1. 추정

1-1. 점추정

① 모집단으로부터 얻은 표본에 기초하여 모수의 가장 그럴듯한 값을 하나의 수치로 구해내는 것이다.
② 확률변수 X가 미지의 모평균 μ와 모분산 σ^2을 갖는 정규분포에 따른다고 가정할 때, 표본평균 $\overline{X}$는 모평균 μ의 하나의 점추정치이다.
③ X의 모집단에서 다섯 개의 표본을 추출하여 25, 30, 29, 31, 28을 얻었다고 가정할 때, 모평균의 점추정치는 28.6이다.
④ 추정량은 확률변수들로부터 만들어진 하나의 확률변수이므로 추출된 표본의 값에 따라 그 값 즉, 추정치가 달라질 수 있다.
⑤ 추정치의 변동은 추정량의 정확도와 관계가 있으며, 이 정확도를 측정하기 위해 추정량의 표준편차를 계산할 필요가 있는데, 바로 표준오차(SE)이다.

$$E(\overline{X}) = \mu,\ SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

⑥ 어떤 알루미늄 합금의 열전도 X는 평균 μ와 표준편차 $\sigma = 0.284$의 정규분포를 따르는 것으로 알려져 있다. μ를 추정하기 위하여 $n = 10$번의 열전도 측정실험을 실시한 결과 42.26, 41.72, 42.18, 41.86, 41.81, 42.04, 41.60, 41.95, 42.34, 41.48을 얻었다. 이때, 표본의 평균이 41.924이고 표본평균의 표준오차는 $\dfrac{0.284}{\sqrt{10}} = 0.0898$이다.
⑦ σ는 모집단의 표준편차이며 하나의 모수이므로 σ를 모르는 경우가 있다. 이러한 경우에는 표본표준편차

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

를 이용하여 σ를 추정할 수 있다.
⑧ 자주 사용되는 추정량의 예

모수	추정량
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
p	$\hat{p} = \frac{X}{n}$

⑨ 추정량의 바람직한 성질 중 하나는 그 추정치의 기댓값이 모수의 참값과 일치하는 것이다. 이러한 성질을 가진 추정량을 불편추정량(unbiased estimator)이라고 한다.

〖예〗 어린 소나무의 성장을 연구하기 위하여 1년생 붉은 소나무 묘목 40그루의 크기를 조사한 자료가 다음과 같다. 해당 자료를 이용하여 전체 1년생 붉은 묘목의 평균 크기에 대한 추정치와 표준오차는 얼마인가?

2.6	1.9	1.8	1.6	1.4	2.2	1.2	1.6	1.6	1.5	1.4	1.6	2.3
1.5	1.1	1.6	2.0	1.5	1.7	1.5	1.6	2.1	2.8	1.0	1.2	1.2
1.8	1.7	0.8	1.5	2.0	2.2	1.5	1.6	2.2	2.1	3.1	1.7	1.2

[풀이]

㉠ 모평균 추정치 $\frac{1}{40}\times(2.6+\ldots+1.2)=1.715$

㉡ 표준분산 $s^2=\frac{1}{40-1}\left\{(2.6-1.715)^2+\ldots+(1.2-1.715)^2\right\}=0.2254$

㉢ 표준오차의 추정치 $\frac{s}{\sqrt{n}}=\frac{\sqrt{0.2254}}{\sqrt{40}}=0.0751$

1-2. 구간추정

1-2-1. 신뢰구간과 해석하기

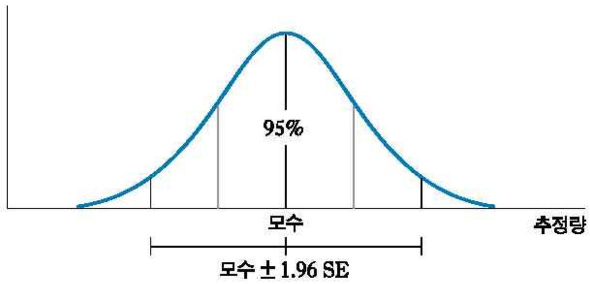
① 신뢰구간이란 모수를 포함한다고 확신하는 구간을 의미한다.

② 여기서 '확신한다' 는 뜻은 '높은 확률을 가진다' 는 의미이며 '신뢰구간이 모수를 포함할 확률' 을 신뢰계수(confidence coefficient) $1-\alpha$ 라 한다.

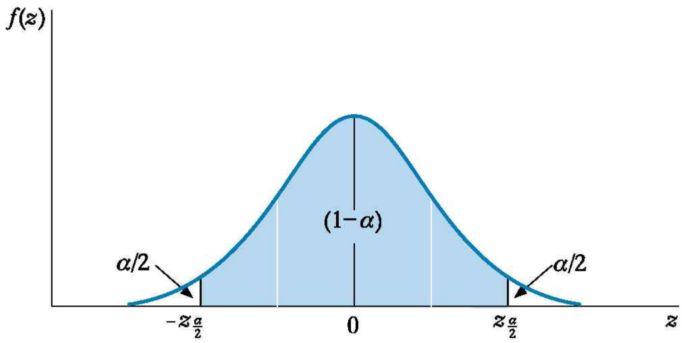
③ 많은 실험자들은 95% 신뢰구간을 자주 구한다.
(신뢰계수 또는 그 구간이 모수를 포함할 확률이 0.95라는 의미)

④ 0.90, 0.95, 0.98, 0.99 등으로 바꾸어 확신의 정도를 증가시키거나 감소시킬 수 있다.

⑤ '신뢰계수가 0.95이다.' 의 의미는?
; 선택 가능한 추정량의 모든 가능한 값들의 95%는 (모수$\pm 1.96SE$) 구간 안에 있을 것이다.
그러나 모수의 값을 모르기 때문에 (추정량$\pm 1.96SE$) 인 구간을 고려한다.



⑥ 표준정규분포에서 신뢰계수 $1-\alpha$는 $-z_{\alpha/2}$(신뢰하한, lower confidence limit :LCL)와 $z_{\alpha/2}$(신뢰상한, upper confidence limit :UCL) 사이의 면적이다.



⑦ 신뢰구간으로 자주 사용되는 Z값들

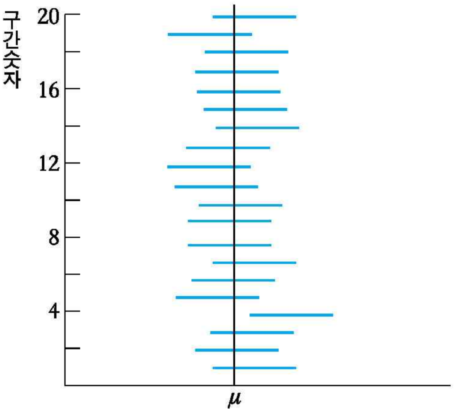
신뢰계수$(1-\alpha)$	α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
0.90	0.10	0.05	1.645
0.95	0.05	0.25	1.96
0.98	0.02	0.01	2.33
0.99	0.01	0.005	2.58

⑧ 성인 남성의 일일 평균 유제품 섭취에서
㉠ 99% 신뢰구간 : 743.23~ 768.77 (g)
㉡ 95% 신뢰구간 : 746.30~ 765.70 (g)
⇒ 구간이 넓어지므로 그 만큼 신뢰도를 증가시키게 됨

⑨ 구간의 너비를 증가시키지 않고 신뢰도를 높이는 유일한 방법은 표본의 크기 n을 증가시키는 것이다.

⑩ 모평균 μ의 참값이 주어진 구간 안에 있다고 95% 확신한다는 것은 무엇을 의미할까?

: 만약 이러한 구간을 20개 만들었고, 각 구간을 서로 다른 표본의 정보를 이용하여 만들었다면, 그 구간들은 20개의 구간 중 95% 즉, 100개 정도 추정했을 때 95번 정도(또는 20개 중에서 19개)는 그들의 하한과 상한 사이에 μ를 포함할 것으로 기대한다는 의미이다.



1-2-2. 모평균 X에 대한 구간추정

① 표본의 크기 n이 클 때, 표본평균 $\overline{X}$는 모평균 μ에 대한 가장 적합한 점추정량(=최우점추정량)이다.

② 모집단의 표준편차가 σ임을 알 때 신뢰구간

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

③ 모집단의 표준편차를 알지 못하고, 표본의 표준편차가 s임을 알 때 근사 신뢰구간

$$\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

〚예〛 어느 과학자가 남성들의 일일 평균 유제품 섭취량에 대하여 알기 위해 $n = 50$명의 남성을 임의표본을 선택했다. 그 결과 유제품의 일일 평균 섭취량은 하루당 $\overline{X} = 756$그램이고 표준편차는 $s = 35$그램이었다. 이 표본정보를 이용하여 남성들의 일일 평균 유제품 섭취량에 대한 95% 신뢰구간을 만들어라.

[풀이]

표본의 크기가 50이므로 충분히 크기에 표본평균 $\overline{X}$는 근사적으로 정규분포를 따른다.
따라서 근사 95% 신뢰구간은

$$\overline{X} \pm 1.96\frac{s}{\sqrt{n}} = 756 \pm 1.96\frac{35}{\sqrt{50}} = 756 \pm 9.70$$

남성들의 일일 평균 유제품 섭취량에 대한 95% 신뢰구간은 하루당 746.30그램에서 765.70그램이다.

④ 모집단의 표준편차를 알지 못할 때에는 근사된 신뢰구간을 구하기 위해 σ의 근삿값으로 s를 대체하였다. 따라서 이때 언급된 신뢰계수는 0.95가 아니라고 볼 수 있다. 그러나 실제로 이 값들의 차이는 극히 미미하기에 통계에서 사용되는 대부분의 구간추정량들은 근사된 신뢰구간들이다. 앞으로 신뢰구간을 언급할 때에는 근사된 것으로 이해하기로 한다.

1-2-3. 모비율 p에 대한 구간추정

① 표본의 크기가 충분히 크다고 가정한다.

② 표본비율 $\hat{p}=\dfrac{X}{n}$가 모비율 p에 대한 최우 추정량이다.

③ 표본비율 $\hat{p}$의 표본분포는 n이 크면 근사적으로 정규분포이며, 평균 p와 표준오차 $SE=\sqrt{\dfrac{pq}{n}}$을 가진다.

④ p와 q를 알지 못하기 때문에, 최우추정량 $\hat{p}$와 $\hat{q}$를 이용하여 추정한다. 즉, $\hat{p}$를 이용한 신뢰구간은

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}<p<\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

이다.

[예] 한 정당에서 다가오는 선거의 예비 유권자 중 985명을 임의 선택하여 전화조사를 실시하였다. 이 조사에서 592명이 A 정당에 투표하겠다고 답하였다면, A 정당에 투표할 모집단의 유권자 비율 p에 대한 90% 신뢰구간을 구하시오. 이 정보에 근거하여 A 정당후보가 선거에서 당선되겠는지 판단하시오.

[풀이]

㉠ p에 대한 점추정치는

$$\hat{p}=\frac{X}{n}=\frac{592}{985}=0.601$$

㉡ 표준오차는

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}=\sqrt{\frac{(0.601)(0.399)}{985}}=0.016$$

㉢ 90% 신뢰구간에 대한 Z값은 Z분포의 상한 꼬리부분에서 면적 $\alpha/2$=0.05를 가지는 값이므로 $z_{0.05}=1.645$이다. 따라서, p에 대한 90% 신뢰구간은

$$\hat{p}\pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}=0.601\pm 0.026$$

또는 $0.575<p<0.627$이다.

㉣ 이 정당에 투표할 유권자의 백분율 추정치는 57.5%와 62.7% 사이에 있다. 그렇다면 이 정당후보가 선거에서 승리할까? 투표의 50% 이상을 얻어야 승리하고 신뢰구간의 상한과 하한이 모두 이 최솟값 50%를 넘기 때문에, 90%의 신뢰도를 가지고 이 후보가 당선될 것이라 말 할 수 있다.