제 5장. 통계적 추론

5.1 점추정

점추정(point estimation) : 하나는 모수를 한 개의 값으로 추정

주요 모수와 추정량

모 수	추정량	표준오차
모평균 : μ	$\overline{X}$	$\sigma/\sqrt{n}$
모비율 : p	$\hat{p}$	$\sqrt{p(1-p)/n}$
모표준편차 : σ (모분산 : σ^2)	S (S^2)	.

정의. **불편 추정량**(unbiased estimator)

$E(\hat{\theta}) = \theta$ 일 때 $\hat{\theta}$을 θ의 **불편추정량** 또는 **비편향추정량**이라 한다.

예. $\overline{X}$는 μ의 불편추정량이며, $\hat{p}$은 p의 불편추정량, S^2은 σ^2의 불편추정량

예) 다음의 표는 어떤 과즙의 당분 함량을 화학분석에 의해 얻은 것이다. 이로부터 당분의 평균함량, 표준편차를 추정해 보자.

14.0	14.2	15.1	13.7	14.5	15.6	14.8	15.1	13.5	15.8

5.2 구간추정

구간추정(interval estimation) : 모수가 포함되리라 기대되는 구간으로 모수를 추정

통계량 L과 U에 대하여 $P(L < \theta < U) = 1 - \alpha$ 일 때,
구간$(L,\ U)$ 또는 (l, u)를 θ의 $100(1-\alpha)\%$ **신뢰구간**(confidence interval)
l과 u : 각각 **신뢰구간의 하한**과 **상한**
$1-\alpha$: **신뢰수준**(confidence level)

정규분포에서 $\overline{X}$의 분포로부터 다음이 성립한다.

$$1-\alpha = P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right)$$

$$= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

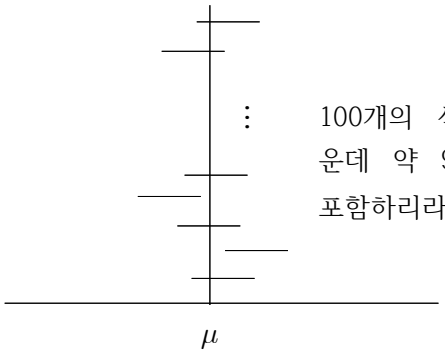
따라서 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

예. $\alpha = 0.05\ (5\%)$일 때, μ의 95% 신뢰구간은 다음과 같다.

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = (\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

95% 신뢰구간의 의미



5.2.1. 신뢰수준의 이해

예) 표준정규분포에서 50개의 난수를 발생시켜 95% 신뢰구간을 구하는 과정을 1000번 반복하자. 1000개의 신뢰구간 중에서 실제로 모수를 포함하는 신뢰구간의 비율을 구하여라.

```
alpha<-0.05
n<-50
mu<-0
sigma<-1
count<-0

for (i in 1:1000){
   x<-rnorm(n, mu, sigma)
   upper<-mean(x)-qnorm(alpha/2)*(sigma/sqrt(n))
   lower<-mean(x)+qnorm(alpha/2)*(sigma/sqrt(n))
   if ( (lower< mu) & (mu< upper) ) count=count+1
}

count/1000
```

▶ qnorm(alpha, mu, sigma) : (mu, sigma)를 모수로 갖는 정규분포의 100(1-alpha) 분위수. 모수를 생략하면 표준 정규분포를 기준으로 한다.

5.3 가설 검정

예) 어느 전구의 평균수명이 평균 $\mu = 1500$ (시간)이고 표준편차 $\sigma = 100$ (시간)인 정규분포를 따른다고 하자. 이 때, 새 공법에 의하면 전구의 평균수명이 증가한다고 할 때, $n = 25$ 개의 전구를 시험 생산한 결과 $\overline{X} = 1550$ (시간)으로 나타났다. 이 결과를 통해 새 공법에 의해 전구의 평균수명이 증가했다고 확신할 수 있는가? 유의수준 5%에서 이를 확인하시오.

[풀이] 주어진 문제를 이용하여 가설을 세우면 다음과 같다.

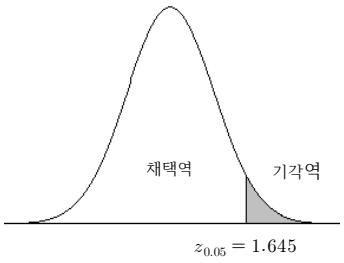
H_0 : $\mu = 1500$ (**귀무가설**, null hypothesis)

H_1 : $\mu > 1500$ (**대립가설**, alternative hypothesis)

검정통계량은 다음과 같다.

$$Z = \frac{\overline{X} - 1500}{20} = \frac{1550 - 1500}{20} = 2.5$$

기각역(critical region)은 $Z \geq 1.645$ 이고 검정통계량은 기각역에 속하므로 유의수준 5%에서 귀무가설을 기각할 수 있다. 즉, 전구의 평균 수명은 증가했다고 말할 수 있다.



가설 검정의 용어

대립가설 : 표본으로부터 입증하고자 하는 가설
귀무가설 : 대립가설에 대한 확실한 근거가 없을 때 받아들이는 가설
검정통계량 : 검정에 사용하는 통계량
유의수준 : 귀무가설이 참일 때 대립가설을 채택하는 오류를 범할 확률
기각역 : 귀무가설을 기각시키는 검정통계량의 관측값의 영역

오류의 종류

		실제 현상	
		H_0 참	H_1 참
검정결과	H_0 채택	옳은 결정	**제2종의 오류**
	H_1 채택	**제1종의 오류**	옳은 결정

유의수준(significance level)
: 귀무가설 H_0가 참일 때 대립가설 H_1을 채택하는 오류를 범할 최대 허용 확률
(즉, 제1종의 오류를 범할 확률)

유의확률(significance probability) 또는 p**-값(**p**-value)**
: 관측값으로부터 H_0를 기각시킬 수 있는 최소의 유의수준
: 따라서, p-값이 작을수록 대립가설 H_1이 참이라는 증거가 강함을 뜻한다.

검정력(power)
: 대립가설의 특정 값에서 귀무가설 H_0를 기각시킬 확률
: 따라서 검정력이 높을수록 좋은 검정법이 된다.

σ를 알 때 μ에 관한 검정법 : (σ를 알 때 : ***Z*-검정**)

검정통계량 : $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

검정법 : ① $H_1 : \mu > \mu_0$일 때 $Z \geqq z_\alpha$
② $H_1 : \mu < \mu_0$일 때 $Z \leqq -z_\alpha$
③ $H_1 : \mu \neq \mu_0$일 때 $|Z| \geqq z_{\alpha/2}$

예) 어느 사탕 캔 제조 공정에서는 생산되는 내용물의 함량을 표준편차 10g이 되도록 생산관리를 하고 있다. 이 공정에서 랜덤하게 15개의 캔을 뽑아서 조사한 결과 내용물의 평균 무게가 294.4g으로 나타났다. 이 캔에 적혀있는 내용물의 함량이 300g이라고 할 때, 이 조사결과에 의해 실제 함량은 300g 미만이라고 할 수 있는가? 유의수준 5%에서 이를 검정하고 유의확률을 구하시오.

[풀이] 귀무가설과 대립가설 : $H_0 : \mu = 300,\ H_1 : \mu < 300$

검정통계량의 값 : $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{294.4 - 300}{10/\sqrt{16}} = -2.24$

유의확률 : $P(Z \leq -2.24) = 0.0125$

유의확률이 유의수준보다 작으므로 유의수준 5%에서 귀무가설 H_0를 기각한다. 따라서 캔의 평균 함량이 300g 미만이라고 결론내릴 수 있다.

5.4 자료를 이용한 예제 (ames.csv)

주어진 자료는 Iowa의 도시 Ames의 2006년부터 2010년 사이의 부동산 거래내역 자료이다. 5년 동안 이 지역에서 발생한 총 2930건의 부동산 거래내역이 모두 기록되어 있다. 본 예제에서는 집의 크기를 나타내는 변수인 Gr.Liv.Area를 모집단으로 사용하도록 한다.

예제 1. 주어진 자료는 전체 부동산에 대한 자료이므로 모집단으로 생각할 수 있다. 거래가 이루어진 전체 부동산의 집의 크기의 평균값(μ)은 얼마인가? 모분산(σ^2)은 얼마인가?
예제 2. 모집단에서 크기가 60인 랜덤 표본을 선택하자. 모집단 평균에 대한 점추정값은 얼마인가?
예제 3. 예제 2에서 선택된 표본을 이용하여 모평균에 대한 95% 신뢰구간을 구해보자. 이 때, 모분산은 예제 1에서 구한 값을 사용하도록 한다. 이 신뢰구간은 모평균을 포함하는가?
예제 4. 예제 3과 동일한 과정을 50번 반복하여 서로 다른 신뢰구간 50개를 구해보자. 이 때, 신뢰구간의 하한값을 lower 벡터에 각각 저장하고 신뢰구간의 상한값은 upper 벡터에 각각 저장하도록 한다. 예제 1에서 구한 모평균의 값은 pop.mean에 저장한다. 그리고 아래의 코드를 실행해보자. 출력된 그래프가 나타내는 실제 신뢰수준은 어떠한가?

```
plot_ci <- function(lo, hi, m) {
  par(mar=c(2, 1, 1, 1), mgp=c(2.7, 0.7, 0))
  k <- length(lo)
  ci.max <- max(rowSums(matrix(c(-1 * lo, hi), ncol=2)))
  xR <- m + ci.max * c(-1, 1)
  yR <- c(0, 41 * k / 40)
  plot(xR, yR, type='n', xlab='', ylab='', axes=FALSE)
  abline(v=m, lty=2, col='#00000088')
  axis(1, at=m, paste("mu = ", round(m, 4)), cex.axis=1.15)

  for(i in 1:k) {
    x  <- mean(c(hi[i], lo[i]))
    ci <- c(lo[i], hi[i])
    if (lo[i]>m | m>hi[i]) {
      col <- "#F05133"
      points(x, i, cex=1.4, col=col)
      lines(ci, rep(i, 2), col=col, lwd=5)
    }
    col <- 1
    points(x, i, pch=20, cex=1.2, col=col)
    lines(ci, rep(i, 2), col=col)
  }
}

plot_ci(lower, upper, pop.mean)          # 그래프 실행 코드
```