통계학 개론

제5장 통계적 추정

5.1 통계적 추정의 이해

우리가 알고 싶은 관심대상은 모든 개체에 대한 관측값의 집합인 모집단

모집단을 안다는 것은 모집단의 분포를 아는 것으로 귀결되는데, 모집단의 분포는 몇 개의 모수에 의해 결정된다.

따라서 모집단을 안다는 것은 모수를 아는 것이다.

모집단을 추정하는 것은 모집단의 분포를 결정하는 모수를 추정하는 것이다.

따라서 모집단의 일부인 표본의 함수 통계량을 만들고 이를 가지고 모수를 추정하게 된다.

통계량으로는 데이터를 요약할 때 이용되는 표본평균, 표본비율, 표본분산 등이 있는데 이들은 각각 모평균, 모비율, 모분산을 추정하는데 이용된다.

통계량은 모수를 추정하는데 이용되므로 추정량(estimator)이라고 부르고, 추정량에 관측값을 대입하여 얻은 추정량의 값을 추정값(estimate)이라 부른다.

추정량과 추정값의 예

모수	추정량	추정값의 예	표본분포
모평균	$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	$\overline{x}=170.1cm$	정규분포
모비율	$\hat{p}=\frac{a}{n}$	$\hat{p}=0.75$	정규분포
모분산	$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$	$s^2=2.54$	카이제곱분포

통계적 추론: 표본으로부터 모집단을 추측하는 것. 추정과 검정으로 구분된다.

❖ 추정(estimation)

표본으로 모집단에 대한 결론을 도출한 것

점추정(point estimation)과 구간(interval estimation)추정으로 나누어진다.

❖ 검정(testing)

모집단에 대한 주장의 타당성을 표본을 통해 점검하는 것

5.2 바람직한 추정량

일반적으로 추정량은 불편성, 일치성, 효율성 등의 특성을 가져야 바람직하다.
불편성: 모든 가능한 통계량값의 평균이 모수와 같아지는 것
일치성: 표본크기가 커질수록 추정량의 값과 모수가 점점 더 가까워지는 것
효율성: 추정량 중 분산이 작은 것

5.3 모평균의 추정

표본평균: 모집단에서 표본 $X_1, X_2, \cdots, X_n$ 을 임의로 추출한 값들의 평균

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

첫째, 표본평균의 기댓값은 모평균이 된다.

$$E(\overline{X}) = \mu$$

둘째, 표본평균의 분포는 표본수가 커질수록 밀집하게 된다.

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

셋째, 중심극한정리에 따라 모집단이 어떠한 분포든지 표본크기가 충분히 크다면 모든 가능한 표본평균은 모평균 주위에 정규분포 모양을 하면서 밀집하게 된다.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

표본평균을 표준화한 통계량 $\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$ 이 ± 1.96 사이에 있을 확률은 95%이다.

$$P(-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$$

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

모든 가능한 표본평균에 대해 구간공식을 적용했을 때 얻어지는 모든 가능한 구간 중 95%의 구간이 모평균 μ를 포함한다.

$$\left[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

그런데 모표준편차 σ를 모르는 경우 구간공식을 쓸 수 없다. 이 경우 표본표준편차 S를 표준편차 σ의 추정량으로 사용하여 모평균 μ의 신뢰구간을 구해야 한다.

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

정규분포 $N(\mu, \sigma^2)$을 따르는 모집단으로부터 얻어진 확률표본을 $X_1, X_2, \cdots, X_n$ 이라 할 때 $T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$는 자유도 n-1인 t분포, 즉 t_{n-1}을 따른다.

100(1-α)% 신뢰구간

$$P(-t_{n-1,\alpha/2} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}) = 1-\alpha$$
$$P(\overline{X}-t_{n-1,\alpha/2} S/\sqrt{n} \le \mu \le \overline{X}+t_{n-1,\alpha/2} S/\sqrt{n}) = 1-\alpha$$

표본수가 적절히 클 때(n≥30) t분포는 정규분포에 근접하므로 $t_{n-1,\alpha/2}$ 대신 우리가 알고 있는 정규분포의 $Z_{\alpha/2}$를 이용할 수 있다.

❖ 모평균의 100(1-α)% 구간추정
모집단이 정규분포이고 모분산 σ^2를 모르는 경우

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

5.4. 모비율의 추정

모집단의 비율(p)을 추정하는 문제는 모평균의 추정을 구하는 문제와 매우 흡사

모집단이 두 개의 배반사건(찬성, 반대)으로 구성되어 있다고 하고, 찬성 모비율이 p이고 반대 모비율이 $1-p$라고 하자. 이때 모집단에서 n개의 표본을 뽑았을 때 찬성자수 X는 이항분포 $B(n,p)$를 따른다.

표본비율($\hat{P}$)은 모집단이 1과 0으로 이루어질 때 1의 비율을 의미하므로 일종의 표본평균이라고 할 수 있다.

$$\hat{p} = \frac{X}{n}$$

표본비율 $\hat{p}$는 표본평균과 마찬가지로 불편추정량이며 표본수가 커지면서 밀집된 정규분포로 근사된다.

$$E(\hat{p}) = E(\frac{X}{n}) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var(\frac{X}{n}) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

한편 분산추정량은 다음과 같다.

$$\widehat{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$$

$\hat{p} \sim N(p, \frac{p(1-p)}{n})$

표본비율을 표준화한 통계량 $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$이 $\pm z_{\alpha/2}$사이에 있을 확률은 100(1-α)%

$$P(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_{\alpha/2}) = 100(1-\alpha)\%$$

$$P(-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \hat{p}-p \le z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 100(1-\alpha)\%$$

$$P(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 100(1-\alpha)\%$$

❖ 모비율 p의 100(1-α)% 구간추정

표본의 크기가 충분히 큰 경우

$$\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

여기서 n은 표본크기이고, $\hat{p}$는 표본비율이다.

5.5 모분산의 추정

모분산은 표본분산(S^2)을 이용하여 추정한다.

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

표본분산의 중요한 특성은 불편성이다. 즉, 표본분산(S^2)은 모분산(σ^2)의 불편추정량이다. 따라서 모분산을 점추정하는데 표본분산이 이용된다.

모분산의 신뢰구간을 추정하려면 표본분산의 분포를 이용해야 한다. 정규분포를 따르는 모집단에서 표본을 추출한 후 구한 표본분산은 자유도가 n-1인 카이제곱(χ^2)분포를 따른다. 즉,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

표본분산의 분포로부터 다음을 도출할 수 있다.

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

$$P\left\{\frac{n-1}{\chi^2_{n-1,\alpha/2}}S^2 \le \sigma^2 \le \frac{n-1}{\chi^2_{n-1,1-\alpha/2}}S^2\right\} = 1-\alpha$$

❖ 모분산(σ^2)의 100(1-α)% 신뢰구간

모집단이 정규분포를 따르는 경우

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

여기서 S^2은 표본분산, $x^2_{k,\alpha}$은 자유도 k인 x^2분포의 (1-α) 백분위수를 뜻한다.