Cheating Sheet

- 필요한 부분만 참고하여 문제를 푸세요.

최빈값 (mode)

: 최빈값은 빈도수가 가장 많이 발생한 괄찰값을 말함

- ex) 1, 3, 6, 6, 6, 7, 7, 12, 12, 19 있을때, 최빈값은 6이다.

중앙값 (median)

: 중앙값은 수치로 된 자료를 크기순서대로 나열할 때, 가장 가운데에 위치하는 관찰값을 말한다.

- ex) 1, 2, 4, 5, 7, 9, 10 있을때, 중앙값은 5이다.

산술평균 (arithmetic mean)

: 우리가 흔히 사용하는 간단한 평균, 그냥 "평균" 이라고도 한다.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X_i}{n}$$

가중평균 (weighted arithmetic mean)

: 같은 모집단에서 표본을 서로 다른 개수로 뽑은 경우(가중치가 존재하는 경우) 평균값을 구할때 사용

$$\bar{X} = \frac{n_1\bar{X_1} + n_2\bar{X_2} + \cdots + n_k\bar{X_k}}{n_1 + n_2 + \cdots + n_k} = \frac{\sum n_i\bar{X_i}}{n_i}$$

분산 (variance)

: 자료가 평균으로부터 얼마나 떨어져 분포하는지를 가늠하는 숫자
: 분산이란 각각의 관찰값에 대한 평균과의 편차를 제곱하여 그 평균을 구한 것

- 모집단의 분산 (σ^2)
- 표본의 분산 (S^2)
  - n 대신 $(n-1)$을 나누는 이유는, $(n-1)$을 나누어줌으로써 모집단의 σ를 추정하는데 더 적절한 표준편차를 구하기 위함이다

$$\sigma^2 = \frac{\sum_1^N (x_i - \mu)^2}{N} = \frac{\sum_1^N x_i^2 - \mu^2}{N}; \qquad S^2 = \frac{\sum_1^n (x_i - \overline{X})^2}{n-1} = \frac{\sum_1^n x_i^2 - n\overline{X^2}}{n-1}$$

여기에서 N: 모집단의 크기

n: 표본의 크기

표준편차 (standard deviation)

: 분산의 양의 제곱근

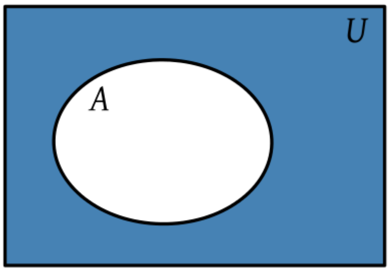
- 모집단의 표준편차 ($\sigma = \sqrt{\sigma^2}$)
- 표본의 표준편차 ($S = \sqrt{S^2}$)

집합이론

- 확률이론을 쉽게 설명하기 위해서는 집합이론의 용어와 부호 사용하는 것이 편리
- 집합 (set) 이란 개체 또는 원소 (element)의 모임이라 정의
- 원소는 { ... } 속에 넣는 것이 관례
    - ex. A = { 남자, 여자 }, B = { 10대, 20대, 30대, ... }

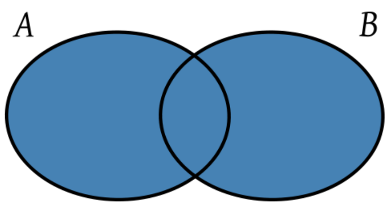
1. 여집합

- A^C = { 전체집합 중에서 집합 A 에 포함되지 않는 원소들 }



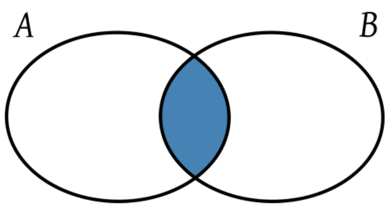
2. 합집합

- $A \cup B$ = { 집합 A 또는 집합 B에 속하는 원소 }



3. 교집합

- $A \cap B$ = { 집합 A와 B의 공통 원소 }



합집합의 계산

- $A \cup B = A + B - A \cap B$
    - if) 집합 A와 B가 서로 배타적(mutually exclusive)일 때 ($A \cap B = \varnothing$)
    - $A \cup B = A + B$

조건부확률

- 사건 B가 발생했다는 조건하에서 사건 A가 발생할 확률

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

베이즈정리 (Bayes' theorem)

- 사전에 알고 있는 정보에 기준을 두고, 어떤 사건이 일어나게 될 확률을 계산하는 이론

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \cdot$$

- Extended form

$$P(B) = \sum_j P(B \mid A_j) P(A_j),$$

$$\Rightarrow P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_j P(B \mid A_j)\, P(A_j)} \cdot$$

기댓값 (Expected Value)

- 확률분포의 평균값 (average, weigthed average)
- 표기법 : $E(X)$ or μ_X
- 기댓값의 계산

$$E(X) = \Sigma X_i \cdot P(X_i)$$

- **기댓값의 특성**

1. 확률변수 X 에 일정한 상수 a 를 곱한 확률변수의 기댓값은 확률변수 X 의 기댓값에 a 를 곱한 것과 같다.
   - $E(aX) = a \cdot E(X)$
2. 확률변수 X 에 일정한 상수 b 만큼을 가감한 확률변수의 기댓값은 확률변수 X 의 기댓값에 b 를 가감한 것과 같다.
   - $E(X + b) = E(X) + b$ or $E(X - b) = E(X) - b$
3. 위의 두 가지 결과를 결합하면 다음 식이 성립된다.
   - $E(aX \pm b) = a \cdot E(X) \pm b$

분산 (Variance)

- 확률분포의 분산
- 표기법 : $Var(X)$ or σ_X^2
- 분산의 계산

$$\begin{aligned} Var(X) &= \Sigma[X_i - E(X)]^2 \cdot P(X_i) \\ &= E[\{X - E(X)\}^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

표준편차 (Standard Deviation)

- 확률분포의 표준편차
- 표기법 : σ
- 표준편차의 계산

$$\sigma_X = \sqrt{\Sigma[X_i - E(X)]^2 \cdot P(X_i)}$$

- **분산과 표준편차의 특성**

  1. 어떤 확률변수에 일정한 상수를 더한 확률변수의 분산은 본래의 확률변수의 분산과 같다. 확률변수에 상수를 더하는 것은 분포의 분산도에는 아무런 영향을 미치지 못하기 때문이다.
     - $Var(X+b) = Var(X)$
       $\sigma(X+b) = \sigma(X)$
  2. 어떤 확률변수에 일정한 상수 a 를 곱한 확률변수의 분산은 본래의 확률변수의 분산에 a^2 를 곱하 것과 같다.
     - $Var(aX) = a^2 Var(X)$
       $\sigma(aX) = a \cdot \sigma(X)$
  3. 위의 두 식을 종합하면 다음과 같은 식이 성립된다.
     - $Var(aX+b) = a^2 Var(X)$
       $\sigma(aX+b) = a \cdot \sigma(X)$

베르누이분포 (Bernoulli Distribution)

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1-p & \text{if } k = 0. \end{cases}$$

This can also be expressed as

$$f(k;p) = p^k (1-p)^{1-k} \quad \text{for } k \in \{0, 1\}.$$

이항확률분포 (Binomial Probability Distribution)

$$\Pr(K=k) = f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\begin{aligned} k &: \text{성공횟수} \\ n &: \text{시행횟수} \\ p &: \text{성공확률} \\ 1-p = q &: \text{실패확률} \\ \binom{n}{k} &: \ {}_nC_k \end{aligned}$$

이항분포의 기댓값과 분산

$$\begin{aligned} \text{기댓값} \quad & \mu = E(X) = np \\ \text{분산} \quad & \sigma^2 = Var(X) = np(1-p) = npq \\ \text{표준편차} \quad & \sigma = \sqrt{np(1-p)} = \sqrt{npq} \end{aligned}$$

다항분포 (Multinomial Distribution)

- 실험의 결과 또는 표본을 뽑는 결과가 상호배타적인 k 개의 사건으로 나타나는 경우
  - ex. 주사위를 던지는 실험 => { 1, 2, 3, 4, 5, 6 }

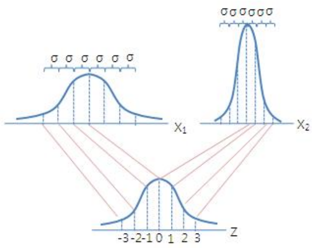
$$p(x_1, x_2, \cdots, x_k; n, p_1, \cdots, p_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

k : 발생가능한결과갯수($k = 2$이면이항분포와같다)
n : 전체시행횟수
x_i : 각결과별발생횟수
p_i : 각결과별확률

표준정규분포

- 표준정규분포는 모든 정규분포를 평균 $\mu = 0$, 표준편차 $\sigma = 1$이 되도록 **표준화**한 것이다. 어떤 확률변수 X의 관찰값이 그 분포의 평균으로부터 표준편차의 몇 배 정도나 떨어져 있는가를 다음과 같이 표준화된 확률변수 Z로 나타내기 때문에 표준정규분포를 Z-분포 라고도 한다.

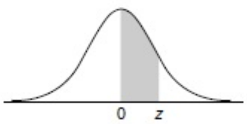
$$Z = \frac{X - \mu}{\sigma}$$



< figure. Standard Normal Distribution >

Table AIV.2 Standard Norms Table

Area between 0 and z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

< figure. 표준정규분포표 >

모집단 평균의 구간 추정 (σ 를 알고 있는 경우)

Z-통계량

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}}$$

Z 값에 대한 신뢰구간

$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1 - \alpha$$

μ 값에 대한 신뢰구간

$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma_{\bar{X}}) = 1 - \alpha$$

신뢰도 $(1-\alpha)$	$Z=0$**에서** $Z_{\alpha/2}$**까지 면적**	$Z_{\alpha/2}$
0.90	0.450	1.64
0.95	0.475	1.96
0.99	0.495	2.57

< table. 신뢰도에 따른 $Z_{\alpha/2}$ 값 >

- 신뢰도 또는 신뢰수준 (confidence level)
  - $1 - \alpha$
  - 신뢰도는 구간으로 추정된 추정값이 실제 모집단의 모수를 포함하고 있을 가능성
- 신뢰구간 (confidence interval)
  - 이때 모수가 포함될 것으로 추정된 구간
- **신뢰도가 높을수록 신뢰구간은 넓어진다.**
  - 이는 범위가 넓을수록 그 속에 모집단의 평균이 포함될 가능성이 더 높아짐을 뜻하며,
  - 반면, 범위가 넓을수록 신뢰구간이 갖는 정보의 가치는 줄어들게 됨을 의미한다.

통계적 가설검정의 순서

1. 귀무가설(H_0) 과 대립가설(H_a) 의 설정
2. 유의수준(α) 의 결정
3. 유의수준을 충족시키는 임계값의 결정
4. 통계량의 계산과 임계값과의 비교
5. 결과의 해석

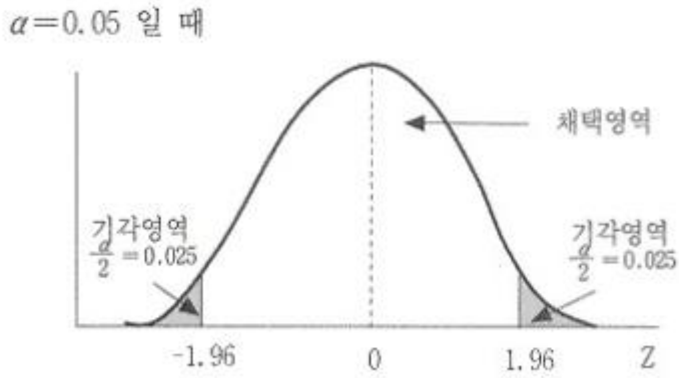
가설검정의 기본용어

귀무가설과 대립가설

- 귀무가설 (H_0 : null hypothesis)
  - 직접 검정대상이 되는 가설
- 대립가설 (H_a or H_1 : alternative hypothesis)
  - 귀무가설이 기각될 때 받아들여지는 가설

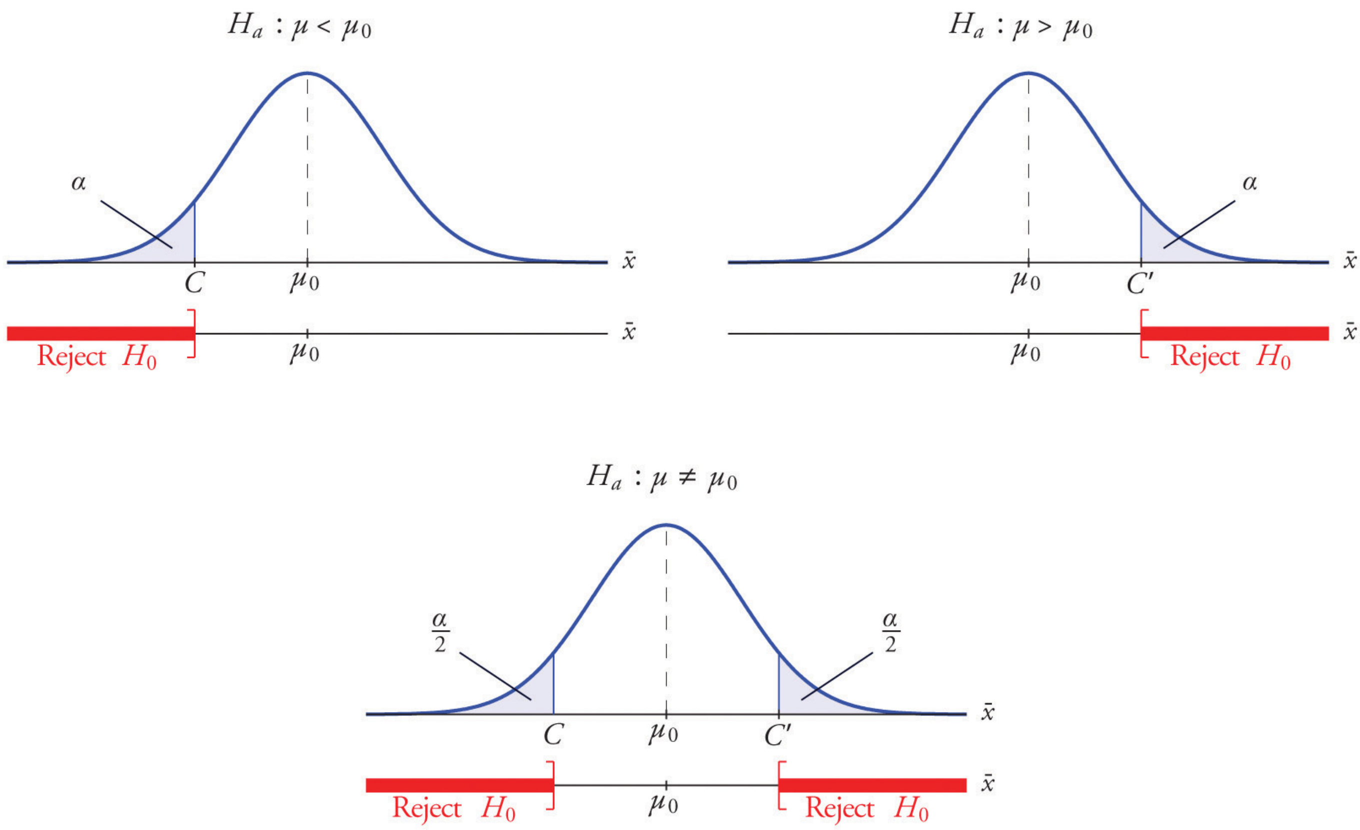
유의수준과 임계값

- 표본에서 계산된 통계량이 가설로 설정된 모집단의 성격과 현저한(significant) 차이가 있는 경우에는 모집단에 대해 설정한 귀무가설을 기각하게 된다
- 이때 명확히 밝혀두어야 할 두가지가 있는데, 첫째는 현저하게 차이가 난다는 것이 무엇을 의미하는지, 둘째는 모집단에 대해 설정한 가설을 채택 또는 기각하는 임계값이 어떤 점이 되어야 하는지이다



- 유의수준 (significance level)
  - 오류를 감수할 확률 (때문에 오류를 범했을 때 손실이 얼만만큼 발생하느냐가 큰 고려요인)
  - 유의수준을 얼마로 할 것인가에 대해서는 연구의 성격, 연구자의 주관 등이 개입되게 되므로 어느 연구에나 적용될 수 있는 보편타당한 기준은 없다
  - 보통 연구에서는 α수준을 $0.01, 0.05, 0.10$ 등으로 정하는 경우가 많다
- 임계값 (critical value)
  - 주어진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준이 되는 값
  - 이 임계값을 중심으로 귀무가설의 기각영역(rejection area)과 채택영역(acceptance area)이 결정된다

양측검정과 단측검정



- 양측검정 (two-tailed test)
  - $H_0 : \mu = \mu_0$
  - 표본 통계량이 μ 보다 현저히 크거나 작으면 기각
  - 기각 영역은 확률분포의 양측에 있게 되므로, 유의수준 α도 양쪽 극단으로 갈려 한쪽의 면적이 $\alpha/2$가 된다

- 단측검정
  - $H_0 : \mu \geq \mu_0 \quad or \quad H_0 : \mu \leq \mu_0$
  - $H_0 : \mu \geq \mu_0$의 경우, 표본 통계량이 μ_0보다 현저히 작으면 기각
  - 기각 영역은 확률분포의 한쪽 극단에만 존재

회귀분석의 개념

- 회귀분석의 목적
  - **회귀분석은 함수적 관계로 알고 있는 두 변수의 관계를 자료를 통해 확인해 보는 것 (인과관계 파악)**
  - **한 변수를 기초로 하여 다른 변수를 예측하는 것 (예측)**
- 회귀분석 종류
  - 단순회귀분석 (simple regression analysis) : 하나의 독립변수와 하나의 종속변수 사이의 관계 분석
  - 다중회귀분석 (multiple regression analysis) : 여러개의 독립변수들과 하나의 종속변수 사이의 관계 분석

단순회귀모형과 회귀식

- 독립변수와 종속변수 간의 1차함수관계 또는 선형관계를 가정할 때 회귀모형은 다음 두 요소를 결합한 형태로 나타낼 수 있다.
  - 확정적 함수관계를 나타내는 부분 : $\alpha + \beta X_i$
  - 확률적 오차항 : ϵ_i
- α, β : 회귀계수 (regression coefficient)
  - 회귀식을 보면 절편을 α 로 하고 기울기가 β 인 직선이 됨을 알 수 있다.
  - 이에 반해서 회귀모형을 그림에서 보면 독립변수와 종속변수가 점으로 나타나 있으며 오차항에 따라서 그 모양이 다양하게 될 수 있다. (독립변수에 대응하는 종속변수의 값이 오차항에 따라서 확률적으로 다르게 나타나기 때문)

모집단의 경우

$$\begin{aligned} \text{단순회귀모형} \quad & Y_i = \alpha + \beta X_i + \epsilon_i \\ \text{단순회귀식} \quad & \mu_{Y \cdot X_i} = \alpha + \beta X_i \end{aligned}$$

- 모집단의 회귀식을 구하는 것은 실제로 불가능한 경우가 대부분이므로 우리는 표본으로부터 회귀식을 구하여 모수를 추정하여야 한다.
  - $\hat{Y}_i$: 회귀식을 통해 구해지는 수치 (예측값)
  - e_i : 예측오차, 추정오차 또는 잔차(residual)

$$e_i = Y_i - \hat{Y}_i$$

표본의 경우

$$\begin{aligned} \text{단순회귀모형} \quad & Y_i = a + bX_i + e_i \\ \text{단순회귀식} \quad & \hat{Y}_i = a + bX_i \end{aligned}$$