

마케터를 위한 기초통계

- 디지털 마케팅 SCHOOL 5기 (17. 4. 24)

수업의 개요와 흐름잡기

Chapter 3. 분석의 기본 단계 (과정) 알기

- 통계적 가설검정
 - 유의수준과 P-value

Chapter 4. 분석 방법론 (모형) 배우기

- 여러가지 상황에 따른 검정의 종류
- ANOVA 분석
- 상관분석
- 회귀분석

Chapter 3. 분석의 기본 단계 (과정) 알기

- 통계적 가설검정
 - 유의수준과 P-value

3.3. 통계적 가설검정

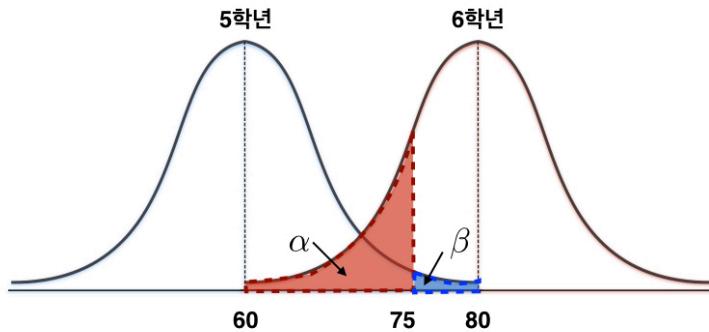
- 가설검정의 개념
- 기본 용어
 - 귀무가설과 대립가설
 - 유의수준 (P-value)
 - 양측검정과 단측검정
- 순서 & 예

3.3.1. 가설검정의 개념

- 통계적 가설검정
 - 표본에서 얻은 사실을 근거로 하여, 모집단에 대한 가설이 맞는지 틀리는지 통계적으로 검정하는 분석 방법
 - 가설의 참 / 거짓을 결정짓는 것은 확신(가능성)의 차이

3.3.1. 가설검정의 개념 (예)

- 어느 초등학교 6학년과 5학년 학생들이 같은 문제를 가지고 시험을 본 결과, 성적의 분포가 아래 그림과 같았다. 6학년 학생들의 평균 성적은 80점이었고, 5학년 학생들의 평균 성적은 60점이었으며, 두 분포는 정규분포를 이룬다고 한다. 어느 교사가 한 학생을 선택하여 학년을 확인하지 않고 그 학생의 점수를 물어보았더니, 시험 성적이 70점이라고 했다. 이때 그 학생은 5학년일까 아니면 6학년일까?



(1)

- 먼저 교사는 다음과 같은 가설을 세웠다.
 - "그 학생은 6학년이다."
- 교사는 이 가설이 틀릴 경우에 대비하여 다른 가설을 세웠다.
 - "그 학생은 5학년이다."

(2)

- 70점을 받은 그 학생이 5학년인지 6학년인지 판단하기 전에, 먼저 몇 점 이상을 6학년으로 보아야 하는지를 결정하여야 한다.
- 만약 75점 이상을 6학년으로 간주한다면
 - 70점을 받은 학생에 대해 "그 학생은 6학년이다"라는 가설은 기각 되고,
 - 대립적으로 설정한 "그 학생은 5학년이다"라는 가설이 채택될 것이다.

(3)

- 문제는 그 학생이 6학년이다, 또는 5학년이다라고 결론을 내릴 때 오류의 위험이 따르고 있다는 것이다.
- 75점 이상이 6학년이다라는 판단기준을 세웠다면
 - α 의 부분만큼은 6학년이면서 5학년이라는 오해를 받을 수 있고,
 - 반면 β 부분만큼은 5학년이지만 6학년으로 오해 받을 수도 있다.

3.3.2. 기본용어 (1) - 귀무가설, 대립가설

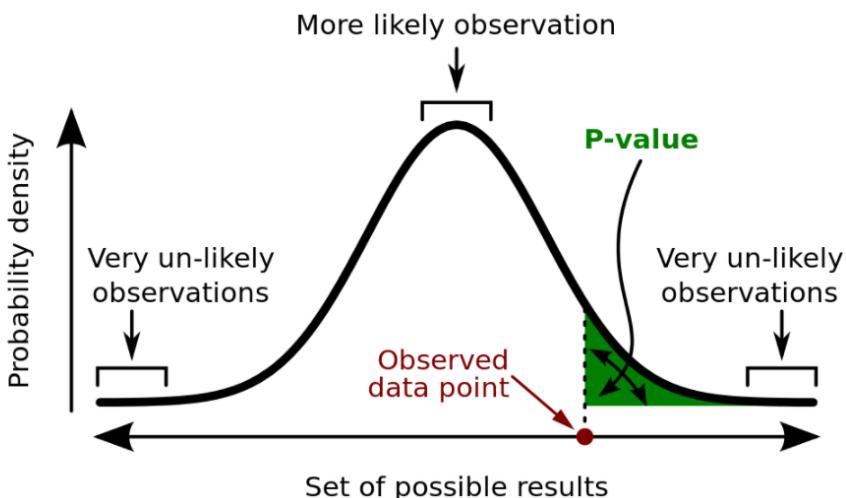
- 귀무가설 (H_0 : null hypothesis)
 - 직접 검정대상이 되는 가설
 - ex. H_0 : 그 학생은 6학년이다.
- 대립가설 (H_a or H_1 : alternative hypothesis)
 - 귀무가설이 기각될 때 받아들여지는 가설
 - ex. H_α : 그 학생은 5학년이다.

3.3.2. 기본용어 (2) - 유의수준 (Significance level)

- 유의수준 (significance level)
 - 오류를 감수할 확률 (따라서 오류를 범했을 때 손실이 얼마나 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 α 수준을 0.01, 0.05, 0.10 등으로 정하는 경우가 많다

3.3.2. 기본용어 (3) - P-value

- 표본에서 계산된 통계량이 가설로 설정된 모집단의 성격과 **현저한(significant)** 차이가 있는 경우에는 모집단에 대해 설정한 귀무가설을 기각하게 된다.

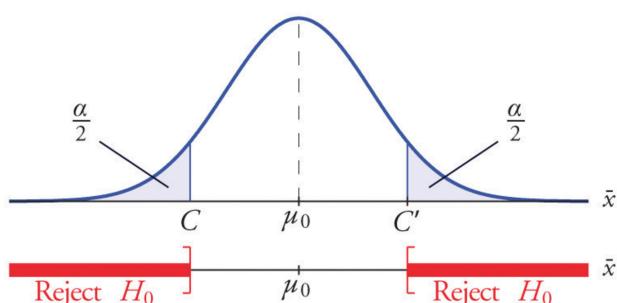


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

3.3.2. 기본용어 (4) - 양측 & 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 α 도 양쪽 극단으로 갈려 한쪽의 면적이 $\alpha/2$ 가 된다

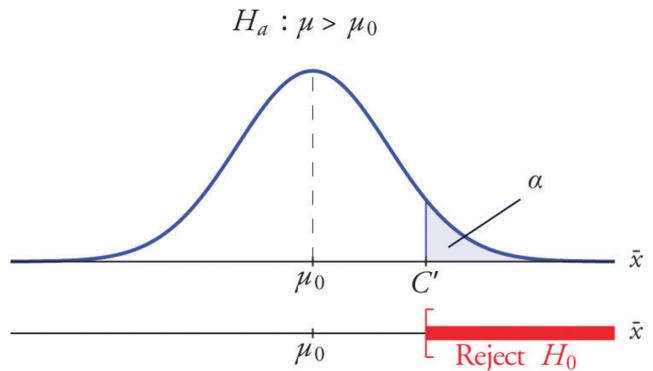
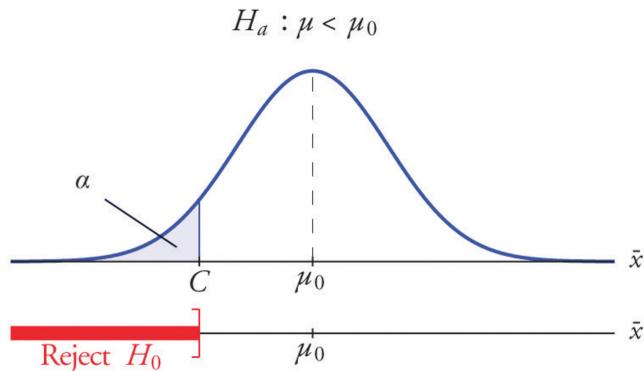
$$H_a : \mu \neq \mu_0$$



3.3.2. 기본용어 (4) - 양측 & 단측검정

- 단측검정 (one-tailed test)

- $H_0 : \mu \geq \mu_0$ or $H_0 : \mu \leq \mu_0$
- $H_0 : \mu \geq \mu_0$ 의 경우, 표본 통계량이 μ_0 보다 현저히 작으면 기각
- 기각 영역은 확률분포의 한쪽 극단에만 존재



3.3.3. 가설검정의 순서

- 귀무가설(H_0)과 대립가설(H_a)의 설정
- 유의수준(α)의 결정
- 유의수준을 충족시키는 임계값의 결정
- 통계량의 계산과 임계값과의 비교
(p-value로 비교시 3, 4번 단계 한번에 가능)
- 결과의 해석

3.3.4. 가설검정 예

- 국내 아이돌그룹 멤버들의 평균 키를 알기 위하여, 16명의 아이돌그룹 멤버의 키를 표본조사하였더니 평균 키가 175cm였다. 국내 아이돌그룹 전체의 평균 키에 대한 표준편차가 5cm라고 하면, 국내 아이돌그룹 멤버의 평균 키가 180cm 이상이라고 할 수 있을까? 유의수준 (α)을 5%로 하여 검정하라.

① 귀무가설 설정 & ② 유의수준 결정

$$\textcircled{1} \quad H_0 : \mu \geq 180\text{cm}$$

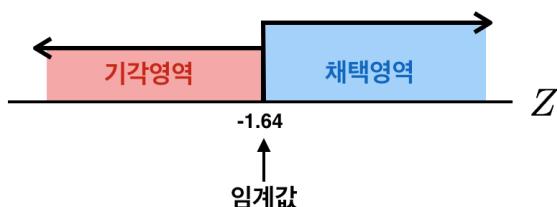
$$H_a : \mu < 180\text{cm}$$

$$\textcircled{2} \quad \alpha = 0.05(5\%)$$

③ 유의수준을 충족시키는 임계값 결정

$$\textcircled{3} \quad \text{채택영역} : Z \geq -1.64$$

$$\text{기각영역} : Z < -1.64$$

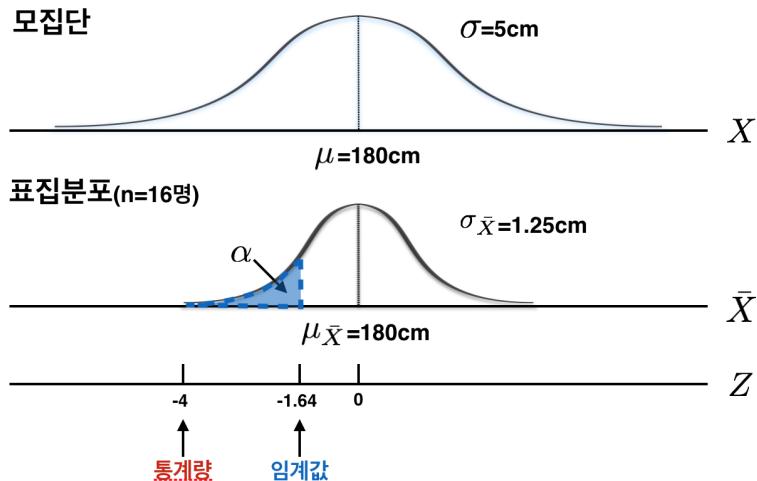


④ 통계량의 계산과 임계값의 비교

④ 175cm에 대응하는 Z값

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{175 - 180}{5/\sqrt{16}} = \frac{-5}{1.25} = -4$$

$Z = -4$ 는 -1.64 보다 작아서 기각영역에 속하므로 H_0 를 기각한다.



⑤ 결과의 해석

- 위의 결과로부터 국내 아이돌그룹 멤버들의 평균 키가 180cm 이상이라고 할 수 없다.

Chapter 4. 분석 방법론 (모형) 배우기

- 여러가지 상황에 따른 검정의 종류
- ANOVA 분석
- 상관분석
- 회귀분석

4.1. 여러가지 상황에 따른 검정의 종류

- 단일 모집단에 대한 검정
 - 평균에 대하여 >>> **Z-test or t-test**
 - 분산에 대하여 >>> **χ^2 -test**
- 두 모집단에 대한 검정
 - 평균에 대하여 >>> **Z-test or t-test**
 - 짝을 이룬 표본의 평균에 대하여 >>> **t-test**
 - 분산에 대하여 >>> **F-test**
- 그 이상 모집단에 대한 검정 >>> **ANOVA test**
- 결론
- 실습

4.1.1 단일 모집단에 대한 검정

- 평균에 대하여

예 (1) - Z-test

- 우리나라 여성 전체의 평균 키는 160cm 이고, 분산은 200 이라고 한다. $10,000$ 명을 표본으로 하여 조사한 결과 평균 169cm 를 얻었다. 우리나라 여성의 평균 키가 160cm 라고 할 수 있을까?
- 모집단 평균을 모르기 때문에 가설검정을 하려는 것인데, 모집단 분산을 알고 있다고 가정하는 것은 비현실적

현실적으로 t-test

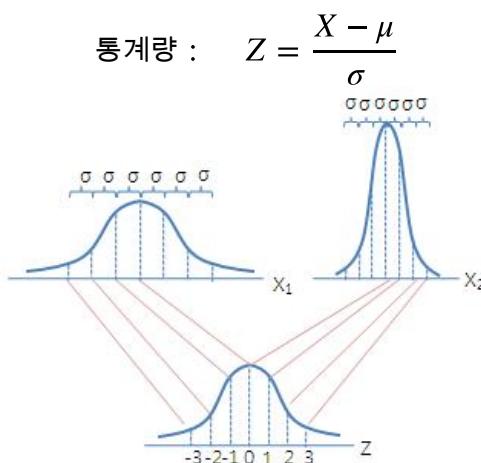
- 우리나라 여성 전체의 평균 키는 160cm 라고 한다. 이에 대한 가설을 검정하기 위하여 $10,000$ 명을 표본으로 하여 조사한 결과 평균 169cm , 분산 300 을 얻었다. 우리나라 여성의 평균 키가 160cm 라고 할 수 있을까?

상황에 따른 Z-test 와 t-test

- 모집단 분포가 정규분포를 이루며, 분산을 알고 있을 때 $\rightarrow Z$ - 분포
- 모집단 분포가 정규분포이긴 하지만, 분산을 모를 때 $\rightarrow t$ - 분포
- 그러나 표본의 크기가 크면, 모집단 분산과 표본에서 뽑은 분산 간의 차이가 작기 때문에 t통계량을 사용하거나 Z통계량을 사용하거나 별 차이가 없다. (중심극한정리)

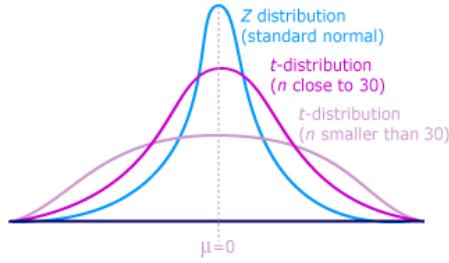
모집단의 분산을 알고 있을 때	표본이 클 때 ($n \geq 30$)	표본이 작을 때 ($n < 30$)
모집단이 정규분포	Z - 분포	Z - 분포
모집단이 비정규분포	Z - 분포	-
모집단의 분산을 알고 있을 때	표본이 클 때 ($n \geq 30$)	표본이 작을 때 ($n < 30$)
모집단이 정규분포	Z - 분포	t - 분포
모집단이 비정규분포	Z - 분포	-

Z-분포 (표준정규분포)



t-분포

$$\text{통계량} \quad t = \frac{(\bar{X} - \mu_{\bar{X}})}{S_{\bar{X}}}, \quad (S_{\bar{X}} = \frac{S}{\sqrt{n}})$$



4.1.2. 단일 모집단에 대한 검정

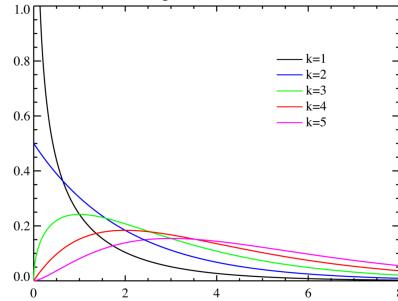
- 분산에 대하여

예 (2) - χ^2 -test

- 어느 고등학교에서는 고교평준화를 시행한 후 이전보다 학생들의 성적이 고르지 않다는 주장을 하고 있다. 평준화 전의 성적의 분산은 $\sigma^2 = 60$ 이라고 하며, 교육과학부에서는 지금도 전과 마찬가지일 것이라는 주장이다. 한 연구자는 교육과학부의 주장을 검정하기 위하여 61명을 선택하여 그 표본의 분산을 계산하여 본 결과 $S^2 = 70$ 이었다. 그 표본이 $\sigma^2 = 60$ 인 모집단에서 나온 것이라고 할 수 있는가를 $\alpha = 0.10$ 에서 검정하려 한다. 고등학교의 성적분포는 정규분포라고 가정한다.

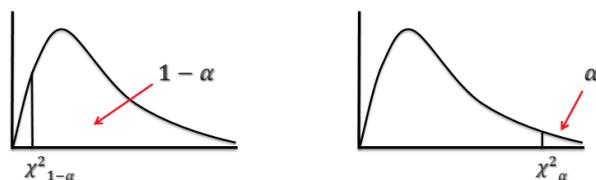
χ^2 -분포(chi-square distribution)

$$\text{통계량} : \quad \chi^2_{n-1} = \frac{(n-1) \cdot S^2}{\sigma^2}, \quad (n-1 = k : \text{자유도})$$



χ^2 -검정 (chi-square test)

- 분산에 대한 가설검정도 평균에 대한 가설검정과 마찬가지로, 계산된 통계량이 표집분포에서 채택영역에 속하는지 또는 기각영역에 속하는지를 알아서 귀무가설을 채택, 거부한다.
- 만일 표본의 자료에서 계산된 χ^2 값이 χ^2 - 분포의 양끝에 있다면(or 계산된 p-value가 유의수준보다 작다면), 그 표본이 귀무가설에서 설정한 모집단에서 뽑혔다고 볼 수 없을 정도로 예외적이라고 볼 수 있기 때문에 귀무가설을 기각한다.



4.1.3. 두 모집단에 대한 검정

- 평균에 대하여 (평균의 차이 : $\bar{X}_1 - \bar{X}_2$)

예 (3) - Z-test or t-test

- 어느 고등학교에서 사교육의 효과여부를 알아보기 위해 사교육을 받고 있는 학생들 중에서 50명의 수학점수를 알아보았더니 평균 85점, 표준편차 5점이었고, 사교육을 받고 있지 않은 학생들 중에서 64명을 뽑아 수학점수를 알아보았더니 평균 80점, 표준편차 8점이었다. 사교육을 받고 있는 학생들의 수학점수가 더 좋다는 가설을 $\alpha = 0.05$ 의 유의수준에서 검정하라. (두 모집단의 표준편차는 같다고 가정한다)

4.1.4. 짹을 이룬 표본에 대한 검정

- 서로 다른 두 개의 모집단에서 뽑는 것이 아닌, 하나의 모집단으로부터 표본을 뽑은 후 그 표본으로부터 쌍으로 된 관찰값들 (paired sample)을 뽑아 이들 간 차이에 대한 가설검정

예 (4) - t-test

- 어느 회사에서 직업훈련이 근로자의 능률향상에 효과가 있는지를 알고 싶다고 하자. 이를 위해 16명의 근로자를 뽑아서 직업훈련을 하기 전과 후의 작업능률의 점수를 알아보았더니 다음 표와 같았다. 이 조사결과로써 훈련전과 훈련후의 능률이 같다고 할 수 있을까? 모집단에서의 차이의 분포는 정규분포라 가정한다.

근로자	훈련후 (X_1)	훈련전 (X_2)	차이 ($D_i = X_1 - X_2$)
A	80	75	5
B	90	83	7
C	92	96	-4
D	75	77	-2
E	86	81	5
F	90	90	0
G	81	82	-1
H	70	67	3
I	89	94	-5
J	88	85	3
K	82	78	4
L	79	82	-3
M	91	96	-5
N	90	80	10
O	78	87	-9
P	89	81	8
합계	-	-	16

4.1.5. 두 모집단에 대한 검정

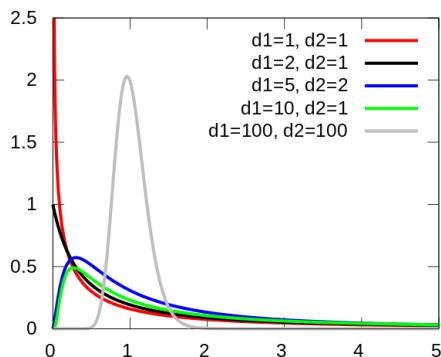
- 분산에 대하여

예 (5) - F-test

- 어느 음료수공장에서는 오란C와 오란D를 생산하고 있다. 오란C공장은 오란D공장보다 품질관리를 더 잘하고 있다고 주장하고 있는데, 이를 검정하기 위하여 오란C 10병과 오란D 10병을 추출하여 용량의 분산을 조사하였더니 각각 18, 20이었다. 두 공장에서 생산되는 음료수의 용량은 정규분포를 이룬다고 가정할 때, 과연 오란C공장의 주장이 옳은지를 $\alpha = 0.05$ 에서 검정하라.

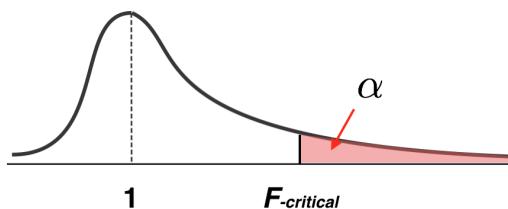
F-분포 (F-distribution)

$$\text{통계량} : F(n_1 - 1, n_2 - 1) = \frac{\chi^2_1 / (n_1 - 1)}{\chi^2_2 / (n_2 - 1)} = \frac{S_1^2}{S_2^2}$$



F-검정 (F-test)

- F 값은 두 분산의 비율로써 계산 되기 때문에, S_1^2 과 S_2^2 이 비슷하면 $F = 1$ 에 가까워진다.
- 그러나 두 표본의 분산으로부터 계산된 F 값이 F -분포표의 임계값보다 매우 크다면(or 계산된 p-value가 유의수준보다 작다면), 이 표본들은 분산 σ^2 이 서로 다른 모집단에서 뽑혔다고 할 수 있다.



4.1.6 결론

- 여러 상황에 따른 검정 방법을 알아 보았으나 결국 검정의 목표는 하나이다.
- 이러한 현상(표본을 통해 들여다 본 평균 혹은 분산 등의 특성)이 일반적인 현상인지 우연에 의한 현상인지를 검정을 통해 알아내는 것이다.

A/B test 실습

- "DMS실습(3)-A/B테스트.xlsx"
- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

4.2. 분산분석 (ANOVA test)

- 들어가며
- 일원분산분석과 이원분산분석
- 분산분석의 예
- 일원분산분석
- 실습

4.2.1. 들어가며

앞에서 우리는 두 모집단의 평균을 비교하기 위해 Z – 검정과 t – 검정을 사용하였다. 그러나 일상생활에서나 학문적인 연구에서 여러 모집단의 평균을 동시에 비교해야 할 경우가 많이 있다. 예를 들면 회장품의 판매촉진을 위하여 광고매체인 신문, 라디오, 텔레비전을 이용할 때 각각의 광고효과가 차이가 있는가를 알아보고자 하는 경우 등이다. 세 매체 간 광고효과의 차이를 비교하기 위하여 t – 검정이나 Z – 검정을 한다고 하면 $\binom{3}{2} = 3$ 번의 검정을 해야 한다. 만일 "이메일 광고" 방법까지 추가한다면 $\binom{4}{2} = 6$ 번의 비교를 해야 할 것이다.

✓ 이렇듯 여러 모집단 평균을 동시에 비교하는 데 사용되는 통계적 연구방법이 분산분석 (analysis of variance) 이며 간단히 ANOVA 라고도 한다.

4.2.2. 일원분산분석과 이원분산분석

- 일원분산분석 (one-way analysis of variance)
 - 독립변수가 하나일 때
 - ex. "광고매체"라는 하나의 독립변수를 여러개의 수준(신문, 라디오, 텔레비전, 이메일광고)으로 나누어 광고매체들 간의 광고효과 차이가 있는가를 알아보는 경우
- 이원분산분석 (two-way analysis of variance)
 - 독립변수가 두 개일 때
 - ex. 만일 "광고매체"뿐만 아니라 "소비자의 나이"도 광고 효과에 어떤 영향을 주는지를 알아보는 경우
- 다원분산분석
 - 독립변수가 3개 이상일 때
 - 다원분산분석은 계산만 복잡할 뿐 기본개념은 이원분산분석과 같음

4.2.3. 분산분석의 예

- 어느 회사에서는 세 개의 서로 다른 기계를 사용하여 제품을 생산하고 있는데, 각각의 기계가 1시간에 생산하는 제품의 양을 다섯 차례 관찰하여 적은 결과가 아래의 표에 나타나 있다.

기계	각 기계의 생산량							\bar{X}_i
1	47	53	49	50	46	49		
2	55	54	58	61	52	56		
3	54	50	51	51	49	51		

- 표를 보면 기계 1을 1시간씩 다섯 번 조사한 결과 시간당 생산량의 평균은 49다. 그리고 같은 방법으로 측정한 기계 2의 평균은 56, 기계 3의 평균은 51이다. 이때 다섯 번의 표본 생산량에 기초하여 세 기계의 평균 생산량은 동일하다고 볼 수 있는가? 이와 같은 문제에 대한 답을 제시하는 것이 분산분석이다

- 이 예의 귀무가설은 "세 기계의 평균 생산량이 모두 동일하다"이며, 대립가설은 "평균 생산량이 모두 동일하지는 않다"가 된다.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{모든 평균이 동일하지는 않다.}$$

(즉, 평균이 서로 다른 기계가 있다.)

- 귀무가설이 기각된다면 표본생산량들이 뽑혀 나온 각 모집단의 평균이 모두 같지는 않음을 말한다. 그러나 세 집단 중에서 어느 집단이 서로 다른지는 알 수 없다. 다시 말하면 대립가설은 " $\mu_1 \neq \mu_2 \neq \mu_3$ " 가 아니다.

이해를 돋기 위한 새로운 예

- 분산분석은 위의 가설을 검정하기 위해 ①생산량의 변동 또는 분산을 요인의 수준차이에 기인한 부분과 ②우연 또는 오차에 의한 부분으로 분해한 다음, 전자가 후자보다 충분히 클 때 요인의 수준에 따라 집단 간 차이가 있는 것으로 판단한다. 이를 이해하기 위해 이번에는 위의 표와 약간 다른 경우를 생각해 보자.

기계	각 기계의 생산량							\bar{X}_i
1	57	32	53	38	65	49		
2	36	49	64	71	60	56		
3	57	69	48	36	45	51		

기계	각 기계의 생산량							\bar{X}_i
1	48	49	49	49	50	49		
2	56	55	56	57	56	56		
3	50	51	51	52	51	51		

- 두 표를 비교해 보면, 세 기계에서 만들어진 생산량의 평균은 같지만 1시간마다 조사한 개별 생산량을 다르게 나타내고 있다. 표A는 1시간마다의 생산량에 차이가 많다. 그러나 표B는 매 시간마다의 생산량이 상당히 고르게 나타나 있다.
- 평균 생산량은 49, 56, 51로 동일하더라도 만일 조사결과가 표B와 같이 나타났다면 세 기계의 차이는 분명히 존재한다고 볼 수 있으며, 따라서 귀무가설은 기각될 것이 분명하다. 왜냐하면 표B에서 나타난 세 기계의 평균 생산량 \bar{X}_i 들의 차이는 우연이라고 볼 수가 없기 때문이다.

4.2.4. 일원분산분석 - 자료의 구성

- 분산분석을 하기 위해 계산을 하려면 자료가 어떻게 구성되어 있고 각 자료가 어떻게 표시되어 있는지를 알아야 한다.

관찰번호	집단1	집단2	집단j	
1	X_{11}	X_{12}	...	X_{1j}
2	X_{21}	X_{22}	...	X_{2j}
3	X_{31}	X_{32}	...	X_{3j}
:	:	:	:	:
i	X_{i1}	X_{i2}	...	X_{ij}
	\bar{X}_1	\bar{X}_2	...	\bar{X}_j
				\bar{X}

< table. 일원분산분석의 자료구성 >

\bar{X} : 전체평균

\bar{X}_j : j번째집단의평균

X_{ij} : j번째집단의 i번째관찰값

4.2.4. 일원분산분석 - 관찰값의 모형

- 관찰값을 X_{ij} 라 하면 X_{ij} 는 다음과 같은 요소로 구성

$$X_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

μ : 전체평균

α_j : j번째집단의영향

ϵ_{ij} : j번째집단에있는관찰값 i의우연적오차

4.2.4. 일원분산분석 - 통계값으로 표현

- 각 관찰값은 전체평균 μ 와, 수준이 다른 집단에 있기 때문에 생기는 전체평균과의 차이 α_j , 그리고 각 집단에 있는 관찰값 i의 개인차 또는 오차 ϵ_{ij} 로 이루어져 있음을 알 수 있다. 위 식은 하나의 모형이며 이를 실제연구에서 얻을 수 있는 통계값으로 표현하면 다음과 같다.

$$X_{ij} = \bar{X} + (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

$(\bar{X}_j - \bar{X})$: j번째집단의평균과전체평균간의차이

$(X_{ij} - \bar{X}_j)$: 각관찰값과각집단평균간의차이

4.2.4. 일원분산분석 - 통계값의 변형

- 위 식에서 \bar{X} 를 왼쪽 항으로 옮겨 다음과 같이 변형시킬 수 있다.

$$(X_{ij} - \bar{X}) = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

- 위 식에서 왼쪽 식과 오른쪽 식을 각각 제곱하여 전체관찰수만큼 합하면 다음과 같다.

$$\sum \sum (X_{ij} - \bar{X})^2 = \sum \sum (\bar{X}_j - \bar{X})^2 + \sum \sum (X_{ij} - \bar{X}_j)^2$$

4.2.4. 일원분산분석 - 제곱합

$$\sum \sum (X_{ij} - \bar{X})^2 = \sum \sum (\bar{X}_j - \bar{X})^2 + \sum \sum (X_{ij} - \bar{X}_j)^2$$

$$SST = SSB + SSW$$

- 제곱합

- SST : 총제곱합(total sum of squares: SST)
- SSB : 집단간 제곱합(sum of squares between groups: SSB)
- SSW : 집단내 제곱합(sum of squares within groups: SSW)

4.2.4. 일원분산분석 - 분산분석표

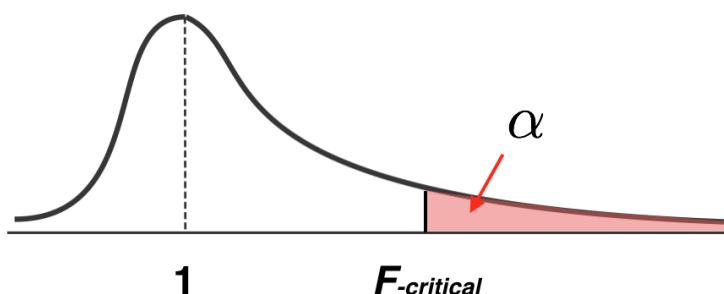
- 분산분석표 (ANOVA table)

분산원	제곱합	자유도	평균제곱	F값
집단간	$SSB = \sum n_i(\bar{X}_j - \bar{X})^2$	$J-1$	$MSB = \frac{SSB}{J-1}$	$\frac{MSB}{MSW}$
집단내	$SSW = \sum \sum (X_{ij} - \bar{X}_j)^2$	$N-J$	$MSW = \frac{SSW}{N-J}$	
합계	$SST = \sum \sum (X_{ij} - \bar{X})^2$	$N-1$		

4.2.4. 일원분산분석 - F 검정

- F-test

- F 값은 두 분산의 비율로써 계산 되기 때문에, S_1^2 과 S_2^2 이 비슷하면 $F = 1$ 에 가까워진다.
- 그러나 두 표본의 분산으로부터 계산된 F값이 F - 분포표의 임계값보다 매우 크다면, 이 표본들은 분산 σ^2 이 서로 다른 모집단에서 뽑혔다고 할 수 있다.



ANOVA test 실습

- "DMS실습(4)-ANOVA분석.xlsx"
- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

4.3. 상관분석과 회귀분석

- 들어가며
- 산포도 (Scatter Plot)
- 상관분석
- 회귀분석

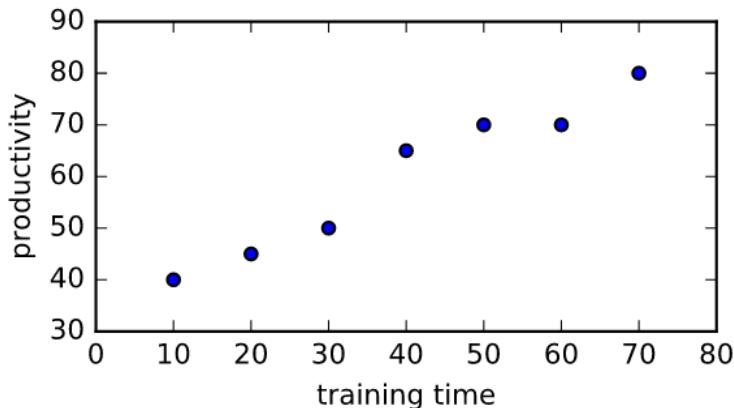
4.3.1. 들어가며

- **상관분석**은 두 변수 간의 관계의 강도, 즉 얼마나 밀접하게 관련되어 있는지를 분석하는 것을 말한다.
- 그러나 때로는 관련성뿐만 아니라 독립변수에 따라 종속변수가 어떻게 변화하는지를 예측하기 원할 때가 있다.
- **회귀분석**에서는 독립변수의 일정한 값에 대응되는 종속변수의 값을 예측하기 위하여 회귀방정식을 구한다.
- 상관분석에 의해 상관계수를 계산하고, 만일 상관계수가 높다면 두 변수 간의 관계를 회귀방정식으로 나타낸다. 따라서 회귀분석을 상관분석과 함께 사용한다면 변수들 간의 관련성에 대한 다양한 정보를 구할 수 있다.

4.3.2. 산포도 (Scatter Plot)

- 두 변수 간 관련성을 분석할 때는 먼저 산포도를 그려보는 것이 좋다.
 - 산포도를 통해 그 자료가 상관분석이나 회귀분석을 할 만한 자료인지 아닌지를 알 수 있다.
 - 만일 분석할 만한 자료가 되지 못한다는 것을 알게 되면 시간과 노력의 낭비를 줄여줌

ex. 어느 조립공장에서는 기능공들의 기술훈련을 자체적으로 실시하고 있는데, 훈련시간이 늘어남에 따라 숙련도가 얼마나 향상되는지를 알아보려고 한다. 훈련시간별로 임의로 1명씩을 표본으로 뽑아 그들의 하루 생산성을 조사하여 다음과 같은 결과를 얻었다.



- scatter plot 을 보면 훈련시간이 늘어남에 따라 생산성도 높아지고 있음을 쉽게 볼 수 있다. 따라서 훈련시간과 생산성 사이에 밀접한 관계가 있음을 알 수 있다.
- 이처럼 scatter plot 만으로도 두 변수 간의 관계를 대략적으로 파악할 수 있다.
- 그러나 두 변수의 관계를 정확히 파악하기 위해서는 두 변수 간의 관련성의 정도를 계수(correlation coefficient)로 알아보는 상관분석,
- 두 변수 간의 함수적 관련성을 나타내는 회귀식(regression equation) 또는 예측식을 구하는 회귀분석을 해야 한다.

4.3.3. 상관분석 - 공분산

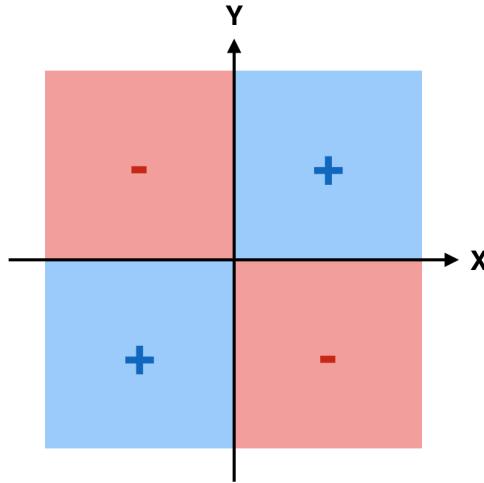
- 상관분석을 하기 위해서는 먼저 공분산(covariance)에 대해 알아야 할 필요가 있다.
 - 상관분석이란 두 변수가 어떻게 함께 움직이는가를 알아 보는 것
 - 공분산 역시 두 변수가 동시에 변하는 정도를 나타냄

$$\text{모집단} \quad \sigma_{XY}^2 = \frac{\sum(X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

$$\text{표본} \quad S_{XY}^2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

4.3.3. 상관분석 - 공분산의 한계

- 공분산에서 X 변수가 증가할 때 Y 변수가 증가하면, 즉 두 변수가 같은 방향으로 변화하면 공분산의 수치는 + 가 된다.
- 만일 두 변수가 변화하는 방향이 서로 다르다면 공분산은 - 의 부호를 가진다.



- 이렇듯 공분산은 두 변수 간의 관계를 말해주지만, 두 변수의 측정단위에 따라서 커다란 차이가 나는 문제점이 있어 상대적인 강도를 나타내는 좋은 지표가 되지 못한다.
 - 예를 들어 변수를 센티미터(cm)로 표시되는 경우보다 미터(m)로 표시되는 경우 공분산의 절대값은 훨씬 작아진다.

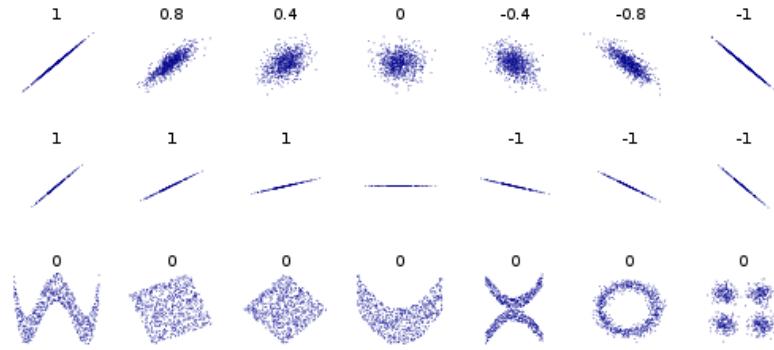
4.3.3. 상관분석 - 상관계수 (r : correlation coefficient)

- 피어슨(K. Pearson, 1857~1936)이 제시한 상관계수 r 은 정규분포를 따르는 두 변수 X 와 Y 가 일직선이라는 선형성을 가정한다.
- r_{XY} 는 두 변수의 공분산 (S_{XY}^2) 을 각 변수의 표준편차인 S_X 와 S_Y 의 곱으로 나누어 상관계수를 구한다.
 - 이로인해 상관계수는 $-1.0 \leq r \leq 1.0$ 의 범위에 있게 된다.
- 어떠한 단위의 측정값을 사용하여도 상관성에 대한 비교와 해석이 용이

$$\text{모집단} \quad \rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$$

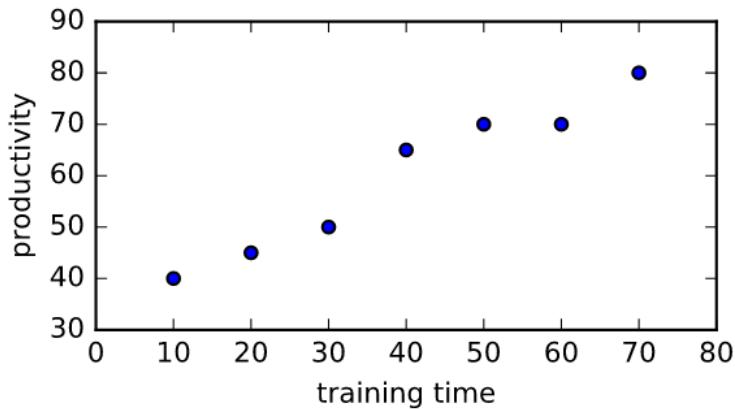
$$\text{표본} \quad r_{XY} = \frac{S_{XY}^2}{S_X S_Y}$$

- $r = +1$: 완전한 선형관계 (완전한 선형관계를 가질 때 공분산 S_{XY} 의 값은 S_X 와 S_Y 를 곱한 값과 같으므로)
- $r = -1$: 완전한 음의 선형관계
- 분포가 원모양, 곡선의 모양 등의 경우 상관계수가 낮거나 관계가 없는 것처럼 보인다.
 - 이는 피어슨 상관계수는 두 변수 간의 관계가 선형적이라는 가정하에 전개된 계수이기 때문

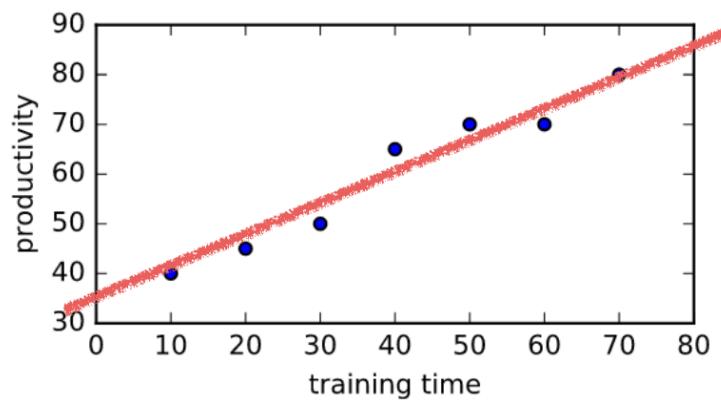


4.3.4. 회귀분석 - 개념

- 위에서 보았던 조립공장의 생산성 예를 다시 보자
- 아래의 산포도를 표현할 수 있는 직선을 찾으면 어떻게 될까?



직선 = 함수 = 인과관계



- 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)

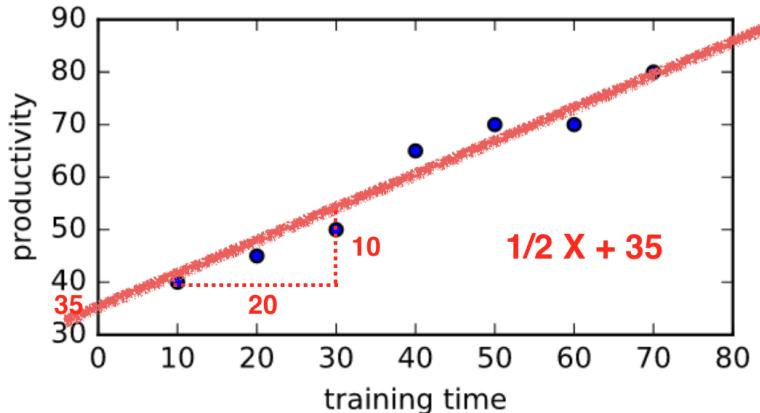
4.3.4. 회귀분석 - 회귀모형과 회귀식

- 독립변수와 종속변수 간의 1차함수관계 또는 선형관계를 가정할 때 회귀모형은 다음 두 요소를 결합한 형태로 나타낼 수 있다.

- 확정적 함수관계를 나타내는 부분 : $a + bX_i$
- 확률적 오차항 : e_i

$$\text{단순회귀모형} \quad Y_i = a + bX_i + e_i$$

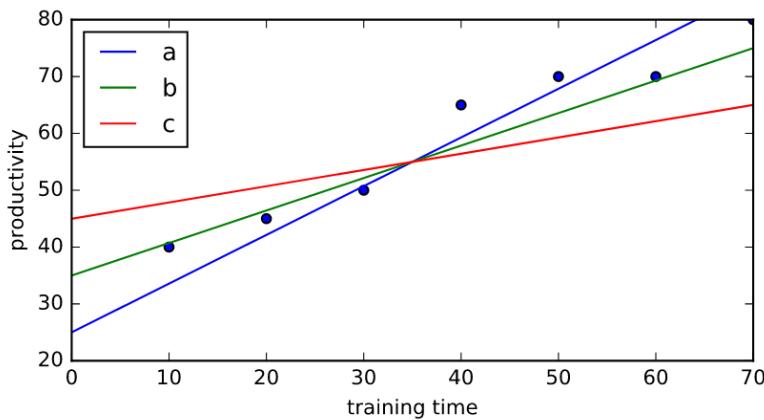
$$\text{단순회귀식} \quad \hat{Y}_i = a + bX_i$$



- a, b : 회귀계수 (regression coefficient)
 - 회귀식을 보면 절편을 a 로 하고 기울기가 b 인 직선이 됨을 알 수 있다.
- \hat{Y}_i : 회귀식을 통해 구해지는 수치 (예측값)
- e_i : 예측오차, 추정오차 또는 잔차(residual)

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

그렇다면 아래 3가지 직선 중에 어떤 직선이 가장 적절한 회귀직선일까?



4.3.4. 회귀분석 - 최소제곱법

- 최소제곱법 (method of least squares) : 잔차의 제곱합이 최소가 되도록 모수를 추정하는 방법
- 회귀식을 결정하는 가장 좋은 방법
- 다른 방법에 의해 구한 회귀식보다 통계학적으로 그 성질이 우수한 α, β 의 추정값을 얻을 수 있다.
 - 통계학적으로 우수하다는 것은 표본에서 통계값을 구할 때 그 통계값이 모집단의 모수를 가장 잘 설명하는 추정값이라는 것을 의미

$$\min \sum e_i^2 = \min \sum (Y_i - \hat{Y}_i)^2$$

4.3.4. 회귀분석 - 최소제곱법에 의한 추정량 (ordinary least square estimator)

- \hat{Y}_i 은 회귀모형의 종속변수에 대한 추정값으로서 다음과 같다.

$$\hat{Y}_i = a + bX_i$$

- 최소제곱법 식에 대입하면 다음과 같다.

$$\min \sum e_i^2 = \min \sum (Y_i - a - bX_i)^2$$

- 잔차제곱합 ($\sum e_i^2$) 을 최소로 하는 a, b 를 찾기 위해 a, b 에 대해 각각 편미분 하면 다음과 같다.

$$\begin{aligned}\sum Y_i &= na + b \sum X_i \\ \sum X_i Y_i &= a \sum X_i + b \sum X_i^2\end{aligned}$$

- 위 두 식을 충족시키는 표본의 회귀계수 a 와 b 는 다음과 같다.

$$\begin{aligned}b &= \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2} \\ a &= \bar{Y} - b \bar{X}\end{aligned}$$

Scatter Plot과 상관분석 그리고 회귀분석 실습

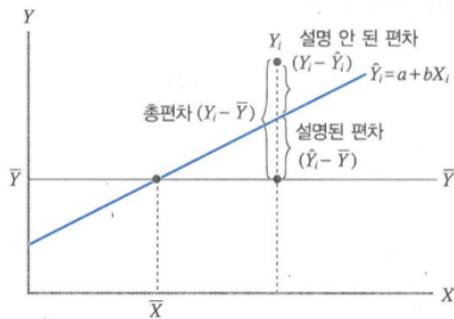
- "DMS실습(5)-상관분석과 회귀분석.xlsx" - "조립공장예제 Sheet"
- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

Q1 - 그렇다면 도출한 회귀식이 몇점짜리 분석인지 어떻게 알까?

- 결정계수 : r^2

4.3.4. 회귀분석 - 편차와 제곱합

그림 15-2 세 편차의 관계



- 총편차 (total deviation) = 설명 안 된 편차 (unexplained deviation) + 설명된 편차 (explained deviation)

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

- 총제곱합 (total sum of squares) = 오차제곱합 (sum of squares error) + 회귀제곱합 (sum of squares regression)

$$\begin{aligned} SST &= SSE + SSR \\ \sum(Y_i - \bar{Y})^2 &= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

4.3.4. 회귀분석 - 결정계수 (r^2)

- 표본회귀식의 적합도를 어느 경우에나 일률적으로 나타내줄 수 있는 방법으로서 가장 많이 사용됨
- 결정계수는 표본회귀식에 의하여 설명된 제곱합이 총제곱합에서 차지하는 상대적 크기를 나타내는 것으로서, r^2 으로 표시
- $r^2 = 1$: 모든 관찰자료가 표본회귀식으로 완전히 설명되는 경우 ($\because SSE = \sum e^2 = 0$)
- $r^2 = 0$: 종속변수가 독립변수의 1차식으로는 전혀 설명이 되지 못함을 의미

$$\begin{aligned} \text{결정계수} &= \frac{\text{회귀제곱합}}{\text{총제곱합}} = 1 - \frac{\text{오차제곱합}}{\text{총제곱합}} \\ r^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \end{aligned}$$

Q2 - 도출한 회귀식이 적절한지 어떻게 판단할까?

- Test (검정)

4.3.4. 회귀분석 - 단순회귀모형에 대한 F – 검정

- 회귀식이 표본자료를 잘 설명하고 있다면 설명된 제곱합 SSR 은 설명 안 된 제곱합 SSE 에 비해 상대적으로 클 것 !

$$\begin{array}{ll} \text{제곱합} & SST = SSE + SSR \\ \text{자유도} & n - 1 = (n - 2) + (1) \end{array}$$

- 단순회귀분석에서 F – 검정식

$$F_{1,n-2} = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

4.3.4. 회귀분석 - β 에 대한 t – 검정

- 회귀선의 기울기를 나타내는 β 가 0 라는 것은 (= 통계적으로 유의하지 않다는 것은) 회귀모형이 성립되지 않는다는 것을 의미 !

- 절편인 α 는 상수이므로 X 의 변화에 따라 Y 가 어떻게 변화하는지에 대해 알려주는 것이 없다.
- 모수 β 에 대한 가설검정은 표본회귀식에서 구한 기울기 b 를 이용해서 모집단 회귀식의 β 에 대한 가설 검정을 하는 것

- 귀무가설

$$\begin{array}{l} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{array}$$

- $\beta = 0$ 을 검정하는 t – 통계량

$$t = \frac{b - 0}{S_b} = \frac{b}{S_b}$$

Scatter Plot과 상관분석 그리고 회귀분석 실습 (1)

- "DMS실습(5)-상관분석과 회귀분석.xlsx" - "조립공장예제 Sheet"
- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

Scatter Plot과 상관분석 그리고 회귀분석 실습 (2)

- "DMS실습(5)-상관분석과 회귀분석.xlsx" - "Analysis Sheet"
- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

수업 끝

- 디지털 마케팅 SCHOOL 5기 (17. 4. 24)