제 4 장 통계적 추정과 가설검정

4.3 통계적 추정

교재 : 사범대생을 위한 확률과 통계, 장세경 지음, 경문사, 2012

- 조사할 대상의 모집단으로부터 임의로 표본을 추출하고 표본의 자료에 함축된 정보를 분석하여 모집단의 특성을 찾아내는 과정을 통계적 추론이라고 한다.

- 통계적 추론의 종류는
  표본을 이용하여 모집단의 미지의 모수를 예측하는 추정과 모집단에 대한 어떤 예상이나 추측의 타당성 여부를 확인하여 채택 또는 기각을 결정하는 가설검정이 있다.

- 추정의 종류는 표본으로부터 모집단의 미지의 모수를 예측하는 점추정과 모집단의 미지의 모수가 포함될만한 구간, 즉 신뢰구간을 예측하는 구간추정이 있다.

4.3.1 점추정

- 점추정 : 추출한 표본으로부터 모집단의 미지의 모수를 하나의 값으로 예측하는 추정방법.
- 모수의 추정에 사용되는 통계량을 추정량이라 하고 추정량에 관측값을 대입하여 얻은 추정량의 값을 추정값이라고 한다.
- 추정량은 여러 관측값을 대입하여 추정값을 구하는 확률변수이고, 추정량에 대입하는 추정값은 실수값이다.

정의 17
모집단으로부터 추출한 표본의 정보를 이용하여 미지의 모수의 참값으로 생각되는 하나의 값을 추측하는 방법을 점추정이라고 한다.

참고. 일반적으로 점추정을 할 때, 예측된 값이 모수의 참값과 일치하는 것은 거의 기대할 수 없다. 그러나 모집단에서 반복하여 표본을 추출한다면 각각의 표본에서 계산된 표본의 통계량은 매번 다를 수 있지만 평균적으로 모수의 주위에 밀집하게 될 것이다. 이때, 모수 주위에 가까울수록 좋은 추정량이 되고 멀수록 나쁜 추정량이 된다. 추정량이 좋고 나쁨을 판별하는 준거로는 불편성, 유효성, 일치성, 충분성이 있다. 이 중에서 표본이 치우침없이 추출되었다고 하고 표본으로부터 계산된 통계량의 평균이 추정하려는 모수가 될 때, 이 통계량을 불편추정량이라고 한다.

정의 18

모수 θ에 대하여,

$$E(\hat{\theta}) = \theta$$

이면, $\hat{\theta}$은 θ의 **불편추정량**이다.

예 : 앞에서 표본평균 $\overline{X}$의 평균과 모평균이 일치하는 것과 표본비율 $\hat{p}$의 평균과 모비율 p가 일치하는 것을 보았다. 즉 $E(\overline{X}) = \mu$이므로 표본평균 $\overline{X}$는 모평균 μ의 불편추정량이고 $E(\hat{p}) = p$이므로 표본비율 $\hat{p}$은 모비율 p의 불편추정량이다.

마찬가지로 표본분산 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$의 기댓값은

$$E(S^2) = \sigma^2$$

이다. 즉 표본분산 S^2은 모분산 σ^2의 불편추정량이다.

$\widehat{\theta}_1$과 $\widehat{\theta}_2$가 모두 모수 θ의 불편추정량일 때, $\widehat{\theta}_1$과 $\widehat{\theta}_2$ 중 표준오차가 더 작은 추정량을 유효추정량이라고 한다. 여기서 표준오차는 각 추정량 $\hat{\theta}$의 표준편차를 말하고 $\widehat{SE}(\hat{\theta})$으로 표기한다.

표본평균 $\overline{X}$의 표준오차는 $\widehat{SE}(\overline{X}) = S(\overline{X}) = \sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$이나 σ가 미지이므로 표본표준편차 S로 대신한다. 즉, $\widehat{SE}(\overline{X}) = \frac{S}{\sqrt{n}}$이다. 그리고 표본비율 $\hat{p}$의 표준오차는

$$\widehat{SE}(\hat{p}) = S(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{\frac{pq}{n}}$$ 이다. (단, $q = 1 - p$)

이다.

정의 19

모수 θ의 불편추정량 $\widehat{\theta}_1$과 $\widehat{\theta}_2$에 대하여

$$\frac{V(\widehat{\theta}_1)}{V(\widehat{\theta}_2)} < 1$$

이면 $\widehat{\theta}_1$은 θ의 유효추정량이고,

$$\frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} < 1$$

이면 $\widehat{\theta}_2$은 θ의 유효추정량이다.

예 : 위에서 정리한 표본평균 $\overline{X}$, 표본분산 S^2, 표본비율 $\hat{p}$은 다른 어떤 불편추정량보다 표분편차가 작은 유효추정량이다.

예제 21. 어떤 모집단의 평균이 μ이고 분산이 σ^2일 때, 임의로 추출한 표본 X_1과 X_2에 대하여 모평균 μ의 추정량이 각각 $\widehat{\mu_1} = \dfrac{X_1 + X_2}{2}$, $\widehat{\mu_2} = \dfrac{2X_1 + 3X_2}{5}$라고 한다. 다음 물음에 답하시오.

(1) $\widehat{\mu_1}$과 $\widehat{\mu_2}$가 불편추정량임을 보이시오.

(2) $\widehat{\mu_1}$과 $\widehat{\mu_2}$ 중에서 유효추정량을 구하시오.

풀이.

정의 20

표본의 크기가 n인 모수 θ의 추정량을 $\hat{\theta}$라 할 때, 임의의 $\epsilon > 0$에 대하여

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \epsilon\right) = 1$$

이면 $\hat{\theta}_n$을 θ의 일치추정량이라고 한다.

예 : 한 개의 동전을 던지는 시행에서 앞면이 나올 확률은 동전을 던지는 횟수가 많아질수록 $\frac{1}{2}$에 가까워진다는 것이 일치추정량의 의미이다.

위에서 정리한 표본평균 $\overline{X}$, 표본분산 S^2, 표본비율 $\hat{p}$은 모두 일치추정량이다.

정의 21

모집단으로부터 추출한 표본의 정보를 모두 사용한 통계량을 충분통계량이라고 한다.

예 : 위에서 정리한 표본평균 $\overline{X}$, 표본분산 S^2, 표본비율 $\hat{p}$은 표본의 모든 정보 X_1, X_2, $\cdots$, X_n을 사용하여 계산하였기 때문에 충분통계량이다.

그러나 산포도의 측도 중 범위는 표본의 최댓값과 최솟값만으로 계산되기 때문에 충분통계량이 될 수 없다.

이상에서 살펴보았듯이 표본평균 $\overline{X}$는 네 가지 추정량의 조건을 모두 만족하므로 모평균 μ에 대한 바람직한 점추정량이라고 할 수 있다.

참고 각 모수의 점추정량

(1) 모평균 μ의 추정량과 표준오차 : $\hat{\mu} = \overline{X}$, $\widehat{SE}(\overline{X}) = \dfrac{S}{\sqrt{n}}$

(2) 모분산 σ^2의 추정량과 표준오차 : $\widehat{\sigma^2} = S^2$, $\widehat{SE}(S^2) = S$

(3) 모비율 p의 추정량과 표준오차 : $\hat{p} = \dfrac{X}{n}$, $\widehat{SE}(\hat{p}) = \sqrt{\dfrac{pq}{n}}$ (단, $q = 1 - p$)

4.3.2 구간추정

정의 22

모집단으로부터 추출한 표본의 정보를 이용하여 미지의 모수의 참값 θ가 속해 있을 것이라고 생각되는 구간 $(L,\ U)$를 추측하는 방법을 구간추정이라고 한다.

- 구간 $(L,\ U)$가 길수록 모수 θ가 속해있을 확률은 높아지고, 짧아질수록 모수 θ에 근접할 확률이 높아진다. 이때, 확률은 신뢰수준 또는 신뢰도라 하고 $1-\alpha$라고 나타낸다.
- 신뢰수준은 0.9, 0.95, 0.99 등이 사용된다.
- 신뢰수준의 여사건의 확률, 즉 오류를 범할 확률의 최대허용한계를 유의수준 또는 위험률이라 하고 α로 나타낸다.
- 구간 $(L,\ U)$를 모수 θ의 $100(1-\alpha)\%$ 신뢰구간이라고 한다.

(1) 모평균의 구간추정

이것은 모평균 μ의 추정량인 표본평균 $\overline{X}$의 분포를 이용한다.

정규분포 $N(\mu,\, \sigma^2)$을 따르는 모집단에서 추출한 크기가 n인 표본을 $X_1,\ X_2,\ \cdots,\ X_n$이라 할 때,

$$\overline{X} \sim N\left(\mu,\, \frac{\sigma^2}{n}\right)\text{이고, 표준화변수 } Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\text{로 변환하면 } Z \sim N(0,\, 1)\text{이다.}$$

모분산 σ^2을 모르더라도 표본의 크기 n이 크면($n \geq 30$) 중심극한정리에 의하여 모분산의 추정량인 S^2으로 구성된 정규분포로 근사하므로

$$\overline{X} \sim N\left(\mu,\, \frac{S^2}{n}\right)\text{이고, 표준화변수 } Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}\text{로 변환하면 } Z \sim N(0,\, 1)\text{이다.}$$

이때, 표준정규분포곡선의 양쪽 꼬리 부분의 곡선 아래의 넓이가 $\frac{\alpha}{2}$가 되는 경계값을 각각 $-z_{\alpha/2}$, $z_{\alpha/2}$라고 하면

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

이고 $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ 또는 $Z = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$이므로 위의 확률변수의 범위는

$$-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \quad \text{또는} \quad -z_{\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < z_{\alpha/2}$$

이다. 두 식을 모평균 μ로 정리하면

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

이고 구간으로 나타내면

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{또는} \quad \left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

이다. 이것을 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간이라고 한다.

참고 모분산 σ^2을 알거나 표본분산 S^2을 알 수 있는 경우(또는 $n \geq 30$)의 모평균의 특별한 신뢰구간

(1) 90% 신뢰구간(즉, 유의수준 $\alpha = 0.1$) :

$$\left(\overline{X} - 1.645 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right) \text{ 또는 } \left(\overline{X} - 1.645 \times \frac{S}{\sqrt{n}},\ \overline{X} + 1.645 \times \frac{S}{\sqrt{n}}\right)$$

(2) 95% 신뢰구간(즉, 유의수준 $\alpha = 0.05$) :

$$\left(\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \text{ 또는 } \left(\overline{X} - 1.96 \times \frac{S}{\sqrt{n}},\ \overline{X} + 1.96 \times \frac{S}{\sqrt{n}}\right)$$

(3) 99% 신뢰구간(즉, 유의수준 $\alpha = 0.01$) :

$$\left(\overline{X} - 2.58 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 2.58 \times \frac{\sigma}{\sqrt{n}}\right) \text{ 또는 } \left(\overline{X} - 2.58 \times \frac{S}{\sqrt{n}},\ \overline{X} + 2.58 \times \frac{S}{\sqrt{n}}\right)$$

예제 22. 어떤 고등학교 3학년 학생들의 평균 수면시간을 알아보기 위해 100명을 조사한 결과 평균이 5시간이었다. 학생들의 평균 수면시간의 95% 신뢰구간을 구하시오. (단, 학생들의 수면시간은 정규분포를 따르고 표준편차는 1시간이다.)

풀이.

예제 23. 어떤 대학교 신입생들의 평균 나이를 알아보기 위해 144명을 조사한 결과 평균이 22세, 표준편차가 2세였다. 신입생들의 평균 나이의 99% 신뢰구간을 구하시오.

풀이.

모분산 σ^2이 알려지지 않고 표본의 크기 n이 작은 경우의 구간추정의 방법을 알아보자. 이 경우는 모분산 σ^2 대신 표본분산 S^2을 사용하고 표본평균 $\overline{X}$를 새로운 변수 $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$로 변환한다. T는 자유도가 $(n-1)$인 t-분포를 따른다. 즉, $T \sim t(n-1)$이다.

t-분포곡선의 양쪽 꼬리 부분의 곡선 아래의 넓이가 $\dfrac{\alpha}{2}$가 되는 경계값을 각각 $-t_{\alpha/2}(n-1)$, $t_{\alpha/2}(n-1)$이라고 하면 t-분포곡선과 두 경계값 사이의 넓이는 $1-\alpha$가 된다. 여기서 넓이는 확률을 의미하므로

$$P\left(-t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

이고 $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$이므로 $-t_{\alpha/2}(n-1) < \dfrac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)$이다. 이것을 모평균 μ로 정리하고 구간으로 나타내면 $\left(\overline{X} - t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}},\ \overline{X} + t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}\right)$이다. 이것을 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간이라고 한다.

참고 모분산 σ^2이 알려지지 않은 경우(또는 $n<30$)의 모평균 μ의 신뢰구간

$$\left(\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

부록의 t-분포표에서 주어진 자유도 $(n-1)$과 유의수준 $\frac{\alpha}{2}$가 교차하는 위치의 값을 찾아서 계수 $t_{\alpha/2}(n-1)$에 대입한다.

예제 24. 어떤 연구실에서는 실험용 쥐의 평균 혈액점도를 알아보기 위해 16마리를 조사한 결과 평균이 4이고 표준편차가 0.8이었다. 실험용 쥐의 평균 혈액점도의 95% 신뢰구간을 구하시오.

풀이.

(2) 모비율의 구간추정

이것은 모비율 p의 추정량인 표본비율 $\hat{p}=\dfrac{X}{n}$의 분포를 이용하는데 이것은 중심극한정리에 의하여 평균이 p이고 분산이 $\dfrac{pq}{n}$인 정규분포로 근사하므로 표준화변수 $Z=\dfrac{\hat{p}-p}{\sqrt{pq/n}}$로 변환하면 $Z \sim N(0,1)$이다. 이때 변환된 확률은

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

이고 확률변수의 범위는 $-z_{\alpha/2} < \dfrac{\hat{p}-p}{\sqrt{pq/n}} < z_{\alpha/2}$이므로 모비율 p로 정리하면

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

이다. 구간으로 나타내면

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{pq}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{pq}{n}}\right)$$

이다. 그러나 여기서 p는 알려지지 않았으므로 표본의 크기 n이 증가함에 따른 일치추정량 $\hat{p}$을 p 대신 사용한다. 따라서 모비율 p의 $100(1-\alpha)$% 신뢰구간은

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

이다.

참고 모비율 p의 특별한 신뢰구간

(1) 90% 신뢰구간(즉, 유의수준 $\alpha=0.1$) : $\left(\hat{p}-1.645\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p}+1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

(2) 95% 신뢰구간(즉, 유의수준 $\alpha=0.05$) : $\left(\hat{p}-1.96\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p}+1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

(1) 99% 신뢰구간(즉, 유의수준 $\alpha=0.01$) : $\left(\hat{p}-2.58\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p}+2.58\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

예제 25. 어떤 정당에서는 당원 중에서 임의로 400명을 추출하여 새로운 정책의 추진여부를 조사하였다. 이 조사에서 240명은 찬성하고 나머지는 반대하였다. 전체 당원에 대한 찬성자의 90% 신뢰구간을 구하시오.

풀이.

4.3.3 표본의 크기 결정

- 표본조사에 의한 여론조사의 결과는 표본의 크기와 추정오차의 범위를 함께 발표한다.
- 표본의 크기가 크면 표본오차가 작아져서 비교적 정확한 결과를 얻을 수 있지만 표본의 크기가 커짐에 따른 자료수집의 시간이나 비용이 많아지기 때문에 표본추출의 의미가 없어진다.
- 표본의 크기가 작으면 시간과 비용은 적게 들지만 모집단과 오차가 커져서 정확한 정보를 제공하지 못한다.
- 이러한 이유에서 표본조사는 추정오차의 범위를 먼저 정해놓고 그것에 적당한 표본의 크기를 결정하여 조사하게 된다.

(1) 모평균의 구간추정에 따른 표본의 크기

모분산 σ^2이 알려지거나 표본의 크기 n이 클 때$(n \geq 30)$ 모평균의 구간추정에서 신뢰수준 $100(1-\alpha)\%$에 대한 신뢰구간은

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \text{ 또는 } \left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

이었다. 여기서 반대로 표본오차를 d라고 할 때, 제시된 표본오차 d에 맞는 표본의 크기 n은 표본평균 $\overline{X}$를 기준으로 하여 구간을 이루는 $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ 또는 $z_{\alpha/2}\dfrac{S}{\sqrt{n}}$에 달려있다.

표본오차가 d 이하인 표본의 크기는

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq d \text{ 또는 } z_{\alpha/2}\frac{S}{\sqrt{n}} \leq d$$

에 의해 결정된다. 따라서

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{d}\right)^2 \text{ 또는 } n \geq \left(z_{\alpha/2}\frac{S}{d}\right)^2.$$

참고

모분산 σ^2이 알려지거나 표본분산 S^2을 알 수 있는 경우(또는 $n \geq 30$)의 표본의 크기

(1) 신뢰수준 90%(즉, 유의수준 $\alpha = 0.1$) : $n \geq \left(\frac{1.645\sigma}{d}\right)^2$ 또는 $n \geq \left(\frac{1.64S}{d}\right)^2$

(2) 신뢰수준 95%(즉, 유의수준 $\alpha = 0.05$) : $n \geq \left(\frac{1.96\sigma}{d}\right)^2$ 또는 $n \geq \left(\frac{1.96S}{d}\right)^2$

(3) 신뢰수준 99%(즉, 유의수준 $\alpha = 0.01$) : $n \geq \left(\frac{2.58\sigma}{d}\right)^2$ 또는 $n \geq \left(\frac{2.58S}{d}\right)^2$

예제 26. 어떤 상표의 과자는 평균 중량이 600g이고 표준편차가 50g이 되게 생산관리를 하고 있다. 현재 생산하고 있는 과자의 평균 중량의 추정에서 신뢰수준 95%, 표본오차 10g이하가 되기 위한 표본의 크기를 구하시오.

풀이.

예제 27. 우리나라 전체 고등학교 3학년 여학생들의 몸무게를 알아보기 위해서 어떤 학교의 학생 36명의 몸무게를 조사한 결과 평균 몸무게가 57kg이고 표준편차가 4.8kg이었다. 학생들의 평균 몸무게의 추정에서 신뢰수준 90%, 표본오차 0.5kg 이하가 되기 위한 표본의 크기를 구하시오.

풀이.

모분산 σ^2이 알려지지 않고 표본의 크기 n이 작을 때($n<30$) 모평균의 구간추정에서 모분산 σ^2의 추정량인 표본분산 S^2을 이용한 t-분포로부터 신뢰수준 $100(1-\alpha)$%에 대한 신뢰구간은

$$\left(\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

이었다. 여기서 반대로 표본오차를 d라고 할 때, 제시된 표본오차 d에 맞는 표본의 크기 n은 표본평균 $\overline{X}$를 기준으로 하여 구간을 이루는 $t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$에 달려있다.

표본오차가 d 이하인 표본의 크기는

$$t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\le d$$

에 의해 결정된다. 따라서

$$n\ge\left(t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)^2.$$

참고

모분산 σ^2이 알려지지 않은 경우(또는 $n < 30$)의 표본의 크기

$$n \geq \left(t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right)^2$$

예제 28. 어떤 상표의 화장품은 무게가 40g으로 표시되어 있다. 이 화장품의 쿠게를 조사하기 위하여 10개의 표본을 추출하여 무게를 측정한 결과 평균이 38g이고 표준편차가 2.5g이었다. 현재 생산하고 있는 화장품의 평균 무게의 추정에서 신뢰수준 90%, 표본오차 1.2g 이하가 되기 위한 표본의 크기를 구하시오.

풀이.

(2) 모비율의 구간추정에 따른 표본의 크기

모비율의 구간추정에서 모비율의 추정량인 표본비율 $\hat{p}$을 이용하여 표준정규분포로부터 신뢰구간을 정의하였다. 이때, 신뢰수준 $100(1-\alpha)$%에 대한 신뢰구간은

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

이었다. 여기서 반대로 표본오차를 d라고 할 때, 제시된 표본오차 d에 맞는 표본의 크기는 표본비율 $\hat{p}$을 기준으로 하여 구간을 이루는 $z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$에 달려 있다.

표본오차가 d 이하인 표본의 크기는

$$z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \le d$$

에 의해 결정된다. 따라서

$$n \ge \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{d}\right)^2.$$

참고 모비율 p 또는 표본비율 $\hat{p}$이 알려진 경우 표본의 크기

(1) 신뢰수준 90%(즉, 유의수준 $\alpha = 0.1$) : $n \geq pq\left(\frac{1.645}{d}\right)^2$ 또는 $n \geq \hat{p}\hat{q}\left(\frac{1.645}{d}\right)^2$

(2) 신뢰수준 95%(즉, 유의수준 $\alpha = 0.05$) : $n \geq pq\left(\frac{1.96}{d}\right)^2$ 또는 $n \geq \hat{p}\hat{q}\left(\frac{1.96}{d}\right)^2$

(3) 신뢰수준 99%(즉, 유의수준 $\alpha = 0.01$) : $n \geq pq\left(\frac{2.58}{d}\right)^2$ 또는 $n \geq \hat{p}\hat{q}\left(\frac{2.58}{d}\right)^2$

예제 29. 어떤 공장에서 생산하는 제품의 불량품을 조사하고자 한다. 지난번 조사에서 제품의 불량품이 4%이었다. 현재 생산하고 있는 제품의 불량품의 추정에서 신뢰수준 95%, 표본오차 0.5% 이하가 되기 위한 표본의 크기를 구하시오.

풀이.

모비율의 구간추정에서 보이$_{\text{끌}}$에 대한 아무런 정보가 주어지지 않은 경우에는 표본비율을 $\hat{p}=\dfrac{1}{2}$로 하여 표본의 크기를 결정한다. 따라서 표본오차가 d 이하인 표본의 크기는

$$n \geq \frac{1}{2}\times\frac{1}{2}\left(\frac{z_{\alpha/2}}{d}\right)^2$$

이다.

참고 모비율 p와 표본비율 $\hat{p}$이 알려지지 않은 경우 표본의 크기

(1) 신뢰수준 90%(즉, 유의수준 $\alpha=0.1$) : $n \geq \left(\dfrac{1}{2}\right)^2\left(\dfrac{1.645}{d}\right)^2$

(2) 신뢰수준 95%(즉, 유의수준 $\alpha=0.05$) : $n \geq \left(\dfrac{1}{2}\right)^2\left(\dfrac{1.96}{d}\right)^2$

(3) 신뢰수준 99%(즉, 유의수준 $\alpha=0.01$) : $n \geq \left(\dfrac{1}{2}\right)^2\left(\dfrac{2.58}{d}\right)^2$

예제 30. 어떤 정당에서는 새로운 정책의 찬성에 대한 지지율을 조사하려고 한다. 새로운 정책에 대한 찬성률의 추정에서 신뢰수준 90%, 표본오차 4% 이하가 되기 위한 표본의 크기를 구하시오.

풀이.

4.3.4 두 모집단의 구간추정

두 개의 모집단을 비교하기 위해서 하나의 모집단의 구간추정을 확대할 수 있다.

예 : 고등학생과 대학생의 한 달 용돈의 평균의 차이(두 모평균의 구간추정 확대)
대도시와 농촌지역의 국회의원 선거의 투표율의 차이(두 모비율의 구간추정 확대)

(1) 두 모평균의 구간추정

두 모집단의 각각의 평균을 μ_1, μ_2, 분산을 σ_1^2, σ_2^2, 표본의 크기를 n_1, n_2, 표본평균을 $\overline{X_1}$, $\overline{X_2}$라 하면 각각 정규분포 $\overline{X_1} \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right)$, $\overline{X_2} \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right)$를 따르고

차 $\overline{X_1} - \overline{X_2}$는 $\overline{X_1} - \overline{X_2} \ \sim \ N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$를 따른다.

비슷하게 두 모분산 σ_1^2과 σ_2^2을 모르더라도 각 표본의 크기 n_1, n_2가 크면(n_1, $n_2 \geq 30$) 모분산의 추정량인 S_1^2과 S_2^2으로 구성된 정규분포에 근사한다. 즉,

차 $\overline{X_1} - \overline{X_2}$는 $\overline{X_1} - \overline{X_2} \ \sim \ N\left(\mu_1 - \mu_2, \dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)$를 따른다.

차 $\overline{X_1} - \overline{X_2}$를 표준화변수 Z로 변환하여 정리하면

두 모평균의 차 $\mu_1 - \mu_2$의 $100(1-\alpha)\%$ 신뢰구간은

$$\left((\overline{X_1} - \overline{X_2}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,,\quad (\overline{X_1} - \overline{X_2}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

또는

$$\left((\overline{X_1} - \overline{X_2}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\,,\quad (\overline{X_1} - \overline{X_2}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

이다.

참고 두 모분산 σ_1^2, σ_2^2이 알려지거나 두 표본분산 S_1^2, S_2^2을 알 수 있는 경우 (또는 $n_1,\ n_2 \geq 30$)의 특별한 신뢰구간

(1) 90% 신뢰구간(즉, 유의수준 $\alpha = 0.1$) :

$$\left((\overline{X_1} - \overline{X_2}) - 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\quad (\overline{X_1} - \overline{X_2}) + 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

또는

$$\left((\overline{X_1} - \overline{X_2}) - 1.645\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\quad (\overline{X_1} - \overline{X_2}) + 1.645\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

(2) 신뢰수준 95%(즉, 유의수준 $\alpha = 0.05$) : 1.645 대신 1.96

(3) 신뢰수준 99%(즉, 유의수준 $\alpha = 0.01$) : 1.645 대신 2.58

예제 31. 어떤 생활연구소에서는 대학교 1학년 학생과 고등학교 3학년 학생의 한 달 용돈의 평균의 차이를 알아보기 위하여 대학생 300명, 고등학생 200명을 조사하였다. 그 결과 대학생은 평균 35만 원, 표준편차 2만 원이고 고등학생은 평균 25만원, 표준편차 1.5만 원이었다. 두 집단의 용돈의 평균의 차에 대한 95% 신뢰구간을 구하시오.

풀이.

참고 두 모분산 σ_1^2, σ_2^2이 알려지지 않은 경우(또는 (또는 n_1, $n_2 < 30$)의 두 모평균의 차 $\mu_1 - \mu_2$의 신뢰구간은

$$\left((\overline{X_1} - \overline{X_2}) - t_{\alpha/2}(n_1 + n_2 - 2) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}(n_1 + n_2 - 2) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

단, S_p^2은 두 모집단의 표본의 크기 n_1과 n_2가 다를 수 있으므로 각각의 자유도 $(n_1 - 1)$과 $(n_2 - 1)$에 대한 가중평균으로 정의한 합동표본분산

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

이다.

앞의 경우와 비슷하게 위 구간을 구할 수 있다.

예제 32. 어떤 보건소에서는 금연클리닉을 찾은 40대 남성 12명과 50대 남성 10명의 흡연량을 조사하였다. 40대의 흡연량은 평균 25개비, 표준편차 8개비이고 50대의 흡연량은 평균 20개비, 표준편차 6개비이었다. 두 집단의 흡연량의 평균의 차에 대한 90% 신뢰구간을 구하시오.

풀이.

(2) 두 모비율의 구간추정

두 모비율에 대한 구간추정은 모비율의 구간추정에서와 같은 조건으로 이루어진다.
두 모집단의 각각의 모비율을 p_1, p_2, 표본의 크기를 n_1, n_2라 할 때,

표본비율의 차 $\hat{p}_1 - \hat{p}_2$를 표준화변수 $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}}$로 변환하여 정리하면

두 모비율의 차 $p_1 - p_2$의 $100(1-\alpha)\%$ 신뢰구간은

$$\left((\hat{p}_1 - \hat{q}_1) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} ,\ \hat{p}_1 - \hat{q}_1) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right)$$

이다.

예제 33. 어떤 정당에서 새로운 정책에 대한 A, B 두 지역의 찬성자 수를 조사한 결과가 다음과 같다. A, B 두 지역의 찬성률의 차에 대한 95% 신뢰구간을 구하시오.

인원수 \ 지역	A	B
찬성자 수	293	341
조사자 수	550	708

풀이.