통계분석

1. 모형 적절성 확인하기

a) 모형이 통계적으로 유의한가?
F통계량을 확인, 유의수준 5% 이하에서 F통계량의 p-값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의

b) 회귀계수들이 유의미한가?
해당 계수의 t통계량과 p-값 또는 이들의 신뢰구간을 확인

c) 모형이 얼마나 설명력을 갖는가?
결정계수를 확인, 결정계수는 0~1 사이의 값을 가지며, 높은 값을 가질수록 설명력이 높다.

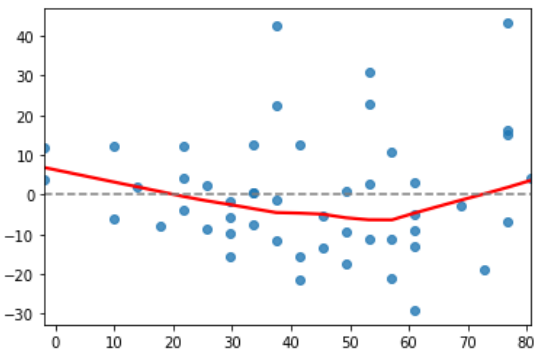
d) 모형이 데이터를 잘 적합하고 있는가?
잔차를 그래프로 그리고 회귀진단을 한다.

e)데이터가 아래의 모형 가정을 만족시키는가?
선형성 : 독립변수의 변화에 따라 종속변수도 일정크기로 변화
독립성 : 잔차와 독립변수의 값이 관련돼 있지 않음
등분산성 : 독립변수의 모든 값에 대해 오차들의 분산이 일정
비상관성 : 관측치들의 잔차들끼리 상관이 없어야 함.
정상성 : 잔차항이 정규분포를 이뤄야 함.

모형의 선형성

- 예측값(fitted)과 잔차(residual)를 비교
- 모든 예측값에서 가운데 점선에 맞추어 잔차가 비슷하게 있어야 한다.
- 빨간 실선은 잔차의 추세를 나타낸다.
- 빨간 실선이 점선에서 크게 벗어난다면 예측값에 따라 잔차가 크게 달라진다는 것으로 선형성이 없다는 것이다.

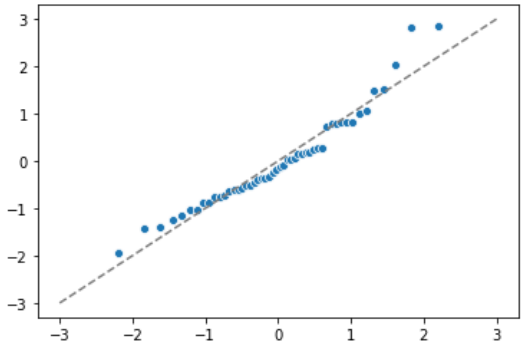
```
import matplotlib.pyplot as plt
import seaborn as sns

fitted = res.predict(df)
residual = df['dist'] - fitted

sns.regplot(fitted, residual, lowess=True, line_kws={'color': 'red'})
plt.plot([fitted.min(), fitted.max()], [0, 0], '--', color='grey')
```



잔차의 정규성

- 잔차가 정규분포를 따른다는 가정을 한다.
- Q-Q Plot로 확인할 수 있다.
- 잔차가 정규분포를 띄면 Q-Q Plot에서 점들이 점선을 따라 배치되어 있어야 한다.

```
import scipy.stats
sr = scipy.stats.zscore(residual)
(x, y), _ = scipy.stats.probplot(sr)
sns.scatterplot(x, y)
plt.plot([-3, 3], [-3, 3], '--', color='grey')
```

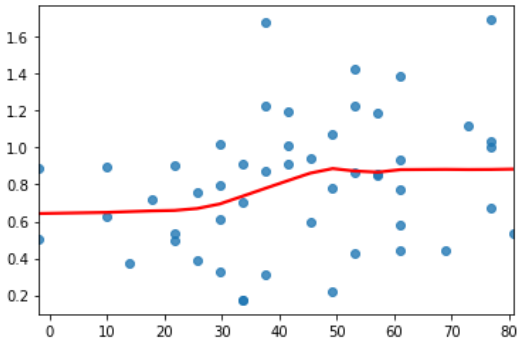


- 잔차의 정규성은 샤피로 검정으로 확인할 수 있다.
- `scipy.stats.shapiro(residual) # ShapiroResult(statistic=0.9450905919075012, pvalue=0.02152460627257824)`
- 위 분석에서 두 번째 값이 p값이다. p값이 0.02이므로 유의수준 5%에서 잔차의 정규성이 위반되었다고 판단한다.

잔차의 등분산성

- 회귀모형을 통해 예측된 값이 어떻든지, 모든 값들에 대하여 잔차의 분산이 동일하다는 가정
- 아래 그래프는 예측값(X축)에 따라 잔차가 어떻게 달라지는지 보여줌
- 빨간색 실선이 수평선에 가까울수록 등분산성이 있다는 것이다.

```
import numpy as np
sns.regplot(fitted, np.sqrt(np.abs(sr)), lowess=True, line_kws={'color': 'red'})
```



잔차의 독립성

- Result.summary의 Durbin-Watson(더빈왓슨, DW검정)으로 확인한다.
- DW검정은 잔차의 독립성을 확인할 수 있는 수치이다. 0이면 잔차들이 양의 자기상관을 갖고, 2이면 자기상관이 없는 독립성을 갖고, 4이면 잔차들이 음의 자기상관을 갖는다고 해석한다.
- 보통 1.5 ~ 2.5사이이면 독립으로 판단하고 회귀모형이 적합하다는 것을 의미한다. DW검정값이 0 또는 4에 가깝다는 것은 잔차들이 자기상관을 가지고 있다는 의미이고, 이는 t값, F값, R제곱을 실제보다 증가시켜 실제로 유의미하지 않은 결과를 유의미한 결과로 왜곡하게 된다.
- 위 회귀분석에서 더빈왓슨검정의 값이 1.676이므로 독립성이 있다고 판단할 수 있다.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dist   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     89.57
Date:                Sat, 19 Dec 2020   Prob (F-statistic):           1.49e-12
Time:                        18:40:59   Log-Likelihood:                -206.58
No. Observations:                  50   AIC:                             417.2
Df Residuals:                      48   BIC:                             421.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -17.5791      6.758     -2.601      0.012     -31.168      -3.990
speed          3.9324      0.416      9.464      0.000       3.097       4.768
==============================================================================
Omnibus:                        8.975   Durbin-Watson:                   1.676
Prob(Omnibus):                  0.011   Jarque-Bera (JB):                8.189
Skew:                           0.885   Prob(JB):                       0.0167
Kurtosis:                       3.893   Cond. No.                         50.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

극단값

- Cook's distance는 극단값을 나타내는 지표이다.
- 48번, 22번, 38번 자료가 특히 예측에서 많이 벗어남을 알 수 있다.

```
from statsmodels.stats.outliers_influence import OLSInfluence
cd, _ = OLSInfluence(res).cooks_distance
cd.sort_values(ascending=False).head()
```

```
48    0.340396
22    0.085552
38    0.068053
44    0.053176
34    0.052576
dtype: float64
```

2. 회귀분석 : 양적변수를 예측하는데 사용.

단순 선형회귀 : 설명변수가 1개이며 반응변수와의 관계가 직선
다중 선형회귀 : 설명변수가 k개이며 반응변수와의 관계가 선형
다항회귀 : 설명변수가 k개이며, 반응변수와의 관계가 1차함수 이상(단 k=1이면 2차함수 이상)
비선형회귀 : 회귀식의 모양이 미지의 모수 Bi들의 선형관계로 이루어져 있지 않은 모형, 예를 들면 지수함수

1) 회귀분석의 4단계

a. 두 변수 간의 선형적 관계를 뒷받침하는 이론의 가정, CLRM이론의 7가지 가정

CLRM(Classical Linear Regression Method)의 7가지 가정

★★1~4를 만족하면 최소자승법 회귀분석은 일관되고, 편향되지 않은 결과를 도출
★★5~6을 만족하면 최소자승법 회귀분석이 최선의 방법
★★4~7은 오차항에 관한 내용으로 해당 검정들을 실시(오차를 모르기 때문에 잔차로 검정)하여 검정 값의 p-value가 충분히 낮으면 회귀분석 시행
(등분산성 검정의 경우 BP test나 White test의 경우 p-value가 높아야 가정이 맞다고 전제함. 잔차로 오차를 예측하는 건 동일)

** 편향되지 않음의 의미
E(θ) = θ
임의의 추정치 모수값
정규분포로 근사하면서 각 추정치가 실제값과 근사해진다.
이 때, 표본평균(추정치)는 모평균, 표본비율은 모비율, 표본표준편차는 모표준편차를 짐작할 수 있다.
하지만 편향되면 위의 식이 성립되지 않는다.

① 선형관계가 있어야 함.
② 무작위 표본추출
③ X값이 2개 이상(직선을 그으려면 2개 이상의 값이 필요)
④ zero-condition Mean : 주어진 오차도를 그리면 그 평균은 0이 된다. 만족 못할 때는 다중회귀분석 진행, 이럴 경우 추정치들이 편향됨.
⑤ 등분산성 : 모든 Xi에 있어 오차들이 같은 정도로 퍼져있다.
⑥ 독립성 : 오차항들끼리는 독립, 어떤 패턴을 가지면 안됨.
⑦ 정규성 : 각 Xi에서 오차들끼리는 정규분포를 이룬다.

b. 두 변수간의 선형관계를 잘 나타내는 직선 찾기, 최소자승법, R-square, SST,SSE,SSR

좌표평면상의 점들과 직선의 거리를 최소하 하는 직선을 찾는 법. 즉, 잔차 제곱합을 최소로 하는 직선
주어진 시그마 식을 b1,b0으로 편미분 했을 때 이를 0으로 만들어주는 b1,b0 값 찾기

★★★ 최소자승법 결과. 매우 중요
회귀직선의 기울기는 공분산/x의 분산, 절편은 회귀식에 평균값을 대입한 것.

$$b_1 = \frac{s_{xy}}{s_x^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$\hat{y}_i = b_0 + b_1 x_i$$

where b_0 and b_1 are chosen to minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Linear Least Square Method

$$y = m_1 x + m_2$$

$$y_n - m_1 x_n - m_2 = R_n$$

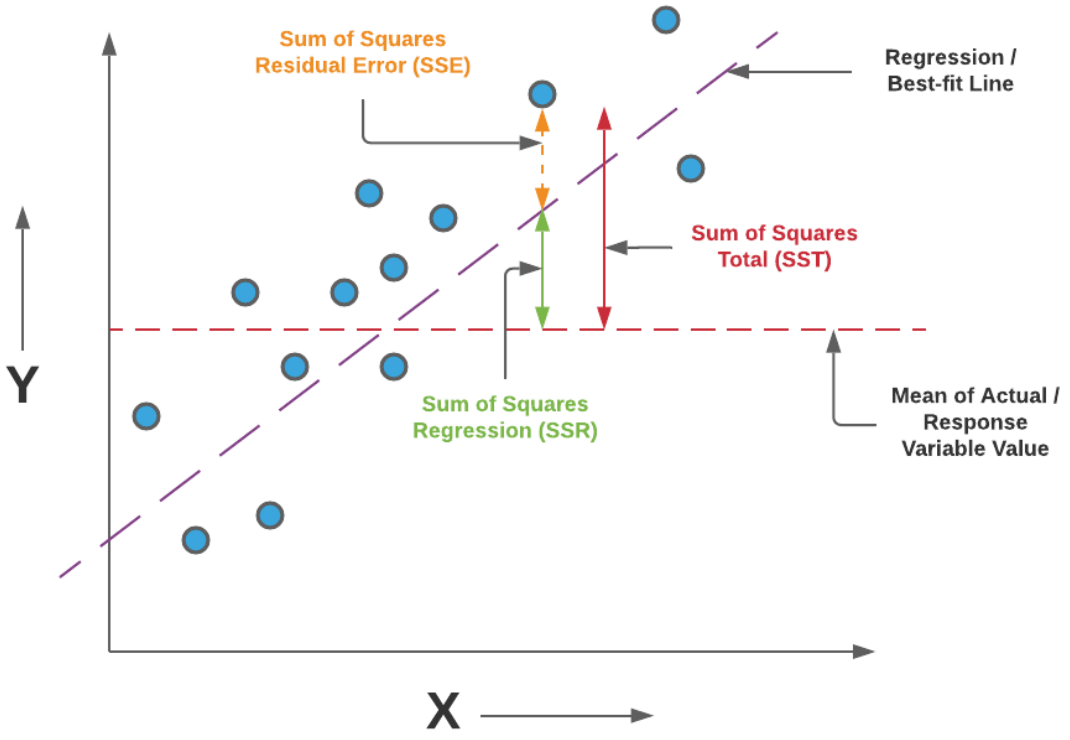
$$y_n^2 - m_1^2 x_n^2 - 2m_1 x_n y_n - 2m_2 y_n + 2m_1 m_2 x_n = R_n^2$$

$$\sum_{n=1}^{N} \left(y_n^2 - m_1^2 x_n^2 - 2m_1 x_n y_n - 2m_2 y_n + 2m_1 m_2 x_n \right) = \sum_{n=1}^{N} R_n^2$$

$$\frac{\partial \sum_{n=1}^{N} R_n^2}{\partial m_1} = 2m_1 \sum_{n=1}^{N} x_n^2 - 2\sum_{n=1}^{N} x_n y_n + 2m_2 \sum_{n=1}^{N} x_n = 0$$

$$\frac{\partial \sum_{n=1}^{N} R_n^2}{\partial m_2} = 2m_2 N - 2\sum_{n=1}^{N} y_n + 2\sum_{n=1}^{N} x_n = 0$$

★★ SSR은 회귀직선으로 설명 가능하지만 SSE는 설명 불가(잔차)



잔차가 최소가 되는 직선을 선택해야 하는데(아래 수식), 그냥 더하면 0이 되므로(잔차의 합은 0) 제곱을 한 뒤 더해준다.

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$
$$= \left(\hat{Y}_i - \bar{Y}\right) + e_i$$

양 변을 제곱해주면 아래의 식이 도출된다. 이 때 교차항 Σ2(Y^-Y-)*e 항이 사라지는 이유는 최소자승법의 전개식과 관련이 있다.

1계도조건을 만족하려면 두 식이 0이 된다.이 식이 교차항에 그대로 등장해 교차항이 0이 된다.

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum e_i^2$$

$$\begin{pmatrix}\text{Total sum}\\ \text{of squares}\end{pmatrix} = \begin{pmatrix}\text{regression}\\ \text{sum of squares}\end{pmatrix} + \begin{pmatrix}\text{error}\\ \text{sum of squares}\end{pmatrix}$$

$$SST = SSR + SSE$$

★★★★결론적으로 R-square는 회귀직선이 설명할 수 있는 비율을 나타낸 것으로 높으면 회귀직선의 설명력이 높다고 할 수 있다.

반대로 R-square가 작으면 잔차가 설명하는 비율이 많은 것으로 회귀직선의 설명력이 떨어진다고 볼 수 있다.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

잔차의 분산 S제곱은 MSE

분산은 편차 제곱의 평균값이다. 잔차의 평균은 0이다. 그러므로 각 잔차의 값을 제곱해서 평균을 내면 그게 잔차의 분산이 된다.

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

c. 모수값에 대한 추정과 가설검정

b0,b1,e를 통해 B0, B1, E를 추정

d. 예측(평균치, 개별값)

Q1. 오차의 평균이 0인 이유

이번에는 수학자 가우스(Gauss)가 정규분포를 유도한 방법을 알아보도록 하자. 가우스는 이항분포에서 정규분포를 유도하는 방법과는 별개로 오차에 대한 고찰을 통해 정규분포를 유도하였는데, 여기서는 앞단원의 나의 실제 키 예제와의 비교를 통해 설명하도록 하겠다. 나의 실제

키 예제의 핵심을 간단히 말하면 **정규분포를 인정한다면, 측정값의 평균을 실제값이라 여기는 우리의 직관은 옳다**는 것이며 좀 더 정확히 표현하면 다음과 같다.

1. 키의 측정값 x이 실제 키의 값인 μ를 평균으로 하는 정규분포를 따른다면 즉, **오차(error) $\epsilon = x - \mu$가 평균 0인 정규분포를 따른다면**
2. 실제 키 μ MLE, 즉 **실제 키일 가능성이 가장 높은 값은 측정값의 평균**이다.

가우스의 논리는 이것을 뒤집으면 된다. 즉, **측정값의 평균을 실제값이라 여기는 우리의 직관이 옳다면, 오차는 정규분포를 따른다**는 것이며 좀 더 풀어서 쓰면 다음과 같다.

1. 실제 키의 MLE, 즉 실제 키일 가능성이 가장 높은 값은 측정값의 평균이라면
2. 오차는 정규분포를 따른다.

가우스는 여기에 오차라면 마땅히 가져야 할 조건 3개를 추가하여 다음과 같은 **오차의 법칙**을 제시하였다.

1. +오차와 -오차가 나올 가능성은 같다. 즉, 오차의 분포를 나타내는 확률밀도 함수 f는 $f(-\epsilon) = f(\epsilon)$인 좌우대칭 함수이다.
2. 작은 오차가 나올 가능성이 큰 오차가 나올 가능성보다 크다. 즉, $f(\epsilon)$는 위로 볼록한 모양이다.
3. $f(\epsilon)$는 2번 미분가능하고, 전체 확률은 1이다. 즉, $\int_{-\infty}^{\infty} f(\epsilon)d\epsilon = 1$
4. 참값의 MLE는 측정값의 평균값이다. 즉, n번 측정하여 측정값을 각각 $x_1, x_2, \cdots, x_n$이라 할 때 가능도 $L = f(x_1 - \mu)f(x_2 - \mu)\dots f(x_n - \mu)$는 $\mu = \frac{x_1+x_2+\cdots+x_n}{n}$에서 최대값을 갖는다.

조건 1,2,3는 직관적으로 오차의 성질로 받아들일 수 있는 조건들로 이들을 포함한 총 4개의 조건에서 정규분포의 확률밀도함수(PDF)를 직접 수학적으로 유도할 수 있고, 결국 정규분포가 세상의 온갖 측정값을 설명하는 중요한 분포라는 결론에 이르게 된다. 혹시 유도 과정이 궁금한 독자는 http://wiki.mathnt.net/index.php?title=정규분포와_그_확률밀도함수 를 참고하기 바란다.

정확도 : 측정값들이 한쪽으로 몰리는 일이 적은 정도. 즉 계통오차가 적은 정도를 나타내는 개념

1-1. 정확도

정확도는 측정값들이 한쪽으로 몰리는 일이 적은 정도를 나타내며, 표기하는 방법은 아래의 세가지 방법 중 하나를 적용하면 됩니다.

– 평균값과 참값의 백분율로 구하는 방법

측정값들의 평균값을 μ, 참값을 X라 할때 정확도는 다음과 같습니다.

$$\text{정확도}(\%) = \frac{\mu}{X} \times 100 \tag{1}$$

– 절대오차

측정값을 x, 참값을 X라 할 때 절대오차 ϵ은 다음과 같습니다.

$$\epsilon = x - X \tag{2}$$

– 상대오차

상대오차 또는 백분율오차는 다음과 같습니다.

$$\text{상대오차}(\%) = \frac{|\epsilon|}{X} \times 100 \tag{3}$$

정밀도 : 측정값들의 퍼짐이 좁은 정도. 즉 우연 오차가 적은 정도를 나타내는 개념

1-2 정밀도

정밀도는 측정값들의 퍼짐이 좁은 정도를 나타내며, 표기하는 방법은 아래의 네가지 방법 중 하나를 적용하면 됩니다.

– 상대표준편차

측정값의 표준편차를 σ, 평균값을 μ라 했을 때 상대표준편차 $\%\mathrm{RSD}$는 다음과 같습니다.

$$\%\mathrm{RSD} = \frac{\sigma}{\mu} \times 100 \tag{4}$$

– 측정값의 범위를 이용한 정밀도 계산과 표기

측정값의 최대값인 x(max)과 최소값인 x(min)으로 range를 구한 후 '평균값±range'로 표기합니다.

$$Range = x(\max) - x(\min) \tag{5}$$

– 평균편차를 이용한 정밀도 계산과 표기

측정값을 x, 측정값의 평균을 μ라 했을 때, 평균편차 $\bar{d}$를 계산한 후 '평균값±$\bar{d}$'로 표기합니다.

$$\bar{d} = \frac{\Sigma|x-\mu|}{n} \tag{6}$$

– 표준편차를 이용한 정밀도 계산과 표기

모집단 표준편차 σ 또는 표본 표준편차 s 를 계산한 후 '평균값±표준편차'로 표기합니다.

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{n}} = \sqrt{Var[x]}\ (\text{모집단표준편차}) \tag{7}$$

$$s = \sqrt{\frac{\Sigma(x-\mu)^2}{n-1}}\ (\text{표본표준편차}) \tag{8}$$

(7)식에서 주어진 $Var[x]$는 모집단의 x에 대한 분산을 뜻합니다.

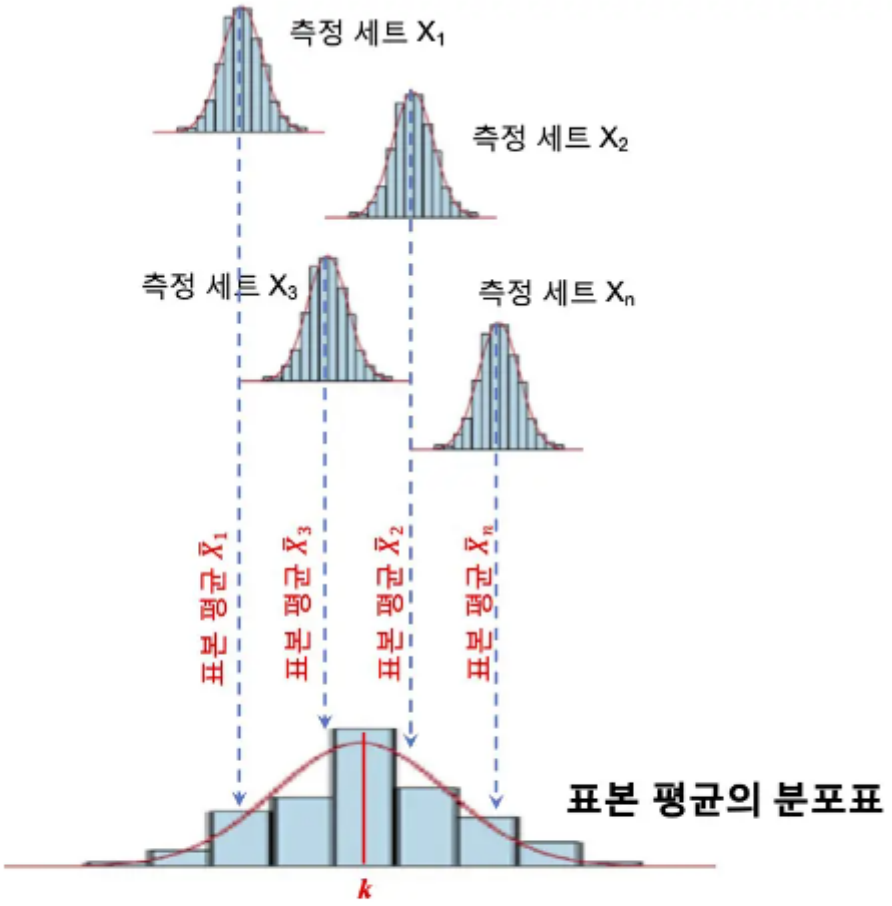
표준오차 : 정확도에서 오차를 구하기 위해서는 참값이 필요한데 참값을 모르는 경우가 많습니다. 결국 우리가 어느 측정값을 표기하기 위해서는 정밀도 뿐만 아니라 오차의 범위도 제시할 필요가 있습니다. 왜냐하면 정확한 참값을 모르기 때문입니다. 이러한 배경에서 출발한 것이 표준오차가 되겠습니다.

계통오차를 제거하여 참값을 도출하는 방법은 무엇일까요? 바로 아래에서 설명할 표본 평균의 평균이라는 값을 이용합니다.

그렇다면 표본 평균의 평균을 참값으로 간주한다면 위에서 측정한 측정값들의 표준편차를 구할 수 있을 것입니다. 이것이 표준 오차(Standard Error of Mean, SEM, 평균오차로 불리기도 함)입니다.

표본 평균의 평균

- (1단계) n명의 사람들이 자신의 저울을 이용하여 각각 물체의 질량을 n번 측정한다고 생각하세요. 그러면 n개의 표본이 생기는 것입니다. 그러면 X1, X2, X3, ..., Xn의 측정값 세트가 존재하겠죠.
- (2단계) 각 측정값 세트의 평균을 내세요. 그러면 X¯1, X¯2, X¯3, ..., X¯n들이 구해질 것입니다. 이것이 표본 평균입니다.
- (3단계) 2단계에서 구한 표본 평균들의 평균을 구하세요. 그 값을 k라고 하고, 이 값이 표본 평균의 평균값입니다.
-



표준 오차 SEM은 표본 평균에 대한 표준편차를 뜻합니다. 각 표본의 측정세트로부터 표본 평균을 구하여 분포표가 만들어집니다. 이 표본 평균 분포표의 평균값이 참값 k로 간주됩니다. 결국 표준 오차는 k로부터 표본 평균들이 어느정도 흩어져 있는가의 척도입니다.

3-2. 표준 오차 정의

표준 오차 SEM은 표본 평균에 대한 표준편차로 정의됩니다.

그러므로 (7)식의 모집단 표준편차 공식을 그대로 사용하는데요. 다만 변수 x 대신에 표본 평균 $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_n$가 적용되고, 평균값 μ 대신에 표본 평균의 평균 k가 적용됩니다.

이를 식으로 표현하면 다음과 같습니다.

$$SEM = \sqrt{\frac{\Sigma(\bar{X}-k)^2}{n}} = \sqrt{Var[\bar{X}]} \tag{9}$$

(9)식에서 $Var[\bar{X}]$를 분산이라고 합니다. 결국 표준 오차는 반복실험으로 구해진 표본 평균들이 k로부터 어느 정도 흩어져 있는가의 척도인 것입니다.

4-2. 표준 오차 SEM을 이용한 측정값의 표기

그렇다면 측정값을 표기할 때 (9) 또는 (14)식으로 구한 표준오차 SEM을 어떻게 활용할까요?

정밀도를 표현할 때와는 약간 다릅니다. 즉 "평균값±(Z-score × SEM)"를 활용하는데요.

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} \tag{15}$$

이 식에서 Z는 90% 신뢰수준의 경우 1.65, 95% 신뢰수준의 경우 1.96, 99% 신뢰수준의 경우 2.58을 적용합니다. 연구에서는 보통 95% 신뢰수준을 적용합니다.

정규분포표의 Z값을 적용하는 것은 실험세트의 측정값이 100개 이상인 경우에 보통 적용합니다. 만일 측정값이 100개 미만인 경우에는 (15)식에서 Z대신에 t분포값이 들어갑니다.

예를 들어 95% 신뢰수준에서 측정값이 5개 뿐인 경우에는 아래 (16)식과 같이 표기합니다. t분포표 보는 방법은 여기를 클릭하세요.

$$\bar{X} \pm 2.776\frac{s}{\sqrt{n}} \tag{16}$$

아울러 사람들이 평균값 $\bar{X}$ 뒤에 쓰여진 것이 정밀도인지 아니면 표준오차인지를 헷갈릴 수 있으므로 반드시 표기해주어야 합니다.

```
library(MASS)
library(ISLR)

attach(Boston)
lm.fit = lm(mdev~lstat)    #medv와 lstat의 상관관계를 선형회귀에 적용
lm.fit

# Call:
#   lm(formula = medv ~ lstat, data = Boston)
#
# Coefficients:
#   (Intercept)        lstat
```

```
#       34.55        -0.95



coef(lm.fit)    #계수 확인
#   (Intercept)        lstat
#       34.55        -0.95

#계수 추정치의 신뢰구간 확인
confint(lm.fit)
#                 2.5 %     97.5 %
#   (Intercept) 33.448457 35.6592247
#   lstat       -1.026148 -0.8739505

#신뢰구간 확인
predict(lm.fit, data.frame(lstat=(c(5,10,15))), interval="confidence" )
#     fit      lwr      upr
# 1 29.80359 29.00741 30.59978
# 2 25.05335 24.47413 25.63256
# 3 20.30310 19.73159 20.87461
# 10일 때의 신뢰구간은 (24.42, 25.63)

#예측구간 확인
predict(lm.fit, data.frame(lstat=(c(5,10,15))), interval="prediction" )
#      fit       lwr      upr
# 1 29.80359 17.565675 42.04151
# 2 25.05335 12.827626 37.27907
# 3 20.30310  8.077742 32.52846
# 10일 때의 예측구간은 (12.82,37.28)

plot(lstat,mdev)
abline(lm.fit)
```