제 5 장. 통계적 추론

5.1 점추정

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

주요 모수와 추정량

모수	추정량	표준오차
모평균: μ	$\overline{X}$	$\frac{\sigma}{\sqrt{n}}$
모비율: p	$\widehat{p}$	$\sqrt{\frac{p(1-p)}{n}}$
모표준편차: σ	S	.
모분산: σ^2	S^2	.

***정의.* 불편 추정량(unbiased estimator)**

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

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

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

당분
14.0 14.2 15.1 13.7 14.5 15.6 14.8 15.1 13.5 15.8

평균, 표준편차를 구하기 위해 numpy의 함수 mean, std를 사용한다.

In [3]:

```
import numpy as np
x = np.array([14.0, 14.2, 15.1, 13.7, 14.5, 15.6, 14.8, 15.1, 13.5, 15.8])
print(x.mean())      # 평균
print(x.std())       # 표준편차
```

```
14.63
0.7430343195303971
```

두 불편 추정량의 비교

표본 $X_1, \ldots, X_n$을 이용해 모평균을 추정하는 불편 추정량:

- $\hat{\mu}_1 = \bar{X} = \sum_{i=1}^{n} X_i/n$
- $\hat{\mu}_2 = X_1$

1. 불편성을 확인

i. 표준 정규 분포 N(0,1)에서 표본을 100개 추출

ii. $\hat{\mu}_1, \hat{\mu}_2$ 계산

iii. 위 과정을 1,000번 반복한 후 각 값의 평균, 분산을 계산

In [2]:

```
from scipy.stats import norm

np.random.seed(1)

mu1,mu2 = [], []
for i in range(1000):
    sample = norm.rvs(loc=0, scale=1, size=100)
    mu1.append(np.mean(sample))
    mu2.append(sample[0])

print('mean of mu1: %.3f' %np.mean(mu1))
print('mean of mu2: %.3f' %np.mean(mu2))

print('std of mu1: %.3f' %np.std(mu1))
print('std of mu1: %.3f' %np.std(mu2))
```

```
mean of mu1: 0.005
mean of mu2: 0.044
std of mu1: 0.100
std of mu1: 1.012
```

5.2 구간추정

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

통계량 L과 U에 대하여 $\Pr(L < \theta < U) = 1 - \alpha$ 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.

$$\begin{aligned} 1 - \alpha &= \Pr\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right) \\ &= \Pr\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right) \end{aligned}$$

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

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

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

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

95% 신뢰구간의 의미

: 100개의 신뢰구간 가운데 약 95개는 μ 를 포함하리라 기대

μ에 대한 확률이 아닌, 표본에 대한 확률임을 유의하세요

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.

In [3]:

```
from scipy.stats import norm
import numpy as np

np.random.seed(1)    # 시드값을 설정. 예제를 반복 가능하게 만든다.

alpha = 0.05
mu = 0
sigma = 1

count = 0
n = norm(loc = mu, scale = sigma)

for i in range(1000):
    x = n.rvs(size=50)
    upper = x.mean() + n.ppf(1-alpha/2) * (sigma/np.sqrt(50))   # ppf(q, loc, scale): 정규분포의 10
    lower = x.mean() - n.ppf(1-alpha/2) * (sigma/np.sqrt(50))

    if (lower < mu) & (mu < upper): count = count + 1

count/1000
```

Out[3]:

0.952

위 과정에서 alpha값을 조절하면 실제로 모수를 포함하는 신뢰구간의 비율이 변화한다.

In [4]:

```
from scipy.stats import norm
import numpy as np

np.random.seed(1)    # 시드값을 설정. 예제를 반복 가능하게 만든다.

alpha = 0.01
mu = 0
sigma = 1

count = 0
n = norm(loc = mu, scale = sigma)

for i in range(1000):
    x = n.rvs(size=50)
    upper = x.mean() + n.ppf(1-alpha/2) * (sigma/np.sqrt(50))   # ppf(q, loc, scale): 정규분포의 10
    lower = x.mean() - n.ppf(1-alpha/2) * (sigma/np.sqrt(50))

    if (lower < mu) & (mu < upper): count = count + 1

count/1000
```

Out[4]:

0.991

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 $\bar{X} = 1550$(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_0 가 참일 때 대립가설 H_1 을 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_1 이 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 (σ 를 알 때 : Z-검정)

검정통계량 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

유의수준 α에서 기각역

$$H_1 : \mu > \mu_0 \Rightarrow Z \geq z_\alpha$$
$$H_1 : \mu < \mu_0 \Rightarrow Z \leq -z_\alpha$$
$$H_1 : \mu \neq \mu_0 \Rightarrow |Z| \geq z_{\alpha/2}$$

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: $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{294.4-30}{10/\sqrt{16}} = -2.24$

유의확률: $\Pr(Z \leq -2.24) = 0.0125$

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