

임상 연구에서 흔히 사용 하는 의학 통계의 실제

김 호

서울대학교 보건대학원

미래세대를 위한 교육 강좌

2007/6/17

Outline

- 통계적 가설 검정의 기본 개념들
 - 가설검정, 통계적 오류, 검정력 및 표본수
- 연속형 자료에서의 통계분석
 - T-test, ANOVA, 회귀분석 (단순회귀, 중회귀)
- 범주형 자료에서의 통계분석
 - 카이제곱 검정, 로지스틱 회귀분석
- 유전자형 자료분석의 기본 개념

기본개념들

- 모집단과 표본 (모수)
- p-value
- 통계적 검정력과 표본수 계산
- 모수적 방법과 비모수적 방법
- 정규성 검정
- 변수의 종류에 따른 분석법
- 통계적 가설 검정

통계적 사고

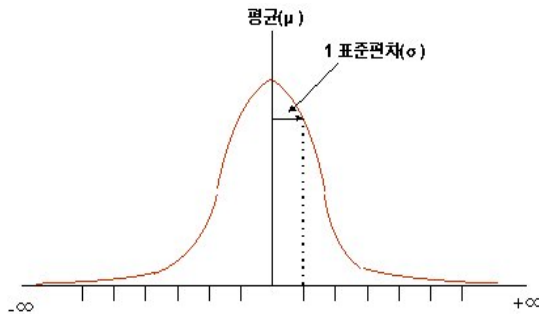
- <-> 결정론적 사고
- 모집단과 표본
- 정규분포를 결정하는 모수 (평균과 분산)
- 평균 : 위치
- 분산 : 산포 (정밀도)

모집단과 표본

- 모집단 : 연구자가 최종적으로 관심을 가지는 집단
- 표본 : 모집단에 대한 통계적 결정을 하기 위하여 모집단으로부터 대표성 있게 뽑은 집단
- 표본이 대표성이 있게 모집단을 반영하여야 함

모집단과 표본

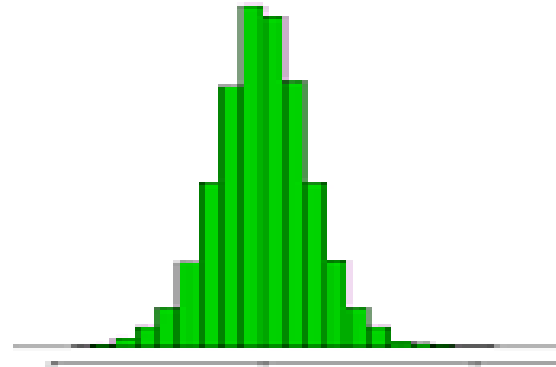
- 모집단
- 모수
 $N(\mu, \sigma^2)$



- 표본 Y_1, \dots, Y_n
- 추정치

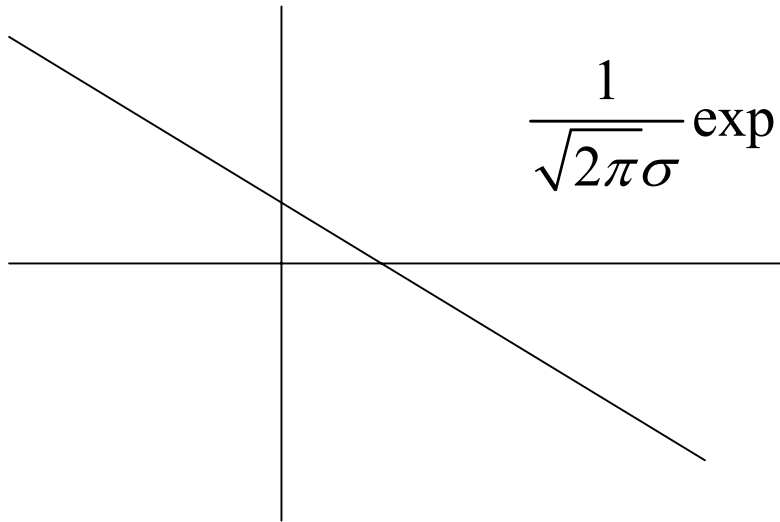
$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$



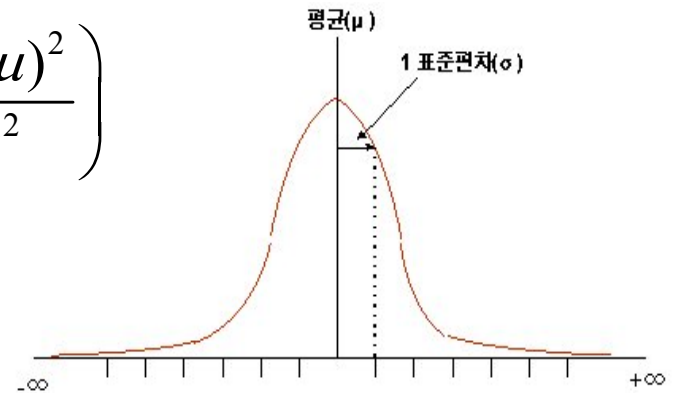
- 모수: 가정한 모형의 통계적 성질을 완전히 결정하는 상수(들)

$$Y = a + b x$$



$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$N(\mu, \sigma^2)$$



- 관심모수 : 연구의 가설을 수학적인 모수로 표시해야 함

- 두 집단에서 평균비교 $d = \mu_1 - \mu_2$

- 두 집단의 비율비교 $r = \frac{p_1}{p_2}$

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

- 연구가설

- 두 집단에서 평균비교

귀무가설

$$H_0 : d = \mu_1 - \mu_2 = 0$$

대립가설 (양측검정)

$$H_a : d = \mu_1 - \mu_2 \neq 0$$

대립가설 (단측검정)

$$H_a : d = \mu_1 - \mu_2 > 0$$

혹은

$$H_a : d = \mu_1 - \mu_2 < 0$$

- 연구가설

- 두 집단의 비율비교

귀무가설

$$H_0 : r = \frac{p_1}{p_2} = 1 \quad H_0 : OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = 1$$

대립가설 (양측검정)

$$H_0 : r = \frac{p_1}{p_2} \neq 1 \quad H_0 : OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \neq 1$$

대립가설 (단측검정)

$$H_0 : r = \frac{p_1}{p_2} > 1 \quad H_0 : OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} > 1$$

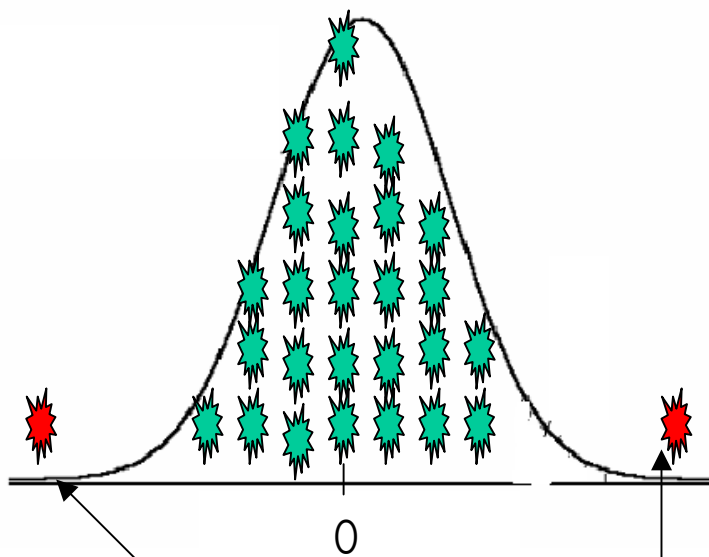
혹은

$$H_0 : r = \frac{p_1}{p_2} < 1 \quad H_0 : OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} < 1$$

P-value (1)

- 연구목적 : 관심변수의 (모)평균이 두 집단에서 다르다.
- \bar{Y}_1 첫 번째 집단에서의 표본 평균
- \bar{Y}_2 두 번째 집단에서의 표본 평균
- 만약 두 집단에서의 모평균이 같다고 하면
- 두 표본 평균은 비슷할 것이다.
- 표본평균의 차이를 반복적으로 구해보면

P-value (2)



통계적으로 대단히 일어나기 어려운 사건

P-value (3)

- P-value = 두 집단의 평균이 같다고 가정했을 때 우리의 자료, 혹은 더 차이가 나는 자료를 얻을 확률
- 작은 p-value : 위의 확률이 작다
 - ➔ 통계적으로 가능하지 않은 일이 일어났다.
 - ➔ 두 집단의 평균이 같다는 가정에 문제가 있다.
 - ➔ 두 집단의 평균은 같지 않다고 결론 내린다.

P-value (3)

- 작지 않은 p-value : 두 집단의 평균이 같다고 가정하면 우리의 자료를 관측할 확률이 작지 않다.
- ➔ 두 집단의 평균이 같다는 가정에 문제가 없다.

양쪽검정, 한쪽검정

- $A(\text{얻은 자료}) \rightarrow B(\text{연구가설})$
- $-B \rightarrow -A$
- 귀무가설 ($-B$) : 두 집단에 차이가 없다. (H_0)
- 대립가설 (B) : 두 집단에 차이가 있다. (H_a)
- 일종의 오류 : 옳은 귀무가설을 기각할 확률
 $= \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) \quad \alpha$
- 이종의 오류 : 틀린 귀무가설을 받아들일 확률
 $= \Pr(\text{Not reject } H_0 \mid H_a \text{ is true}) \quad \beta$
- Power = $1 - \beta$ (있는 차이를 발견할 확률)

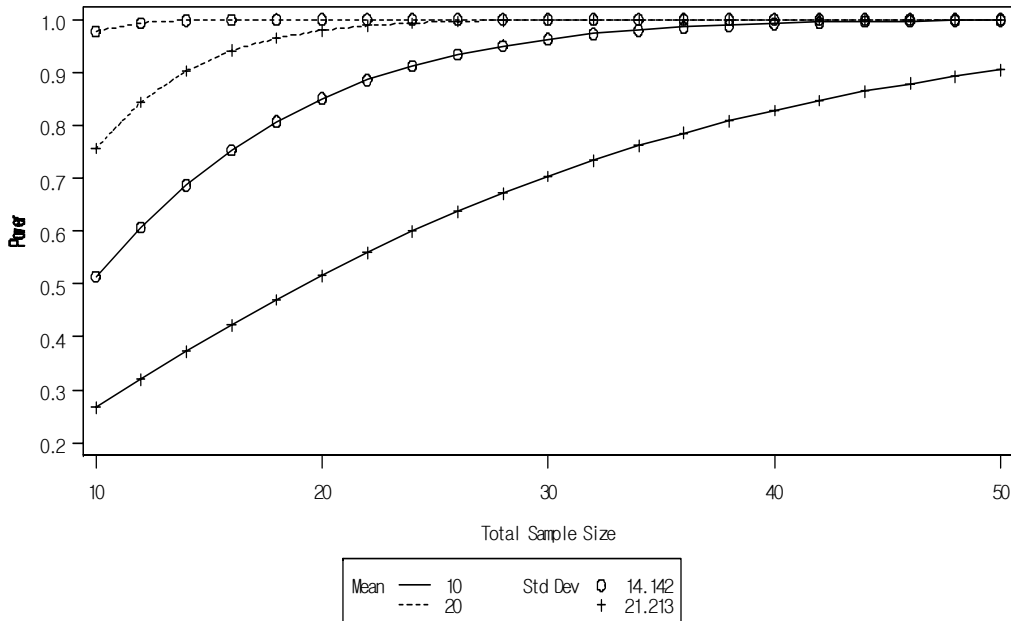
- 가설검정시 발생 가능한 4가지 상황

표본을 이용한 가설검정결과	모집단의 진실	
	H_0 참	H_0 거짓 (H_a 참)
H_0 채택	옳음 ($1 - \alpha$)	제2종 오류(β)
H_0 기각	제1종 오류(α)	옳음 (검정력= $1 - \beta$)

❖ 검정방법 A(모수적인 방법, 검정력=90%)의 검정방법 B(비모수적인 방법, 검정력=70%)보다 더 큰 검정력을 주었다.

-> 실제로 차이가 있을 때 A 방법을 100번 실시했을 때 90번의 경우 차이가 있다고(귀무가설 기각) 결정하였도 B 방법을 실시하였을 때는 100번 중 70번 귀무가설을 기각하였다.

-> A가 더 좋은 방법 ! (실제차이=?)



표본수 계산, Get Motivated <예시1>

	<i>Trt A</i>	<i>Trt B</i>	
<i>+</i>	n_{11} 52	n_{12} 48	n_{1+}
<i>-</i>	n_{21} 48	n_{22} 52	n_{2+}
	n_{+1}	n_{+2}	n_{++}

$$\chi^2 = \frac{(n_{11} - n_{+1}n_{1+} / n_{++})^2}{v_{11}}, \quad v_{11} = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n_{++}^2(n_{++} - 1)}$$

$$\chi^2 = \frac{(52 - 100 \times 100 / 200)^2}{(100 \times 100 \times 100 \times 100) / (200^2 \times 199)} = 0.32, \quad p > 0.05$$

- $n_{ij}^* = 100 \times n_{ij}$ 라고 하고, χ^2 를 다시 계산하면

$$\chi^{*2} = 100^2 / 100 \times \chi^2 = 32.00, p < 0.01$$

- 두 예에서 비율은 정확히 같음에도 불구하고 통계적 유의성은 상당히 다르다. ???
- 전통적 통계적 가설 검정의 유의성은 표본수에 크게 의존한다.
- 통계적 유의성이 없었던 경우라도 표본수를 크게 하면 유의성을 볼 수 있다.
- 표본수(실험의 비용)와 통계적 유의성(실험의 효용성)의 균형을 맞추는 것이 요구됨
- 최소의 비용으로 효과를 증명하고 싶다.

통계학에서의 표본수 계산

- 표본조사의 경우
 - 목적 : 추정 (estimation)
 - 도구 : 표본오차
 - 예 : 여론조사
- 임상시험의 경우
 - 목적 : 검정 (testing)
 - 도구 : 제1종의 오류, 제2종의 오류
 - 예 : 임상시험

- 단순임의 추출(simple random sampling)에서
- N : 모집단의 크기, n : 표본의 크기라면

$$\hat{\mu} = \bar{y} = \sum_{i=1}^n y_i / n$$

$$Var(\bar{y}) = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1} \right)$$

$$1.96\sqrt{Var(\bar{y})} \cong 2\sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}} = B: 95\% \text{ 신뢰구간 (표준오차)}$$

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}, D = B^2 / 4$$

- 만약 y_i 가 0 혹은 1의 값을 가지게 되면,
 \bar{y} 는 비율이 되고,

이 경우 $n = \frac{Npq}{(N-1)D + pq}$ 가 된다.

예1) $N=2000$, 95% 신뢰수준, $B=0.05$ 이라면 n 은 ?

>> 사전정보가 없다면 $p=q=0.5$ 대입

$$D = B^2 / 4 = 0.05^2 / 4 = .000625$$

$$n = \frac{2000 \times 0.5 \times 0.5}{1999 \times .000625 + 0.5 \times 0.5} = 333.56$$

최소한 334명의 표본이 필요하다.

연속형 변수의 비교

예) 새로운 관절염 치료제의 치료효과에 대한 임상실험을 실시한다고 하자.

치료효과는 2주간 치료 후 혈중 Prostag-landing 양이 평균 10, 표준편차 2 이면 치료가 된 것으로 간주한다. 치료 후 두 집단의 혈중 Prostaglandin 양의 변동이 20% 미만이면 두 치료제의 효과는 동등한 것으로 간주한다. 단측 검정으로 연구 대상수를 구하시오. 또 동일한 가정으로 양측 검정의 결과와 비교하시오. 검정력=90%,
결과의 척도 :Prostaglandin농도 (연속형)

$$n_c = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{(\mu_c - \mu_t)^2}$$

$$\Delta_A = \mu_1 - \mu_2 = 2 \mu gm / dl (10 \times 0.2), \sigma = 2.0 \mu gm / dl$$

$$Z_\alpha = 1.645, Z_\beta = 1.282$$

$$n_t = n_c = \frac{2 \times (1.645 + 1.282)^2 \times 2.0^2}{2^2} = 17.13 \cong 18$$

$$\text{Effective Size (=E/S)} = (\mu_1 - \mu_2) / \sigma$$

```
proc power;
```

```
twosamplemeans test=diff
```

```
meandiff = 2
```

```
stddev = 2
```

```
power=0.90
```

```
sides=1
```

```
npergroup=. ;
```

```
run;
```

Two-sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	1
Mean Difference	2
Standard Deviation	2
Nominal Power	0.9
Null Difference	0
Alpha	0.05

Computed N Per Group

Actual	N Per
Power	Group
0.902	18

```
proc power;
```

```
  onesamplemeans
```

```
  means = 10
```

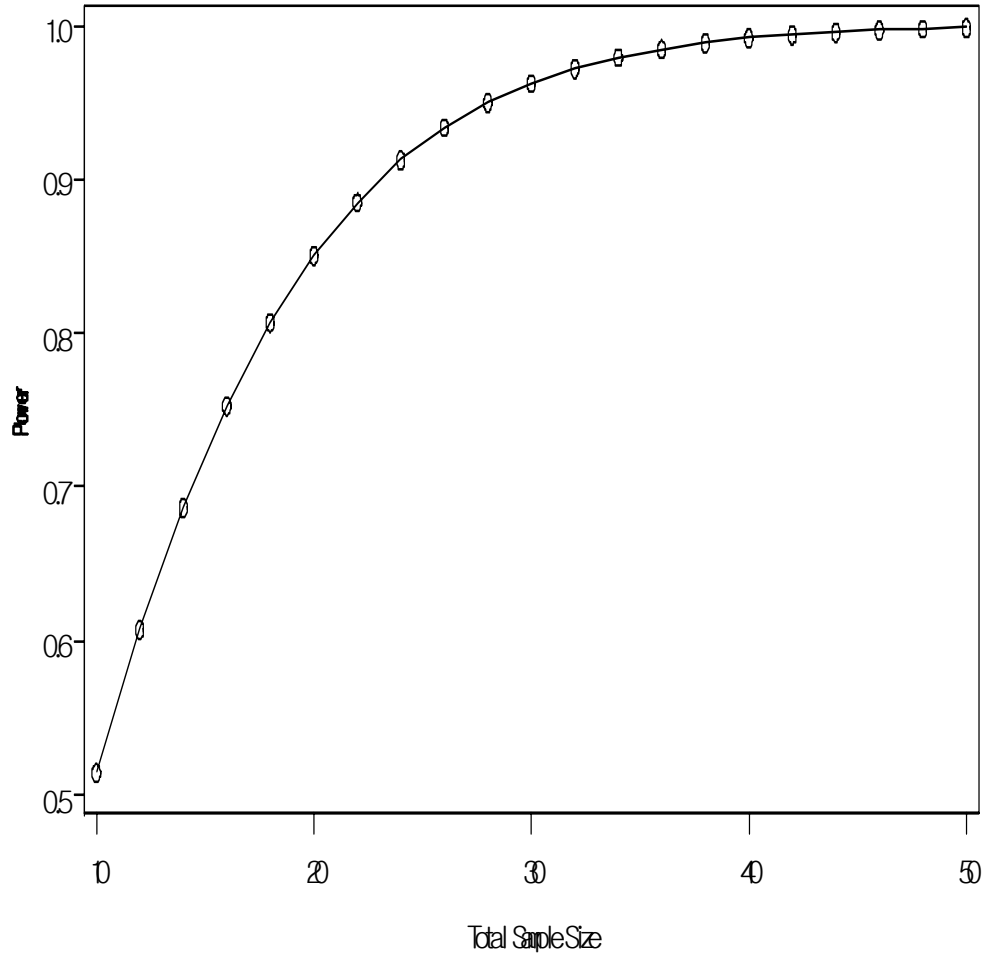
```
  stddev =14.142
```

```
  power=.
```

```
  ntotal=10 ;
```

```
  plot x=n min=10 max=50 ;
```

```
run;
```



```
proc power;
```

```
onesamplemeans
```

```
means = 10 20
```

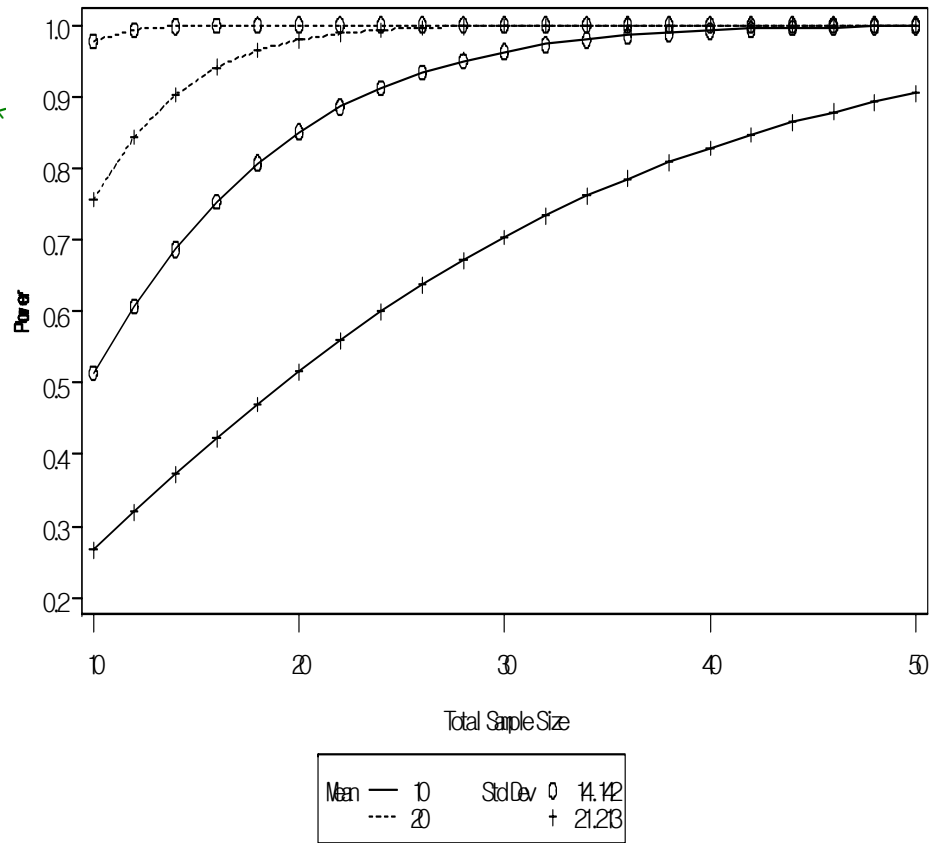
```
stddev =14.142 21.213 /*
```

```
power=.
```

```
ntotal=10 ;
```

```
plot x=n min=10 max=50 ;
```

```
run;
```



모수적 방법과 비모수적 방법 (1)

자료	평균	중앙값
1,2,3,4,5	3	3
1,2,3,4,5,100	19	3.5

- 중앙값(median)은 평균에 비하여 이상치에 대해서 둔감(robust)하다.
- 자료의 정규성 분포가정을 하면 평균과 분산을 통하여 모집단의 성질을 완전히 파악할 수 있다. (모수적 방법)

모수적 방법과 비모수적 방법 (2)

비모수적 방법은 자료의 (정규성) 분포가정을 하지 않는다

자료의 평균과 분산이 아닌 순위를 이용한 방법을 사용한다.

자료의 분포가정 (eg 정규성)이 만족되면 효율이 떨어진다.

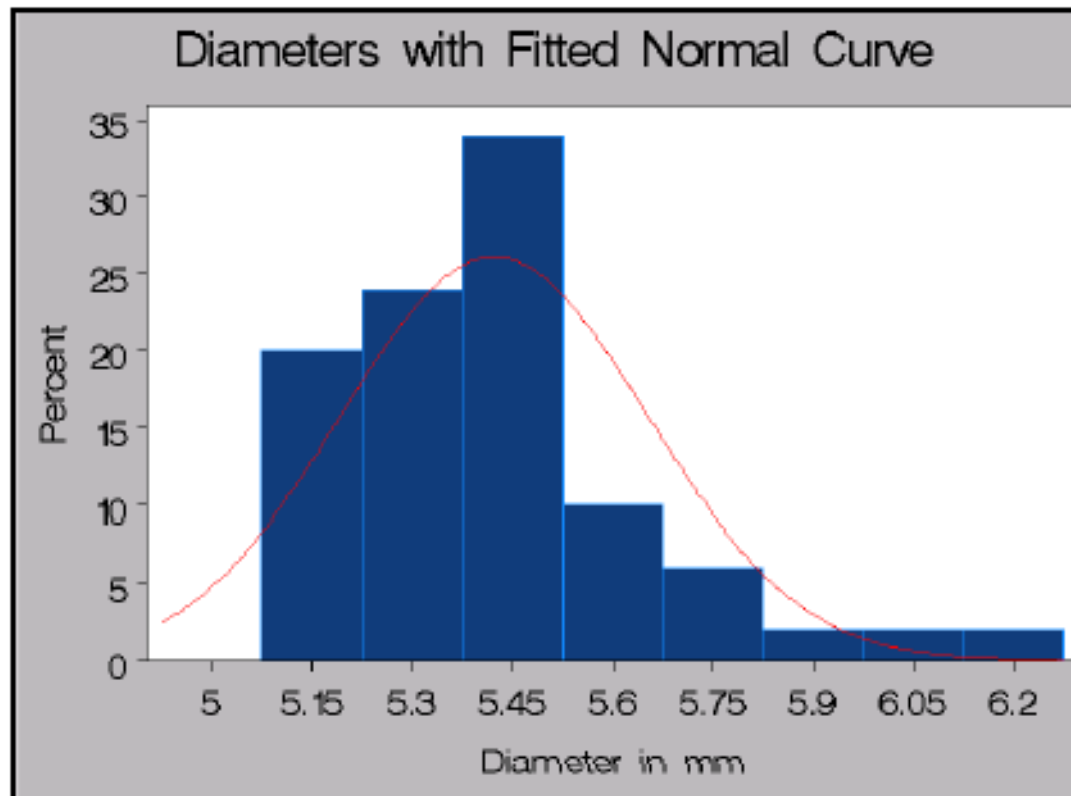
Robust 한 결과를 준다. (outlier에 둔감)

자료의 정규성 검정 (SAS 예제)

```
data ;  
input diameter @@;  
label diameter='Diameter in mm';  
datalines;  
5.501 5.251 5.404 5.366 5.445  
5.576 5.607 5.200 5.977 5.177  
...  
;  
run;  
proc univariate data=rods normal;  
histogram diameter /  
normal (mu=est sigma=est)  
midpoints = 5 to 6.30 by 0.15;  
run;
```

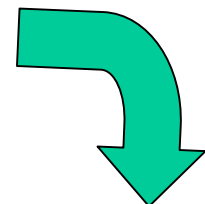
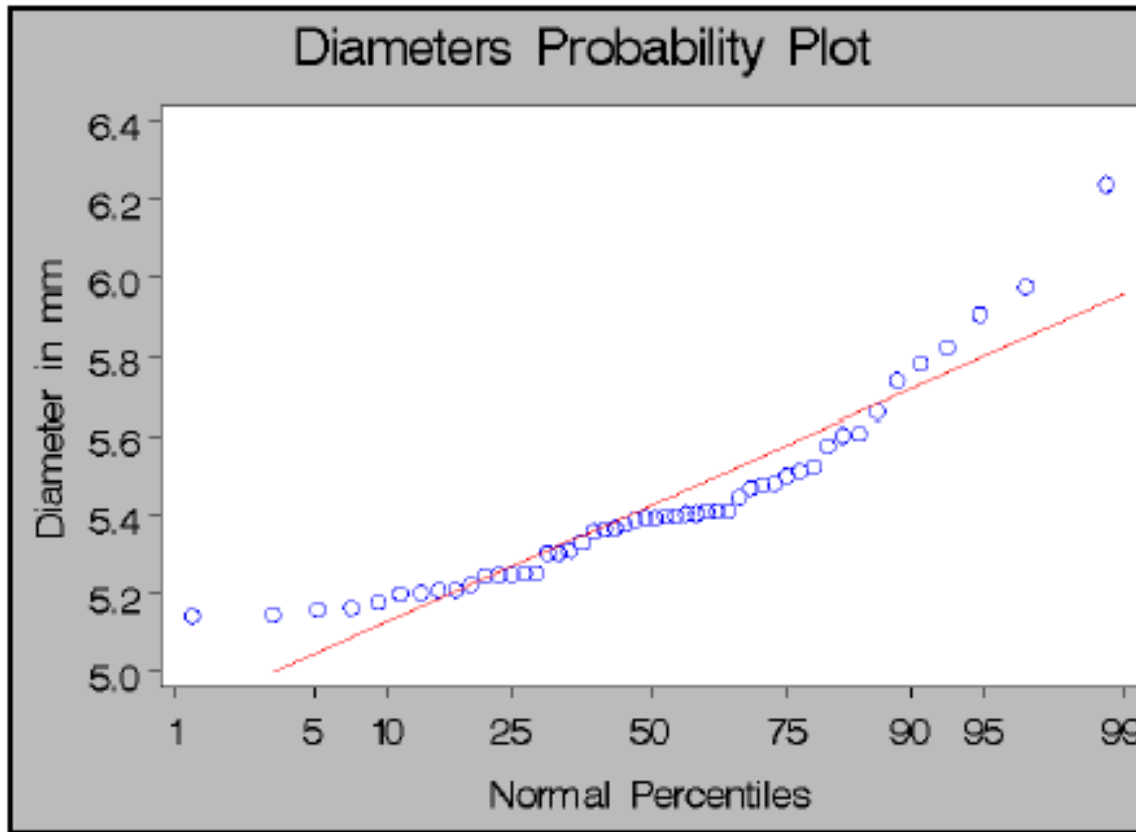
Tests for Normality

Test	--Statistic--	-----p Value-----
Shapiro-Wilk	W 0.87927	Pr < W 0.0001
Kolmogorov-Smirnov	D 0.181735	Pr > D <0.0100
Cramer-von Mises	M-Sq 0.27774	Pr > M-Sq <0.0050
Anderson-Darling	A-Sq 1.668626	Pr > A-Sq <0.0050


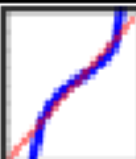
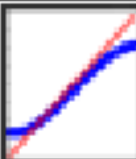
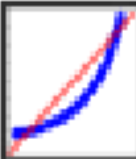
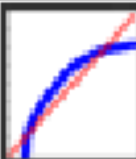
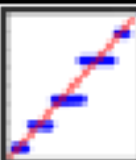


귀무가설: 자료가 정규분포를 따른다

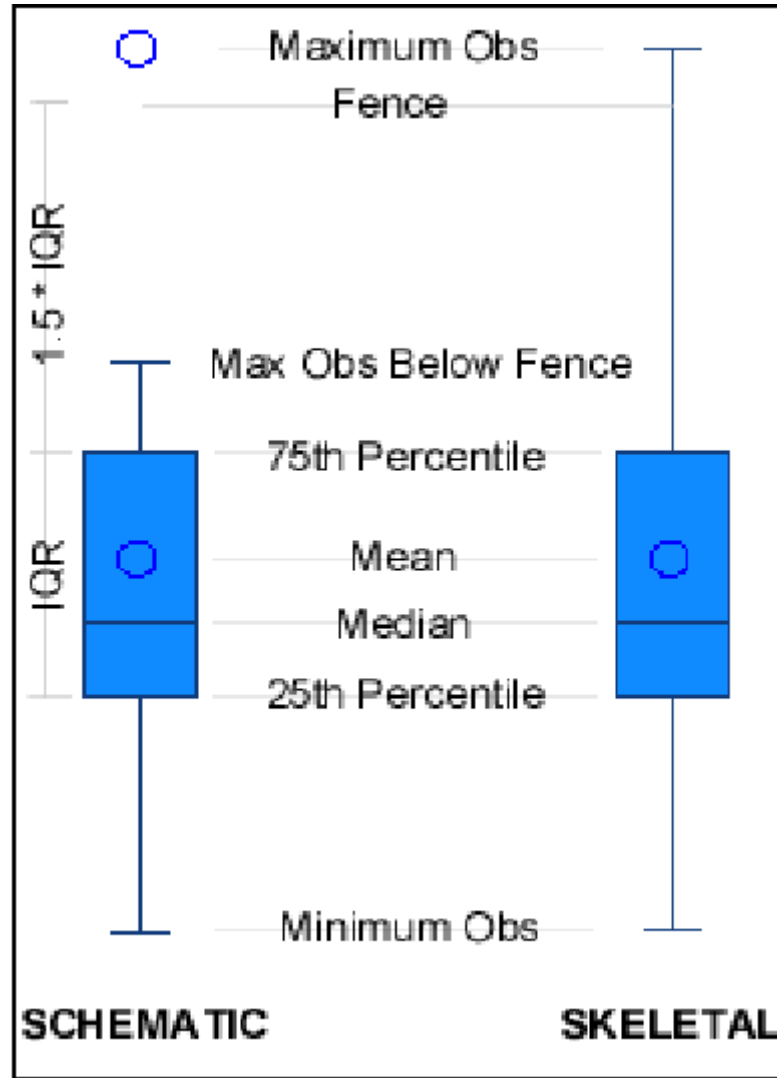
```
proc univariate data=rods noprint;  
  probplot diameter /  
  normal (mu=est sigma=est);  
run;
```



Skewed to the right

Pattern		Interpretation
	<p>All but a few points fall on a line</p>	<p>Outliers in the data</p>
	<p>Left end of the pattern is below the line while the right end of the pattern is above the line</p>	<p>Symmetric, long tails at both ends</p>
	<p>Left end of the pattern is above the line while the right end of the pattern is below the line</p>	<p>Symmetric, short tails at both ends</p>
	<p>Curved pattern with slope increasing from left to right</p>	<p>Skewed to right</p>
	<p>Curved pattern with slope decreasing from left to right</p>	<p>Skewed to left</p>
	<p>Staircase pattern</p>	<p>Data have been rounded or may be discrete</p>

Box plot



변수의 분류

- 수학적 개념(척도)에 의한 분류

- ① 명칭 or 명목척도 (nominal scale)

- 범주로만 의미

- ex) 성별, 혈액형

- ② 순위척도 (ordinal scale)

- 명목 + 대소관계

- 가감승제와 같은 수학적 조작은 불가능

- ex) 교육정도 (국졸/중졸/고졸/대졸)

- 사회경제수준 (상/중/하)

- 특정사항에 대한 의견 (아주찬성/찬성/중립/반대/아주반대)

③ 간격척도 (interval scale)

- 측정치 간의 간격에 의미가 있는 경우

ex) 병리소견에 -0 의 간격과 $0+$ 의 간격이 같은가?

온도의 경우 20°C , 30°C 와 10°C , 20°C 의 10°C 는 본질적으로 같다.

- 가감은 가능, 승제는 불가능

즉, 비(ratio)의 개념은 가지지 못함.

ex) $100^{\circ}\text{C}/50^{\circ}\text{C} \neq 212^{\circ}\text{F}/122^{\circ}\text{F}$

왜냐하면 0°C , 0°F 는 인위적인 영점을 정한 것이기 때문

④ 비척도 (ratio scale)

- 절대 영점을 가지게 되므로 수학적으로 가장 완전한 형태의 변수

ex) 40세는 20세에 비해 20살 많고(간격), 2배(비) 더 살았다

- 어떤 변수를 어떤 분류로 할 것인가를 미리 정해야 함

ex) 연령

11, 12, 13

← 비척도

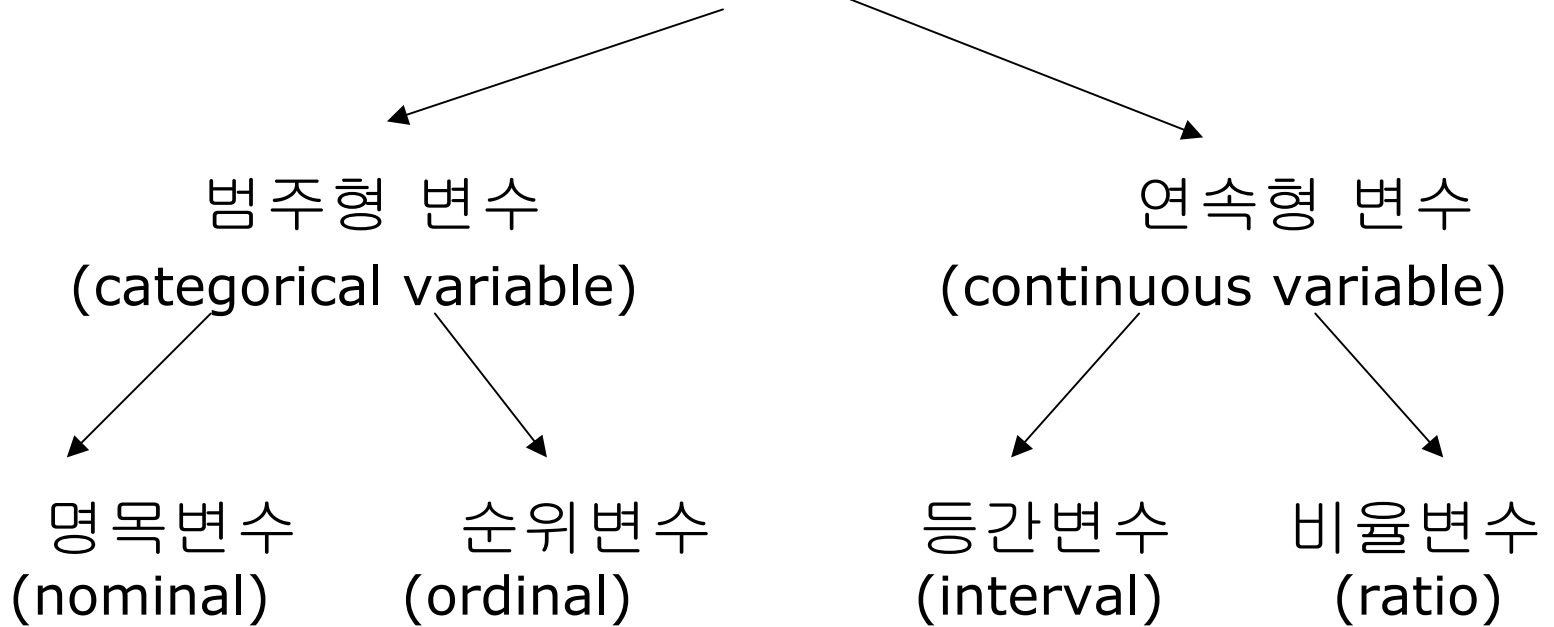
~9, 10~19, 20~

← 순위척도

성년 / 미성년

← 명칭척도 (순위척도 with 2 categories)

DATA



- 인과관계에 따른 변수

독립변수(설명변수) *independent (explanatory) variable*: 원인

종속변수 (반응변수) *dependent (response) variable* : 결과

- 전산입력 형식에 따른 변수

숫자변수

문자변수

날짜변수

변수 종류에 따른 통계분석법

종속변수	독립변수	통계분석법
연속변수 (혈압)	명목척도(2개 범주)	T 검정, paired T검정
연속변수 (혈압)	범주형 (3개 이상)	분산분석(ANOVA)
범주형 (병 발생 여부)	범주형 (투약여부)	카이제곱검정 (하나의 독립변수) 로지스틱 회귀분석(둘 이상의 변수)
연속형 (아기의 체중)	연속형 (재태 임신기간)	회귀분석
연속형 (출생 시 체중)	연속형 + 범주형 (재태기간 smoking 여부)	공분산분석 (ANCOVA)
생존시간 (연속형, >0)	연속형 + 범주형 나이 smoking 여부	생존분석

자료의 성격	모수적 방법	비모수적 방법
종속변수가 범주형	카이제곱검정	Fisher's exact test Ncnemar test Cochran's Q
종속변수가 연속형 두개의 독립된 집단	T-test	Wilcoxon rank sum test Man-whitney median test
두개의 짝 지은 집단	Paired t-test	Wilcoxon signed rank test
세 개 이상의 집단	ANOVA	Kruscal-Wallis test
제3의 변수의 영향고려	2-way ANOVA	Friedman's 2-way ANOVA
상관분석	Pearson correlation	Spearman's correlation Kendall's tau Stuart's tau

t-test

(연속변수의 두 집단 평균비교)

T-test

- 관심 변수가 연속일 때 (정규분포를 따를 때) 두 집단 간에 평균의 차이를 보는 검정 : 두 개의 독립적인 집단간의 차이
- Paired(짝지은) t-test : 한 개체에서 짝지은 관찰치들의 동질성을 볼 때 : 처치 전의 값과 후의 값을 비교할 때 (처치전과 후에 상관관계가 존재한다는 가정을 고려)
- 표본수가 적은 경우에는 정규분포 가정을 확인하기가 곤란하다. -> 비모수적 방법
- 두 개 이상의 집단 혹은 다른 변수로 보정을 할 때 -> ANOVA (분산분석)

Single Sample Analysis

Peppers Dataset

- dataset: peppers

Obs	angle
1	3
2	11
3	-7
4	2
5	3
6	8
7	-3
8	-2
9	13
10	4
11	7
12	-1
13	4
14	7
15	-1
16	4
17	12
18	-3
19	7
20	5
21	3
22	-1
23	9
24	-7
25	2
26	4
27	8
28	-2

```
proc means data=peppers mean std stderr t probt;  
run;
```

ptions

- . stderr: the standard error of the mean
- . t: $H_0: \mu = 0$ 을 검정하는 t test
- . probt: the significance probability of the t test

The MEANS Procedure

분석 변수 : angle

Pr > t	평균값	표준편차	표준오차	t값
3.1785714	5.2988718	1.0013926	3.17	0.0037

Two Independent Samples

Bullets Dataset

- dataset: bullets

Obs	powder	velocity
1	1	27.3
2	1	28.1
3	1	27.4
4	1	27.7
5	1	28.0
6	1	28.1
7	1	27.4
8	1	27.1
9	2	28.3
10	2	27.9
11	2	28.1
12	2	28.3
13	2	27.9
14	2	27.6
15	2	28.5
16	2	27.9
17	2	28.4
18	2	27.7

```
proc ttest data=bullets;
var velocity;class powder;
run;
```

The TTEST Procedure

Variable	powder	N	Lower CL Mean	Upper CL Mean	Lower CL Std Dev
velocity	1	8	27.309	27.638	0.2596
velocity	2	10	27.841	28.06	0.2106
velocity	Diff (1-2)		-0.771	-0.422	0.2582

Variable	powder	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
velocity	1	0.3926	0.799	0.1388	27.1	28.1
velocity	2	0.3062	0.5591	0.0968	27.6	28.5
velocity	Diff (1-2)	0.3467	0.5276	0.1644		

Variable	Method	Variances	DF	t Value	Pr > t
velocity	Pooled	Equal	16	-2.57	0.0206
velocity	Satterthwaite	Unequal	13.1	-2.50	0.0267

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
velocity	Folded F	7	9	1.64	0.4782

For H0: Variances are equal, F = 1.64 DF = (7,9)

Two Related Samples : paired t-test

Pulse Dataset

- dataset: pulse

Obs	pre	post	d
1	62	61	1
2	63	62	1
3	58	59	-1
4	64	61	3
5	64	63	1
6	61	58	3
7	68	61	7
8	66	64	2
9	65	62	3
10	67	68	-1
11	69	65	4
12	61	60	1
13	64	65	-1
14	61	63	-2
15	63	62	1

d = pre-post
(difference in rate)

```

proc means data=pulse mean std stderr t
  probt;
var d;
run;

```

The MEANS Procedure

분석 변수 : d

평균값	표준편차	표준오차	t값	Pr > t
1.4666667	2.3258383	0.6005289	2.44	0.0285

Two-sided p-value ←

One-sided p-value = $0.0285/2 = 0.0143$ for $H_0 : d = 0$ vs. $H_1 : d > 0$

Table 1. Blood Pressure (BP) Evolution (24-Hour Continuous Monitoring [CM] and Office BP)

	Doxazosin (Dx)	Enalapril (Ena)	<i>P</i> (Dx) <i>P</i> (Ena)Doxazosin (Dx)
24-Hour CMBP			
Systolic BP (mm Hg)	I: 152 ± 14	I: 149 ± 19	.004
	F: 142 ± 15	F: 140 ± 14	.03
Diastolic BP (mm Hg)	I: 86.5 ± 8	I: 86 ± 11	.01
	F: 83 ± 7	F: 82 ± 8	.09
Mean BP (mm Hg)	I: 108 ± 8	I: 107 ± 12	.002
	F: 103 ± 8	F: 101 ± 8	.03
Office BP			
Systolic BP (mm Hg)	I: 152 ± 15	I: 155 ± 15	.05
	F: 146 ± 21	F: 147 ± 18	.06
Diastolic BP (mm Hg)	I: 93 ± 7	I: 93 ± 9	.09
	F: 89 ± 11	F: 89 ± 9.6	.01
Mean BP (mm Hg)	I: 109 ± 14	I: 109.6 ± 13	NS
	F: 108 ± 14	F: 108 ± 12	NS

I, initial BP (baseline); F, final BP (12 months); NS, not significant.

TABLE 3. *Voiding diary variables at baseline and after 10 weeks of treatment with tolterodine or oxybutynin for overactive bladder symptoms*

Variable	Tolterodine	Oxybutynin
Voids/24 hrs.:		
No. evaluable pts.	190	188
Mean baseline \pm SD	11.3 \pm 3.2	11.4 \pm 3.5
Mean wk. 10 \pm SD	9.6 \pm 3.4	9.7 \pm 3.3
Mean change from baseline	-1.7	-1.7 (ANOVA p = 0.0001)
Estimated difference in mean change (95% CI)		0.0086 (-0.41-0.43)
p Value		0.97
Urge incontinence episodes/24 hrs.:		
No. evaluable pts.*	104	102
Mean baseline \pm SD	2.4 \pm 2.6	2.9 \pm 3.4
Mean wk. 10 \pm SD	1.1 \pm 2.1	1.1 \pm 2.1
Mean change from baseline	-1.3	-1.8 (ANOVA p = 0.0001)
Estimated difference in mean change (95% CI)		0.50 (-0.03-1.03)
p Value		0.065
Voided vol./void:		
No. evaluable pts.	190	188
Mean baseline \pm SD (ml.)	149 \pm 46	148 \pm 48
Mean wk. 10 \pm SD (ml.)	182 \pm 70	182 \pm 70
Mean change from baseline (ml.)	33	34 (ANOVA p = 0.0001)
Estimated ml. difference in mean change (95% CI)		-0.6 (-9.2-8.1)
p Value		0.90
Pads/24 hrs.:		
No. evaluable pts.†	43	47
Mean baseline \pm SD	3.1 \pm 1.9	2.8 \pm 1.5
Mean wk. 10 \pm SD	2.0 \pm 1.9	1.9 \pm 1.5
Mean change from baseline	-1.1 (ANOVA p = 0.0003)	-0.9 (ANOVA p = 0.0001)
Estimated difference in mean change (95% CI)		-0.19 (-0.65-0.28)
p Value		0.43

* Patients not reporting incontinence at baseline were excluded from analysis.

† Patients not using pads at baseline were excluded from analysis.

ANOVA (Analysis of Variance)

분산분석

세 집단 이상에서의 연속변수
평균들의 비교

ANOVA (Analysis of Variance)

변수

brand: 5개의 비닐 Brand → 전체평균

wear: 얼마나 약한가 → i 수준의 효과

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = \begin{matrix} ACME \\ AJAX \\ CHAMP \\ TUFFY \\ XTRA \end{matrix}$$

$$= \mu_i + \varepsilon_{ij}$$

↑
 i 수준의 평균

model

Obs	brand	wear
1	ACME	2.3
2	ACME	2.1
3	ACME	2.4
4	ACME	2.5
5	CHAMP	2.2
6	CHAMP	2.3
7	CHAMP	2.4
8	CHAMP	2.6
9	AJAX	2.2
10	AJAX	2.0
11	AJAX	1.9
12	AJAX	2.1
13	TUFFY	2.4
14	TUFFY	2.7
15	TUFFY	2.6
16	TUFFY	2.7
17	XTRA	2.3
18	XTRA	2.5
19	XTRA	2.3
20	XTRA	2.4

ANOVA for One-Way Classification

```
proc anova data=venerer; class brand;  
model wear=brand; run;
```

The ANOVA Procedure

Dependent Variable: wear

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.61700000	0.15425000	7.40	0.0017
Error	15	0.31250000	0.02083333		
Corrected Total	19	0.92950000			

R-Square	Coeff Var	Root MSE	wear Mean
0.663798	6.155120	0.144338	2.345000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
brand	4	0.61700000	0.15425000	7.40	0.0017

Least Significant Difference Comparisons of BRAND Mean

```
proc anova data=veneer;  
class brand;  
model wear=brand;  
means brand/lsd;  
run;
```

```
                The ANOVA Procedure  
                t Tests (LSD) for wear  
Alpha                                0.05  
Error Degrees of Freedom              15  
Error Mean Square                      0.020833  
Critical Value of t                    2.13145  
Least Significant Difference            0.2175
```

Means with the same letter are not significantly different.

T Grouping	Mean	N	brand
A	2.6000	4	TUFFY
B	2.3750	4	XTRA
B	2.3750	4	CHAMP
B	2.3250	4	ACME
C	2.0500	4	AJAX

Sas 예제

```
options pageno=1 nodate ls=130 ps=60 nocenter;
filename inbrakes 'c:\myweb\int\taillite.dat';
data one;
  infile inbrakes ;
  input  id vehtype group positn speedzn resptime follotme folltme;
if group=1;
  label vehtype='Vehicle Type'
        group='Group - Light On=1      Light Off=2'
        positn='Light Position'
        speedzn='Speed Zone'
        resptime='Response Time'
        follotme='Following Time in Video Frames'
        folltme='Following Time in Categories'
        ;run;
proc sort; by vehtype;
/* Let's do one-way ANOVA to see the effect of vehicle type */
proc anova;
class vehtype;
model resptime=vehtype;
title 'Parametric ANOVA analysis';
run;
/* What's wrong with this ?
   We didn't check the normality assumption.
   Let's do proc univariate to check the normality
```

The ANOVA Procedure

Class Level Information

Class	Levels	Values
vehtype	4	1 2 3 4

Number of Observations Read 733
 Number of Observations Used 733

Parametric ANOVA analysis
 2

The ANOVA Procedure

Dependent Variable: resptime Response Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3886.5377	1295.5126	3.75	0.0109
Error	729	252025.5278	345.7140		
Corrected Total	732	255912.0655			

R-Square 0.015187
 Coeff Var 41.91847
 Root MSE 18.59339
 resptime Mean 44.35607

Source	DF	Anova SS	Mean Square	F Value	Pr > F
vehtype	3	3886.537689	1295.512563	3.75	0.0109

```
proc univariate data=one normal plot;  
var resptime;  
by vehtype;  
histogram resptime /cfill=blue kernel(color=red) normal(color=black);  
probplot resptime / normal (mu=est sigma=est);  
title 'Normality Check';  
run;  
  
proc boxplot data=one ;  
plot resptime*vehtype  
  /boxstyle =SCHEMATIC  
  cboxes =blue  
  cboxfill =gray  
  idcolor=red ;  
run;
```

Vehicle Type=1

UNIVARIATE 프로시저

변수: resptime (Response Time)

적률

N	157	가중합	157
평균	42.9617834	관측치 합	6745
표준편차	17.6402386	분산	311.178017
왜도	1.79175355	첨도	5.20936144
제곱합	338321	수정 제곱합	48543.7707
변동계수	41.0603033	평균의 표준오차	1.40784431

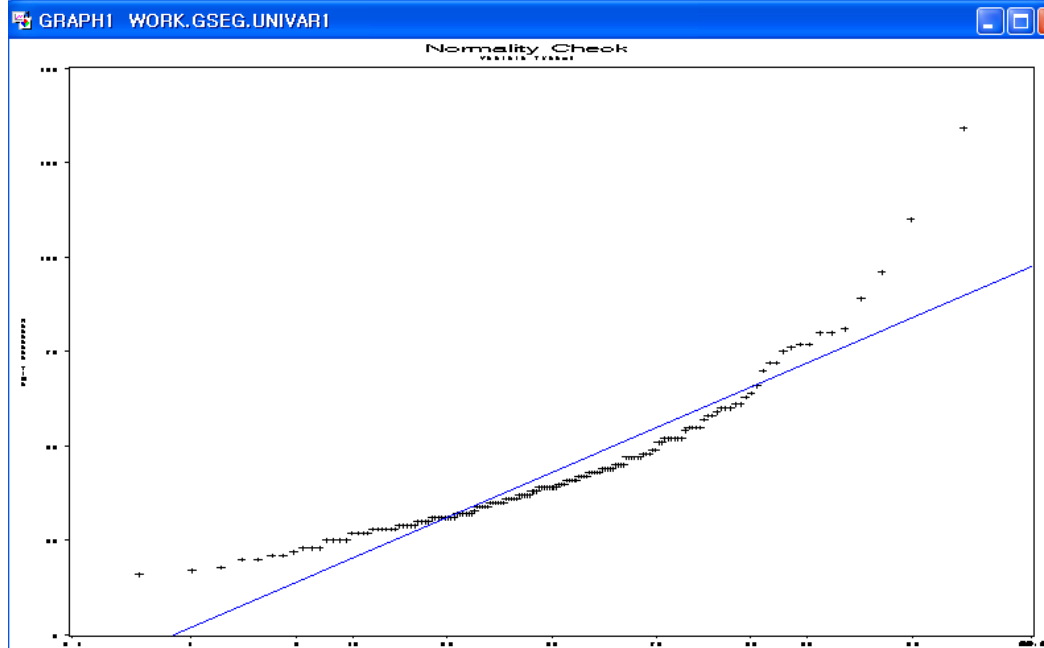
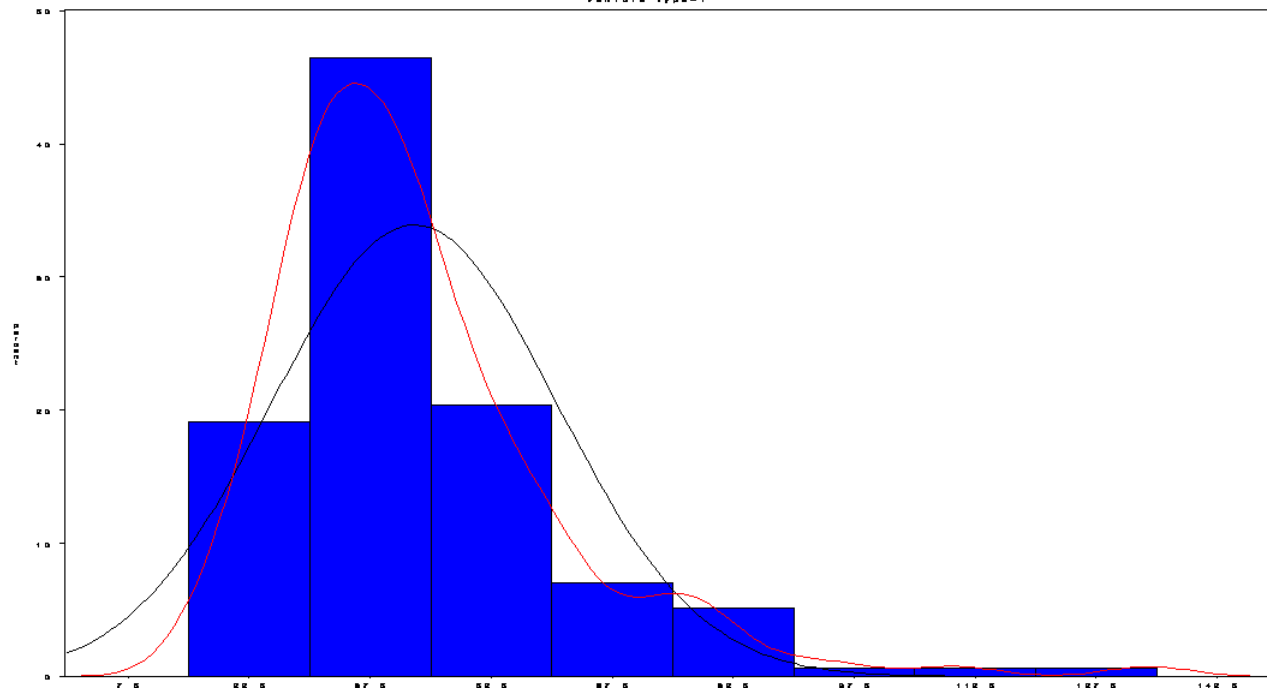
.....

정규 분포에 대한 적합도 검정

검정	-----통계량-----	-----p-값-----
Kolmogorov-Smirnov	D 0.13553580	Pr > D <0.010
Cramer-von Mises	W-Sq 0.78799637	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 4.60747650	Pr > A-Sq <0.005

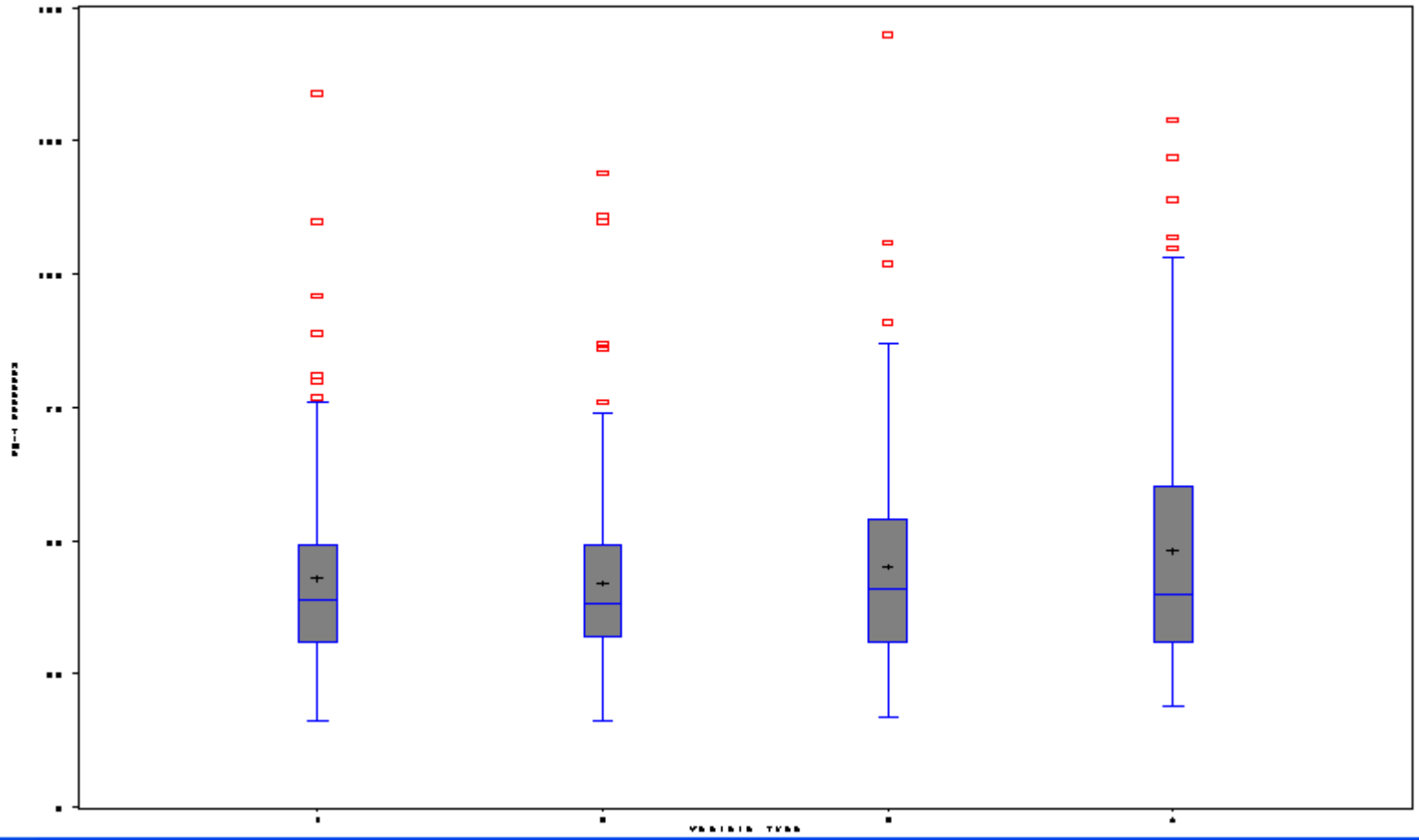
Normality Check

Variable: Type=1





Normality Check



```

proc npar1way wilcoxon;
class vehtype;
var resptime ;
title 'Nonpara One-Way ANOVA for Tail Light Study';
run;

/* The other way is transformation.
   Let's take log transformation so that we have normal
   distribution.
*/

data t;
set one;
t=log(resptime);
label t='ln (response time)';
run;

proc sort; by vehtype;

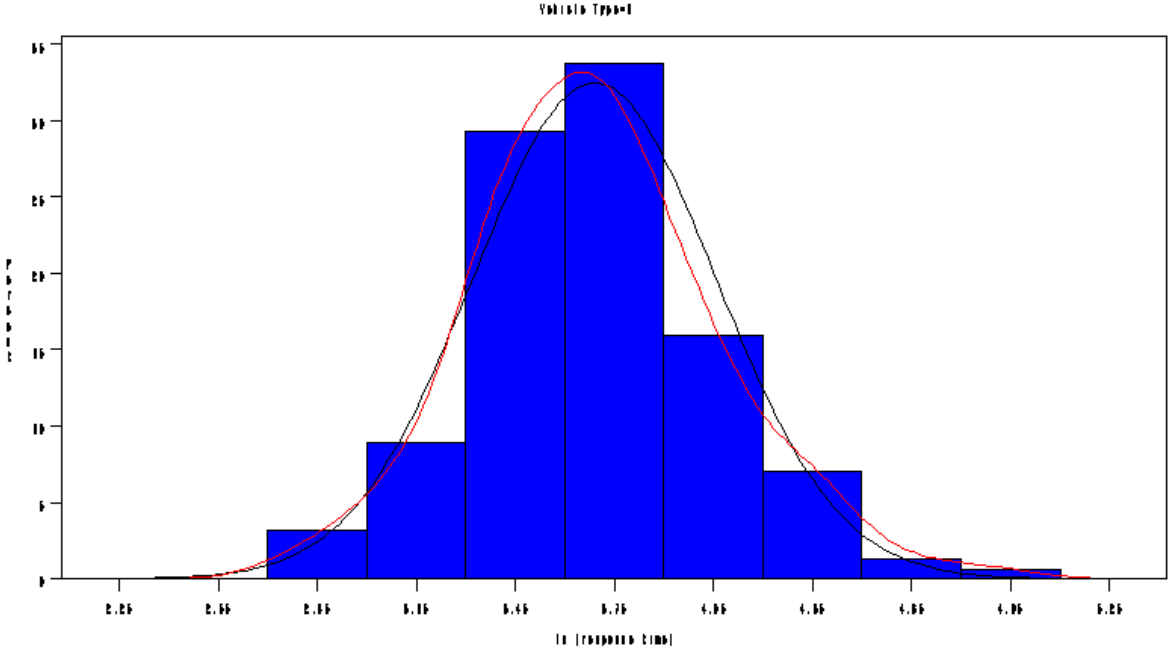
proc univariate data=t normal plot;
var t;
by vehtype;
histogram t /cfill=blue kernel(color=red) normal(color=black);
probplot t / normal (mu=est sigma=est);
title 'Normality Check for transformed variable';
run;

```

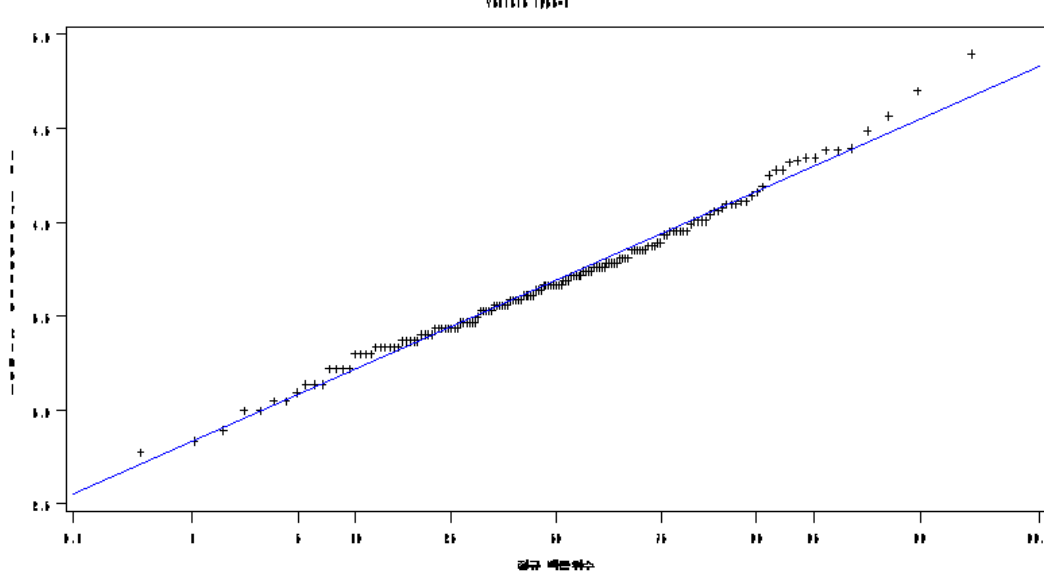
정규 분포에 대한 적합도 검정

검정	-----통계량-----	-----p-값-----
Kolmogorov-Smirnov	D 0.05711477	Pr > D >0.150
Cramer-von Mises	W-Sq 0.08162639	Pr > W-Sq 0.204
Anderson-Darling	A-Sq 0.49951454	Pr > A-Sq 0.215

Normality Check for transformed variable



Normality Check for transformed variable



```

/* The transformed variable seems to normally distributed.
   Then we can do parametric ANOVA with normality assumption
*/

```

```

proc anova;
class vehtype;
model t=vehtype;
title 'ANOVA for the log transformed response time';
run;

```

```

proc boxplot data=t ;
plot t*vehtype
     /boxstyle =SCHEMATIC
     cboxes   =blue
     cboxfill =gray
     idcolor=red ;
run;

```

The ANOVA Procedure

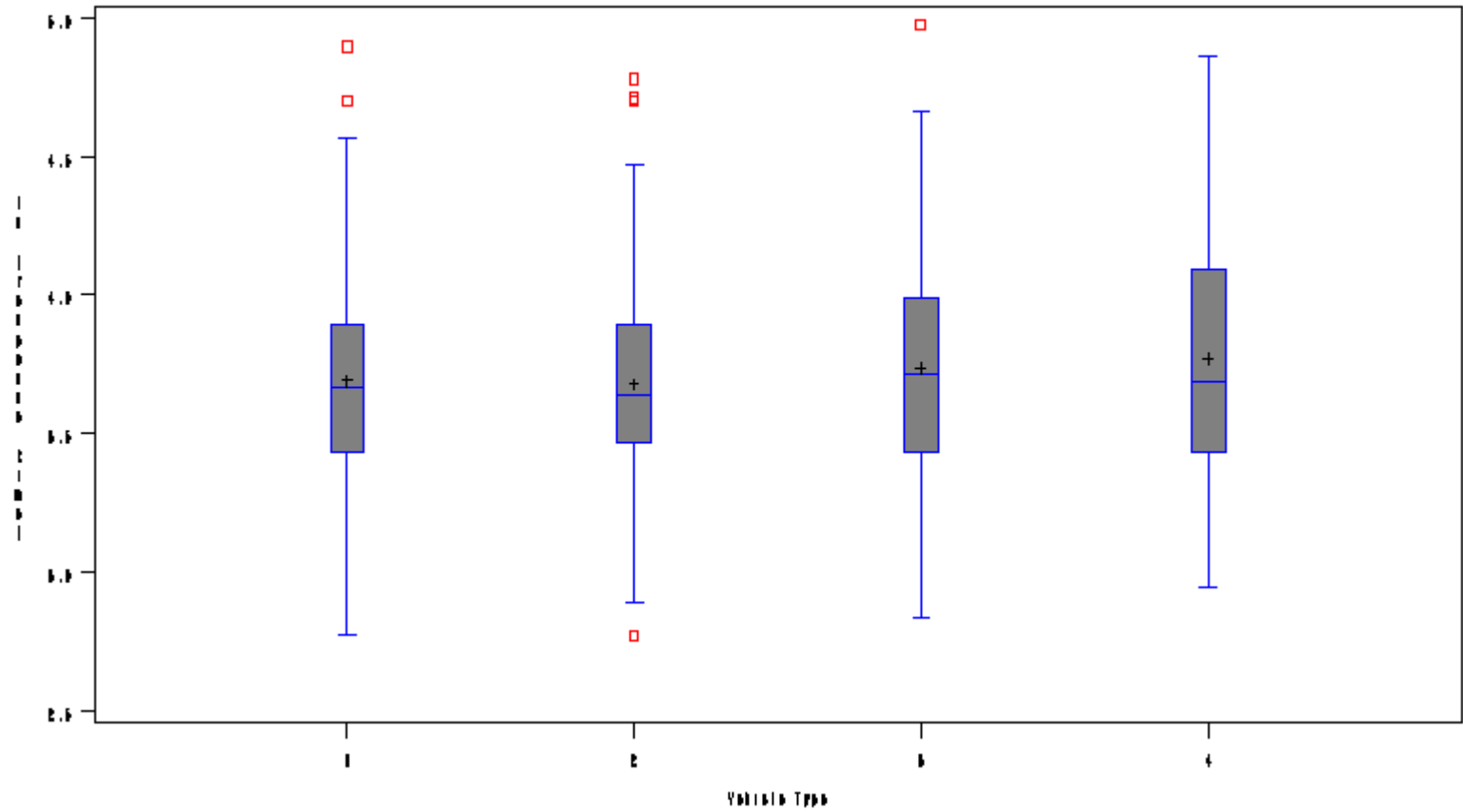
Dependent Variable: t In (response time)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.0313778	0.3437926	2.44	0.0633
Error	729	102.7458858	0.1409409		
Corrected Total	732	103.7772636			

R-Square	Coeff Var	Root MSE	t Mean
0.009938	10.09838	0.375421	3.717634

Source	DF	Anova SS	Mean Square	F Value	Pr > F
vehtype	3	1.03137782	0.34379261	2.44	0.0633

ANOVA for the log transformed response time



회귀분석 (Regression Analysis)

연속형 설명변수가 연속형 종속
변수에 미치는 영향을 분석

회귀 계수의 의미

단순회귀 : X 1단위 증가시 Y 증가분의 기대치

$$\text{Let } Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E(Y | X = x+1) = \beta_0 + \beta_1(x+1)$$

$$- E(Y | X = x) = \beta_0 + \beta_1 x$$

$$= \beta_1$$

$$\beta_0 = E(Y | X = 0)$$

단순회귀와 중회귀에서 회귀 계수들의 의미 차이

중회귀: 다른 X 들이 일정한 값으로 남아있을 때
관심 X 가 1단위 증가 시 Y 의 기대치의 증가분

$$\text{Let } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$E(Y | X_1 = x_1 + 1, X_2 = x_2) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2$$

$$- E(Y | X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$= \beta_1$$

한편 $Y = \beta_0 + \beta_1^* X_1 + \varepsilon$ 이라면

β_1^* 은 X_2 와는 아무런 관계없이 X_1, Y 의 그림에서의 기울기이다.

만약 β_1 과 β_1^* 이 다르다면 X_1 의 효과를 보는데 있어서 X_2 를 고려하느냐 마느냐 하는 것에 따라서 결론이 다르게 된다.

이러한 경우 X_2 를 혼란변수(confounder)하고 한다.

→ 혼란변수를 고려하지 않은 모형에서의 결론은 올바른 결론이라 할 수 없다.

→ 연구 설계 시부터 혼란변수로 작용할 수 있는 모든 변수들을 고려해야 한다.

가상적 예제

Y =수축기 혈압 X_1 =고혈압 여부, X_2 =연령

$$Y = \beta_0 + \beta_2^* X_2 + \varepsilon$$

의 모형에서 β_2^* 는 고혈압 여부와 상관없이 단순히 자료에서 연령이 증가함에 따라 혈압이 얼마나 증가하는가를 나타내고 있다. 하지만

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

의 모형에서 β_2 는 가상의 사람이 고혈압 상태가 같다고 할 때의 연령과 혈압과의 관계를 나타낸다.

어떠한 모수에 우리가 더 관심이 있는가 ?

중회귀

자료

Y : 가축시장을 운영하는 비용 (COST)

\underline{X} : 각 가축의 수

CATTLE

CALVES

HOGS

SHEEP

모형

$$COST = \beta_0 + \beta_1(CATTLE) + \beta_2(CALVES) + \beta_3(HOGS) + \beta_4(SHEEP) + \varepsilon,$$

$$\varepsilon \sim \text{iid } N(0, \sigma^2)$$

```
Proc reg data=auction;
```

```
model cost=cattle calves hogs sheep;
```

SAS 시스템

Model: MODEL1

Dependent Variable: cost

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7936.73649	1984.18412	52.31	<.0001
Error	14	531.03865	37.93133		
Corrected Total	18	8467.77514			
Root MSE		6.15884	R-Square	0.9373	
Dependent Mean		35.29342	Adj R-Sq	0.9194	
Coeff Var		17.45040			

Parameter Estimate						
Variable	DF	Estimate	Parameter Error	Standard t Value	Pr > t	
Intercept	1	2.28842	3.38737	0.68	0.5103	
cattle	1	3.21552	0.42215	7.62	<.0001	
calves	1	1.61315	0.85168	1.89	0.0791	
hogs	1	0.81485	0.47074	1.73	0.1054	
sheep	1	0.80258	0.18982	4.23	0.0008	

중회귀방정식을 위한 독립변수의 선정

- Forward selection
- Backward elimination
- Stepwise selection

분석 -> 회귀분석 -> 선형회귀분석 ->
 방법:단계선택

housize	income	aircapac	applidx	family	peak	변수
3.20	34.99	7.00	7.79	4.00	7.52	
1.30	14.16	.50	3.65	4.00	2.35	
4.10	22.96	3.00	5.85	1.00	5.06	
2.30	24.54	5.00	4.97	2.00	5.01	
1.90	20.61	3.00	4.82	6.00	4.51	
1.90	20.68	1.00	4.66	1.00	2.98	
3.30	30.02	6.50	6.05	1.00	6.85	
2.40	26.34	3.50	7.35	4.00	5.83	

선형 회귀분석

종속변수(D): peak

블록 1 / 1

이전(V) 다음(N)

독립변수(I): applidx, family

방법(M): 단계 선택

선택 변수(E):

케이스 설명(C):

WLS 가중값(H):

통계량(S)... 도표(L)... 저장(A)... 옵션(O)...

확인 명령문(P) 재설정(R) 취소 도움말

모형	진입된 변수	제거된 변수	방법
1	income		단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).
2	aircapac		단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).
3	applidx		단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).
4		income	단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).
5	housesize		단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).
6	family		단계선택 (기준: 입력할 F의 확률 <= .050, 제거할 F의 확률 >= .100).

모형 요약

모형	R	R 제공	수정된 R 제공	추정값의 표준오차
1	.930 ^a	.865	.863	.52986
2	.972 ^b	.944	.942	.34401
3	.980 ^c	.960	.958	.29185
4	.979 ^d	.959	.958	.29306
5	.983 ^e	.966	.964	.27075
6	.984 ^f	.969	.966	.26245

- a. 예측값: (상수), income
- b. 예측값: (상수), income, aircapac
- c. 예측값: (상수), income, aircapac, applidx
- d. 예측값: (상수), aircapac, applidx
- e. 예측값: (상수), aircapac, applidx, housesize
- f. 예측값: (상수), aircapac, applidx, housesize, family

a. 종속변수: peak

분산분석^g

모형	제곱합	자유도	평균제곱	F	유의확률
1 선형회귀분석 잔차 합계	104.305 16.284 120.589	1 58 59	104.305 .281	371.518	.000 ^a
2 선형회귀분석 잔차 합계	113.844 6.746 120.589	2 57 59	56.922 .118	480.983	.000 ^b
3 선형회귀분석 잔차 합계	115.819 4.770 120.589	3 56 59	38.606 .085	453.255	.000 ^c
4 선형회귀분석 잔차 합계	115.694 4.895 120.589	2 57 59	57.847 .086	673.550	.000 ^d
5 선형회귀분석 잔차 합계	116.484 4.105 120.589	3 56 59	38.828 .073	529.681	.000 ^e
6 선형회귀분석 잔차 합계	116.801 3.788 120.589	4 55 59	29.200 .069	423.941	.000 ^f

- a. 예측값: (상수), income
- b. 예측값: (상수), income, aircapac
- c. 예측값: (상수), income, aircapac, applidx
- d. 예측값: (상수), aircapac, applidx
- e. 예측값: (상수), aircapac, applidx, housize
- f. 예측값: (상수), aircapac, applidx, housize, family
- g. 종속변수: peak

제외된 변수⁹

모형	진입-베타	t	유의확률	편상관	공선성 통계량	
					공차한계	
1	housesize	.213 ^a	2.327	.024	.295	.259
	aircapac	.500 ^a	8.978	.000	.765	.316
	applidx	.135 ^a	.997	.323	.131	.127
	family	.012 ^a	.254	.800	.034	.993
2	housesize	.188 ^b	3.301	.002	.404	.259
	applidx	.372 ^b	4.816	.000	.541	.118
	family	.016 ^b	.512	.611	.068	.993
3	housesize	.149 ^c	3.009	.004	.376	.251
	family	.033 ^c	1.230	.224	.164	.977
4	housesize	.154 ^d	3.283	.002	.402	.275
	family	.036 ^d	1.335	.187	.176	.985
	income	.116 ^d	1.214	.230	.160	.077
5	family	.053 ^e	2.145	.036	.278	.951
	income	.033 ^e	.354	.725	.048	.071
6	income	.003 ^f	.034	.973	.005	.069

- a. 모형내의 예측값: (상수), income
- b. 모형내의 예측값: (상수), income, aircapac
- c. 모형내의 예측값: (상수), income, aircapac, applidx
- d. 모형내의 예측값: (상수), aircapac, applidx
- e. 모형내의 예측값: (상수), aircapac, applidx, housesize
- f. 모형내의 예측값: (상수), aircapac, applidx, housesize, family
- g. 종속변수: peak

카이제곱 검정

두 이산변수 간의 관련성 검정

카이제곱 검정 (1)

	Satisfied	Not	Total
Drug A	16 (45.2%)	15	31
Drug B	9 (36.4%)	3	22

카이제곱 검정 (2)

두 사건 A와 B가 독립

$$\leftrightarrow P(A \text{ and } B) = P(A) P(B)$$

만약 약제와 반응이 독립이라면 기대값은

	Satisfied	Not	Total
Drug A	$31/53 *$ $25/53 *$ $53=14.6$		31
Drug B			22
Total	25	28	53

카이제곱 검정 (3)

- 카이제곱 통계량은 이 기대치(14.6)와 실제값(16)의 차이의 제곱의 함수이다.
- 카이제곱 통계량이 크다 (작은 p-value) → 기대치와 실제값이 다르다 → 기대치를 계산하기 위한 가정 (귀무가설: 두 변수가 독립이다)이 틀리다 → 두 변수간에 상관성이 있다 (약품에 따라 반응이 다르다)는 대립가설을 채택한다.

2 × 2 table

Chi-square statistics

Mantel-Haenszel Chi-square

	1	2	
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	N

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}}$$

Pearson chi-square

$$Q_P = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

```
data respire;
  input treat $ outcome $ count ;
cards;
test      f 40
test      u 20
placebo   f 16
placebo   u 48;
proc freq;
  weight count;
  tables treat*outcome/chisq;
run;
```

SAS 시스템

FREQ 프로시저

treat * outcome 교차표

treat	outcome		
칼럼	백백백백	백백백백	총합
행	f	u	
placebo	16 12.90 25.00 28.57	48 38.71 75.00 70.59	64 51.61
test	40 32.26 66.67 71.43	20 16.13 33.33 29.41	60 48.39
총합	56 45.16	68 54.84	124 100.00

treat * outcome 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	1	21.7087	<.0001
우도비 카이제곱	1	22.3768	<.0001
연속성 수정 카이제곱	1	20.0589	<.0001
Mantel-Haenszel 카이제곱	1	21.5336	<.0001
파이 계수		-0.4184	
분할 계수		0.3860	
크라머의 V		-0.4184	

Fisher의 정확 검정

(1,1) 셀 빈도(F)	16
하단측 p값 Pr <= F	2.838E-06
상단측 p값 Pr >= F	1.0000

테이블 확률 (P)	2.397E-06
양측 p값 Pr <= P	4.754E-06

표본 크기 = 124

```
data severe;
  input treat $ outcome $ count ;
cards;
Test      f 10
Test      u  2
Control   f  2
Control   u  4
;
proc freq order=data;
  tables treat*outcome / chisq
  nocol;
  weight count;
run;
```

SAS 시스템

FREQ 프로시저

treat * outcome 교차표

treat	outcome		
행	백분율	빈도	총합
	f	u	
Test	10 55.56 83.33	2 11.11 16.67	12 66.67
Control	2 11.11 33.33	4 22.22 66.67	6 33.33
총합	12 66.67	6 33.33	18 100.00

treat * outcome 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	1	4.5000	0.0339
우도비 카이제곱	1	4.4629	0.0346
연속성 수정 카이제곱	1	2.5313	0.1116
Mantel-Haenszel 카이제곱	1	4.2500	0.0393
파이 계수		0.5000	
분할 계수		0.4472	
크라머의 V		0.5000	

경고: 셀들의 75%가 5보다 작은 기대도수를 가지고 있습니다.
 카이제곱 검정은 올바르지 않을 수 있습니다.

Fisher의 정확 검정

(1,1) 셀 빈도(F)	10
하단측 p값 Pr <= F	0.9961
상단측 p값 Pr >= F	0.0573
테이블 확률 (P)	0.0533
양측 p값 Pr <= P	0.1070

표본 크기 = 18

Exact Test

Table Cell				
(1,1)	(1,2)	(2,1)	(2,2)	probabilities
12	0	0	6	.0001
11	1	1	5	.0039
10	2	2	4	.0533
9	3	3	3	.2370
8	4	4	2	.4000
7	5	5	1	.2560
6	6	6	0	.0498

Table Probabilities

- One-tailed p-value

$$p = 0.0533 + 0.0039 + 0.0001 = 0.0573$$

- Two-tailed p-value

$$p = 0.0533 + 0.0039 + 0.0001 + 0.0498 = 0.1071$$

Difference in Proportions

$$E\{p_1 - p_2\} = \pi_1 - \pi_2$$

$$v_d = \frac{p_1(1 - p_1)}{n_{1+} - 1} + \frac{p_2(1 - p_2)}{n_{2+} - 1}$$

$$d \pm \left\{ z_{\alpha/2} \sqrt{v_d} + \frac{1}{2} \left(\frac{1}{n_{+1}} + \frac{1}{n_{+2}} \right) \right\}$$

Odds Ratio and Relative Risk

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

$$f = \log \{OR\} = \log \left\{ \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \right\}$$
$$= \log \{p_1 / (1 - p_1)\} - \log \{p_2 / (1 - p_2)\}$$

$$v_f = \left\{ \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right\}$$

$$\exp(f \pm z_{\alpha/2} \sqrt{v_f})$$

$$RR = \frac{p_1}{p_2}$$

$$RR = OR \times \frac{(1 + n_{21}/n_{22})}{(1 + n_{11}/n_{12})}$$

if n_{11} and n_{21} are small relative to n_{12} and n_{22}
rare outcome assumption

	Yes	No	total	Proportion Yes
Group1	n_{11}	n_{12}	n_{1+}	$p_1 = n_{11}/n_{22}$
Group2	n_{21}	n_{22}	n_{2+}	$p_2 = n_{21}/n_{12}$
total	n_{+1}	n_{+2}	N	

```
data stress;  
    input stress $ outcome $ count ;  
    cards ;  
low f 48  
low  u 12  
high f 96  
high u 94  
;  
proc freq order=data;  
    tables stress*outcome / chisq  
    measures nocol nopercnt;  
    weight count;  
run ;
```

stress * outcome 교차표

stress 빈도 행 백분율	outcome		총합
	f	u	
low	48 80.00	12 20.00	60
high	96 50.53	94 49.47	190
총합	144	106	250

stress * outcome 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	1	16.2198	<.0001
우도비 카이제곱	1	17.3520	<.0001
연속성 수정 카이제곱	1	15.0354	0.0001
Mantel-Haenszel 카이제곱	1	16.1549	<.0001
파이 계수		0.2547	
분할 계수		0.2468	

Fisher의 정확 검정

(1,1) 셀 빈도(F) 48
 하단측 p값 Pr <= F 1.0000
 상단측 p값 Pr >= F 3.247E-05

테이블 확률 (P) 2.472E-05
 양측 p값 Pr <= P 4.546E-05

통계량	값	점근표준오차
감마	0.5932	0.1147
Kendall의 타우-b	0.2547	0.0551
Stuart 타우-c	0.2150	0.0489
Somers D C R	0.2947	0.0631
Somers D R C	0.2201	0.0499
Pearson 상관계수	0.2547	0.0551
Spearman 상관계수	0.2547	0.0551
람다 비대칭 C R	0.0000	0.0000
람다 비대칭 R C	0.0000	0.0000
람다 대칭	0.0000	0.0000
불확실 계수 C R	0.0509	0.0231
불확실 계수 R C	0.0630	0.0282
불확실 계수 대칭	0.0563	0.0253

상대위험도의 추정값(행1/행2)

연구 유형	값	95% 신뢰한계	
사례대조연구 (오즈비)	3.9167	1.9575	7.8366
코호트 (칼럼1 리스크)	1.5833	1.3104	1.9131
코호트 (칼럼2 리스크)	0.4043	0.2389	0.6841

표본 크기 = 250

```
data respire;
    input treat $ outcome $ count ;
    cards;
test      yes      29
test      no       16
placebo   yes      14
placebo   no       31
;
proc freq order=data;
    tables treat*outcome / measures
    chisq nocol nopercnt;
    weight count;
run ;
```

FREQ 프로시저

treat * outcome 교차표

treat 행	outcome		총합
	빈도 백분율 yes	no	
test	29 64.44	16 35.56	45
placebo	14 31.11	31 68.89	45
총합	43	47	90

상대위험도의 추정값(행1/행2)

연구 유형	값	95% 신뢰한계	
사례대조연구 (오즈비)	4.0134	1.6680	9.6564
코호트 (칼럼1 리스크)	2.0714	1.2742	3.3675
코호트 (칼럼2 리스크)	0.5161	0.3325	0.8011

표본 크기 = 90

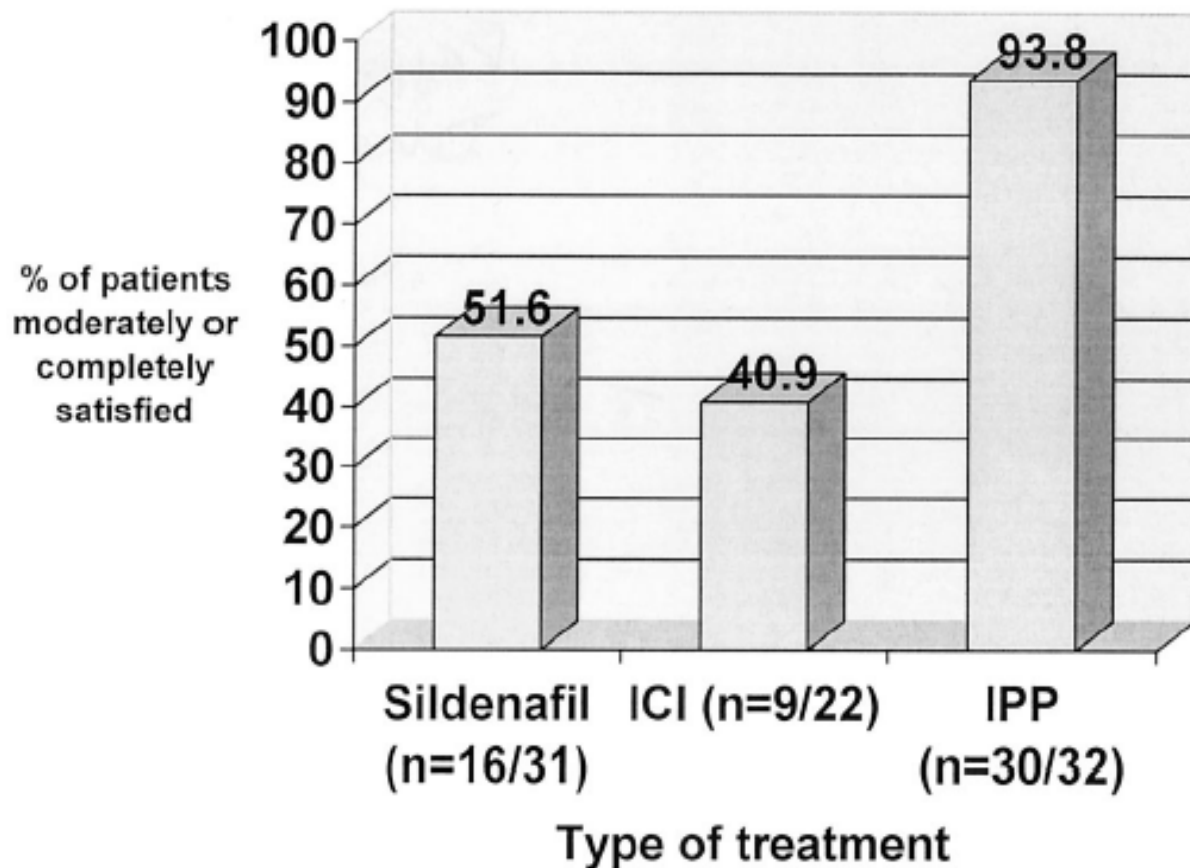


FIG. 3. In response to EDITS questionnaire question 1 on overall satisfaction with treatment 30 of 32 patients (93.75%) who underwent IPP were moderately or completely satisfied (response 3 or 4) compared with 16 of 31 (51.61%) on sildenafil and 9 of 22 (40.91%) on ICI.

Table 1. Correlation of US Grade with Histologic Grade According to Total Fat Content *

US Grade	Total Fat Content					Total
	Normal (0)**	Minimal (≤10)	Mild (11-30)	Moderate (31-60)	Severe (>60)	
Normal	0 (0)	25 (27)	23 (25)	12 (13)	3 (3)	63 (68)
Mild	0 (0)	5 (5)	13 (14)	5 (5)	0 (0)	23 (25)
Moderate	0 (0)	0 (0)	0 (0)	3 (3)	4 (4)	7 (7)
Total	0 (0)	30 (32)	36 (39)	20 (22)	7 (8)	93 (100)

US grade of fatty liver correlates significantly well with percent total fat-containing hepatocytes (X^2 : $p < .001$, linear by linear association: $p < .001$).

*Total fat content = the percentage of both macrovesicular and microvesicular fat-containing hepatocytes.

**Numbers in parentheses are percentage.

McNemar Test : Matched pairs

	Response 1		
Response 2	Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

The question is whether

$$p_1 = \frac{n_{+1}}{n} \quad \text{and} \quad p_2 = \frac{n_{1+}}{n} \quad \text{are the same.}$$

$$Q_M = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \sim \chi_1^2$$

	yes	no	Total
yes	20	5	25
	44.44	11.11	55.56
	80.00	20.00	
	66.67	33.33	
no	10	10	20
	22.22	22.22	44.44
	50.00	50.00	
	33.33	66.67	
Total	30	15	45
	66.67	33.33	100.00

Statistics for Table of hus_resp by wif_resp

McNemar's Test	
Statistic (S)	1.6667
DF	1
<u>Pr > S</u>	<u>0.1967</u>

“Ho : husband and wife
의 approval rates는 같다”
를 기각하지 못함.

Simple Kappa Coefficient

Kappa	0.3077	
ASE	0.1402	
95% Lower Conf Bound		0.0329
95% Upper Conf Bound		0.5825

신뢰구간이 0을 포함하지 않으므로 =0 이라는 귀무가설을 95% 신뢰수준에서 기각한다.

Sample Size = 45

- Kappa=1 >> perfect agreement,
- Kappa > 0.8 >> excellent agreement
- Kappa > 0.4 >> moderate agreement

Logistic Regression

로지스틱 회귀분석

설명변수 (연속, 혹은 이산)가
이산형 종속변수에 미치는 영
향 분석

$Y = 1$ for disease $X_1 = 1$ for male $X_2 = \text{age}$
 0 for non-disease 0 for female

불연속 연속

$p(Y = 1) \propto \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (linear predictor)

$\log \frac{p(Y = 1)}{1 - p(Y = 1)} \rightarrow$ logit link function

$\text{logit}[p(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$$\begin{aligned}
 p(Y = 1) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)} \\
 &= \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2)}
 \end{aligned}$$

$$\beta_0 = \log \left(\frac{P(Y=1 | X_1 = X_2 = 0)}{1 - P(Y=1 | X_1 = X_2 = 0)} \right), X_1 = X_2 = 0 \text{ 일 때의 log odds 값}$$

$$\log \left(\frac{P(Y=1 | X_1 = 1, X_2 = x)}{1 - P(Y=1 | X_1 = 1, X_2 = x)} \right) = \beta_0 + \beta_1 + \beta_2 x$$

$$\log \left(\frac{P(Y=1 | X_1 = 0, X_2 = x)}{1 - P(Y=1 | X_1 = 0, X_2 = x)} \right) = \beta_0 + \beta_2 x$$

β_1 = 연령으로 보정한 후 (연령이 같은 값으로 남아있을 때) 성별

이 남일 때(여에 비하여) 병 걸릴 확률이 log odds ratio의 증가분

$\exp(\beta_1) = \dots\dots\dots$ odds ratio의 증가분

$$\beta_2 = \log \frac{P(Y=1 | X_1=a, X_2=x+1)}{1-P(Y=1 | X_1=a, X_2=x+1)} - \log \frac{P(Y=1 | X_1=a, X_2=x)}{1-P(Y=1 | X_1=a, X_2=x)}$$

: 다른 x들이 일정한 값으로 남아 있을 때(성별이 일정할 때, 성별의 효과를 보정한 후) 연령이 한 단위 증가할 시 병 걸릴 확률의 log odds ratio의 증가분

$\exp(\beta_2)$ =..... odds ratio의 증가분

```
proc logistic data=esr descending;  
  model response=fibrin globulin;  
  title 'ESR Data';  
run;
```

y=0,1 인 경우 default는 작은 값(0)을 기준으로,
큰 값을 기준으로
하려면 descending option이 필요

The LOGISTIC Procedure

Model Information

Data Set	WORK.ESR
Response Variable	response
Number of Response Levels	2
Number of Observations	32
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	response	Total Frequency
1	1	6
2	0	26

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	32.885	28.971
SC	34.351	33.368
-2 Log L	30.885	22.971

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.9138	2	0.0191
Score	8.2067	2	0.0165
Wald	4.7561	2	0.0927

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-12.7920	5.7964	4.8704	0.0273	
fibrin	1	1.9104	0.9710	3.8708	0.0491	0.6710
globulin	1	0.1558	0.1195	1.6982	0.1925	0.3936

Analysis of Maximum Likelihood Estimates

Variable	Odds Ratio
Intercept	
fibrin	6.756
globulin	1.169

유전자형 자료분석의 기본 개념

Genotype Freq, Allele Freq,

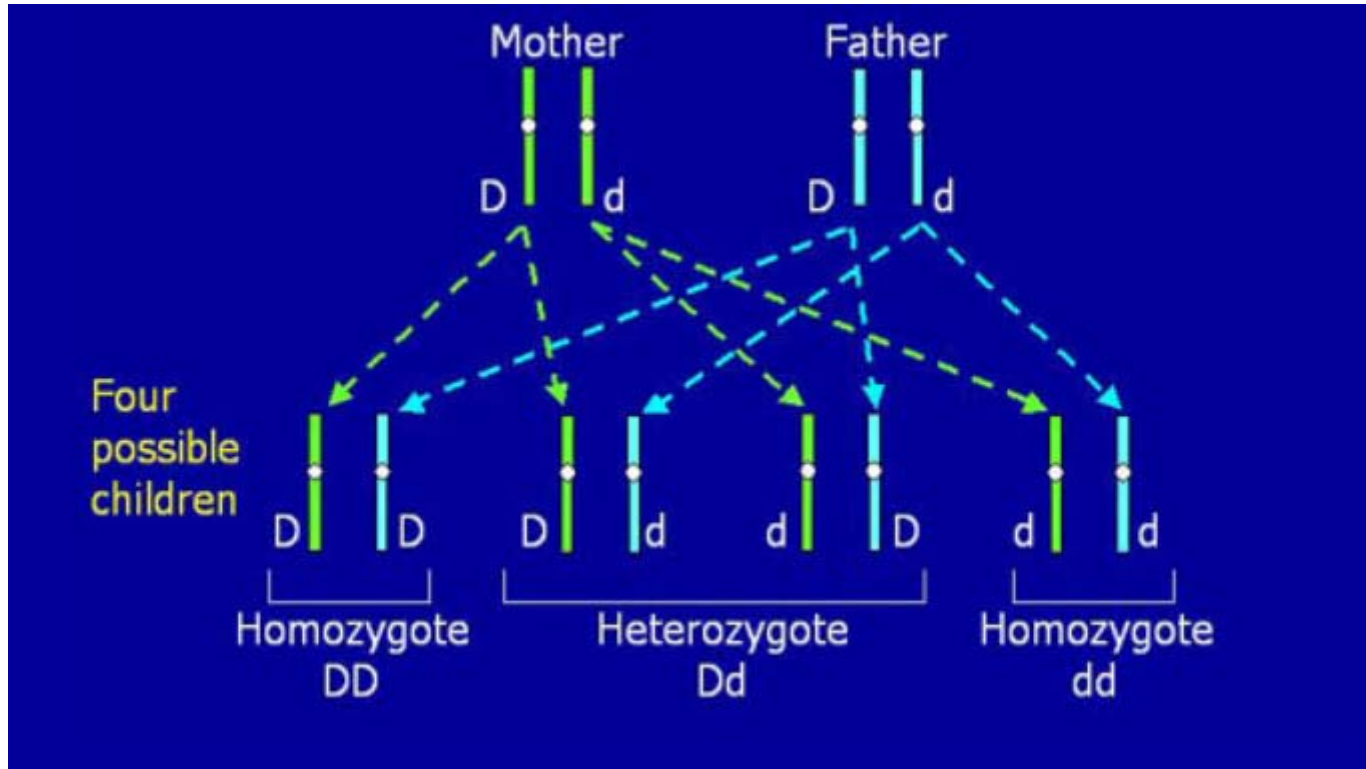
Hardy-Weinberg Disequilibrium,

(Mendelian) Genetic Models,

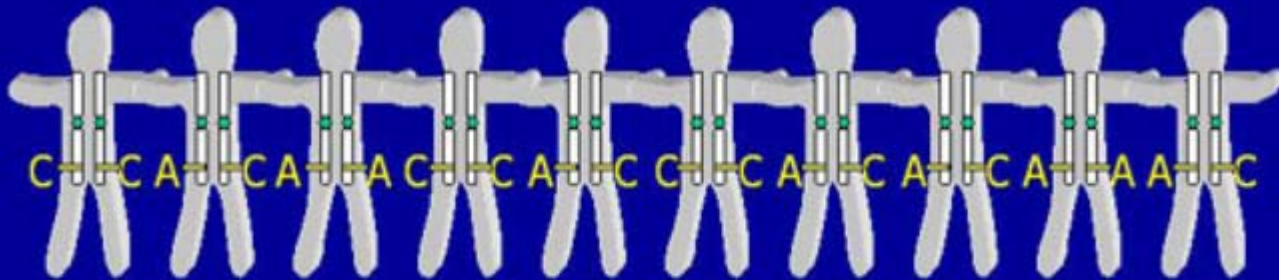
LD, Haplotype, Heritability, SNP Association Study,

Multiple Comparisons

Genotype



Allele frequency

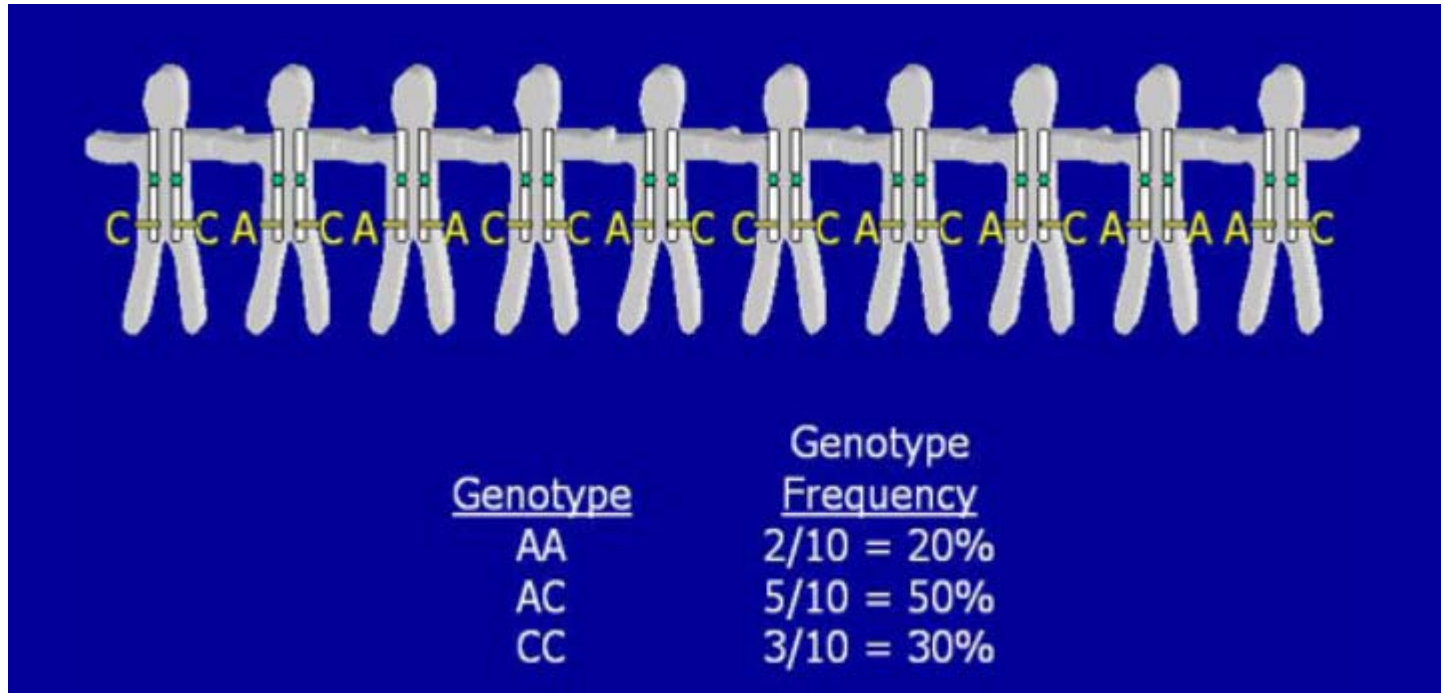


$$\text{Allele Frequency} = \frac{\text{Number of alleles}}{2 \times (\text{number of people})}$$

$$\text{Frequency of A allele} = \frac{9}{20} = .45$$

$$\text{Frequency of C allele} = \frac{11}{20} = .55$$

Genotype frequency



Hardy-Weinberg

In a stable population with random mating, allele frequency predicts genotype frequency.

Goodness-of-fit can be applied to test H-W Equilibrium

probability of **A** allele = P_A

probability of **G** allele = P_G

probability of **AA** genotype = $P_A \times P_A$

probability of **AG** genotype = $2P_A \times P_G$

probability of **GG** genotype = $P_G \times P_G$

		father	
		P_A	P_G
mother	P_A	P_A^2	$P_A P_G$
	P_G	$P_G P_A$	P_G^2

Deviation from this relationship is called Hardy-Weinberg Disequilibrium.

- **Chi-square Test**

Ho: 우리의 자료가 특정모형(HWE)을 따른다

$$\sum \frac{(\text{관찰값} - \text{기대값})^2}{\text{기대값}} \sim \chi_{df}^2$$

자유도 = 범주의 개수 - 1 - 추정된 모수의 수

• HWE 예제 1

	개수		빈도	
	관찰값	기대값	관찰값	기대값
AA	298	294.3063	0.2980	0.2943
Aa	489	496.3875	0.4890	0.4964
aa	213	209.3063	0.2130	0.2093
total	1000	1000	1	1

- $p = (2 \times 298 + 489) / (2 \times 1000) = 0.5425$
 $q = (489 + 2 \times 213) / (2 \times 1000) = 0.4575$
- $P(A) = p^2 = (0.5425)^2 = 0.2943$
 $P(Aa) = 2pq = 2 \times (0.5425) \times (0.4575) = 0.4964$
 $P(aa) = q^2 = (0.4575)^2 = 0.2093$
- 기대 값 (**expected frequency**)
 $AA = P(AA) \times 1000 = 294.3064$
 $Aa = P(Aa) \times 1000 = 496.3875$
 $aa = P(aa) \times 1000 = 209.3063$

- 검정통계량

$$\chi^2 = \sum \frac{(298 - 294.3063)^2}{294.3063} + \frac{(489 - 496.3875)^2}{496.3875} + \frac{(213 - 209.3063)^2}{209.3063}$$
$$= 0.2215$$

$$\text{자유도} = \mathbf{3-1-1=1}$$

∴ 자유도가 **1**인 카이제곱 분포에 근거가 **p**값이 **0.6379**이므로 관찰된 값은 **Ho (HWE 상태)**를 기각할 수 있는 충분한 근거가 없다. 즉 **HWE** 상태라고 결론 내린다.

실무에서는 **genotype error check**의 방법으로 많이 사용된다.

• **Test of association (Odds ratio, Chi-square test)**

	1	2	Total
Case	n_{11}	n_{12}	n_{1+}
Control	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	N

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{(n_{11} / n_{1+}) / (n_{12} / n_{1+})}{(n_{21} / n_{2+}) / (n_{22} / n_{2+})} = \frac{n_{11} / n_{12}}{n_{21} / n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Chi-square test with $df=(\#col-1)(\#row-1)$:

Ho: OR=1

Expected cell freq is bigger than 5, if not use Fisher's Exact test

- Chi-square 예제 : Genotype-based (Co-dominant model)**

	MM	Mm	mm	Total
Case	n_{2A}	n_{1A}	n_{0A}	n_{+A}
Control	n_{2O}	n_{1O}	n_{0O}	n_{+O}
Total	n_{2+}	n_{1+}	n_{0+}	N

$$OR_{MM/mm} = \frac{n_{2A}n_{0O}}{n_{2O}n_{0A}}$$

Co-dominant model

$$OR_{Mm/mm} = \frac{n_{1A}n_{0O}}{n_{1O}n_{0A}}$$

MM Mm mm 간의 관계를 가정하지 않음

자유도 2인 검정

- Chi-square 예제 : Genotype-based (Dominant Model)**

	MM or Mm	mm	Total
Case	$n_{2A} + n_{1A}$	n_{0A}	n_{+A}
Control	$n_{2O} + n_{1O}$	n_{0O}	n_{+O}
Total	$n_{2+} + n_{1+}$	n_{0+}	N

OR_{MM or Mm /mm}

Dominant model
 (MM = Mm) > mm
 자유도 1인 검정

- Chi-square 예제 : Genotype-based (Recessive model)**

	MM	Mm or mm	Total
Case	n_{2A}	$n_{1A} + n_{0A}$	n_{+A}
Control	n_{2O}	$n_{1O} + n_{0O}$	n_{+O}
Total	n_{2+}	$n_{1+} + n_{0+}$	N

OR_{MM/Mm or mm}

Recessive model

MM > (Mm=mm)

자유도 1인 검정

- Chi-square 예제 : Genotype-based (Additive Model)**

	MM	Mm	mm	Total
Case	n_{2A}	n_{1A}	n_{0A}	n_{+A}
Control	n_{2O}	n_{1O}	n_{0O}	n_{+O}
Total	n_{2+}	n_{1+}	n_{0+}	N

$$OR_{MM/Mm} = OR_{Mm/mm}$$

$$OR_{MM/mm} = 2 OR_{Mm/mm} = 2 OR_{MM/Mm}$$

Additive model

$(MM - Mm) = (Mm - mm)$ Dose-Response 가정

자유도 1인 검정

Linkage Disequilibrium

Alleles at different sites should occur in a combinations relative to their SNP allele freq

SNP 1

Probability of **A** allele = P_A

Probability of **G** allele = P_G

SNP 2

Probability of **T** allele = P_T

Probability of **C** allele = P_C

Probability of **AT** haplotype = $P_A P_T$

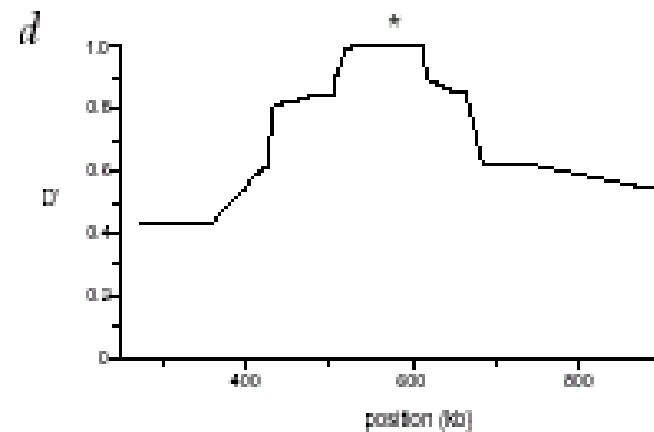
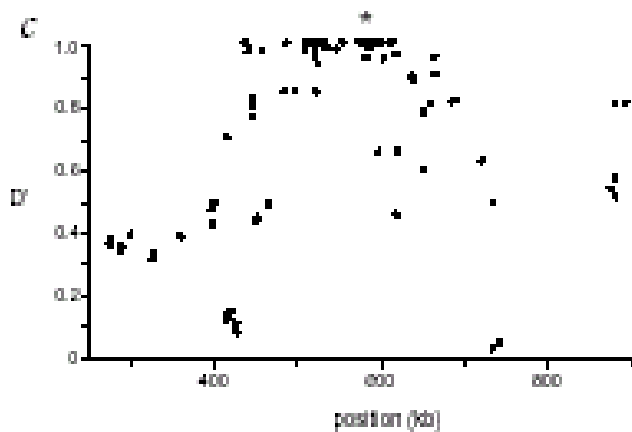
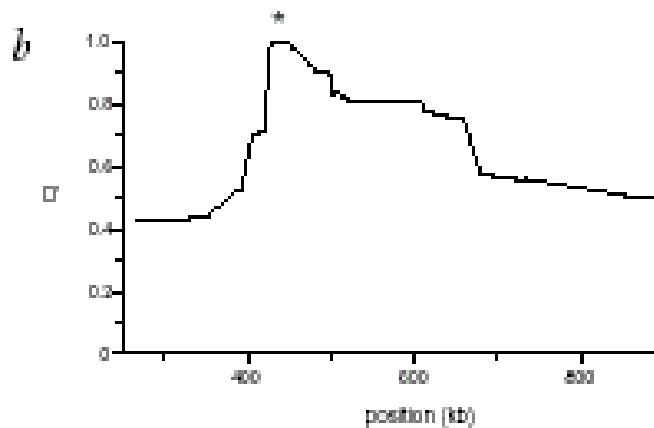
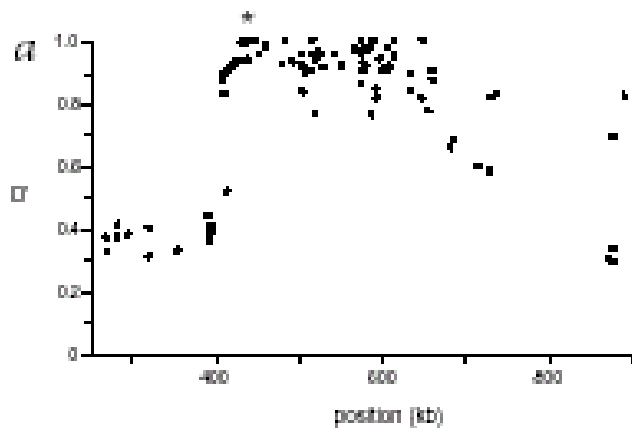
Probability of **AC** haplotype = $P_A P_C$

Probability of **GT** haplotype = $P_G P_T$

Probability of **GC** haplotype = $P_G P_C$

Deviations from this relationship indicates linkage disequilibrium.

LD Block



Linkage Disequilibrium (D')

Marker 1

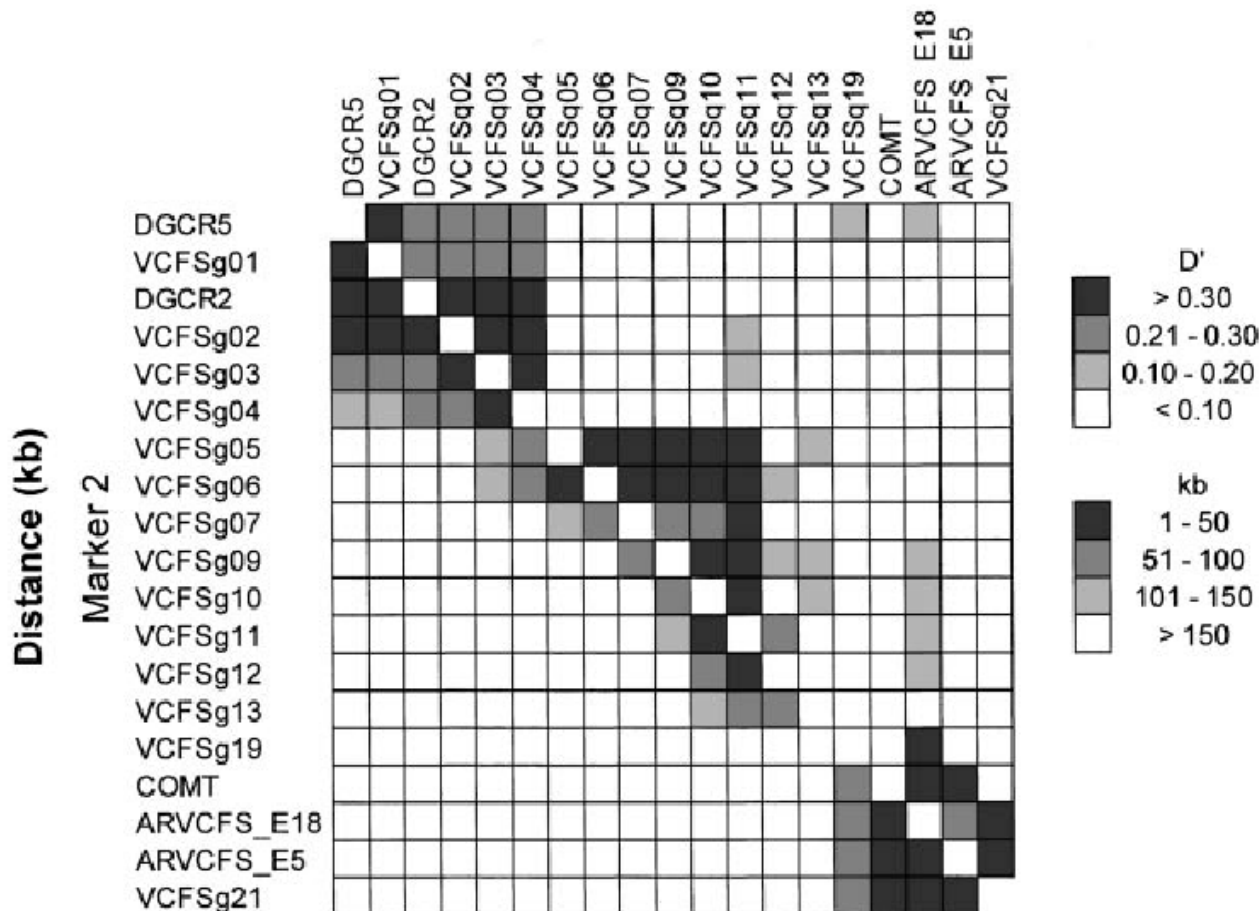
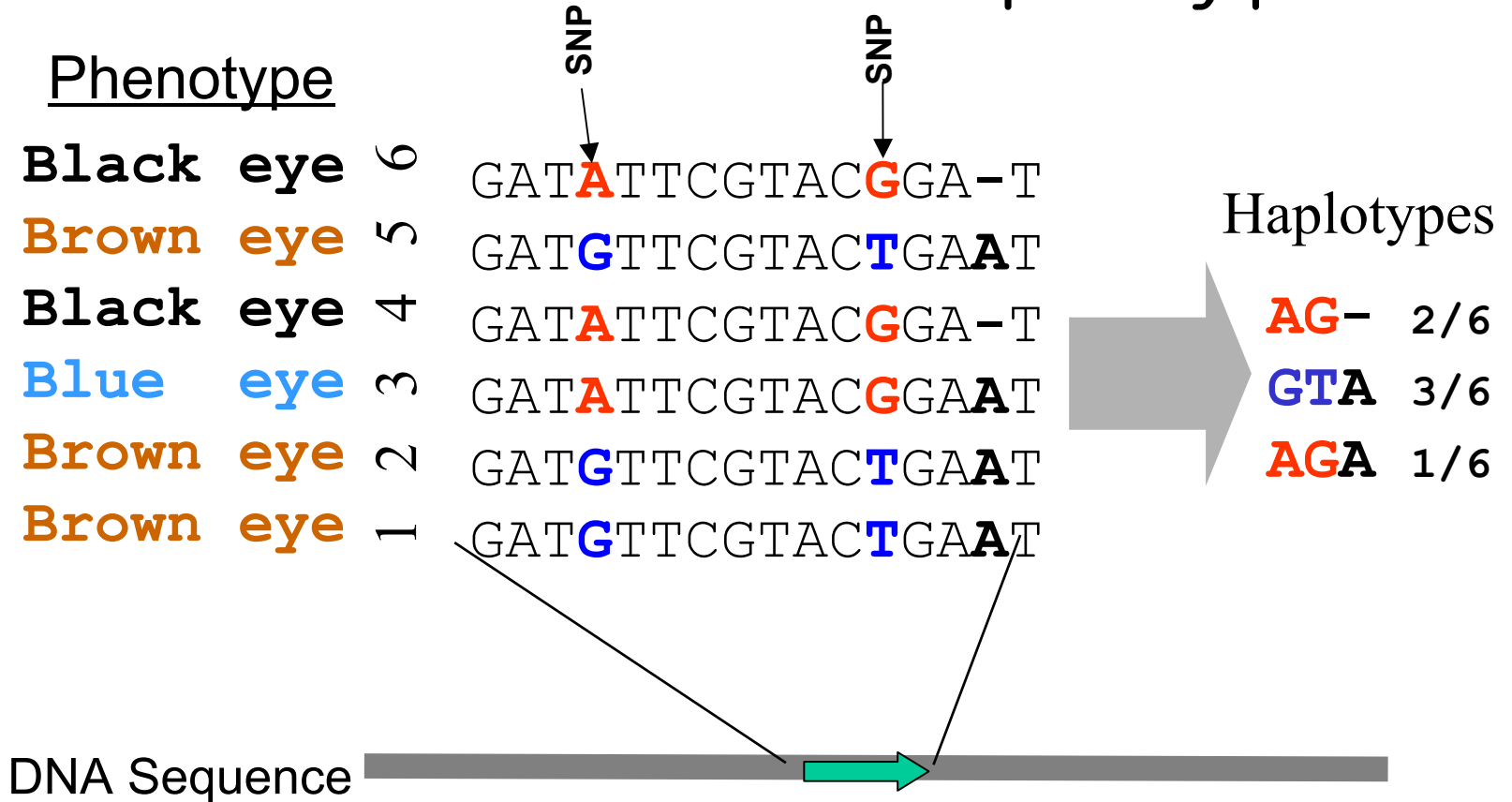


Fig. 4. Matrix of linkage disequilibrium and physical distance interval between all 171 marker pair combinations. Markers are arranged in map order as shown in Figure 2. Marker to marker LD as measured by D' is shown above the diagonal. Dark squares denote $D' > 0.30$, medium squares denote D' between 0.21–0.30, light squares denote D' between 0.10–0.20,

and white squares denote $D' < 0.10$. Markers to marker physical distance intervals in kb are shown below the diagonal. Dark squares denote interval distance between 1–50 kb, medium squares denote interval distance between 51–100 kb, light squares denote interval distance between 101–150 kb, and white squares denote interval distance greater than 150 kb.

From SNP to Haplotype



Association study using haplotype

Hap	AG-	GTA	AGA	Total
Case				
Control				
Total				2N

Hap Pair	AG-/AG-	AG-/GTA	AG-/AGA	AGA/AGA	Total
Case					
Control					
Total					N

만약 AGA가 risk hap

Hap Pair	Else	Else/AGA	AGA/AGA	Total
Case				
Control				
Total				N

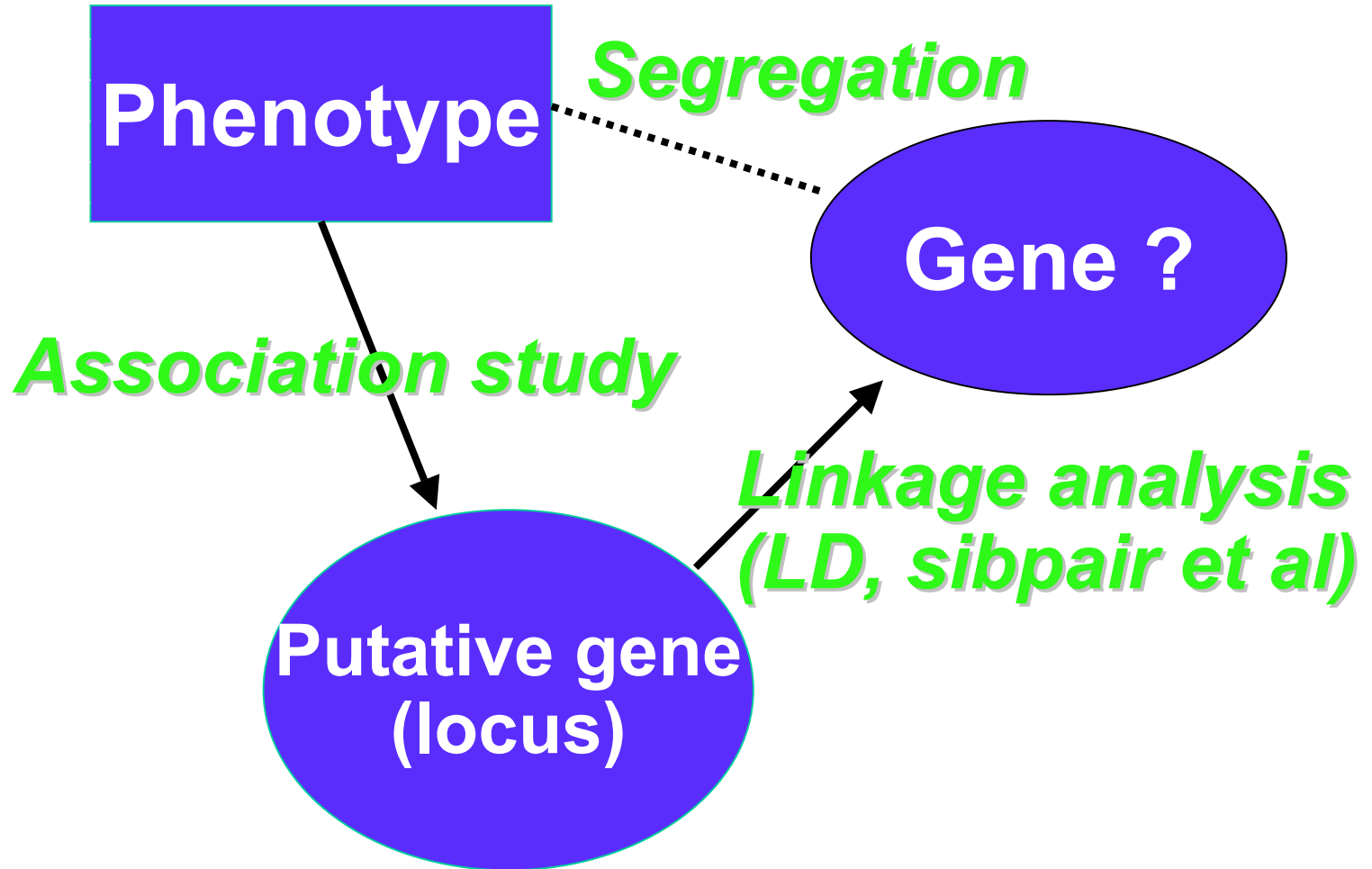
만약 AGA가 risk hap이고 Dominant Model을 적용한다면

Hap Pair	Else	Else/AGA or AGA/AGA	Total
Case			
Control			
Total			N

How to identify the genes

- Family study
 - Linkage analysis: pedigree 필요
 - Sib pair analysis: oligogenic, multigenic
- Population study
 - Case-control association study

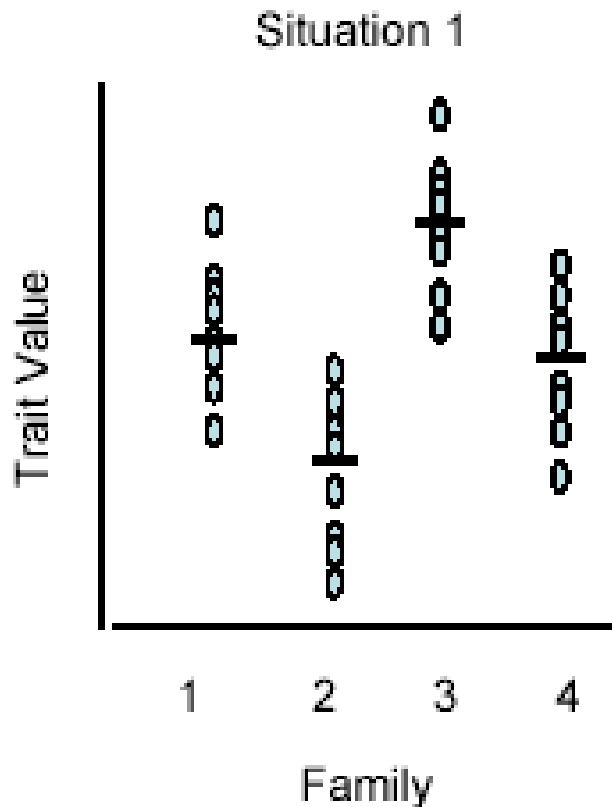
New Gene Discovery



Heritability

형질(Trait)	유전율 (%)	형질(Trait)	유전율 (%)
수명	29	언어능력	63
키	85	최대맥박수	84
몸무게	63	계산능력	76
아미노산 분비	72	기억력	47
혈중지질 농도	44	사회적응력	66
혈중최대 젖산 농도	34	감성	58

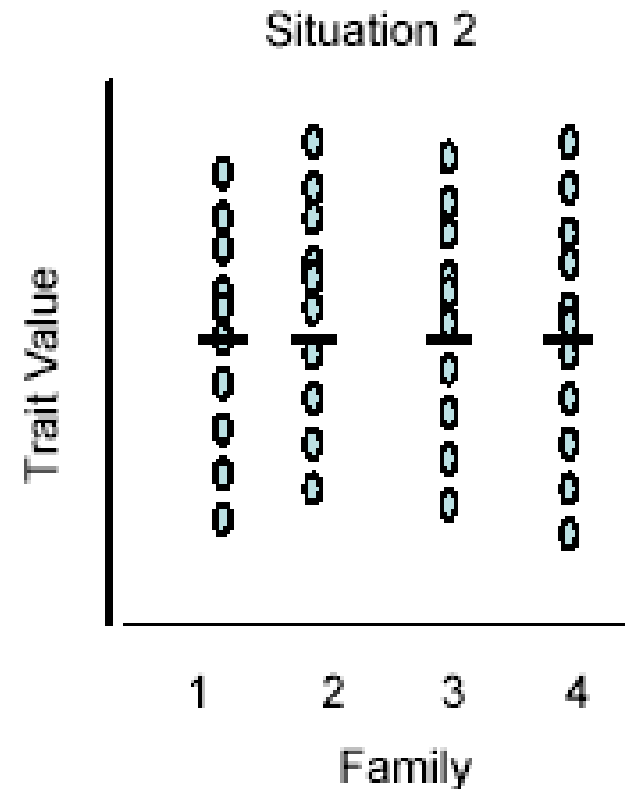
Which Trait has Higher Heritability?



$$\text{Var}(B)=2.5$$

$$\text{Var}(W)=.5$$

$$\text{Vp}=\text{Var}(B)+\text{Var}(W)=3$$



$$\text{Var}(B)=0$$

$$\text{Var}(W)=3$$

$$\text{Vp}=\text{Var}(B)+\text{Var}(W)=3$$

Partitioning Variation

Between vs within Family

Case 1

$$\frac{Var(B)}{V_P} = \frac{2.5}{3} = .83$$

Case 2

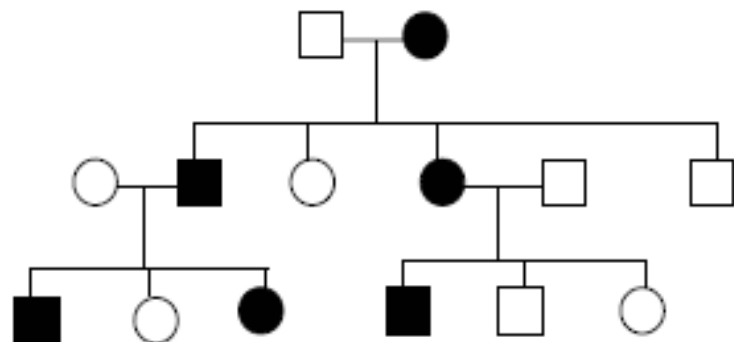
$$\frac{Var(B)}{V_P} = \frac{0}{3} = 0$$

Which of these situations is indicative of a high genetic component?

Var(B)=variation among group, group are full sib groups

Var(B)=Cov(full sibs)

Linkage Analysis
(family)



Single Gene
Mendelian Inheritance
Rare (High Penetrance)
 ~300-400 Short Tandem Repeat (MS) Markers

Association Studies*
(population)

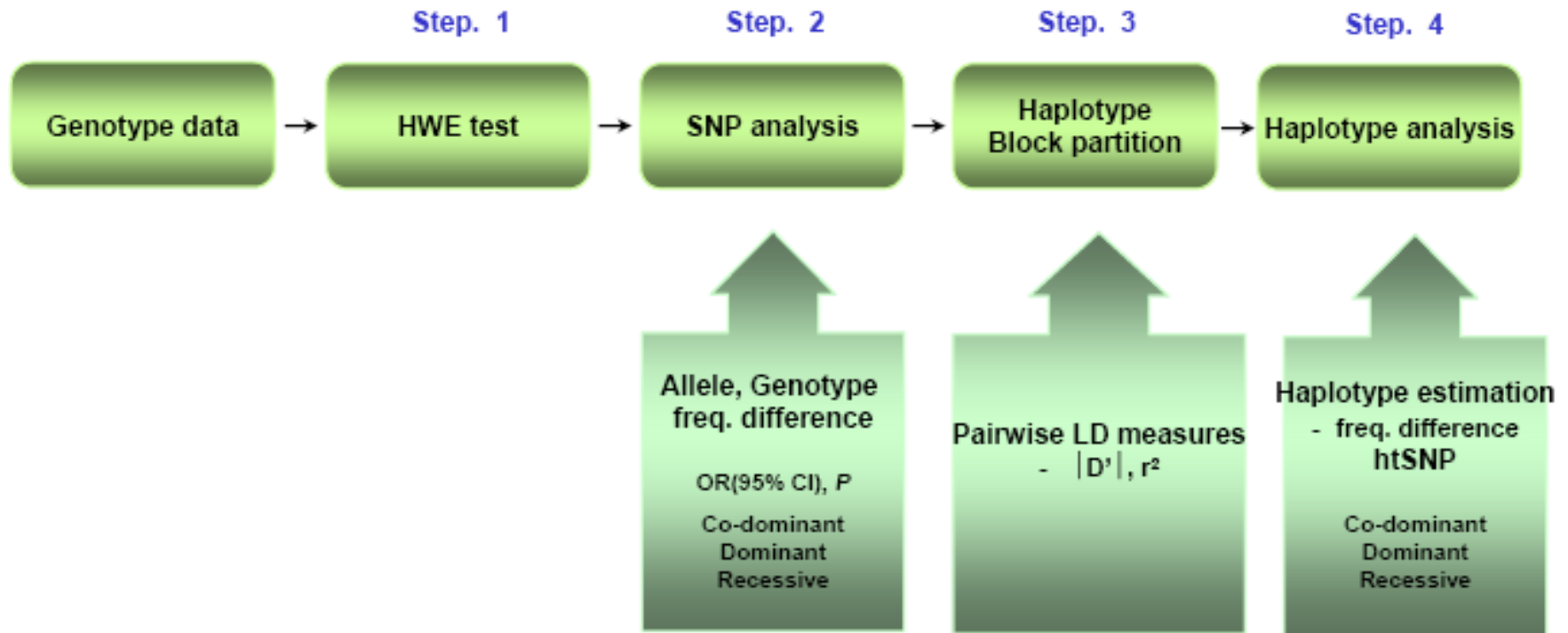
Cases	Controls
40% T 60% C	15% T 85% C

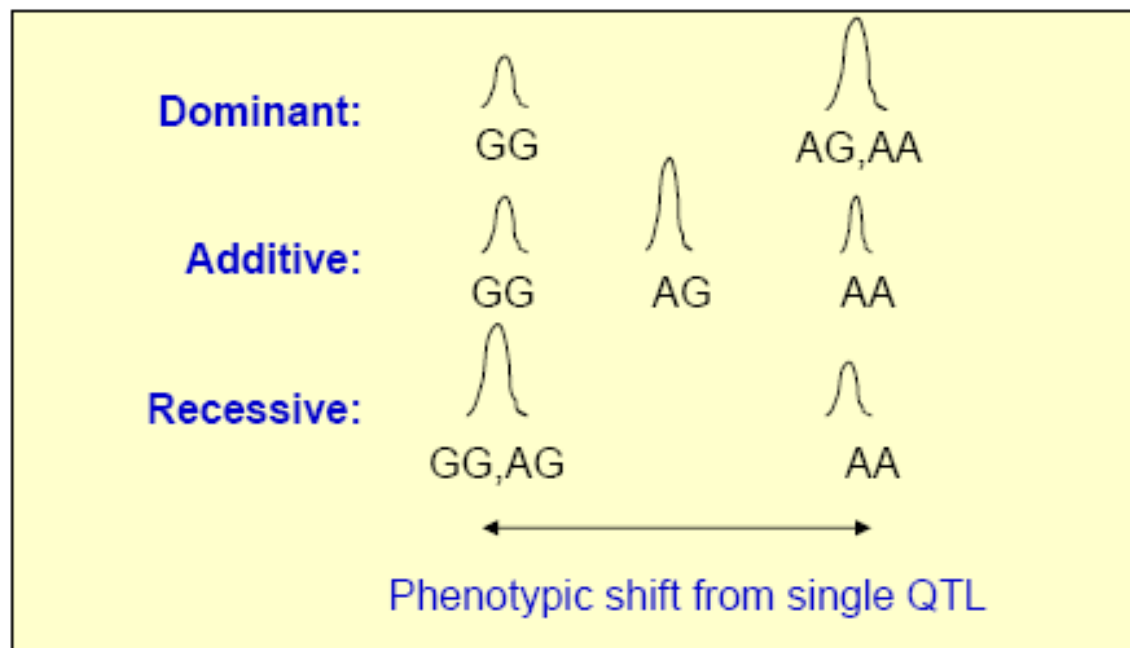
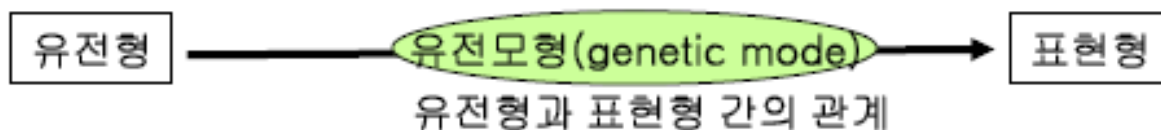
Polygenic (also G x E)
Complex Inheritance
Common
 ~30,000 to 50,000 Polymorphic SNP Markers
 * Example: APOE4 and Alzheimer's Disease

SNP Association Study

1. Study design
 1. Select target disease
 2. Case-control criteria
 3. Determine # of samples
2. Sample and Data Collection
 1. Genetic materials
 2. Clinical information/phenotypic classification
 3. Environmental Information
3. Genotyping
 1. Select candidate genes/SNP
 2. Whole genome screening
 3. Select appropriate method of genotyping
4. Statistical Analysis

- Statistical analysis scheme of SNP Genotyping Data





유전모형별 유전형 재구성 방법

Genotype	Dominant	Additive	Recessive
AA	2	3	2
AG	2	2	1
GG	1	1	1

Multiple Comparisons (다중비교)

ex) 한 test 에서 유의수준이 α 인 test가 있다고 하자.

$$\text{Let } H_{01} : \alpha_1 = 0, \quad \Pr(\text{do not reject } H_{01} | H_{01} \text{ is true}) = 1 - \alpha$$

$$H_{02} : \alpha_2 = 0, \quad \Pr(\text{do not reject } H_{02} | H_{02} \text{ is true}) = 1 - \alpha$$

then $\Pr(\text{do not reject } H_0 | H_0)$ where $H_0 = H_{01}$ and H_{02}

$$= \Pr(\text{do not reject } H_{01} \text{ and do not reject } H_{02} | H_0)$$

$$= (1 - \alpha) (1 - \alpha) = (1 - \alpha)^2$$

일반적으로 $\alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_k = 0$ 를

multiple comparison을 한다면 $(1 - \alpha)^k \leq (1 - \alpha)$

$$1 - 0.1855 = 0.8145 = (.95)^4 \leq .95$$

\therefore overall α 는 0.05가 아니라 0.1855가 되므로 type I error가 Inflation 되었다.

Multiple Comparisons

Bonferroni Correction : 만약 m 개의 multiple comparison을 한다면 각각의 유의수준을 α/m 로 하면 전체의 유의수준을 α 에 가깝게 할 수 있다.

예) m 이 4인 경우 $(1 - \frac{0.05}{4})^4 \cong 0.95 = 1 - 0.05$

응용) 10개의 mean을 비교하는 경우

p값의 기준을 0.05로 하면 overall p값을 유지할 수 없으므로 각각의 경우 $\frac{0.05}{10} = 0.005$ 를 기준으로 test를 실시한다.

이를 “Bonferroni corrected p-value”라고 한다.

Multiple Comparisons: FDR

False Discovery Rate

$$\text{FDR} = \text{False Positive} / \text{Total Positive}$$

Edit section: Classification of m

	# declared non-significant	# declared significant	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

m_0 is the number of true [null hypotheses](#)

$m - m_0$ is the number of false [null hypotheses](#)

V is the number of [false positives](#)

T is the number of [false negatives](#)

$H_1 \dots H_m$ the null hypotheses being tested

In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S , T , U , and V are all unobservable [random variables](#).

The false discovery rate is given by $E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$ and one wants to keep this value below a threshold α .

Multiple Comparisons: FDR (independent test)

Benjamini and Hochberg (1995)

1. Order p-values by $P_{(1)}, P_{(2)}, \dots, P_{(m)}$

2. Find the largest k such that

$$P_{(k)} \leq \frac{k}{m} \alpha.$$

3. $1, 2, \dots, k$ 까지는 유의하다.

(예) $m=500K$,

$0.05/500K = 10^{-7}$: Bonferroni correction

$0.05/500K * 2$,

$0.05/500K * 3 \dots\dots$ 해서

$P_{(2000)} < 2000 * 10^{-7}$ 이고 $P_{(2001)} > 2001 * 10^{-7}$ 이라면 2000개 뽑는다.

Multiple Comparisons: FDR (dependent test)

Benjamini and Yekutieli (2001)

1. Order p-values by $P_{(1)}, P_{(2)}, \dots, P_{(m)}$
2. Find the largest k such that

$$P_{(k)} \leq \frac{k}{m \cdot c(m)} \alpha$$

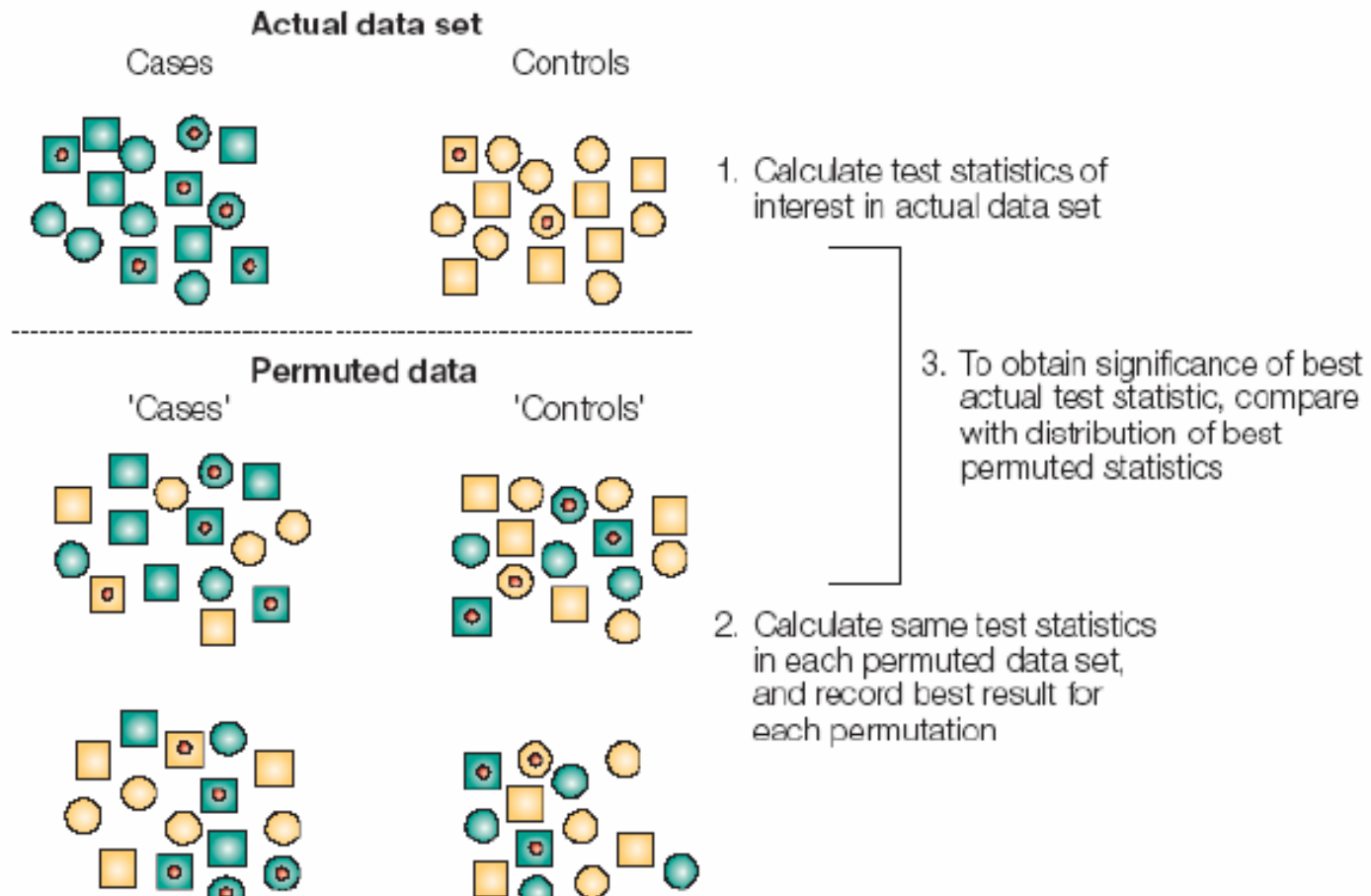
3. $1, 2, \dots, k$ 까지는 유의하다.

If tests are indep or positively correlated then

$$c(m) = 1$$

If tests are negatively correlated then $c(m) = \sum_{i=1}^m \frac{1}{i}$

Permutation test



Statistical Models for SNP Association Study

Response Var	Group	Statistical Methods
연속변수 (BMI, BP, etc)	2 groups 2 groups (N<5 per group) 3 groups or more 보정변수	T-test Wilcoxon test ANOVA ANCOVA, regression
이항변수 (case-control)	2 groups 2 groups (N<5 per group) 보정변수	Chi-square test Fisher's Exact test Logistic regression



ELSEVIER

Association of *UCP2* and *UCP3* gene polymorphisms with serum high-density lipoprotein cholesterol among Korean women

Min Ho Cha^a, Il Chul Kim^b, Kil Soo Kim^c, Byoung Kab Kang^a, Sun Mi Choi^a, Yoosik Yoon^{d,*}

^aDepartment of Medical Research, Korea Institute of Oriental Medicine, Daejeon 305-811, Korea

^bDepartment of Biology (the Second Stage BK21 Program), Chonnam National University, Gwangju 500-757, Korea

^cObesity Clinic, Kirin Oriental Hospital, Seoul 137-905, Korea

^dDepartment of Microbiology, School of Medicine, Chung-Ang University, Seoul 156-756, Korea

Received 6 December 2006; accepted 8 January 2007

Table 1

General characteristics of study subjects

Variables	n	Mean \pm SD
Age (y)	658	27.72 \pm 7.48
Weight (kg)	658	66.42 \pm 10.89
BMI (kg/m ²)	658	25.69 \pm 3.99
SBP	607	115.08 \pm 12.71
DBP	608	72.20 \pm 10.06

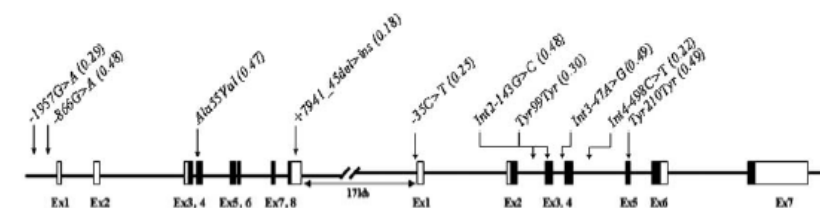
SBP indicates systolic blood pressure; DBP, diastolic blood pressure.

Table 2

Polymorphisms of *UCP2* and *UCP3* genes used for genotyping of the subjects

Gene	Locus	Position	Locus from start codon	Minor allele frequency	HWE	rs no.
<i>UCP2</i>	-1957G>A	Promoter	-6422G>A	0.29	0.80	rs649446
	-866G>A	Promoter	-5331G>A	0.48	0.16	rs659366
	Ala55Val	Exon 4	+320C>T	0.47	0.05	rs660339
	+7941_45del>ins	3' UTR	+3473_45del>ins	0.18	0.08	-
<i>UCP3</i>	-35C>T	Promoter	-2078C>T	0.25	0.99	rs1800849
	Int2-143G>C	Intron 2	+521G>C	0.48	0.13	rs2075576
	Tyr99Tyr	Exon 3	+834C>T	0.30	0.99	rs1800006
	Int3-47A>G	Intron 3	+1063A>G	0.49	0.37	rs1685325
	Int4-498C>T	Intron 4	+1811C>T	0.22	0.74	rs2734827
	Tyr210Tyr	Exon 5	+2546C>T	0.49	0.16	rs2075577

HWE indicates Hardy-Weinberg equilibrium; rs, reference no. of each SNP for NCBI's database; UTR, untranslated region.

Map of *UCP2* and *UCP3* on chromosome 11q13

A

LDs among *UCP2* and *UCP3* polymorphisms

		r ²									
		-1957G>A	-866G>A	Ala55Val	+7941_45del>ins	-35C>T	Int2-143G>C	Tyr99Tyr	Int3-47A>G	Int4-498C>T	Tyr210Tyr
	-1957G>A	-	1	1	1	0.83	0.84	0.84	0.84	0.91	0.94
	-866G>A	0.44	-	0.99	0.98	0.84	0.83	0.84	0.83	0.55	0.93
	Ala55Val	0.45	0.97	-	0.99	0.84	0.83	0.84	0.83	0.54	0.93
	+7941_45del>ins	0.10	0.23	0.24	-	0.81	0.81	0.82	0.80	0.81	0.91
	-35C>T	0.68	0.31	0.31	0.06	-	1.00	1.00	0.99	0.97	0.95
	Int2-143G>C	0.30	0.68	0.66	0.15	0.43	-	0.99	0.99	0.62	0.92
	Tyr99Tyr	0.70	0.32	0.32	0.07	0.97	0.43	-	0.99	0.97	0.95
	Int3-47A>G	0.30	0.67	0.65	0.15	0.42	0.95	0.42	-	0.62	0.90
	Int4-498C>T	0.10	0.09	0.09	0.54	0.11	0.11	0.11	0.11	-	0.62
	Tyr210Tyr	0.37	0.83	0.80	0.19	0.38	0.81	0.38	0.80	0.11	-

B

Haplotypes of *UCP2* and *UCP3*

Hap.	-1957G>A	-866G>A	Ala55Val	+7941_45del>ins	-35C>T	Int2-143G>C	Tyr99Tyr	Int3-47A>G	Int4-498C>T	Tyr210Tyr	Freq.
ht1	G	G	C	del	C	G	T	A	C	C	0.425
ht2	A	A	T	ins	T	C	C	G	C	T	0.256
ht3	G	A	T	ins	C	C	T	G	T	T	0.147
ht4	G	G	C	del	C	G	T	A	T	C	0.032
ht5	G	G	C	del	T	C	C	G	C	T	0.020
ht6	A	A	T	del	C	G	T	A	C	T	0.014
Others ⁽²⁾	-	-	-	-	-	-	-	-	-	-	0.105

C

Fig. 1. Gene maps of the *UCP2* and *UCP3* on chromosome 11q13. Coding exons are marked by black blocks and 5' and 3' UTRs are marked by white blocks. A, Polymorphisms in *UCP2* and *UCP3* genes. The minor allele frequencies shown in parentheses were determined via the genotyping of 658 subjects. B, Linkage disequilibrium coefficients among *UCP2* and *UCP3* polymorphisms. C, Haplotypes constructed from 10 polymorphisms of *UCP2* and *UCP3* and their frequencies.

Table 3

Analyses of covariance of *UCP* polymorphisms with HDL cholesterol controlling for age and BMI among Korean female subjects

Gene	Locus	C/C	C/R	R/R	Genetic power	<i>P</i>		
						Codominant	Dominant	Recessive
<i>UCP2</i>	-1957G>A	330 (54.16 ± 17.60)	266 (53.05 ± 16.31)	55 (49.75 ± 15.48)	0.350	.198	.177	.121
	-866G>A	186 (54.40 ± 16.40)	313 (54.46 ± 17.54)	159 (49.74 ± 16.02)	0.781	.014	.295	.003
	Ala55Val	192 (54.72 ± 17.07)	306 (54.29 ± 17.23)	157 (49.56 ± 15.94)	0.814	.011	.133	.003
<i>UCP3</i>	+7941_45del>ins	433 (53.48 ± 15.82)	175 (53.71 ± 18.44)	30 (46.07 ± 12.13)	0.549	.104	.490	.033
	-35C>T	327 (54.20 ± 17.48)	175 (52.41 ± 15.51)	53 (52.83 ± 20.76)	0.165	.358	.169	.979
	Int2-143G>C	184 (53.60 ± 15.58)	312 (54.38 ± 17.19)	158 (50.84 ± 18.02)	0.471	.118	.693	.041
	Tyr99Tyr	319 (54.24 ± 17.58)	280 (52.34 ± 15.48)	54 (53.09 ± 20.65)	0.214	.300	.146	.924
	Int3-47A>G	175 (53.66 ± 15.65)	318 (54.39 ± 16.76)	160 (50.79 ± 18.66)	0.493	.108	.715	.038
	Int4-498C>T	402 (54.10 ± 16.81)	223 (52.00 ± 16.33)	31 (52.00 ± 22.75)	0.262	.215	.081	.752
	Tyr210Tyr	178 (54.07 ± 16.17)	310 (54.62 ± 17.05)	166 (49.96 ± 17.43)	0.753	.017	.417	.005
Haplotype	<i>UCP2-UCP3-h1</i>	222 (51.50 ± 17.27)	276 (53.72 ± 15.77)	126 (55.37 ± 16.73)	0.462	.111	.087	.077
	<i>UCP2-UCP3-h2</i>	356 (53.97 ± 17.32)	228 (52.74 ± 15.24)	40 (49.93 ± 16.52)	0.261	.356	.226	.268
	<i>UCP2-UCP3-h3</i>	465 (53.55 ± 16.24)	142 (53.12 ± 17.82)	17 (46.53 ± 12.73)	0.303	.205	.378	.083

Number of subjects (mean ± SD) and *P* values of 3 alternative models (codominant, dominant, and recessive) are shown. *P* values of less than .05 are set in boldface type. C/C, C/R, and R/R represent homozygotes for the common allele, heterozygotes, and homozygotes for the rare allele, respectively.

Table 4

Correction of the significance level according to multiple comparison

Gene	Polymorphisms	Observed <i>P</i> values	Rank	Bonferroni threshold	FDR (BH) threshold	FDR (BL) threshold
<i>UCP2</i>	-866G>A	.003	1	0.0038	0.0038	0.0038
<i>UCP2</i>	Ala55Val	.003	2	0.0038	0.0077	0.0045
<i>UCP3</i>	Tyr210Tyr	.005	3	0.0038	0.0115	0.0054
<i>UCP2</i>	+7941_45del>ins	.033	4	0.0038	0.0154	0.0065
<i>UCP3</i>	Int3-47A>G	.038	5	0.0038	0.0192	0.0080
<i>UCP3</i>	Int2-143G>C	.041	6	0.0038	0.0231	0.0102
Haplotype	<i>UCP2-UCP3-h1</i>	.077	7	0.0038	0.0269	0.0133
Haplotype	<i>UCP2-UCP3-h3</i>	.083	8	0.0038	0.0308	0.0181
<i>UCP2</i>	-1957G>A	.121	9	0.0038	0.0346	0.0260
Haplotype	<i>UCP2-UCP3-h2</i>	.268	10	0.0038	0.0385	0.0406
<i>UCP3</i>	Int4-498C>T	.752	11	0.0038	0.0423	0.05
<i>UCP3</i>	Tyr99Tyr	.924	12	0.0038	0.0462	0.05
<i>UCP3</i>	-35C>T	.979	13	0.0038	0.05	0.05

The list of the observed *P* values of the recessive models is sorted from smallest to largest. *P* values that pass multiple comparison thresholds are set in boldface type. The rightmost 3 columns demonstrate 3 different multiple comparison procedures: Bonferroni procedure, FDR controlling procedures of Benjamini and Hochberg (BH), and FDR of Benjamini and Liu (BL).

Table 5

Analyses of covariance of UCP polymorphisms with atherogenic index controlling for age and BMI among Korean female subjects

Gene	Locus	C/C	C/R	R/R	Genetic power	P		
						Codominant	Dominant	Recessive
<i>UCP2</i>	<i>-1957G>A</i>	328 (3.63 ± 1.28)	266 (3.62 ± 1.34)	54 (3.94 ± 1.29)	0.307	.392	.472	.183
	<i>-866G>A</i>	185 (3.60 ± 1.24)	312 (3.55 ± 1.24)	158 (3.91 ± 1.48)	0.733	.060	.518	.018
	<i>Ala55Val</i>	191 (3.59 ± 1.24)	305 (3.55 ± 1.24)	156 (3.91 ± 1.49)	0.733	.066	.364	.020
	<i>+7941_45del>ins</i>	431 (3.64 ± 1.22)	174 (3.61 ± 1.47)	30 (4.23 ± 1.34)	0.584	.195	.54	.070
<i>UCP3</i>	<i>-35C>T</i>	325 (3.62 ± 1.27)	276 (3.66 ± 1.34)	52 (3.75 ± 1.34)	0.088	.746	.449	.910
	<i>Int2-143G>C</i>	183 (3.65 ± 1.21)	311 (3.57 ± 1.26)	157 (3.82 ± 1.49)	0.395	.225	.823	.093
	<i>Tyr99Tyr</i>	317 (3.62 ± 1.28)	280 (3.66 ± 1.34)	53 (3.74 ± 1.33)	0.084	.747	.453	.935
	<i>Int3-47A>G</i>	174 (3.64 ± 1.21)	317 (3.55 ± 1.25)	159 (3.83 ± 1.49)	0.488	.145	.802	.055
	<i>Int4-498C>T</i>	401 (3.59 ± 1.21)	221 (3.74 ± 1.45)	31 (3.86 ± 1.37)	0.289	.178	.064	.503
	<i>Tyr210Tyr</i>	177 (3.62 ± 1.22)	309 (3.54 ± 1.25)	165 (3.89 ± 1.48)	0.704	.050	.604	.016
Haplotype	<i>UCP2-UCP3-h1</i>	221 (3.78 ± 1.39)	274 (3.62 ± 1.26)	126 (3.52 ± 1.22)	0.372	.206	.214	.100
	<i>UCP2-UCP3-h2</i>	354 (3.63 ± 1.25)	228 (3.65 ± 1.38)	39 (3.92 ± 1.33)	0.199	.679	.576	.414
	<i>UCP2-UCP3-h3</i>	463 (3.64 ± 1.22)	141 (3.65 ± 1.55)	17 (4.20 ± 1.30)	0.320	.214	.445	.083

Number of subjects (mean ± SD) and P values of 3 alternative models (codominant, dominant, and recessive) are shown. P values of less than .05 are set in boldface type. C/C, C/R, and R/R represent homozygotes for the common allele, heterozygotes, and homozygotes for the rare allele, respectively.

감사합니다.

hokim@snu.ac.kr

<http://plaza.snu.ac.kr/~hokim>