을 활용하여 모수의 신뢰구간을 구하는 방법
  - $T = T(X_1, \dots, X_n; \theta)$ 가 추측변량이면, 고정된  $0 < \alpha < 1$ 에 대하여  $P(t_1 < T < t_2) = 1 - \alpha$ 를 만족하는  $t_1, t_2$ 를 구한다.
  - $T$ 의 분포가 모수에 의존하지 않으므로  $t_1$ 과  $t_2$ 는  $\alpha$ 의 값이 주어지면 상수값이 될 것이다.
  - 부등식  $t_1 < T(x_1, \dots, x_n; \theta) < t_2$ 를  $L(x_1, \dots, x_n) < \theta < U(x_1, \dots, x_n)$ 의 꼴로 바꿔쓸 수 있다면,  $P(L(X_1, \dots, X_n) < \theta < U(X_1, \dots, X_n)) = 1 - \alpha$ 가 된다.
  - $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ 은  $\theta$ 에 대한  $100(1 - \alpha)\%$  신뢰구간이 된다.

# 정규분포의 모수에 대한 신뢰구간

- 모평균에 대한 신뢰구간
- 모분산에 대한 신뢰구간

# 모평균에 대한 신뢰구간 (모분산을 알 때)

- $X_1, \dots, X_n$ 을  $N(\mu, \sigma^2)$ 으로부터의 랜덤표본이라고 하고, 모평균  $\mu$ 에 대한 신뢰구간을 구해 보자.
  - 모분산  $\sigma^2$ 이 알려져 있으면, 변량  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ 의 분포는  $N(0,1)$ 으로서 모수  $(\mu, \sigma)$ 에 의존하지 않으므로 추측변량이다.
  - $P\left[-z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$ 이다.
  - $P\left[\bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$ 이다.
  - 결과적으로 모평균  $\mu$ 에 대한  $100(1 - \alpha)\%$  신뢰구간  $[\bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ 을 구할 수 있다.

# 모평균에 대한 신뢰구간 예제(모분산을 알 때)

- 우리나라 대학교재 가운데서 64권을 표본 추출하여 한 문장을 이루고 있는 글자 수를 조사하였다. 표본조사 결과에 의하면 한 문장은 평균 60자로 되어 있다. 단 전체 대학교재에서 한 문장을 이루는 글자 수의 표준편차는 32자라고 한다. 그렇다면 우리나라에서 대학교재로 쓰이는 모든 책은 한 문장이 평균 몇 글자로 되어 있을까? 95%를 신뢰계수로 구간 추정하라.

# 모평균에 대한 신뢰구간 예제(모분산을 알 때)

- $1 - \alpha = 0.95$  ,  $\sigma = 32$  ,  $\bar{x} = 60$  ,  $n = 64$
- $P \left[ -z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq z_{\frac{\alpha}{2}} \right] = P \left[ -z_{0.05} \leq \frac{\sqrt{64}(60 - \mu)}{32} \leq z_{0.05} \right] = 0.95$
- $P \left[ \bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = P \left[ 60 - 1.96 \frac{32}{\sqrt{64}} \leq \mu \leq 60 + 1.96 \frac{32}{\sqrt{64}} \right] = 0.95$
- $\left( 60 - 1.96 \frac{32}{\sqrt{64}} \leq \mu \leq 60 + 1.96 \frac{32}{\sqrt{64}} \right) = (52.16 \leq \mu \leq 67.84)$
- 따라서 95%를 신뢰계수로 할 때, 대학교재에서 한 문장을 이루는 평균 글자 수는 약 52자에서 68자 사이일 것이다.

# 모평균에 대한 신뢰구간 (모분산을 모를 때)

- $X_1, \dots, X_n$ 을  $N(\mu, \sigma^2)$ 으로부터의 랜덤표본이라고 하고, 표본이 적은 경우 모평균  $\mu$ 에 대한 신뢰구간을 구해 보자.
  - 모분산  $\sigma^2$  이 알려져 있지 않은 경우에는,  $[\bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$  이 모수인  $\sigma^2$  의 값에 의존하기 때문에 사용할 수 없다.
  - $\sigma^2$  를  $S_n^2 = \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{(n-1)}$  으로 추정하게 되면 추측변량  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$  은 자유도가  $(n-1)$  인  $t$  분포를 따르게 된다.
  - $P \left[ -t_{\frac{\alpha}{2}}(n-1) \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \leq t_{\frac{\alpha}{2}}(n-1) \right] = 1 - \alpha$ 이다.
  - $P \left[ \bar{X}_n - t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \right] = 1 - \alpha$ 이다.
  - 결과적으로 모평균  $\mu$  에 대한  $100(1 - \alpha)\%$  의 신뢰구간  $\left[ \bar{X}_n - t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \right]$ 을 구할 수 있다.

# 모평균에 대한 신뢰구간 예제(모분산을 모를 때)

- 외부세계로부터 얻는 정보 중에서 85%는 시각을 통해서 들어온다고 한다. 그래서 같은 정보량일지라도 귀로 듣기보다 활자로 읽는 편이 빠르다. 일반적으로 보통 책 한쪽에는 600자 글자가 적혀있다. 대학생 16명을 표본으로 뽑아 책 한 쪽을 읽는데 걸리는 시간을 조사하였다. 조사 결과 한 쪽을 읽는 데 걸리는 시간은 평균 90초이고 표준편차는 24초이다. 그렇다면 대학생 모든 학생들이 책 한 쪽을 읽는 데 걸리는 평균 시간에 대하여 95%를 신뢰계수로 구간추정하라.

# 모평균에 대한 신뢰구간 예제(모분산을 모를 때)

- $1 - \alpha = 0.95$  ,  $S_n = 24$  ,  $\bar{x} = 90$  ,  $n = 16$
- $P \left[ -t_{\frac{\alpha}{2}}(n-1) \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \leq t_{\frac{\alpha}{2}}(n-1) \right] = P \left[ -2.131 \leq \frac{\sqrt{16}(90 - \mu)}{24} \leq 2.131 \right] = 0.95$
- $P \left[ \bar{X}_n - t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\frac{\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \right] = P \left[ 90 - 2.131 \frac{24}{\sqrt{16}} \leq \mu \leq 90 + 2.131 \frac{24}{\sqrt{16}} \right] = 0.95$
- $\left( 90 - 2.131 \frac{24}{\sqrt{16}} \leq \mu \leq 90 + 2.131 \frac{24}{\sqrt{16}} \right) = (77.21 \leq \mu \leq 102.79)$
- 따라서 95%를 신뢰계수로 할 때, 책 한쪽을 읽는데 걸리는 평균 시간은 약 77.21초에서 102.79초 사이일 것이다.
- 표본의 수가 더 많은 경우는  $z$  확률변량을 추측변량으로 사용하고,  $\sigma$ 를 대신하여  $s$ 를 사용한 다.



# 신뢰구간의 의미

- 관찰값이  $X_1 = x_1, \dots, X_n = x_n$  으로 주어지면, 이들의 평균을  $\bar{x}_n$  라 했을 때, 신뢰구간은  $[\bar{x}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$  의 값이 알려진 구간이 된다.
  - $\bar{x}_n$  에 대해 대칭
  - 구간의 길이는 모분산  $\sigma^2$  이 작을수록 표본의 크기  $n$  이 커질수록 짧아진다.
- 주어진 표본에 대해 모수(여기서는 모평균)의 신뢰구간은 그 모수를 포함하거나 포함하지 않거나 둘 중의 하나이다. 독립된 여러 개의 표본에 근거하여 모수의  $100(1 - \alpha)\%$  신뢰구간들을 구해보면, 그 신뢰구간들 중에서  $100(1 - \alpha)$  개의 신뢰구간들이 모수를 포함한다는 의미이다.

# 모분산에 대한 신뢰구간 (모평균을 알때)

- $X_1, \dots, X_n$  을  $N(\mu, \sigma^2)$  으로부터의 랜덤표본이라고 할 때, 모분산  $\sigma^2$  에 대한 신뢰구간을 구해보자.
- 모평균  $\mu$  가 알려진 경우  $\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}$  는 자유도가  $n$  인 카이제곱분포를 따르는 추측변량이 된다.
- $P \left[ \chi_{1-\frac{\alpha}{2}}^2(n) \leq \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}}^2(n) \right] = 1 - \alpha$  이다.
- $P \left[ \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\frac{\alpha}{2}}^2(n)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\frac{\alpha}{2}}^2(n)} \right] = 1 - \alpha$  이다.
- $\left[ \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\frac{\alpha}{2}}^2(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\frac{\alpha}{2}}^2(n)} \right]$  이 모분산  $\sigma^2$  에 대한  $100(1 - \alpha)\%$  신뢰구간이 된다.

# 모분산에 대한 신뢰구간 (모평균을 모를 때)

- $X_1, \dots, X_n$  을  $N(\mu, \sigma^2)$  으로부터의 랜덤표본이라고 할 때, 모분산  $\sigma^2$  에 대한 신뢰구간을 구해보자.
- 모평균이 알려지지 않는 경우에는,  $\mu$  를 표본평균  $\bar{X}_n$  로 추정한 통계량  $\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2}$  을 추측변량으로 사용한다. 이 추측변량은 자유도가  $n - 1$  인 카이제곱분포를 따른다.
- $P \left[ \chi_{1-\frac{\alpha}{2}}^2(n-1) \leq \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}}^2(n-1) \right] = 1 - \alpha$
- $P \left[ \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} \right] = 1 - \alpha$
- 따라서  $\left[ \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}, \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} \right]$  는 모분산  $\sigma^2$  에 대한  $100(1 - \alpha)\%$  신뢰구간이 된다.

# 모분산에 대한 신뢰구간 예제 (모평균을 모를 때)

- A은행에서 새로운 시스템을 도입한 이후 이용고객들의 불편신고가 오히려 옛날보다 많이 증가하였다. 이에 은행에서 실제 이용고객 10명을 대상으로 은행 업무를 보기 위하여 기다리는 시간(단위 : 분)을 측정하여 다음 자료를 얻었다. 대기시간의 분산에 대한 90% 신뢰구간을 구하라. 대기시간은 정규분포를 따른다고 가정한다.
  - 3.00 5.50 2.07 7.30 4.97 5.90 10.35 4.10 6.13 5.83

# 모분산에 대한 신뢰구간 예제 (모평균을 모를 때)

- $\bar{x}_n = 5.515$  ,  $S^2 = 5.298$  ,  $\chi_{0.05}^2(9) = 16.92$  ,  $\chi_{0.95}^2(9) = 3.33$  ,  $\alpha = 0.1$
- $P \left[ \chi_{1-\frac{\alpha}{2}}^2(n-1) \leq \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}}^2(n-1) \right] = P \left[ 3.33 \leq \frac{9 \times 5.298}{\sigma^2} \leq 16.92 \right] = 0.90$
- $P \left[ \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{1-\alpha/2}^2(n-1)} \right] = P \left[ \frac{9 \times 5.298}{16.92} \leq \sigma^2 \leq \frac{9 \times 5.298}{3.33} \right] = 0.90$
- $\left( \frac{9 \times 5.298}{16.92} \leq \sigma^2 \leq \frac{9 \times 5.298}{3.33} \right) = (2.82 \leq \sigma^2 \leq 14.32)$
- 따라서, 90%를 신뢰계수로 할 때, 모분산의 신뢰구간은  $(2.82 \leq \sigma^2 \leq 14.32)$ 이다.

# 모분산 비에 대한 신뢰구간 (모평균을 알 때)

- 두 개의 독립인 랜덤표본  $X_1, \dots, X_n$ 과  $Y_1, \dots, Y_m$ 의 분포가 각각  $N(\mu_X, \sigma_X^2)$ ,  $N(\mu_Y, \sigma_Y^2)$ 이라고 할 때, 모분산 비에 대한 신뢰계수  $100(1 - \alpha)$ 의 신뢰구간을 구하여라.

- 모평균  $\mu_X$  와  $\mu_Y$  를 아는 경우의  $\frac{\sigma_X^2}{\sigma_Y^2}$ , 확률변량  $\frac{\frac{\sum_{i=1}^n (X_i - \mu_X)^2}{n\sigma_X^2}}{\frac{\sum_{i=1}^m (Y_i - \mu_Y)^2}{m\sigma_Y^2}} = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F(n, m)$ 는 추측변량이다.

- $P \left[ F_{1-\frac{\alpha}{2}}(n, m) \leq \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \leq F_{\frac{\alpha}{2}}(n, m) \right] = 1 - \alpha$ 이다.

- $\left[ \left( \frac{S_X^2}{S_Y^2} \right) F_{1-\frac{\alpha}{2}}(m, n), \left( \frac{S_X^2}{S_Y^2} \right) F_{\frac{\alpha}{2}}(m, n) \right]$  가 모분산 비의  $100(1 - \alpha)\%$  신뢰구간이다.

# 모분산 비에 대한 신뢰구간 (모평균을 모를 때)

- 두 개의 독립인 랜덤표본  $X_1, \dots, X_n$ 과  $Y_1, \dots, Y_m$ 의 분포가 각각  $N(\mu_X, \sigma_X^2)$ ,  $N(\mu_Y, \sigma_Y^2)$ 이라고 할 때, 모분산 비에 대한 신뢰계수  $100(1 - \alpha)$ 의 신뢰구간을 구하여라.

- 모평균  $\mu_X$  와  $\mu_Y$  를 모르는 경우  $\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_X^2} \sim \chi(n - 1)$ ,  $\frac{\sum_{i=1}^m (Y_i - \bar{Y}_m)^2}{\sigma_Y^2} \sim \chi(m - 1)$ 이고  $\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F(n - 1, m - 1)$ 을 추측변량으로 사용한다.

$$\bullet P \left[ F_{1-\frac{\alpha}{2}}(n-1, m-1) \leq \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \leq F_{\frac{\alpha}{2}}(n-1, m-1) \right] = 1 - \alpha$$

- 모분산 비  $\frac{\sigma_X^2}{\sigma_Y^2}$  에 대한 신뢰도  $100(1 - \alpha)\%$  신뢰구간은  $\left[ \frac{S_X^2}{S_Y^2} F_{1-\frac{\alpha}{2}}(m-1, n-1), \frac{S_X^2}{S_Y^2} F_{\frac{\alpha}{2}}(m-1, n-1) \right]$

# 모평균의 차이에 대한 신뢰구간

- 독립인 두 개의 랜덤표본  $X_1, \dots, X_n$  과  $Y_1, \dots, Y_m$  의 분포가 각각  $N(\mu_X, \sigma_X^2)$  와  $N(\mu_Y, \sigma_Y^2)$  일 때, 모평균의 차이  $\mu_X - \mu_Y$ 에 대한 신뢰구간을 구하라.

- 모분산  $\sigma_X^2$  ,  $\sigma_Y^2$  를 아는 경우 확률변량  $Z = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$  는 표준정규분포  $N(0,1)$  를 따르고 추측변량이 된다.

- $$P \left[ -z_{\frac{\alpha}{2}} \leq \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \leq z_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

- $$\left[ (\bar{X}_n - \bar{Y}_m) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{X}_n - \bar{Y}_m) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right]$$
 이 모평균의 차  $\mu_X - \mu_Y$  에 대한  $100(1 - \alpha)\%$  의 신뢰구간이다.



# 모평균의 차이에 대한 신뢰구간

- 모분산  $\sigma_X^2$ ,  $\sigma_Y^2$  를 모르는 경우
  - $\sigma_X^2 = \sigma_Y^2$ 의 경우
    - 합동추정량(pooled estimator)를 사용한다.  $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$  ( $S_X = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)}{n-1}$ ,  $S_Y = \frac{\sum_{i=1}^m (Y_i - \bar{Y}_m)}{m-1}$ )
  - $\sigma_X^2 \neq \sigma_Y^2$ 의 경우
    - 웰치에 의해 제안된 방법을 사용한다.

# 근사신뢰구간

- 모분포가 정규분포를 따르지 않더라도 중심극한정리를 사용하여 모수에 대한 신뢰구간을 근사적으로 구할 수 있다.
- 랜덤표본  $X_1, X_2, \dots, X_n$ 의 분포의 평균과 분산이 각각  $\mu, \sigma^2$ 이라고 할 때 모평균  $\mu$ 에 대한 신뢰구간을 구하자.
  - 표본의 크기가 크면 확률변량  $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$ 의 분포는 표준정규분포  $N(0,1)$ 로 수렴한다. 이것을 추측변량으로 사용한다.
  - $P\left[-z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq z_{\frac{\alpha}{2}}\right] \approx 1 - \alpha$ 이다.
  - $P\left[\bar{X}_n - z_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}\right] \approx 1 - \alpha$ 이다.
  - 결과적으로 모평균  $\mu$ 에 대한  $100(1 - \alpha)\%$  신뢰구간  $[\bar{X}_n - z_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}]$ 을 구할 수 있다.

# 표본의 크기

- 모분산  $\sigma^2$ 가 알려진 경우에 모평균  $\mu$ 의 추정량  $\hat{\mu} = \bar{X}$ 에 대한  $100(1 - \alpha)\%$  오차한계를  $d$  이하로 만드는 표본크기  $n$ 은?
  - $\left[ \bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$ 의 길이가  $2d$  이하가 될 것이다.
  - $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq d$  즉,  $n \geq \left( z_{\frac{\alpha}{2}} \frac{\sigma}{d} \right)^2$
  - 이것은 모집단의 분포가 정규분포인 경우에는 정확한 것이고, 표본크기가 큰 경우에는 임의의 모집단에 대하여 근사적으로 적용될 수 있다.

