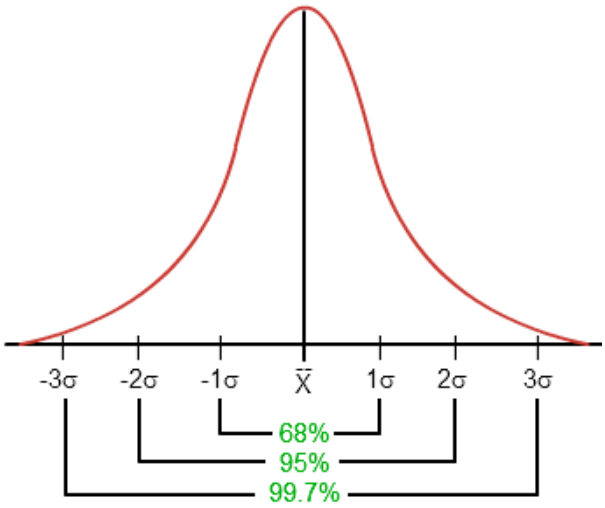
주요법칙

실증적 규칙

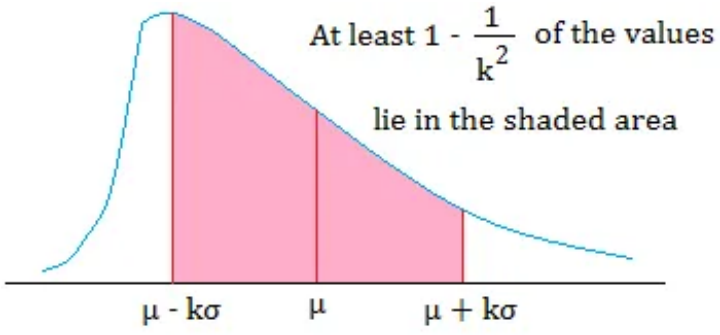


평균 μ, 표준편차 σ인 확률분포함수가 좌우 대칭인 벨 모양인 경우 다음이 성립한다.

$$P(|X-\mu| \le k\sigma) \approx \alpha$$

- $k = 1$: 평균을 중심으로 $\pm$1 표준편차 구간에는 적어도 68% 데이터가 포함되어 있음
- $k = 2$: 평균을 중심으로 $\pm$2*표준편차 구간에는 적어도 95% 데이터가 포함되어 있음
- $k = 3$: 평균을 중심으로 $\pm$3 표준편차 구간에는 적어도 99.8%(대부분) 데이터가 포함되어 있음
- 만약 $k = 6$인 경우, 6시그마 품질, 100만개 중 2개만 범위 밖으로 나간다. 불량

Chebychev 부등식



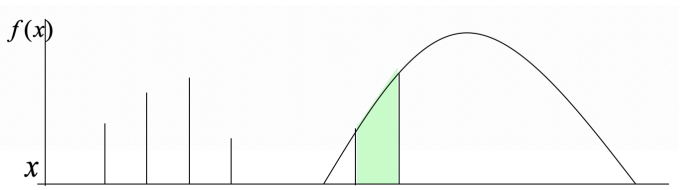
확률변수 X 가 평균 μ , 분산 σ^2 을 갖는 경우 분포의 형태와 상관 없이 양의 상수 k 에 대하여 다음이 성립한다.

$$P(|X-\mu| \le k\sigma) \ge 1 - \frac{1}{k^2}$$

- $k = 2$: 평균을 중심으로 $\pm$1 표준편차 구간에는 적어도 75% 데이터가 포함되어 있음
- $k = 3$: 평균을 중심으로 $\pm$1 표준편차 구간에는 적어도 89% 데이터가 포함되어 있음
- $k = 4$: 평균을 중심으로 $\pm$1 표준편차 구간에는 적어도 95% 데이터가 포함되어 있음

확률 모형(probability model)이란

확률 모형이란 확률 변수의 확률 분포 함수(probability density function)를 의미한다. 확률 분포 함수는 확률 변수(x)가 가질 수 있는 각 값을 정의역(domain) 확률($f(x)$))을 치역으로(range) 한 함수이다. 아래 그림은 이산형 확률 밀도 함수와 연속형 확률 분포(밀도) 함수의 예이다.



이산형 경우에는 막대의 높이가 (히스토그램에서는 바의 높이) 연속형인 경우에는 면적이 확률이다. 그러므로 연속형 확률 분포 함수에서는 x의 단일 값에서 확률은 0이다.

모집단 분포 가정

다음은 모집단의 분포 형태 $f(x)$를 가정하는 예제이다.

- 모집단 전체를 조사한 경우 전체 자료로부터 히스토그램을(이것이 확률 분포 함수이다. 물론 정확한 함수 식은 알 수 없지만) 그리거나 상대 빈도 개념으로 관심 구간의 확률을 구할 수 있다. (예) 한남대학교 학생 중 용돈이 30,000(원)~35,000(원)인 학생의 비율(확률)은? 전체 12,000명 중 용돈이 이 구간에 속하는 학생 수가 확률이 된다.
- 자료를 시뮬레이션(simulation: 모집단 가정) 할 때 사용한다.
- 회귀분석이나 분산 분석에서 오차항에 대한 정규 분포 가정이 있다. 이 가정이 무너지면 회귀 계수(t-검정), 모형의 유의성(F-검정) 검정이 불가능하다.
- 소표본(표본의 크기 n<20~30)일 경우 모평균에 대한 가설 검정 시 모집단은 정규 분포임을 가정한다.

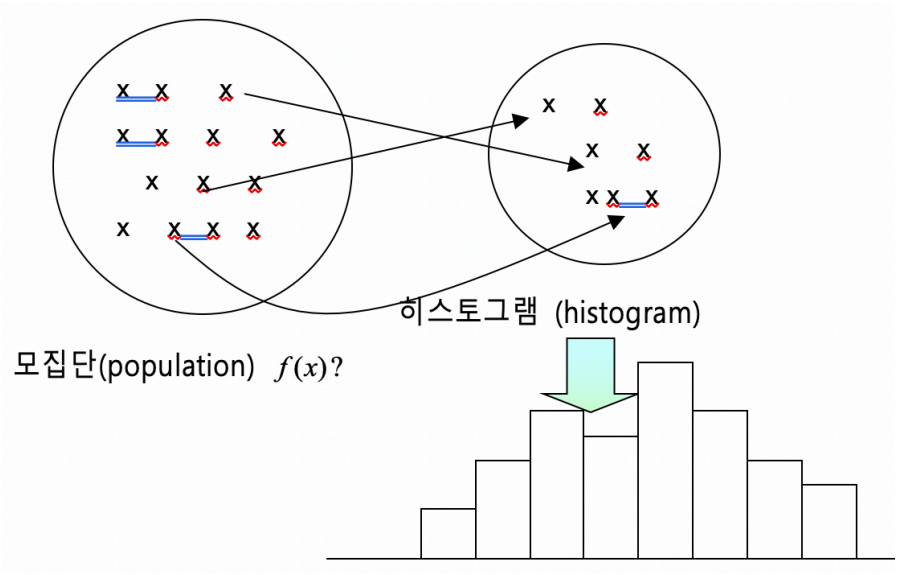
모집단 분포 가정이 불가능한 경우

모집단의 분포를 가정하지 않으면 어떻게 확률 분포를 알 수 있나? 표본 자료의 분포(즉 표본 분포)를 이용할 수 밖에 없다.

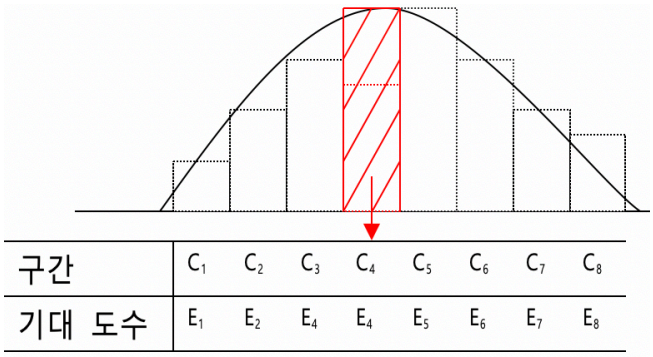
- (Goodness-of-fits:적합성 검정) : 빈도표 활용 검정통계량 방법 χ^2 -검정
- 그래프 이용 방법: P-P plot 방법, 시각적 방법, rule of thumb

Frequency table (빈도표) 개념 이용하기 ~ χ^2-분포 적합성 검정

표본 자료로부터 모집단 확률 밀도 함수를 어떻게 구할 수 있는가? 히스토그램으로부터 확률분포함수를 구한다? 불가능하다. 접근 방법은 표본 자료로부터 빈도표를(이를 관측 빈도) 만들고 (histogram과 동일) 모집단이 따를 것 같은 분포로부터(예:정규분포) 빈도표(이를 기대 빈도)를 만들어 비교하면 빈도의 차이가 거의 없으면 모집단은 그 분포를 따른다고 하자 그렇지 않으면 기각한다.



(이산형)



(연속형)

표본 분포가 설정한 모집단 분포와 동일하다면 관측 도수와 (observed frequency) 기대 도수는 (expected frequency) 비슷한 값일 것이다. 즉 $O_1 \approx E_1$, $O_2 \approx E_2$, ..., $O_k \approx E_k$ (위 예에서는 $k = 8$)

$$TS = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i} \sim \chi^2 (df = k - c - 1)$$

검정통계량 (test statistics) ?

- c = 모수 추정 개수

Probability Plot: 시각적 방법

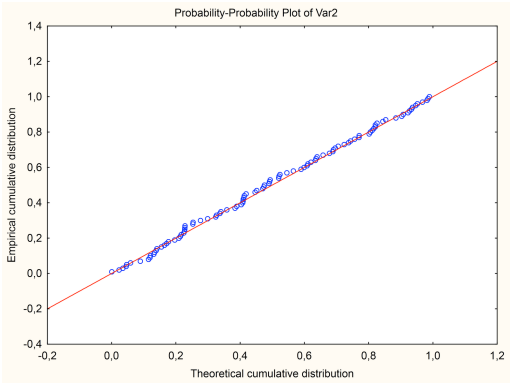
- 두 데이터의 실증적 empirical 분포함수는 동일한가?
- 이론적 분포함수와 데이터의 분포함수는 동일한가? 보여주는 시각적 그래프

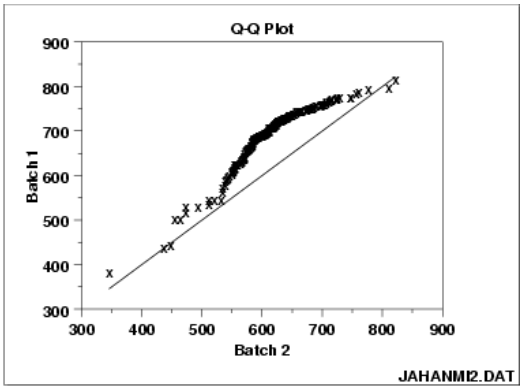
Probability-Probability plot

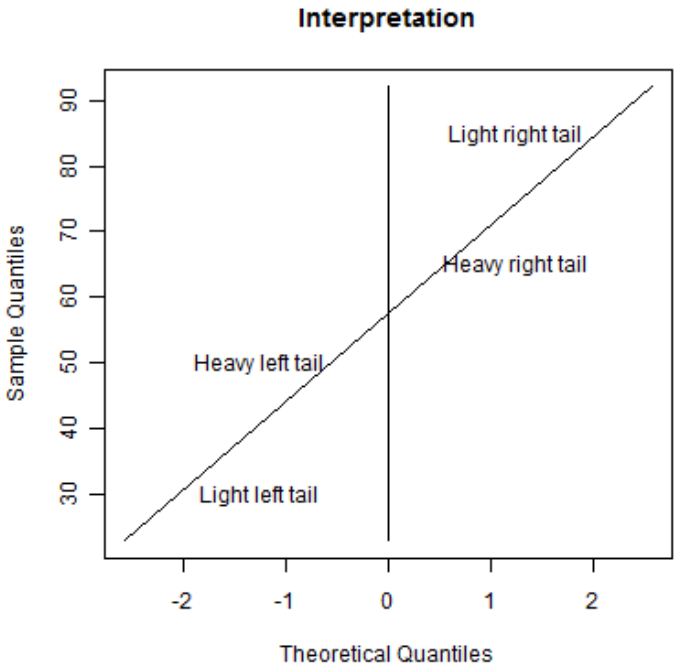
두 데이터의 누적분포함수를 2차원 그래프에 표현함, X-축에는 정규분포의 누적분포함수 $F(z)$, Y-축에는 데이터 누적분포함수 $F(x)$

Quantile-Quantile plot

두 데이터의 누적분포함수를 2차원 그래프에 표현함, X-축에는 한 데이터의 p-백분위 값 $F^{-1}(p_i = (i-0.5)/n)$, Y-축에는 다른 데이터 p-백분위 값 $F^{-1}(p)$







통계량 활용 - 치우침 판단

평균 왜도 pearson moment skewness

- 모집단 왜도 $\tilde{\mu_3} = E(\frac{X-\mu}{\sigma})^3 = \frac{\mu_3}{\sigma^3}$
- 표본왜도 : $skew = \frac{\sum (x_i - \bar{x})^3/n}{(\sum (x_i - \bar{x})^2/n)^{3/2}}$
- 정규분포의 왜도는 0이고 지수분포는 2이다.

중앙값 활용

pearson first skewness (mode skewness)

$$skew = \frac{mean - mode}{std}$$

Pearson's second skewness coefficient (median skewness)

$$skew = \frac{3(mean - median)}{std}$$

사분위 기반 왜도

$$skew = \frac{(Q_3 + Q_1 - 2Q_2)}{IQR}$$

Groeneveld & Meeden's coefficient

$$skew = \frac{mean - median}{E(|X - median|)}$$

- 정규분포=0, 우로 치우침 +, 좌로 치우침 : 이는 통계량의 분포를 모르므로 정규분포 가설을 검정할 수 없어 시각적 판단 수준임

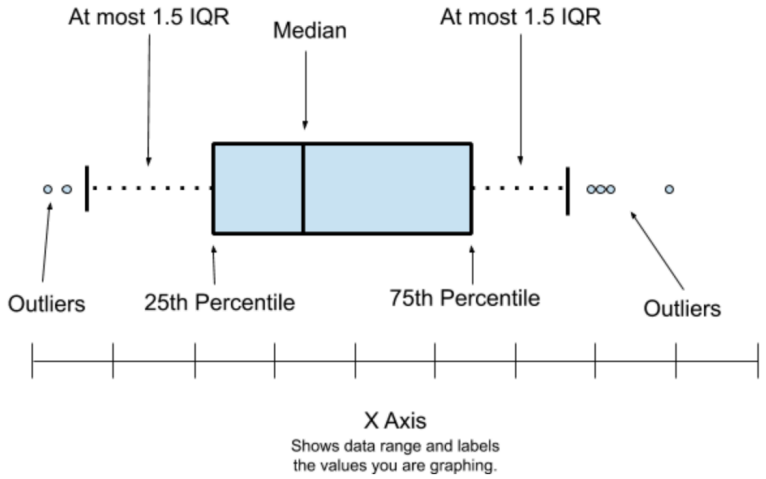
이상치 진단 anomaly detection

벨모양 좌우대칭 분포 z-score

실증적 법칙 empirical rule 에 의하면 $z = \frac{(x - m(x))}{s(x)}$ 값은 ± 2 범위 내에 95%가 속하므로 z-점수가 2를 초과하면 이상치로 판단한다. $m = mean,\ s = standard\ deviation$

모든 분포 형태 적용

상자 그림 box plot



Robust z-점수

$$robust - z = \frac{|x - med(x)|}{mad(x)},\ MAD = 1.4826 \times median(|x_i - med(x)|)$$

$$med = median,\ mad = median\ absolute\ deviation$$

[참고 시험 숙련도 Proficiency test]

$$rz = \frac{x_i - MD}{0.7413 * IQR}$$

- ▷ $|rz| \cong 0$: 매우 일치
- ▷ $|rz| < 2$: 양호
- ▷ $2 \leqq |rz| < 3$: 주의
- ▷ $3 \leqq |rz|$: 미흡

이산형 확률모형

공정 주사위 판단

보유한 주사위가 공정한지 fair 알아보기 위하여 1,000번을 던저 나온 결과를 정리한 것이다. 주사위가 공정한지 유의수준 5%에서 검정하시오.

눈금	1	2	3	4	5	6
빈도	150	160	165	155	170	200

통계적 가설

귀무가설 : 주사위는 공정하다. 각 눈금이 나올 확률은 $\frac{1}{6}$이다.

대립가설 : 주사위는 공정하지 않다.

기대빈도 계산

귀무가설이 옳다는 가정 하에 계산되는 빈도이다.

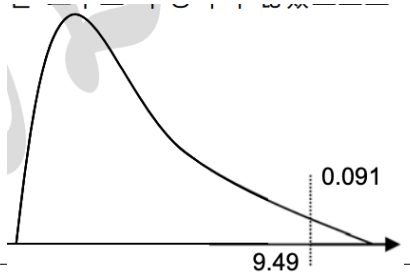
눈금	1	2	3	4	5	6
빈도 O_i	150	160	165	155	170	200
가대빈도 E_i	1000/6	1000/6	1000/6	1000/6	1000/6	1000/6

검정통계량 및 샘플링 분포

$$ts = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(df = k - c - 1), k = \text{셀의 크}$$

기, c = 추정된 모수의 개수

본 예제의 경우 $k = 6, c = 0$이다.



유의확률 p-value 계산

$$p - value = P(\chi^2(df, \alpha) \geq ts)$$

```
import numpy as np
x=[1,2,3,4,5,6]
f=[150,160,165,155,170,200]
e=np.repeat(1000/6,6)
E
```

```
array([166.66666667, 166.66666667, 166.66666667, 166.66666667,
       166.66666667, 166.66666671])
```

```
import scipy.stats as st
ts=sum((f-e)**2/e)
print('Test statistic=%.2f | p-value=%.3f'%(ts,1-st.chi2.cdf(ts,5)))
```

```
Test statistic=9.50 | p-value=0.091
```

유의확률이 유의수준 5%보다 크므로 귀무가설을 기각할 수 없어 주사위는 공정하다.

성비는 공평한가?

우리나라 출생 아이의 성비가 동일한지 알아보기 위하여 자녀가 3인인 1,000 가구의 남아 수를 조사한 자료이다. 이를 이용하여 성비가 동일한지 검정하시오.

남자 아이 수	0	1	2	3
빈도	100	350	400	150

통계적 가설

귀무가설 : 남아, 여아 비율은 동일하다.

대립가설 : 동일하지 않다.

기대빈도 계산

귀무가설이 옳다는 자녀가 3인인 경우 남자의 수 X는 이항분포 $B(n=3,p=0.5)$을 따른다.

그러므로 남자아이 수가 i명 일 빈도=$P(X=i \mid X \sim B(3,0.5)) * 1000, i=0,1,2,3$

검정통계량 및 샘플링 분포

$$ts=\sum_i \frac{(O_i-E_i)^2}{E_i} \sim \chi^2(df=k-c-1), k=\text{셀의 크기}, c=\text{추정된 모수의 개수}$$

본 예제의 경우 $k=4, c=0$이다.

유의확률 p-value 계산

$$p-value=P(\chi^2(df,\alpha) \geq ts)$$

```
import numpy as np
import scipy.stats as st
x=[0,1,2,3]
f=[100,350,400,150]
e=st.binom.pmf(x,3,0.5)*1000; e
```

```
array([125., 375., 375., 125.])
```

```
import scipy.stats as st
ts=sum((f-e)**2/e)
print('Test statistic=%.2f | p-value=%.3f'%(ts,1-st.chi2.cdf(ts,3)))
```

```
Test statistic=13.33 | p-value=0.004
```

유의확률이 0.004로 유의수준 5%보다 작으므로 기각되고 2명, 3명 남자 빈도가 이론빈도보다 높으므로 우리나라 남아 비율이 높다고 할 수 있다.

연습문제

다음은 한남대학교 정문을 통과하는 차량의 수가 Poisson 분포를 따르는지 알아보기 위하여 1 분마다 차량 통과 회수를 300 회 조사하였다. 아래 자료를 이용하여 Poisson 분포를 따르는지 검정하시오. (유의수준=0.05)

통과 차량	0	1	2	3	4	5	6	7
관측 빈도	20	54	74	67	45	25	11	4

연속형 확률모형

연속형 확률변수에 대한 확률모형 적합성은 정규분포를 따르는지 보는 것이 가장 중요하다. 왜냐하면 모든 통계적 방법론이 정규분포(적어도 벨모양의 좌우 대칭에 근사)를 가정하거나 치우침이 없다는 가정을 근거로 개발되었다.

통계적 가설

데이터의 분포가 이론적 정규분포를 따르는지 검정하는 적합성 검정임

- 귀무가설 : 데이터 모집단 분포는 정규분포이다
- 대립가설 : 정규분포를 따르지는 않는다 <=> 그러나 어떤 분포인지는 모른다.

```
import numpy as np
data=np.random.exponential(0.5,100)
```

평균이 0.5인 지수분포를 따르는 난수 데이터 100개를 표본추출하였다.

Shapiro Wilk W-통계량

$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, 상수 a_i는 분산행렬을 이용하여 구함

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \le n \le 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } 12 \le n \le 2000 \end{cases}$$

유의확률이 0.05보다 작으므로 귀무가설이 기각되어 데이터는 정규분포를 따르지 않는다.

```
import scipy.stats as stats
stats.shapiro(data)
```

```
(0.7728214859962463, 3.929337064056959e-11)
```

Kolmogorov D-통계량

$D = \max_x \left| F_n(x) - \Phi(x) \right|$, $\Phi(x)$ 누적정규분포, $Fn(x)$ 데이터누적분포함수

유의확률이 매우 작아 귀무가설이 기각되어 정규분포를 따르지 않는다. 모수 추정 결과를 매개변수로 넣을 수 있다.

```
import scipy.stats as stats
stats.kstest(data,'norm',args=(data.mean(),data.std()))
```

```
KstestResult(statistic=0.1846542541771728, pvalue=0.001858923062673114)
```

적합성 검정 가능한 분포

dist{‘norm’, ‘expon’, ‘logistic’, ‘gumbel’, ‘gumbel_l’, ‘gumbel_r’, ‘extreme1’} 가능한 함수

귀무가설 : 데이터는 지수분포를 따른다.

대립가설 : 데이터는 지수분포를 따르지 않는다.

유의확률이 0.629로 귀무가설을 기각하지 못하므로 지수분포를 따른다.

```
import scipy.stats as stats
stats.kstest(data,'expon',args=(0,data.mean()))
```

```
KstestResult(statistic=0.0744020062140947, pvalue=0.6297010879529599)
```

Anderson-Darling AD 통계량

$$A^2 = n\int (Fn(x) - \Phi(x))^2 \left|\Phi(x)\Phi(1-x)\right|^{-1} d\Phi(x)$$

통계량 값이 6.69로 가장 큰 기각역 값 1.053 (여기에 해당하는 유의수준은 0.01)보다 크므로 귀무가설이 기각되어 정규분포를 따르지 않는다.

```
import scipy.stats as stats
stats.anderson(data,dist='norm')
```

```
AndersonResult(statistic=6.699223575977072, critical_values=array([0.555, 0.632, 0.759, 0.885, 1.053]),
 significance_level=array([15. , 10. ,  5. ,  2.5,  1. ]))
```

대부분의 분퐁에 대한 적합성 검정은 가능하다. scipy.stats 분포

귀무가설 : 데이터는 지수분포를 따른다.

대립가설 : 데이터는 지수분포를 따르지 않는다.

유의수준 15% 기각값보다 작으므로 귀무가설을 기각하지 못하므로 지수분포를 따른다.

```
import scipy.stats as stats
stats.anderson(data,dist='expon')
```

```
AndersonResult(statistic=0.5542469038363862, critical_values=array([0.917, 1.072, 1.333, 1.596, 1.945])
```

해결방안 : 정규변환 Normal Transformation

(간단한 정규변환 normal transformation 방식

우로 치우침 : $\sqrt{X} \rightarrow ln(X) \rightarrow \frac{1}{X}$ -> 치우침의 정도가 극심

좌로 치우침 : $\sqrt{max(X+1)-X} \rightarrow ln(max(X+1)-X) \rightarrow \frac{1}{max(X+1)-X}$

Modified Tukey Ladder of Power

$$Y = \begin{pmatrix} X^{\lambda} \ if \ \lambda > 0 \\ ln(X) \ if \ \lambda = 0 \\ -X^{\lambda} \ if \ \lambda < 0 \end{pmatrix}$$

Box-Cox tranformation : George Box and David Cox

$Y = \frac{X^{\lambda}-1}{\lambda}$, 단 $\lambda = 0$이면 $Y = ln(X)$

Tukey 변환가 동일하지만 최적의 λ 값은 MLE 방법에 의해 찾음

Box-Cox 함수를 사용하면 최적 λ(0.22) 값과 정규변환된(xt) 값이 출력된다. 정규변화된 값에 대한 S-W 정규성 검정을 하면 정규분포를 따른다. 당연한 결과이다.

```
import scipy.stats as stats
xt,alpha = stats.boxcox(data)
```

```
   alpha
0.22562423703607506
```

```
import scipy.stats as stats
stats.shapiro(xt)
```

```
(0.9932992458343506, 0.9054390788078308)
```

함수 적합

선형함수 linear finction

모형 & 데이터

$$Y_i = \alpha + \beta X_i + e_i, i = 1$$

```
def func(x, a, b):
    return a*x + b
```

함수 설정

```
import numpy as np
x = np.linspace(0, 10, 100)
```

설명변수 X [0~10] 정수를 100개 동일 구간으로 나누어 저장

```
Out[26]: array([ 0.        ,  0.1010101 ,  0.2020202 ,  0.3030303 ,  0.4040404 ,
                 0.50505051,  0.60606061,  0.70707071,  0.80808081,  0.90909091,
                 1.01010101,  1.11111111,  1.21212121,  1.31313131,  1.41414141,
                 1.51515152,  1.61616162,  1.71717172,  1.81818182,  1.91919192,
```

```
y=func(x, 1, 2)+0.9*np.random.normal(size=len(x))
```

$y = 1 + 2 * x$ 함수값에 $0.9 * N(0,1)$ 난수 생성(e_i) 값을 더한다.

가장 적합한 함수 구하는 규칙 Least Sqaure Methods

관측치 y_i와 적합치 $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$ 의 차이의 제곱합이 최소가 되는 모수 α, β 구하는 방법

$$min_{\alpha,\beta} \sum_{i}^{n} (y_i - \alpha - \beta x_i)^2$$

```
from scipy.optimize import curve_fit
beta, est_cov = curve_fit(func,x,y)
```

beta에는 OLS 추정값 $\hat{\alpha}, \hat{\beta}$가 저장되고 est_cov에는 추정분산이 출력된다. 대각행렬이 각 모수의 추정분산이 된다.

```
beta
```

```
array([1.06267426, 1.55151878])
```

```
est_cov
```

```
array([[ 0.00096514, -0.00482569],
       [-0.00482569,  0.03233374]])
```

회귀분석과 결과 비교

```
from scipy import stats
import numpy as np
slope, intercept, r_value, p_value, std_err=stats.linregress(x,y)
```

```
slope, intercept, r_value, p_value, std_err
```

```
Out[33]: (1.0626742610849165,
          1.5515187834324884,
          0.9605816298688206,
          2.7442134254151623e-56,
          0.031066665621587178)
```

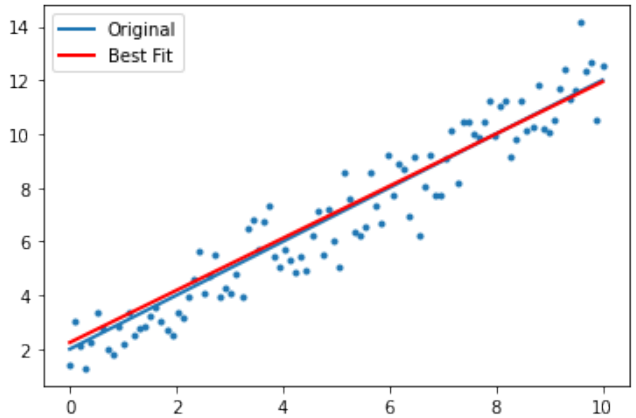
```
import statsmodels.api as sm
model=sm.OLS(y,sm.add_constant(x))
fit=model.fit()
```

```
fit.summary()
```

Dep. Variable:	y	**R-squared:**	0.923
Model:	OLS	**Adj. R-squared:**	0.922
Method:	Least Squares	**F-statistic:**	1170.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5515	0.180	8.628	0.000	1.195	1.908
x1	1.0627	0.031	34.206	0.000	1.001	1.124

```
import matplotlib.pyplot as plt
plt.scatter(x, y, marker='.')
plt.plot(x, 2+1*x, linewidth=2)
plt.plot(x, func(x,*popt), color='red', linewidth=2)
plt.legend(['Original', 'Best Fit'], loc=2)
plt.show()
```



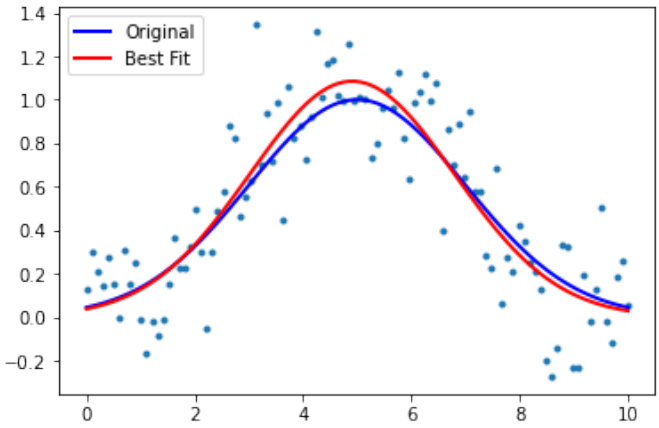
비선형함수

$$y_i = a exp(\frac{-(x-b)^2}{2c^2})$$

```
def func(x, a, b, c):
    return a*np.exp(-(x-b)**2/(2*c**2))
x = np.linspace(0, 10, 100)
y = func(x, 1, 5, 2) # 답인 y들과
y_gen = y + 0.2*np.random.normal(size=len(x)) # noise
beta,est_cov=curve_fit(func, x, y_gen)
beta
```

```
Out[40]: array([1.08475072, 4.90685069, 1.88868034])
```

```
plt.scatter(x, y_gen, marker='.')
plt.plot(x, y, linewidth=2, color='blue')
plt.plot(x, func(x, *beta), color='red', linewidth=2)
plt.legend(['Original', 'Best Fit'], loc=2)
plt.show()
```



연습문제

$$a_0 \exp\left(\frac{-(x-b_0)^2}{2c_0^2}\right) + a_1 \exp\left(\frac{-(x-b_1)^2}{2c_1^2}\right)$$

```
x = np.linspace(0, 20, 200)
```