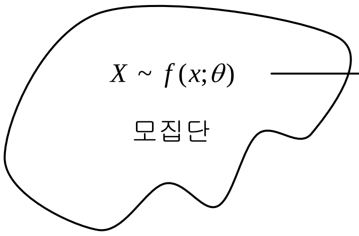
모집단 모수(parameter, θ)를 값을 추정(estimation)하거나 가설검정(hypothesis testing)을 위하여 확률 표본(random sample)을 추출하게 된다. 표본으로부터 모수에 **"가장"** 적절한 통계량(statistic)을 계산하고 이를 이용하여 추론(inference)을 하게 된다.



확률표본 $(x_1, x_2, ..., x_n) \overset{iid}{\sim} f(x;\theta)$

iid =independently and identically distributed

서로 독립이고 동일한 분포를 갖는다.

표본 크기 n 인 확률표본 $(X_1, X_2, ..., X_n)$ 의 결합분포함수는 다음과 같다.

$$f(x_1, x_2, ..., x_n) = f(x_1) f(x_2) ... f(x_n)(\text{독립}) = f(x) f(x) ... f(x)(\text{동일분포})=[f(x)]^n$$

이제까지는 확률변수의 함수(이것이 통계량)에 대한 확률분포함수(pdf)를 얻는 방법에 대해 살펴보았다. 예를 들면 표본평균($\bar{X} = \sum X_i / n$, 확률표본의 함수)의 분포는 평균 μ , 분산 σ^2 / n 인 정규분포를 따른다.($\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$) 한편 표본 분산의 경우 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 임을 알았다. $\bar{X}$ 와 S^2 이 서로 독립이므로 t-분포의 정의에 의해 $\dfrac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} / \sqrt{\dfrac{(n-1)S^2}{\sigma^2} / (n-1)} = \dfrac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$ 이다.

7.1 추정 개념

휴대폰 제조업체는 생산된 제품이 1년 이내 고장 날 확률(p)을 알고자 한다. 학교 농협에서 번호표를 뽑은 후 창구까지 가는데 걸리는 평균 시간(μ)을 알고자 한다. 지름이 10mm인 파이프를 만드는 공장장은 파이프 지름의 분산(σ^2)을 알고자 한다. p, μ, σ^2 을 "(목표) 모수" (target parameter)라 하고 이것을 θ 라고 나타낸다.

θ 에 근사할 것이라고 생각하는 하나의 값으로 제시한다면 이를 점추정(point estimate), θ 을 포함하고 있을 가능성이 높은 구간을 제시하는 것은 구간추정(interval estimate)이라 한다. $\theta = \mu$ 인 경우 확률표본(데이터) $(X_1, X_2, ..., X_n)$ 으로부터 계산된 표본평균($\bar{X}$)을 θ 에 근사한 값으로 제시한다면 $\bar{X}$ 은 점추정이다. 구간추정의 예는 $(\bar{X} - t(n-1)\dfrac{s}{\sqrt{n}}, \bar{X} + t(n-1)\dfrac{s}{\sqrt{n}})$ 이다.

DEFINITION (추정량, 추정치)

목표 모수(θ)에 대한 추정 값을 얻기 위하여 확률표본(데이터) 값들로부터 계산하는 식을 추정량 (estimator)이라 한다. 실제 측정값으로부터 추정량에 의해 계산된 추정 값을 추정치 (estimate)라 한다.

EXAMPLE 7.1

모집단 평균(μ)에 대한 추정 값으로 사용되는 표본평균을 생각해 보자.

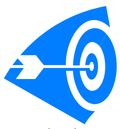
$\bar{X} = \sum X_i / n$ 가 추정량이 되고 실제 데이터(측정) 값에 의해 얻어진 값을 추정치라 한다.

"목표 모수"(θ)에 대한 추정치는 무한히 많이 존재한다. 예를 들어 모수가 모집단 평균(μ)인 경우 추정량으로 사용될 수 있는 것은 표본평균, 표본 중앙값, $(X_{(1)} + X_{(n)})/2 \ldots$ 많다. 그럼 어느 추정량이 좋을까(good)?

EXAMPLE 7.2

10명의 학생들에게 학생들의 평균 용돈을 조사시켜 보라. 다를 것이다. 각 학생들이 자기 나름대로 얻은 계산 방식이 추정량이 된다. 당신의 어느 학생 조사 결과를 믿겠는가?

7.2 점 추정치의 Bias, MSE

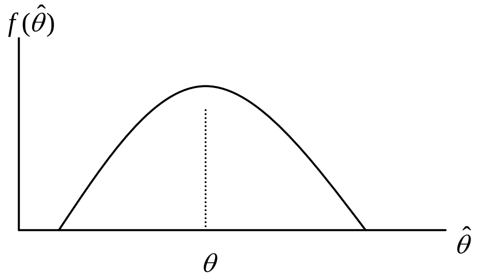


점추정은 과녁에 화살을 쏘는 것과 같다. 모집단으로부터 확률표본을 얻고 이로부터 모수에 대한 예측 값(추정치)을 얻는다. 즉 과녁에 한 발의 화살을 쏘는 것과 같다. 과연 bull-eye일까? 좋은 추정치라면 한 번에 bull-eye일까?

내가 화살을 한 발 쏘아 bull-eye를 했다고 하자. 명궁이라 할 수 있나? 아닐 것이다. 2발, 아니 20발쯤 연속 명중시킨다면 명궁이라 할 수 있을 것이다. 이처럼 한 번의 추정치로는 그 추정치가 좋은지를 판단할 수는 없다. 추정치 good 여부를 판단하려면 추정치를 여러 번 구해야 한다. 즉 추정치의 평균과 분산이 필요하게 되는 것이다.

"목표 모수(θ)"에 대한 추정량을 $\hat{\theta}$으로 표현하자. 추정량 $\hat{\theta}$을 여러 번 얻는다면 그 추정 값은 모수 θ을 중심으로 흩어져 있을 것이다. 모수 부근에 있을 가능성은 높고 멀어질수록 가능성은 떨어질 것이다.

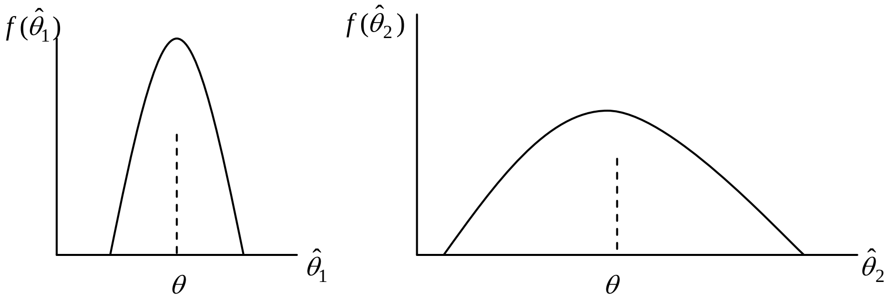
추정량의 분포



DEFINITION (Bias, 편의)

만약 $E(\hat{\theta})=\theta$ 이면 (점) 추정량 $\hat{\theta}$ 는 불편 추정량(unbiased estimator)이라 한다. 추정량의 편의(Bias)는 $B=E(\hat{\theta})-\theta$ 로 정의된다.

$E(\hat{\theta})\neq\theta$ 인 추정량을 편의 추정량(biased estimator)이라 한다. 다음 두 추정량은 모두 불편 추정량이다. 그럼 어떤 추정량이 더 좋은가? 당연히 분산이 적어야 좋은 추정량이다. 즉 $E(\hat{\theta}-\theta)^2$ 을 최소화 하는 추정량이 좋은 추정량일 것이다. $E(\hat{\theta}-\theta)^2$ 은 추정량의 Mean Square Error(평균자승오차)라 정의한다. $MSE(\hat{\theta})=E(\hat{\theta}-\theta)^2$.



THEOREM

$$MSE(\hat{\theta})=E(\hat{\theta}-\theta)^2=V(\hat{\theta})+B^2$$

위의 정리를 보면 불편 추정량인 경우 추정분산($V(\hat{\theta})$)과 $MSE(\hat{\theta})$ 이다

HOMEWORK #12-1 DUE 11월 9일

위의 Theorem을 증명하시오.

EXAMPLE 7.3

$f(x)=\frac{1}{\theta}e^{-x/\theta}$ 으로부터 표본 크기 3인 확률표본 (X_1,X_2,X_3) 을 얻었다. 모수 θ에 대한 추정량으로 다음 4개를 생각해 보자.

① $\hat{\theta}_1=X_1$ ② $\hat{\theta}_2=(X_1+X_2)/2$ ③ $\hat{\theta}_3=(X_1+2X_2)/3$ ④ $\hat{\theta}_4=\overline{X}=(X_1+X_2+X_3)/3$

⑴불편 추정량인 것은?

⑵추정량 중 추정 분산이 가장 작은 추정량은?

EXAMPLE 7.4

$f(x)\sim Uniform(\theta,\theta+1)$ 으로부터 표본 크기 n 인 확률표본 $(X_1,X_2,...,X_n)$ 을 얻었다.

⑴ $\overline{X}$ 이 모수 θ 의 편의 추정량임을 보이시오. 그리고 편의(bias)를 구하시오.

⑵ $\overline{X}$ 의 함수로 모수 θ 의 불편추정량을 구하시오.

⑶ $\overline{X}$ 의 평균자승오차, $MSE(\overline{X})$ 을 구하시오.

EXAMPLE 7.5

$f(x) \sim Normal(\mu,\sigma^2)$ 으로부터 표본 크기 n 인 확률표본 $(X_1, X_2,..., X_n)$ 을 얻었다.

(1)표본분산 $S^2 = \sum_{i=1}^{n}(X-\bar{X})^2/(n-1)$ 은 모집단 분산 $\theta=\sigma^2$ 의 불편 추정량임을 보이시오.

(2)표본 표준편차 $S=\sqrt{S^2}$ 는 모집단 표준편차의 편의 추정량임을 보이시오.

(1)

(2)$\boxed{\text{TIP}}$ ①감마함수의 경우 $E(X^b)=\dfrac{\beta^b\Gamma(\alpha+b)}{\Gamma(\alpha)}$ 이 성립한다. $Gamma(\alpha=n/2, \beta=2)\sim\chi^2(n)$

② $\Gamma(n/2)=\dfrac{(n-2)!\sqrt{\pi}}{2^{(n-1)/2}}$ ③ $E(S)=\dfrac{\sigma}{\sqrt{n-1}}E[(\dfrac{(n-1)S^2}{\sigma^2})^{1/2}]$

HOMEWORK #12-2

$f(x)\sim Poisson(\theta=\lambda)$ 으로부터 표본 크기 n 인 확률표본 $(X_1, X_2,..., X_n)$ 을 얻었다.

(1)표본평균 $\bar{X}$ 가 모수 $\theta=\lambda$ 의 불편 추정량임을 보이시오.

(2) $\bar{X}, \bar{X}^2$ 을 이용하여 $4\lambda+\lambda^2$ 불편 추정량을 구하시오.

HOMEWORK #12-3

$f(x;\theta)=\dfrac{1}{\theta}e^{-x/\theta}$ 표본 크기 n 인 확률표본 $(X_1, X_2,..., X_n)$ 을 얻었다.

(1)표본평균 $\bar{X}$ 가 모수 θ 의 불편 추정량임을 보이시오.

(2) $nX_{(1)}$ 가 모수 θ 의 불편 추정량임을 보이시오.

$\boxed{\text{TIP}}$ 최대값 $X_{(n)}: f_{X_{(n)}}(x)=n[F_X(x_n)]^{n-1}f_X(x_n)$, 최소값 $X_{(3)}: f_{X_{(1)}}(x)=n[1-F_X(x_1)]^{n-1}f_X(x_1)$

HOMEWORK #12-4

$f(x) = Binomial(n, p)$. $\hat{\theta}_1 = \hat{p}_1 = Y/n$, $\hat{\theta}_2 = \hat{p}_2 = (Y+1)/(n+2)$ 두 추정량을 생각해 보자.

(1)두 추정량은 불편 추정량인가?

(2) $MSE(\hat{p}_1), MSE(\hat{p}_2)$ 을 구하고 어는 추정량의 MSE가 적은지 보이시오.

(*)불편 추정량의 MSE가 편의 추정량의 MSE에 비해 항상 적은 것은 아니다.

7.3 불편 추정량...

목표 모수 θ	표본의 크기	점 추정량 $\hat{\theta}$	추정량 평균 $E(\hat{\theta})$	추정량 표준오차 $\sigma_{\hat{\theta}}$
μ	n	$\overline{X}$	μ	$\sigma / \sqrt{n}$
p	n	$\hat{p} = Y/n$	p	$pq / \sqrt{n}$
$\mu_1 - \mu_2$	n_1, n_2	$\overline{X}_1 - \overline{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\sigma_1^2 / n_1 + \sigma_1^2 / n_2}$
$p_1 - p_2$	n_1, n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}$

EXAMPLE 7.6

크기 n 인 확률표본 $(X_1, X_2, ..., X_n)$ 을 평균 μ , 분산 σ^2 인 모집단에서 얻었다고 하자.

$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$ 인 분산의 불편 추정량임을 보이시오.

$$E[\sum (X_i - \overline{X})^2] = E(\sum X_i^2 - n\overline{X}^2) = (n-1)\sigma^2$$

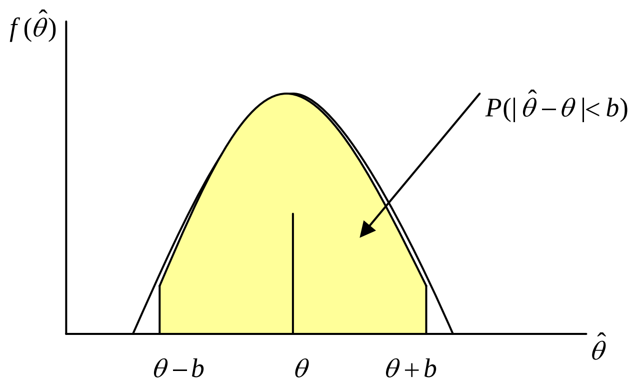
7.4 추정치의 GOODNESS 평가

좋은 추정량이란 (목표) 모수와의 차이가 적은 추정량을 의미한다.

DEFINITION (추정 오차)

추정 오차 (estimation error) ε는 추정량과 모수의 차이(거리)이다. $\varepsilon = |\hat{\theta} - \theta|$

추정량은 확률표본 $(X_1, X_2, ..., X_n)$의 함수이므로 확률변수이다. 예를 들면 모평균 $\theta = \mu$에 대한 추정량 $\overline{X}$는 확률변수이고 확률분포함수로 정규분포를 갖는다.(CLT) 그러므로 추정 오차 ε도 random quantity이므로 추정량 $\hat{\theta}$이 불편 추정량이라면 다음을 생각할 수 있다.



추정 오차에 대한 한계 값으로 b을 생각할 수 있을 것이다. $P(|\hat{\theta}-\theta|<b)$은 추정 오차가 항상 b보다 작다는 것을 의미하는 것은 아니지만 “그럴 가능성 매우 높다”는 것을 의미한다.

$P(|\hat{\theta}-\theta|<b) = \int_{\theta-b}^{\theta+b} f(\hat{\theta})d\hat{\theta} = 0.95$의 의미는 확률 표본을 추출을 100번 정도 했을 때 추정 오차가 b 이하인 것이 95개임을 의미한다. 만약 추정량 $\hat{\theta}$의 확률분포함수(pdf)을 알고 있다면 오차 한계 b을 구할 수 있다. 이 예제에서 얻은 구간이 95% 신뢰구간이 된다.

물론 $\hat{\theta}$의 확률분포함수를 모르더라도 Empirical Rule 혹은 Tchebysheff's Theorem을 이용할 수 있다. $k \ge 1$에서 $b = k\sigma_{\hat{\theta}}$라 놓으면 Tchebysheff's Theorem에 의해 추정 오차가 $b = k\sigma_{\hat{\theta}}$보다 작을 확률은 적어도 $1-1/k^2$ 정도이다. 예를 들어 $k=2$이면 확률은 0.75이다. 만약 추정량의 분포가 종모양이면 0.95이다.

EXAMPLE 7.7

○○대학교 학생들의 cheating 경험 비율을 알아보기 위하여 학생 1000명을 임의 추출(확률

표본)하여 조사하였더니 cheating 경험이 있다고 한 학생은 560이었다. 추정 오차가 추정량의 표준오차 (standard error) 2배가 되도록 설정하시오.

모비율에 대한 추정량은 $\hat{\theta} = \hat{p} = Y/n = 560/1000 = 0.56$ 이다.

추정량의 표준오차 $\sigma_{\hat{p}}$ 은 $\sigma_{\hat{p}} = \sqrt{pq/n}$ 이므로 추정오차 $b = 2\sigma_{\hat{p}} = 2\sqrt{pq/n}$.

$b = 2\sqrt{pq/n} \approx 2\sqrt{(0.56)(0.44)/1000} = 0.03$ (significance? 0.95, why?)

EXAMPLE 7.8

두 회사 타이어 수명을 비교하고자 타이어 100개씩 임의 추출하여 주행 거리를 측정하였다.

$\bar{x}_1 = 26{,}400, s_1^2 = 1{,}440{,}000, n_1 = 100$
$\bar{x}_2 = 25{,}100, s_1^2 = 1{,}960{,}000, n_2 = 100$
두 회사 주행 거리 평균 차이($\mu_1 - \mu_2$)를 추정할 때 추정량의 추정 오차가 표준오차 2배가 되도록 설정하시오.

($\mu_1 - \mu_2$)에 대한 추정량은 $\bar{y}_1 - \bar{y}_2 = 26400 - 25100 = 1300$ 이다.

추정량의 표준오차 $\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이므로

$b = 2\sigma_{\bar{y}_1 - \bar{y}_2} \approx 2\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2\sqrt{\frac{1440000}{100} + \frac{1960000}{100}} = 368.8$ (significance? 0.95, why?)

95% 신뢰구간: $(1300 - 368.8, 1300 + 368.8)$

HOMEWORK #12-5

확률변수 X 을 전구 수명이라 하자. 모집단이 $f(x;\theta) \sim \exp(\theta = \beta)$ 평균이 β 인 지수분포를 따른다고 하자. 확률표본 10개를 얻어 평균 수명을 조사하였더니 1020이었다. 모수 θ 의 점 추정치를 구하고 추정오차가 표준오차 2 배가 되도록 하시오.

7.5 신뢰구간 (confidence interval)

구간 추정이란 확률 표본의 측정치를 이용하여 구간의 한계 값을 계산하는 규칙이다. 추정된 구간은 2가지 성질을 갖는다.

- 구간은 목표 모수를 포함하고 있다.
- 상대적으로 좁은 구간이다.

구간 추정량(interval estimator)을 신뢰구간(confidence interval이라 하고 구간의 극 값을 상한(upper limit), 하한 (lower limit)이라 한다. 신뢰구간이 모수를 포함할 확률을 신뢰계수 혹은 신뢰수준 (confidence level)이라 한다.

$\hat{\theta}_L$ 을 신뢰구간 하한, $\hat{\theta}_U$ 을 신뢰구간 상한이라 하자. 만약 $P(\hat{\theta}_L \le \theta \le \hat{\theta}_U) = 1-\alpha$ 이라면 $1-\alpha$ 는 신뢰수준이다. $[\hat{\theta}_L, \hat{\theta}_U]$ 을 양측 신뢰구간(two-sided confidence interval)이라 한다.

다음은 각각 단측 신뢰구간이라 한다. $P(\theta \le \hat{\theta}_U) = 1-\alpha$ (상한), $P(\hat{\theta}_L \le \theta) = 1-\alpha$ (하한)

Pivotal method

신뢰구간을 구하는 가장 유용한 방법으로 다음 조건을 만족하는 pivotal 통계량을 구한다.

①표본 측정값과 모수 θ 을 함수이다.

②pivotal 통계량의 확률분포함수는 θ 에 의존하지 않는다.

EXAMPLE 7.8

평균이 $\theta = \beta$ 인 지수분포를 따르는 모집단으로부터 모수 θ 을 추정하기 위하여 크기 1인 확률 표본(X_1)을 추출하였다고 하자. 신뢰수준 90% 신뢰구간을 구하시오.

Pivotal 통계량: $U = X_1/\theta$ 조건 만족? 무슨 분포?

$P(U < a) = 0.5$ 만족하는 $a = 0.051$, $P(U < b) = 0.5$ 만족하는 $b = 2.996$

$0.9 = P(X_1/2.996 \le \theta \le X_1/0.051)$ ➔90% 신뢰구간의 ? 표본 추출 100번 하면 이 구간이 모수를 포함할 가능성이 90%

EXAMPLE 7.9

구간이 $[0,\theta]$인 uniform 분포를 모집단으로부터 모수 θ을 추정하기 위하여 크기 1인 확률표본(X_1)을 추출하였다고 하자. 신뢰수준 95% 하한 신뢰구간을 구하시오.

Pivotal 통계량: $U = X_1/\theta$ 조건 만족? 무슨 분포?

$P(U \le a) = 0.95$ 만족하는 $a = 0.95$ 그러므로 $0.95 = P(\frac{X_1}{0.95} \le \theta)$

HOMEWORK #12-6

확률변수 X는 평균이 μ, 분산이 1인 정규분포로부터 추출한 표본이다.

①모집단 평균 μ의 95% 신뢰구간을 구하시오.

②모집단 평균 μ의 95% 하한 신뢰구간을 구하시오.

TIP $(X-\mu)/\sigma \sim Normal(0,1)$

HOMEWORK #12-7

확률변수 X는 평균이 0, 분산이 σ^2인 정규분포로부터 추출한 표본이다.

①모집단 분산 σ^2의 95% 신뢰구간을 구하시오.

②모집단 분산 σ^2의 95% 상한 신뢰구간을 구하시오.

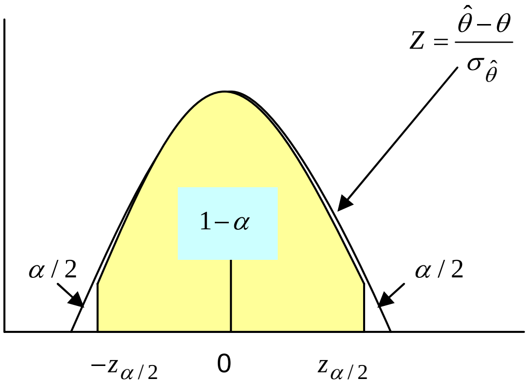
TIP $X^2/\sigma^2 \sim \chi^2(1)$

7.6 대표본 신뢰구간

모수가 $\mu, p, (\mu_1-\mu_2), (p_1-p_2)$인 경우 표본 크기가 크면 $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$는 표준 정규분포를 따른다.

EXAMPLE 7.10

추정량 $\hat{\theta}$가 평균 θ, 표준편차 $\sigma_{\hat{\theta}}$인 정규분포를 따른다고 할 때 $100(1-\alpha)\%$ 신뢰구간?



$$P(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}) = 1-\alpha \;\Rightarrow\; P(\hat{\theta}-z_{\alpha/2}\sigma_{\hat{\theta}} \le \theta \le \hat{\theta}+z_{\alpha/2}\sigma_{\hat{\theta}}) = 1-\alpha$$

$100(1-\alpha)\%$ 신뢰구간은 $[\hat{\theta}-z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta}+z_{\alpha/2}\sigma_{\hat{\theta}}]$

$100(1-\alpha)\%$ 하한 신뢰구간은 $\hat{\theta}-z_{\alpha}\sigma_{\hat{\theta}}$, $100(1-\alpha)\%$ 상한 신뢰구간은 $\hat{\theta}+z_{\alpha}\sigma_{\hat{\theta}}$

EXAMPLE 7.11

고객 64명을 임의 추출하여 쇼핑 시간 평균은 33, 분산은 256이었다. 고객들의 평균 쇼핑 시간 μ 에 대한 90% 신뢰구간을 구하시오.

29.71 / 36.29

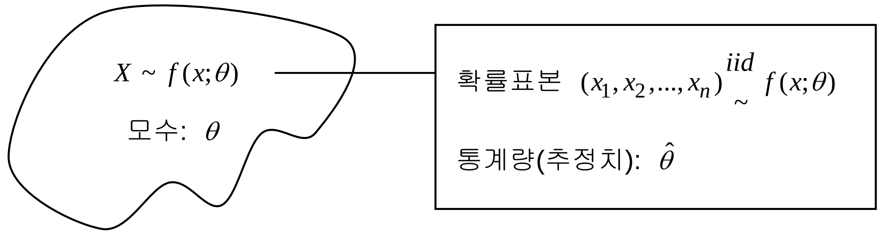
EXAMPLE 3.12

남자 50명, 여자 60명을 임의 추출하여 cheating 경험 여부를 물었더니 남녀 모두 12명이 각각 경험이 있다고 응답하였다. 남녀별 경험 비율의 차이에 대한 98% 신뢰구간을 구하시오.

-0.1451, 0.2251

HOMEWORK #12-8

평균이 10인 지수분포로부터 표본 크기 100인 표본을 추출하고 95% 신뢰구간을 구하시오. 이 과정을 100번하여 95% 신뢰구간을 의미를 해석하시오.



점 추정치: $\hat{\theta}$, 확률표본의 함수이므로 $\hat{\theta}$의 분포함수는 Sampling distribution이다.

- 불편성(unbiasedness): $E(\hat{\theta}) = \theta$
- 최소 분산(variance): $MSE(\hat{\theta}) = V(\hat{\theta}) + B^2$, 편의: $B = E(\hat{\theta}) - \theta$

구간 추정: $(\hat{\theta}_L, \hat{\theta}_U)$

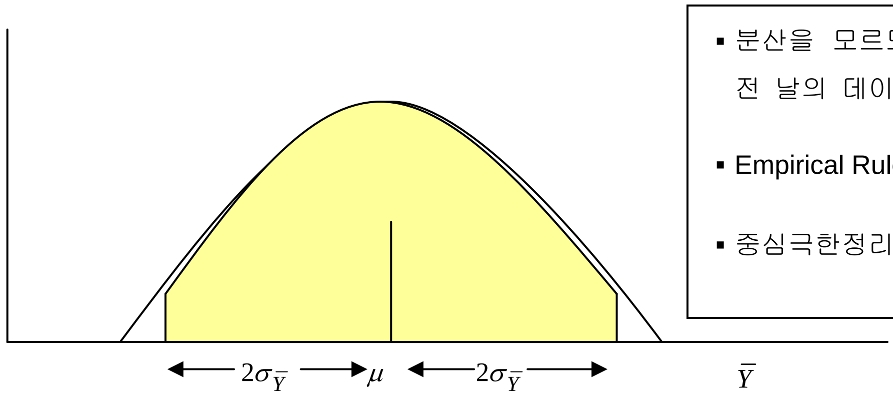
- Pivotal 통계량: $P(\hat{\theta}, \theta) \sim pdf\ does\ not\ depend\ \theta$ (예) $X \sim \exp(\theta) \Rightarrow \dfrac{X}{\theta} \sim \exp(1)$
- 대표본 분포: $\dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim Normal(0,1)$ (예) $\dfrac{\overline{X} - \mu(=\theta)}{\sigma / \sqrt{n}} \sim Normal(0,1)$

7.7 표본 크기 설정

실험 설계란 정보(상품)를 구매하는 행위이다. 원하는 만큼의 정보를 최소의 가격(표본 데이터 개수)으로 얻으면 좋은 구매 행위이다.

예를 들어 철강 생산 공장 하루 평균 생산량 μ을 대한 추정을 하려고 한다고 하자. 신뢰수준 95%에서 추정 오차(Δ)가 5톤 이내가 되게 하려면 표본의 크기를 얼마로 하면 될까?

추정 오차가 $2\dfrac{\sigma}{\sqrt{n}}$이므로 $2\dfrac{\sigma}{\sqrt{n}} = 5$을 풀면 $n = \dfrac{4\sigma^2}{25}$ (최적 표본 크기).



- 분산을 모르므로 사전 정보 이용. 이전 날의 데이터 이용 분산 추정.
- Empirical Rule: $2\sigma_{\overline{Y}}$
- 중심극한정리: $1.96\sigma_{\overline{Y}}$

대표본 추정의 경우 $Z = \dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim Normal(0,1)$ 이므로 추정 오차는 $z_{\alpha/2}\sigma_{\hat{\theta}}$ 이다.

EXAMPLE 7.13

철강 생산 공장 하루 평균 생산량 μ 을 대한 추정을 하려고 한다고 하자. 신뢰수준 95%에서 추정 오차(Δ)가 5톤 이내가 되게 하려면 표본의 크기를 얼마로 하면 될까? (대표본) 그리고 모집단의 표준오차가 21이라는 사전 정보가 있다고 하자.

$1.96\dfrac{\sigma(=21)}{\sqrt{n}} = 5$ ➔ $n = 67.8$ 이므로 표본의 크기를 68개로 한다.

EXAMPLE 7.14

행정수도 법안 찬성률을 조사하고자 한다. 추정 오차가 0.04미만이고 신뢰수준을 90%로 하고자 한다. 모집단 찬성률(p)이 0.6이라는 사전 정보가 있다고 한다. 혹은 모집단 비율 p가 0.6 근처에 놓이게 하려고 한다.

$z_{\alpha/2}\sigma_{\hat{\theta}} = 0.04$ ➔ $1.645\sqrt{\dfrac{pq}{n}} = 0.04$ ➔ $1.645\sqrt{\dfrac{0.6 \times 0.4}{n}} = 0.04$ ➔ $n = 406$

EXAMPLE 7.15

은행 A에서 기다리는 시간 평균과 은행 B에서 기다리는 시간 평균의 차이를 조사하고자 한다. 각 은행의 기다리는 시간을 조사하였더니 범위는 8분이었다. 신뢰수준 95%에서 추정 오차가 1분 미만이 되게 하려고 한다. 표본의 크기를 얼마로 하면 되나?

$z_{\alpha/2}\sigma_{\hat{\theta}} = 1$ ➔ 두 집단의 표본의 크기가 동일하다면 ➔ $n = 30.73$ 즉 31로 한다.

HOMEWORK #13-1

지난번 투표율은 0.65였다. 이번이 있을 투표율 조사에서 추정 오차가 0.02 이내이고 신뢰수준이 95%가 되게 하려면 표본의 크기를 얼마로 해야 하나?

HOMEWORK #13-2

철강 하루 생산량은 최저 10통, 최대 30톤이라 한다. 신뢰수준 90%에서 추정오차가 2톤 이내가 되게 하려면 표본의 크기를 얼마로 해야 하나?

HOMEWORK #13-3

두 지역의 투표율 차이를 추정하고자 한다. 90% 신뢰수준에서 추정 오차가 0.05 이내가 되도록 하고자 할 때 표본의 크기는 얼마로 해야 하나? 표본의 크기는 동일하게 뽑는다고 가정하자. 지난 선거의 투표율은 각각 55%, 60%였다고 하자.

7.8 μ 와 $\mu_1-\mu_2$ 의 소표본(small sample) 신뢰구간

대표본의 경우 모집단의 분포와 상관 없이 μ 의 좋은 추정량 $\overline{X}$ 는 정규분포를 따른다. 그러므로 대표본 신뢰구간에 의해 모집단 평균 μ 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{V(\hat{\theta})} \Rightarrow \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

같은 방법으로 $\mu_1-\mu_2$ 의 $100(1-\alpha)\%$ 신뢰구간은 $(\overline{X}_1-\overline{X}_2)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}}$

각각의 분산을 모를 때는 표본 분산으로 대체(추정)하면 된다.

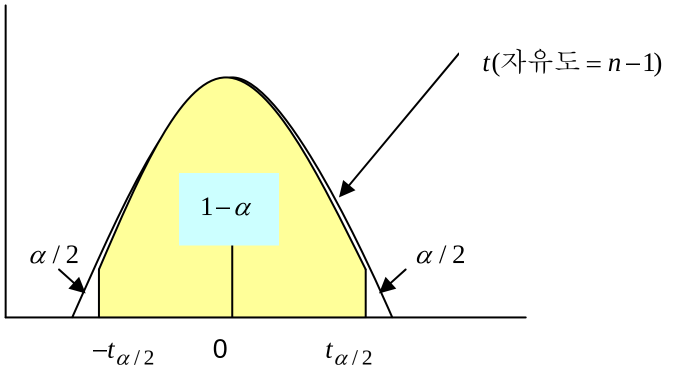
만약 대표본이 아니라면 이것이 성립하지 않는다. 그럼 소표본($n<20\sim30$)일 때는?

모집단이 정규분포임을 가정하면 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim Normal(0,1)$ 이고 $\frac{(n-1)S_2}{\sigma^2} \sim \chi^2(n-1)$ 이고 $\overline{X}$ 와 S^2 은 서로 독립이다.(2.4절, 2.5절 참고) 그러므로 다음이 성립한다.

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

그러므로 소표본인 경우 모집단 평균 μ 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$$



EXAMPLE 7.16

철강 하루 생산량 평균을 추정하기 위하여 16일을 조사하여 다음 결과를 얻었다. 표본 평균은 25톤이고 분산은 1.6톤이었다. 하루 생산량 평균에 대한 95% 신뢰구간을 구하시오. 단 하루 철강 생산량의 분포는 정규분포를 따른다고 한다.

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \Rightarrow 25 \pm 2.131 \times \frac{0.4}{\sqrt{16}} \Rightarrow (24.79, 25.21)$$

EXAMPLE 7.17

철강 하루 생산량 평균을 추정하기 위하여 25일을 조사하여 다음 결과를 얻었다. 표본 평균은 25톤이고 분산은 1.6톤이었다. 하루 생산량 평균에 대한 95% 신뢰구간을 구하시오.

$$\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}} \Rightarrow 25 \pm 1.96 \times \frac{0.4}{\sqrt{25}} \Rightarrow (24.843, 25.157)$$

독립인 두 모집단 평균 차이 $\mu_1 - \mu_2$ 에 대한 신뢰구간을 구해보자. 각 집단으로부터 표본의 크기 n_1, n_2 인 확률표본(random sample)을 얻어 평균($\overline{X}_1, \overline{X}_2$)과 분산($S_1^2, S_2^2$)을 얻었다고 하자.

좋은 점 추정치로 $\overline{X}_1 - \overline{X}_2$ 을 생각할 수 있다. 각 표본평균의 정규분포를 따르므로 다음도 성립한다.

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n2}}} \sim Normal(0,1)$$

만약 두 모집단의 분산이 동일하다고 가정하면(등분산 equal variance 가정, $\sigma_1^2 = \sigma_2^2 = \sigma^2$),

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n2}}} \sim Normal(0,1)$$

다음을 생각해보자. 이를 통합분산(pooled variance)라 한다.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \quad \text{why?}$$

그러므로

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n2}}} \sim t(n_1 + n_2 - 2)$$ 이로부터 $\mu_1 - \mu_2$ 의 $100(1-\alpha)\%$ 신뢰구간은

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

EXAMPLE 7.18

새로운 교육 방법의 효과를 보기 위하여 조립 시간(분)을 조사하였다. 두 방법의 조립 시간 평균의 차이에 대한 95% 신뢰구간을 구하시오. 모집단은 정규분포를 따른다.

Standard:	32	37	35	28	41	44	35	31	34
New:	35	31	29	25	34	40	27	32	31

$$\overline{X}_1 = (32 + 37 + ... + 34)/9 = 35.22$$
$$\overline{X}_2 = (35 + 31 + ... + 31)/9 = 31.56$$

$$S_1^2 = [(32 - 35.22)^2 + (37 - 35.22)^2 + ... + (34 - 35.22)^2]/8 = 195.56/8$$
$$S_2^2 = [(35 - 31.56)^2 + (31 - 31.56)^2 + ... + (31 - 31.56)^2]/8 = 160.22/8$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{195.56+160.22}{16} = 22.24 \text{ 그러므로 } S_p = 4.71$$

$$(35.22-31.56) \pm 2.12 * 4.71 * \sqrt{\frac{1}{9}+\frac{1}{9}} \Rightarrow 3.66 \pm 4.71$$

신뢰구간이 0을 포함하고 있으므로 두 모집단의 평균 차이는 유의하다고 할 수 없다.

in SAS

```
data a;
    input x g $ @@;
    cards;
    32  s 37 s  35 s  28 s  41 s  44 s  35 s 31 s  34 s
    35  t 31 t 29 t  25 t  34 t  40 t  27 t  32 t  31 t
run;

proc ttest data=a alpha=0.05;
    class g;
    var x;
run;
```

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
x	Pooled	Equal	16	1.65	0.1185
x	Satterthwaite	Unequal	15.8	1.65	0.1187

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
x	Folded F	8	8	1.22	0.7849

Variable	g	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev
x	s	9	31.422	35.222	39.023	3.3395	4.9441	9.4718
x	t	9	28.116	31.556	34.996	3.0228	4.4752	8.5735
x	Diff (1-2)		-1.046	3.6667	8.379	3.512	4.7155	7.1767

HOMEWORK #13-4

철강 생산 시 식히는 과정에서 소금 물을 사용하는 방법과 오일을 사용하는 방법 중 어느 것이 강도를 높이는지 알아보기 위하여 다음과 같이 측정 자료를 얻었다. 강도는 정규 분포를 따른다고 하자.

소금물:	145	150	153	148	141	152	146	154	139	148
오일:	152	150	147	155	140	146	158	152	151	143

두 방법의 강도 평균 차이에 대한 90% 신뢰구간을 구하시오.

(1)수 작업으로 계산하시오.

(2)SAS를 이용하여 구하시오.

HOMEWORK #13-5

HOMEWORK #13-5에서 오일을 이용하였을 경우(소금물 집단에 대한 데이터가 없다고 가정하자) 강도의 평균에 대한 90% 신뢰구간을 구하시오. SAS를 이용하여 구하시오.

```
data b;
    input y @@;
    cards;
    145   150   153   148   141   152   146   154   139   148
run;

proc univariate data=b alpha=0.1 cibasic;
    var y;
run;
```

(소금물)

적률

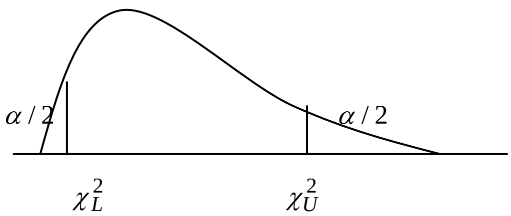
N	10	가중합	10
평균	147.6	관측치 합	1476
표준편차	4.97102717	분산	24.7111111
왜도	-0.4822006	첨도	-0.6379006
제곱합	218080	수정 제곱합	222.4
변동계수	3.36790459	평균의 표준오차	1.57197682

정규성 가정하 기본 신뢰 한계

모수	추정값	90% 신뢰한계	
평균	147.60000	144.71839	150.48161
표준편차	4.97103	3.62560	8.17832
분산	24.71111	13.14500	66.88495

7.9 모분산 σ^2 신뢰구간

모집단이 정규분포($Normal(\mu,\sigma^2)$)로부터 확률표본 $(X_1, X_2,..., X_n)$ 에 대해 다음이 성립함을 알고 있다. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$



사실 카이제곱 분포는 좌우 대칭이 아니므로 자유도에 따라 최소 구간이 달라지게 된다. 그러나 이렇게 구하는 것이 매우 복잡하므로 양쪽에 $\alpha/2$를 할당하게 된다.

$$P[\chi_L^2 \le \frac{(n-1)S^2}{\sigma^2} \le \chi_U^2] = 1-\alpha \;\; \Rightarrow \;\; P[\frac{(n-1)S^2}{\chi_U^2} \le \sigma^2 \le \frac{(n-1)S^2}{\chi_L^2}] = 1-\alpha$$

그러므로 모분산 σ^2 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left(\frac{(n-1)S^2}{\chi^2_{(\alpha/2)}}, \frac{(n-1)S^2}{\chi^2_{(1-\alpha/2)}}\right)$$

EXAMPLE 7.18

페이지 70의 예제 프로그램에서 소금물 방법의 경우 강도에 대한 95% 모분산 신뢰구간을 구하시오.

$$\left(\frac{(n-1)S^2}{\chi^2_{(1-\alpha/2)}}, \frac{(n-1)S^2}{\chi^2_{(\alpha/2)}}\right) \Rightarrow \left(\frac{(10-1)24.71}{\chi^2_{(\alpha/2)}(\text{자유도}=9)=19.02}, \frac{(10-1)24.71}{\chi^2_{(1-\alpha/2)}=2.7}\right)$$ 은 (13.15, 66.88)

HOMEWORK #13-6

HOMEWORK 13-5에서 오일 방법의 경우 강도에 대한 **90%** 모분산 신뢰구간을 구하시오.