Chapter 5. 가설검정

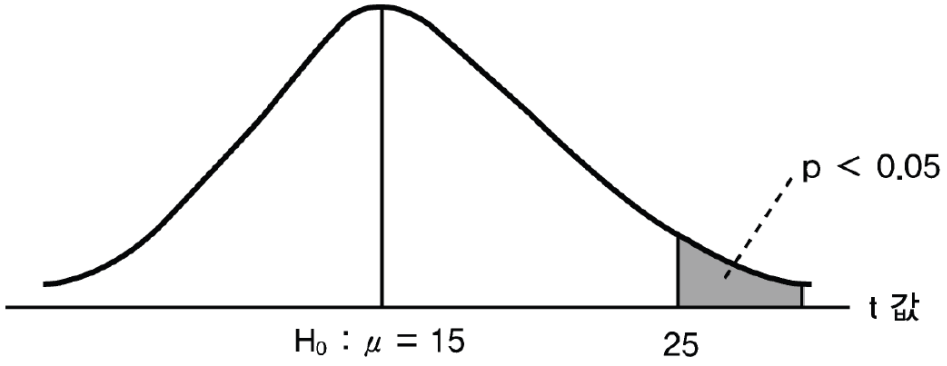
1. 가설검정

1.1. 가설이란

- 가설
  - 표본으로부터 주어지는 정보를 이용하여, 모수에 대한 예상, 주장 또는 단순한 추측을 기술하는 것
  - 대립가설(alternative hypothesis: H_1): 데이터로부터 얻은 강력한 증거에 의해 연구자가 입증하고자 하는 가설
  - 귀무(영)가설(null hypothesis: H_0): 대립가설에 상반되는 가설로서 분석 이전에 자연현상에 가까운 사실에 대한 가설

1.2. 가설검정

- 가설검정
  - 설정된 가설 중에 어느 것이 맞는 지를 검정
  - 검정통계량을 활용
  - 분포: 정규분포, *t*-분포, χ^2-분포, *F*-분포
- 기각역
  - 귀무가설을 기각하게 되는 검정통계량의 관측값 영역
- 유의수준
  - 5%, 1%를 주로 사용
- 오류
  - α와 β 오류



검정결과 \ 실제현상	귀무가설이 사실	대립가설이 사실
귀무가설을 채택	옳은 결정	β
귀무가설을 기각	α	옳은 결정

1.3. 가설검정의 종류

표본의 개수	검정 대상	모분산 파악여부	분석 구분
1개	평균	알고 있음	한 표본에서 평균에 대한 Z-검정
		모름	한 표본에서 평균에 대한 t-검정
	비율	관계없음	한 표본에서 비율에 대한 비율검정
	분산	관계없음	한 표본에서 분산에 대한 모분산검정
2개	평균	관계없음 독립된 표본	두 표본에서 평균에 대한 t-검정
		관계없음 쌍체 표본	두 표본에서 평균에 대한 쌍체 t-검정
	비율	관계없음	두 표본에서 비율에 대한 비율검정
	분산	관계없음	두 표본에서 분산에 대한 모분산검정

2. 한 표본에 대한 가설검정

2.1. Z-검정

(1) 분석개요

- 모집단에 대한 분산을 알고 있는 경우 가설 검정 방법
- 음료수 병의 함량에 대한 조사
- 분산은 1ml로 알려져 있음
  - H_0 : 음료수병 함량은 $350ml$이다$(\mu_1 = \mu_0)$.
  - H_1 : 음료수병 함량은 $350ml$이 아니다$(\mu_1 \neq \mu_0)$.

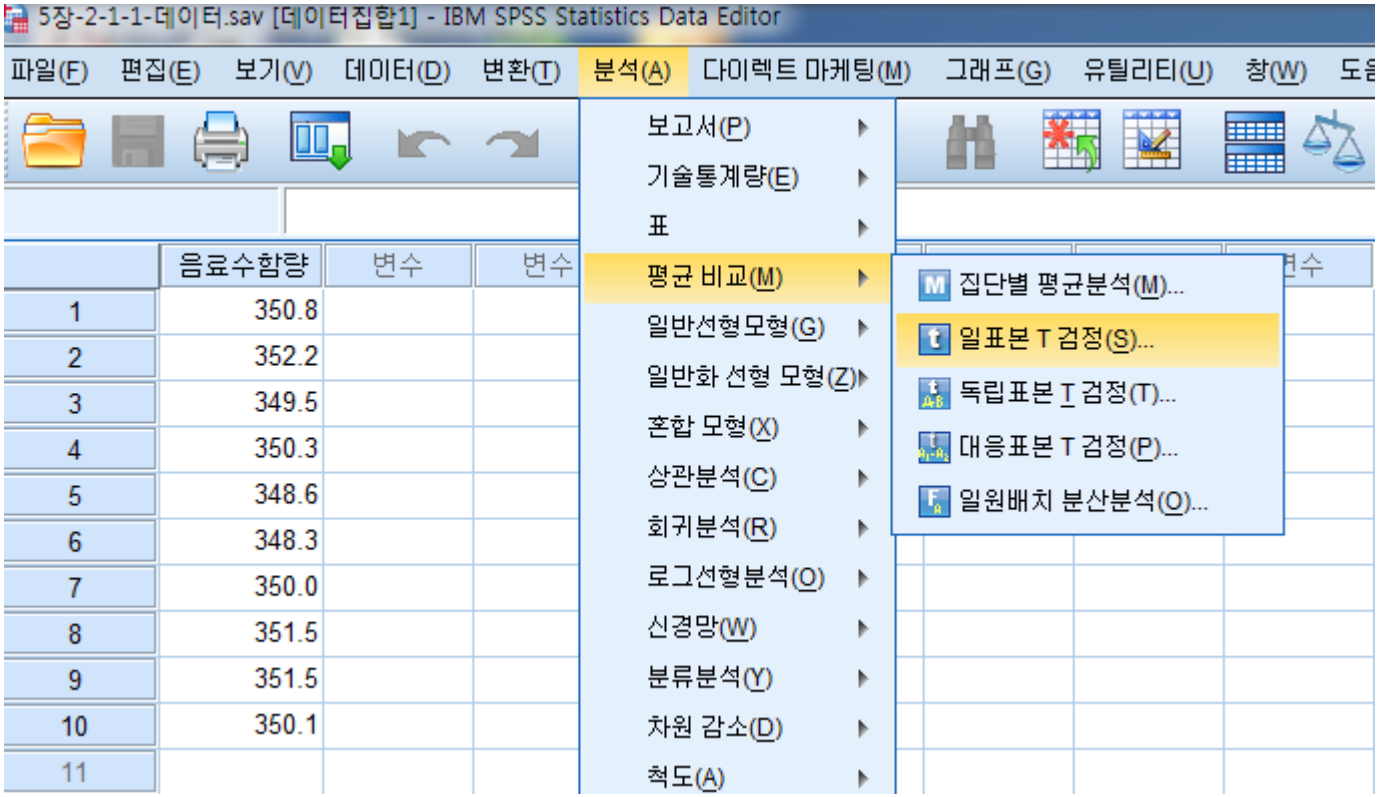
(2) 분석데이터

- 10개의 자료를 수집함

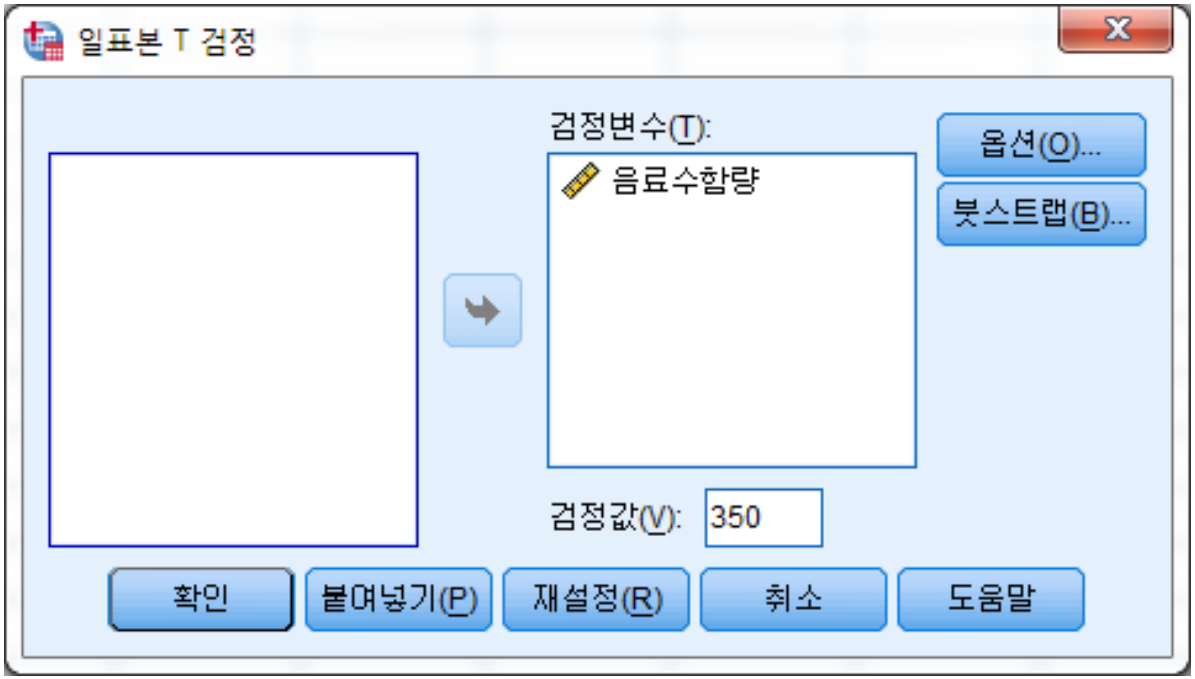
(3) 분석과정

– 기술통계 확인

- STEP 01: [평균비교] [일표본 T검정]
- STEP 02: 분석 변수 지정



검정변수 : 음료수함량
검정값 : 350 입력



(4) 결과해석

일표본 통계량

	N	평균	표준편차	평균의 표준오차
음료수함량	10	350.280	1.2647	.3999

일표본 검정

	검정값 = 350					
					차이의 95% 신뢰구간	
	t	자유도	유의확률 (양쪽)	평균차	하한	상한
음료수함량	.700	9	.502	.2800	-.625	1.185

검정 통계량 0.700,
p-value=0.502 로서 유의수준 5% 하에서 귀무가설을 채택함.

2.2. t-검정

(1) 분석개요

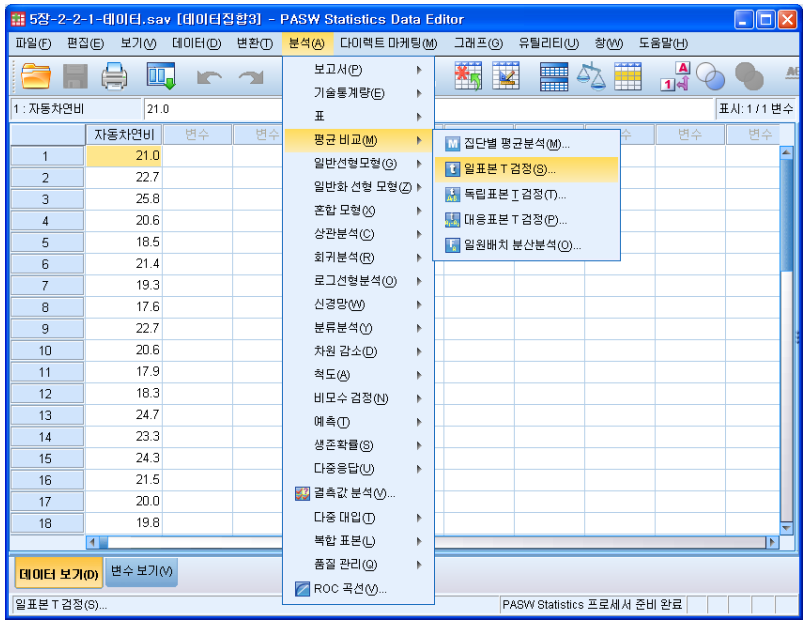
- 모집단에 대한 분산을 모르고 있는 경우 가설 검정 방법
- 자동차 연비에 대한 조사
  - H_0 : 자동차의 연비는 $20km/l$ 미만이다 $(\mu_1 < \mu_0)$.
  - H_1 : 자동차의 연비는 $20km/l$ 이상이다$(\mu_1 \geq \mu_0)$.

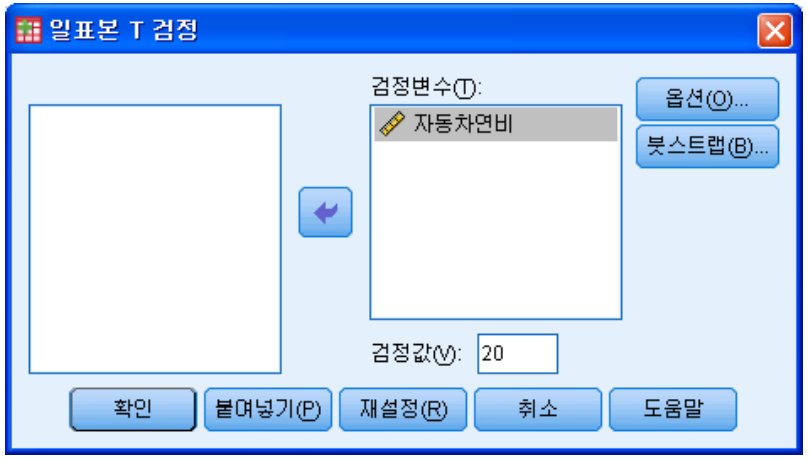
(2) 분석데이터

- 20대 자동차 데이터를 수집함

(3) 분석과정

- STEP 01: [일표본 T-검정] 메뉴 클릭
- STEP 02: 검정변수 및 검정값 지정





(4) 결과해석

일표본 통계량

	N	평균	표준편차	평균의 표준오차
자동차연비	20	21.140	2.3383	.5229

일표본 검정

	검정값=20					
					차이의 95% 신뢰구간	
	t	자유도	유의확률 (양쪽)	평균차	하한	상한
자동차연비	2.180	19	.042	1.1400	.046	2.234

2.3. 비율검정

(1) 분석개요

- 한 표본에 대한 비율검정은 성공, 실패 또는 불량률과 같이 비율에 대한 검정
- 대학 졸업자의 전공분야 취직에 대한 조사
  - H_0 : 대학졸업자 중 자신의 전공을 살릴 수 있는 직장에 입사하는 비율이 20%이다 $(p_1 = p_0)$.
  - H_1 : 대학졸업자 중 자신의 전공을 살릴 수 있는 직장에 입사하는 비율이 20%가 아니다 $(P_1 \neq p_0)$.

(2) 분석데이터

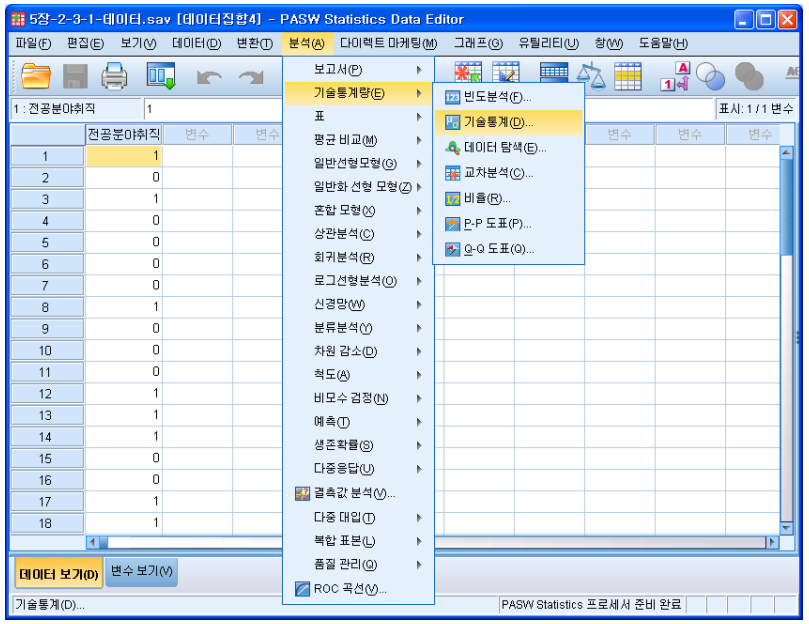
- 30명의 대학졸업자 데이터를 수집함

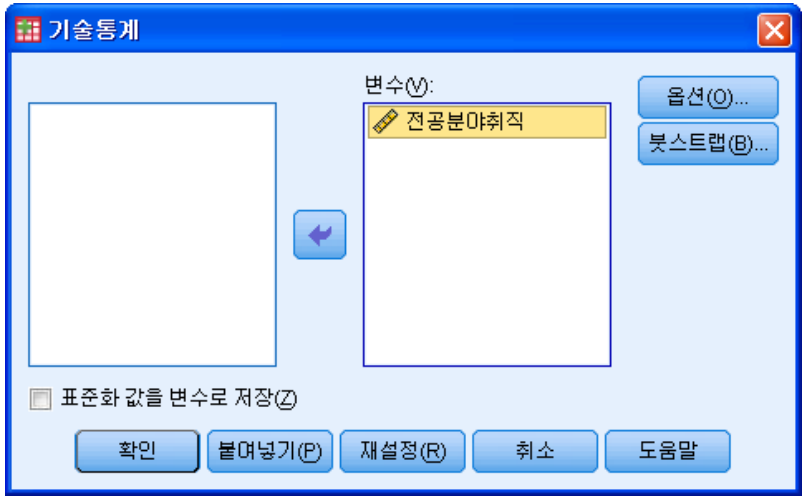
(3) 분석과정

(5장-2-3-1.sav)

– 기술통계 확인

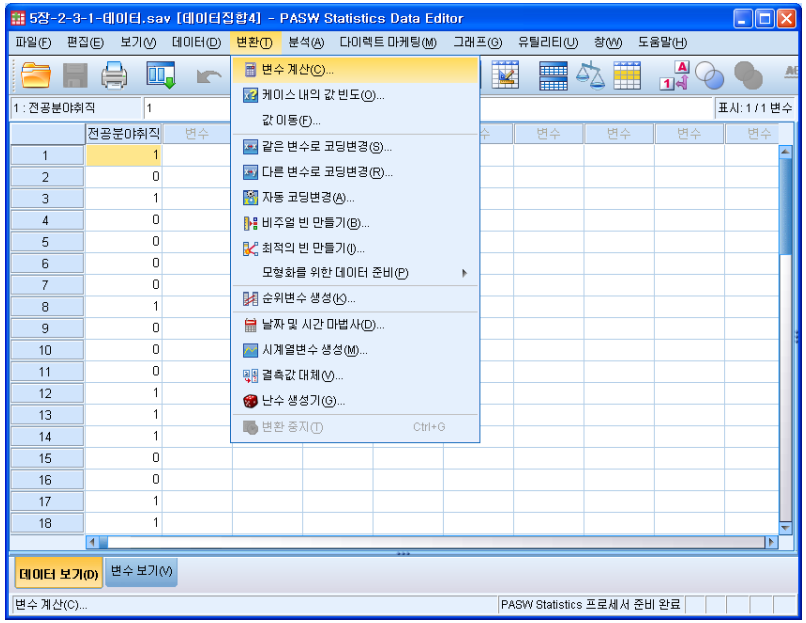
- STEP 01: [기술통계] 메뉴 클릭
- STEP 02: 분석 변수 지정

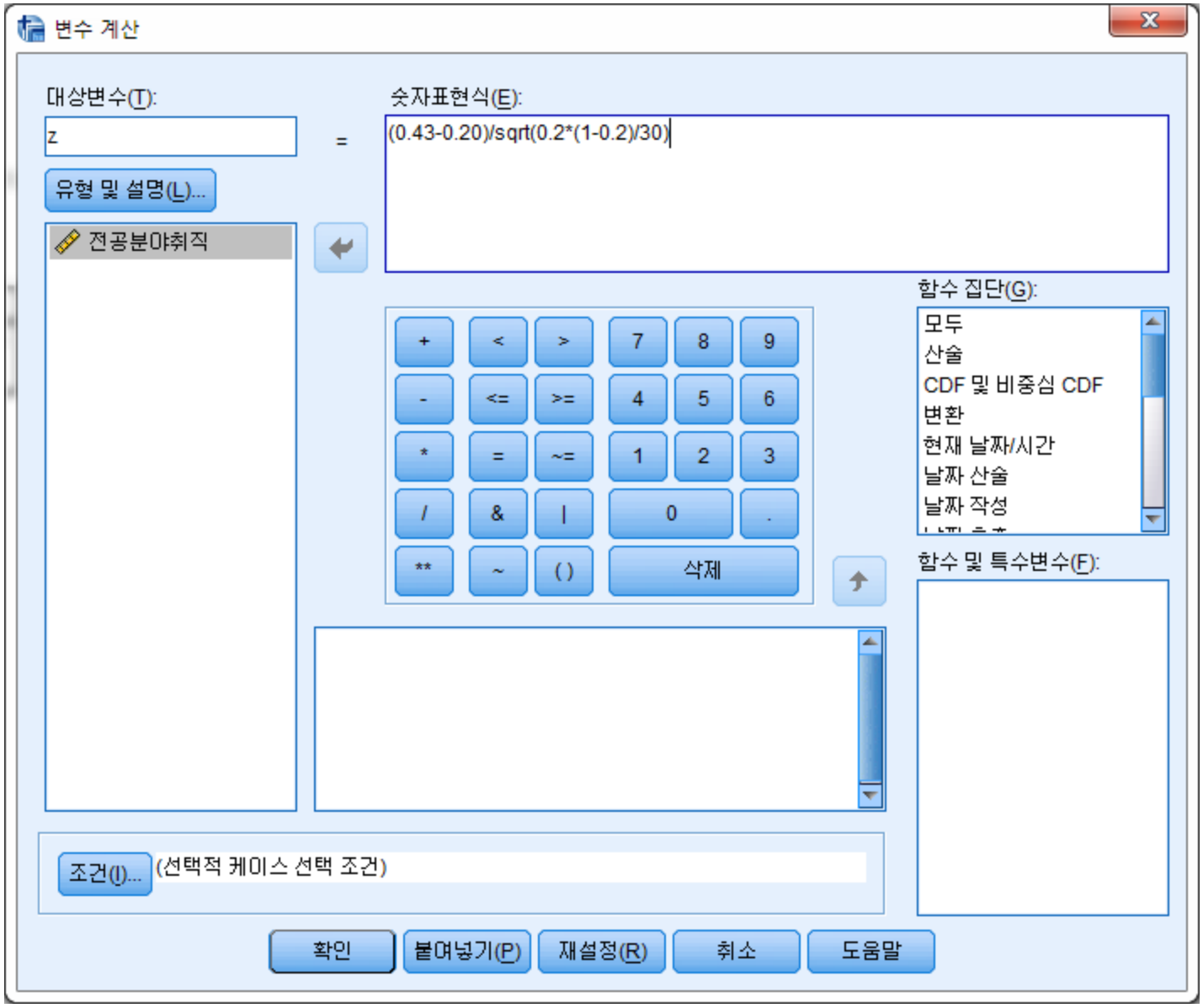




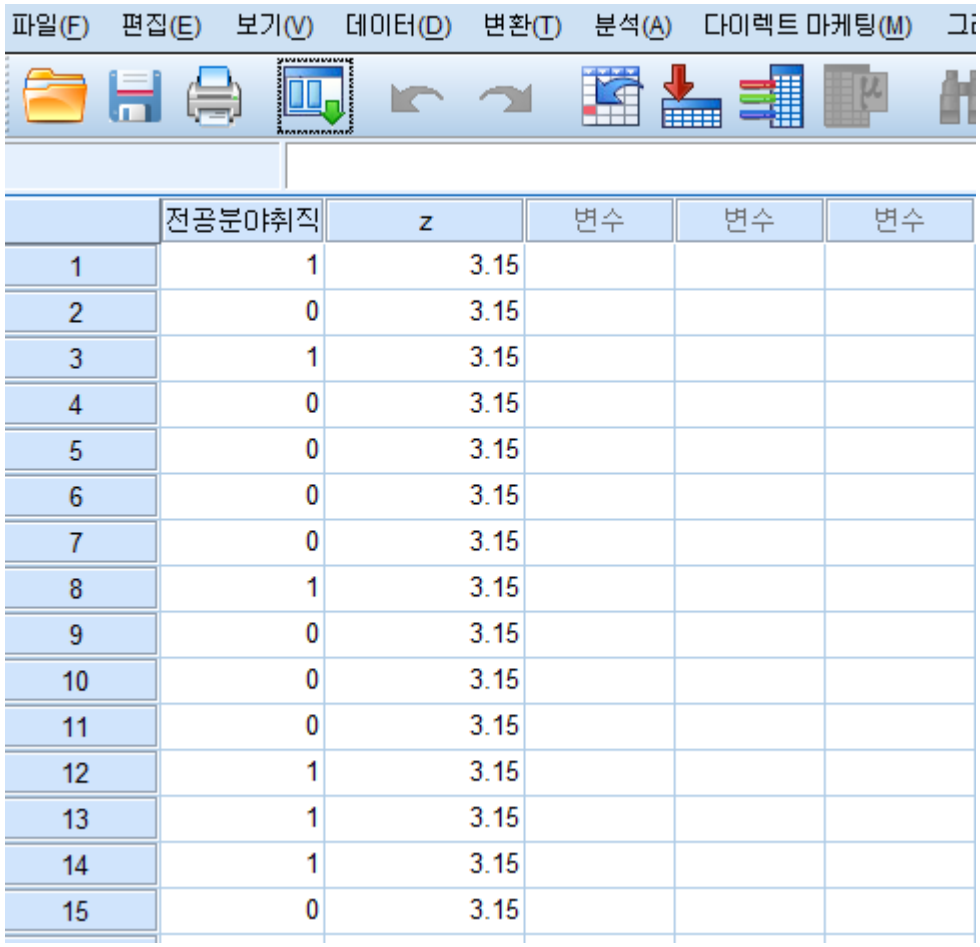
– Z값 계산

- STEP 01: [변수 계산] 메뉴 클릭
- STEP 02: Z값 계산식 입력
- STEP 03: 계산된 Z값 보기




변수 계산
대상변수(T):
z
=
숫자표현식(E):
(0.43-0.20)/sqrt(0.2*(1-0.2)/30)
유형 및 설명(L)...
전공분야취직
함수 집단(G):
모두
산술
CDF 및 비중심 CDF
변환
현재 날짜/시간
날짜 산술
날짜 작성
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) 삭제
함수 및 특수변수(F):
조건(I)...
(선택적 케이스 선택 조건)
확인
붙여넣기(P)
재설정(R)
취소
도움말

Z값 계산



	전공분야취직	z	변수	변수	변수
1	1	3.15			
2	0	3.15			
3	1	3.15			
4	0	3.15			
5	0	3.15			
6	0	3.15			
7	0	3.15			
8	1	3.15			
9	0	3.15			
10	0	3.15			
11	0	3.15			
12	1	3.15			
13	1	3.15			
14	1	3.15			
15	0	3.15			

1) 유의확률 계산(p.216)

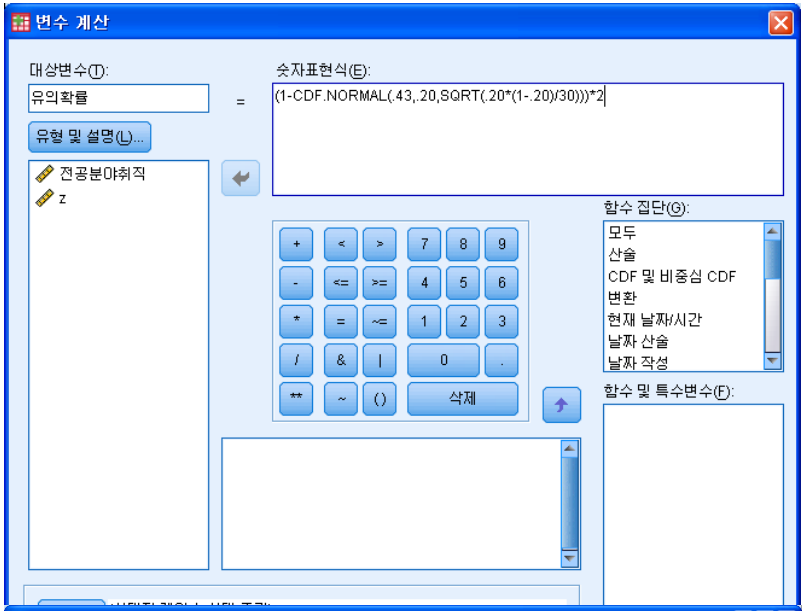
- STEP 04: 다시 [변수계산] 메뉴 클릭
- STEP 05: 계산식 입력
- STEP 06: 유의확률 확인

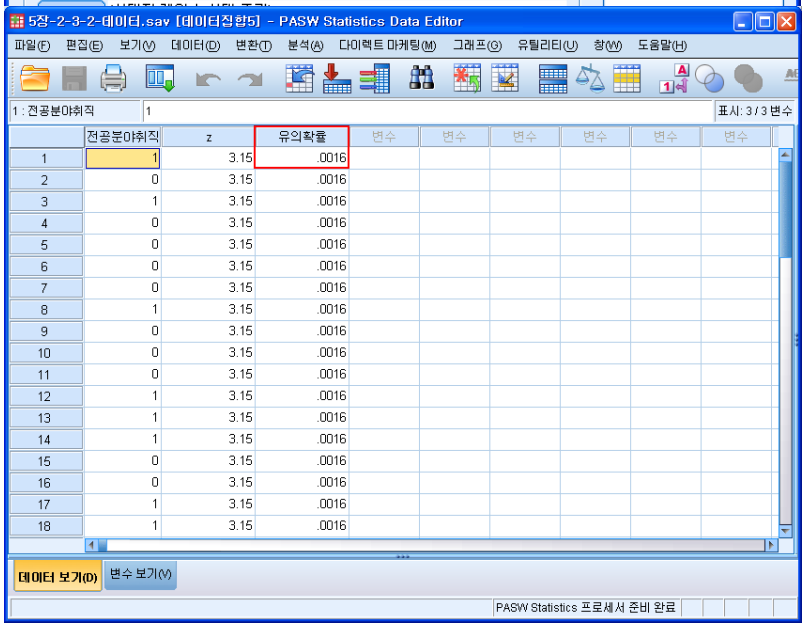
2) α = 0.05일 경우 기각역 설정

*양측검정의 경우

z <-1.96 or z > 1.96이면 유의수준 5%하에서 귀무가설을 기각함.

(4) 결과해석





	전공분야취직	z	유의확률
1	1	3.15	.0016
2	0	3.15	.0016
3	1	3.15	.0016
4	0	3.15	.0016
5	0	3.15	.0016
6	0	3.15	.0016
7	0	3.15	.0016
8	1	3.15	.0016
9	0	3.15	.0016
10	0	3.15	.0016
11	0	3.15	.0016
12	1	3.15	.0016
13	1	3.15	.0016
14	1	3.15	.0016
15	0	3.15	.0016
16	0	3.15	.0016
17	1	3.15	.0016
18	1	3.15	.0016

2.4. 모분산검정

(1) 분석개요

- 한 표본에 대한 분산검정은 표본추출을 통해 나온 표본분산에 대해 기대하고 있는 모분산과 같은지를 검정
- 음료수 함량 조사의 모분산에 대한 조사
  - H_0 : 모분산은 1이다$(\sigma_1^2 = \sigma_0^2)$.
  - H_1 : 모분산은 1이 아니다 $(\sigma_1^2 \neq \sigma_0^2)$.

(2) 분석데이터

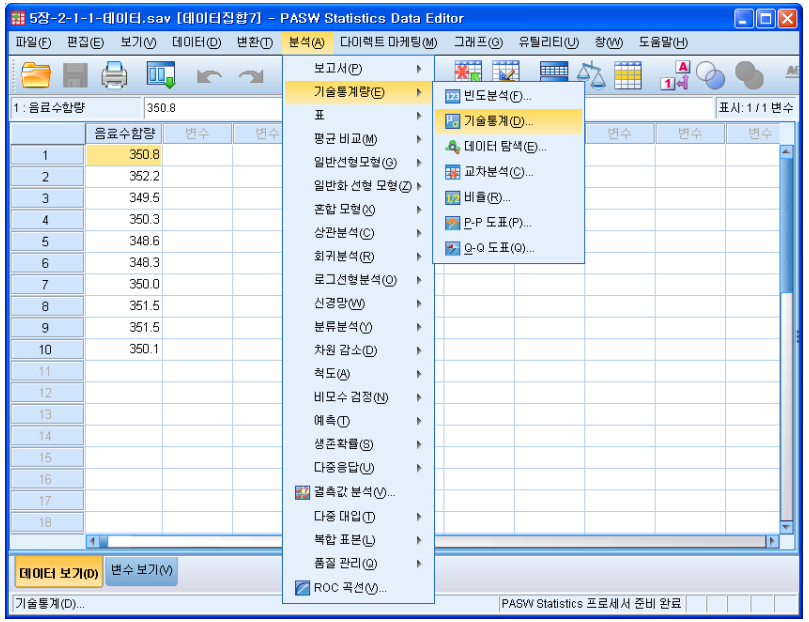
- 10개의 음료수 병 함량 데이터를 수집함

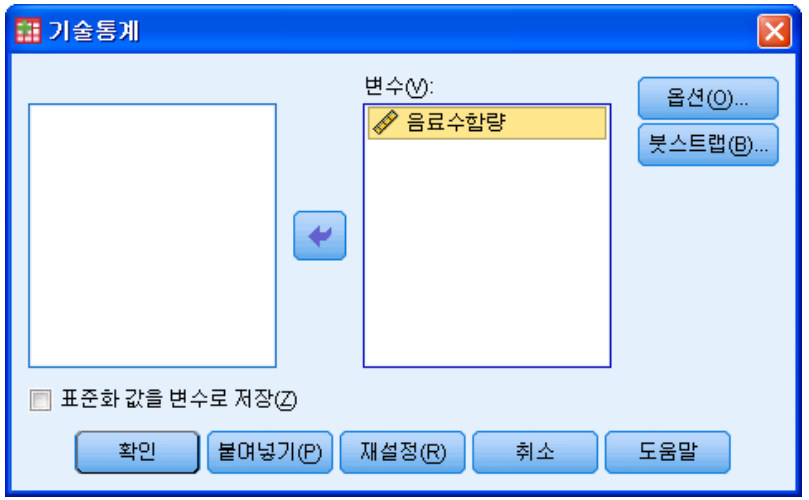
(3) 분석과정

(5장-2-1-1.sav)

– 기술통계 확인

- STEP 01: [기술통계] 메뉴 클릭
- STEP 02: 분석 변수 지정





기술통계량

	N	최소값	최대값	평균	표준편차
음료수함량	10	348.3	352.2	350.280	1.2647
유효수 (목록별)	10				

– X^2 값 계산

- STEP 01: [변수 계산] 메뉴 클릭
- STEP 02: X^2값 계산 식 입력
- STEP 03: 계산된 X^2 값 보기

$X^2 = (n-1)S^2 / \sigma^2$

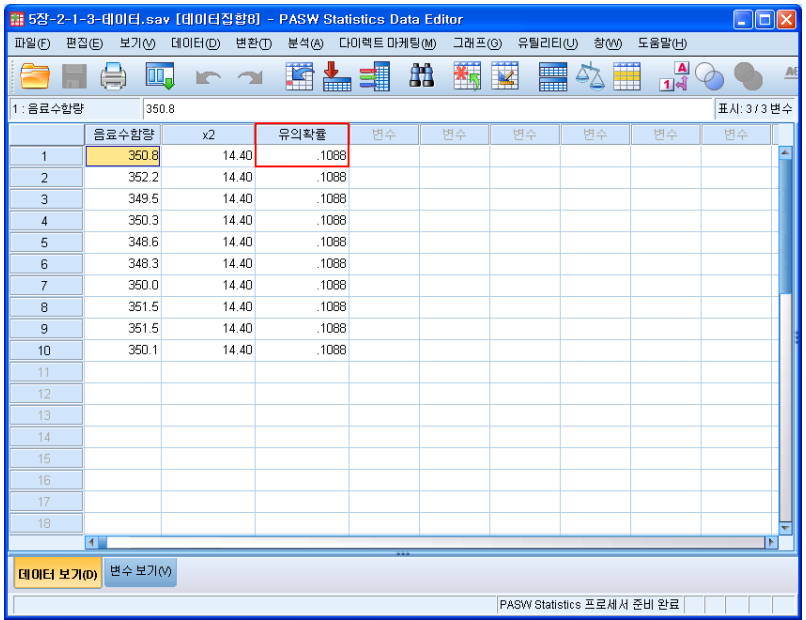
1)유의확률 계산(p.221)

- STEP 04: 다시 [변수계산] 메뉴 클릭
- STEP 05: 계산식 입력
- STEP 06: 유의확률 확인

유의확률
=(1-Cdf.Chisq(X^2 값,(n-1)))

2) $X^2 > X^2(n-1, \alpha)$이면
귀무가설 기각

(4) 결과해석



PART 3 다변량통계분석

Chapter 8. 회귀분석

1. 회귀분석의 개요

1.1. 회귀분석이란

- 정의
  - 1개 또는 그 이상의 독립(또는 설명) 변수들과 1개의 종속변수들의 선형관계를 파악하기 위한 기법
- 회귀분석의 주요 목적
  - 독립변수와 종속변수간의 선형 상관관련성 여부
  - 상관관계가 있다면 관계의 크기 및 유의도
  - 변수들간의 종속관계의 성격(+ 또는 -)
  - 회귀분석은 독립변수들과 종속변수와의 "선형결합관계"를 유도

<Note> 회귀분석의 기초지식

1. 회귀분석의 개요

■ **(1) 회귀분석이란? :**

어떤 하나의 변수 값(종속변수, 목적변수)을,
다른 변수의 값(독립변수, 설명변수)을 이용해서
ⓐ 예측하고 ⓑ 제어하고자 할 때 사용하는 분석방법.

< 참고 > 회귀분석의 종류
1) 단순회귀분석 : 독립(설명)변수가 1개
2) 다중회귀분석 : 독립(설명)변수가 2개 이상

< 참고 > 더미(dummy) 변수 :
자료의 형태가 질적변수일 때 사용, 0과 1만을 취함.

■ **(2) 회귀분석의 절차 :**

1단계 : 두 변수의 산점도 작성 → 선형관계의 존재여부 확인

2단계 : 최소제곱법으로 최적의 함수값을 구함.

3단계 : 회귀계수에 대한 유의성 검정을 실시.

H_0 : $\boldsymbol{\beta} = \mathbf{0}$ (회귀모형은 의미가 없다.)

4단계 : 의사결정을 실시.

■ **(3) 회귀분석의 가정 :**

1. 기본가정 : 자료분석에서 회귀분석을 실시하기에 알맞은 환경.
   ① 종속변수에 영향을 미친다고 생각되는 독립변수는 모두 포함.
   ② 독립변수들은 비교적 독립이어야 함. ⇒ 다중공선성의 문제.

2. 고려해야 할 사항들

① 다중공선성 : 한 독립변수의 값이 증가할 때, 다른 독립변수의 값이 증가하거나 감소하는 현상.

(check 하는 방법)

ⓐ 분산팽창지수(VIF) = $1 / (1-R_i^2)$

* R_i^2 : x_i 이외의 변수를 독립변수로, x_i 변수를 종속변수로 하는 회귀분석에서의 결정계수값.
* VIF ≥ 10 이면 다중공선성의 문제가 있다고 봄.

ⓑ 공차한계(다중공선 허용치, tolerence) : VIF 의 역수.

* $TOL_i = 1 / VIF_i$
* 만약 $TOL_i < 0.1$ 이면 다중공선성의 문제가 있음.

ⓒ 고유값과 상태지수.

* 고유근이 1에 비해 매우 작은 경우.
* 상태지수(condition number : c_i) :
  $10 < c_i < 30$ 이면 의심, $c_i \geq 30$이면 심각.

② 오차의 자기상관(autocorrelation) 문제 :
ⓐ 주요 독립변수가 누락되었을 때,
ⓑ 적용한 함수의 형태가 자료에 부적합할 때 흔히 발생.

<참고> 더빈-왓슨(Durbin-Watson) 통계량
㉠ D-W 값이 2에 가까우면 ⇒ 자기상관 무시.
㉡ D-W 값이 0에 가까우면 ⇒ 정(positive)의 자기상관.
㉢ D-W 값이 4에 가까우면 ⇒ 부(negative)의 자기상관.
즉, 0이나 4에 가까우면 모형이 부적합 함.

③ 다른 유사한 분석과의 차이점.
㉠ 판별분석 : 독립변수(metric척도), 종속변수(명목척도).
㉡ 회귀분석 : 독립, 종속 모두 metric척도. – 대부분의 경우.
㉢ 분산분석 : 독립변수가 명목변수이나, 구간척도의 경우.
㉣ 요인분석, 집락분석 : 독립/종속으로 구분하지 않고, 변수들의 관계를 이용하여 분석.

■ **회귀분석의 목표 :** 독립변수와 종속변수 사이의 관계식을 도출.

1) 단순회귀분석 : $y = b_0 + b_1 x$

2) 다중회귀분석 : $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

이 때, $b_0, b_1, \cdots, b_p$를 회귀계수라 함.

<Note> 회귀식의 추정 ⇒ 회귀계수 β_0 및 β_1 의 추정

y의 추정값 $\hat{y}$을 회귀계수 β_0와 β_1의 추정량 b_0와 b_1을 이용하여 나타내면

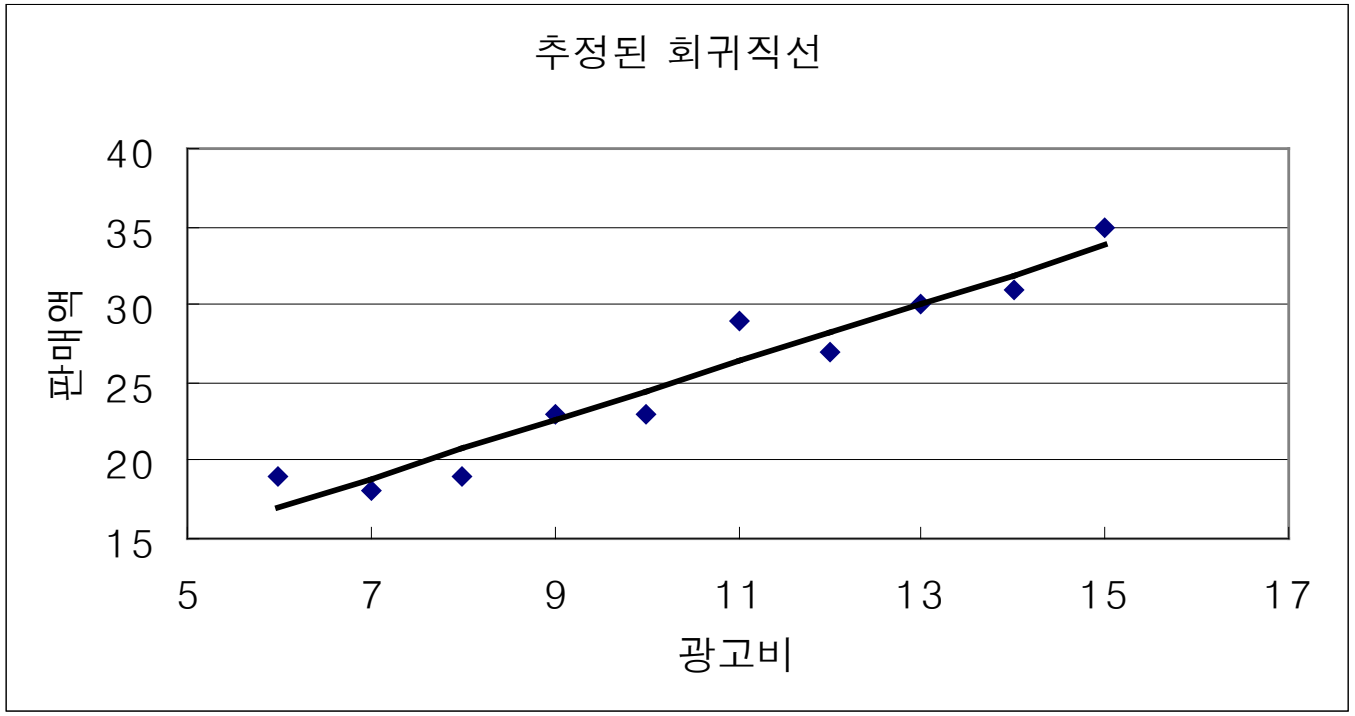
$$\hat{y} = b_0 + b_1 x$$

이 되며, 이 식을 표본회귀식(sample regression equation) 또는 추정된 회귀직선(estimated regression line)이라 한다. 여기서 b_0는 추정된 회귀직선의 절편(intercept)으로 $x=0$에서 $\hat{y}$의 값이며, b_1은 추정된 회귀직선의 기울기(slop)로 x가 한 단위 증가할 때마다 증가하는 $\hat{y}$의 증가량을 나타내고 있다.

(예) 관측 값과 추정된 회귀직선

⇒ ① 종속변수 : 판매액

② 독립변수 : 광고비



1) 잔차

⇒ 관측값과 추정값과의 차이를 잔차(residual)라 하고 e_i로 표시하면

$$e_i = y_i - \hat{y}_i$$

이다.

2) 최소제곱법 ⇒ 효과적인 회귀계수 추정방법

표본자료 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$에서 오차의 제곱합

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

을 최소로 만드는 β_0와 β_1의 값을 추정치 b_0와 b_1로 택하는 것이다.

오차의 제곱합 Q를 β_0와 β_1에 대하여 각각 편미분하고, 편미분한 방정식을 0으로 만드는 β_0와 β_1을 b_0와 b_1이라 하여 정리하면 다음과 같은 정규방정식 (normal equation)

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

을 얻을 수 있고, 이 식을 b_0와 b_1에 대하여 풀면

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

을 얻는다.

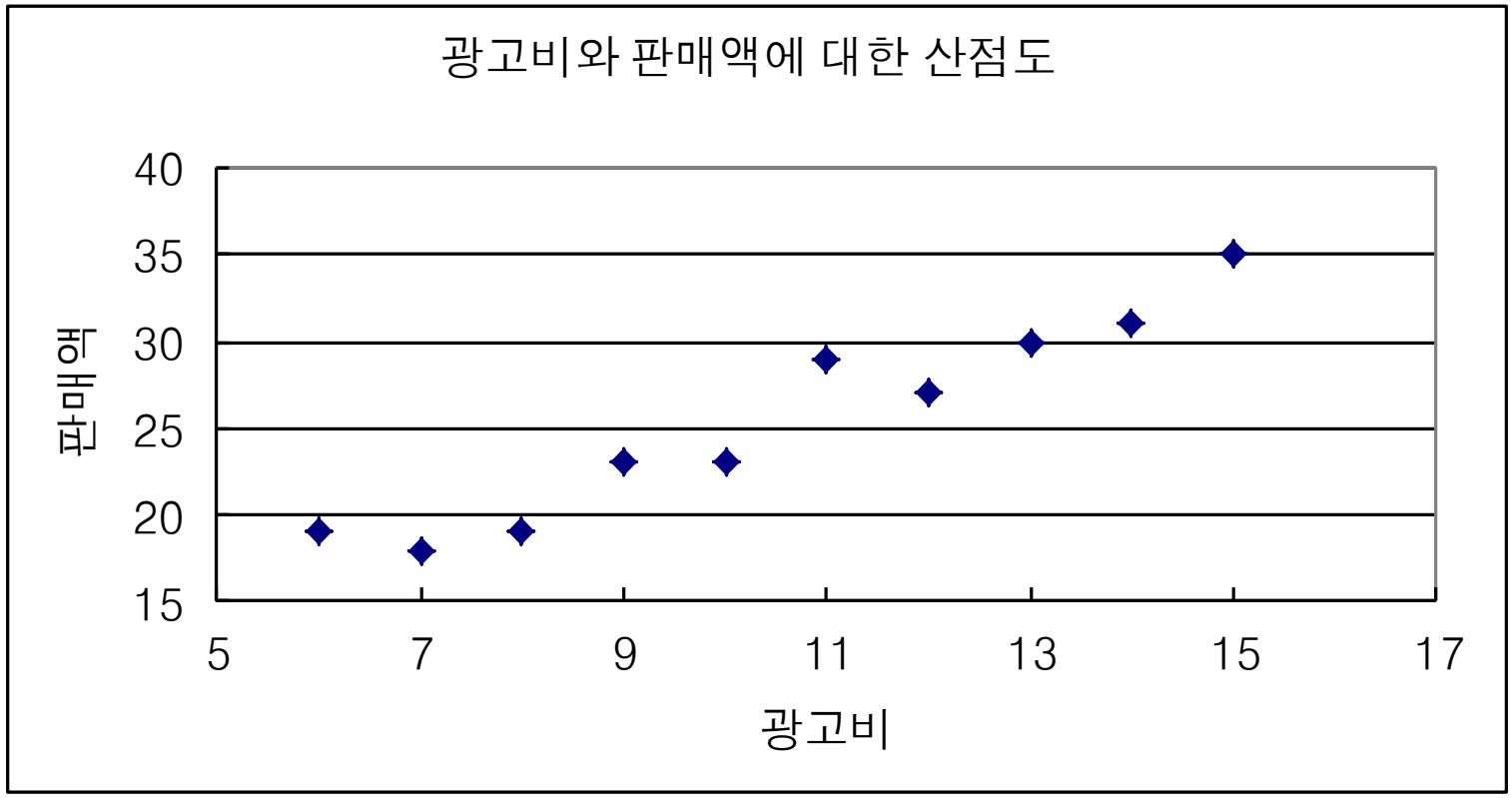
따라서 최소제곱추정량을 이용한 추정된 회귀직선

$$\hat{y}_i = b_0 + b_1 x_i, \quad (i = 1, 2, \cdots, n)$$

을 얻을 수 있다.

예제 1 어떤 제품을 판매하는 제조업체에서 제품의 판매액(단위: 백만원)에 광고비(단위: 10만원)가 미치는 영향을 조사하기 위해서 10개 대리점을 조사하여 다음과 같은 자료를 얻었다. 광고비와 판매액에 대한 산점도를 그려라.

광고비	13	10	9	15	11	14	12	7	8	6
판매액	30	23	23	35	29	31	27	18	19	19


광고비와 판매액에 대한 산점도
40
35
30
25
20
15
판매액
5
7
9
11
13
15
17
광고비

예제 4 **예제 1**에 주어진 자료를 사용하여 최소제곱법으로 회귀계수 β_0와 β_1을 추정하여 추정값 $\hat{y}_i$을 구하고, 잔차 $e_i = y_i - \hat{y}_i$를 구하라. 산점도 위에 추정된 회귀직선을 그려라.

풀이 광고비의 평균은 $\bar{x} = 10.5$, 판매액의 평균은 $\bar{y} = 25.4$, $S_{xx} = 82.5$ 그리고 $S_{xy} = 154$ 이다. 따라서 회귀계수의 추정치는

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{154}{82.5} = 1.8667$$

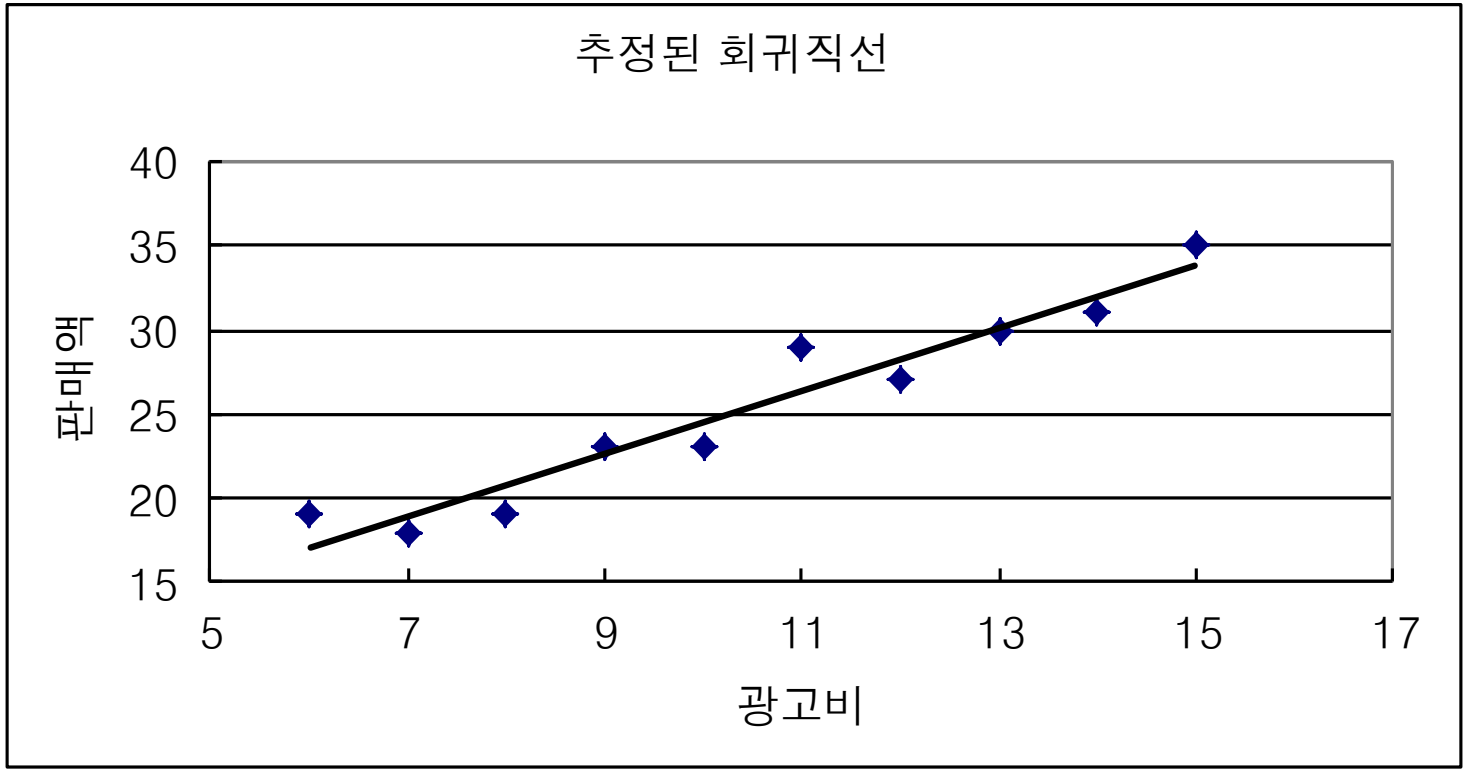
$$b_0 = \bar{y} - b_1\bar{x} = 25.4 - 1.8667 \times 10.5 = 5.8$$

이므로 추정된 회귀직선은

$$\hat{y} = 5.8 + 1.8667\,x$$

이 된다. 따라서 추정값 $\hat{y}_i$과 잔차 $e = y_i - \hat{y}_i$는 아래 표와 같다.

x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
13	30	30,0671	−0,0671
10	23	24,4670	−1,4670
9	23	22,6003	0,3997
15	35	33,8005	1,1995
11	29	26,3337	2,6663
14	31	31,9338	−0,9338
12	27	28,2004	−1,2004
7	18	18,8669	−0,8669
8	19	20,7336	−1,7336
6	19	17,0002	1,9998


추정된 회귀직선
40
35
30
25
20
15
판매량
5
7
9
11
13
15
17
광고비

1.2. 회귀분석의 종류

- 단순회귀분석 $y_i = \beta_0 + \beta_1 x_i - \epsilon_i$
- 다중회귀분석 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$
- 더미회귀분석 $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \beta_{d1} d_{d1i} + \cdots + \beta_{dl} d_{dli} + \epsilon_i$
- 로지스틱회귀분석 $P_z = \frac{1}{1+e^{c-z}}$
- 다항회귀분석 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$
- 비선형회귀분석 $y_i = \beta_0 (1 - e^{-\beta_1 x_i})$

2. 단순회귀분석

2.1. 모형의 개요

- 개요
  - 한 개의 종속변수와 한 개의 독립변수와의 선형관계를 파악하는 방법
- 모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- 모형추정
  - 최소 제곱법(OLS: Ordinary Least Squares)

$$\min \sum (y_i - \hat{y_i})^2 = \min \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

<Note> 단순회귀모형의 유의성 검정1

⇒ 단순회귀모형에서 특히 관심이 되는 것은 기울기 β_1이다.

→ 만일 β_1이 0이라면 독립변수와 종속변수사이에 직선적인 관계가 없다는 것을 의미

⇒ 회귀계수의 유의성 검정 : t-분포 이용

따라서 두 변수가 직선적인 관계를 가지고 있는지를 판단하는 검정으로 가설

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

을 검정하는 것은 단순회귀분석에서 중요하다.

◀ 1단계 : 가설설정

H_o : $\beta_1 = 0$ (추정한 회귀직선은 의미가 없다)

◀ 2단계 : 검정통계량 계산 ⇒ SPSS를 이용하여 계산

위 가설을 검정하기 위한 검정통계량은

$$t_0 = \frac{b_1}{\sqrt{MSE / S_{xx}}}$$

이다.

여기서 MSE는 오차의 분산 σ^2의 추정량으로

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

이다.

◀ 3단계 : 유의수준 α 하에서 기각역 결정
⇒ SPSS를 이용할 때는 p-값 이용

유의수준 α에서 기각역은

$$|t_0| > t_{\alpha/2}(n-2)$$

이 된다. 즉, 검정통계량이 기각역을 만족하면 귀무가설 H_0을 기각한다.

◀ 4단계 : 만약 귀무가설

$H_o : \beta_1 = 0$ (추정한 회귀직선은 의미가 없다)

가 기각되면, 추정된 회귀직선은 유의 하다고 결론 내린다.