가설검정이란?

- 과학적 실험이나 연구관찰등을 통해 얻어진 자료에 근거하여 연구대상이 되는 현상들에 대한 가설이나 이론등에 대한 타당성을 검정
- 확률변수들의 분포에 대한 단정이나 추측등의 표현을 통계적 가설(statistical hypothesis)
  - 단순가설(simple hypothesis)
  - 복합가설(composite hypothesis)

	H_0 참	H_0 거짓
H_0 채택	옳은 결정	제 2종 오류
H_0 기각	제 1종 오류	옳은 결정

집단의 평균 비교 (T 검정)

어떤 모집단으로부터 크기 m 인 랜덤표본을 얻고 또 독립적으로 다른 모집단으로부터 크기 n인 랜덤표본을 추출하였다고 했을 때...

표본 A : $x_1^A, x_2^A, \cdots, x_m^A$ (표본평균 $\overline{x_A}$, 모평균 μ_A, 모분산 σ_A^2)

표본 B : $x_1^B, x_2^B, \cdots, x_n^B$ (표본평균 $\overline{x_B}$, 모평균 μ_B, 모분산 σ_B^2)

우리의 관심은 다음의 가설

$$H_0 : \mu_A = \mu_B \quad \text{vs} \quad H_1 : \mu_A \neq \mu_B$$

이때 판단 기준인 표본평균의 차이 $\overline{x_B} - \overline{x_A}$ 는 어떤 확률분포?

통계학이론에 의하면 미리 σ_A^2, σ_B^2 를 알고 있는 경우

그러나 일반적으로 미리 σ_A^2, σ_B^2 를 알고 있는 경우가 흔치 않으므로

$$t = \frac{\overline{x_B} - \overline{x_A} - (\mu_B - \mu_A)}{s_p\sqrt{(1/m + 1/n)}} \quad , \quad \text{단} \quad s_p^2 = \frac{(m-1)s_A^2 + (n-1)s_B^2}{m+n-2}$$

를 이용하게 된다.(단, 두 집단의 모분산이 같다는 가정 하에서

이때 통계량 t 는 자유도 (m+n-2)인 t 분포를 따르고 마찬가지로 귀무가설이 옳다는 가정 하에서 관측된 사건이 발생할 확률 p-값을 계산

이 검정을 이표본 t- 검정이라고 함.

모분산의 동일성(등분산성) 검정

두 독립 그룹의 검정은 모분산이 동일해야 한다는 가정 $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$이 전제되어 있음.

검정을 수행하기 전에 두 분산의 동일성을 먼저 검정해야 함.

만일 검정결과 분산이 유의한 차이를 보이면 검정 대신 수정 검정 (자유도가 조정되고 결합표준편차 대신 각 그룹별 표준편차를 이용)을 해야한다.

[예제 4.4] 어떤 A 지역의 월 평균 가계소득을 알아보기 위해 임의로 30 가구를 추출하여 조사한 결과 다음과 같다고 하자.

월 평균 가계소득(단위: 만원)
240, 320, 290, 380, 390, 250, 320, 370, 290, 270, 300, 290, 300, 420, 320, 310, 340, 250, 210, 200, 230, 380, 290, 380, 450, 400, 220, 310, 320, 430

[예제 4.4]에서 알아보고자 하는 것은 A 지역의 월 평균 가계소득이 300만원인가에 관심이 있다. 따라서 검정하고자 하는 가설은 다음과 같다.

H_0 : $\mu = 30$ (A 지역의 월 평균 가계소득이 300만원이다.)

H_1 : not H_0 (A 지역의 월 평균 가계소득이 300만원이 아니다.)

```
salary <- c(240, 320, 290, 380, 390, 250, 320, 370, 290, 270, 300, 290,
            300, 420, 320, 310, 340, 250, 210, 200, 230, 380, 290, 380,
            450, 400, 220, 310, 320, 430)
t.test(salary, mu = 300)
```

[예제 4.5] 보건사회부에서는 각 개인의 소득수준에 따라 하루에 섭취하는 단백질의 양에 차이가 있는가를 파악하고자 한다. 사람들을 소득수준에 따라 두 집단으로 구분하고, 이들 각각의 집단에서 임의로 표본을 추출하여 단백질 섭취량을 조사한 결과가 다음과 같다. 소득이 높은 집단이 낮은 집단보다 단백질 섭취량이 많은가에 대하여 유의수준 5%에서 가설 검정하라.

소득수준이 높은 집단	87 86 59 68 98 69 80 78 69 77
소득수준이 낮은 집단	51 76 73 66 65 49 65 75 62 72 55 58 65 73

출처: 김규곤 등(2016)

```
# 독립표본 t-검정
# 두 집단의 데이터 입력
x <- c(87, 86, 59, 68, 98, 69, 80, 78, 69, 77)
y <- c(51, 76, 73, 66, 65, 49, 65, 75, 62, 72, 55, 58, 65, 73)
# 등분산성 검정(두 집단인 경우의 F-검정)
var.test(x, y)
# 등분산성 검정 결과 확인 후 옵션 지정
t.test(x, y, alternative="greater", var.equal=T)
```

대응 Paired T-검정

다음과 같이 n명의 대상으로부터 짝을 이룬 데이터가 있다고 가정

번호	처리전	처리후	차이
1	x_1	y_1	d_1
2	x_2	y_2	d_2
.	.	.	.
.	.	.	.
.	.	.	.
n	x_n	y_n	d_n

$$d_i = x_i - y_i$$

이제 처리전과 처리 후의 모평균을 각각 μ_1 , μ_2 라고 했을 때

이들의 차이를 $\mu_d = \mu_1 - \mu_2$ 라고 하면

우리의 관심: $H_o : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$

검정 절차

각 관측치간의 차이 $d_1, d_2, \cdots, d_n$ 이 서로 독립적이고 동일한 정규분포를 따른다는 가정 하에서 검정통계량은

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \qquad \text{단,} \quad s_d^2 = \sum (d_i - \bar{d})^2 / (n-1)$$

으로 귀무가설하에서 자유도 (n-1)인 t 분포를 따름.

그러므로 이를 이용하여 p-값을 계산 가능.

[예] 한 연구에서 비만인 9명의 여성에게 12주 동안 극저칼로리 식이요법(VLCD)을 실시한 다음 프로그램 참여 전 체중과 식이요법 후의 체중을 비교하였다.

이 프로그램은 체중감소 효과가 있다고 할 수 있는가?

번호	참여전 체중	참여후 체중	차이(전-후)
1	117.3	83.3	34.0
2	111.4	85.9	25.5
3	98.6	75.8	22.8
4	104.3	82.9	21.4
5	105.4	82.3	23.1
6	100.4	77.7	22.7
7	81.7	62.7	19.0
8	89.5	69.0	20.5
9	78.2	63.9	14.3
평균	98.53	75.94	22.59
표준편차			5.32

이때 귀무가설과 대립가설은 다음과 같다

$$H_o : \mu_d = 0 \quad \text{vs} \quad \mathrm{H}_1 : \mu_\mathrm{d} > 0$$

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = 12.74$$

한편 자유도 8인 t-분포의 임계치는 1.86이므로

t > $t_{0.05} = 1.86$ 가 되어 5% 유의수준에서 귀무가설은 기각.

즉 이 프로그램은 체중감소효과가 있다고 할 수 있음.

[예제 4.6] 한글 타자속도를 빠르게 하기 위한 교육을 8명의 타자수에서 실시하여 교육전과 교육후의 타자속도(글자/분)를 조사하였더니 아래와 같다. 단 타자속도는 정규분포를 따른다. 타자교육이 속도를 증가시켰는지 유의수준 5%에서 검정하여라.

타자수 번호	1	2	3	4	5	6	7	8
교육전	52	60	63	43	46	56	62	50
교육후	58	62	62	48	50	55	68	57

```
x <- c(52, 60, 63, 43, 46, 56, 62, 50)
y <- c(58, 62, 62, 48, 50, 55, 68, 57)
t.test(x, y, alternative="less", paired=T)
```

분산분석 (Analysis of Variance)

- 분산분석의 개념?
- 3개 이상의 모집단의 평균이 같은지를 검정
- 모집단의 분포가 정규분포이며 분산이 같다는 가정하에서 그 분산을 토대로 하여 평균이 같은지를 검정하는 방법

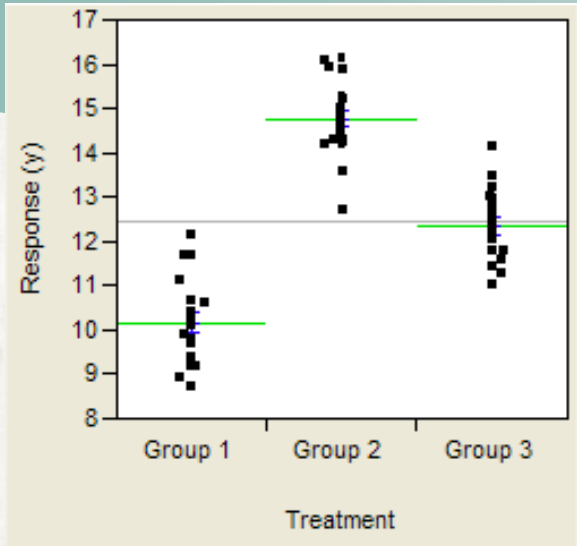
  ➔ 분산분석법

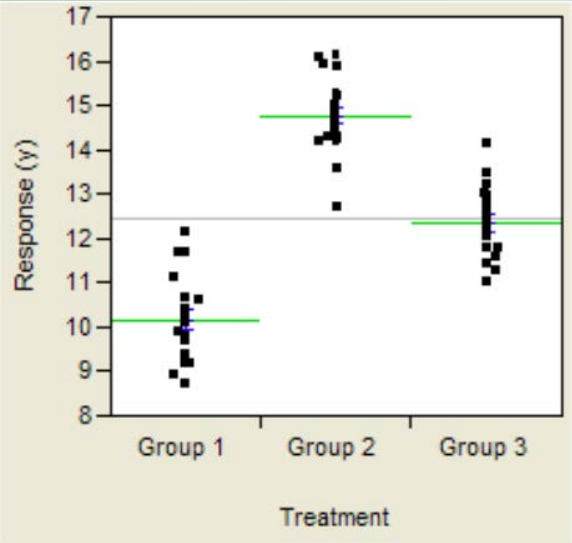
- K 개의 수준(level)을 가지는 요인의 처리효과에 관심이 있다고 하고 이러한 K개 그룹의 평균을 비교하는 문제를 고려.

처리	1	2	$\cdots$	K
	X_{12}	X_{21}	$\cdots$	X_{k1}
	X_{12}	X_{22}	$\cdots$	X_{k2}
	.	.	$\cdots$	.
	.	.	$\cdots$	.
	X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}
평균	$\overline{X_1}$	$\overline{X_2}$	$\cdots$	$\overline{X_k}$
모평균	μ_1	μ_2	$\cdots$	μ_k

$$n = n_1 + n_2 + \cdots + n_k$$

관심 : $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs H_1 : not H_0





$$\sum_i \sum_j (X_{ij} - \overline{X})^2 = \sum n_j(\overline{X_j} - \overline{X})^2 + \sum\sum(X_{ij} - \overline{X_j})^2$$

SS_T(전체평방합) = SS_A(처리평방합) + SS_E(오차평방합)

자유도 $n-1$ = $k-1$ $n-k$

분산분석표

변 인	평방합	자유도	평균평방합	F 비
처리(그룹)간	SS_A	$k-1$	$MS_A = SS_A/k-1$	$F = \frac{MS_A}{MS_E}$
오차	SS_E	$N-k$	$MS_E = SS_E/N-k$	
전체	SS_T	$N-1$		

분산분석을 하는데 전제되는 가정

(1) 관측치들은 각 그룹(처리수준)내에서 정규분포를 따른다.

(2) 각 그룹(처리수준별)의 모분산은 동일하다.

(3) 관측치들은 서로 독립적이다.

(1) 튜키의 HSD 절차(Tukey's HSD Procedure)

튜키의 검정 혹은 튜키의 HSD 검정은 쌍별 비교에만 적용되는 기법으로서 Bonferroni t와 같이 처리수준의 수와 오차평방합의 자유도를 고려한 승수(multiplier)를 이용한다. HSD 통계량은

$$HSD = \text{승수} \times \sqrt{MS_E/n}$$

(2) 셰페의 절차(Sheffe's Procedure)

셰페의 방법은 모든 사후 비교방법 중 가장 다용도적인 방법으로서 쌍별 비교뿐만 아니라 모든 형태의 비교를 할 수 있게 한다.

$$S = \sqrt{(k-1)F_{\alpha,df}}\sqrt{MS_E\sum C_j^2/n_j}$$

[예제 4.7] 한 백화점의 세 매장 A, B, C에서 5개월 동안 판매량을 정리한 자료의 결과가 다음과 같다. 분산분석표를 작성하고, 세 매장별로 판매량의 차이가 있는지를 유의수준 $\alpha = 0.05$로 검정하여라(출처: 김규곤 등, 2016).

세 매장의 판매량(단위 : 1,000,000원)

	A	B	C	총 평균
	67	58	70	
	59	59	64	
	60	60	68	
	60	61	62	
	64	62	71	
평균	$\overline{y_1} = 62.0$	$\overline{y_2} = 60.0$	$\overline{y_3} = 67.0$	$(\overline{y}) = 63.0$
제곱합	$\sum_{j=1}^{5}(y_{1j} - \overline{y_1})^2 = 46$	$\sum_{j=1}^{5}(y_{2j} - \overline{y_2})^2 = 10$	$\sum_{j=1}^{5}(y_{3j} - \overline{y_3})^2 = 60$	

```
oneway <- read.table("D://data//oneway_anova.txt", header=T)
resultone <- aov(oneway$selling~oneway$shop)
resultone
summary(resultone)
boxplot(oneway$selling~oneway$shop)
TukeyHSD(resultone)
```

교차(분할표) 분석

예 1) 완두콩 데이터 : 멘델의 유전법칙

품 종	1	2	3	4
도 수	150	80	70	20

H_0 : $p_1 = 9/16$, $p_2 = 3/16$ $p_3 = 3/16$, $p_4 = 1/16$

(자료는 멘델의 법칙을 따른다.)

H_1 : not H_0 (자료는 멘델의 법칙을 따르지 않는다.)

```
x <- c(150, 80, 70, 20)
p0 <- c(9/16, 3/16, 3/16, 1/16)
chisq.test(x, p = p0)
```

교차(분할표) 분석

- 동일성검정

성별에 따라 영화선택기준이 다른지 조사하고자 한다. 남학생 100명과 여학생 100명을 랜덤 추출하여 가장 선호하는 기준을 택하게 하여 분류한 결과가 다음과 같다. 과연 남학생과 여학생의 영화선택기준이 다르다고 할 수 있는가?

	배 우	감 독	영화내용
여 자	10	5	85
남 자	20	10	70

출처: 김규곤 등(2016)

homogeneity - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

no	sex	category
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11	1	2

```
homo <-
read.table("C://data//homogeneity.txt",
header=TRUE, sep="")

homo$sex
homo$category
table(homo$sex, homo$category)
 summary(table(homo$sex,
homo$category))
 chisq.test(homo$sex, homo$category)
```

독립성검정 : 월 수입과 학력간의 관계 조사

		월수입		
		상	중	하
학 력	고 졸	50	50	100
	전문대졸	70	200	50
	대 졸	120	60	50

월 수입과 학력은 연관이 있는가?

```
category1 <- c(50, 50, 100)
category2 <- c(70, 200, 50)
category3 <- c(120, 60, 50)
indepm <- rbind(category1, category2, category3)
chisq.test(indepm)
```

- [예 1]은 약종류에 따라 감기발생의 분포가 동일한지를 검정하는 경우:

  ➔ **동일성검정(test of homogeneity).**

- [예 2]는 월 수입과 학력간의 연관관계를 알아보는 경우 : 두 요인간에 관계가 있는지를 검정

  ➔ **독립성검정(test of independence)**

이론적배경

2차원 분할표에서 각 칸의 관측도수(O_{ij})와 주어진 가설이 옳다는 가정 하에서의 기대도수(E_{ij})를 비교 --> 차이를 양적으로 평가 : 카이제곱 통계량(χ^2)

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

이때 카이제곱 통계량은 주어진 가설이 옳다는 가정 하에서

χ^2-분포라고 알려진 분포를 따른다. ==> p-값을 계산.

상식적으로 카이제곱 통계량값이 크면(p-값이 작으면)

즉, 관측도수와 기대도수 차이가 크면 주어진 귀무가설을 기각.

감기발생여부

		감기걸림	감기안걸림	합
약종류	위약	p_{11}	p_{12}	1
	비타민C	p_{21}	p_{22}	1

$H_0: p_{11} = p_{21}, \quad p_{12} = p_{22}$ $\qquad$ $H_1: H_0$가 아니다.

검정통계량 : $X^2 = \sum \frac{(\text{관측도수} - \text{기대도수})^2}{\text{기대도수}}$

drug * cold 교차표

			cold		전체
			no	yes	
drug	C	빈도	122	17	139
		drug의 %	87.8%	12.2%	100.0%
	placebo	빈도	109	31	140
		drug의 %	77.9%	22.1%	100.0%
전체		빈도	231	48	279
		drug의 %	82.8%	17.2%	100.0%

카이제곱 검정

	값	자유도	점근 유의확률 (양측검정)	정확한 유의확률 (양측검정)	정확한 유의확률 (단측검정)
Pearson 카이제곱	4.811[b]	1	.028		
연속수정[a]	4.141	1	.042		
우도비	4.872	1	.027		
Fisher의 정확한 검정				.038	.021
유효 케이스 수	279				

a. 2x2 표에 대해서만 계산됨

b. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 23.91입니다.

학력 * 월수입 교차표

			월수입			전체
			상	중	하	
학력	고졸	빈도	50	50	100	200
		전체 %	6.7%	6.7%	13.3%	26.7%
	대졸	빈도	120	60	50	230
		전체 %	16.0%	8.0%	6.7%	30.7%
	전문대졸	빈도	70	200	50	320
		전체 %	9.3%	26.7%	6.7%	42.7%
전체		빈도	240	310	200	750
		전체 %	32.0%	41.3%	26.7%	100.0%

카이제곱 검정

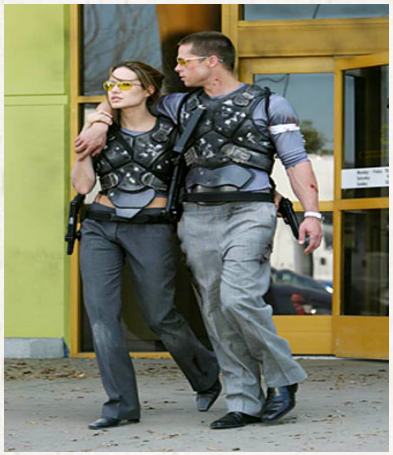
	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	160.653[a]	4	.000
우도비	151.111	4	.000
유효 케이스 수	750		

a. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 53.33입니다.

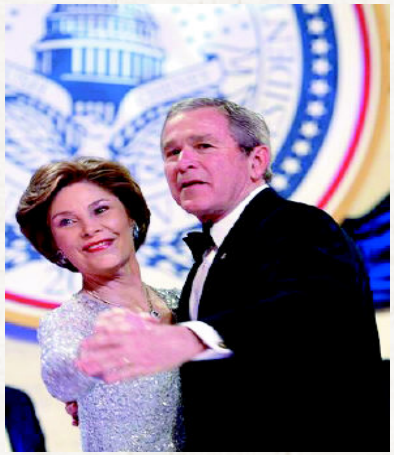
Regression

Example:

David Beckham: 1.83m
Victoria Beckham: 1.68m



Brad Pitt: 1.83m
Angelina Jolie: 1.70m



George Bush :1.81m
Laura Bush: ?

- To predict height of the wife in a couple, based on the husband's height

Response (out come or dependent) **variable** (**Y**): height of the wife
Predictor (explanatory or independent) **variable** (**X**): height of the husband

- 회귀분석이란?

=> 한 변수가 또 다른 하나이상의 여러 변수들에 의해 어떻게 설명(explanation) 또는 예측(forecasting)되는지를 알아보기 위해 적절한 함수식으로 표현하여 자료분석을 하는 통계적 방법

■ 단순 선형 회귀 모형

반응변수 y 와 설명변수 x 사이에 다음과 같은 선형관계가 있다고 가정

$$y = \beta_0 + \beta_1 x + \epsilon$$

==> 변수 y를 변수 x 에 대한 일차식으로 설명하려는 선형모형

관찰값 번호	y	x
1	y_1	x_1
2	y_2	x_2
.	.	.
.	.	.
n	y_n	x_n

선형회귀모형에서 모수 β_0, β_1을 추정하기 위하여

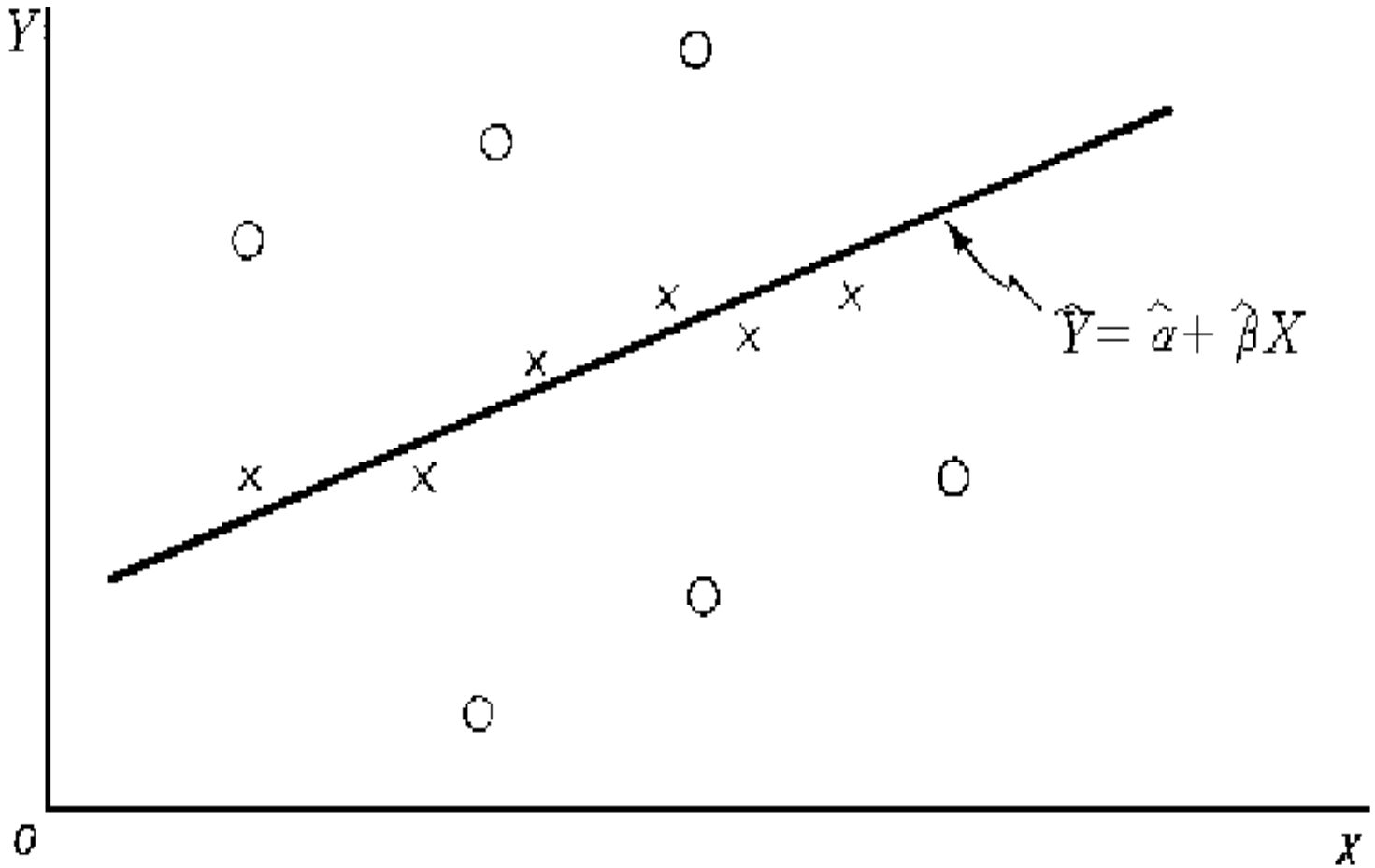
보통 최소제곱법(least squares method)을 사용.

최소제곱법이란 오차들의 제곱합을 최소로 하는 회귀계수를 추정하는 방법.

$$\hat{\beta}_0,\ \hat{\beta}_1 \quad ==> \quad \min \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$


Y
O
X
$\hat{Y} = \hat{\alpha} + \hat{\beta}X$

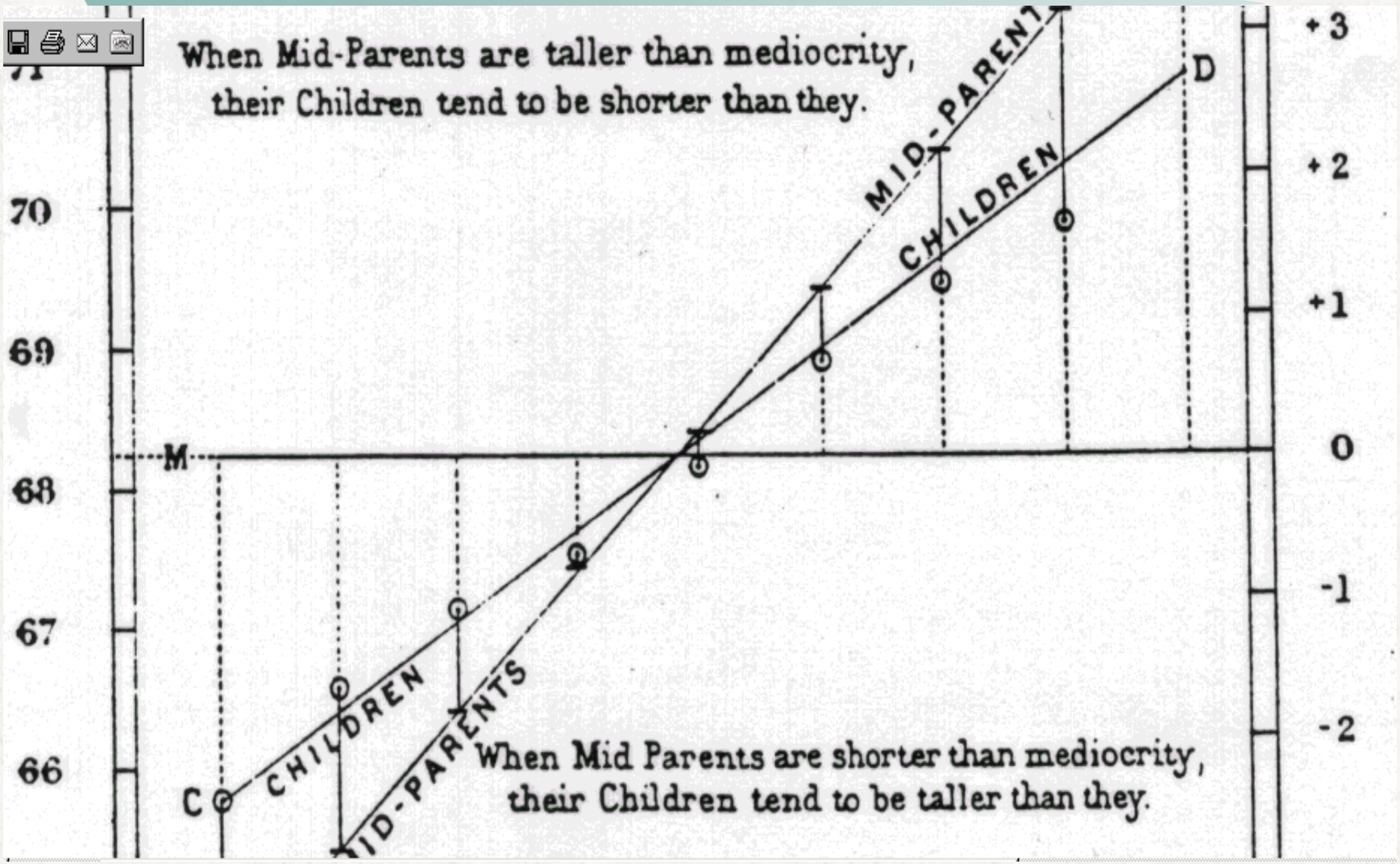
- History

Galton (1886) presented these data in a table, showing a cross-tabulation of 928 adult children born to 205 fathers and mothers,

by their height and their mid-parent's height.

He visually smoothed the bivariate frequency distribution

and showed that the contours formed concentric and similar ellipses, thus setting the stage for correlation, regression and the bivariate normal distribution.


When Mid-Parents are taller than mediocrity,
their Children tend to be shorter than they.
MID-PARENTS
CHILDREN
D
M
C
CHILDREN
MID-PARENTS
When Mid Parents are shorter than mediocrity,
their Children tend to be taller than they.
70
69
68
67
66
+3
+2
+1
0
-1
-2

```
f <- read.table("c://bigdata/galton.txt")
f
919 919 71.5 73.7
920 920 71.5 73.7
921 921 70.5 73.7
922 922 70.5 73.7
923 923 70.5 73.7
924 924 69.5 73.7
925 925 69.5 73.7
926 926 69.5 73.7
927 927 69.5 73.7
928 928 69.5 73.7
```

```
names(f) <- c("id", "parent","child")
plot(f$child, f$parent)
 summary(fit)
Call:
lm(formula = f$child ~ f$parent)
Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
f$parent     0.64629    0.04114  15.711   <2e-16 ***
---
```

에어로빅데이터 예

m_regression.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

no	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse
1	44	89.47	44.609	11.37	62	178	182
2	40	75.07	45.313	10.07	62	185	185
3	44	85.84	54.297	8.65	45	156	168
4	42	68.15	59.571	8.17	40	166	172
5	38	89.02	49.874	9.22	55	178	180
6	47	77.45	44.811	11.63	58	176	176
7	40	75.98	45.681	11.95	70	176	180
8	43	81.19	49.091	10.85	64	162	170
9	44	81.42	39.442	13.08	63	174	176
10	38	81.87	60.055	8.63	48	170	186
11	44	73.03	50.541	10.13	45	168	168
12	45	87.66	37.388	14.03	56	186	192
13	45	66.45	44.754	11.12	51	176	176
14	47	79.15	47.273	10.6	47	162	164
15	54	83.12	51.855	10.33	50	166	170
16	49	81.42	49.156	8.95	44	180	185
17	51	69.63	40.836	10.95	57	168	172
18	51	77.91	46.672	10	48	162	168
19	48	91.63	46.774	10.25	48	162	164
20	49	73.37	50.388	10.08	67	168	168
21	57	73.37	39.407	12.63	58	174	176
22	54	79.38	46.08	11.17	62	156	165
23	52	76.32	45.441	9.63	48	164	166
24	50	70.87	54.625	8.92	48	146	155
25	51	67.25	45.118	11.08	48	172	172
26	54	91.63	39.203	12.88	44	168	172
27	51	73.71	45.79	10.47	59	186	188

Ln 1, Col 1

```
mreg <- read.table("D://co//m_regression.txt", header=T)
result_mreg <-
lm(oxy~age+weight+runtime+rstpulse+runpulse+maxpulse)
result_sreg
summary(result_mreg)
```